Data Misinterpretation: A Consequence of Data Structure?

A Cognitive Imperfection and Its Economic Implications

BACHELOR THESIS WITHIN: *Economics*
NUMBER OF CREDITS: *15 HP*
PROGRAMME OF STUDY: International Economics
AUTHORS: *Joakim Ben David, Balázs Faragó*
JÖNKÖPING, May 2023

Bachelor Thesis in Economics

Title: Data Misinterpretation: A Consequence of Data Structure?

Authors: Balázs Faragó and Joakim Ben David

Tutor: Lina Bjerke

Date: 2023-05-21

Key terms: Data Misinterpretation, Behavioural Economics, Policy Decision-Making, Cognitive Bias, Convex Hull, Agent-Based Computational Economics

---

Abstract

This study examines the claim that individuals misinterpret the mean of a dataset (displayed as a scatterplot) more when the convex hull of the dataset is less representative of the data. In addition, this study also tests whether outliers in the data can predict the magnitude of error that individuals make in interpreting the mean of the dataset. Lastly, the study investigates whether individuals' interpretations are predicted better by the mean of the convex hull than by the full dataset's mean. The method used to conduct these investigations is through a survey, followed by several linear regression analyses. Applications of this study include improving the communication of data in economic policy and business contexts, along with broader applications in extending models that heavily rely on agents' interpretations of information: especially bounded rationality and social norm-based models. The results show that convex hull unrepresentativeness correlates positively with error in mean interpretation; however, that the convex hull mean is not predictive of the interpretations' direction. Overall, the study contributes to the field of visual information interpretation by investigating the effect of data structure on its interpretation – an unexplored area of research. This is done while initiating the concretization of bounded rationality in economics, by exploring the idea that individuals perceive a general shape of the information presented to them rather than a detailed, full picture. This can lead to misinterpretations whenever the general shape (convex hull) is not representative of the dataset.

Table of Contents

1. Introduction 1

2. Literature Review 5

2.1 Background 5

2.2 Theoretical Framework 10

2.2.1 Terms and Concepts 10

2.2.2 Data Interpretation Literature 19

3. Methodology 22

3.1 Research Method 22

3.2 Data Collection: Survey Design 22

4. Analysis, Results and Interpretation 27

4.1 Dummy Variable: Statistics Background 27

4.2 Regressions 29

4.2.1 Regression 1 31

4.2.2 Regression 2 35

4.2.3 Regression 3 36

4.2.4 Regression 4 39

5. Conclusion 41

5.1 Limitations 42

5.2 Suggestions for Future Research 44

6. Bibliography 47

7. Appendices 52

Tables

Table 1…… 24

Table 2…… 30

Table 3…… 32

Table 4…… 32

Table 5…… 35

Table 6…… 37

Table 7…… 40

Figures

Figure 1…… 4 & 14

Figure 2……... 11

Figure 3……... 12

Figure 4……... 15

Figure 5……... 16

Figure 6……... 25

Figure 7……... 28

Figure 8……... 28

Figure 9……... 29

Figure 10…… 34

Figure 11…… 38

Figure 12…… 40

1. Introduction

Data misinterpretation is the primary concept in our study. We consider the general definition of data misinterpretation to be: Gathering an inaccurate understanding of what a dataset shows. But in the context of this study, the more precise definition is: Error in interpreting what the mean of a visual dataset (on a scatter plot) is.

Misinterpreting data is something we all have probably done at some point. After all, it can be very easy to misinterpret data. At first glance, this seems to be a relatively innocent habit. What if we do slightly misunderstand a piece of information? Most often, we are not in the position to make heavy decisions that affect others anyway. Consider this, we have all probably fallen into the rabbit hole of internet medicine when sick and (mis)self-diagnosed with a dozen different conditions: a common but relatively innocent and harmless habit (other than to ourselves in some cases). However, we do have strong expectations of doctors and medical professionals to accurately interpret our health and correctly diagnose us. That is because their interpretation has a direct impact on us and those in our environment. Certain individuals and groups have the power to make decisions that carry an immense level of influence (over groups, the public and society), so their perception and interpretation needs to be as accurate and unbiased as humanly possible. Similarly, it is expected of governments, authorities, and those in leadership positions of companies and organizations to interpret data in a reliable manner and make decisions accordingly. So, how can we ensure that data is interpreted correctly? The first step is in understanding the root of why and how data is misinterpreted. Correspondingly, the purpose of this paper is to investigate whether individuals misinterpret data because of the data's (general) structure. More specifically, do outliers and unrepresentative convex hulls worsen our perception of data?

Data is prevalent across various fields and industries today. This makes the misinterpretation of data an integral issue of our society, with many implications. This paper focuses on the implications of misinterpretation in the field of economics. This topic is highly relevant to economic policy and theory as well as behavioral decision theory, including models related to behavioral finance and business. From an economic standpoint, it is imperative to study and understand data misinterpretation as policy decisions rely on data. We, as a society, depend on policymakers to make decisions based

on an accurate interpretation of the data they are presented with, and that they present to us. Furthermore, from the perspective of a firm, it is extremely important to make the right business decisions when evaluating data. Some of the relevant models and concepts discussed in this paper include P-beauty Contests, the BCG matrix, the ADL Matrix, economic models of social norms, etc.

Furthermore, as economists we are interested in the interpretations of the crowd, as systematic misinterpretation can result in different emergent behavior of the population - than when under perfect information and (more importantly) interpretation of that information. In fact, several use cases of information cascades as well as social norms and their interpretations, exist in economic and finance literature. Social norm and herding reliant models may account for agents interpreting averages and incorporate this information into their behavior (such as the one described by Banerjee (1992)). Overall, this study helps in developing models that include agent heterogeneity, especially as it relates to individuals' information processing capabilities. This is especially helpful for agent-based modelling in economics which can be used for modelling network behavior, herding and similar phenomena; as they may relate to economic crises, certain markets, collective decision making such as voting, systematic risks or clustered volatility in housing markets, and in general, many kinds of emergent behavior (Farmer & Axtell, 2022).

Previous explorations of data interpretation have leaned into external (contextual) factors that may incline individuals to develop a biased perception of data. Often, research has concentrated on biases resulting from the data's context or its formatting, such as how heuristics affect perception when data represents a particular context: for example, individuals interpreting sales differently when displayed across time or space (Hutchinson & Alba, 1997). What has been neglected is whether factors that are innate to data structure itself could lead to misinterpretation. Our purpose was to fill this hole in the current literature by investigating how visualized data misinterpretation may arise from the innate structure of the data, while connecting it to economic implications. We hypothesized that innate characteristics of a dataset's structure might have a fundamental influence on how individuals interpret it. Specifically speaking, data has two innate (interconnected) characteristics that we believe may skew individuals' perception due to how they affect its structure: **Outliers** and **Convex Hull Unrepresentativeness**. We will shortly explain

what data structure is and what these two structural characteristics are, to provide a baseline level of understanding before delving into the more complex side of the paper.

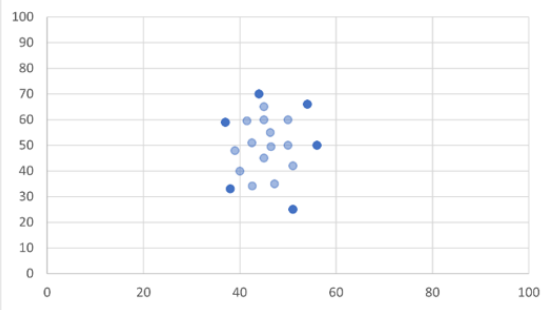
Data misinterpretation has a very broad range of application, but this study investigates it in a very narrow way in order to achieve clear results and a high level of control for certain variables. To do so, we chose to work with visual data on scatter plots. Therefore, in the context of this study, a dataset's structure is the visual representation of the data. The concept used that encapsulates this type of data structure is *convex hulls*. When looking at a dataset on a scatterplot, the convex hull refers to the data points furthest out in all directions that create the "shape" of the dataset. An easy way to imagine the shape is as "an elastic band [that is] stretched around the geometry [of the data]" (Gillies, 2023). To make it clear, the convex hull is the subset of points that touch this imaginary elastic band. Nevertheless, we use the convex hull as a proxy for (what we call) data structure. One could say that in other words, data structure is the "big picture". The idea behind this is that individuals develop an interpretation of data in response to its general structure (shape, outline, or big picture = convex hull) rather than the smaller parts within it that it is made up of (smaller pictures = rest of the datapoints).

Outliers are observations (datapoints) that lie at an abnormal distance from the general clusters of points in a dataset. We use what we call *outlierness* as a measure for outliers and their effect on data structure. The *outlierness* of a dataset is encapsulated by the number of outliers there are and how far (distance) they are from the mean of the dataset. Outlying datapoints affect data structure because they are most often inherently part of the convex hull, strongly influencing the outline/shape of the data.

Convex Hull Representativeness describes whether the convex hull is a reliable representation of the full dataset. The convex hull datapoints visually represent the shape of the dataset, but the rest of the datapoints within it can be distributed in an endless number of ways (the dataset can have clusters, or different levels of density and skewness). In Figure 1 Graph B below, most of the dataset's points are to the left within the convex hull, but the convex hull itself does not reflect that: so, it is unrepresentative of the full dataset. In contrast, Figure 1 Graph A shows that the convex hull has high representativeness since it portrays a more accurate reflection of how the rest of the dataset points are distributed.

Graph A

Representative Convex Hull



Graph B

Unrepresentative Convex Hull

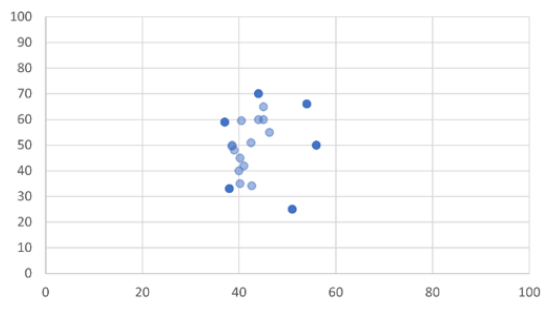


***Figure 1**: Explanation of representative and unrepresentative convex hulls*

On that account, we intend to test outliers and convex hull representativeness (as two innate structural characteristics) in data and determine whether they may bear responsibility in influencing individuals' interpretation. In a survey format, datasets on scatterplots were generated to represent various data structures that encompass the effect of outliers and convex hull unrepresentativeness. Participants of a survey were tasked with guessing where on the datasets they believe the mean of the dataset is – this represented their interpretation of each dataset. Having gathered participants' guesses, regression analyses were carried out to investigate whether outlierness and convex hull unrepresentativeness played a role in their error in interpreting the mean.

To ensure a comprehensive flow, the paper is structured according to the following order. To begin with, a literature review is conducted and separated into two sections. First a background segment to further expound on topic of data misinterpretation in relation to economic policy and business. Then an overview of the theoretical framework from cognitive sciences that support our hypothesis and research method, which also offers a more thorough description of the relevant terms/concepts and their use. Thereafter, the study's methodology is explained, elaborating on the data collection method and design. This is followed by an analysis and results section in which we elucidate our regression models, interpret our results, and discuss them. Finally, we conclude the paper with an anchoring of the results to applications and implications, by acknowledging its limitations, and lastly offering suggestions for future research.

2. Literature Review

2.1 Background

In this section we intend to explain, in detail, the applications and implications of data misinterpretation within economics using real-world examples.

Exploring the implications of data misinterpretation on the macro level quickly directs attention toward the political and economic scene, as they are two segments of a country that carry immense power and influence over society. Data misinterpretation can be described as an event that develops in the process of perception, opinion formation and decision-making. Accordingly, in the context of economic and societal influence, data misinterpretation is a highly relevant factor in policy decisions. To realize the impact potential of data misinterpretation, one needs to grasp its role in economic and political decision making.

Economic decisions and policy applications are plenty. Examples include the use of data by think tanks, international organisations (such as CEF) and consultancy firms that aid governmental decision making regarding the following types of phenomena: monetary, financial, macroeconomic, and fiscal. For example, the European International Organization called CEF (2022) claims to aim central banks (among others) using data. Whenever one stakeholder communicates data to another, the correct interpretation of data - and especially its central tendency - is important. Misinterpretation can lead to policy that is implemented for false reasons and the more channels of communication there are, the higher is the likelihood of error: since individuals are an increasing number of steps removed from the source of the data and receive a more and more general version of the information on which they then base their policy decisions.

Misinterpretation of data (especially averages = means, which are commonly referred to) is a valid concern when it comes to government policy – which includes economic policy. Horsley, Liu and Levenshus (2010) discuss the nature of communication with regards to US government, and one of the key takeaways from this publication - in relation to our study - is that even within government decision making, there are many layers of bureaucracy (not to mention, external stakeholders and the public). Therefore if, for

example, someone presents data in a central part of government, by the time it gets to the local authorities the information is much "distilled", and it is reasonable to think that only the most salient parts of the data are passed on. This is dangerous as it facilitates the spread of misinformation once the data has been misinterpreted – for example because of outliers or convex hull unrepresentativeness. Therefore, it is a valuable effort to aid the correct interpretation of data from the start, as such spirals of communication through decision-making bodies make the implications of misinterpretation even worse. Data misinterpretation resulting from a misperception of outliers' influence on the mean is also an issue that makes instances of the *availability heuristic*[1] worse. Outliers make it into news and public information more than average events, because outliers inherently serve well as shocking news since they are simply more unusual (The Decision Lab, 2023). The attention outliers draw here relates to the kind of misinterpretation that is described in this study, which goes beyond the selection bias of information expressed by availability heuristics. Notably, this can still result in outcomes similar to the availability heuristic - namely a bias from outliers in the interpretation of central tendency. This is highly relevant to the public perception of economic events such as financial market crises or application of certain policies, which is important as public behavior reflects onto the economy. For example, (theoretically) if news is released alluding to a potential financial market crisis - in response to an unusual event (outlier) that does not necessarily indicate a crisis - the public might engage in a bank run, withdrawing all their money from the bank thinking the bank could fall which can then lead to bank failures, liquidity crisis, and general recessionary economic implications (Mooman, 2022).

The (economic) repercussions of the COVID-19 pandemic can be traced to possible cases of data misinterpretation. While the virus spread all over the world, the approach undertaken to handle the infectious disease differed across governing authorities, reflecting the different ideologies and interpretations held in relation to the virus. In many countries, governments imposed eminently restrictive policies to impede the virus's spread (Bucholz, 2020). One can argue however, that the restrictiveness imposed was a(n)

[1] Availability heuristic is a tendency of people to associate memories that they can easily retrieve, to general tendencies in the world. This means that people may think that something is common because they see it in the news a lot, however, the news only broadcasts events that are special or interesting and so, this reliance on memorable information can result in a kind of selection bias (The Decision Lab, 2023b).

(over)reaction unproportional to its statistical danger (in certain countries) (Buchholz, 2020). The global attention and social media virality of COVID-19 most likely played a part in this reaction (Venegas -Vera et al., 2020). Could it also have been an unreasonable (over)weighting of outliers leading to a misinterpretation of the data on COVID-19? The idea is that the outlying severe cases of the virus might have led to individuals perceiving the virus to be more dangerous than it really was. Nevertheless, it is widely recognised that the restrictions imposed in many countries have led to economic losses, as a result of (economic, transitionary) inactivity (money stopped circulating in the economy); economic damage that has been speculated to advance to a recession in the near future (Economist, 2023). Small to medium-sized businesses suffered greatly from short-term economic damage due to Covid-19 (Sarker et.al., 2022). Moreover, Sweden took an unpopular approach in being very minimalistic and flexible with its restriction policies (BBC, 2020). As statistics show, Sweden's mild approach better reflects the true (general, mean) danger of the virus: about as dangerous as the Spanish Flu (Liang et al., 2021). Resultantly, the country's economy suffered significantly less than others (Government Offices of Sweden, 2021; Miltimore, 2020). Would the outcome of the virus's spread be any different if restrictions were less extreme, minimizing the future economic damage? Studying the impact that outliers have on individuals' interpretation of data is vital because understanding their role enables us to avoid misinterpretation and, most importantly, make sure that our and others' actions reflect correct perceptions about the information we are presented and we ourselves present. It is important to acknowledge that in all real-world circumstances such as the ones above, data has context. Therefore, it is inevitable that misinterpretation to some degree must be related to the external factors, as suggested by previous literature (subsection 2.2.2). As previously stated however, we are contributing to the literature by filling a hole in the research of data misinterpretation's origin that we believe is missing, through the investigation of data structure. Our argumentation is thus not that external factors (such as context) do not play a role in data misinterpretation, but that innate characteristics do carry a part of that responsibility.

In the greater scheme of economic and financial theory, the implications of data misinterpretation are relevant in mean-field like models and models in which agents incorporate the average of some public information into their behaviour. In behavioural

finance, there exists a model with several variations called *p-beauty contests* (or rather a general class of models called K-Level models) (Dhami, 2016). P-beauty contests were the discovery of J.M. Keynes. John Duffy and Rosemarie Nagel (1997) summarize his findings: "As Keynes observed, much of professional investment behaviour is concerned with attempting to keep one step ahead of average behaviour. The average investor, or in Keynes' view, 'the ignorant individual left to himself', is unlikely to disappear, as it is precisely the behaviour of this type of investor that market professionals are attempting to forecast" (Duffy & Nagel, 1997). From this, we can take away that investors' general perception of information about the market is essential to understand the nature of said market. Accordingly, we seek to investigate what structural factors in data might alter perceptions. Through its potential implications to these kinds of models, our results may contribute to the fields of behavioural finance, economics, and decision theory (Dhami, 2016). Other, mean-field like problems in economic theory include models of social norms, herding, co-ordination games, peer effects, and public goods provision. As an example, Matthew O. Jackson discusses variations and implications of the *friendship paradox* and other social norm-based network models for political economy which often rely on agents' understanding of a mean of some agents' behaviour, or the behaviour of the collective (Jackson, 2019). Additionally, there are behavioural biases such as the misinterpretation of *regression to the mean*[2] (more generally called "regressive fallacy", a case of post-hoc fallacy), which are only made worse (increased bias) when individuals misinterpret what the mean is; especially since this can be traced back to outliers (Caroll, 2015). Hence, this is highly relevant to our research, as it involves having participants of a survey guessing the mean of datasets.

Data misinterpretation is a common issue when it comes to visual data (Statistics Canada, 2011; The Assessment Network, 2020). Scatterplots are of common use in many business practices and are often interpreted by non-statisticians without direct access to the raw data. This, therefore, may render interpretation to be very surface level. A common example of what is essentially a scatterplot in strategic planning is the - long in use –

[2] Regression to the mean is a statistical phenomenon that as one keeps on adding events, or observations, if there is an outlier observation, it is likely that the next observation will be closer to the general tendency or the mean. People tend to forget about this and that an outlier is an outlier. Therefore, they attach some causational interpretation to the movement of the observations while it is purely a statistical phenomenon (Caroll, 2015).

BCG (Boston Consulting Group) matrix (Madsen, 2017). In general, business and management analytics scenarios often make use of theoretical "matrices" such as the BCG. As long as one can measure the labels on a matrix, these matrices can be made into scatterplots by simply dividing the axes based on the matrix sections. Further examples include GE (General Electric) multifactorial analysis developed by McKinsey & Company. Another example is the ADL (Arthur D. Little) Matrix used for portfolio management (McKinsey & Company, 2008; Strong, 2014). While these matrix-based frameworks offer structure and guidance, simple scatterplots can be used to analyze any variables of interest in a business setting as well. Chengyee Janie Chang and Yan Luo (2019) summarize the previous literature on semantic and visualisation-based biases such as framing or the salience attached to certain visualisations and how this affects reporting, decision making and communication in an auditing environment, specifically for the company KPMG. However, the authors do not discuss the inherent characteristics of data that may make it more difficult to interpret. The closest they come is discussing previous literature on the effects of data preparation – as will also be mentioned in relation to the studies of McNutt and the pipeline regarding visualisation mirages (Chang & Luo, 2019). Nevertheless, the missing part to be investigated by our study is still evident and seemingly relevant for auditing/consulting firms. Furthermore, the increasing use of data in business practices is evident and can potentially be of use to researchers of economics in different fields such as regional economics (Schintler & Chen, 2017). Additionally, it can give insight into population dynamics, and at an increasingly high definition (Schintler & Chen, 2017). Importantly, this data must be interpreted by policymakers, and an understanding of when interpretation is easier or more difficult based on the structure of such data is a useful insight which this study aims to provide. It is also argued by some that understanding events and patterns is difficult since it is on an individual level, and that it is even more difficult in an organization as certain events are given different weights by different people (Mandl, 2019). This is a more general observation that is not necessarily about visual data. However, visualizing data is not necessarily better: especially with the use of big data. Those who present it must choose certain dimensions to present (the variables on the graph axes), which now enforces two sets of biases on the rest of the organization. One: the dimensions considered, and two: now outliers on the dimensions considered may take up a disproportionate amount of attention, while the outliers may give a wrong impression of the dimensions presented. This

essentially means that from one interpreter, the misconceptions about the subject that the data is about may spread throughout the organization. This could happen as the original interpreter is convinced that some significant tendency is true while he might have misperceived this tendency right at the beginning. Essentially, the complexity of any decision-making body such as a company or management board – similarly to the aformentioned example of the government – means that misinterpretation can quickly spread and become misinformation that leads to erroneous conclusions and decision-making.

2.2 Theoretical Framework

Next, it is important to discuss the basis of data misinterpretation literature from which this research is motivated. Since interpretation of data contains influences from perception, an interdisciplinary turn towards the lands of cognitive sciences must be taken - only to eventually return to economic considerations.

2.2.1 Terms and Concepts

We would be remiss not to acknowledge the vagueness or complexity of some of the terms used in this research. In clarification, these are the terms and concepts utilized to carry out the study, what they mean, and why they are used.

In this research we used visual datasets. Visual data can be presented in a variety of forms. We chose to display datasets on scatter plot graphs. As briefly mentioned in the introduction to this paper, we identified and aimed to test two (interconnected) innate structural characteristics that are part of data: **Outliers** and **Convex Hull Representativeness**. Basic definitions of these terms were included in the Introduction section, but in the context of the rest of the paper, more elaborate denotations are necessary.

These characteristics are the two variables influencing a dataset's structure which were hypothesized to play a role in the way individuals interpret the data. As mentioned, the convex hull of a dataset was used as the representation the dataset's structure. The convex hull is a specific subset of points within the entire dataset: To be specific, the exact subset of points that encloses the full dataset. It is easiest to understand convex hulls visually. Consider Figure 2 below. In the figure, there is a dataset with a number of data points:

the points that are farthest out in all directions create a certain shape (in this case, a relatively rigid circle called a polygon). These datapoints create the smallest possible shape (using the data) while keeping the rest of the dataset inside; these points are the convex hull.

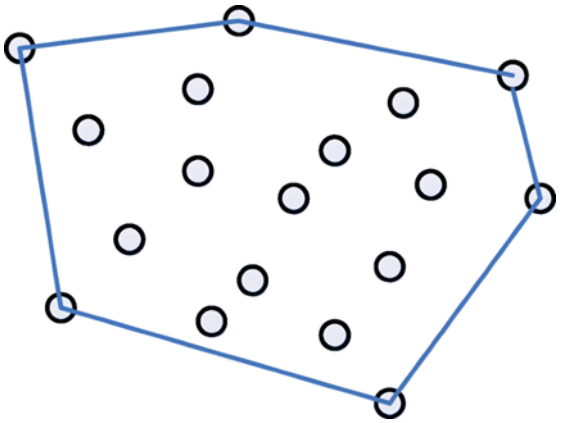


***Figure 2**: Convex Hull of a Dataset*

(Liu et al., 2018)

An easy way to imagine the shape they are responsible for is as "an elastic band [that is] stretched around the geometry [of the data]" (Gillies, 2023). A more precise definition, however, is that the convex hull is the smallest possible polygon – traced by outer data points (furthest outward) - that fully encloses the dataset (Sommer, 2020). It is important to remember that the convex hull is not the shape that these points make, but the subset of points themselves. We used the algorithm called Graham´s Scan[3] to identify the convex hull of datasets.

The convex hull of data was used as a proxy for (what we call) its structure. Relating this back to **Outliers** and **Convex Hull Representativeness**; these two characteristics influence the data's structure via their role in shaping the convex hull.

[3] In our case we used the convex hull algorithm in the Shapely library for Python. This algorithm relies on the more general algorithm called Graham´s scan (JTS, 2020). Briefly, the algorithm relies on taking the lowest value datapoint in the y-dimension, and connecting the datapoint to the datapoint which can be connected with a line which has the smallest angle of inclination that creates a turn of less than 180^0. The algorithm repeats until it connects back to the original starting point (Krishnan, 2020). This is grossly simplified, but gives a good general picture of the functioning of the algorithm.

To discern how they shape the convex hull, it is important to understand them individually and how they are interconnected.

Outliers are observations (datapoints) that lie at an abnormal distance from the general clusters of points in a dataset. Analysts subjectively choose what distances are to be considered abnormal. In this research, we used an algorithm called DBSCAN[4] to detect outliers.

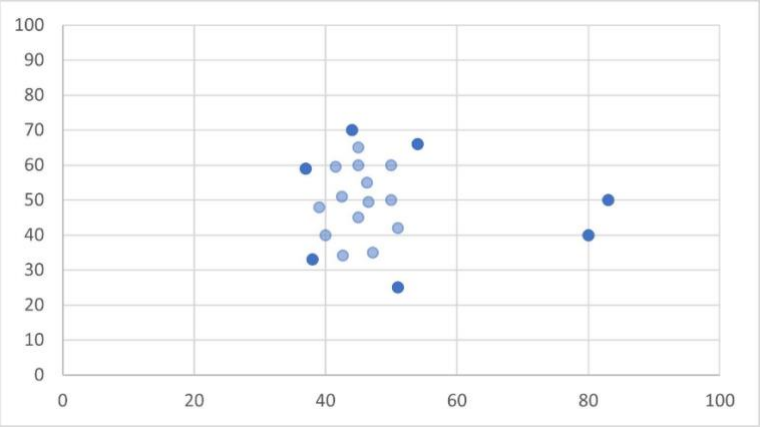


***Figure 3**: Outliers in a Dataset*

In the context of their impact on data structure, it is important to not only attend to the number of outliers a dataset has, but also consider the magnitude of each outlier. This is why, instead of just the number of outliers, the *outlierness* of a dataset was measured.

To repeat, the impact of **Outliers** on data structure (convex hull) is measured using *outlierness*. A dataset's *outlierness* is measured by the number of outliers it has, as well as the outliers' magnitude: the extent of their distance from the mean of the full dataset.

[4] DBSCAN is a clustering algorithm that separates datapoints according to density (Ellis, 2022). It is able to identify where datapoint are clustered and can detect outliers: datapoints that do not belong to any clusters (Ellis, 2022). The algorithm was chosen for its ability to handle irregular data and outliers (Ellis, 2022). We obtain outliers by finding all the points that are not considered part of any of the clusters detected. DBSCAN has two hyperparameters to consider: "epsuse" and "min_values_use" (Valeti, 2021). We restrict the range of meaningful values of these hyperparameters through visual inspection since the elbow method did not produce interpretable results - possibly due to the large heterogeneity in the datasets used, and that we wished to use the same hyperparameters on all our datasets. The set of combinations used is therefore the cartesian product of "epsuse"={2,3,4,5,6} and "min_values_use"={2,3,4,5,6,7,8}.

Additionally, it is important to mention that a measure for distance is required for this. We are using Mahalanobis[5] distance, because it is the distance measure that is most suitable for measuring *outlierness* in our case.

The *outlierness* of a dataset is a value that we calculate. The equation[6] for *outlierness* is:

$$\Omega = \frac{k}{n} d_{mean} + \frac{n-k}{n} d_{max} \tag{1}$$

$$k = number\ of\ outliers$$

$$n = sample\ size, total\ number\ of\ datapoints$$

$$d_{mean} = average\ mahalanobis\ distance\ of\ the\ outliers\ from\ dataset\ mean$$

$$d_{max} = largest\ mahalanobis\ distance\ of\ the\ outliers\ from\ dataset\ mean$$

We apply this formula for each dataset that we use, to find its *outlierness*.

[5] Mahalanobis distance is a distance metric that accounts for the standard deviation within the dataset used. This metric is particularly useful for outlier detection and quantifying the degree of outlying in this study. To illustrate the importance of using Mahalanobis distance: in contrast to the Euclidean distance, Mahalanobis distance can differentiate between data points that are far away in a spread-out dataset and those that are far away in a dataset with low variance. More strictly the definition of Mahalanobis distance is $d^2 = (x_1 - x_0)^T \Lambda (x_1 - x_0)$ where $x_1$is a datapoint, x_0 is where one measures it´s distance from (often the mean) and Λ is the inverse of the covariance matrix of the dataset (Bishop, 2006).

[6] This gives us a weighted average, which prefers the d_{max} when there are less outliers and d_{mean} when there are more. This implementation balances out the weighing of outliers where there may be multiple outliers, outliers in different directions, clusters of outliers, as well as single outliers.

As explained before, **Convex Hull (Un)Representativeness** is a characteristic that shows whether the convex hull of a dataset is a reliable representation of the full dataset. View Figure 1 again below. In Figure 1 Graph B the convex hull is unrepresentative of the full dataset. In Figure 1 Graph A, the convex hull is relatively representativeness of the full dataset since it portrays a more accurate reflection of how the rest of the dataset points are distributed.

Graph A

Representative Convex Hull

Graph B

Unrepresentative Convex Hull

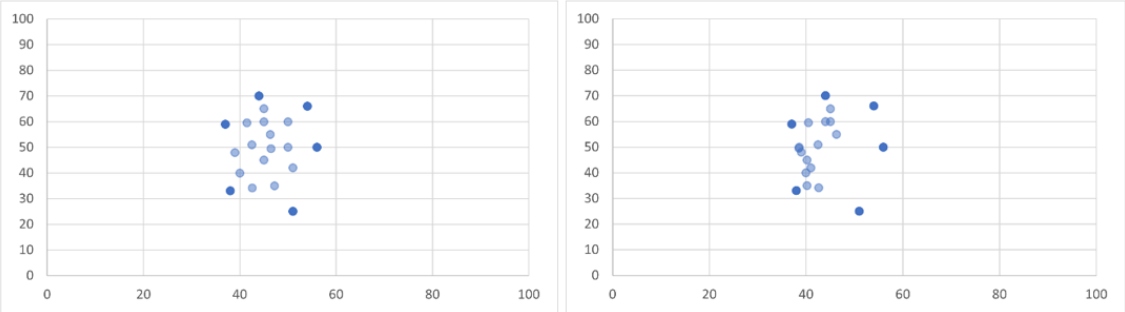


Figure 1: *Explanation of representative and unrepresentative convex hulls*

Now, just as *outlierness* is used as the measure for **Outliers**, a statistical measure is needed to determine **Convex Hull (Un)Representativeness**. For this, we utilize means (averages). The way we measure how representative the convex hull is, is by comparing the convex hull's mean to the full dataset's mean (distance). The further the convex hull mean is from the dataset mean, the more unrepresentative the convex hull is. Furthermore, here is where the two characteristics are interconnected. Convex hulls capture *outlierness* because they describe the outline of the data, which is highly elastic to the presence of outliers. This is developed in the next subsection.

Important to mention, is that Euclidian distance is used here, not Mahalanobis distance.

The Convex Hull and Outliers: Use and Motivation

This subsection is purely intended to provide motivation for the use of convex hulls and outliers in representing data structure, as well as to slightly elaborate on their use. This subsection provides a thorough exploration of the literature that supports the use of these structural characteristics for those who are interested, otherwise it is possible to skip.

To reiterate, the use of convex hulls is an effective way to capture *outlierness* because they describe the outline of the data, which is highly elastic to the presence of outliers that may have a biasing effect. To put it simply: when outliers are in the data, the mean of the convex hull changes more than the mean of the entire dataset. Figure 4 portrays the difference between how much they change (distance wise) in response to the outliers' presence, for this example. As one can see, the convex hull mean changes nearly twice as much as the full dataset mean (twice the distance). Furthermore, the bar chart in Figure 5 explicitly shows this difference in movement.

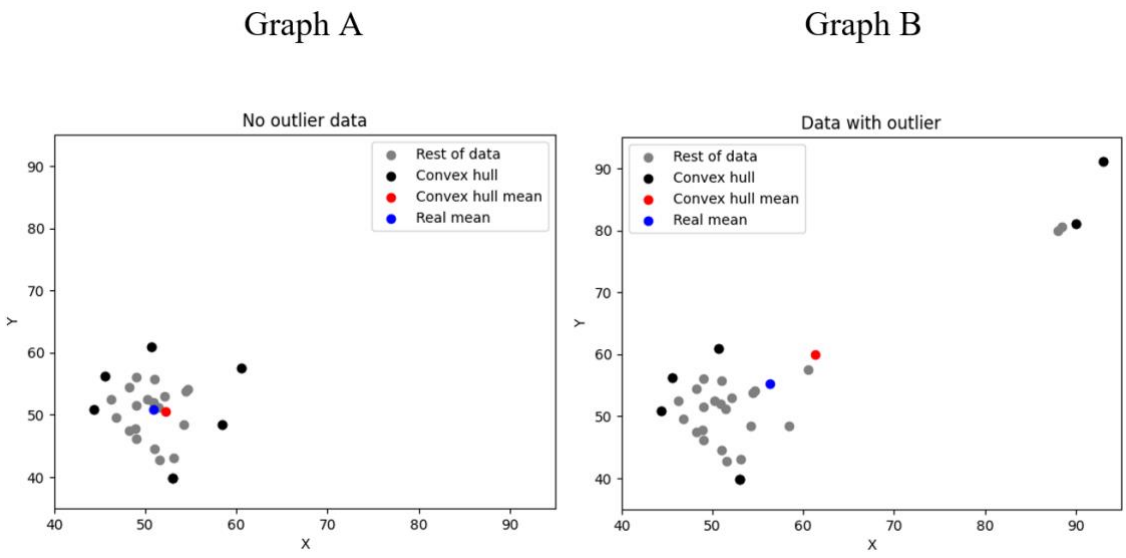


Figure 4: *The effect of outliers on means in a scatterplot.*

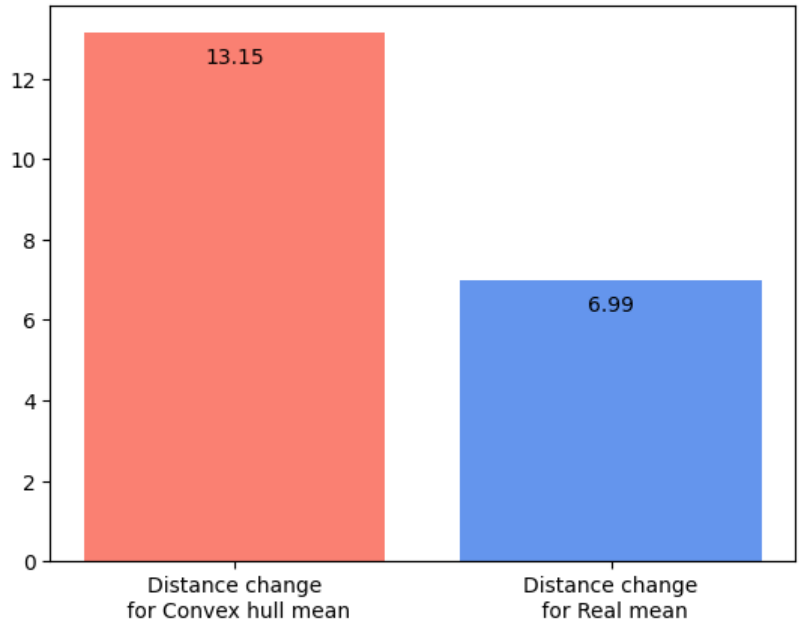


Figure 5: *Distances moved by real mean and convex hull mean as a result of adding outliers in Figure 4.*

The convex hull's elasticity to outliers makes it optimal to use when exploring the effect of data structure on interpretation, because if outliers are to have a biasing effect, the difference between the statistical mean of the convex hull and the full dataset mean will encapsulate it.

To explain how this works in further detail, consider the following. When calculating the mean of the entire dataset, the outliers are a small(er) fraction of the datapoints involved. Whereas when taking the mean of the convex hull vertices (datapoints), they take up a larger fraction of the datapoint involved (since they, most of the time, are inherently part of the convex hull). This can be a mathematical analogy for individuals perceiving datapoints to have an unequal weight distribution; under their perception, outliers hold more weight (value/influence) in comparison to the other datapoints.

But what reasons do we have to believe that individuals perceive outliers to hold more weight than the rest of the dataset's points?

The root of this phenomenon is explained by literature in the fields of cognitive science. Research in these fields provide us with the idea that individuals' primary interpretation is fundamentally influenced by the "big picture" as opposed to the many small pieces that make up an image. This, among other relevant phenomena are discussed subsequently.

Under Gestalt Perception for example, there exists a notion known as priority of wholistic level processing: which means that when looking at visual stimuli that is the sum of smaller parts, we tend to notice the macro phenomenon first (Lamberts & Goldstone, 2006). The idea transfer here, is that in the case of this study, individuals tend to notice the general shape or outline (convex hull) of the data before the particularities within it. Thus, according to theory, the convex hull of the data (its structure) might have a stronger influence on individuals' interpretation than the full dataset with all its values (datapoints). This is important since the data's convex hull can be unrepresentative of its true, full value (what it really shows in totality). This notion bolsters the use of convex hulls as representations of the general shape (outline) that is perceived from a visualized dataset.

Within Gestalt Psychology, the wider, more general "Law of Prägnanz" supports this study's use of convex hulls as it states that we tend to perceive closed figures as being of "minimal area, smooth, continuous maximally symmetric and maximally convex" (Lamberts & Goldstone, 2006). This is a very fitting description of a convex hull, with the note that of course, smoothness and symmetry are not necessarily properties of convex hulls.

Additionally, research from fields such as computational neuroscience also support the idea that people perceive data that is presented to them by its general shape. For example, a study by Billings et. al. (2021) called "Simplicial and Topological Descriptions of Human Brain Dynamics" suggests that the use of topological data analysis in processing fMRI data is useful for this same reason. More specifically, the idea is that the brain stores a simplified general shape (described by a simplex structure) of perceived information. In the context of this study, the convex hull (a, related, structure representing the data) is used to estimate what is perceived by the participants in the study (as a proxy for the general shape of the data portrayed to them). This enforces that this approach is well motivated by neuroscience literature but could also be extended since using the convex hull does not consider the homological[7] properties of the data. This extension is recommended for future research.

[7] Homology is a fundamental topological property of a shape. More specifically, it concerns how many

Furthermore, other notions in cognition outside of Gestalt Psychology literature also support our use of the convex hull. Literature on the visual cognitive system suggests that it aims at metabolic efficiency, and that it therefore does not intake information in full and equal resolution (Lamberts & Goldstone, 2006). An example of this is that when one focuses on visual stimulus, the background becomes blurry. Specifically, this causes us to be more sensitive to low spatial frequency (which means details get blurry). The idea, in the context of this study, is that the details are all the individual datapoints, and that we therefore perceive a blurry or more general image of the visualized data (structure = convex hull) more readily than the details. A similar story is told by literature on perception being the taking of "smooth derivatives[8]" (Lamberts & Goldstone, 2006).

In the groundwork stages of this study, the original, more general objective from which the idea of convex hulls being better predictors of perceived data than the entire dataset was not limited to visual perception. The decision to focus on the visual perception of data (especially via scatterplots) came from its ease of capturing the phenomenon that individuals would assign disproportionate amount of cognitive importance to outliers.

Additionally, this is supported by literature on the Peak-End Rule concept which is often used in behavioural economics. The peak-end rule can be described as a cognitive bias that effects the perception/interpretation of experiences in memory. The concept states that the most intense part of an experience - and how the larger experience ends - forms the memory of the experience the most (Kahneman & Tversky, 2000, p. 673-692). The peak-end rule explains that one's memory of an experience does not correspond to the average level of feelings (positive or negative) throughout the experience, but to its most extreme point(s) as well as its end (Kahneman & Tversky, 2000, p. 693-708). The focus is on the intense part here, drawing the similarity to outliers and extreme values (in how they are perceived with more attention than the other datapoints). To determine whether individuals misinterpret data as a result of its structure, a method is needed to capture

"holes" an object has. This is where the commonly discussed idea of topologists claiming that a mug is the same as a donut comes from. Of course, this is a generalized idea into N-dimensions. In the case of this study, the convex hull of the data is used as a proxy for its general shape and therefore, any holes in the shape of the data is neglected. This is a simplification which perhaps can be avoided in future research on this topic.

[8] In this field of research, several ways to think about visual cognition exist, but they seem to address the same underlying idea- the "smooth derivative" is just another approach.

structure as it pertains to the way individuals observe the data. This means that the concept must take care of both data structure and how it is perceived. That, precisely, is what the notion of the convex hull encapsulates. We support this use case with the cognitive science literature in this theoretical framework section.

Moreover, using the convex hull in addition to *outlierness*, more general claims can be made than what could be made with outliers purely. The cause of increased guess error can be called an unrepresentative convex hull. Of course, this unrepresentativeness may be due to *outlierness* (outliers), but by using the convex hull we are able to establish claims that are robust even if the characteristics of outliers and the magnitude of their distance are not fully covered in our measure: *outlierness*.

2.2.2 Data Interpretation Literature

In existing literature on data misinterpretation, researchers have heavily leaned into studying extrinsic factors regarding how they influence perception, while characteristics that are intrinsic to the data remain unexplored. Why this is the case may be because it could come across counterintuitive to think that characteristics within data itself can lead individuals to misinterpret it. Nonetheless, a plethora of literature has been published concerning biases involved in data interpretation. These studies can generally be sorted into two main categories, according to the source of misinterpretation they seek to understand. Historically, the investigated sources of misinterpretation have been framing (also referred to as semantic context) and visual context (Fogacci, 2020). Hutchinson and Alba (1997) conducted a study to investigate the eyeballing of correlation in a table format. Their study evinces the violation of "semantic invariance": interpreting two sets of data differently due to what the data represents, despite them having identical statistical properties (Hutchinson, Alba, 1997). Although it is in a different visual format than this study, this is consistent with the literature on biases, behavioural economics and cognitive psychology that motivate our use of data structure (as discussed in subsection 2.2.1). Examples of misinterpretation arising from visualization can be labelling, colors, size etc. (McNutt et.at., 2020). Furthermore, there are plenty of books and papers that discuss the idea of "lying with statistics", which focus on the visual representation of statistics. One example is the book "How Charts Lie: Getting Smarter about Visual Information" by Alberto Cairo (2019).

Seemingly, previous literature on data interpretation has lacked consideration for the impact of innate structural factors in data. McNutt et. al. (2020) breaks down the potential sources of "visualization mirages" into a pipeline: curating, wrangling, visualizing, reading. The phrase "visualization mirages" is used to refer to misleading data visualisation and where the origins of such problems lie in the process making them, i.e., the pipeline. Visualizing is a self-evident title, and reading refers to mostly behavioural biases such as the well-known anchoring effect (The Decision Lab, 2023a). Curating regards issues that may arise in sampling or simulated data, and wrangling has to do with encoding data or errors in normalizing, standardizing and data preparation in general (McNutt et.at., 2020). Overall, even this paper, which uses much of the previous literature on visualisation misinterpretation, does not speak of any inherent biasing factor that may arise from the structure of the data itself. The purpose of this paper, again, is to fill this hole in the current literature by investigating the causes of visualized data misinterpretation that arise from the innate structure of data and connect it to economic implications, as opposed to the previously mentioned explanations for misinterpretation: visualization and semantic biases. This would complement all other explorations of biasing factors, as any guidelines or measures provided to medicate bias would have to integrate this paper's results to fully fulfil their purpose.

One might argue that this paper takes a step back in this field since previous literature implicitly implies that it is external, contextual factors which lead to misinterpretation. The issue, however, is that this relies on the assumption that there is no heterogeneity in the datasets themselves lending to differing interpretations of their statistical properties. This assumption was never verified before moving to external effects. Having operated under this assumption, one can also argue that the literature up to today cannot be fully reliable, as any misinterpretation that the literature has linked to external variables could also have been partially provoked by the data's innate characteristics. Therefore, an additional goal of this study is to provide stronger, more thorough fundamental assumptions and basis of knowledge for future research. Resultantly, a full understanding of data interpretation can then promote more effective economic policymaking and more all-encompassing economic theory.

As has now been discussed, there is strong motivation for this paper's hypotheses from cognitive science, neuroscience and psychology literature. Within cognitive sciences, however, there are additional fields of inquiry that provide further basis to investigate data structure: one is the concept of Gist Perception. The concept states that individuals remember a general idea/structure of what was presented to them, not the entirety of that visual information (in our case, the data). For example, the paper titled "Gist Perception Requires Attention" by Arien Mack and Jason Clarke (2012) shows that for this general picture to form, attention is needed. Conveniently, that study also used visual stimuli. Overall, this (that attention affects gist formation, where gist could be the general shape of a dataset like in the case of this study) is decent motivation for the idea that outliers may be important in the formation of gist, given that there is a disproportional amount of attention paid to outliers.

Furthermore, additional research gives reason to believe that people would be paying more attention to outliers or that outliers have a larger cognitive impact. A well-known phenomenon that captures this is what is often referred to as "salience bias"[9] or "outlier effect". The phenomenon was first mentioned by Anne Treisman (1980) in her seminal work on the Feature Integration Theory of Attention, and has been studied further ever since, even by behavioural economists such as Paul Dolan et. al. (2012) who have used the idea in their MINDSPACE framework for choice architecture. A study by Jarvenpaa (1990) also supports that in the context of visually displayed data, parts of visuals with higher salience are perceived as more important by interpreters. This of course has implications in decision making as although it may be the case that more salient parts of the visual, such as outliers, are more important, sometimes the case may be the exact opposite: that the salient parts are anomalies that are not descriptive of for example central tendency (such as the mean). This further motivates the idea that outliers are key in understanding the interpretation of data, which may lead to certain kinds of economic/financial decision making. Shelley Taylor and Susan Friske (1978) also seem to confirm that salience is achieved by either brightness, motion complexity or novelty, or most relevantly, statistical novelty. This paper connects visual interpretation to social

[9] This, is the term for the previously mentioned cognitive impact used in the psychology literature. Some stimuli have a larger cognitive impact than others.

behaviour which motivates us to also follow a similar path by connecting data misinterpretation to economic and social behaviour which couples with works of economists such as Mathew O. Jackson (2019) that deal with social norms, herding, homophily and similar topics. In our case, the connection is made through the incorporation of bounded rationality (as suggested by the results) in the interpretation of averages by agents in these models.

3. Methodology

The methodology section of this paper consists of two main segments. First a short introduction to the research method of the study, then a thorough description of the data collection and its design.

3.1 Research Method

The method used to collect data for this paper was surveys. To investigate the role of data structure in data misinterpretation, information needed to be gathered that describes how individuals interpret data in response to different data structures in a controlled environment. No previous studies were found that had investigated this specific subject, so no data existed to make use of. Hence, to achieve the most relevant results it was crucial to collect new data. To obtain such data, a specific set-up was required. It was necessary to find a medium through which it was possible to expose participants to datasets displayed on scatterplots. In addition, a method to capture and represent the participants' interpretation of the datasets was needed – this will be explained further in the latter part of the Methodology section when describing the survey in greater detail. Nevertheless, it was decided surveys would be the most effective and efficient way to collect such data as they can be quickly distributed and are capable of producing the necessary data. Moreover, contextual variables that needed control could be covered by the design of the survey (also explained later in the Methodology section).

3.2 Data Collection: Survey Design

A copy of the survey is included as Appendix 6. The participants of the survey were 92 students at Jönköping University, from various faculties. This choice of sample was decided upon since similar studies had also used students as participants, such as

Hutchinson and Alba (1997). The survey consisted of 20 scatter plot graphs, each displaying a (unique) dataset. Throughout the survey, three types of data structure were included: Basic, Type 1 and Type 2. These datasets were designed to capture the effects of **Outliers** and **Convex hull (Un)Representativeness** in different ways. Additionally, prior to being exposed to the graphs, each participant was asked whether they have passed a university level statistics course. This was added as a dummy variable to control for whether statistics education would make a difference in interpretation accuracy (size of their guess errors).

Participants' Interpretation

The participants of the survey were asked to provide their interpretation of the datasets by guessing where they think the mean of each dataset is on the graph. They did this by marking where they thought the mean was with a dot or an X. Originally, the survey was planned to be online. However, online survey services that offered a feature which would allow participants to draw on an image were outside of this paper's budget. Ultimately, the survey was made paper based. Surveys we distributed personally, and their results were later manually transferred from paper to computer.

Dependent Variable

Accuracy of Participant's Guess of the Mean – measured by the Euclidian distance between their guess and the true (full dataset) mean.

Independent Variables

Outliers – measured by dataset *outlierness* (**Ω**) using formula .

Convex Hull (Un)Representativeness– measured by the (Euclidian) distance between the convex hull mean and full dataset ("real") mean. (As the reader will see, these are referred to with C and R - for convex hull mean and real mean respectively)

Statistics Background (Dummy Variable)– measured by whether the participant has passed* a university level statistics course. This is captured in a binary, dummy variable.

Later-on, this paper often uses mathematical notation for distances instead of the whole formula. We make use of mahalanobis distance only in the measure of *outlierness* but not in any of the regressions presented later: Therefore, no differentiation in notation for it is necessary. All other distances are Euclidian. A summary of this notation is presented in Table 1 below:

Table 1: Explanation of distances to be used as variables in regression analyses.

Distance variable[10]	Explanation	*Note: All distances here are Euclidian.* *1. The "guessed mean" G is guessed by survey participants.* *2. The "real mean" R is the mean of the entire dataset presented in any given scatterplot.* *3. The "convex hull mean" C is the mean of the datapoints that are in the convex hull of a dataset used for any given scatterplot.*
$d(R,G)$	Distance between real mean (R) and guessed mean (G).	
$d(R,C)$	Distance between real mean (R) and the mean of the convex hull (convex hull mean = C).	
$d(C,G)$	Distance between convex hull mean (C) and guessed mean (G).	

[10] This text called this notation a variable, however it is well known that this notation denotes a function of two variables, namely the two objects which define the distance one measures between. We use this notation just like the more well know notation for a function such as $f(x)$. However, when distance is referred to in this text, it is referring to the output of the distance function. This is done because the input variables are relevant information.

As mentioned before, the survey contained three different main categories of datasets. These are explained further below with the help of Figure 6.

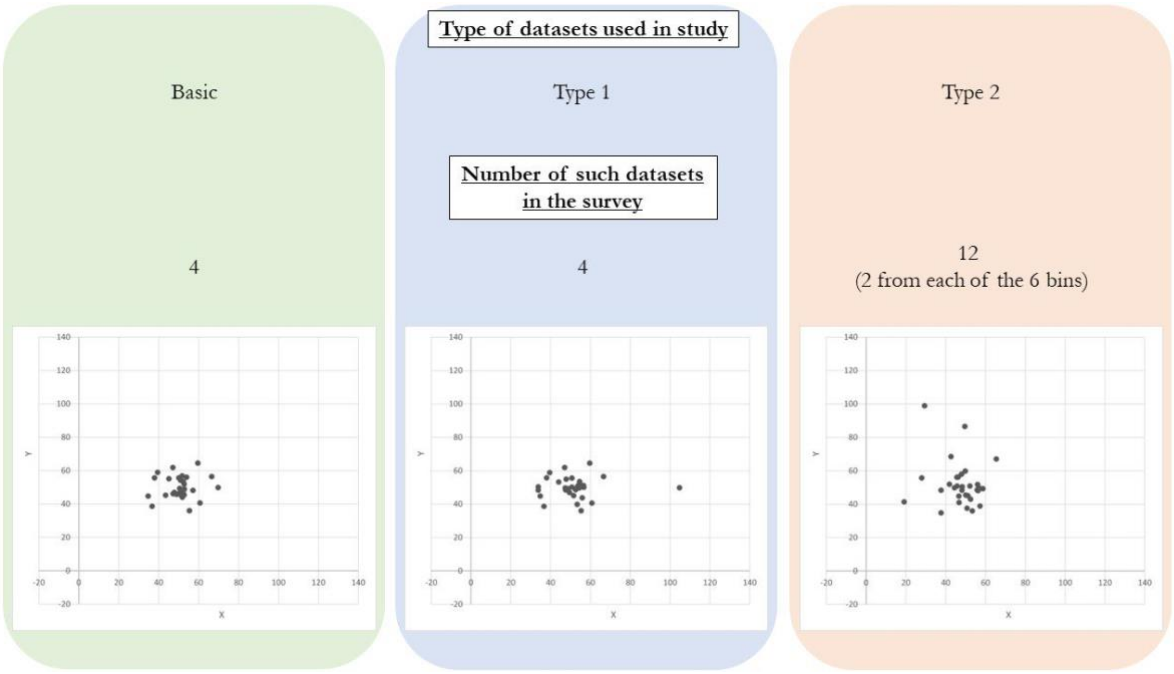


***Figure 6**: Types of datasets used in the study.*

Basic

The basic dataset type speaks for itself. In this structure, the full dataset mean and convex hull mean are set to the same number: the convex hull is highly representative. Furthermore, there were no obvious outliers (although DBSCAN may still have detected certain datapoints as such). This means that the independent variables should theoretically not lead to misinterpretation. Hence, the datasets with the Basic structure are meant to capture a level of baseline error that participants make - that is, if they made errors here, it is expected of them to make larger errors on the other dataset types. There were four graphs of this category in the survey.

Type 1

With these graphs, the primary goal was to test the effect of having an outlier in the dataset in isolation. To generate this structure, a Basic dataset was created, and an outlier was added. After doing so, a compensation was needed to keep the convex hull mean (C) and real mean (R) the way they were without the outlier. The compensation was done by

adding 2-3 extra datapoints. These graphs were intended to make results more robust. They aim to separate the *outlierness* effect from the effect of an unrepresentative convex hull by having an outlier but also having the same means (C & R) as it did without the outlier. Different versions of this structure were created where the outlier is in different directions relative to the main cluster of points. Among them, there was also variation in the distance of the outlier from the main cluster. Altogether, four such datasets were used in the survey.

Type 2

The Type 2 dataset made up the majority of the survey graphs. These datasets were generated randomly but the number of datapoints was the same as in the Basic graphs. This type of structure clearly displayed an unrepresentative convex hull and degree of *outlierness*. After generating a large number of such datasets, a distribution of them emerged with varying levels of distance between R and C. Binning these datasets into six bins according to distance, two datasets were chosen from each bin to make sure there is variation in the graphs used. Overall, there were twelve of these graphs in the survey.

The order of the graphs was randomized so that each participant was exposed to a unique order. This was done to avoid order-bias: when participants learn how to answer questions as they complete the survey. By randomizing the order for each participant, the learning effect is cancelled out in the aggregate. Aside from this, two important factors were controlled for using the survey design: having no contextual or visual formatting effects. The graphs presented had no context or labelling, which excludes any semantic biases. Axes were simply named “X” and “Y” with no meaning attached to them, and there were no titles. This was important to control for as previous studies have shown evidence of semantic bias in interpreting data, and we wished to examine the effect of data structure in isolation from the effects previously studied (Hutchinson & Alba, 1997; Jarvenpaa,1990; Chengyee & Yan, 2021). Secondly, visualisation methods such as the use of different colours, the ways the axes are labelled, and other similar properties of data visualization were previously shown to be associated with biases in the way that data is interpreted (Arunachalam, Pei & Steinbart, 2002; Fogacci, 2020). Thus, all graphs in this study were formatted in the same way with exception to the actual data presented in them. Dataset values were restricted between 140 and -20 for both axes. The additional

and last control was the aforementioned statistics knowledge dummy variable, which was the first question in the survey.

Additionally, it should be mentioned that the convex hull and means (C & R) were never displayed in any way on any of the graphs: the convex hull was not traced in any way.

Again, an example of the survey is attached as Appendix 6.

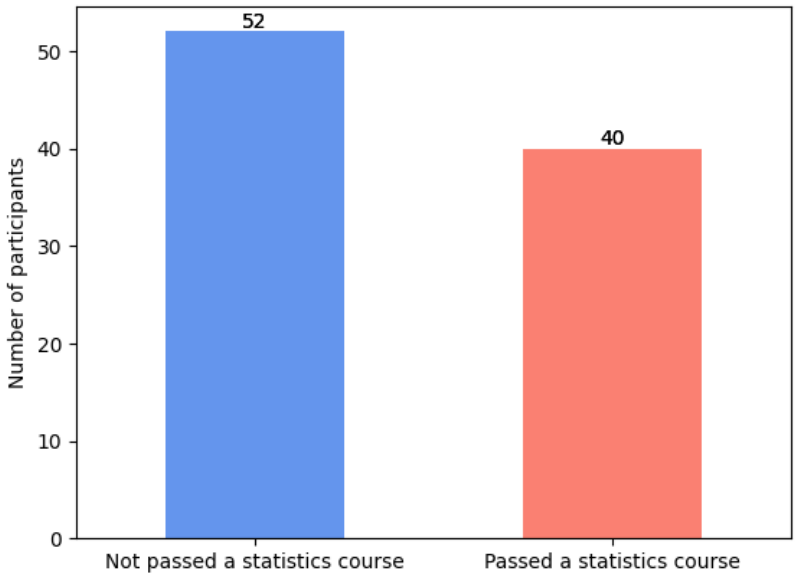
4. Analysis, Results and Interpretation

As described in the Survey Design, participants were tasked with guessing the mean of the datasets. To analyse the data (guesses), three regressions with different purposes of testing each independent variable and a fourth to investigate the relationship between the independent variables were run.

4.1 Dummy Variable: Statistics Background

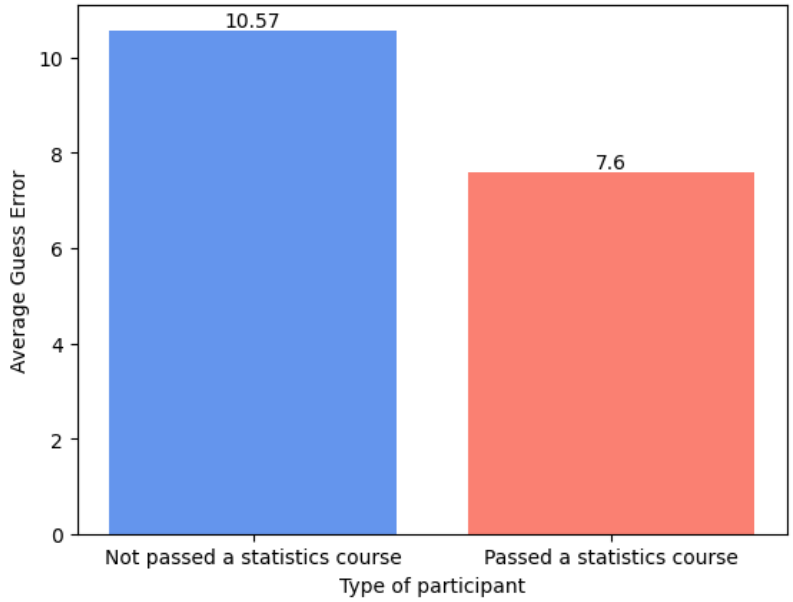
Before discussing the regressions and their results, it may be of interest to cover the qualitative/descriptive statistics on the statistics education variable (dummy variable in the coming regressions). This may provide some interesting results and observations as well as give us a clearer picture of the nature of this dummy variable.

On average, participants who had passed a statistics course made smaller errors (d(R, G)) compared to those who had not. See Figure 8. The number of students that had passed and not passed a statistics course prior to participating in the survey was approximately the same, with 12 more students who had not passed. See Figure 7.



Source: Survey results

Figure 7: *Distribution of participants with and without passing a university level statistics course.*



Source: Survey results

Figure 8*: Average guess errors for participants who have/ have not passed a university level statistics course.*

The full distributions for guess error magnitude between the two participant groups (have passed & have not passed) are displayed below in Figure 9.

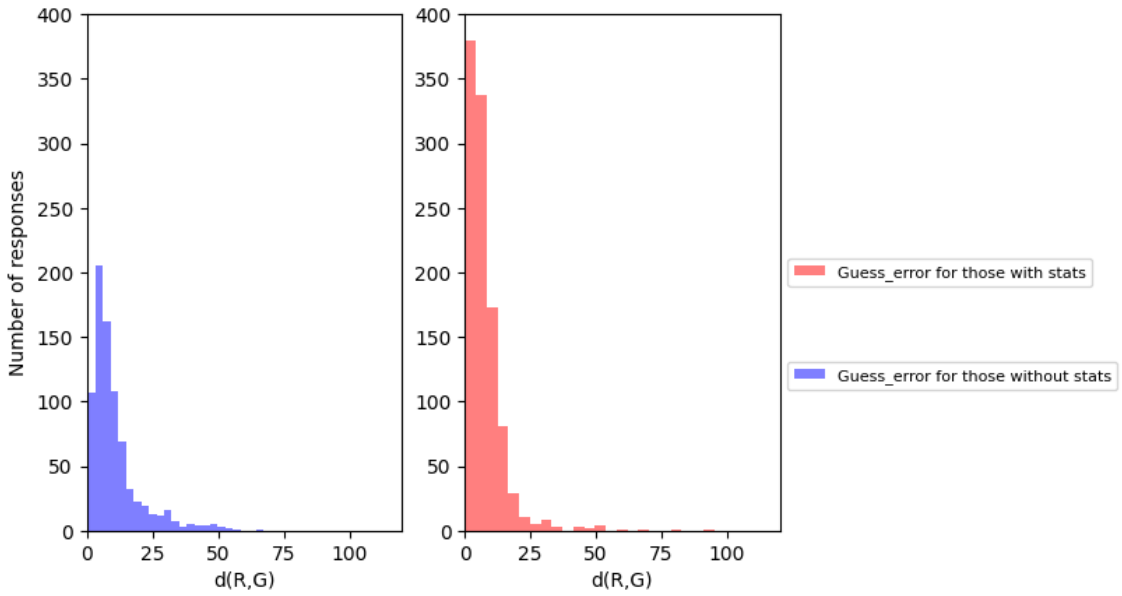


***Figure 9**: Distribution of guess errors for the two types of survey participants.*

Figure 9 shows that, in general, the distributions are skewed towards smaller errors. Additionally, one can see that the very smallest of mistakes were made mostly by those with statistics education: as the peak of the distribution is slightly to the right in values of d(R, G) for those who had not passed a statistics course.

Overall, statistics knowledge seems to matter but as to be seen in the regressions coming, this is not so much the case in combination with other independent variables.

4.2 Regressions

This section interprets the results of the regression analyses as well as the underlying models. In addition, the implications of the results and the conclusions reached from them are discussed. We wish to achieve statistical power of 0.8. 80% is the probability of correctly rejecting the null hypothesis with this criterion. The critical R-Squared is with this condition (Faul et.al., 2009)[11]. This means that for the first three regressions (the ones with the dummy variable), a minimum r-squared of ≈ 0.065 is needed for statistical power.

[11] This R-squared was calculated using the software G Power, given "N" & # of independent variables.

Prior to discussing each regression, another point is important to mention. If only Type 2 datasets were included, the distribution of convex hull representativeness in the graphs (distances between C & R) would be approximately uniform. However, the dataset types Basic and Type 1, which always have a low distance between the convex hull mean (C) and the real mean (R) (higher representativeness) are also included. Therefore, there are more datasets that have a low C & R distance than ones that have a large distance. Essentially, this means that many of the distributions we use as dependent or independent variables are skewed. To combat this, versions of the regressions where we *unskewed* or transformed these variables were also run so that they more closely resemble a normal distribution, which is more fitting to the use of OLS regression analysis. To make this transformation, the skewed variable was raised to the power of 1/k. Values of "k" that minimized skew while keeping "k" an integer (for simplicity) were chosen. The chosen values of "k" for each of the variables used in the coming regression analyses are displayed in Table 2 below.

Table 2: Variables in regression analyses and the constant used to transform them.

Variable to be transformed ("$\boldsymbol{x}$")	Transformation constant **k** where $\boldsymbol{x^{\frac{1}{k}}}$
d(R, C)	2
d(R, G)	10
d(C, G)	5
Outlierness (Ω)	2

Although interpretating the regressions that use transformed variables may be more complicated and less direct, it is not a large limitation in this case because the main concern is about whether the regression function (rearranged to equal the dependent variable) has a mostly positive slope with respect to the independent variable. In other words, we are concerned with whether *outlierness* or convex hull representativeness has a positive or negative relationship with the magnitude of participants' errors, and not so much the functional form. The same is true later, when the attention is on whether the relationship between d(C, G) and d(R, G) is negative and not necessary the exact function which describes the relationship.

4.2.1 Regression 1: Predicting Error Magnitude (of guess, G) by Degree of Convex Hull Unrepresentativeness:

$$d(R,G) = \theta_0 + \theta_1 d(R,C) + D_1 + \varepsilon_1$$

(2)

With the hypothesis:

$$H_0: \theta_1 > 0$$

$$H_1: \theta_1 \leq 0$$

(2.1)

The model underlying Regression 1 is that which is displayed in equation 2 with the hypothesis 2.1. This regression investigates whether larger distances between convex hull mean and real mean (larger d(R, C) = lower representativeness) was positively correlated with a larger guess error (distance between guess and real mean = d(R, G)) from the respondents of the survey. In the first three regressions, the dummy variable D_1 was included. This is the binary variable that indicated whether participants had passed a university level statistics course, where 1 indicated yes and 0 indicated no.

Even for the untransformed version of regression 1, we can see that the estimated coefficient for d(R, C) (i.e. θ_1) is positive and significant. The r-squared value is also sufficient for statistical power as it is 0.082>0.065. This holds for the transformed version of the regression as well with an even better fit (see Appendix 1). The dummy variable for statistics knowledge is not significant at the α=0.05 nor the α=0.1 significance level, for the non-transformed regression. For the transformed version of the regression, it is significant at the α=0.1 level with a slightly negative coefficient. This means that the effect of statistics education is slightly negative on error magnitude if at all significant. This implies a slightly larger error without statistics education. We can state that although different statistics knowledge resulted in different average error sizes as mentioned above, it is not quite significantly predictive of the dependent variable (guess error magnitude). As the reader will see later, this continues to be true for all regressions where the statistics dummy variable was involved. The exact results of the not transformed regression can be found in Table 3 below.

Table 3: OLS Regression results of regression 1 (not transformed)

OLS Regression Results 1						
R-Squared	**Adj.R-Squared**	**F-statistic**	**N**	**Dependent Variable**	**Kurtosis:**	**Skew**
0.082	0.081	81.65	1840	d(R,G)	16.721	2.942
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	6.2241	0.378	16.482	0.000	5.483	6.965
d(C,G)	0.3890	0.031	12.734	0.000	0.329	0.449
STATS	-0.3632	0.400	-0.908	0.364	-1.148	0.421

Source: Survey results

Table 4: Breausch-Pagan Test (BP) results for regression 1

Test	Value (transformed)	Value
Breusch-Pagan (BP) test statistic	≈2.50	≈52.62
BP - p-value	≈0.03	≈0.00

Source: Survey results

The diagnostic test (BP) of the latter regression with transformation, points to no significant presence of heteroscedasticity. This is not the case for the untransformed model, which is one more indication that the transformation was indeed useful (Appendix 1). Since the fit of the transformed version of the model is low but not too low (so although there is a non-linear function and it is not overfitting), it is possible that the functional form of this model is actually to some extent descriptive of participants' perception (Appendix 1&4).

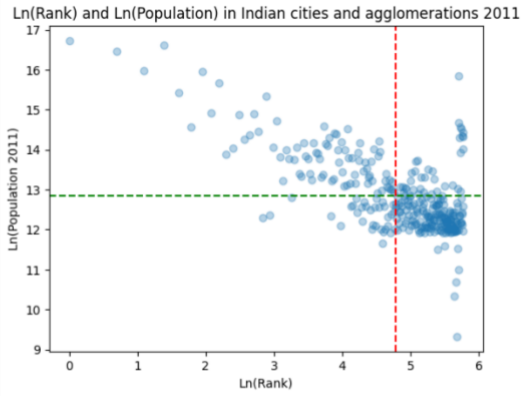
Overall, the results of this study successfully show that individuals find it more difficult to pinpoint the mean of a dataset (displayed on a scatterplot) when the convex hull is unrepresentative (distance between R & C is large), as expected.

In economic theory, and policy as well, the implications are in describing a kind of behaviour that falls under the umbrella of bounded rationality[12]. Models in which agents act on (some) mean can now incorporate the finding that agents are less accurate in doing so when the distance between the mean and the convex hull mean increases (unrepresentativeness of the convex hull increases). Although this study is focused on visual interpretation, it can be argued that this is as applicable, if not more, to non-visual cases of data interpretation because, ideally, visuals only make information easier to interpret. In terms of policy however, more direct use cases can also be found. As mentioned in earlier sections, publicly displayed information is vital in many cases, and those presenting such information should know that certain kinds of data, naturally, make it more difficult to interpret a measure of central tendency such as the mean. If not medicated, this would mean that there could be more heterogeneity in the understanding of some information than of others (by the public), not accounting for what the information is about or how it is displayed. Some practical implications could be advising to display the mean in some non-intrusive way whenever the presenter of data is concerned about possible misinterpretation. This could be done in some similar way to what is displayed in Figure 10 below. Here, the dotted lines are the means on the two dimensions displayed, and their intersection corresponds to what we have previously referred to as the real mean (full dataset mean, R).

[12] "The term 'bounded rationality' is used to designate rational choice that takes into account the cognitive limitations of the decision-maker — limitations of both knowledge and computational capacity" (Simon, 1990). This term was popularized by Herbert A. Simon, when investigating the nature of economic models and decision making of agents that have limited memory or information processing capabilities.

Example

This is displaying means on an example scatterplot via dotted lines. The data is an unrelated example to the study.



Data source: (Brinkhoff, 2011)

***Figure 10**: Example of displaying the mean on a scatterplot.*

Implications on visualisation such as what was mentioned above can also influence business performance if implemented in business decision making contexts. Presenters of information should be aware of the difficulty to interpret data with the structural characteristic of an unrepresentative convex hull and pay extra attention to fully and correctly conveying such information. This could be true not only within company, but to external stakeholders and customers. Overall, within a business context, examples of implications can be found in strategic planning, market research, and/or performance evaluation.

Additionally, in cases when researchers have participants of a study interpret data on scatterplots, they must be careful with individuals' interpretation of central tendency and how that may differ based on the structural nature of the data. Clearly, however, when such surveys are designed, the researcher must also be aware of the other (external) factors from previous research such as certain kinds of semantic bias and visualisation-based factors leading to misinterpretation.

4.2.2 Regression 2: Predicting Error Magnitude with *Outlierness*

$$d(R, G) = \omega_0 + \omega_1 \Omega + D_1 + \varepsilon_2 \quad (3)$$

With the hypothesis:

$$H_0: \omega_1 > 0$$
$$H_1: \omega_1 \leq 0 \quad (3.1)$$

Next is regression 2 which investigates the effect of the *outlierness* measure (Ω) on the size of the error respondents made (distance between guess and real mean = d(R, G)). The results of this regression - using non-transformed variables - showed a positive and significant relationship. However, the R-squared value was not large enough to consider this result statistically powerful. More precisely, we are less than 80% sure that we can make this conclusion: that Ω is positively correlated with d(R, G). Transforming the variables resulted in a slightly better fit but not enough to achieve the sought-after statistical power (see Appendix 2).

Table 5: OLS Regression results of regression 2 (not transformed)

OLS Regression Results 2						
R-Squared	**Adj.R-Squared**	**F-statistic**	**N**	**Dependent Variable**	**Kurtosis:**	**Skew**
0.021	0.020	19.83	1840	d(R,G)	17.866	3.137
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	7.674	0.396	19.220	0.000	6.831	8.384
d(C,G)	6.4939	1.045	6.213	0.000	4.444	8.544
STATS	-0.2773	0.414	-0.670	0.503	-1.089	0.534

Source: Survey results

4.2.3 Regression 3: Predicting the Direction of Guesses, i.e., the negative correlation between d(R, G) & d(C, G) - distance between guess and real mean & guess and convex hull mean:

$$d(R,G) = \beta_0 + \beta_1 d(C,G) + D_1 + \varepsilon_3 \quad (4)$$

And the following is expected for heteroscedasticity:

$$|\varepsilon_3| = \alpha_0 + \alpha_1 \mathrm{d(R,C)} + \mathrm{D_1} + \varepsilon_4 \quad (4.1)$$

With the hypothesis:

$$H_0\colon\ \beta_1,\ \alpha_1 < 0$$
$$\mathrm{H_1}\colon \beta_1,\ \alpha_1 \geq 0 \quad (4.2)$$

The last of the three main regression analyses addresses the direction of the error individuals made. We expected that d(R, G) - the distance from the guess to the real mean – would be negatively correlated with d(C, G) – the distance from the guess to the convex hull mean. This would mean that individuals are biased towards the direction of the convex hull mean. This would also mean that individuals' guesses are better predicted by the convex hull mean (data structure) than the real mean (full dataset). The results of this regression did not support that hypothesis, however. The coefficient of d(C, G) (distance of guess to convex hull mean or β_1) is positively correlated with the error magnitude and this is statistically significant at α=0.05. This indicates that unlike expected, the guesses were not predicted by the convex hull mean: i.e., the direction of guess errors was not necessarily in the direction towards which the convex hull mean was displaced from the real mean. Nevertheless, the R-squared value is too low for these results to have statistical power. With the transformed variables, the fit is even weaker (see Appendix 3).

Table 6: OLS Regression results of regression 3 (not transformed)

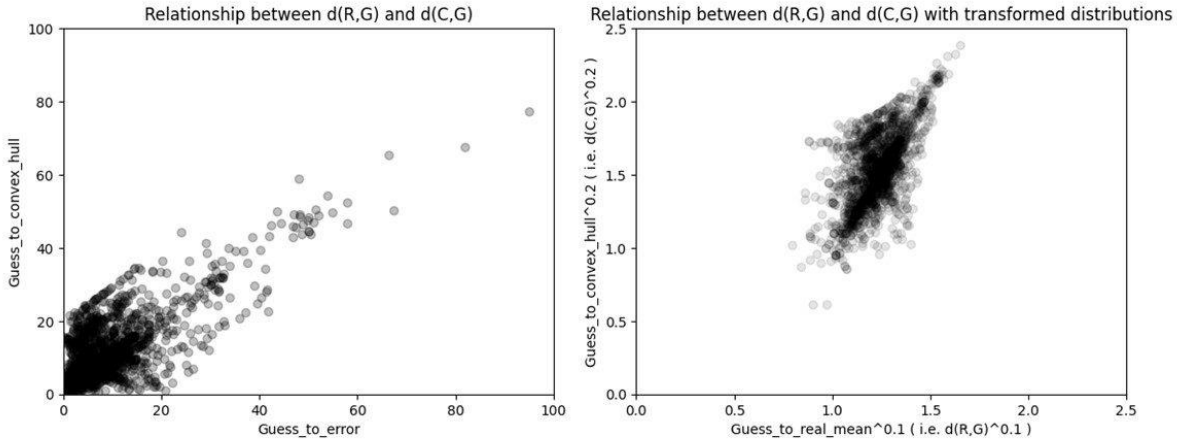
OLS Regression Results 3						
R-Squared	**Adj.R-Squared**	**F-statistic**	**N**	**Dependent Variable**	**Kurtosis:**	**Skew**
0.556	0.555	1149	1840	d(R,G)	5.902	0.708
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	0.7989	0.272	2.937	0.003	0.265	1.332
d(C,G)	0.7282	0.015	47.912	0.000	0.698	0.758
STATS	0.1157	0.278	0.415	0.678	-0.430	0.662

Source: Survey results

Additionally, it was expected that heteroscedasticity follows a certain pattern, displayed in equation 4.1. This would mean that the model of equation 4 fits the data better when d(R, C) is high (high unrepresentativeness). This is because it was expected that convex hull unrepresentativeness is visually noticeable, resulting in more bias towards the convex hull mean (although, of course, participants did not know what this is) whenever the distance between the real and convex hull means are large (high unrepresentativeness). However, this expectation about the model's heteroscedasticity was never investigated since the results showed that the relationship between d(R, G) and d(C, G) was not negative (as hypothesized) in the first place.

The conclusion from this regression and its transformed version is that we cannot tell whether the directions of individuals' errors are as expected with any statistical power. This means that its results do not confirm the hypothesis that the guessed means (i.e., individuals' perception of the mean) are better predicted by the convex hull mean than the mean of the full dataset. The takeaway from this regression is that when individuals make errors in interpreting the mean, the errors do not necessarily happen in the expected direction. In fact, the non-powerful results point in the other direction. Also, if one inspects a scatterplot of the relationship without the dummy variable (Figure 11), it also shows no negative relationship. At most, one can notice a slight positive relationship. The possibly positive relationship exists because if the bias is not necessarily in the direction that was expected, then any guess that is further away from the real mean is also likely to

be further away from the convex hull. This indicates that participants made errors but in no specific direction in any consistent way. This slight positive relationship could mean that hypothesis 4.2 was not accepted because d(R, C) was simply not large enough.



Source: Survey results

***Figure 10**: Relationship between d(R, G) & d(C, G) - distance between real mean and guess and distance between convex hull mean and guess.*

If changes are made to theoretic models in economics, any consistent bias in mean-interpretation by economic agents cannot, according to our results, be built-in. However, it can still be considered that agents become worse at interpreting the means of datasets whenever the data structure (convex hull) representativeness decreases. Similarly, in policy, the application of these results means that presenters of data cannot predict any bias just based on the characteristics of data structure that we have considered. However, given the results of the first regression and its transformed version, they can expect that some data will be more difficult to interpret central tendency such as the mean for.

4.2.4 Regression 4 - Exploring the Relationship Between Convex Hull Unrepresentativeness and *Outlierness*:

$$d(R, C) = \gamma_0 + \gamma_1 \Omega + +\varepsilon_5 \quad (5)$$

With the hypothesis:

$$H_0: \gamma_1 > 0$$
$$H_1: \gamma_1 \leq 0 \quad (5.1)$$

Finally, a fourth regression was conducted that does not incorporate any survey data. This regression was simply to test whether *outlierness* **Ω** is positively related to convex hull unrepresentativeness (d(R, C)). This would confirm the claim that outliers affect the convex hull mean more than the mean of the full dataset, and that this is also true for the data used in the survey. The regression was conducted on more datasets than used in the survey (the entire distribution from which datasets were picked uniformly based on their d(R, C)). This is simply because it is more reliable than a regression with only 20 observations (the number of datasets used in the survey). The regression confirms the positive relationship, as the coefficient of **Ω** is positive. This is even more clear when looking at it displayed on graphs (see Figure 12). The phenomenon seems to also be mostly true for the datasets used (Appendix 5). Perhaps, it can be argued that if datasets that had a stronger relationship between these variables were chosen, the regression using **Ω** (*outlierness*) as a predictor of d(R, C) (convex hull unrepresentativeness) would have a better fit. Nevertheless, this is not necessarily a limitation of the study as it is valuable to be able to highlight the difference in the predictive power of **Ω** and d(R, C) in predicting d(R, G) (the error magnitude, i.e., misinterpretation). This can mean that either…

1. The convex hull is a better predictor of error magnitude than **Ω** because it includes some additional information about the dataset's structure which is relevant in predicting the size of errors individuals make in guessing the mean.

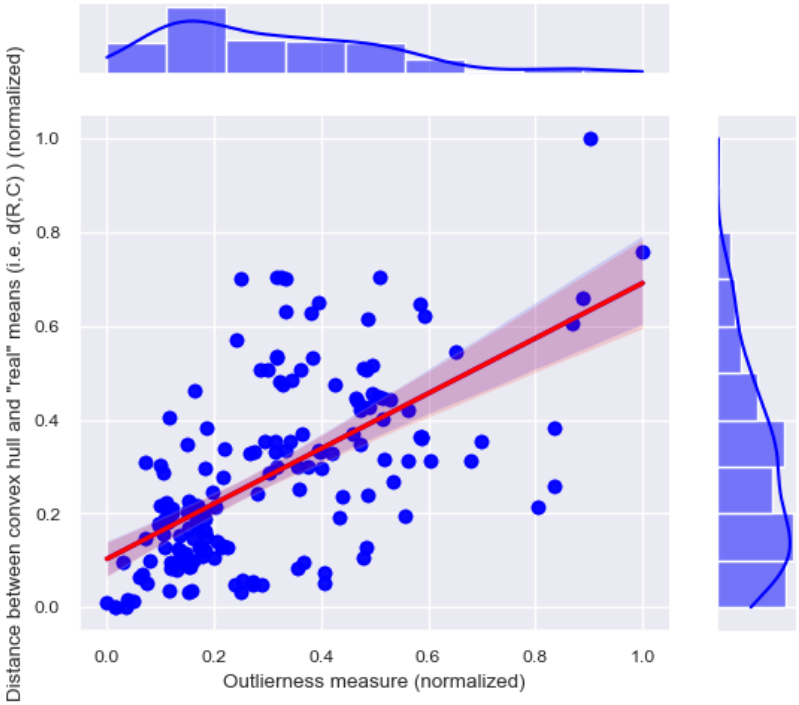
   …or…

2. A better measure of the effect of outliers on data structure is needed than *outlierness* (Ω) to have sufficient predictive power.

Table 7: OLS Regression results of regression 4 (not transformed)

OLS Regression Results 4						
R-Squared	**Adj.R-Squared**	**F-statistic**	**N**	**Dependent Variable**	**Covariance Type**	**Skew**
0.379	0.375	78.44	164	d(R,C)	HC3	0.567
	Coefficients	**Std. Err.**	**z**	**P>|t|**	**0.025**	**0.975**
Constant	0.1032	0.020	5.168	0.000	0.064	0.142
Ω	0.567	0.066	8.857	0.000	0.458	0.719

Source: Survey results



Source: Generated datasets

***Figure 11**: Difference in real means and convex hull means with varying levels of outlierness.*

5. Conclusion

This paper aimed to develop a greater understanding of what leads to individuals misinterpreting data by testing the effect that the structure of the data itself might have on perception. Going into the study, it was expected that with less representative convex hulls there would be more misinterpretation, and that the guesses in the survey would be closer to the convex hull mean than the real mean. To test this on the conducted survey's results, several regression analyses were run. Regressions 1 & 2 had the respective hypotheses that convex hull unrepresentativeness and *outlierness* would positively correlate with the error of guesses d(R, G). Regression 3 tested the hypothesis that the distance between the real mean and the guess d(R, G) would be negatively correlated with the distance between the guess and the convex hull mean d(C, G).

The study has shown evidence of an increasing error in the interpretation of means in datasets when the convex hull representativeness of said datasets decreases. Concurrently, there was no conclusive evidence of the convex hull means being a better predictor of the interpreted means than of the real means. Furthermore, having passed a university level statistics course was associated with less misinterpretation; however, this effect was not significant when coupled with the structural properties of the dataset (*outlierness* and convex hull unrepresentativeness), which carry the focus of this study as independent variables. As discussed earlier, the results imply that economic models which have agents interpreting some kind of average tendency can now incorporate the extent to which they can do that based on the structure of the information being interpreted. However, unlike expected, the results cannot contribute to such models by predicting a certain direction of the bias that individuals make in their interpretations. This study relates to policy implementation that is reliant on visual information. More broadly, this relates to economic theories in which information interpretation plays a large role, and especially whenever a central tendency is an important aspect of the interpretation. Essentially, this implies that the results of this study can find relevance in applications related to certain bounded rationality models by explaining how bounded the rationality of the economic agents are, given different levels of convex hull representativeness of the information they perceive. In policy implications, the results give insight into when it may be difficult to communicate information to the public. In these situations, the presenters of information must make sure to provide extra (perhaps visual) aid in making the averages clear when

they are relevant. Similar use cases of this study may be applied in the business sector, where the results can lead to better communication and hence decision-making within firms. A deeper understanding of how agents interpret information and averages can provide additional detail into a growing field of economics called agent-based computational economics. This field of economics differs in methodology from others as it aims to simulate economic events and environments, and can do that with agent-level detail – meaning it can incorporate bounded rationality, strategic and stochastic behaviour on the micro level (Farmer & Axtell, 2022). It has been discussed that such models can help with understanding the spread of financial crises and in general, the reaction of an economic environment to some event (Buchanan, 2009). Herding phenomena are also common in an economic environment: since individuals conform to norms, they may sway towards what is average or majority behaviour (Banerjee, 1992). Clearly, for this they need some interpretation of what average is. Essentially, broadly applying the results of this study in this framework addresses the difficulty of perceiving true social norms depending on how representative the convex hull is of the true mean or social norm. There have also been applications of herding models in finance especially, as herd behaviour has been blamed for facilitating financial crises (Sharma, & Bikhchandani, 2001).

5.1 Limitations

A primary limitation of the study is the 92 university participants used. Firstly, 92 may not be a large enough number to generalize to large populations such as the entire stock market. Additionally, perhaps university students differ in some unexplored way from the general population.

One limitation of this study is that we may not have fully covered the effects of clustering. Although our measure of *outlierness* attempts to encapsulate outlying clusters as well as single datapoints, if the clusters are of similar size and especially if there are more than two clusters, it is difficult to even define which of the clusters should be called an outlier, if any. A solution to this could be having outliers for each cluster within the dataset, if the clusters are large and dense enough to be considered true clusters (determined using DBSCAN). However, this method may too present challenges as the threshold at which something is considered a cluster is very sensitive (Ellis, 2022), and the set of datasets used in the survey had a lot of variation. Ideally, a more extensive and quantitative

hyperparameter optimization routine, more advanced than the visual inspection, could have been used.

Another potential limitation is how statistics education was quantified in the study: which was done using a simple yes or no question. This does not fully encapsulate statistical competence, as the students that had passed may have passed with different grades, after any number of tries, just one or multiple courses, under different course structures, at different times in the past, and could remember less or more of their curriculum. Alternatively, a small test of possibly relevant statistics knowledge or their grades weighed by the time since they had passed the course could have been used as a more representative quantification of their knowledge. On the other hand, one can argue that the used alternative gives a better light to the relevance of statistics courses in the subject at hand, since if statistics knowledge was relevant, it may be the case that it also could have been acquired outside of "traditional" higher education. For example, participants could have been asked if they have statistics-relevant work experience or if they have ever self-studied statistics, and how advanced these studies were.

Another limitation of the study is format. Since this study aimed at eliminating the effect of visualisation, thereby considering only structural effects, there was a limit to the size of the datapoints used, as too large variation may have come with difficult-to-read graphs due to too much overlapping of points. We wished to avoid this as it would mix in too much of the visualisation effect. This may have led to the survey being restricted to only small effects of data structure (i.e., *outlierness* & convex hull representativeness) being detected, when the effects could be more obvious with some even larger levels of *outlierness* for example. In hindsight, what could have been done is an effort to encapsulate visualisation effects quantitatively based on previous literature as well as techniques developed for this purpose, and to simply account for them statistically.

Another limitation may be that, although we believed we had (either) removed or accounted for uninvestigated but significant factors through the survey design or dummy variable, this may not have been the case. The relevant factors that we had accounted for were data context, visual representation of the data, and the knowledge of participants in statistics as measured by whether they had passed a university level statistics course. However, others may argue that other characteristics of the participants may have been

influential to the results. Factors such as age or gender for example. On the other hand, there are other variables we could have accounted for such as time of day the result had been submitted, and/or day of the week it was submitted on etc… We simply did not control for most of these variables since we did not think they would be significantly influential to the results. We also seem not to be alone in having thought this, as similar studies such as that of Hutchinson and Alba (1997) also only controlled for similar variables (for example, order of questions and visualisation factors when they investigated semantic context). Still, it is arguable that they should have been controlled for. Some personal characteristics may have implicitly been included by design or the dummy variable such as age, since after all one can expect that university students are of a certain age (mostly 17 and above). On the other hand, there are some variables that we thought could perhaps be more relevant but were not possible to measure or control for, such as the time it took to answer the questions: this could have measured the mental effort exerted by the survey participants.

5.2 Suggestions for Future Research

To consider a broader interpretation of our hypotheses, one of the main motivations behind this research was to spark a line of interest in the field of behavioural economics and neuroeconomics. This paper's topic can be researched farther with a heavier emphasis on the general, neural representation of information. To be more specific, it would lend beneficial to explore the topic outside of visual information, using neuroscientific methods for example. This would continuously have many applications in and outside economics. It is applicable in any mean-field problem where the agents have public information of the decision-making of the population and take this into consideration. There are plenty of examples in economics. As mentioned previously, one could say that the financial market is one of these examples, but some models of competition in markets within the field of industrial organisation may be affected as well to account for this behavioural phenomenon. Another question that may arise, even if this hypothesis is confirmed, is whether it scales, when it scales, and what the consequences are for it when scaling? This is true for many topics in behavioural economics. In other words, given that there are these tendencies for "strange" behavioural phenomena on the individual level, do they remain on average or on aggregate?

However, before considering such a large picture, it would be ideal to test the robustness of our findings for different visualisations and environments. Additionally, as mentioned in the limitations section, it could be reasonable to check the robustness of the structural effects' magnitude when combined with other biasing factors statistically accounted for, instead of excluded totally. In general, now that this phenomenon specific to data structure has been examined in an isolated fashion, the reasonable next step would be to examine whether it holds when combined with the other (external) factors discussed by the previous literature on data interpretation and related cognitive biases. This would essentially conclude research on data visualisation which can aid economic policy by providing an even more all-encompassing framework for how to display information to stakeholders and the public.

Going further, interpretation can be tested-for more broadly in how it is measured. One could consider testing interpretations of variance or skewness for example, in addition to the mean. This was considered here as well; however, the idea was abandoned to make the research more feasible given the time restriction. One of the interesting extra hypotheses that could be tested is whether individuals think of standard deviation as mean absolute deviation (and wrongly so). This would have wider applications in economics. The current investigation only provides results regarding the mean. Although this is a commonly used property of datasets, if the convex hull is a satisfactory proxy of interpreted structure in other ways too, we can define a broad and widely applicable rule of thumb on the information processing capabilities of economic agents described under bounded rationality i.e., we can talk about what the bounds are more widely.

A subtle but perhaps useful change one could make to this study is to use another algorithm for the definition of the convex hull in the datasets use. In most cases this would not make any difference for the outcome. However, there has been research that developed convex hull finding algorithms which aim for similarity to human visual perception (Liu et al., 2018). This kind of algorithm (called VIACH in Liu et al.(2018)) would be interesting to use for our case as the research context involves human visual perception. Additionally, it may be the case that although the regression addressing direction of errors was unable to confirm or deny the respective hypothesis with any statistical power, more sophisticated methods may be able to notice relevant patterns in the participants' responses. Here, some cluster analysis of the survey responses may be

interesting. This cluster analysis would be even more relevant if other characteristics of the datasets could be included as predictors and if perhaps other inputs are given by respondents such as skew or correlation.

Lastly, perhaps a market experiment could be done, where the flow of information (for instance about prices) is controlled, and the experimenters could check if the behavior of the participants is as if they acted based on all data provided to them or the just convex hull. The result of such an experiment could tell us about the nature of information use in markets and could have major implications in economic theory and policy. Other experiments of this nature have too been able to inform us about real markets in the past, such as the ones conducted by Vernon L Smith in the 1950s-70s which have become widely used ever since (The Nobel Foundation, 2002).

6. Bibliography

Arunachalam, Pei, B. K., & Steinbart, P. J. (2002). Impression Management with Graphs: Effects on Choices. *The Journal of Information Systems*, 16(2), 183–202. https://doi.org/10.2308/jis.2002.16.2.183

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

BBC. (2020, August 5). Coronavirus: Sweden's economy hit less hard by pandemic. *BBC News*. https://www.bbc.com/news/business-53664354

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics. *Wiley Series in Probability and Statistics*., Chapter 3. https://doi.org/10.1002/0471725153

Billings, J., Saggar, M., Hlinka, J., Keilholz, S., & Petri, G. (2021). Simplicial and topological descriptions of human brain dynamics. *Network Neuroscience*, 1–20. https://doi.org/10.1162/netn_a_00190

Brinkhoff, T. (2011). *Asia*. Population Statistics, Maps, Charts, Weather and Web Information. http://www.citypopulation.de/en/india/cities/

Bishop, C. M. (2006). Pattern recognition and machine learning (1st ed.). *Springer Science+Business Media, LLC*.

Buchanan, M. (2009). Meltdown modelling. *Nature*, *460*(7256), 680-682. http://proxy.library.ju.se/login?url=https://www.proquest.com/scholarly-journals/meltdown-modelling/docview/204550690/se-2

Buchholz, K. (June 18, 2020). Strict or Lenient? COVID-19 Lockdowns Compared [Digital image]. Retrieved May 19, 2023, from https://www.statista.com/chart/22048/university-of-oxford-coronavirus-containment and-health-index-selected-countries/

Cairo, A. (2019). How charts lie: Getting smarter about visual information. *Gildan Media*.

Caroll, R. T. (2015, November 21). "Regressive fallacy". *The Skeptic's Dictionary - Skepdic.com*. https://www.skepdic.com/regressive.html

CEF. (2022, August 8). Data and analysis for designing policies - center of excellence in finance. *CEF*. https://www.cef-see.org/data-and-analysis-for-designing-policies

Chengyee, J. C., & Luo, Y. (2021). Data visualization and cognitive biases in audits. *Managerial Auditing Journal*, *36*(1), 1-16. https://doi.org/10.1108/MAJ-08-2017-1637

Dhami, S. S. (2016). The Foundations of Behavioral Economic Analysis (First). *Oxford University Press*.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277. https://doi.org/10.1016/j.joep.2011.10.009

Duffy, & Nagel, R. (1997). On the Robustness of Behaviour in Experimental `Beauty Contest' Games. *The Economic Journal (London)*, 107(445), 1684–1700. https://doi.org/10.1111/1468-0297.00250

Government Offices of Sweden. (2021, September 21). Economic measures on account of the pandemic 2020–2021. https://www.government.se/articles/2021/09/economic-measures-on-account-of-the-pandemic-20202021/

Economist: Expect inevitable‘ recession by 2024. (2023, Jan 19). *Charleston Regional Business Journal* http://proxy.library.ju.se/login?url=https://www.proquest.com/newspapers/economist-expect-inevitable-recession-2024/docview/2768938625/se-2

Ellis, C. (2022, June 8). When to use DBSCAN. Retrieved April 21, 2023, from https://crunchingthedata.com/when-to-use-dbscan/.

Farmer, J. D., & Axtell, R. L. (2022). Agent-Based Modeling in Economics and Finance: Past, Present, and Future. *Federal Reserve Bank of St Louis*.

Faul, F., Erdfelder, E., Buchner, A. et al. (2009, November). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160 (2009). https://doi.org/10.3758/BRM.41.4.1149

Fogacci, Borghi, C., & Cicero, A. F. G. (2020). Misinterpreting data in lipidology in the era of COVID-19. *Journal of Clinical Lipidology*, 14(4), 543–544. https://doi.org/10.1016/j.jacl.2020.07.004

Gillies, S. (2023, May 1). The shapely user manual. *The Shapely User Manual - Shapely 2.0.1 documentation*. https://shapely.readthedocs.io/en/stable/manual.html

Horsley, J. S., Liu, B. F., & Levenshus, A. B. (2010). Comparisons of U.S. government communication practices: Expanding the Government Communication Decision Wheel. *Communication Theory*, *20*(3), 269–295. https://doi.org/10.1111/j.1468-2885.2010.01363.x

Hutchinson, J. W., & Alba, J. W. (1997). Heuristics and biases in the "eyeballing" of data: The effects of context on intuitive correlation assessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 591–621. https://doi.org/10.1037/0278-7393.23.3.591

Jackson (2019). The Friendship Paradox and Systematic Biases in Perceptions and Social Norms. *The Journal of Political Economy*, 127(2), 777–818. https://doi.org/10.1086/701031

Jarvenpaa, S.L. (1990). Graphic displays in decision making — the visual salience effect. *Journal of Behavioral Decision Making*, 3: 247-262. https://doi.org/10.1002/bdm.3960030403

JTS. (n.d.). Class ConvexHull. *Org.locationtech.jts:JTS-core* 1.18.0 API. https://locationtech.github.io/jts/javadoc/index.html?org/locationtech/jts/algorithm/ConvexHull.html, Copyright: 2020

Kahneman, D., & Tversky, A. (Eds.). (2000). Choices, values, and frames. Russell sage Foundation ; Cambridge University Press.

Krishnan Muthu T. (2020, January 26). Understanding Graham scan algorithm for finding the convex hull of a set of points. *Muthukrishnan*. https://muthu.co/understanding-graham-scan-algorithm-for-finding-the-convex-hull-of-a-set-of-points/

Lamberts, K., & Goldstone, R. L. (2006). Handbook of Cognition. *Sage*.

Liang, S. T., Liang, L. T., & Rosen, J. M. (2021). COVID-19: a comparison to the 1918 influenza and how we can defeat it. *Postgraduate medical journal*, *97*(1147), 273–274. https://doi.org/10.1136/postgradmedj-2020-139070

Liu, R., Tang, Y. Y., & Chan, P. P. (2018). A fast convex hull algorithm inspired by human visual perception. *Multimedia Tools and Applications*, *77*(23), 31221–31237. https://doi.org/10.1007/s11042-018-6185-0

Mack, A., & Clarke, J. (2012). Gist perception requires attention. *Visual Cognition*, *20*(3), 300–327. https://doi.org/10.1080/13506285.2012.666578

Madsen, D. Ø. (2017). Not dead yet: The rise, fall and persistence of the BCG matrix. *Problems and Perspectives in Management*, *15*(1), 19–34. https://doi.org/10.21511/ppm.15(1).2017.02

Mandl, C.E. (2019). Perceiving Events, Patterns, and Structure. In: Managing Complexity in Social Systems. Management for Professionals. *Springer, Cham*. https://doi.org/10.1007/978-3-030-01645-6_2

McKinsey & Company. (2008, September 1). Enduring ideas: The Ge–mckinsey nine-box matrix. *McKinsey & Company*. Retrieved March 14, 2023, from https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/enduring-ideas-the-ge-and-mckinsey-nine-box-matrix

McNutt, A., Kindlmann, G., & Corell, M. (2020). Surfacing Visualization Mirages. Proceedings of the *2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831

Miltimore, Jon. (2020, 12 August). BBC: Sweden's Economy Is Doing Way Better Than the Rest of the EU During the COVID-19 Pandemic. FEE Stories. *Foundation for Economic Education*. https://fee.org/articles/bbc-sweden-s-economy-is-doing-way-better-than-the-rest-of-the-eu-during-the-covid-19-pandemic/

Mooman, M. (2023, April 10). Understanding bank run, consequences for economy. The Express Tribune. https://tribune.com.pk/story/2410923/understanding-bank-run-consequences-for-economy

Sarker, Md. R., Rahman, S. M., Islam, A. K., Bhuyan, Md. F., Supra, S. E., Ali, K., & Noor, K. M. (2022). Impact of covid-19 on small- and medium-sized enterprises. *Global Business Review*, *0*. https://doi.org/10.1177/09721509221093489

Sharma, & Bikhchandani, S. (2001). Herd Behavior in Financial Markets. *IMF Staff Papers*, 47(3), 279 https://www.imf.org/external/pubs/ft/staffp/2001/01/pdf/bikhchan.pdf

Schintler, L.A., & Chen, Z. (Eds.). (2017). Big Data for Regional Science (1st ed.). *Routledge*. https://doi.org/10.4324/9781315270838

Simon, H.A. (1990). Bounded Rationality. In: Eatwell, J., Milgate, M., Newman, P. (eds) Utility and Probability. The New Palgrave. *Palgrave Macmillan, London*. https://doi.org/10.1007/978-1-349-20568-4_5

Sommer, P. (2020, June 19). A gentle introduction to the convex hull problem. *Medium*. Retrieved April 21, 2023, from https://medium.com/@pascal.sommer.ch/a-gentle-introduction-to-the-convex-hull-problem-62dfcabee90c

Statistics Canada, Government of Canada. (2011). Statistics: Power from data! misinterpretation of statistics. *Statistics: Power from Data! Misinterpretation of statistics*, https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch6/misinterpretation-mauvaiseinterpretation/5214805-eng.htm

Strong, H. (2014). Chapter 1: Arthur D Little Matrix. In Marketing and management models: A guide to understanding and using business models (1st ed., pp. 2–8). essay, *Business Expert Press*.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*, *2*, 249–288. https://doi.org/10.1016/s0065-2601(08)60009-x

The Assessment Network. (2020). Cambridge. *Cambridge University Press & Assessment*, https://www.cambridgeassessment.org.uk/blogs/the-misinterpretation-of-data-why-we-all-need-to-be-data-literate/

The Decision Lab. (2023a, April 4). *Anchoring bias*. The Decision Lab. https://thedecisionlab.com/biases/anchoring-bias

The Decision Lab. (2023b, May 2). Availability heuristic - Why do we tend to think that things that happened recently are more likely to happen again?. *The Decision Lab*. https://thedecisionlab.com/biases/availability-heuristic

The Nobel Foundation. (2002). Vernon L. Smith Biographical - The Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2002. *NobelPrize.org*. https://www.nobelprize.org/prizes/economic-sciences/2002/smith/biographical/

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Valeti, D. (2021, October 18). DBSCAN algorithm for Fraud Detection & Outlier detection in a data set. *Medium*. Retrieved April 21, 2023, from https://medium.com/@dilip.voleti/dbscan-algorithm-for-fraud-detection-outlier-detection-in-a-data-set-60a10ad06ea8

Venegas -Vera, V. A., Colbert, G. B., & Lerma, E. V. (2020). Positive and negative impact of social media in the COVID-19 ERA. *Reviews in Cardiovascular Medicine*, *21*(4), 561–564. https://doi.org/10.31083/j.rcm.2020.04.195

7. Appendices

Appendix 1

OLS Regression Results 1 (transformed variables)						
R-Squared	**Adj.R-Squared**	**F-statistic**	**Df-Model**	**Df-Residuals**	**AIC**	**BIC**
0.091	0.090	91.57	2	1837	-3243	-3226
N	**Dependent Variable**	**Kurtosis:**	**Skew**	**Omnibus**	**Jarque-Bera**	**Cond.No.**
1840	d(R,G)^0.1	3.535	-0.043	15.325	22.507	8.09
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	1.1495	0.006	202.087	0.000	1.138	1.161
d(R,C)^0.5	0.0247	0.002	13.403	0.000	0.021	0.028
STATS	-0.0083	0.005	-1.764	0.078	-0.018	0.01

Source: Survey Results

Appendix 2

OLS Regression Results 2 (transformed variables)						
R-Squared	**Adj.R-Squared**	**F-statistic**	**Df-Model**	**Df-Residuals**	**AIC**	**BIC**
0.032	0.031	30.10	2	1837	-3127	-3111
N	**Dependent Variable**	**Kurtosis:**	**Skew**	**Omnibus**	**Jarque-Bera**	**Cond.No.**
1840	d(R,G)^0.1	3.611	0.014	18.190	28.669	6.30
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	1.1873	0.005	254.851	0.000	1.178	1.196
Ω^0.5	0.0928	0.012	7.544	0.000	0.069	0.117
STATS	-0.0067	0.005	-1.370	0.171	-0.016	0.003

Source: Survey Results

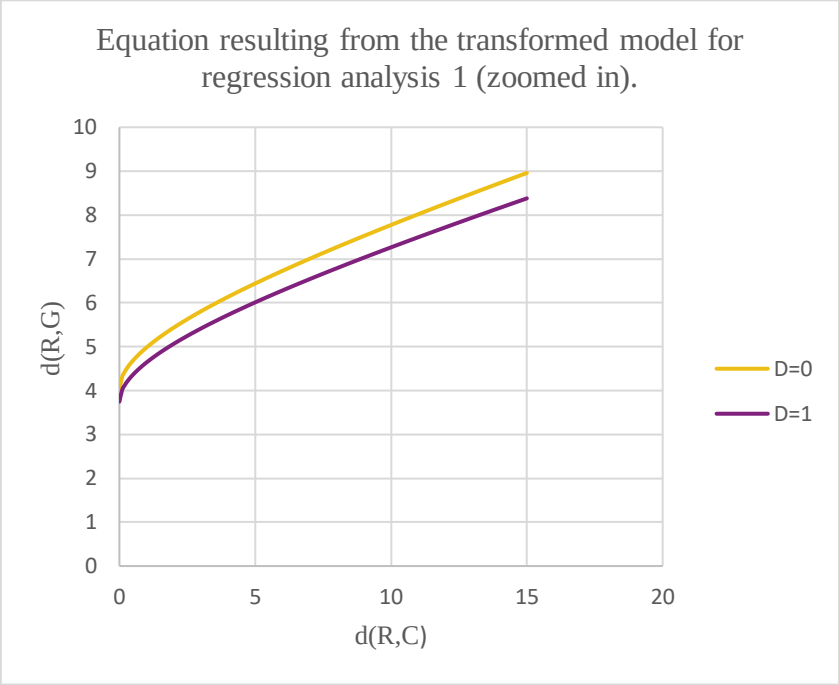
Appendix 3

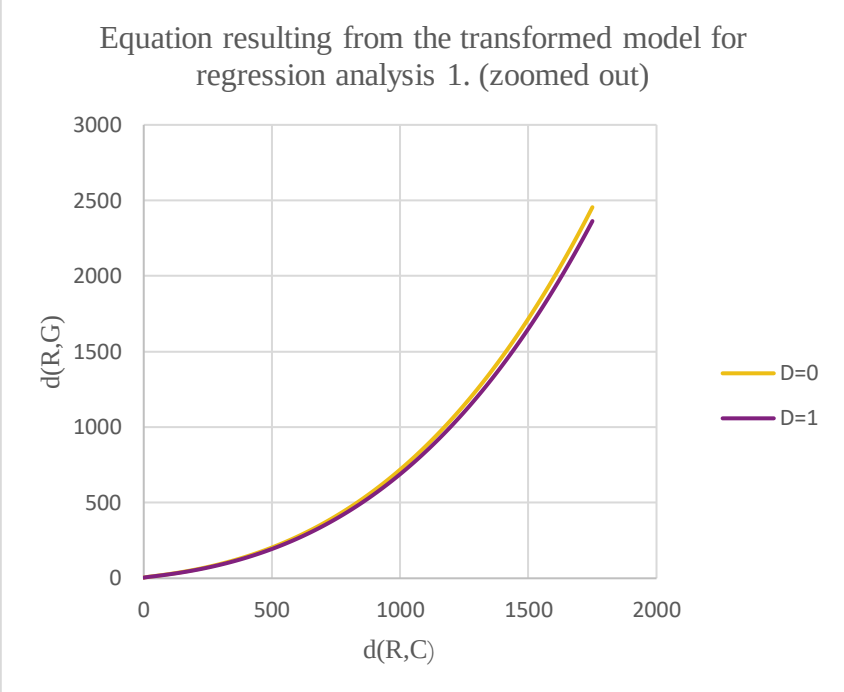
OLS Regression Results 3 (transformed variables)						
R-Squared	**Adj.R-Squared**	**F-statistic**	**Df-Model**	**Df-Residuals**	**AIC**	**BIC**
0.403	0.403	620.6	2	1837	-4018	-4001
N	**Dependent Variable**	**Kurtosis:**	**Skew**	**Omnibus**	**Jarque-Bera**	**Cond.No.**
1840	d(R,G)^0.1	4.217	-0.494	116.221	188.387	14.3
	Coefficients	**Std. Err.**	**t**	**P>|t|**	**0.025**	**0.975**
Constant	0.8019	0.012	67.204	0.000	0.779	0.825
d(C,G)^0.2	0.2617	0.007	35.154	0.000	0.247	0.276
STATS	0.0002	0.004	0.044	0.965	-0.007	0.008

Source: Survey Results

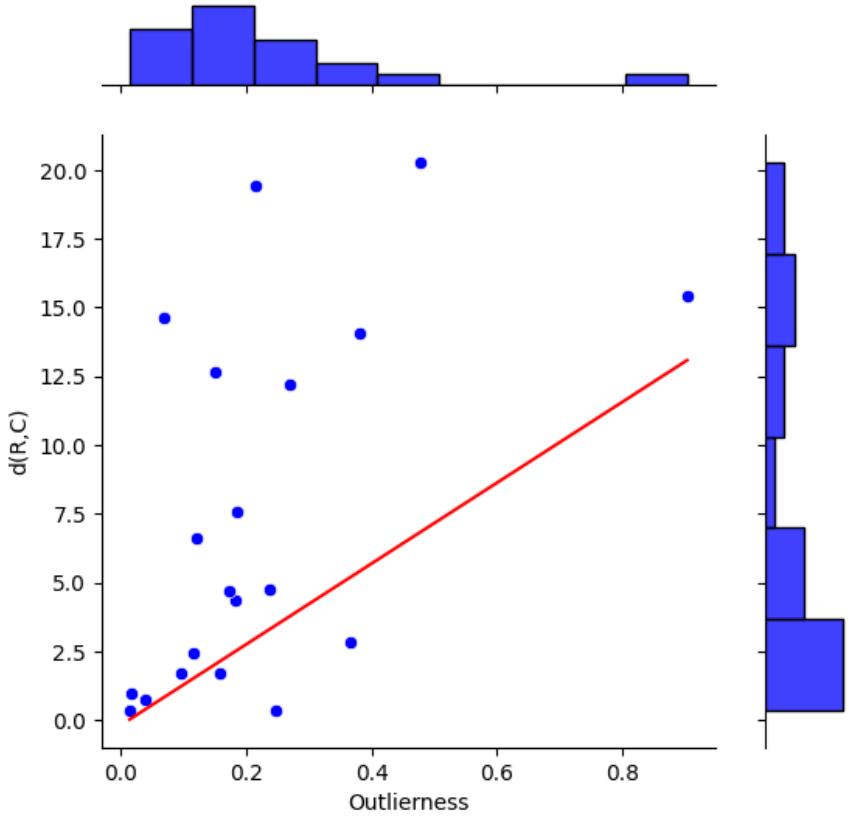
Appendix 4

$$d(R,G) = \left(1.1495 + 0.0247 d(R,C)^{\frac{1}{2}} - 0.0083 D_1 + \varepsilon_1\right)^{10}$$





Appendix 5



Source: Generated datasets

Appendix 6

GDPR Thesis Study Consent Form

Required by European Union General Data Protection Regulation 2016/679

GDPR Consent for Survey about Data Interpretation and Data Structure

Please tick the appropriate boxes **Yes**

Taking part in the study

I consent to JIBS processing my personal data in accordance with current data protection legislation and the data delivered. ☐

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason. ☐

My signature below indicates that I choose to take part in the thesis study and consent to JIBS treating my personal data in accordance with current data protection legislation and the data delivered.

_____ _____ _____

Name of participant [IN CAPITALS] Signature Date

Participant Information Sheet template

This is for guidance only, You MUST, however, include contact details and the Data Protection Privacy Notice.

Invitation paragraph

You *are being invited to take part in a thesis study. Before you decide whether or not to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully.*

What is the purpose of the study collecting personal data?
This survey is used to collect data for a bachelor thesis. The survey should take approximately 5-10 minutes. Participation in the experiment will help us develop a greater understanding of how people interpret data, and how that relates to data structure.

It is entirely up to you to decide whether or not to take part. If you decide to do so, you will be given this information sheet to keep and will be asked to give your consent. All the information that we collect about you during the course of the research will be kept strictly confidential. You will not be able to be identified in any ensuing reports or publications.

Under GDPR you have the following rights over your personal data:

- ***The right to be informed**. You must be informed if your personal data is being used.*
- ***The right of access**. You can ask for a copy of your data by making a 'subject access request'.*
- ***The right to rectification.** You can ask for your data held to be corrected.*
- ***The right to erasure**. You can ask for your data to be deleted.*
- ***The right to restrict processing**. You can limit the way an organisation uses your personal data if you are concerned about the accuracy of the data or how it is being used.*
- ***The right to data portability**. You have the right to get your personal data from an organisation in a way that is accessible and machine-readable. You also have the right to ask an organisation to transfer your data to another organisation.*
- ***The right to object**. You have the right to object to the use of your personal data in some circumstances. You have an absolute right to object to an organisation using your data for direct marketing.*
- ***How your data is processed using automated decision making and profiling**. You have the right not to be subject to a decision that is based solely on automated processing if the decision affects your legal rights or other equally important matters; to understand the reasons behind decisions made about you by automated processing and the possible consequences of the decisions, and to object to profiling in certain situations, including for direct marketing purposes.*

You should also know that you may contact the data protection officer if you are unhappy about the way your data or your participation in this study are being treated at *dpo@ju.se*

Thank you for reading this information sheet and for considering whether to take part in this research study.'

Contact details for further information

Thesis supervisor _____Lina Bjerke_____ E-mail: **Lina.Bjerke@ju.se**

Thesis student _____Joakim Ben David_____ E-mail: **bejo20vl@student.ju.se**

Thesis student _____Balazs Farago_____ E-mail: ***faba20ku*****@student.ju.se**

Survey

Bachelor Thesis
by Joakim Ben David & Balazs Farago

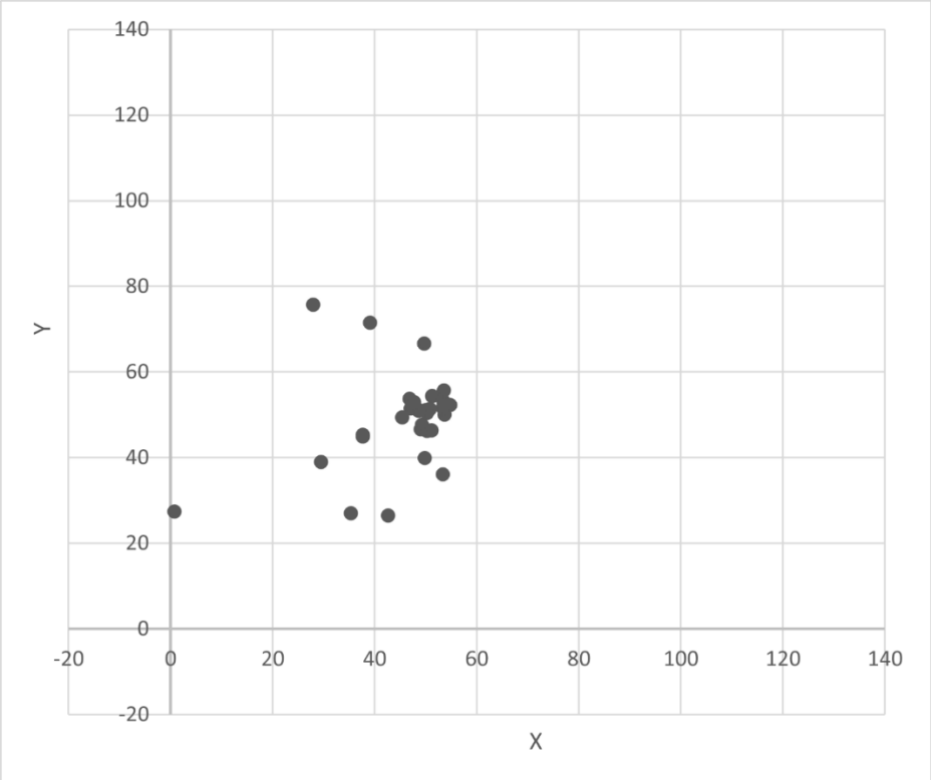
Instructions
You will be exposed to a series of scatter plot graphs with data points. On each graph, you are asked to estimate where you personally think the mean of the dataset is; that is, the average of all the data points. You are asked to show us your guess by drawing a dot where you believe the mean is on the graph.

Please tick the following box if it applies to you: Yes

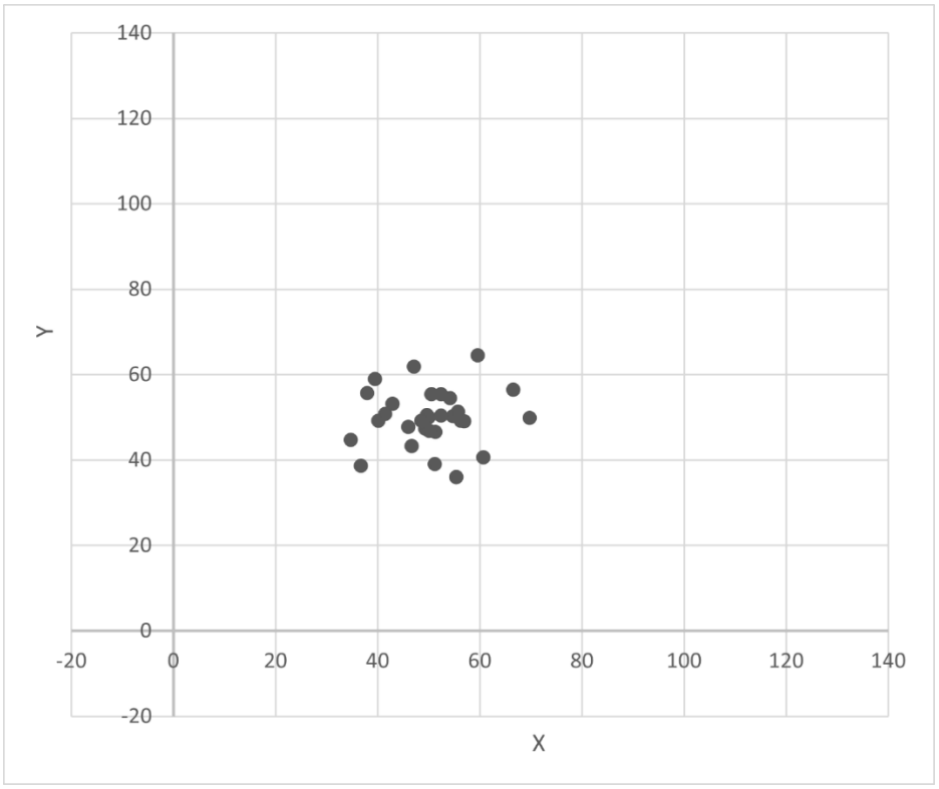
Have you passed any university level statistics course? o

Please interpret where the mean is on the following graphs:

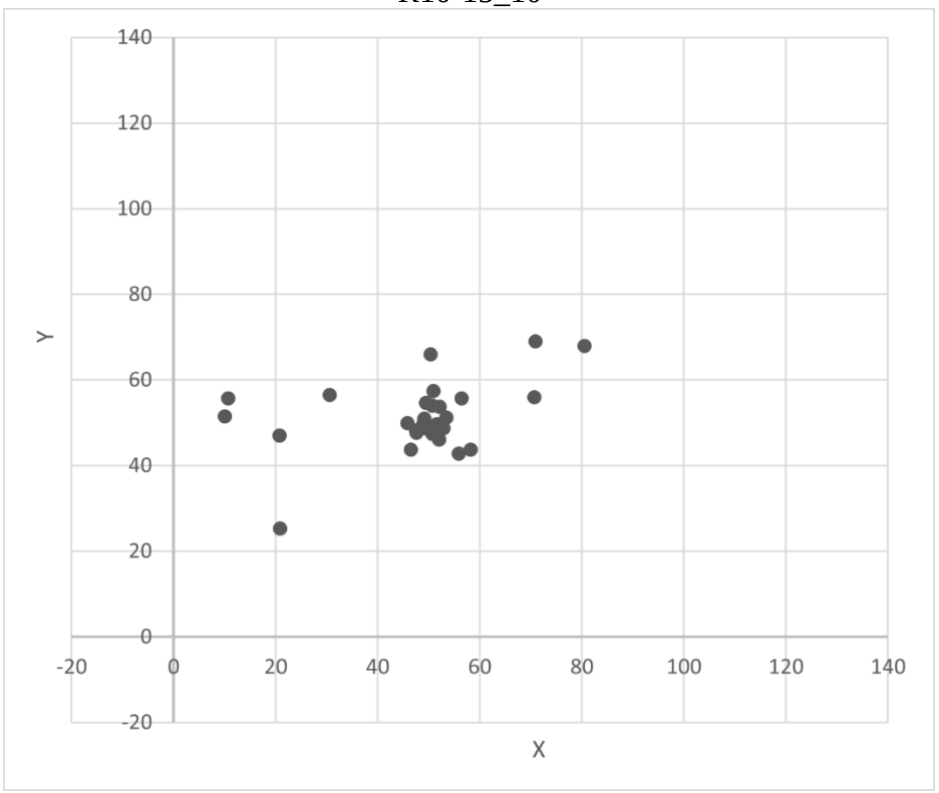
R15-20_29



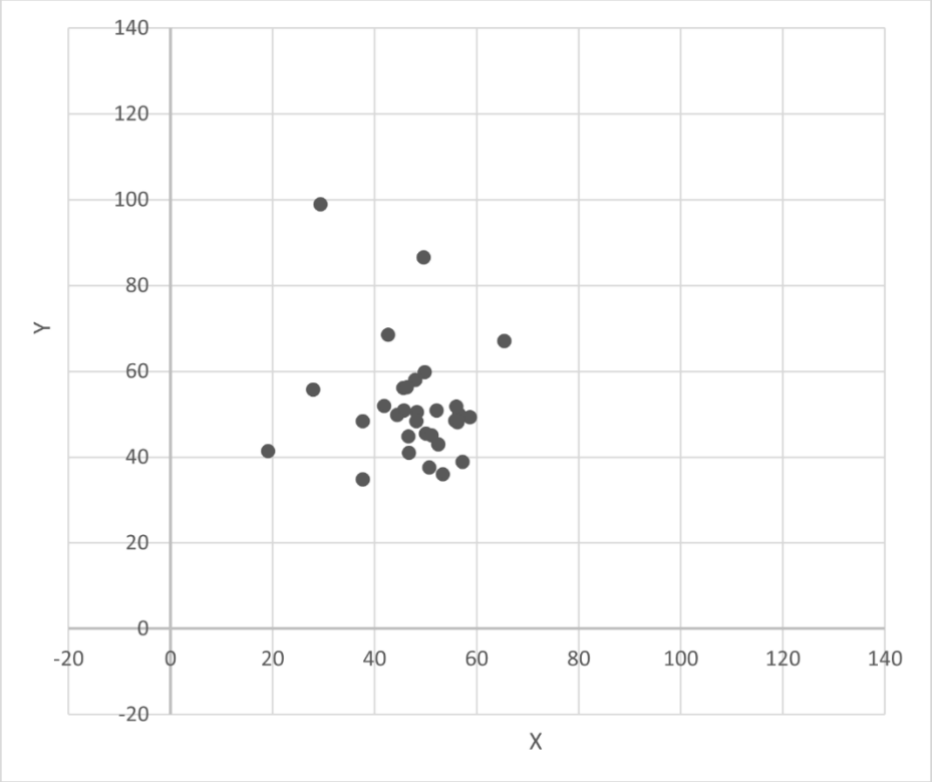
o3


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

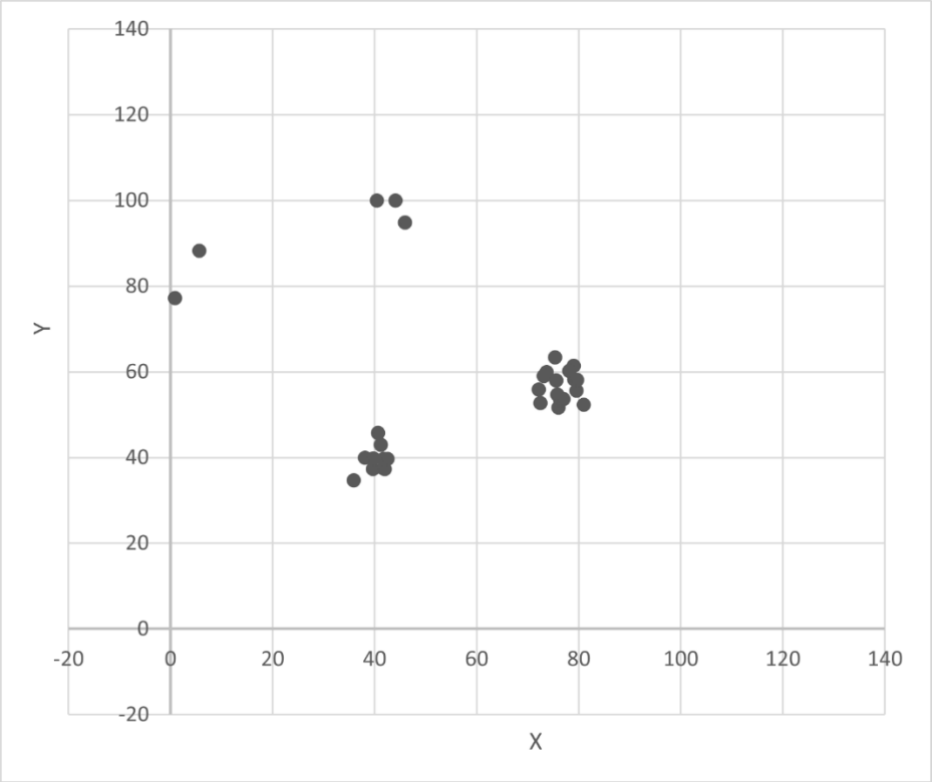
R10-15_10


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

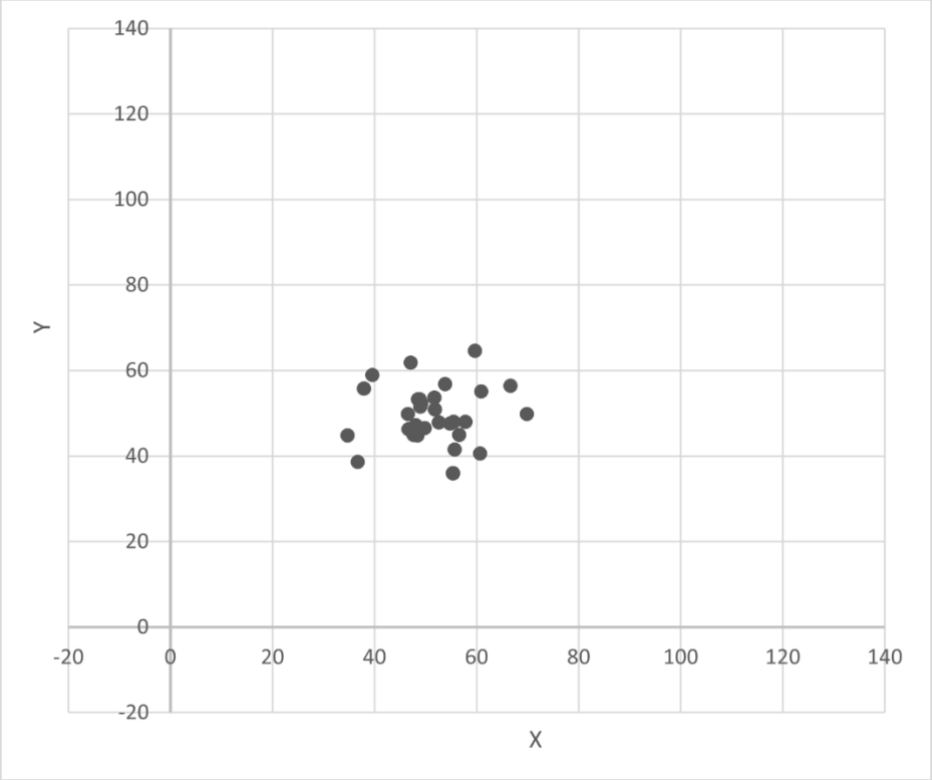
R25-30_6



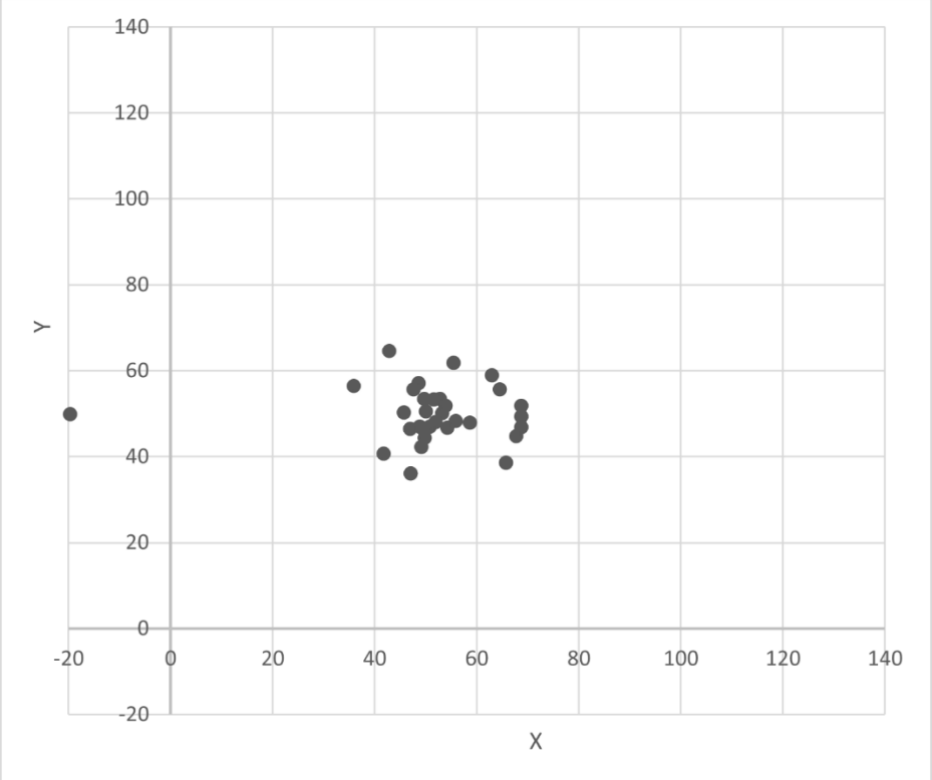
R25-30_7



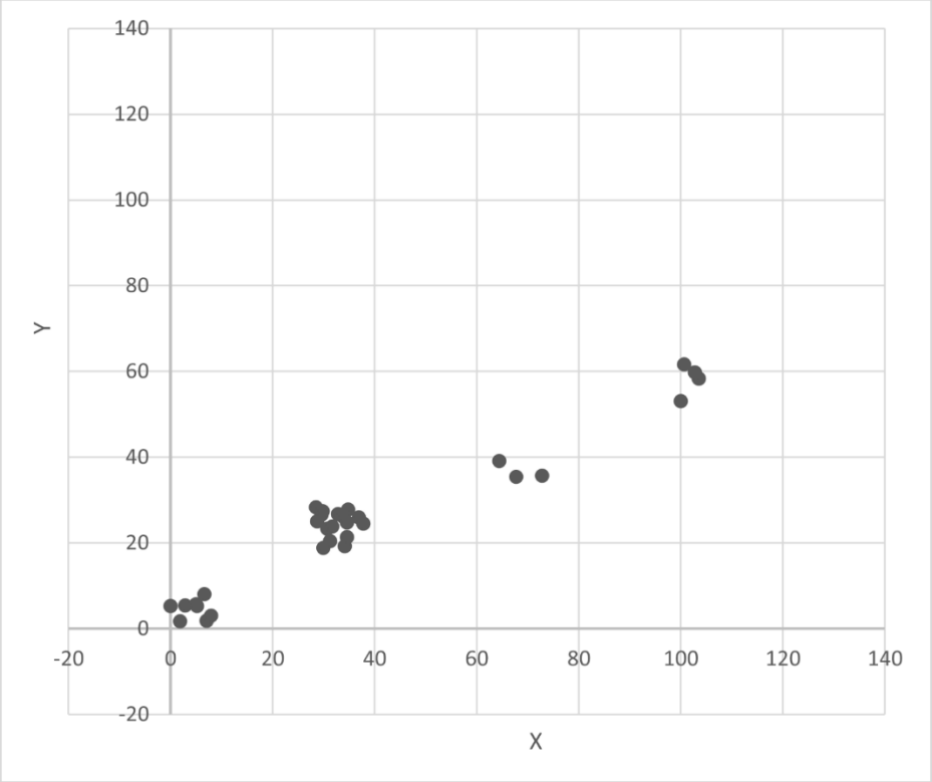
o5


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

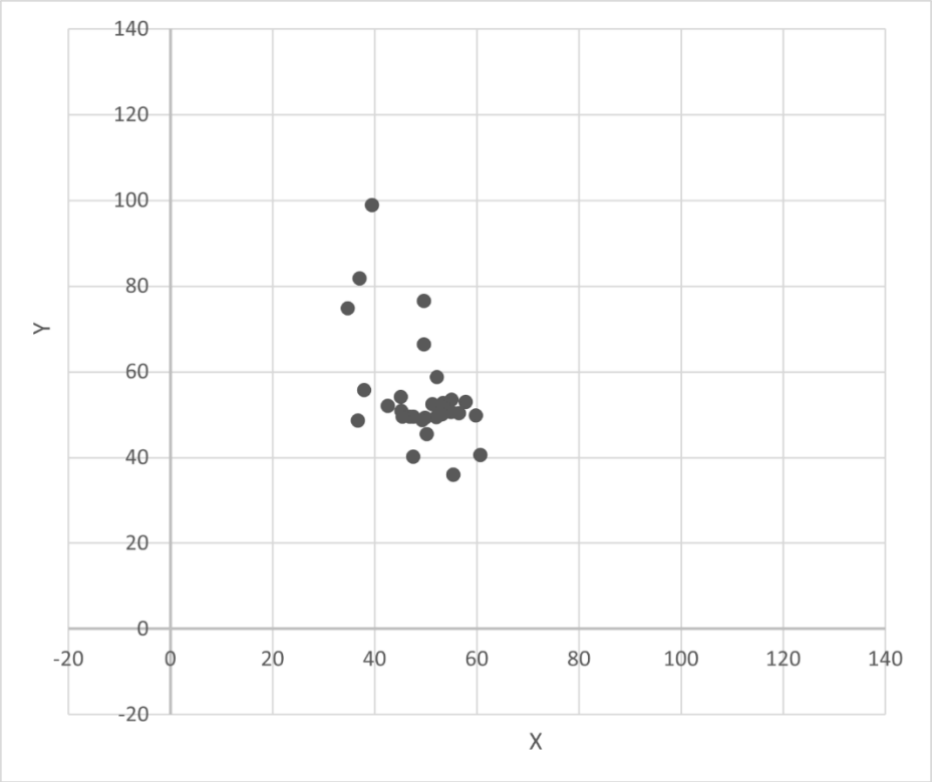
bl8


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

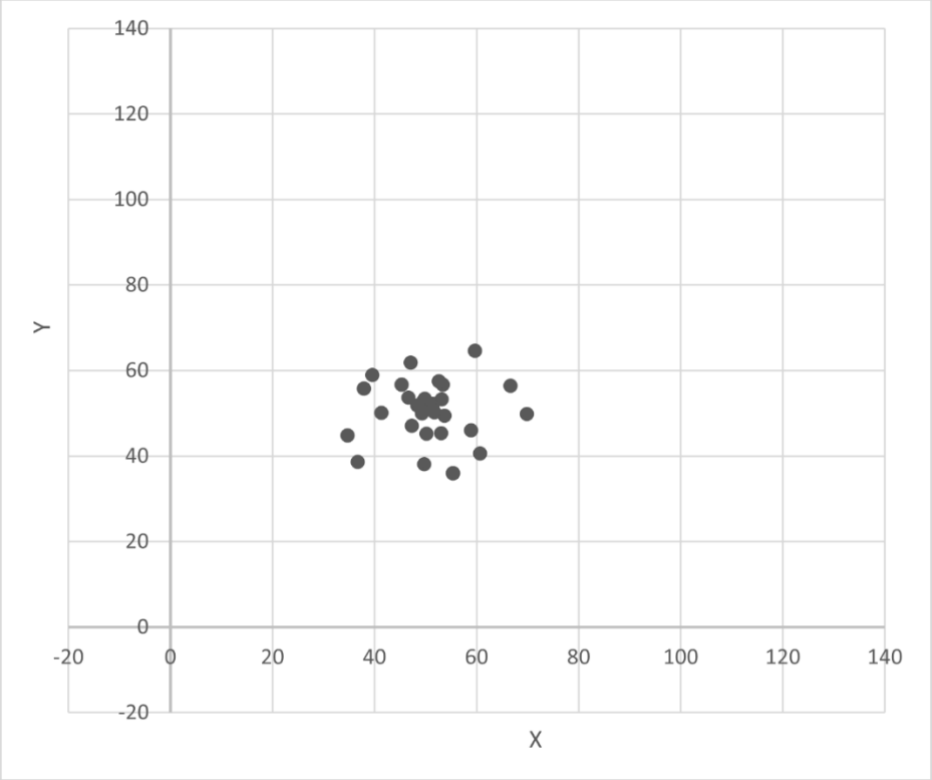
R10-15_110



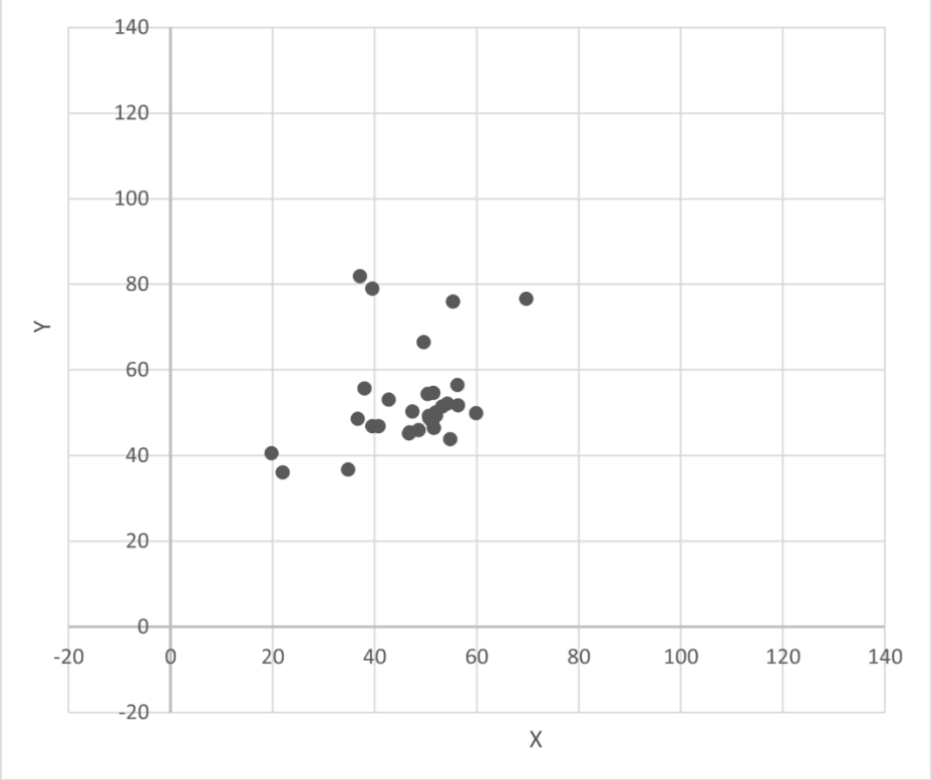
R0-5_1



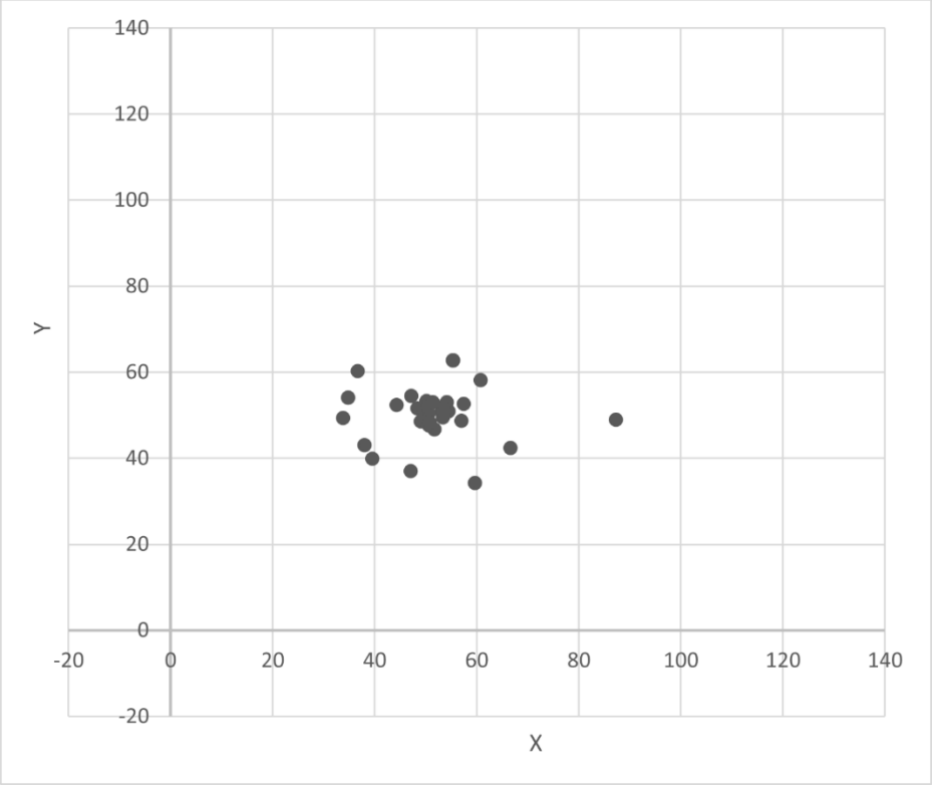
o4


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

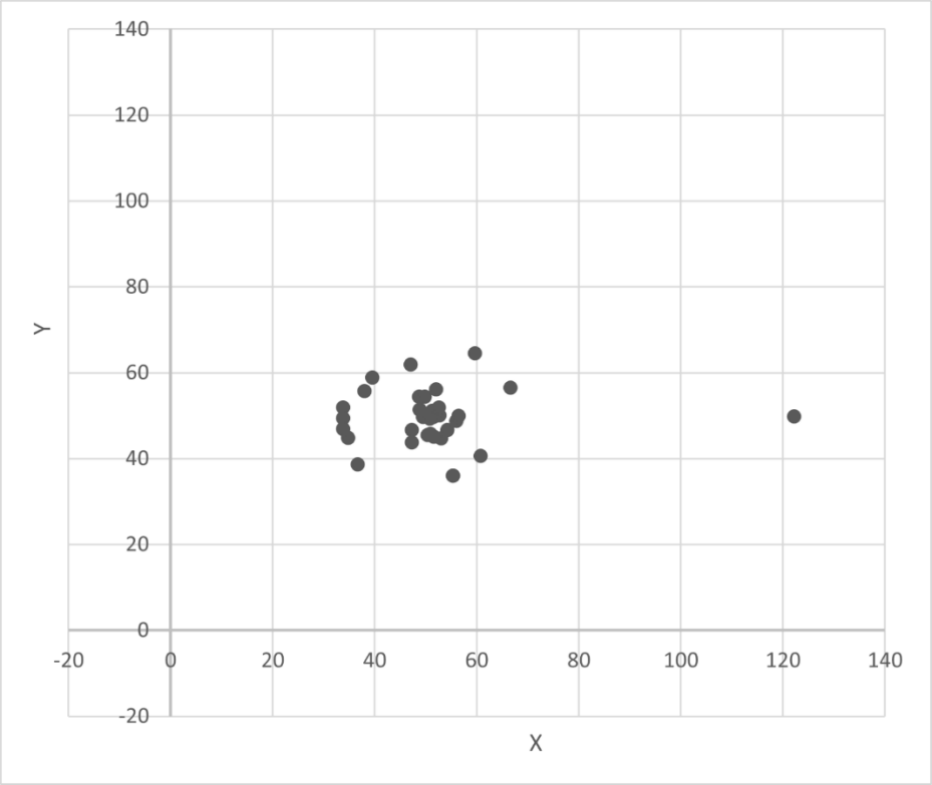
R15-20_2


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

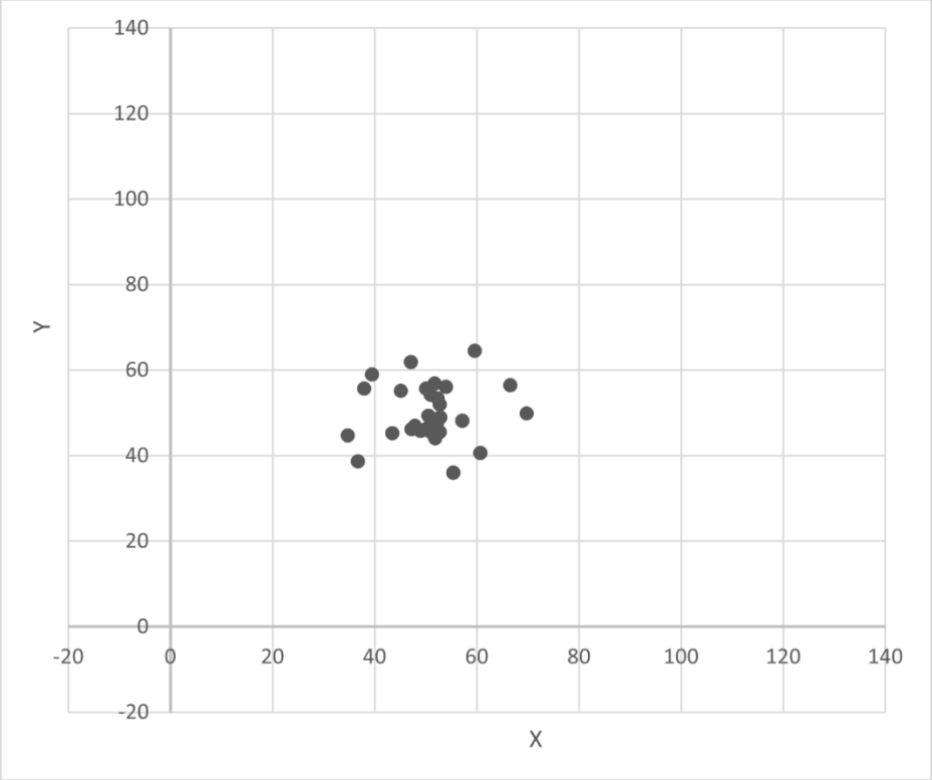
bl7


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

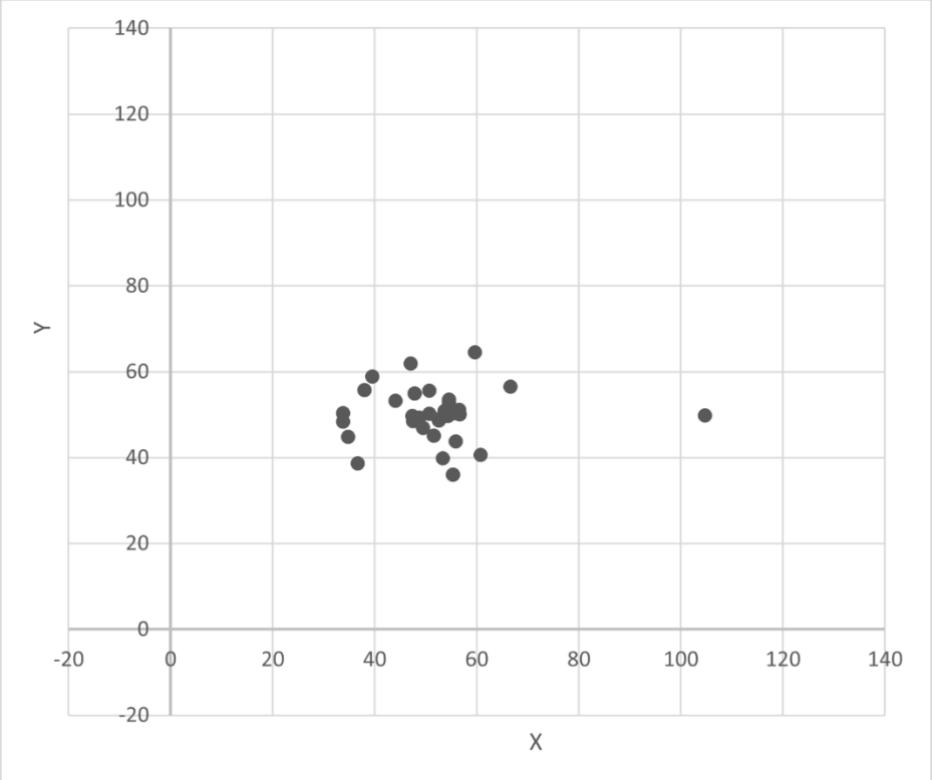
bl3


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

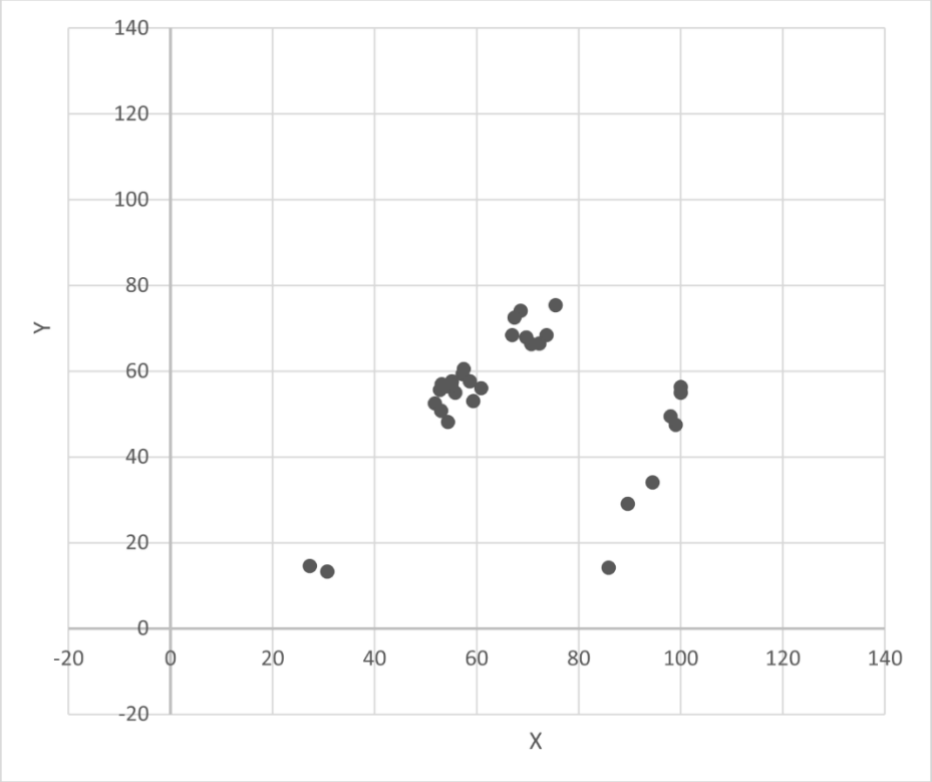
o1


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

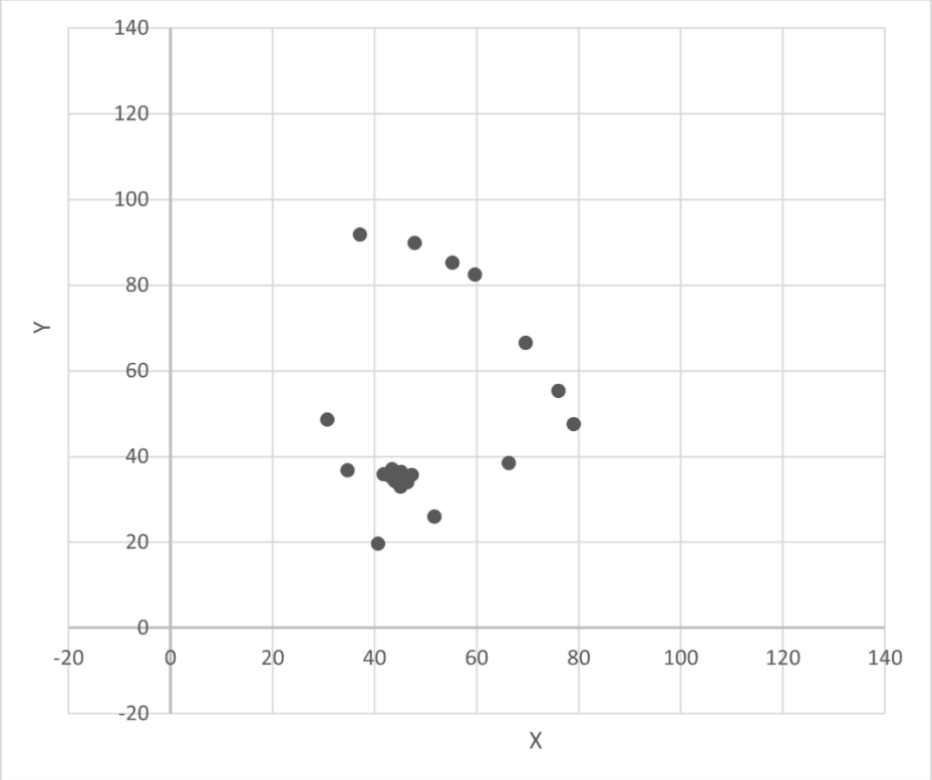
bl1


140
120
100
80
60
40
20
0
-20
Y
-20
0
20
40
60
80
100
120
140
X

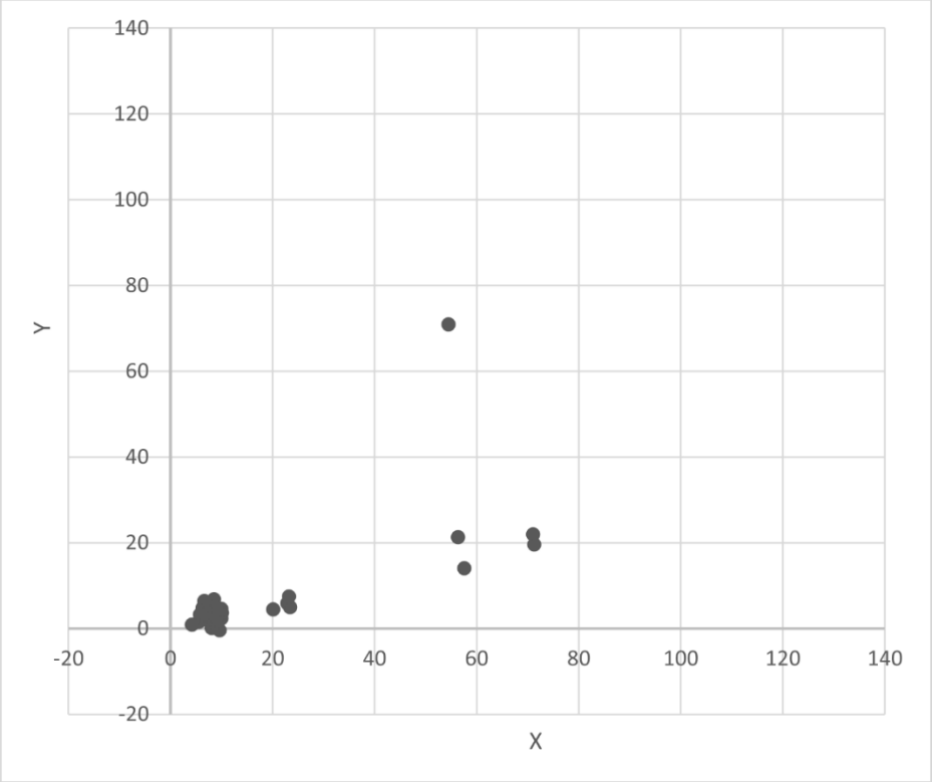
R5-10_64



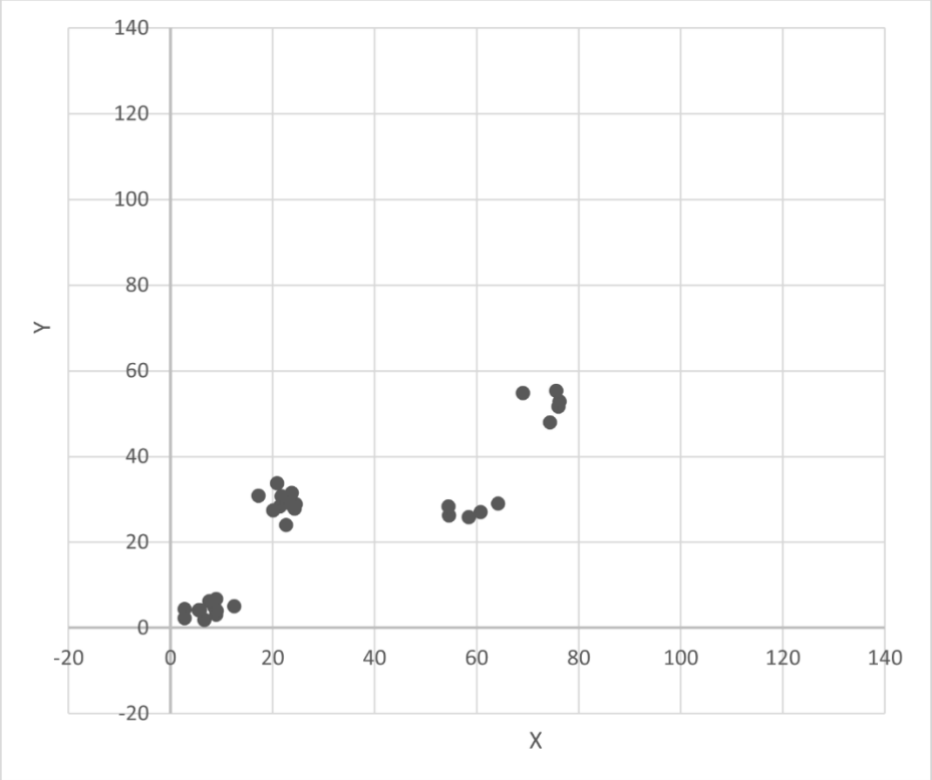
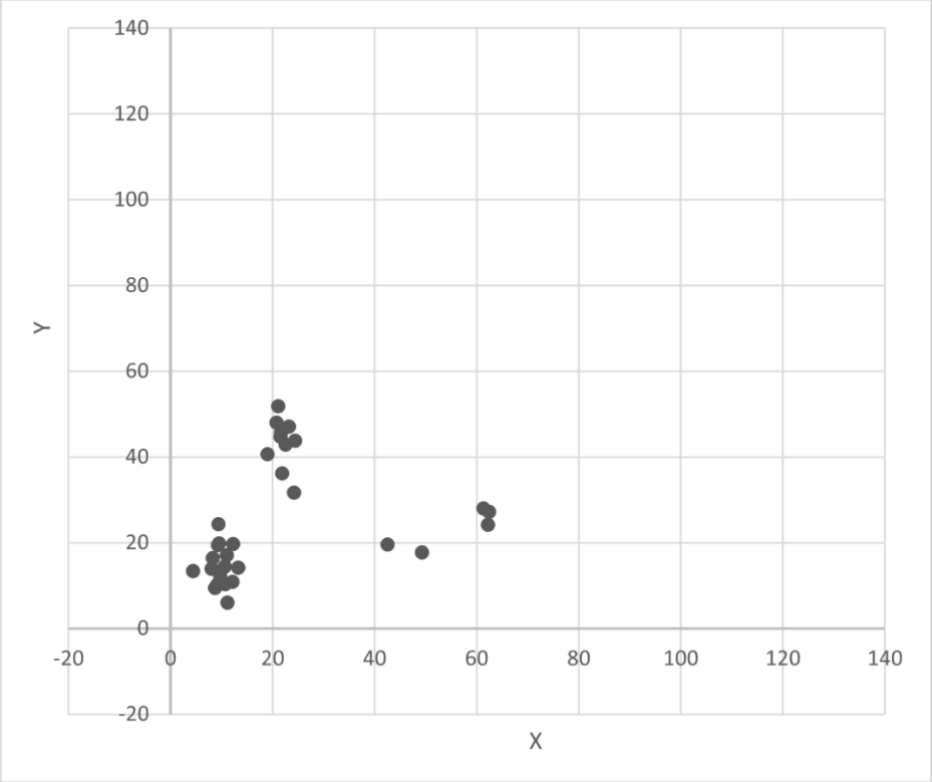
R20-25_46



R0-5_62



R20-25_131



R5-10_70