City University London

MSc in Business Systems Analysis and Design

Project Report

2013-2014

Investigating Educational Performance in Relation to Crime in London Boroughs through Quantitative and Visual Analysis Methods

SANDRA OBIANWUZIA

Supervised by:

Dr Cagatay Turkay

September 2014

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Sandra Obianwuzia

Date: 26th September, 2014

Abstract

Innovative approaches that combine both human and computer capabilities are needed as a guide to analysts, to help in analysing the growing amount of multivariate data sets today. Analysing multiple multivariate data sets is seen as a challenge to most analysts. This is because they are usually measured in different scales, and contains different numbers of variables. In this research, the author proposes a data mining framework, called Multidimensional Temporal Data Explorer (MTDE), for analysing multiple temporal multivariate datasets which can be easily interpreted. The research focuses on the two multivariate data sets – Gcse result by pupil of residence and crime rates statistics by London Borough and by time. MTDE was applied to each datasets in a bid to find similar patterns of correlation of which were successful.

Keywords: Information Visualisation, Visual Analysis, Principal component analysis, R programming Language, Open data, Education, Crime, Visual Perception

Acknowledgement

I would like to express my profound gratitude to my parents, Mr. and Mrs. Obianwuzia, for their constant belief in my abilities. I am eternally grateful for everything that you have done for me.

My completion of this project could not have been possible without the support of my project supervisor, Dr. Cagatay Turkay. Thank you very much for always putting me on the right track during the duration of this project. Your words of advice and constructive suggestions were sources of encouragement and contributed a great deal towards the actualization of this project. "Tessekur Ederim,Hocam".

Finally, I wish to thank all my loved ones for their care, love and support while I was working on my project. To my elder sister, Miss Sarah Obianwuzia – thank you for always providing me with some much needed comic relief during this hectic period. You are very much appreciated.

Contents

LIST OF FIGURES

LIST OF TABLES

APPENDICES

A- Project Proposal

B: R Programming Commands used in Performing Data Analysis in RStudio

- **B1:** Commands Used in Performing the Standardization process For Both Data Sets.
- **B2:** Commands used for Performing Pearson's Correlation and Plotting the Scatter Diagrams for Both Datasets
- **B3**: Commands Used In Performing PCA (Calculating The Eigenvalues, The Factor Loadings, And The PCA Scores)

C: DETAILED DESCRIPTIONS OF THE DATA SET

- **C1**: Detailed description of the raw GRLPR data set
- **C2:** Part of the filtered Dataset combined. Originally have 192 rows
- **C3**: Combined PCA scores for both variables.

D: SCATTER PLOT MATRICES

- **D1:** Scatter plot matrices for years 2010, 2011, 2012, and 2013.
- **D2:** Scree plot of GRLPR PCA values

CHAPTER 1: INTRODUCTION AND OBJECTIVES

Information Visualization techniques have been shown to be relevant in different domains, but have not been broadly studied for the application in identifying the degree of crime and education attainment relationship both locally and globally. It is as a result of the data integration problems faced by analyst looking at multiple datasets that are of different scales. It is relevant for education and crime analyst to retrieve, understand and analyze the correlations effectively and efficiently. In this chapter, the author gives a project background knowledge, aims and objectives, project scope, project beneficiaries, work plan, and limitations.

1.1 PROJECT BACKGROUND

Education in its general sense is a form of learning (Schoolkidsfun.com, 2014) in which individuals or a set of people transfer information, knowledge, and ability from one era to the next via various means such as instructing or exploration. It is viewed as one of the most vital things in life because without it; one may not be able to contribute positively to the world. Hence wise, to make these contributions, an individual need to attain certain grades at various levels of education. It is referred to as Educational Performance (EP) or Educational Attainment (EA) - these titles will be used interchangeably throughout this dissertation.

With more and more educational institutions collecting and storing huge amounts of multivariate data, pertaining to distinctive processes, present datasets often get to be intricate and excessively extensive to handle or comprehend. As a result, its sets off a requirement for solutions that makes the investigation and analysis of these datasets more sensible. One particular opportunity made by such multivariate information is the possibility to search for relationships and explore connections between these datasets. Consequently, researchers are constantly discovering distinctive techniques to which these connections are analysed, visualized and communicated to the right bodies. Data visualization or information visualization techniques are increasingly becoming the method used for interpreting huge amounts of data in visual form. Card et al(1999) provided an extensive overview of research that has

been done whereas Tufte(1983, 2010) wrote comprehensively on information visualization and the development of the field.

Enhancing student's educational performance is the key component of the educational process. For this reason, Social scientists have over the years researched on factors that may be influenced by how well students perform academically. Factors including but not limited to health, parental involvement, classroom designs, teacher's involvement, gender, and race(Hamnett, Ramsden and Butler, 2007; Sabahat, 2012; Wilder, 2014) have been found to be influenced by or have an influence on student's educational Attainment.

However, there is a relatively small literature linking out-of-school crime rates as a factor that may have a relationship with the educational performance of students. A vast majority of research in the social sciences is concerned with the correlation between unemployment rate and crime rate through statistical analysis methods(Becker, 1974; Hale and Sabbagh, 1991; Elliott and Ellingworth, 1996; Paternoster and Bushway, 2001; Raphael and Winter- Ebmer, 2001), of which a certain degree in the association has been found. Also, these research studies do not apply the use of information visualization to communicate results, gain insights and possibly find if a pattern/trend exists. Therefore making it hard for individuals to comprehend.

Spatiotemporal datasets on crime rates and student grades are been publicized annually and monthly by the open data websites, to be used for various purposes. As a result, the author was motivated to contribute to the body of knowledge comprising the support of visual analytics to the social sciences with an improved research technique. Detailed review of visual analytics was discussed in the next chapter.

This dissertation considers how information visualization may be designed and applied to identify the degree to which out-of-school crime rates is correlated to educational performance over a period, in different coverage location. As a result of the author been based in London, it was decided to focus the research on each London borough. To begin, the aims and objectives are stated in the following section.

1.2 AIMS AND OBJECTIVES

The overall aims of this research are fourfold:

- To identify the degree of relationship between out-of-school crime rates and students Gcse performance.
- To develop information visualization design appropriate for exploring the temporal relationships.
- To discover if a pattern/trend exists over the period and by borough.
- To propose a framework for analysing (quantitative and visually) multiple temporal multivariate data sets.

To successfully fulfil the aims described above, goals or steps were described to guide the author in achieving the desired outcome. These are stated below:

- Data sets will be extracted from the London Datastore.
- Data sets will be filtered, covering the same period and location, and normalised to the same format.
- The data from both sets will be synthesised to establish if correlation points exists between datasets.
- A data visualization design will be developed to explore the temporal correlations over the same period in order to discover possible patterns/trends.
- A "recipe" for analysing multiple multivariate temporal data sets will be proposed.

1.3 PROJECT PRODUCTS AND BENEFICIARIES

The intended project will contribute to the already existing knowledge by providing the beneficiaries with an underlying understanding of the correlations between the educational performance of students and crime, as well as possibly discovering a pattern, through the use of statistical and visualization techniques. Most importantly, providing a framework for analysing more than one dataset with different scales, that can be adopted for use by researchers in other field of study.

The intended beneficiaries of this project are policy makers, who set plans pursued by the government. Local education authorities, who are the local councils in England and Wales responsible for the implementation of educational laws in a system; social scientists in the educational field, who will want to adapt this technique in discovering more patterns. Analyst interested in data with a temporal dimension locally or globally. Parents, who together with the other beneficiaries, would need this information to be able to make more informed decisions regarding education, like providing facilities and environment that will support learning and subsequently enhance students' academic achievements.

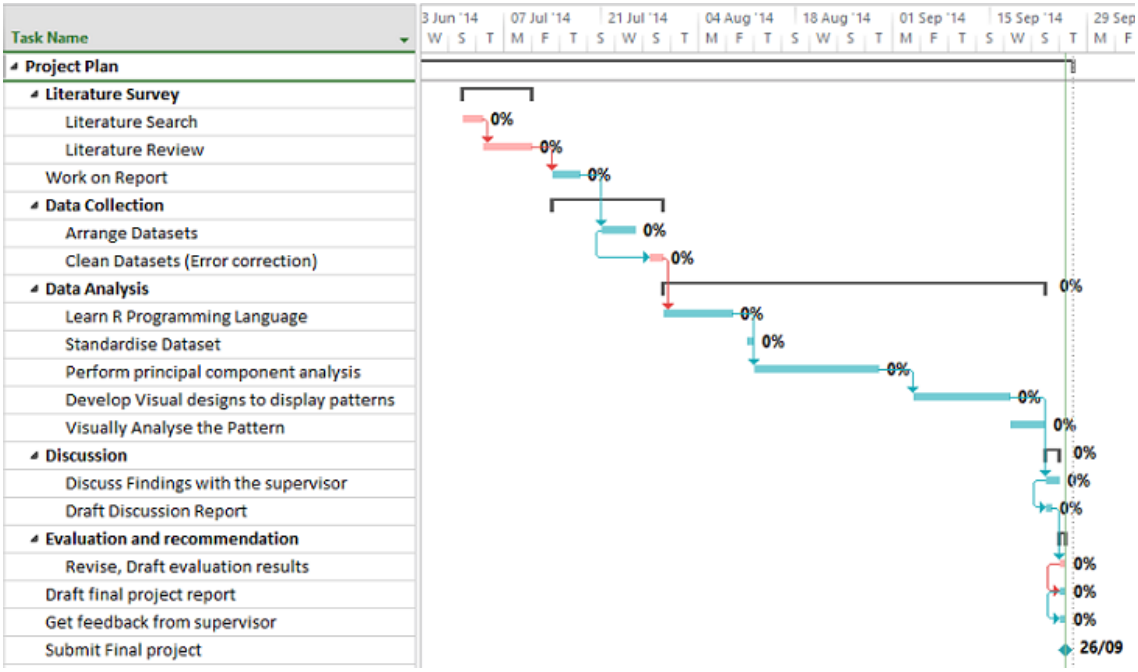
1.4 WORK PLAN/CHANGES MADE

This research was originally focused in finding a relationship between educational performances over London borough profiles, hence the title of the project proposal. These profiles included demography statistics, labour market statistics, environment statistics, economy statistics, and so on. It was felt that the focus was very vague. Therefore, following a consultation with the project supervisor, it was decided to focus on the relationship between crime and educational performance.

The original aim that was proposed was to find spatial and temporal relationships between each data set. This was later changed to focus on the temporal relationship because of the capabilities of the researcher. Considering the fact that the researcher was bounded by time, understanding

the technicalities involved in Spatio-temporal analysis was not deemed feasibly. Please see figure 1 for a change in work plan.

Figure 1: New Work Plan



1.5 STRUCTURE OF THE DISSERTATION

To achieve the objectives of this project, I will use the data from two datasets mentioned later. The dissertation will be structured as follows. First, a review of the relevant literature to ground my understanding of the subject matter will be carried out. Second, I will provide a detailed description of the methods and tools used for data analysis and visualization. Third, the results will be stated, in response to the methods applied. Fourth, a discussion on if the results retrieved matched with the project objective will be stated. Finally, an evaluation of the project as a whole will be provided.

CHAPTER 2: CONTEXT

This chapter provides a thorough review of literature which informed the project.

2.1 CRIME IN ENGLAND

Crime in the United Kingdom is depicted as acts of violent and non-violent crime that takes place in the United Kingdom. In Greater London, crime is controlled by three police forces; the Metropolitan Police (widely known informally as "**The Met**") which is responsible for law enforcement in the vast majority of the capital, divided into 32 boroughs, the City of London Police responsible for law enforcement within the city and British Transport Police that safeguards the rail network and London underground.

The Metropolitan Police publishes figures on crimes monthly and annually for the public to make use of. These figures are based on two categories; property crime and violent crime.

Property crime is defined as "*incidents where individuals, households or corporate bodies are deprived of their property by illegal means or where their property is damaged*"(Office of National Statistics, 2013). It does not involve force or threat of force against a victim, and includes burglary, theft, motor vehicle theft, fraud or forgery, and arson. On the other hand, violent crime includes sexual offences, robbery and violence against a person that involves the use of force.

The effective response to crime is one of the crucial priorities faced by any government and it's high on the public policy agenda. Then again, in order to provide effective crime control, it is vital to clearly comprehend the interaction between crime and various factors. As stated earlier, most research studies are concerned with the effect of unemployment rates on crime, and vice versa. However, there are few researches that look on the direct connection between educational attainment and crime especially in the UK. Data on educational outcomes and crime do not exist in the same place; hence, researchers will have to combine both datasets, needing suitable matching variables in each data source.

2.2 SECONDARY EDUCATIONAL QUALIFICATION IN THE UK

This section provides background knowledge on the academic qualification the author intends to make use of.

For a pupil to graduate from secondary school, a school leaving certificate is required. This is an academic qualification awarded for the completion of high school ("high school" and "secondary school" will be used interchangeably throughout the rest of the dissertation). In the UK, the General Certificate of Secondary Education (GCSE) is awarded to students as they finish secondary education. It is generally taken by students aged 14-16 in a number of subjects. Students attend a compulsory two years secondary education which incorporates GCSE referred to as Key Stage 4. This term is defined in section 82 of the Educational Act (2002) as "*the period beginning at the same time as the school year in which the majority of pupils in this class attain the age of fifteen and ending at the same time as the school year in which the majority of pupils in this class cease to be of compulsory school age*".

2.3 RESEARCH LINKING CRIME AND EDUCATIONAL DATA

After rigorous exploration of studies, the author found just two researches linking crime data and education data, both authored by Steve Machin, Olivier Marie and Sucica Vujic. In the first publication, Machin et al (2011) attempts to obtain the causal effect of education on crime by using the raising of the school leaving age(RoSLA) in the 1970s . Data was analysed in a regression-discontinuity setting. Their results demonstrate that the amount of crime reduces (for every 1000 of the population) by 2.1% when there is a 10% increase in the age at which individuals leave school. Additionally, it was found that a 1% decrease in the proportion of individuals leaving school with no qualifications reduces the number of property crime convictions by between 0.85% - 1%. Furthermore, their results on violent crime were to a great extent statistically irrelevant because they believe emotion motivates people to commit violent crime rather than economic reasons. Although this analysis does not get into the causal factors behind the numbers, it is important to note that there are many possible factors at play.

The second publication focuses on the connection between youth crime (young people aged 16-21) and education. They employed the same

principles that applied to RoSLA. They discovered that a 1% increase in male students decreases the number of male crime rates by around 1.9%, while the same amount of increase of male students in post-secondary education decreases male crime rates by 1.7%. Female students had less impact in reducing crime, at 1.15% and 1.3% respectively (2012).

What is clear from previous work is that there is evidence for some impact of education on crime through the use of secondary school leaving age and gender, but neither the connection between student grade performance and crime rates, nor whether patterns exist over a period for the crime types in different geographical locations are known. The author also found no studies that used visualization to explore and communicate these patterns.

In order to understand how visualization can provide insight for making valuable decisions, a review of information visualization literature was undertaken by the author to guide the design process of this project. This is described in the next section

2.4 INFORMATION VISUALIZATION – IMPROVING UNDERSTANDING AND INSIGHT

Modern computer systems are allowed to store large amounts of data due to the steady development of technology. Data are increasingly generated and stored by organisations for different business processes daily. This can be done automatically by sensors and monitoring systems or entered into the system manually. Even simple daily transactions such as making an online payment, or paying physically by credit card, and using your mobile phone are recorded by computers. These stored datasets usually consists of many parameters, resulting in multidimensional data. Individuals, organisations, and the society as a whole believe that there is a potential source of valuable information that can be derived from analysing these data visually, of which can lead to the respected bodies gaining a competitive advantage. However, it is an intricate task exploring these data textually. Additionally, as different organisations store different data, researchers are constantly looking for relationships between these datasets from multiple sources that can be analysed visually and be beneficial for the public and the society at large.

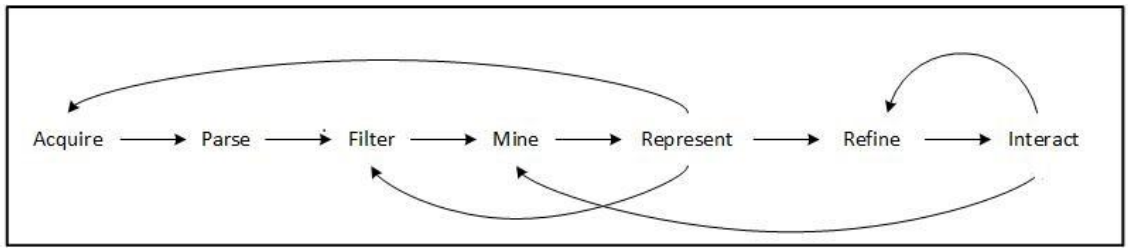
Gathering data is getting better and better but finding what can be done with it is the area the society lags behind. Information Visualization is increasingly becoming the preferred technique by researchers and organisations for gaining insight through visual representations. Gershon et al (1998) defines information visualization as "*the process of transforming data, information and knowledge into visual form making use of humans' natural visual capabilities*". They also state that it acts as an interface between the human perception and the computer, and these visual interfaces helps to interpret large volumes of data effectively to discover underlying characteristics, patterns and trends which are all the necessary outcomes of this dissertation.

Information visualization combines aspects of human-computer interfaces, data mining, scientific visualizations, imaging, and graphics (Robertson, Card and Mackinlay, 1993; Gershon and Eick, 1995). Rather than focusing on the data, it focuses on the information that can be derived through the use of various visual designs. Therefore, providing insights to the data is considered one of the major purposes of information visualization. Card et al (1999) declares that "*the purpose of visualization is insight, not pictures*".

2.4.1 HOW DO WE VISUALIZE DATA

Fry (2003) described seven stages that should be followed when visualizing data in his book titled "*Visualizing Data*". These are processes or steps that are necessary in order to convey information to individuals, however not all the steps shown in the figure below are important for every research because each research has its own unique requirements.

Figure 2: Interaction between Seven Stages Adopted From Few (2003, P.15)

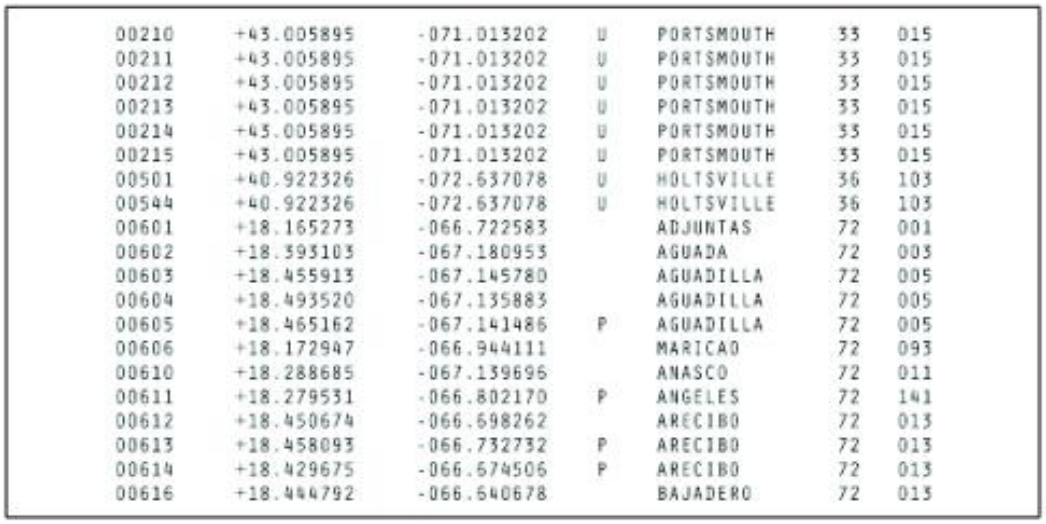


The following steps are described below in detail with an example that was adapted by Fry to illustrate the process.

Acquire

This is the stage where all the data necessary for visualization are obtained, whether from an internet source or a file stored in a hard drive. Fry (2003) made use of the zip code data found on the U.S. Census Bureau web site because of its availability to the public. The figure below shows only a tiny proportion of dataset as shown in his book.

Figure 3: Zip Codes in the Format Provided By U.S. Census Bureau Adopted By Fry (2003)

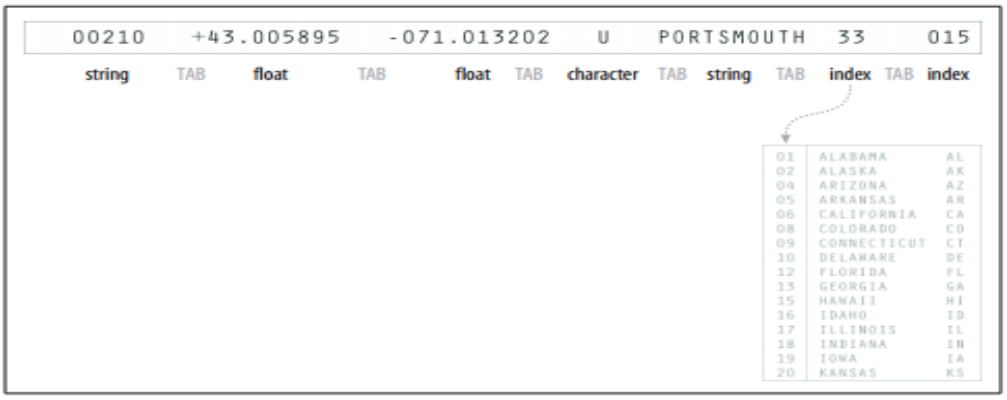


00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00544	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00601	+18.165273	-066.722583		ADJUNTAS	72	001
00602	+18.393103	-067.180955		AGUADA	72	003
00603	+18.455913	-067.145780		AGUADILLA	72	005
00604	+18.493520	-067.135883		AGUADILLA	72	005
00605	+18.465162	-067.141486	P	AGUADILLA	72	005
00606	+18.172947	-066.944111		MARICAO	72	093
00610	+18.288685	-067.139696		ANASCO	72	011
00611	+18.279531	-066.802170	P	ANGELES	72	141
00612	+18.450674	-066.698262		ARECIBO	72	013
00613	+18.458093	-066.732732	P	ARECIBO	72	013
00614	+18.429675	-066.674506	P	ARECIBO	72	013
00616	+18.444792	-066.640678		BAJADERO	72	013

Parse

Fry (2003) states this stage is where the data is parsed into the respected formats for analysis and visualization. Each line of the dataset are broken down into separate parts and changed into a useful format- in other words, giving it an understandable structure, following a set of rules. This is to make datasets more easily interpreted, managed, or transmitted by a computer. Figure 4 shows the layout or format of each line showed in figure 3, in their respected data type to make it understandable for parsing and deriving information. The data types shown in the figure below are string, float, character and index.

Figure 4: Structure of Acquired Data Adopted from Fry (2003)

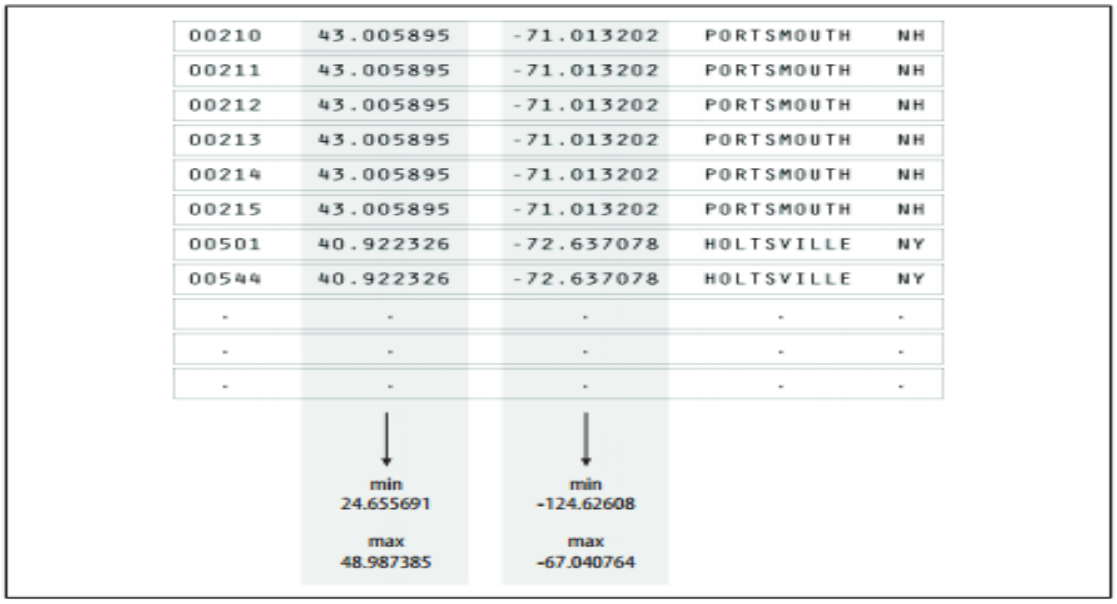


Filter

Fry (2003) stated that this stage involves cleansing the data – identifying incomplete, incorrect, inaccurate, and irrelevant parts of the data and replacing, modifying, or deleting the dirty data. Hence this is otherwise known as the data cleansing stage. This process makes datasets consistent with other similar datasets in the system, of which data inconsistencies may have been as a result of user entry errors, or corruption in data transmission or storage. As regards to the example used, Fry decided to focus on 48 states in the U.S. so as to reduce the amount of mathematical work needed to analyse the data.

Mine

Figure 5: Mining the Data: Comparing Minimum and Maximum Values

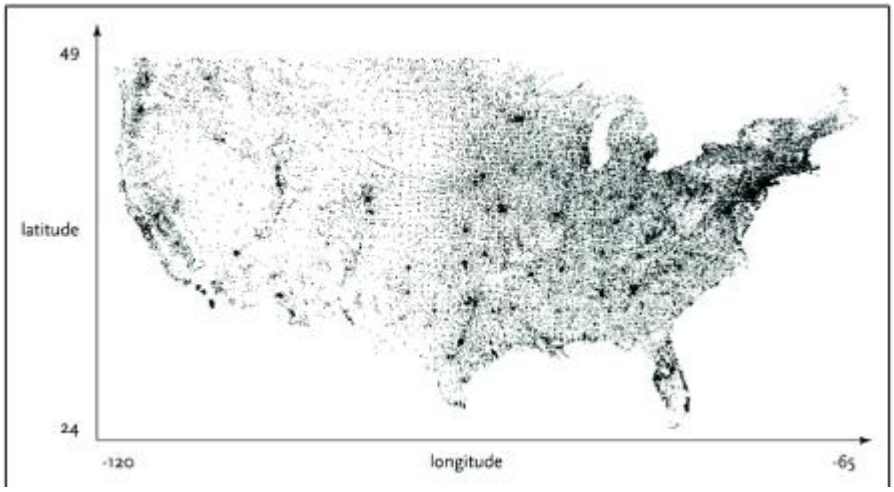


In this phase, simple or complex data analysis methods, involving maths, statistics and data mining are carried out. The data analysis method chosen depends on whether the data is quantitative, qualitative or both. Oates (2006) describes quantitative data as "*data, or evidence, based on numbers…generated by experiments, surveys and other research strategies*", where as qualitative data are "…*all non-numeric data- words, images, sounds, …found in interview tapes, researchers' diaries, company documents, and websites*". Seeing as the data used in the example are mostly quantitative, a simple data analysis was used – figuring the lowest and highest longitude and latitude. This is shown in figure 5.

Represent

The analysed data or rather results from the analysis are represented visually in this stage. The kind of visualization technique used depends on the goal you want to achieve and the kind of data. Fry (2003) points out that represent phase of the overall process is a fundamental challenge as the researcher need to make the most important decision of how to present the data, because the method used in representing the data can influence changes in data acquired and the data filtered. In this example, Fry mapped the minimum and maximum longitude and latitude in a two-dimensional plot. This is shown in figure 6.

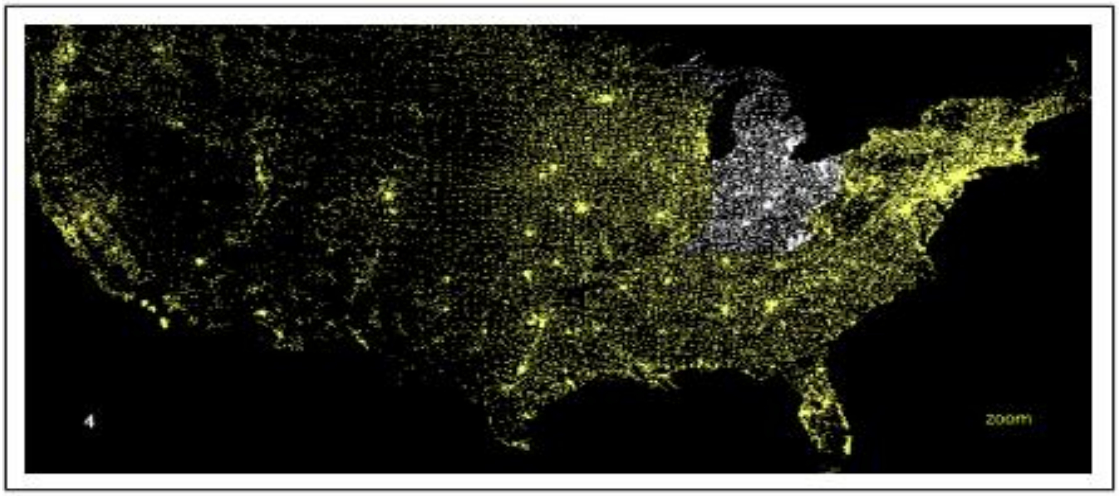
Figure 6: Basic Visual Representation of Zip Data Adopted from Fry (2003)



Refine

Fry (2003) states that colors should be used to call more attention to a specified data thereby establishing hierarchy, making it easier to gain insights into the data and view patterns. The background was colored deep gray and the part of the information viewed was in white. While the area not viewed was in yellow shown in the figure below.

Figure 7: Using Color to Refine Representation

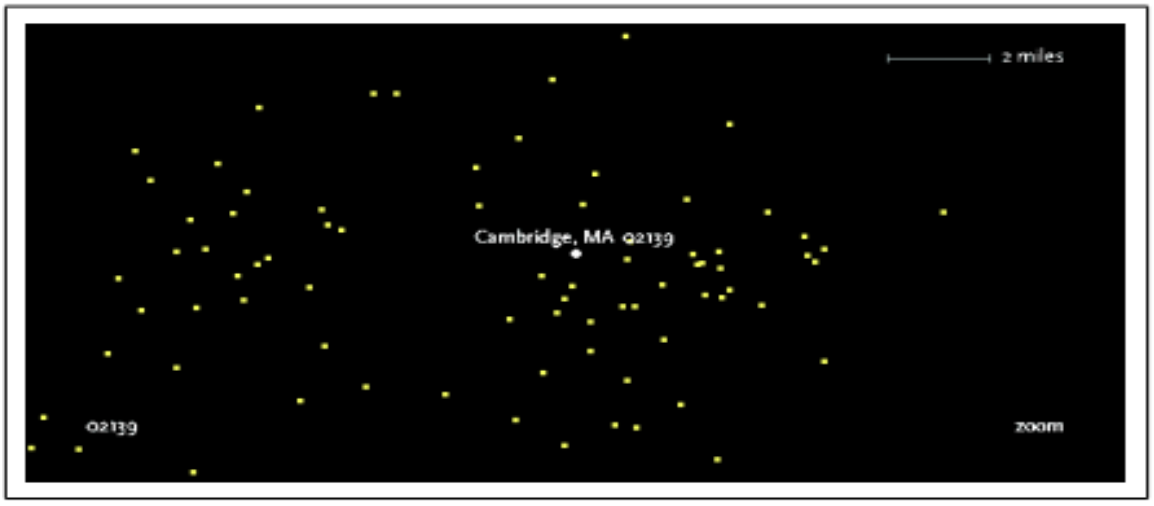


Interact

This is when interaction is added to the visual designs to enable the user to have control or explore the data. It may be in form of selecting a part of the data made possible by visual analytic tools. It is used to present and manage information that is large in size, complex and requires both human and machine analysis. The main goal of Visual analytics research is to transform information into an opportunity (Keim et al., 2006).

Fry (2003) states that this step can influence changes in the refinement stage as changes in the selected part of the data may result to the data been designed differently. Figure 8 shows an area that was selected by a user - by typing a zip code; the respected area on the map is viewed.

Figure 8: User Entering the Full Zip Code(021369) Adopted from Fry (2003)



The following section provides an in depth study on visual analytics.

2.5 VISUAL ANALYTICS

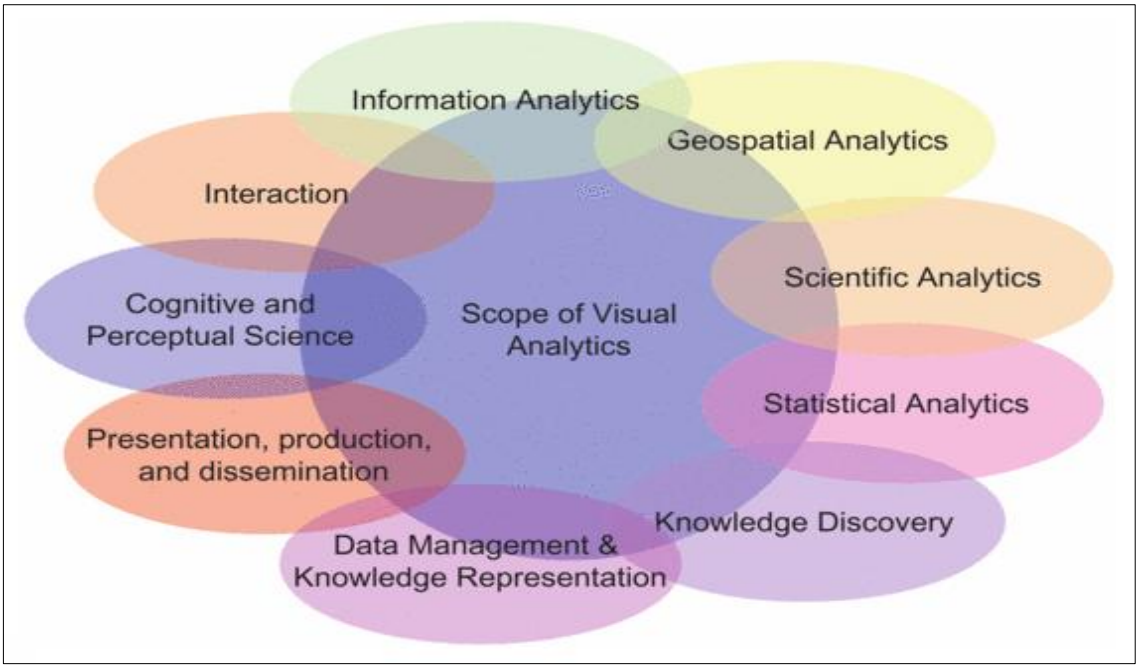
Visual analytics is described as "*an outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces*"(Wong and Thomas, 2004). Of concern here, is information visualization. It is an iterative process that involves the seven stages of visualizing data. To be more precise, it involves collecting data from heterogeneous sources, analysing it, visually representing the information, interacting with it and making better decisions.

Visual analytics has some overlapping goals and techniques with information visualization, however, visual analytics defers from information visualization as an integrated approach joining information visualization (handles abstract data structures) techniques with statistical transformation and data analysis techniques. Nonetheless, information visualization acts as the part of the interface between the user and the computer system, as a result, increasing human cognitive capabilities. There are various visual analytic tools available for free, such as Tableau public, and R. The tool used by the author for this dissertation will be described in the methods section. Figure 9 shows the detailed scope of visual analytics adapted from Keim et al. paper on the challenges of visual Analytics (2006).

Visual Analytics combines techniques from information analytics, geospatial analytics, scientific analytics and statistical analytics. Most especially cognitive

and perceptual science which deals with human factors play a significant role in communicating visual representations, as well as support decision-making process (Keim et al., 2006).

Figure 9: Visual Analytics as a Highly Interdisciplinary Field of Research



2.6 HOW DO YOU GAIN INSIGHTS FROM VISUALIZING DATA

Insight has been stated as the main purpose of visualizing data by many researchers, however, North (2006) states that the definition of insight, is very difficult and challenging, either as a result of other researchers finding their description to be too restrictive to capture its value or too ambiguous to be helpful. Nonetheless, he identified essential characteristics that are useful in gaining an understanding of insights/knowledge (insight and knowledge will be used interchangeably from this point forward). Therefore, insight should be complex, deep, qualitative, unexpected, and relevant.

To understand how to gain insights, Yi et al. suggests that there are paths to be followed. They identified four distinctive processes through which knowledge can be gained for novel findings, These are " *1) Provide Overview, 2) Adjust, 3) Detect Pattern, and 4) Match Mental Model*"(2008).

The researcher points out that the aforementioned distinctive procedures identified by Yi et al(2008) can be achieved by following the seven stages of

visualizing data suggested by Fry(2003), hence merging the both ideas. These different processes are described below.

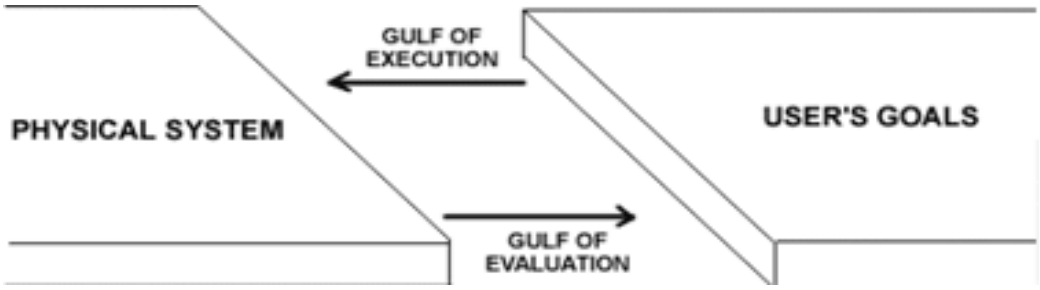
The representing stage of visualizing data suggested by Fry(2003) is a process where knowledge can be gained when providing an overview of the big picture of dataset, which encourages further exploration of the part of the dataset that needs to be investigated more. As argued by Fry (2003), this representing stage is the most important stage of the process and an example is shown in figure 5 above.

Another way to gain insights is through adjusting the range of selection when exploring the dataset. It is difficult to achieve significant knowledge from a large, unstructured dataset. This enables the beneficiaries to make sense of various parts of the visualized large amount of data by employing the use of a filtering interaction technique. Gonzalez and Kobsa stated that "*data analysts appreciated being able to quickly include or exclude data for visualization, so they could concentrate on certain portion of data*"(2003). This can be achieved in the interaction stage proposed by Fry (2003).

The third process of gaining knowledge, which is one of the aims of this dissertation, is to detect patterns. This can be achieved in the represent, refine and interact stages of visualizing data(Fry, 2003). However, patterns or trends can be seen more clearly in the interact stage. It includes the detection of specific relationships, distributions or outliers that could lead to new knowledge been discovered or further questions and hypotheses been asked. In this case, finding relationships is the one of the main objectives of this dissertation. Card et al. (1999) argues that "*visually organizing data by structural relationships (e.g., by time) enhances patterns*.", Also they go on to state that "*visualization can allow for the monitoring of a large number of potential events if the display is organized so that these stand out by appearance or motion*.". Finding trends in visualized information is common with research studies that use the exploratory data analysis approach (This will described more in detail in the methods chapter). Gonsalez and Kobsa stated that finding novel connections, patterns and outliers are one of the benefits that are attained when visually exploring the data (2003).

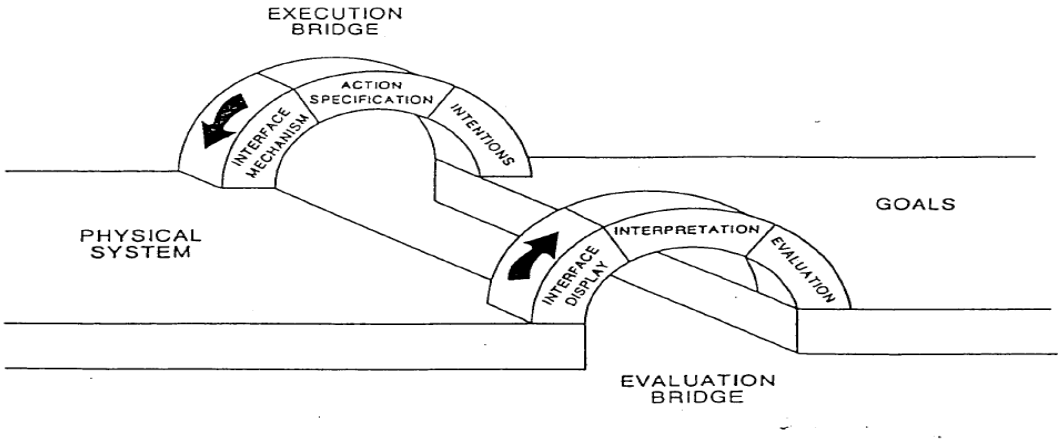
The final procedure of how knowledge can be attained is by matching people's mental model to the visual representations of the datasets. Yi et al. argues that "*One of the benefits of information visualization is that a visual representation of data can decrease the gap between the data and user's mental model of it, thereby reducing cognitive load in understanding, amplifying human recognition of familiar presences, and linking the presented visual information with real-world knowledge*."(2008). The gap that exists between the user's mind and the visual representation is described as the gulf of execution and gulf of evaluation(Hutchins, Hollan and Norman, 1985; Norman, 1986). The gulf of execution describes the degree to which the interaction conceivable outcomes of an artefact, systems interface or likewise matches what an individual perceives is the likely outcome of interacting with the object. On the other hand the gulf of evaluation describes the degree to which the artefact or system give outcomes that could be straightforwardly seen and translated regarding the perception of the user. In other words, it refers to the distance from the visual interface to the user (Norman, 2013). This is illustrated in the figure below adapted from Norman (1986)

Figure 10: The Gulf of Execution and Evaluation Adopted from Norman (1986)



In essence, to match your perceptions with your outcomes, these gulfs must be bridged – that is, the time taking to match your perceptions with the outcomes in both directions must be reduced. This is illustrated below, adapted from Norman (1986). This helps in reducing the cognitive effort required to understand the outcomes.

Figure 11: Bridging the Gulfs of Execution and Evaluation



In the context of this dissertation, the author is concerned with bridging the gulfs to recognizing and understanding patterns through visual representations. Hence the shorter the gap, the easier for trends to be spotted. This can be achieved in the refine and interact stages of visualization.

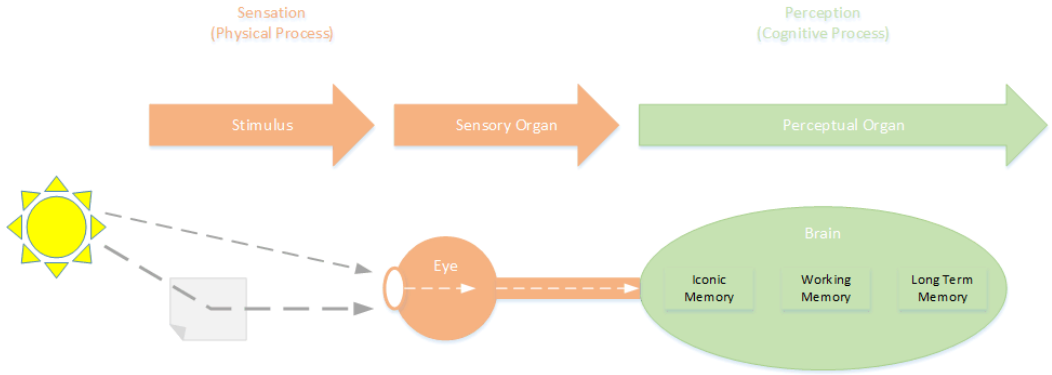
2.7 VISUAL PERCEPTION AND GRAPHICAL COMMUNICATION

Few stated that "*because graphical information is visual, it must express information in ways that human eyes can perceive and brains can understand*"(2012, p.61). There are various graphical means of presenting quantitative information such as line graphs, bar graphs, and maps. To be able to perceive patterns easily, the graphs have to be presented in ways that they are visible. This can be done through the use of the right graphs and colors (Ware, 2004; Few, 2012).

Ware went on to state that "We *can easily see patterns presented in certain ways, but if they are presented in other ways, they become invisible*…" (2004, p. xxi). Understanding what we see starts with the sensation in our eyes and gradually progresses to the perceptual organ (brain) (Few, 2012). The sequence of stages involved is illustrated in the figure 12

The visual image (represented by the sun) is seen by the eyes responsible for absorbing and translating what we see to the perceptual organ (brain). The brain then converts what we see into meaningful information, "*which is where images are actually perceived…and produces what we experience as visual perceptio*n"(Few, 2012, p. 65).

Figure 12: Mechanics Of Sight Adopted From Few (2012, P. 65)



Few (2010, p.65) points out that the process of converting visual data into information requires three types of memory - iconic memory, working memory, and long-term memory. Iconic memory registers the image that has just been seen for a very short time(less than a second) before it fades away into working memory. The image is then transformed into meaningful information and stored for a short time (few seconds to a few hours) before fading away into long term memory.

2.7.1 PREATTENTIVE PROCESSING OF VISUAL IMAGES

In order for the eyes to recognize visual designs (iconic memory) effectively, Ware stated that it has to contain some preattentive attributes. This was grouped into four categories; form, color, motion, and spatial position(2004, pp. 151–152). In the context of this project, form and color attributes were focused on.

- Form
  - Length
  - Width
  - Size
- Color
  - Hue
  - Intensity

According to Ware, the form of a graph determines how it is perceived. For instance, bars in a bar graph uses length, width, and size to represent quantitative values. To make similar set of bars to be seen as a group, he also suggests the use of color attributes.

2.8 GOALS, VALUE AND CHALLENGES OF ANALYSING VISUAL DATA

Card et al stated that the main objective of visualizing large datasets is to open up human cognition; to view patterns, trends and inconsistencies easily, and as a result gain knowledge and enhance decision making (1999). To be more specific, there are three significant goals of visualizing data according to Keim et al. (2006); presentation, confirmatory analysis, and exploratory analysis. In the case of presentation, the researcher is concerned with communicating the information of an analysis in a fixed format effectively and efficiently, of which the kind of visual design depends largely on the researcher. In confirmatory analysis, the researcher is focused on proving if his/her hypothesis is right or wrong through the use of visualization techniques, therefore, it is a goal-oriented research. Finally, in exploratory analysis, the researcher begins by having a vague idea of what to achieve, but depending on the visual analysis to find useful information, such as patterns, trends and possibly novel results. According to keim et al., it is a difficult task.

However, there are challenges involved when analysing visual data. According to Keim et al. the "…*ability to collect and store data is growing faster than our ability to analyse it*"(2006). In the context of this dissertation, the main challenges would be synthesizing heterogeneous types of data from heterogeneous information sources to carry out effective analysis. An example of an application domain is computational biology - sometimes referred to as bioinformatics, that requires the integration of "…*real-valued gene expression data, functional annotation of genes, genotyping information, a graph of interacting proteins, equations describing the dynamics of a system, localization of proteins in a cell, and natural language text*…"(Keim et al., 2006). Keim et al., goes further to state that combining these datasets from multiple sources with different data types, is one of the primary problems in decision theory, information theory, and statistics, which clearly pose a challenge in visual analytics.

The other vital challenge in visualizing data is the ability for the intended audience to interpret or recognize and understand the information presented. According to Keim et al., this is the biggest challenge in visual analytics (2006). It all depends on the quality of the used data and methods. Most often than

not, the raw data are sourced with missing values, and wrong data amongst other problems. Hence, pre-processing the data, such as data reduction, data cleaning, data migration and parsing, aggregation and combining etc. is essential. However, pre-processing the data still poses quality problem because it is a challenge to minimize these errors and to provide a flexible visual design to cope with them.

In conclusion, visually analysing data comes with a lot of challenges but that does not conceal the fact that visual analytic techniques are needed today. Visual analytic techniques are needed in many sectors, such as environmental sciences, geo sciences, social sciences, engineering and economics according to Keim (2010). Keim goes further to state that "*visual analytics techniques are essential to deal with data sets which are growing fast in size and complexity to gain understanding, to discover patterns, and to optimize and steer complicated processes*."(2010).

CHAPTER 3: METHODS

This project was aimed at exploring the temporal relationship between out-of-school crimes and student educational attainment from different spatial coverage locations, through the use of quantitative and visual analysis methods, and also to find if a novel pattern or trend exists.

To fulfil the project objectives, thus answering the research question, Oates (2006) suggests that a research strategy must be adopted. In the context of this research, the experiment strategy was adopted. This approach is defined by Oates as a "*strategy that investigates cause and effect relationships, seeking to prove or disprove a causal link between a factor and an observed outcome…*" (2006). I experimentally examined and visualized the extent to which student GCSE grades from each London borough are influenced by or influences the crime rates of their respective borough.

This chapter provides detailed description of how the research objectives were carried out. It presents the methodology adopted and a detailed description of the methods used in gathering, analysing and visualizing the data. It also describes the method used for evaluating the project outcomes.

3.1 METHODOLOGY

In order to adopt the best methodology suitable for carrying out the activities necessary for achieving the project objectives, the first step was to have a broad and thorough understanding of the project topic area, the different methodology techniques, and the project aims and objective. Much of this activity was carried out during the literature review phase. It was then decided to adopt the process of visualizing data (using an exploratory data analysis approach) described as the "seven stages of visualizing data" by Ben Fry (2003) to achieve the objectives. This was as a result of the author needing a methodology that encompassed the data gathering phase, data analysis phase, and visual analysis phase, of which the chosen methodology provides. Please refer to section 2.4.1 for a review of the process.

The following sections provide the detailed descriptions of how each stage was applied and carried out by the researcher, as well as how the proposed

framework of methods used in analysing both data sets was created. Issues encountered (if any), limitations or adjustments made were also stated.

3.2 ACQUIRING OF DATA

This stage was carried on the 21th July, 2014. The researcher made use of open government data. A Nominet Trust report, *Open Data and Charities* (Hall et al., 2012) defines datasets as "open data" if it meets the following three criteria:

> *"It is made accessible online.*
>
> *It is published in an open machine readable format.*
>
> *It is licensed to allow others to re-use it."* (Hall et al., 2012)

As of late, the British government appears to be starting to grasp the thought of open data. Websites such as **data.gov.uk** provide publicised datasets on diverse topics such as educational attainment statistics, household income, crime statistics, road safety and hospital finance. Their aim is to "*help people understand how the government works and how policies are made*" (Data.gov.uk, 2014). Davies (2010) points out that exploring open government data "*can give individuals greater understanding of the state and direct access to facts and information can empower individuals in their interactions with the state*".

Public datasets can be assessed in various formats and analysed in different number of ways. These datasets are usually retrieved in the form of a CSV (comma delimited), or XLS (Ms Excel Spreadsheet) file. They are then analysed using tools such as (but not limited to) SPSS, SAS, and R, and also applications such as (not limited to) Tableau and D3 for visualizing the outputs.

For this particular research, public data on the statistics of crime and education achievement was already available via the data.london.gov.uk (London Datastore) website. To be precise, temporal data on the crime rates and GCSE results by location of pupil residence in London. Both datasets were identified as having relevant data to inform the project. Oates describes this data gathering or generation method as *Documents – "documents that already exist prior to the research…"* (2006).

Both datasets were retrieved from the London data store in Microsoft Excel spreadsheet format and are mostly made up of quantitative data (data based on numbers). The links to each dataset are http://data.london.gov.uk/datastore/package/crime-rates-borough and http://data.london.gov.uk/datastore/package/gcse-results-location-pupil-residence-borough . This website is provided by the Greater London Authority (GLA – a recognized administrative body for Greater London) with their sole aim to allow "*citizens to be able to access the data that the GLA and other public sector organisations hold, and to use data however they see fit – free of charge…Raw data often doesn't tell you anything until it has been presented in a meaningful way. We want to encourage the masses of technical talent that we have in London to transfer rows of text and numbers into apps, websites or mobile products which people can actually find useful*." (Data.london.gov.uk, 2014). A detailed description of each datasets in its original layout is discussed in chapter 5.

3.2 PARSING DATA

This stage of the research was carried out on the 21st July, 2014. Not much was done on parsing both datasets as they were retrieved in an already understandable structure. However, in an effort to understand the data types contained in each datasets, the author (adopting the method used by Fry (2003) in this stage) designated suitable data types to each piece of the data, in other words tagging each part of the data for its intended use and consequently making it more useful to the data analysis tool that will manipulate or represent it in some form. One thing of note is the change of data type of the variables of the different datasets to "float" as a result of the standardization process performed in the filtering stage.

3.3 FILTERING DATASETS

In this stage(performed between the 28th and 29th July, 2014) , the author removed parts of the data which was thought of as not been useful in accomplishing the project objectives, as suggested by Fry (2003). This was one of the intricate stages carried out by the researcher. The basis for this research was focused on two data sets that were found to be inconsistent for data analysis, as regards to the number of variables, units, different time periods covered, and empty data contained in a dataset. The empty data were

denoted with 'x' as shown for the variable *Boy Pupils at the End of KS4 Achieving 5+ A*-C Including English and Mathematics* for the borough of Kensington and Chelsea in the year 2012-13. It was suggested by the project supervisor to find the mean of the other values for the respected variable and substitute the missing value with the average. However, the researcher found over 15 missing values for the selected time period for all variables.

Hence, the author found that using this proposed method will artificially create average data points, which may change the values of correlation and possibly contain biased results. In that respect, the author made a decision to keep the variables that were needed for this research. These variables are *All Pupils at the End of KS4 Achieving 5+ A* - C Including English and Mathematics* and *All Pupils at the End of KS4 Achieving 5+ A* - G Including English and Mathematics.* This is otherwise similar to Listwise deletion of missing data. Listwise deletion and Pairwise deletion are the most common techniques for handling missing data (Peugh and Enders, 2004, p.525). Peugh and Enders went ahead to describe Listwise deletion as a process which removes all the data for a case that has one or more missing values, whereas Pairwise deletion attempts to reduce the number of data lost from Listwise deletion by carrying out a correlation matrix and selecting pairs of variables for which data is available – this is on an analysis-by-analysis basis(2004, p. 528). Although Little and Rubin (1987) do not generally recommend these traditional approaches, the researcher found the Listwise deletion method to be appropriate in the context of this project. The subsequent paragraphs explain the reason.

All pupils are required to have completed year 10 and 11, otherwise known as Fourth Form or Fifth Form of secondary school in England to qualify for GCSE. This is referred to as Key Stage 4, which is the legal term for two years of secondary school education that incorporates GCSEs, and other examinations, and pupils are of ages between 14 and 16.

The aforementioned variables were chosen (of which there was no missing data) because at the end of the two-year GCSE course, candidates receive a grade for each subject that they have written. The pass grades, from highest to lowest, are: A*(pronounced 'A-star'), A, B, C, D, E, F and G. A student that achieved nothing deserving of credit, gets awarded with a grade U

(ungraded/unclassified), therefore no GCSE is giving to the student in that subject. A GCSE at grades D-G is a level 1 qualification, while a GCSE at A*-C is a level 2 qualification. As expected, level 2 qualifications are much more desirable and insisted by employers and educational institutions. These variables were changed to "LevelTwo" and "LevelOne" correspondingly to make it less wordy. Additionally, the author needed the general statistic for all pupils that took GCSE exams based on their respective residence and not by their gender for a selected time period. Therefore, there was not any need to analyse the other indicators.

Data for the time period of 2007 to 2013 were retained, as well as data for the 32 London Boroughs within the Greater London area for each datasets. These boroughs vary in sizes illustrated in the figure below.

Figure 13: The 32 London Boroughs in the study area. With exception of 'city of London' (no crime rates recorded by the metropolitan police and not part of the London borough of greater London)

The time period was selected because the variables needed from the GCSE results by location of pupil residence were consistent from the year 2007 to 2013. In conclusion, a total of nine and two independent variables were derived from the crime rates and GRLPR dataset respectively.

3.3.1 NORMALIZING DATASETS

In order to visually identify patterns of correlation between the crime rates and GRLPR dataset, the data needed to be standardized to the same format. The process was needed because both datasets were multivariate; the variables derived from the crime rates dataset were in rates per thousand populations (using the mid-year population estimates), where as the GCSE result indicators were in percentages. Additionally, the standard deviation of each variable for a particular year, were very different. For instance, the standard deviation for the variable "TheftandHandling" was found to be 27.7 compared to that of "SexualOffences", which was 0.4, and similarly, "LevelTwo" qualifications was 7.88 and "LevelOne" was 1.82, for the year 2008, calculated using Ms Excel.

In that manner, the author needed to standardise the each variable for comparison, in order for the data to have standard deviation of 1 and a standard mean of 0 as suggested by Coghlan(2014). Everitt also stated that "*…the solution most often suggested to deal with the problem of different units of measurement is to simply standardize each variable to unit variance prior to any analysis*" (2011).This process is seen as one of the primary challenges faced by researchers working with multiple datasets according to Keim et al(2006).

Stolcke et al stated that feature scaling normalization method also called unity-based normalization should be used to bring the data value dimensions into a standardized and comparable range of independent variables and is generally performed during the data pre-processing step (2008). Feature scaling has three different methods - rescaling, standardization, and scaling to unit length. The author adopted the standardization method (also known as Standard score or Z-score) "*…that uses the same metric as other standard scores so they can be compared to one another easily*"(Salkind and Rasmussen, 2007), to normalize each dataset to the same scale for the purpose of data analysis. Everitt also suggested this method for use in multivariate or

cluster analysis (2011,). Stolke et al (2008) mathematically described the standard score formula as:

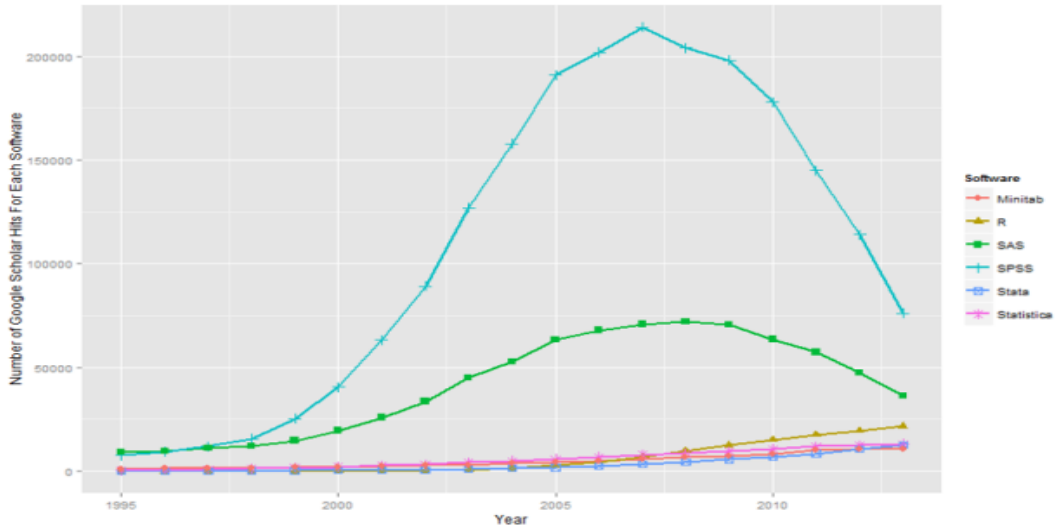
$$x_i' = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$$

, where x_i is an original value of the ith feature, $\bar{x}_i$ is the feature average, σ_{x_i} is the standard deviation, and x_i' the standardize feature value.

3.3.2 DATA ANALYSIS TOOL

The standardization process was carried out using a data analysis and visualization programming language known as R (Ihaka and Gentleman, 1996) in the RStudio tool environment . It is described by Fox and Andersen as a "*…free, cooperatively developed, open-source implementation of S, a powerful and flexible statistical programming language and computing environment that has become the effective standard among statisticians*" (2005). R usage increased significantly over the years (Smith, 2012; Rexer, Allen and Gearan, 2011; Muenchen, 2014). In his report "*The Popularity of Data Analysis software",* Muenchen (2014) demonstrated the rapid growth of R in the number of scholarly articles published which according to him, means that R application is rising accordingly. This he found to be different for other popular data analysis tools such as SPSS, and SAS. Figure 14, adopted from Muenchen (2012) report illustrate the difference between the number of scholarly articles found for the top five data analysis tool from 1995 to 2013.

Figure 14: Number of scholarly articles found for the top five classic statistics packages (adopted from Muenchen (2012))



As shown in the figure above, SPSS had a dominant steady increase up until the year 2007, where there was a sharp decline afterwards. SAS on the other hand had a steady increase (not comparable to SPSS) up until the year 2008. R is shown to have a steady increase that may likely surpass that of SPSS and SAS in the future. Hence, part of reason for choosing to use this tool for data analysis.

To perform the standardization process in R, the "scale()" function was called, according to Coghlan (2014), which performs the formula above. An example of the function called to standardise the crime rates dataset for the year 2008 is

```
crimeStandard08 <- as.data.frame(scale(crimeSorted[1:32,3:10]))
```

Where as.data.frame is the command used in changing the layout of the dataset, crimeSorted[1:32,3:10] is the dataset name and required column index of the data set sent to the scale() function, and crimeStandard08 is the vector variable created to store the standardised values.

The standardised datasets were further sorted to have a panel-like data structure.

Please refer to appendix B1 for the detailed commands used in performing the standardization process for both data sets.

The milestone for the filtering stage was reached at this point and the data mining stage began.

3.4 MINING DATA

This stage began on the 30th July, 2014. The author carried out over half the percentage of data analysis needed, for the effective visualization and communication of the degree of correlation, with the possibility of recognizing novel patterns.

The author made use of data analysis and visualization software RStudio, to analyse the degree of correlation between each dataset. As stated earlier, nine and two independent variables were derived from the crime rates and GRLPR datasets respectively. As a result, visualizing and identifying the pattern of correlation between the two datasets over time would likely produce a very busy graph. The researcher needed a method that would reduce the number of variables for each data set, at the same time capture the important features with smallest amount of information lost. Principal component analysis(PCA) was the preferred method selected because it " *…provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it*"(Shlens, 2014). Mankin also summarised the main goals of applying PCA as the ability to

> *"identify how different variables work together to create the dynamics of the system*
>
> *reduce the dimensionality of the data*
>
> *decrease redundancy in the data*
>
> *filter some of the noise in the data*
>
> *compress the data*
>
> *prepare the data for further analysis using other techniques" (2010)*.

In the context of this research, PCA was used as an intermediate step in the data mining stage to prepare the data for visual analysis. Mankin (2010) also pointed out that in order to apply PCA, the data have to abide to the two PCA principles-that are, variables should be highly correlated for an accurate PCA and the most significant PCA value must be one with the highest variance. Furthermore, she argued that a high correlation between variables meant it

contained redundant data. Hence, PCA transforms highly correlated indicators into a set of linearly uncorrelated indicators.

The computation of PCA was textually described as extracting orthogonal factors that sums to a maximum variance rotation of the original variable dimension, of which the first factor (variable) capture the maximum variance and subsequent factors capture variance lower than the previous factor, and so on (Statsoft.com, 2014) . These following factors are uncorrelated to each other because the remaining variability not caught by the preceding factor is calculated (Statsoft.com, 2014).

The researcher carried out PCA adopting the method specified by Coghlan in his booklet titled "*A Little book of R for Multivariate Analysis*" (2014). This decision was made because she gave an accurate, understandable and detailed description, in comparison to that of Mankin (2010), and also because R programming language was used.

As a first step to mining the data, Coghlan (2014) advices that the unstandardised data is plotted to see the degree of correlations between each combination of variables and are certain PCA is needed. For this reason, a scatter plot was plotted for each pair of crime rates variables and GRLPR variables by year. The functions and packages available through the RStudio and needed to plot the scatter diagrams are shown in the table below.

Table 1:Functions and Package used to plot the scatter diagrams

Package	**What is it used for**
• Gclus	• To cluster variables that has a high correlation towards the principal diagonal. It also color codes the correlation depending on the degree, with pink and lighter boxes for strong and weaker correlations respectively.
Functions	**What are they used for**
• dmat.color() • order.single()	• It accepts a dissimilar matrix and returns different colors. • It reorders objects so that similar objects are grouped together towards the principal diagonal.

• cpairs()	• It draws a scatter plot matrix,

Calculating the correlation coefficient was the second step carried out by the author to find the exact correlation coefficient. Pearson correlation coefficient - developed by Karl Pearson, which is a measure of the linear correlation between two variables (Few, 2012) was used. It calculates the correlation coefficient where values fall between +1 and -1. According to Few, "*…a value of 0 indicates that there is no linear correlation…a value of +1 indicates that there is a perfect positive linear correlation…a value of -1 indicates that there is a perfect negative linear correlation…the greater the value, either positive or negative, the stronger the linear correlation*" (2012). It is mathematically calculated using the equation below(Kline, 1994, p. 19)

$$r = \frac{N\sum X_1 X_2 - \sum X_1 X_2}{[N\sum X_1^2 - (\sum X_1)^2]^{1/2}\ [N\sum X_2^2 - (\sum X_2)^2]^{1/2}}$$

The "*mosthighlycorrelated()*" function written by Coghlan (2014) described below, was adopted in calculating the correlation values between variables in the crime rates dataset because it shows the variables in descending order of their correlated values.

"
```
> mosthighlycorrelated <- function(mydataframe,numtoreport)
{
# find the correlations
cormatrix <- cor(mydataframe)
# set the correlations on the diagonal or lower triangle to
zero,
# so they will not be reported as the highest ones:
diag(cormatrix) <- 0
cormatrix[lower.tri(cormatrix)] <- 0
# flatten the matrix into a dataframe for easy sorting
fm <- as.data.frame(as.table(cormatrix))
# assign human-friendly names
names(fm) <- c("First.Variable",
"Second.Variable","Correlation")
# sort and print the top n correlations
head(fm[order(abs(fm$Correlation),decreasing=T),],n=numtoreport
)
```

}"

Whilst *"cor.test()"* function was used for deriving the coefficients for the GRLPR dataset by year.

Please refer to appendix B2 for the detailed commands used for performing pearson's correlation and plotting the scatter diagrams for both datasets.

The researcher then carried out the PCA on both datasets still applying Coghlan's process as a guideline. The function *"prcomp()"* was called to find the components for each dataset. However, seeing as both datasets were temporal, the author derived factors for each time period.

The number of principal components extracted was based on two decisive factors, the scree test (Cattell, 1966) and the Guttman-Kaiser criterion. The scree test is a type of plot that graphically illustrates the factors to be retained. The factors preceding the most evident slope was selected as suggested by Cartell (1966). The Guttman-Kaiser criterion states that the principal component(s) eigenvalues that is/are above 1.0 be retained as it shows that the factor(s) encapsulate most of the total variance amongst a number of variables (Yeomans and Golder, 1982). Eigenvalues represents the sum of variance for each component and it is derived by squaring the total of the factor loadings (the degree of "…*correlations of a variable to a factor"(*Kline,1994)) for each principal component (Kline,1994). This technique was originally proposed by Guttman (1954) but made popular by Kaiser(1960, 1961). Although both techniques can be distinctively used to make a decision on how many components to retain, the researcher employed both techniques to be certain of the consistency.

As a final step in the data mining stage, the author calculated the factor loadings to find the variable that was highly correlated to the component and the newly derived values of the selected principal component for each borough and year. This was achieved by applying the function below written by Coghlan (2014)

```
"> calcpc <- function(variables,loadings)
{
# find the number of samples in the data set
as.data.frame(variables)
```

```
numsamples <- nrow(variables)
# make a vector to store the component
pc <- numeric(numsamples)
# find the number of variables
numvariables <- length(variables)
# calculate the value of the component for each sample
for (i in 1:numsamples)
{
valuei <- 0
for (j in 1:numvariables)
{
valueij <- variables[i,j]
loadingj <- loadings[j]
valuei <- valuei + (valueij * loadingj)
}
pc[i] <- valuei
}
return(pc)
}"
```

Please refer to appendix B3 for the detailed commands used in performing PCA (calculating the eigenvalues, the factor loadings, and the PCA scores).

3.5 REPRESENT ANALYSED DATA SET

In this stage (began on the 4th September, 2014), the decision on the best medium to communicate the pattern of correlation was made. It was stated earlier that it is one of the most difficult decision to make in a visualization project. Few justified this by pointing out that "*choosing whether to display data in one or more tables, one or more graphs, or some combination of the two, is a fundamental challenge of data presentation*"(2012, p.42).

To begin with, the researcher went back to the mining stage to analyse the correlation of the newly reduced variables gotten from applying PCA. This was done by designing a scatter plot using Tableau Public Software. Which is a "*…groundbreaking data visualization software created by Tableau Software…allows for instantaneous insight by transforming data into visually appealing, interactive visualizations…*" (InterWorks, 2014). The process of designing the scatter plot, involved represent, refine and interact stages proposed by Fry (2003). The researcher then proceeded with the process of

visually analysing the data. Following the discussion with the project supervisor and gaining knowledge from relevant literature(Tufte, 1983; Ware, 2004; Few, 2012), the researcher initially selected two different graphs for communicating the pre-analysed dataset – a bar graph and line graph, in Tableau Public environment, for the following reasons.

Table 2: Mediums of Communication

Bar Graph	Line Graph
• To compare patterns over time for each borough and datasets • To group similar boroughs with similar patterns for both datasets	• To see the overall trends of both datasets for each borough, for a particular year • To compare trends for each year

However, the researcher found the combination of both graphs to be redundant in visually analysing the data. It was then decided to make use of the bar graph as it was found to be sufficient enough.

3.6 REFINE VISUALIZED DATA SET

The decision on the colors to use for making patterns readily perceivable was carried out in this stage (30th August, 2014). Light shades of colors were chosen to be soothing and easy on the eyes (Tufte, 1990, p. 90).

3.7 INTERACT WITH REFINED DATA SET

Seeing as the data sets deal with 32 London boroughs for each year from 2008-2013 , it was a difficult task to compare both patterns from each datasets, hence, the bar graphs were sorted by boroughs, where as the line graph by year. This was done using the filter interactive option on Tableau Public. The bar graphs were later grouped by boroughs having similar patterns.

CHAPTER 4: RESULTS

This chapter outlines the results derived from adopting the methodology and applying the methods described above. The following section provides the detailed description of datasets gathered as well as issues encountered (if any), limitations, or adjustments made.

4.1 DESCRIPTION OF DATA GATHERED

In this research, I look at annual crime rates (per thousand populations) publicly provided for free by the metropolitan police in London and the GCSE results by location of pupil residence (GRLPR) in London, both retrieved from the London Datastore.

The crime rates data includes data of the rates per thousand population of both property and violent crime types, as well as all recorded and other notifiable offences from the period of 1999 to 2014. These variables are categorised according to their geospatial location and time period. The following section lists the types of crime recorded by the metropolitan police by their groups.

4.1.1 PROPERTY CRIME

The metropolitan police and the Home Office recognize the following as property crimes and their description are stated.

- *Burglary*: This is theft, or attempted theft, from a building or premises where access is not approved. Damage to a building or premises that seems to have been brought about by an individual attempting to enter to commit a burglary, is also considered as burglary.
- *Theft and Handling*: This is otherwise known as Handling Stolen Goods. This is when a purchase from a thief is made or merely being in possession of stolen goods.
- *Fraud or Forgery*: It is a deception deliberately practiced in order to secure unfair or unlawful gain.
- *Criminal Damage*: This is described in the Criminal Damage Act (1965), section 1(1) as "*A person who without lawful excuse destroys or damages any property belonging to another intending to destroy or damage any such property or being reckless as to whether any such property would be destroyed or damaged shall be guilty of an offence*."

- *Drugs*: It is a term used to describe various offences identified with the utilization of restrained substances.

4.1.2 VIOLENT CRIME

The metropolitan police and the Home Office recognize the following as violent crimes and their descriptions are stated.

- *Violence Against the Person*: This refers to a crime which is committed by direct physical harm or force being applied to another person. It includes a range of offences such as Murder, Harassment, Common Assault, Actual Bodily Harm, Possession of Offensive Weapon and other Violence.
- *Sexual Offences*: The United Kingdom's legislation recognizes Rape, Sexual Assault, and causing a person to engage in sexual activity without consent, among others in the Sexual Offences Act(2003).
- *Robbery*: This is a crime that involves theft with the use of force.

Crime rates of all recorded offences and other notifiable offences, recorded by the metropolitan police do not belong to any group of crime that were aforementioned. All other notifiable offences are the rates of all offences that are statutorily notifiable by the home office. The Metropolitan Police suggests that extreme caution should be taken when comparing crime figures from 2012/13 with earlier years because changes were made in the classification of police recorded crime.

The screenshot of a tiny proportion of the raw datasets retrieved are shown in the table below.

Table 3: Raw Data On Crime Rates Retrieved In The Format Provided By The London Datastore

B	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
	Violence Against the Person										
Borough	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-	2012-	2013-
Barking & Dagenham	29.5	31.5	31.8	30.8	29.3	30.6	29.4	24.5	22.0	19.7	22.0
Barnet	18.1	22.2	19.5	16.7	15.6	15.3	14.0	12.6	11.6	11.2	12.5
Bexley	18.4	18.7	19.7	16.8	14.4	15.3	16.6	13.3	11.7	12.1	11.8
Brent	27.5	34.0	30.5	22.5	20.9	22.3	23.5	23.3	23.3	19.7	20.4
Bromley	16.8	19.4	18.4	18.9	16.7	18.3	17.4	15.7	14.7	13.1	14.1
Camden	31.4	36.9	34.9	31.2	27.4	28.6	27.3	27.4	25.2	22.4	20.6
Croydon	23.6	25.5	22.9	19.8	18.5	19.8	20.0	18.5	18.3	17.7	17.7
Ealing	23.2	25.8	25.2	24.2	24.2	23.6	24.4	24.3	21.0	20.4	18.5
Enfield	18.2	18.4	18.5	18.6	14.5	14.2	15.4	14.4	13.2	13.7	15.0
Greenwich	30.6	34.6	33.5	32.0	30.7	28.1	25.1	21.8	20.8	20.3	20.8
Hackney	33.6	34.1	34.5	32.5	31.4	28.0	28.1	24.6	21.8	22.6	24.0
Hammersmith & Fulham	25.0	26.9	29.1	28.9	28.2	27.9	26.2	27.1	23.8	23.7	22.4
Haringey	24.4	27.9	30.2	24.2	22.7	21.7	20.4	19.4	18.4	18.2	18.6
Harrow	13.4	14.1	13.7	12.8	11.4	13.6	14.8	13.7	11.9	11.9	13.2
Havering	17.4	18.6	18.4	15.9	13.6	13.5	14.5	15.7	13.8	13.8	14.3
Hillingdon	21.1	23.8	24.9	23.2	22.9	23.5	22.1	21.1	19.5	17.3	17.9
Hounslow	32.1	30.9	28.8	24.1	22.9	23.3	23.0	22.8	21.8	20.1	20.1
Islington	36.5	42.1	38.2	33.9	28.4	30.0	29.7	29.6	24.6	25.6	24.3
Kensington & Chelsea	20.4	19.7	19.7	21.8	19.4	19.4	18.4	18.4	17.4	17.3	17.0
Kingston upon Thames	23.0	23.3	21.2	19.5	16.7	14.0	14.4	14.1	13.8	13.5	12.5
Lambeth	34.8	35.8	32.4	29.7	26.9	27.1	26.5	25.8	23.4	22.0	23.6
Lewisham	26.5	31.1	33.2	31.3	32.1	33.0	28.3	25.1	22.6	20.5	22.4
Merton	19.1	20.1	19.2	17.5	18.9	19.1	17.1	14.4	12.6	12.6	12.5
Newham	32.0	31.4	33.2	29.4	29.0	25.9	26.1	23.4	20.7	20.6	21.3
Redbridge	18.9	18.9	15.8	16.9	16.3	15.7	15.9	15.0	13.8	14.5	16.0
Richmond upon Thames	14.1	14.3	12.9	11.6	10.6	11.0	12.1	11.5	10.6	10.8	10.2
Southwark	34.9	36.3	34.6	31.5	32.5	32.2	31.5	27.1	22.9	22.2	21.8
Sutton	17.7	19.6	17.4	16.3	13.9	14.2	14.1	12.0	12.8	13.3	13.1
Tower Hamlets	37.0	37.4	34.9	35.4	29.7	26.2	25.8	25.4	22.7	23.3	25.2
Waltham Forest	26.8	27.7	30.1	26.2	23.0	23.2	23.5	21.5	19.0	19.0	20.5
Wandsworth	19.3	21.2	21.7	19.6	18.0	17.3	17.5	16.1	13.3	13.3	13.7
Westminster	45.3	48.7	42.3	37.8	38.2	39.2	40.7	40.2	37.7	31.4	32.7
Heathrow											
Inner London	30.8	32.9	32.2	29.6	28.0	27.3	26.5	25.0	22.3	21.5	22.0
Outer London	21.6	23.4	22.4	20.4	18.9	19.1	19.0	17.7	16.4	15.7	16.2
Met Police Area	25.2	27.2	26.2	24.0	22.5	22.3	22.0	20.6	18.8	18.0	18.5
England and Wales	18.1	19.5	19.5	19.1	17.4	16.2	15.5	14.5	14.6	10.6	

The GRLPR data set contains GCSE and equivalent results of pupils at the end of Key Stage 4 in maintained schools (state or public schools) in England and who are resident in England 2001 to 2013. The GCSE results are categorised by the local and regional areas of England, and counts, percentages and averages of GCSE result indicators for a time period. The local regions of England are the 32 London Boroughs and City of London, while the regions of England includes the North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East , and South West regions. The data also focuses on England as a whole. The key GCSE results indicators are listed below. Please find the detailed list of indicators in appendix C1.

- All Pupils at the End of KS4 Achieving 5+ A* - C.
- All Pupils at the End of KS4 Achieving 5+ A* - G.
- All Pupils at the End of KS4 Achieving 5+ A* - C including English and Mathematics.

- All Pupils at the End of KS4 Achieving 5+ A* -G including English and Mathematics.

The table below illustrates the structure of the GRLPR dataset.

Table 4: Raw data on GRLPR retrieved in the format provided by the London Datastore

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Area	All Pupils at the End of KS4	All Pupils at the End of KS4 Achieving 5+A*-C	All Pupils at the End of KS4 Achieving 5+A*-G	All Pupils at the End of KS4 Achieving 5+A*-C Including English and Mathem	All Pupils at the End of KS4 Achieving 5+A*-G Including English and Mathem	All Pupils at the End of KS4 With Any Passes	All Pupils at the End of KS4 With No Passes	Average GCSE and Equivalent Point Score Per Pupil at the End of KS4	Average Capped GCSE and Equivalent Point Score Per Pupil at the End of KS4	All Boy Pupils at the End of KS4	Boy Pupils at the End of KS4 Achieving 5+A*-C	Boy Pupils at the End of KS4 Achieving 5+A*-G	Boy Pupils at the End of KS4 Achieving 5+A*-C Including English and Mathem	Boy Pupils at the End of KS4 Achieving 5+A*-G Including English and Mathem	Boy Pupils at the End of KS4 With Any Passes	Boy Pupils at the End of KS4 With No Passes	Average GCSE and Equivalent Point Score Per Boy Pupil at the End of KS4
	Count	rcentage	rcentage	rcentage	rcentage	rcentage	rcentage	Score	Score	Count	rcentage	rcentage	rcentage	rcentage	rcentage	rcentage	Score
City of London	20	75	100	60	100	100	0	417.4	339.5	10	x	100	60	100	100	0	400.6
Barking and Dagenham	2,229	59.3	94.3	41.4	93.2	98.7	1.3	371.9	299.3	1,071	55.1	92.5	37.7	91.3	98.3	1.7	357.9
Barnet	3,130	73.1	94.6	61.7	93.9	98.9	1.1	418.6	326.6	1,572	70.2	94	59.5	93.2	98.6	1.4	406.1
Bexley	2,935	64.2	93.5	51	93.2	98.9	1.1	391	311.8	1,479	60.8	91.2	46.9	90.9	98.5	1.5	379
Brent	2,868	67.6	94.1	53.6	93.1	98.6	1.4	399	318	1,423	62.1	92.5	47.4	91	98.2	1.8	379.7
Bromley	3,287	75.2	94.4	62.7	93.6	98.6	1.4	424.9	330.4	1,677	72.6	93.2	60.1	92.4	98.4	1.6	403.1
Camden	1,183	60.9	93	43.8	91.5	98	2	365.7	302.2	567	54.5	90.7	39.7	88.7	97.7	2.3	346.9
Croydon	3,980	64.8	92.2	49.5	91.2	98.4	1.6	397.1	310.4	2,011	60.8	89.6	44.8	88.2	97.8	2.2	376.1
Ealing	3,004	69.3	96	54	95.1	99.4	0.6	426.9	327.9	1,502	64.3	95.1	49.3	94.1	99.4	0.6	407.5
Enfield	3,670	60.3	92.2	48.9	90.8	97.4	2.6	365.7	302.5	1,891	55.5	90.6	45.6	89.2	97.1	2.9	347.2
Greenwich	2,557	60.3	92.8	44.8	91.1	98.8	1.2	383.4	301.3	1,283	55.3	90.5	40.8	88.5	98.6	1.4	359.5
Hackney	1,941	57.7	91.7	41.5	90.2	98.3	1.7	357.1	294	892	53	88.6	36.5	86.7	97.8	2.2	337.3
Hammersmith and Fulham	991	63.2	91.9	43.5	90.7	98.7	1.3	377.7	304.5	510	59.8	90.6	40.2	88.6	98.6	1.4	366.9
Haringey	2,313	60.8	90.5	44.2	88.6	96.6	3.4	361.1	296.7	1,224	56.5	88.2	40.6	85.9	95.6	4.4	340.3
Harrow	2,477	74.1	95.5	62.9	94.7	98.7	1.3	417.8	335.1	1,252	68.6	94.3	57.7	93.2	98.6	1.4	396.1
Havering	2,997	66.2	94.7	53.9	93.4	98.8	1.2	405	314.9	1,520	63	93.8	50.5	92.4	98.6	1.4	391.6
Hillingdon	3,101	65.6	92.6	49.3	91.2	98.7	1.3	397	310.1	1,609	61.5	91.5	46.2	89.7	98.4	1.6	382
Hounslow	2,473	66.3	92.8	49.9	91.6	98.8	1.2	400.6	314.1	1,226	60.4	90.8	44.4	89.3	98	2	383.4
Islington	1,417	57.6	90.2	41.4	89	97.4	2.6	356	292.5	710	56.9	89.2	38	87.9	97	3	356.8
Kensington and Chelsea	518	67.6	94.4	51	93.6	98.5	1.5	390.7	315.9	265	x	94.3	46	93.2	98.1	1.9	374.7
Kingston upon Thames	1,424	72.3	93.4	60.4	92.6	98.4	1.6	424.6	331.6	719	67.5	91	56.5	89.8	97.5	2.5	396.6
Lambeth	2,203	62.3	90.9	45.6	89.7	98.9	1.1	382.3	300	1,076	57.1	87.9	40.5	86.3	98.6	1.4	360.4
Lewisham	2,665	61.1	91.7	45.9	90.4	98.4	1.6	376	301.3	1,306	56.3	89.5	42.3	88.4	98.2	1.8	349.9
Merton	1,732	67.4	90.8	53.3	89.9	97.7	2.3	386.1	309.7	939	63.7	89.4	49.1	88.2	97.3	2.7	372.8
Newham	3,243	58	94	45.7	92.1	99	1	363.4	300.7	1,611	51	93.2	39.4	91.3	99.1	0.9	340.9
Redbridge	3,196	72.7	95.7	62.7	94.8	98.5	1.5	418.4	330.8	1,617	67.6	94.4	56.8	93.3	98.3	1.7	401
Richmond upon Thames	1,169	73.1	92.5	64.5	91.8	98.5	1.5	426.2	332.3	533	69	93.1	58.7	92.5	99.2	0.8	403.4
Southwark	2,287	59.3	88.8	45.5	87.7	97.6	2.4	363.6	291.5	1,159	54.1	85.5	39.8	84.6	97	3	338.1
Sutton	2,209	72.6	95.3	57.5	94.1	99.1	0.9	417.9	329.9	1,129	71.4	94	54.2	92.6	98.9	1.1	415.3
Tower Hamlets	2,057	59.9	94.5	42.5	93.4	98.8	1.2	382.1	303.2	1,050	56.8	93.9	39.2	92.8	98.4	1.6	373
Waltham Forest	2,693	62.3	94.6	47.6	93.9	98.4	1.6	377.4	306.5	1,321	56.6	93.9	42.7	93	98.2	1.8	355.3
Wandsworth	1,622	62.9	90.7	48.3	89.3	98	2	377.7	304.2	797	60.5	88	44.4	86.8	97.2	2.8	369
Westminster	944	63.7	91.1	46.5	90.4	98.1	1.9	370.2	303.6	488	60.5	90.2	42.8	89.1	98	2	354.5
North East	32,059	66.5	91.8	45	89.7	98	2	404.8	305.9	16,418	62.2	90.1	41.6	87.7	97.5	2.5	384.4
North West	87,067	65.5	92.2	47.5	90.7	98	2	390.4	306.8	44,446	61.4	90.5	43.8	88.9	97.6	2.4	373.8
Yorkshire and The Humbe	63,382	62.2	91.2	44.5	89.5	97.8	2.2	388	298.7	32,233	58	89.4	40.7	87.5	97.5	2.5	369
East Midlands	53,312	63.2	92.6	47.1	90.7	98.4	1.6	391	305.1	27,266	58.9	91.2	43.5	89.2	98.1	1.9	375.1
West Midlands	67,115	64.2	92.5	46.2	90.9	98.4	1.6	389.2	305.7	34,297	59.2	90.6	41.6	88.8	98.1	1.9	377
East of England	65,794	64.7	92.9	50.2	91.7	98.4	1.6	391.2	309.9	33,472	60.3	91.3	46	89.8	98.1	1.9	373
London	74,535	65.2	93.2	51	92.1	98.5	1.5	391.6	312.2	37,439	60.9	91.6	46.8	90.3	98.1	1.9	373.9
South East	92,114	66.1	93.3	51.8	92.2	98.6	1.4	399.5	313.5	47,004	62.4	92	48.1	90.7	98.3	1.7	383.5
South West	57,154	63.6	93	49.2	91.2	98.6	1.4	387.9	308.6	29,206	58.8	91.4	44.9	89.5	98.3	1.8	369.1

The next section describes how the datasets were parsed, issues encountered (if any), limitations, and adjustments made.

4.2 UNDERSTANDING DATA TYPE AND DATA STRUCTURE

A row from each datasets were analysed and data types allocated to each column. This is shown in the table below.

Table 5: Structure of acquired datasets

CRIME RATES			
00AB	Barking & Dagenham		120.5
Index	*String*		*Float*
GCSE RESULTS BY PUPIL'S RESIDENCE			
00AB/E09000002	Barking and Dagenham	2280	81.9
Index	*String*	*Integer*	*Float*

Index

This data type is commonly used as a string or integer. In this case, it is a string because it is alphanumeric. It maps a location to a column or data in the table.

Crime Rates: The index code 00AB is used to locate crime rate data of the Borough of Barking & Dagenham.

GCSE Results by Pupil's Residence: The index code 00AB; which was the old index for locating the GCSE results (counts, percentages, and scores) of the Borough of Barking and Dagenham, was changed in the year 2010 to E09000002. The author retained the use of the old code for this research to provide consistency between both datasets for smooth integration.

String

The simply refers to a set of characters that forms a sentence or word. In both datasets, the Borough name is designated as a string.

Float

This data type is used to represent a number with a decimal point which can be used for the latitudes and longitudes of each location. However, it is used to represent the crime rates per thousand population and the percentages of pupils achieving a specific indicator for a specific borough in the GRLPR dataset.

After going through the process of understanding the data contained in both dataset, the author proceeded to the next stage which is the filtering stage described in the following section.

4.3 DATA PRE-PROCESSING

The filtering process produced new data structure for each data set, where crime rates data generated a set of independent variables. The variables were divided into three groups according to the type of crime stated in the data gathering stage. The numbers of variables derived were 9 shown in the table below.

Table 6: Variables filtered from the crime rates dataset.

Property Crime Variables	**Violent Crime Variables**	**Area**
Burglary	Violent against the Person	Borough Name
Theft and Handling	Sexual Offences	
Fraud or Forgery	Robbery	
Criminal Damage		
Drugs		

The GRLPR dataset generated independent variables that were divided into two groups shown in the table below

Table 7: Variables filtered from the GRLPR dataset

Area	**GCSE Result Indicators**
Borough Name	All pupils at the end of KS4 achieving 5+A*-C including English and Mathematics. All Pupils at the End of KS4 achieving 5+A*-G including English and Mathematics.

As stated earlier, Gcse indicators were renamed to "LevelTwo" and "LevelOne" correspondingly. Please find the filtered datasets in appendix C2.

Performing the stated standardization technique changed the values for each dataset. Pictures of the standardised dataset are shown in the table below.

Table 8: combined standardised dataset

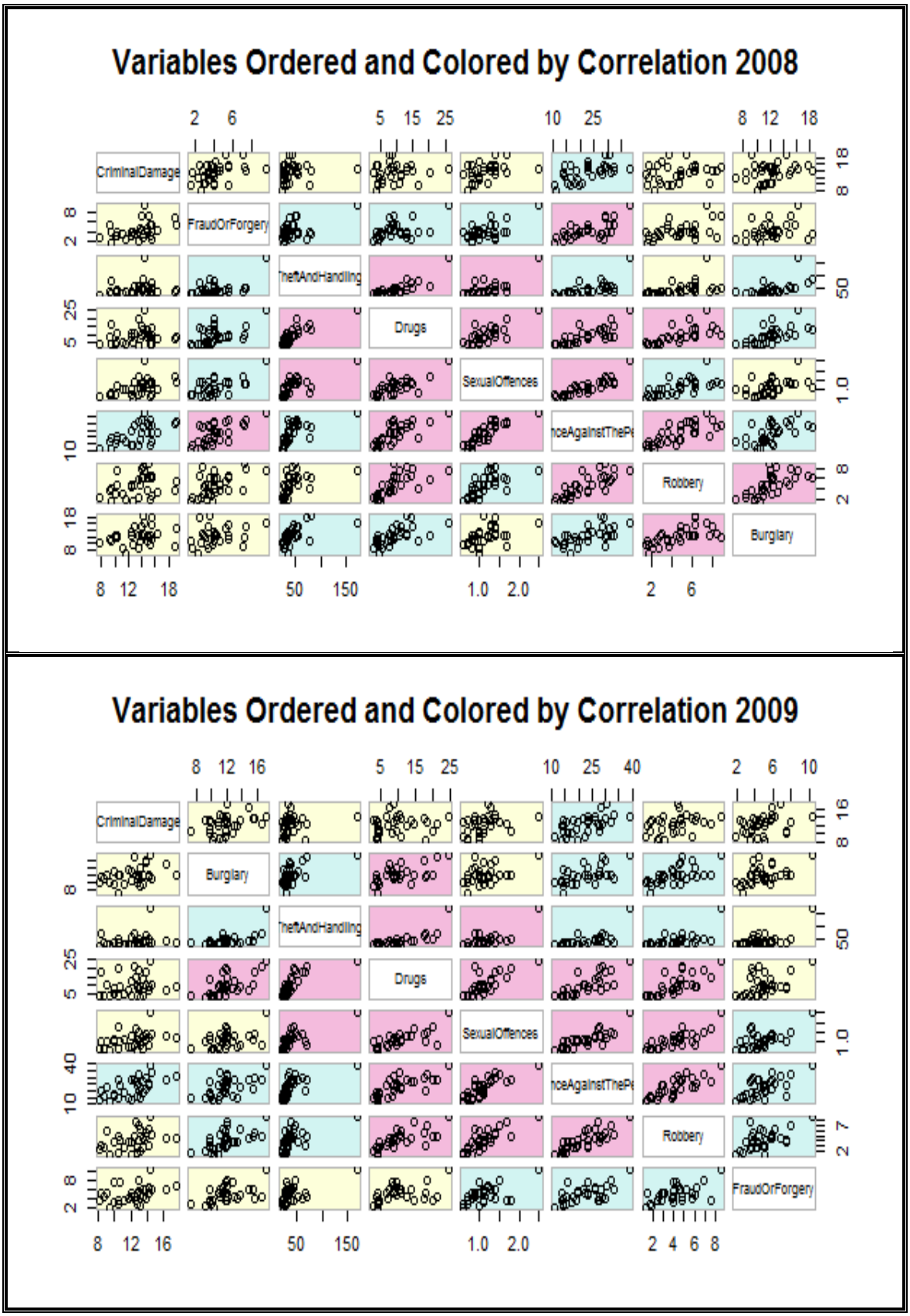
A	B	C	D	E	F	G	H	I	J	K	L
Area	Year	ViolenceAgainstThePerson	SexualOffences	Robbery	Burglary	TheftAndHandling	FraudOrForgery	CriminalDamage	Drugs	LevelTwo	LevelOne
BarkingAndDagenham	2008	0.950748658	1.423625336	-0.46648	-1.11553	-0.276995088	1.656041471	2.073620772	-0.30075	-1.27589	0.678358
Barnet	2008	-0.955095931	-0.892781991	-0.83619	0.022524	-0.497506675	-0.755355861	-1.24977369	-1.04696	1.580139	1.035053
Bexley	2008	-1.122031224	-1.150160583	-1.39076	-0.92585	-0.696328598	-0.583113194	0.729326383	-0.98956	0.074742	0.678358
Brent	2008	-0.217798389	0.136732377	1.382097	-0.01541	-0.475817011	-0.123799417	-1.137749158	0.311517	0.44054	0.627401
Bromley	2008	-0.802071913	-0.892781991	-0.97483	-0.4327	-0.55534578	-0.296042083	0.766667894	-1.1235	1.72083	0.882184
Camden	2008	0.686434445	-0.120646215	0.735096	2.336562	1.349729736	-0.410870527	0.281228253	0.732454	-0.93823	-0.1879
Croydon	2008	-0.551668974	-0.120646215	0.088096	-0.5465	-0.616799829	0.278100139	0.13186221	-0.41556	-0.13629	-0.34077
Ealing	2008	0.241273665	-0.378024807	0.088096	0.439809	-0.352908913	-0.123799417	-0.353577431	0.330651	0.496816	1.64653
Enfield	2008	-1.10811995	-0.892781991	-0.14298	0.136329	-0.515581396	-0.697941638	-0.31623592	-0.24335	-0.22071	-0.5446
Greenwich	2008	1.145506499	0.651489561	0.365382	0.89503	-0.042023725	0.622585472	2.148303794	-0.16682	-0.79754	-0.39173
Hackney	2008	1.24288542	1.423625336	0.504025	0.060459	0.312240793	-0.123799417	0.094520699	2.052665	-1.26182	-0.85034
HammersmithAndFulham	2008	0.79772464	0.651489561	0.088096	1.04677	0.326700569	-0.46828475	-0.204211387	0.349784	-0.98044	-0.59555
Haringey	2008	0.032604549	0.394110969	0.735096	1.350251	0.084499317	0.737413916	0.841350916	0.062782	-0.88196	-1.66564
Harrow	2008	-1.539369455	-1.150160583	-1.11348	-0.77411	-0.775857368	-0.870184305	-2.071286928	-1.06609	1.748968	1.442704
Havering	2008	-1.233321419	-1.407539174	-1.52941	-0.92585	-0.609569941	-0.181213639	-0.241552898	-0.83649	0.482747	0.780271
Hillingdon	2008	0.060427098	-0.378024807	-0.74376	0.363939	-0.38544341	-0.008970972	1.140083002	-0.35816	-0.16443	-0.34077
Hounslow	2008	0.060427098	-0.120646215	-0.51269	-0.50857	-0.360138801	0.220685917	0.542618829	-0.26249	-0.08002	-0.13695
Islington	2008	0.825547189	0.651489561	0.781311	2.108952	1.006310051	-0.985012749	0.804009405	1.000323	-1.27589	-1.46181
KensingtonAndChelsea	2008	-0.426467505	-0.635403399	-0.32783	-0.47063	1.190672198	-0.123799417	-1.436481244	1.402127	0.074742	0.882184
KingstonUponThames	2008	-0.802071913	-0.378024807	-1.34455	-1.98803	-0.378213522	-1.042426971	-0.689651028	-0.95129	1.39724	0.372619
Lambeth	2008	0.616878073	0.651489561	1.751811	0.326004	-0.02033406	-0.812770083	0.281228253	0.732454	-0.68499	-1.10512
Lewisham	2008	1.340264341	0.651489561	0.781311	-0.01541	-0.302299697	0.737413916	0.691984872	-0.10942	-0.64278	-0.74842
Merton	2008	-0.496023877	-0.378024807	-0.8824	-1.45694	-0.674638934	-0.46828475	-0.839017071	-0.79823	0.398332	-1.00321
Newham	2008	0.909014835	0.394110969	1.936669	-0.12922	0.189332695	1.713455693	0.654643361	-0.07115	-0.67092	0.117837

4.4 DATA ANALYSIS

4.4.1 SCATTERPLOT MATRICES

This section identifies if there is a correlation between variables in individual datasets. This was done by applying the commands written in appendix B2 for plotting the scatter diagrams. The scatter plot matrices for year 2008 and 2009 are shown in the figure 15. Please refer to appendix D1 to view the scatter plot matrices for years 2010, 2011, 2012, and 2013.

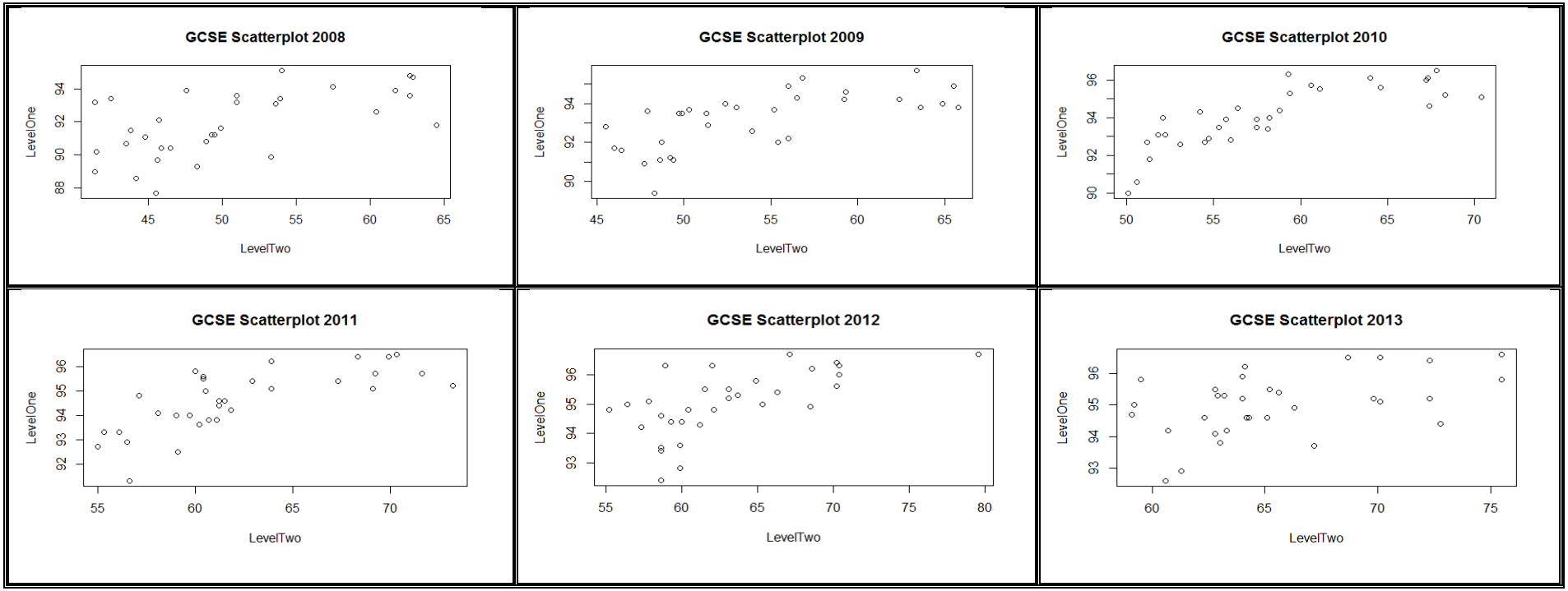
Figure 15: Scatter plot matrices of the variables in the Crime Rates dataset, for the years 2008 and 2009



As can be seen in the figures above and in appendix D1, there seems to be a positive high correlation between variables. For instance, the variables "SexualOffences" and "ViolenceAgainstThePerson" have a positive strong correlation, because the small circles are closely gathered together and the box is coloured pink. In essence, all pink boxes are seen to have a positive high correlation. Overall, violent crime types were observed to be highly correlated with each other over time.

The following scatter plot diagrams identifies if "LevelOne" and "LevelTwo" Gcse indicators are highly or poorly correlated.

Figure 16: Scatter plot matrices on "LevelOne" and "LevelTwo" correlations per year.



In figure 17, it was observed that a correlation between the two variables existed. However, it was rather weak over time, with the exception of the years 2010, 2011, and 2012. These aforementioned years depicted a correlation that was strong but not perfectly so.

Following the observations made for both datasets, the correlation coefficient was calculated to identify the extent to which these variables are related to each other (as stated in chapter 3). The results for the degree of correlation between the crime rates indicators are shown below arranged in descending order of the variables with the highest correlation. The assumption made from understanding the scatter plot matrices was verified. Correlations between violent crime, specifically "ViolenceAgainstThePerson" and "SexualOffences", were indeed the most highly correlated. In addition, it was noticed that the correlation matrix between each pair of variables were high. The least correlated variables ("ViolenceAgainstThePerson", "FraudOrForgery") had a coefficient of 0.669668247, which is still close to +1, hence still highly correlated. Therefore, the crime rates indicators met with the first PCA principle suggested by Mankin(2010).

Figure 17: Most highly correlated variables by year

Year 2008		
First.Variable	**Second.Variable**	**Correlation**
ViolenceAgainstThePerson	SexualOffences	0.872265328
TheftAndHandling	Drugs	0.828814886
ViolenceAgainstThePerson	Robbery	0.766171583
ViolenceAgainstThePerson	Drugs	0.762142144
SexualOffences	Drugs	0.761592961
SexualOffences	TheftAndHandling	0.732220794
Robbery	Drugs	0.678094662
ViolenceAgainstThePerson	FraudOrForgery	0.669668247

Year 2009		
First.Variable	**Second.Variable**	**Correlation**
ViolenceAgainstThePerson	SexualOffences	0.820044
ViolenceAgainstThePerson	Robbery	0.81422
SexualOffences	Robbery	0.786236
ViolenceAgainstThePerson	Drugs	0.773347
TheftAndHandling	Drugs	0.75331
SexualOffences	Drugs	0.744113
Robbery	Drugs	0.739696
SexualOffences	TheftAndHandling	0.733631

Year 2010		
First.Variable	Second.Variable	Correlation
ViolenceAgainstThePerson	SexualOffences	0.871226
ViolenceAgainstThePerson	Drugs	0.828087
TheftAndHandling	Drugs	0.807503
SexualOffences	Drugs	0.779544
SexualOffences	Robbery	0.767164
ViolenceAgainstThePerson	Robbery	0.761519
Robbery	Drugs	0.714401
ViolenceAgainstThePerson	CriminalDamage	0.689677

Year 2011		
First.Variable	Second.Variable	Correlation
TheftAndHandling	Drugs	0.873424
ViolenceAgainstThePerson	SexualOffences	0.867269
ViolenceAgainstThePerson	Drugs	0.824567
ViolenceAgainstThePerson	Robbery	0.785636
SexualOffences	Robbery	0.776518
SexualOffences	Drugs	0.77107
ViolenceAgainstThePerson	TheftAndHandling	0.764763
SexualOffences	TheftAndHandling	0.737337

Year 2012		
First.Variable	Second.Variable	Correlation
ViolenceAgainstThePerson	SexualOffences	0.881046
TheftAndHandling	Drugs	0.868261
ViolenceAgainstThePerson	Drugs	0.844234
SexualOffences	Drugs	0.839701
SexualOffences	TheftAndHandling	0.800045
ViolenceAgainstThePerson	TheftAndHandling	0.773282
SexualOffences	Robbery	0.769139
ViolenceAgainstThePerson	Robbery	0.756078

Year 2013		
First.Variable	Second.Variable	Correlation
ViolenceAgainstThePerson	SexualOffences	0.90885
TheftAndHandling	FraudOrForgery	0.899596
ViolenceAgainstThePerson	CriminalDamage	0.846267
TheftAndHandling	Drugs	0.841378
ViolenceAgainstThePerson	Drugs	0.834721
FraudOrForgery	Drugs	0.823288
SexualOffences	CriminalDamage	0.820931
SexualOffences	Drugs	0.808627

The results of the degree of correlation between GRLPR indicators are below.

Year 2008

```
        Pearson's product-moment correlation

data:  GcseSorted[1:32, 3] and GcseSorted[1:32, 4]
t = 3.9155, df = 30, p-value = 0.0004814
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2920890 0.7734109
sample estimates:
     cor
0.581552
```

Year 2009

```
Pearson's product-moment correlation

data:  GcseSorted[33:64, 3] and GcseSorted[33:64, 4]
t = 4.8378, df = 30, p-value = 3.681e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4073413 0.8211480
sample estimates:
      cor
0.6620057
```

Year 2010

```
        Pearson's product-moment correlation

data:  GcseSorted[65:96, 3] and GcseSorted[65:96, 4]
t = 7.3891, df = 30, p-value = 3.119e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6315867 0.8999514
sample estimates:
     cor
0.803357
```

Year 2011

```
        Pearson's product-moment correlation

data:  GcseSorted[97:128, 3] and GcseSorted[97:128, 4]
t = 5.8095, df = 30, p-value = 2.377e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5077338 0.8584925
```

```
sample estimates:
      cor
0.7276091
```

Year 2012

```
        Pearson's product-moment correlation

data:  GcseSorted[129:160, 3] and GcseSorted[129:160, 4]
t = 4.5843, df = 30, p-value = 7.515e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3776574 0.8093828
sample estimates:
      cor
0.6418307
```

Year 2013

```
        Pearson's product-moment correlation

data:  GcseSorted[161:192, 3] and GcseSorted[161:192, 4]
t = 2.9171, df = 30, p-value = 0.006632
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1451767 0.7034624
sample estimates:
      cor
0.4700745
```

The results of the correlation coefficient tallied with what was observed form the scatter plot diagrams. The years 2010, 2011, and 2012 are shown to be highly correlated. Having values as 0.803357, 0.7276021, and 0.6418307 respectively, which are close to +1. In summary, it was also found to meet the first PCA principle.

4.4.2 PRINCIPAL COMPONENT ANALYSIS

The PCA values returned when the *"prcomp()"* was called for both datasets by year, are described below using the *"summary()"* function.

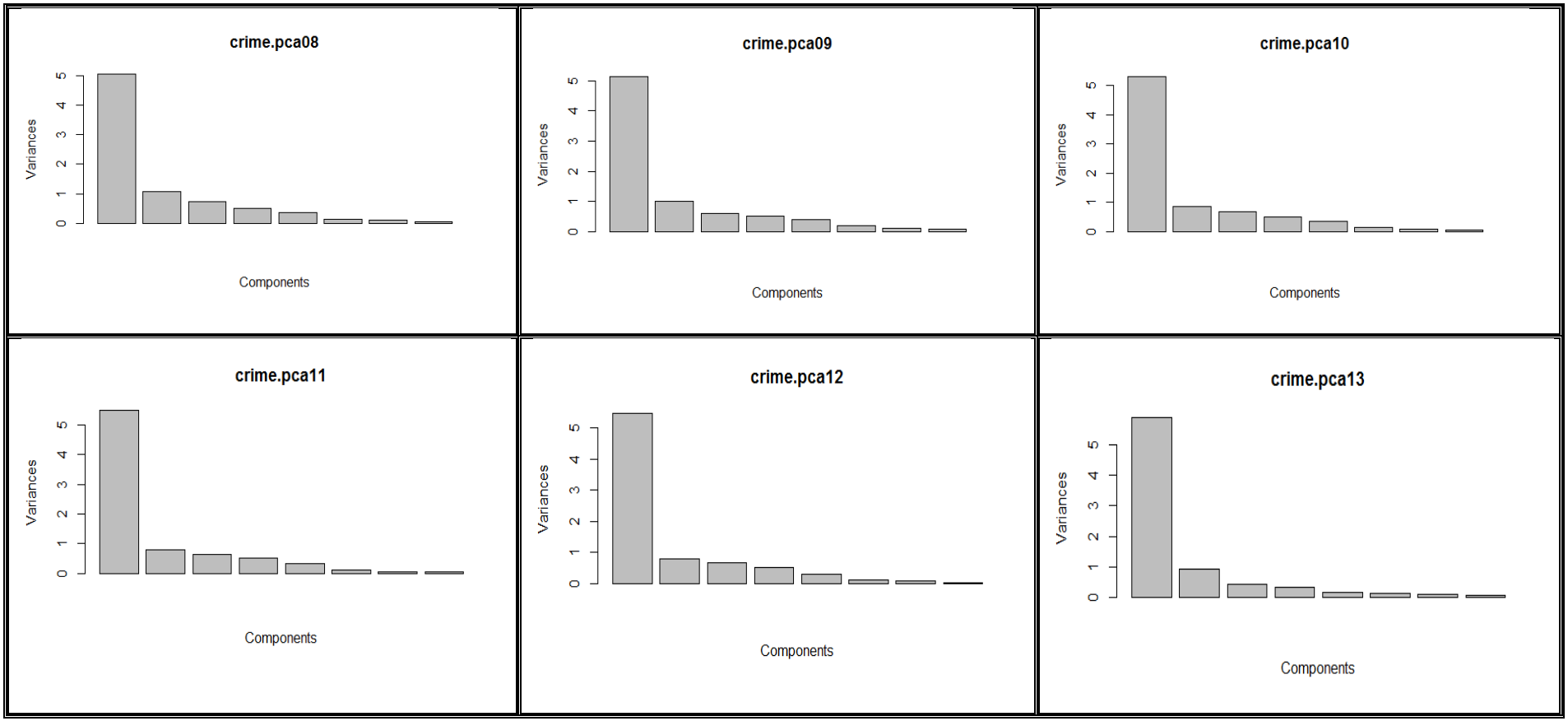
The crime rates dataset returned eight PCA values with the first PCA accounting for most of the variance in each year. However, to be certain of

```
> summary(crime.pca08)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8
Standard deviation     2.2451 1.0318 0.85545 0.70106 0.6006 0.37391 0.33750 0.23904
Proportion of Variance 0.6301 0.1331 0.09147 0.06144 0.0451 0.01748 0.01424 0.00714
Cumulative Proportion  0.6301 0.7631 0.85461 0.91605 0.9611 0.97862 0.99286 1.00000

> summary(crime.pca09)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.265 0.9961 0.77153 0.71938 0.62302 0.45337 0.32263 0.26253
Proportion of Variance 0.641 0.1240 0.07441 0.06469 0.04852 0.02569 0.01301 0.00862
Cumulative Proportion  0.641 0.7651 0.83947 0.90416 0.95268 0.97837 0.99138 1.00000

> summary(crime.pca10)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.303 0.9341 0.82340 0.69886 0.60461 0.38007 0.31187 0.22355
Proportion of Variance 0.663 0.1091 0.08475 0.06105 0.04569 0.01806 0.01216 0.00625
Cumulative Proportion  0.663 0.7720 0.85679 0.91784 0.96354 0.98160 0.99375 1.00000

> summary(crime.pca11)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.3435 0.89403 0.79454 0.72235 0.56625 0.34261 0.25980 0.22420
Proportion of Variance 0.6865 0.09991 0.07891 0.06522 0.04008 0.01467 0.00844 0.00628
Cumulative Proportion  0.6865 0.78639 0.86530 0.93053 0.97061 0.98528 0.99372 1.00000

> summary(crime.pca12)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.3371 0.88310 0.82227 0.71807 0.56045 0.36103 0.28630 0.19943
Proportion of Variance 0.6828 0.09748 0.08452 0.06445 0.03926 0.01629 0.01025 0.00497
Cumulative Proportion  0.6828 0.78026 0.86477 0.92923 0.96849 0.98478 0.99503 1.00000

> summary(crime.pca13)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8
Standard deviation     2.4276 0.9659 0.65352 0.57162 0.41734 0.33125 0.28816 0.2297
Proportion of Variance 0.7367 0.1166 0.05339 0.04084 0.02177 0.01372 0.01038 0.0066
Cumulative Proportion  0.7367 0.8533 0.90669 0.94753 0.96931 0.98302 0.99340 1.0000
```

how many PCA values to retain for each year, the Guttman-Kaiser criterion and scree test were used.

The scree plots of PCA values for each year were plotted shown in the table below.

Figure 18: The scree plots of PCA values for each year (crime data)



It was very obvious that the elbow occurred at the second PCA value in every year; hence the first PCA value was shown to account for the most variance. To be thoroughly sure of the consistency, the Guttmans-Kaiser criterion was used by calculating the eigenvalues for each year.

Table 9: Eigenvalues of PCA values for Crime rates Dataset

PCA	PCA 2008	PCA2009	PCA2010	PCA2011	PCA2012	PCA2013
1	5.04055	5.128312	5.303788	5.491834	5.462204	5.893388
2	1.064542	0.992204	0.872567	0.799289	0.77986	0.933049
3	0.731789	0.595259	0.677988	0.631299	0.676125	0.427093
4	0.49149	0.517511	0.488409	0.52179	0.515618	0.326746
5	0.360777	0.388159	0.365559	0.320644	0.314107	0.174175
6	0.139806	0.205546	0.144453	0.117382	0.130346	0.109728
7	0.113905	0.104089	0.097263	0.067497	0.081968	0.083035
8	0.05714	0.068921	0.049973	0.050265	0.039771	0.052785

Using the criterion, the first PCA satisfied Mankin's (2010) second principle as it was above 1.0 for each year.

The same process was carried out for the GRLPR dataset and results are shown below.

```
> summary(Gcse.pca08)
Importance of components:
                          PC1    PC2
Standard deviation     1.2576 0.6469
Proportion of Variance 0.7908 0.2092
Cumulative Proportion  0.7908 1.0000
> summary(Gcse.pca09)
Importance of components:
                          PC1    PC2
Standard deviation     1.289 0.5814
Proportion of Variance 0.831 0.1690
Cumulative Proportion  0.831 1.0000
```

```
> summary(Gcse.pca10)
Importance of components:
                          PC1     PC2
Standard deviation     1.3429 0.44344
Proportion of Variance 0.9017 0.09832
Cumulative Proportion  0.9017 1.00000
> summary(Gcse.pca11)
Importance of components:
                          PC1    PC2
Standard deviation     1.3144 0.5219
Proportion of Variance 0.8638 0.1362
Cumulative Proportion  0.8638 1.0000
> summary(Gcse.pca12)
Importance of components:
                          PC1    PC2
Standard deviation     1.2813 0.5985
Proportion of Variance 0.8209 0.1791
Cumulative Proportion  0.8209 1.0000
> summary(Gcse.pca13)
Importance of components:
                          PC1   PC2
Standard deviation     1.212 0.728
Proportion of Variance 0.735 0.265
Cumulative Proportion  0.735 1.000
```

As can be seen, PC1 accounted for much of the variance. The result of the Guttmans-kaiser criterion for selecting the PCA values to retain is shown below. Please see appendix C2 for the scree plots.

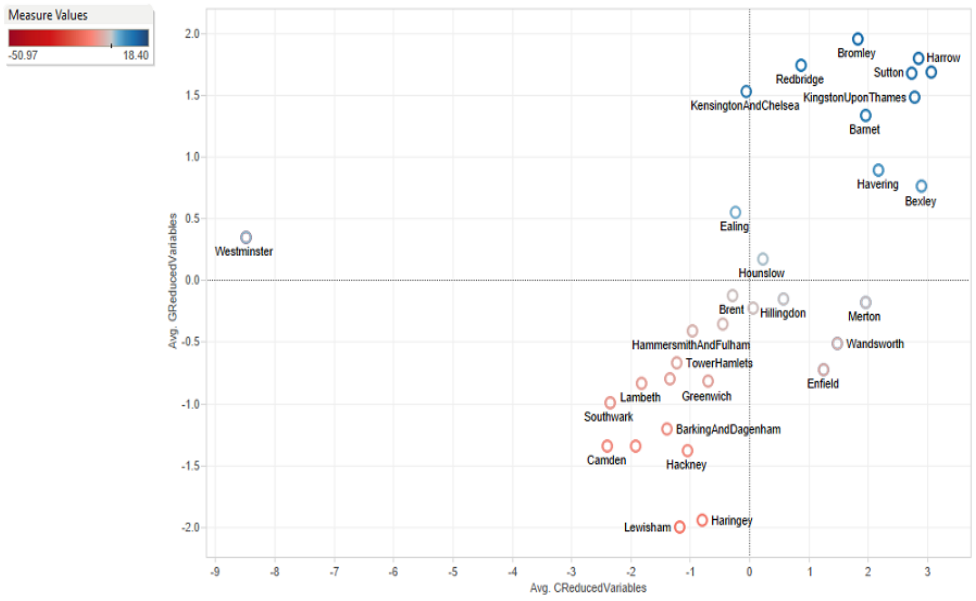
Table 10: Eigenvalues of PCA values for GRLPR dataset

PCA	PCA2008	PCA2009	PCA2010	PCA2011	PCA2012	PCA2013
1	1.581552	1.662006	1.803357	1.727609	1.641831	1.641831
2	0.418448	0.337994	0.196643	0.272391	0.358169	0.358169

Both PCA values for each data frame were also identified to meet the second principle of PCA as well (Mankin, 2010). Please refer to appendix C3: for the result of the first PCA scores for each year and for both datasets. The names of the PCA scores were changed to "CReducedVariables" and "GReducedVariables" for crime rates and GRLPR datasets in that order.

In summary, principal component analysis was applied to each datasets, and at the end both datasets were reduced into one indicator each. A scatter plot diagram below illustrated the degree of correlation between the new set of reduced variables.

Figure 19: Scatter plot diagram for PCA scores



The scatter plot is based on the average of each variable. It was found that there existed a strong positive correlation between crime rates and GCSE attainment. It was also noticed that the borough of Westminster was located very far apart from the other borough, therefore the researcher decided to examine its pattern more. This is carried out in the next section.

However, one of the sole aim of this project was to find if a pattern or trend existed between these two by borough and year. At this point, the researcher proceeded to visualize the data.

4.5 INFORMATION VISUALIZATION

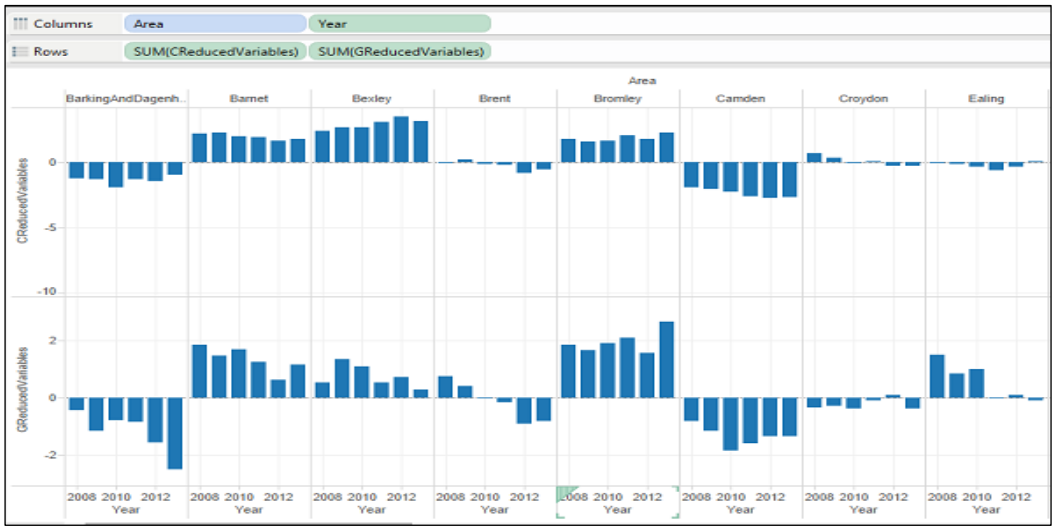
As stated earlier, Tableau Public was used in visualizing the data. The visualization process was broken down into three phases.

4.5.1 PHASE 1: Initial Representation of Data

The data was formatted to the default settings of Tableau Public as it was read. The bar graph of the first eight boroughs is shown in figure 20. The bars are arranged in a vertical table lens manner described by Few(2012) as when "*two sets of bars, one for each variable, arranged in this manner make it possible to see correlations*". The bars represent the overall sum of values for each variable, each borough, and each year.

Since the adopted methodology is followed using an exploratory data analysis approach, the researcher found it difficult to find a pattern or trend with the initial visual design. Therefore, the need to refine and interact with the data was carried out.

Figure 20: Initial Representation of the first eight boroughs overtime

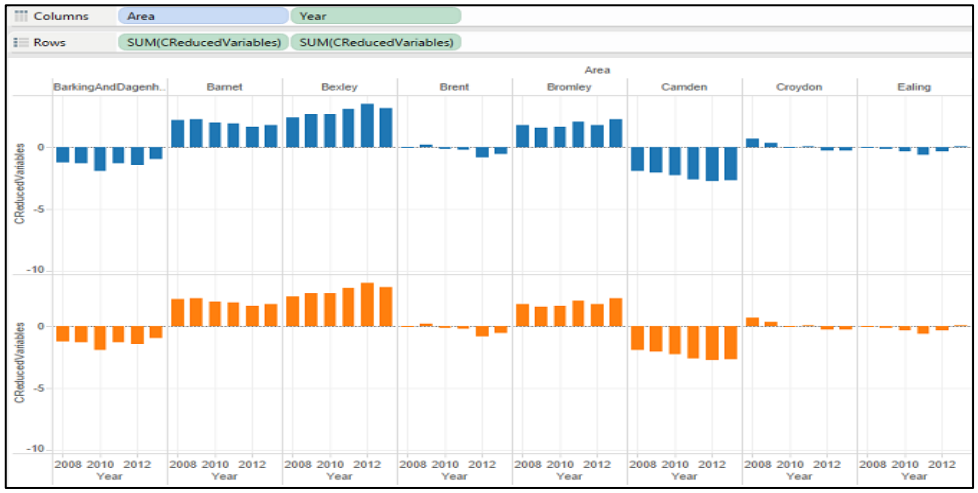


4.5.2 PHASE 2: Refining the data

Iteration 1

The initial bar graphs represented the two variables ("CReducedVariables" and "GReducedVariables") in the same color (blue) which made patterns invisible to the eyes. Light hues were used to differentiate the two variables; orange for GReducedVariables and blue for CReducedVariables. Figure 20 was refined to Figure 21 shown below.

Figure 21: First Iteration Refined bar graph of figure 20

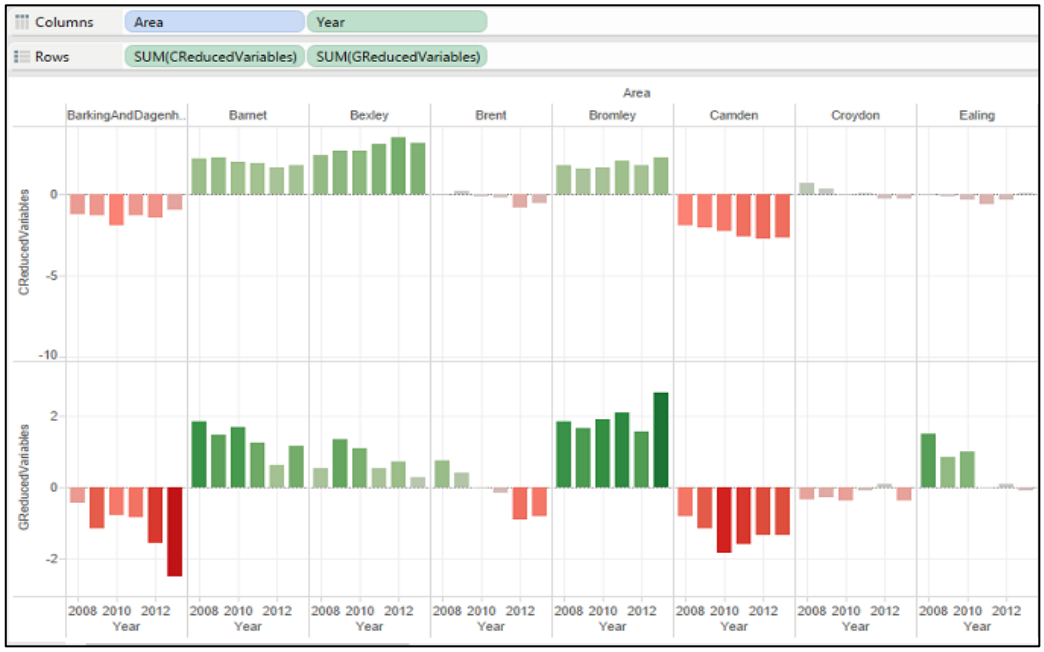


Despite designing each variable to be distinct, patterns were still hard to find, hence the need to refine it more with distinctive hues.

Iteration 2

The bars were further refined to have a color diverging from red to green, where the red end of the spectrum indicated the lower values and the green end of the spectrum indicated the higher values. The refined figure is shown below.

Figure 22: Second Iteration Refined Bar Graph of figure 21



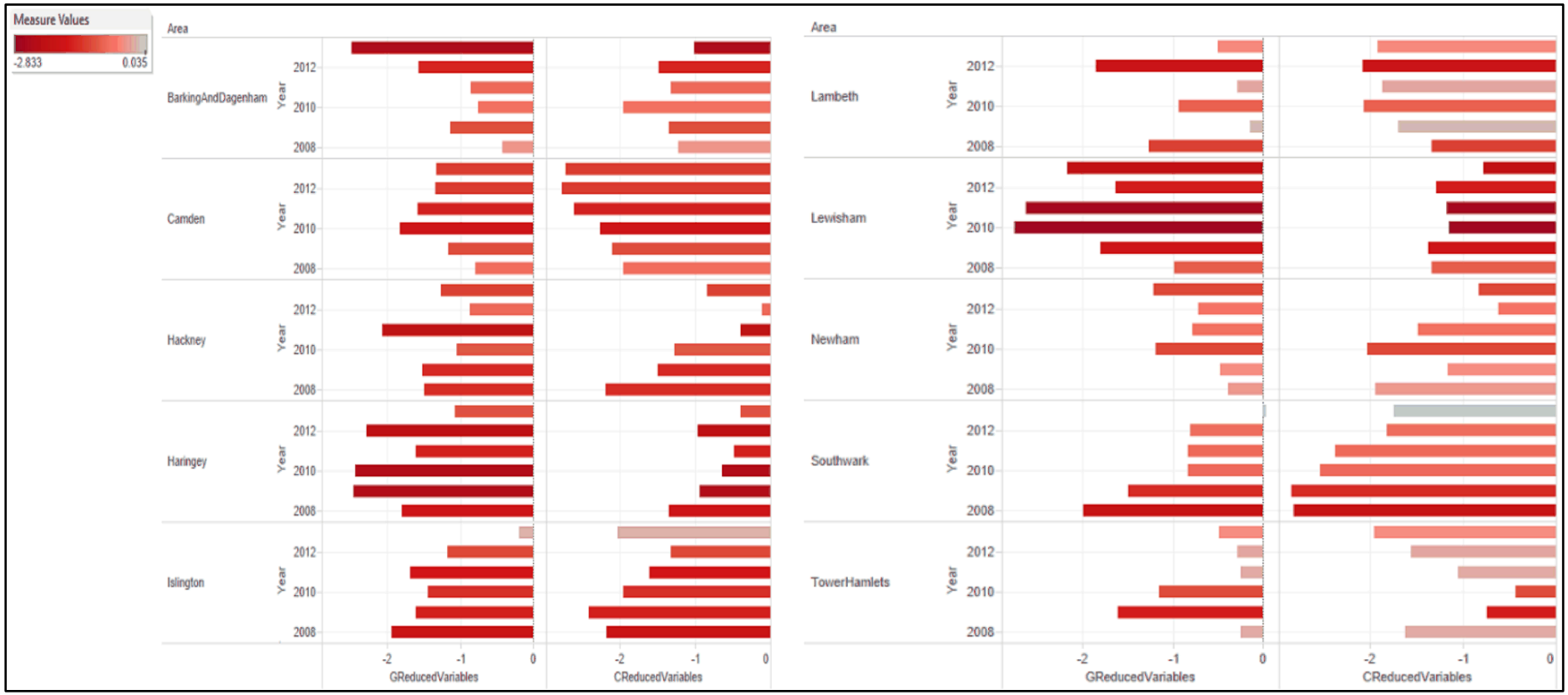
At this point, the researcher could see some patterns in correlation between the two variables such as bars of Barnet and Bexley boroughs. The researcher decided that bars with similar colors needed to be explored more in depth. This was done in phase 3.

4.5.3 PHASE 3: Interacting with the Data

By using the interactive filter feature in Tableau public, the researcher filtered the bar graphs by borough having similar colors. Three groups of boroughs with similar correlation patterns were identified and represented in a set of horizontal bars stacked above one another for the time period of 2008-2013.

Group 1: Bars colored in diverging shades of red, where the darker shades and lighter shades of red indicated the lower values and higher values respectively.

Figure 23: Pattern of correlation where both crime rates (CReducedVariables) and GCSE attainment (GReducedVariables) tend to decrease together



As shown in figure 23, the lowest and highest values were -2.833 and 0.035 respectively. The researcher identified a correlation pattern – although not perfect, where as GCSE attainment (GReducedVariables) of students decreased, the crime rates (CReducedVariables) for that specific borough had a tendency to decrease in a corresponding manner. This was true for the borough of BarkingAndDagenham, Camden, Hackney, Haringey, Islington, Lambeth, Lewisham, Newham, SouthWark, and TowerHamlets. Figure 23, illustrates the correlation pattern for the aforementioned boroughs.

GROUP 2: Bars colored in diverging shades of green, where lighter shades and darker shades of green represent lower and higher values respectively. As shown in figure…the lowest and highest values are 0.287 and 3.496 respectively. The researcher identified a correlation pattern –although not perfect, where as GCSE attainment (GReducedVariables) of students increased, the crime rates (CReducedVariables) for that specific borough had a tendency to increase in a corresponding manner. This was exact for the borough of Barnet, Bexley, Bromley, Harrow, Havering, KingstonUponThames, Redbridge, RichmondUponThames, and Sutton, This is illustrated in figure 24 ."Gcse attainment for students" and "GReducedVariables" will be used interchangeably from this point. "Crime rates" and "CReducedVariables" will be used interchangeably from this point.

GROUP 3: Bars colored in diverging colors from red to green, where the red end of the spectrum represents the lower values and the green end of the spectrum represents the higher values. As shown in figure 25 the lowest value is -9.64, while the highest value is 3.24. The correlation pattern that was perceived by the researcher was that of irregular patterns for this group of boroughs. The boroughs perceived to have an obvious correlation pattern by the researcher are discussed in the following paragraphs.

To begin with, Westminster had a trend where there was a gradual steady decrease in crime rates (CReducedVariables) and a gradual unsteady increase in Gcse attainment (GReducedVariables) through most of the time period. Consequently, as CReducedVariables tend to decrease, GReducedVariables increased. This demonstrated a negative correlation.

Figure 24: … Pattern of correlation where both crime rates (CReducedVariables) and GCSE attainment(GReducedVariables) tend to increase together

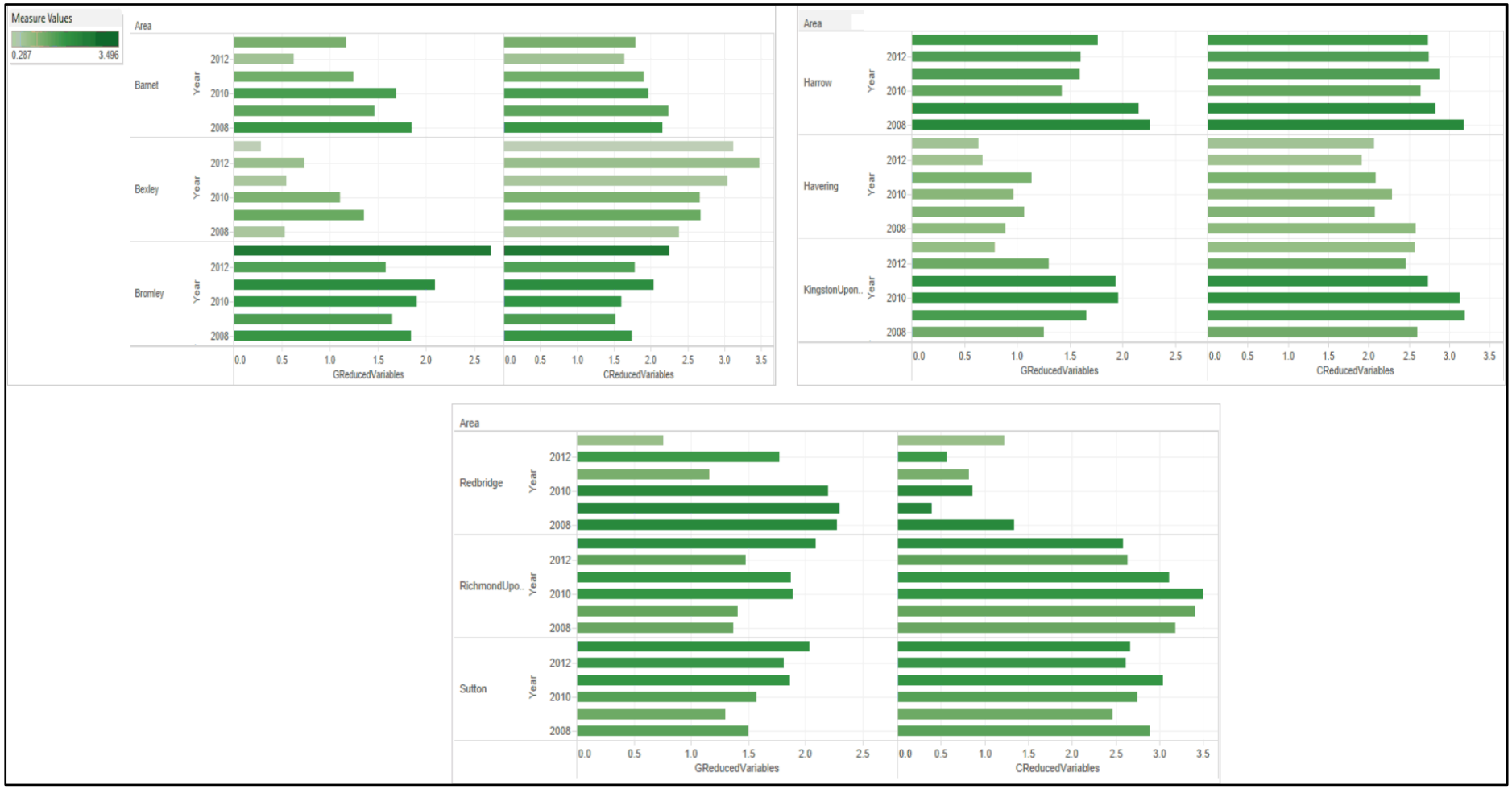
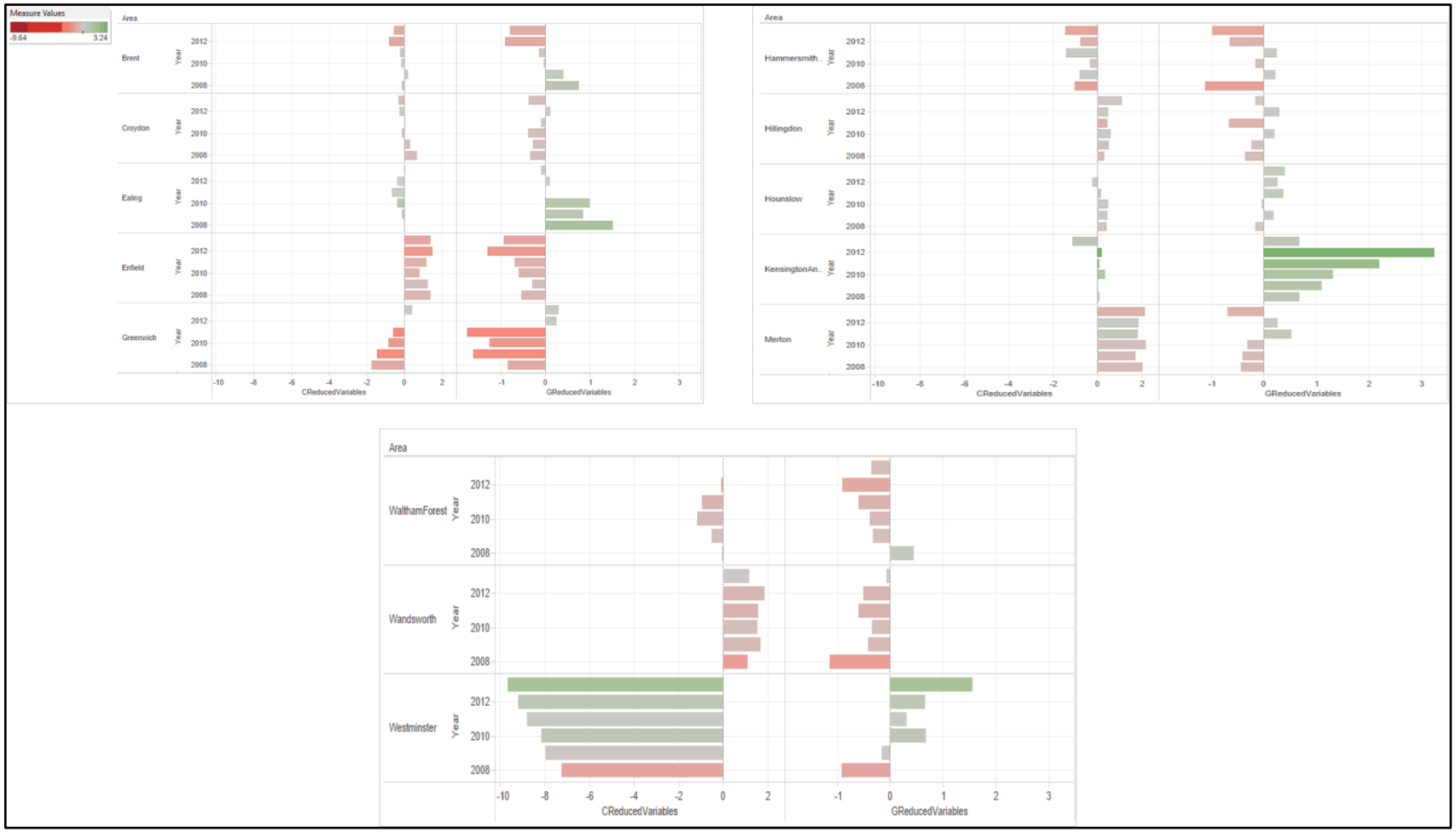


Figure 25: Pattern of correlation where both crime rates (CReducedVariables) and GCSE attainment(GReducedVariables) tend to increase together to fluctuate through time



The borough of Enfield, and Wandsworth had a correlation pattern where an increase in crime rates, led to a decrease in Gcse attainment of students overtime. Whereas, in the borough of KensingtonAndChelsea, a major increase in GReducedVariables up until the year 2012, led to a minor increase in CReducedVariables. In the year 2013, there was a major decrease in Gcse attainment of students, which led to a corresponding major decrease in crime. Overall, this group of boroughs exhibited a pattern that was fluctuated (high or less) through time.

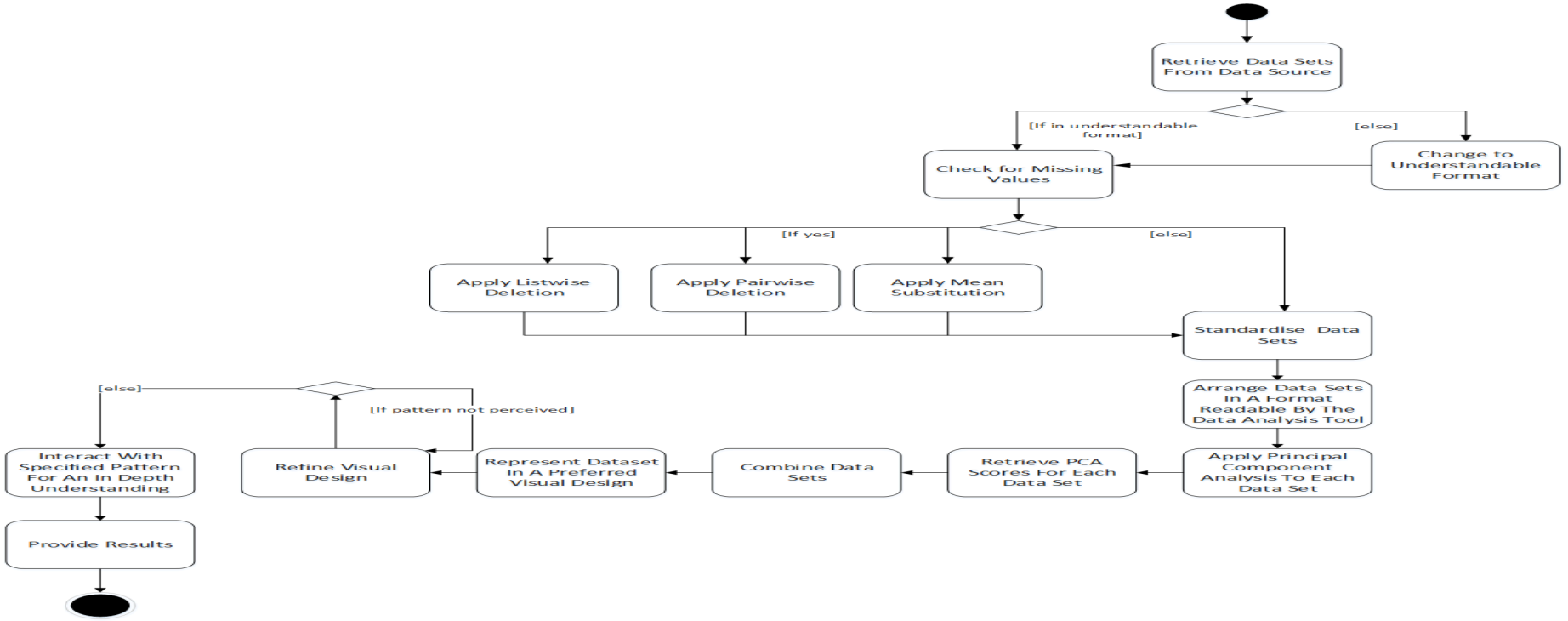
*All the visual designs can be seen more clearly via the following link

https://public.tableausoftware.com/profile/juliet6259#!/vizhome/CRIMEANDGRADES/Group1

4.6 ACTIVITY DIAGRAM OF THE PROPOSED FRAMEWORK FOR ANALYSING MULTIPLE TEMPORAL MULTIVARIATES DATA SETS

Following its application to perceive patterns, the processes used in achieving the objectives were put together and described using an activity diagram. The process illustrated in the diagram describes the direction to be applied by other researchers in similar projects. In general, it was assigned a name known as Multidimensional Temporal Data Explorer.

Figure 26: Multidimensional Temporal Data Explorer



CHAPTER 5: DISCUSSION

This chapter discusses if the project objectives were met, including what the project findings were, areas for future research, and provides recommendations. The discussion is broken down into five sections, one per project objectives.

5.1 Data sets will be extracted from the London Datastore

This project focused on two different datasets – crime rates by borough and Gcse result by location of pupil residence, borough. These data sets were successfully retrieved in an easy readable format from the London datastore. Retrieving all the datasets needed for a research from an open data source is not that common. Yet the researcher was lucky enough to find these data sets publicised by one data source in the similar structure. They were both spatio-temporal data sets. This made it easy and less time consuming in understanding their structure and content. Overall, the objective was successfully met.

5.2 Data sets will be filtered, covering the same period and location, and normalised to the same format

Both data sets were retrieved in different format and time period as explained in section 3.2 of which the GRLPR data set contained missing values. These were some of the challenges faced when conducting an analysis of multiple multivariate data sets, stated in section 2.8. Both data sets were easily filtered to contain variables needed for this research. However, to proceed with the analysis, both datasets had to be normalized to the same format as shown in section 3.3.1.

Achieving this objective was quiet stressful. This was because the researcher noticed it required a lot of iteration to structure each data set into a readable

format that was recognizable by RStudio. Nonetheless, the objective was successfully achieved eventually.

5.3 The data from both sets will be synthesised to establish if correlation points exists between datasets.

Identifying the degree of correlation both statistically and visually was one of the main aims of this research. Bearing in mind that both data sets produced an unequal number of variables after the filtering process, it was decided, as a first step, to reduce the number of variables by performing principal component analysis proposed by schlens (2014). It was a challenge to perform principal component analysis on temporal dataset because the researcher still needed to retain the temporal nature of each data sets, hence PCA was carried out by year. This made it possible to find the correlation between each data set through the use of a scatter plot. By looking at the plot, the researcher was confident that there existed a strong positive correlation between crime rate and Gcse attainment statistics by borough on average. On the whole, this objective was successfully met.

5.4 A data visualization design will be developed to explore the temporal correlations over the same period in order to discover possible patterns/trends.

A bar graph represented in a vertical table lens manner, was found to be the best medium for communicating the relationship between the variables. The preattentive attributes described Ware by (2004) made it easy to perceive temporal patterns of correlation between the variables over the same period. The bar graph allowed more effective use of color expression in representing the weight of values. This resulted in finding groups of boroughs with similar temporal patterns of correlation .In the end, the objective was considered to be successfully met.

5.5 A "recipe" for analysing multiple multivariate temporal data sets will be proposed.

In general, the researcher was able to identify relevant sources to feed into the methods used in this research, to the point that a framework was developed. This is a framework proposed by the researcher to solve the challenge of analysing heterogeneous datasets from multi data sources stated by Keim et al (2003). It can also be applied by analysts in similar research.

5.6 Limitations of this project

The research was only focused on the temporal relationship between crime and educational performance by London borough. This results derived deserve further investigation to see if these patterns are consistent over larger coverage regions and time.

CHAPTER 6: EVALUATION, REFLECTION AND CONCLUSION

The overall aims of this research were fourfold – to identify the degree of relationship between crime rates and educational performance in London Boroughs, to develop a visual design suitable for exploring the temporal relationship, to determine if a pattern can be perceived and to propose a structure that can be applied in analysing multiple temporal heterogeneous datasets. These aims were broken down into specific objectives that provided a researcher with a guideline for achieving the project goals.

A rigorous literature review was carried out which included a general information on crime and educational performance in London, with a critical analysis of previous researches finding the connection between the two. It also justified why information visualization and visual analysis techniques are necessary when carrying out research to find a pattern. Overall, the researcher was satisfied with the level of review completed.

The chosen methods used in this research were provided in a systematic manner. The researcher made sure to justify why a specific method was chosen, and made sure it was understood before application. The two data sets were retrieved, cleaned, and analysed. In addition the results derived clearly showed which objectives were achieved. Overall, the researcher felt the project was a success.

In conclusion, the research contributed to already existing knowledge in the connection between crime and educational performance. It also proposed a "recipe" that can be applied by other researchers dealing with similar project aims. Since the researcher was not concerned about the causal reasons for the resulting patterns (out of the project scope), it is hoped that the use of quantitative analysis and visual analysis techniques, in finding patterns of correlation between multivariate temporal data sets, can direct social scientist or the beneficiaries at large to areas of future research and in the end aid in making better decisions.

REFERENCES

Allison, P. (2002). Missing data. 1st ed. Thousand Oaks, Calif.: Sage Publications.

Azeezullah, I., Pambudi, F., Shyy, T., Azeezullah, I., Ward, N., Hunter, J. and Stimson, R. J. (2012) 'Statistical analysis and visualization services for Spatially Integrated Social Science datasets', in *2012 IEEE 8th International Conference on E-Science (e-Science)*, pp. 1–8. doi: 10.1109/eScience.2012.6404421.

Becker, G. S. (1974) *Essays in the economics of crime and punishment*. New York, NY: National Bureau of Economic Research.

Bendix, F., Kosara, R. and Hauser, H. (2005) 'Parallel sets: visual analysis of categorical data', in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 133–140. doi: 10.1109/INFVIS.2005.1532139.

Bretaña, G. (1965) *Firearms. Act 1965*. Her majesty's stationery office.

Card, S. K., Mackinlay, J. D. and Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.

Cattell, R. B. (1966) 'The Scree Test For The Number Of Factors', *Multivariate Behavioral Research*, 1(2), pp. 245–276. doi: 10.1207/s15327906mbr0102_10.

Coghlan, A. (2014). A little book of R for Multivariate Analysis. 1st ed. Cambridge, UK: Wellcome Trust Sanger Institute, p.24.

Data.gov.uk, (2014). About | data.gov.uk. [online] Available at: http://data.gov.uk/about [Accessed 21 Sep. 2014]

Data.london.gov.uk, (2014). Welcome to the London Datastore | London DataStore. [online] Available at: http://data.london.gov.uk/ [Accessed 21 Sep. 2014].

David Smith (2012); *R Tops Data Mining Software Poll,* Java Developers Journal, May 31, 2012.

Davies, T. (2010) 'Open data, democracy and public sector reform', *A look at open government data use from data. gov. uk*. Available at: http://www.opendataimpacts.net/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf (Accessed: 8 September 2014).

Dickson, G. W., DeSanctis, G. and McBride, D. J. (1986) 'Understanding the Effectiveness of Computer Graphics for Decision Support: A Cumulative Experimental Approach', *Commun. ACM*, 29(1), pp. 40–47. doi: 10.1145/5465.5469.

Educational Act (2002) *Education Act 2002*. Available at: http://www.legislation.gov.uk/ukpga/2002/32/section/82 (Accessed: 4 August 2014).

Elliott, C. and Ellingworth, D. (1996) 'The relationship between unemployment and crime: A cross-sectional analysis employing the British Crime Survey 1992', *International Journal of Manpower*, 17(6/7), pp. 81–88. doi: 10.1108/01437729610149358.

Everitt, B. 2011, *Cluster analysis*, Wiley-Blackwell, Oxford.

Few, S. (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press.

Fox, J. and Andersen, R. (2005) 'Using the R statistical computing environment to teach social statistics courses', *Department of Sociology, McMaster University*. Available at: http://socserv.socsci.mcmaster.ca/jfox/Teaching-with-R.pdf (Accessed: 13 September 2014).

Fry, B. (2003) *Visualizing data*. O'Reilly. Available at: https://www.dawsonera.com/abstract/9780596519308 (Accessed: 7 August 2014).

Galili, T. (2010) *Correlation scatter-plot matrix for ordered-categorical data*, *R-statistics blog*. Available at: http://www.r-statistics.com/2010/04/correlation-scatter-plot-matrix-for-ordered-categorical-data/ (Accessed: 13 June 2014).

George, M. K. (1948) *Rank correlation methods*. Oxford, England: Griffin.

Gershon, N. and Eick, S. G. (1995) 'Visualization's new tack: making sense of information', *IEEE Spectrum*, 32(11), pp. 38–40, 42, 44–7, 55–6. doi: 10.1109/6.469330.

Gershon, N., Eick, S. G. and Card, S. (1998) 'Information Visualization', *interactions*, 5(2), pp. 9–15. doi: 10.1145/274430.274432.

Gonzalez, V. and Kobsa, A. (2003) 'Benefits of information visualization systems for administrative data analysts', in *Seventh International Conference on Information Visualization, 2003. IV 2003. Proceedings*, pp. 331–336. doi: 10.1109/IV.2003.1217999.

Guttman, L. (1954) 'Some necessary conditions for common-factor analysis', *Psychometrika*, 19(2), pp. 149–161. doi: 10.1007/BF02289162.

Hale, C. and Sabbagh, D. (1991) 'Testing the Relationship between Unemployment and Crime: A Methodological Comment and Empirical Analysis Using Time Series Data from England and Wales', *Journal of Research in Crime and Delinquency*, 28(4), pp. 400–417. doi: 10.1177/0022427891028004002.

Hall, W., Shadbolt, N., Tiropanis, T., O'Hara, K. and Davies, T. (2012) 'Open data and charities'. Available at: http://eprints.soton.ac.uk/341346/ (Accessed: 8 September 2014).

Hamnett, C., Ramsden, M. and Butler, T. (2007) 'Social Background, Ethnicity, School Composition and Educational Attainment in East London', *Urban Studies*, 44(7), pp. 1255–1280. doi: 10.1080/00420980701302395.

Holtz, S., Valle, G., Howard, J. and Morreale, P. (2011) 'Visualization and pattern identification in large scale time series data', in *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 129–130. doi: 10.1109/LDAV.2011.6092333.

Hutchins, E. L., Hollan, J. D. and Norman, D. A. (1985) 'Direct manipulation Interfaces', *Human-Computer Interaction*, 1(4), p. 311.

Ihaka, R. and Gentleman, R. (1996) 'R: A Language for Data Analysis and Graphics', *Journal of Computational and Graphical Statistics*, 5(3), pp. 299–314. doi: 10.2307/1390807.

InterWorks, Inc., (2014). Why Tableau?. [online] Available at: http://www.interworks.com/services/business-intelligence/why-tableau [Accessed 20 Sep. 2014].

Kaiser, H. F. (1960) 'The application of electronic computers to factor analysis', *Educational and Psychological Measurement*, 20, pp. 141–151. doi: 10.1177/001316446002000116.

Kaiser, H. F. (1961) 'A Note on Guttman's Lower Bound for the Number of Common Factors1', *British Journal of Statistical Psychology*, 14(1), pp. 1–2. doi: 10.1111/j.2044-8317.1961.tb00061.x.

Karl Rexer, Heather Allen, & Paul Gearan (2011); 2011 Data Miner Survey Summary, presented at Predictive Analytics World, Oct. 2011.

Keim, D. A., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006) 'Challenges in Visual Data Analysis', in *Tenth International Conference on Information Visualization, 2006. IV 2006*, pp. 9–16. doi: 10.1109/IV.2006.31.

Keim, D. ., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006) 'Challenges in Visual Data Analysis', in *Tenth International Conference on Information Visualization, 2006. IV 2006*, pp. 9–16. doi: 10.1109/IV.2006.31.

Kline, P. (1994) *An Easy Guide to Factor Analysis*. Psychology Press.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. 1st ed. New York: Wiley.

Machin, S., Marie, O. and Vujić, S. (2011) 'The Crime Reducing Effect of Education*', *The Economic Journal*, 121(552), pp. 463–484. doi: 10.1111/j.1468-0297.2011.02430.x.

Machin, S., Marie, O. and Vujic, S. (2012) 'Youth Crime and Education Expansion', *German Economic Review*, 13(4), pp. 366–384. doi: 10.1111/%28ISSN%291468-0475.

Mancebon, M. J. and Molinero, C. M. (2000) 'Performance in primary schools', *Journal of the Operational Research Society*, 51(7), pp. 843–854. doi: 10.1057/palgrave.jors.2600980.

Mankin, E. (2010) *PCA How To*, *Scribd*. Available at: http://www.scribd.com/doc/39718072/PCA-How-To-1 (Accessed: 16 September 2014).

Matsui, K., Yamanouchi, M. and Sunahara, H. (2011) 'A Proposal of Framework for Information Visualization in Developing of Web Application', in *2011 IEEE/IPSJ 11th International Symposium on Applications and the Internet (SAINT)*, pp. 457–462. doi: 10.1109/SAINT.2011.85.

May, R., Hanrahan, P., Keim, D. ., Shneiderman, B. and Card, S. (2010) 'The state of visual analytics: Views on what visual analytics is and where it is going', in *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 257–259. doi: 10.1109/VAST.2010.5649078.

Muenchen, R. A. (2014) 'The Popularity of Data Analysis Software', *r4stats.com*. Available at: http://r4stats.com/articles/popularity/ (Accessed: 13 September 2014).

Norman, D. A. (1986) 'Cognitive engineering', *User centered system design*, pp. 31–61.

Norman, D.A. 2013, The design of everyday things, MIT Press, Cambridge, Mass; London.

North, C. (2006) 'Toward measuring visualization insight', *IEEE Computer Graphics and Applications*, 26(3), pp. 6–9. doi: 10.1109/MCG.2006.70.

Oates, B. J. (2006) *Researching Information Systems and Computing*. London: SAGE.

Obradovic, S. (2009) 'Education and Economic Growth', *Lex ET Scientia International Journal*, 16, p. 377.

Office of National Statistics (2013) *Chapter 1: Property Crime - Overview*, *Office for National Statistics*. Available at: http://www.ons.gov.uk/ons/rel/crime-stats/crime-statistics/focus-on-property-crime--2011-12/rpt-chapter-1-overview.html#tab-Overview-of-property-crime (Accessed: 1 August 2014).

Paternoster, R. and Bushway, S. D. (2001) 'Theoretical and Empirical Work on the Relationship Between Unemployment and Crime', *Journal of Quantitative Criminology*, 17(4), pp. 391–407. doi: 10.1023/A:1012593805457.

Peugh, J. L. and Enders, C. K. (2004) 'Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement', *Review of Educational Research*, 74(4), pp. 525–556.

Rainey, D. V. and Murova, O. (2004) 'Factors Influencing Education Achievement', *Applied Economics*, 36(21), pp. 2397–2404.

Raphael, S. and Winter- Ebmer, R. (2001) 'Identifying the Effect of Unemployment on Crime', *Journal of Law and Economics*, 44(1), pp. 259–283. doi: 10.1086/jle.2001.44.issue-1.

Robert A. Muenchen (2012). '*The Popularity of Data Analysis Software*'

Robertson, G. G., Card, S. K. and Mackinlay, J. D. (1993) 'Information Visualization Using 3D Interactive Animation', *Commun. ACM*, 36(4), pp. 57–71. doi: 10.1145/255950.153577.

Sabahat, U. (2012) 'Factors Influencing Students' Academic Performance at Higher Secondary Level: Teachers' Perception', *Communication & Mass Media Complete*, 12(9), pp. p524–548.

Salkind, N. J. and Rasmussen, K. (2007) *Encyclopedia of Measurement and Statistics*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. Available at: http://0-srmo.sagepub.com.wam.city.ac.uk/view/encyclopedia-of-measurement-and-statistics/n431.xml (Accessed: 12 September 2014).

Savikhin, A., Maciejewski, R. and Ebert, D. S. (2008) 'Applied visual analytics for economic decision-making', in *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08*, pp. 107–114. doi: 10.1109/VAST.2008.4677363.

Schoolkidsfun.com, (2014). Education in general is form of learning - school kids fun. [online] Available at: http://schoolkidsfun.com/Education.html [Accessed 26 Sep. 2014].

Sexual Offences Act 2003 (2003). Available at: http://www.legislation.gov.uk/ukpga/2003/42/contents (Accessed: 20 August 2014).

Shlens, J. (2014) 'A tutorial on principal component analysis', *arXiv preprint arXiv:1404.1100*. Available at: http://arxiv.org/abs/1404.1100 (Accessed: 16 September 2014).

Speier, C. and Morris, M. G. (2003) 'The Influence of Query Interface Design on Decision-Making Performance', *MIS Quarterly*, 27(3), pp. 397–423.

Statsoft.com, (2014). How to Reduce Number of Variables and Detect Relationships, Principal Components and Factor Analysis. [online] Available at: http://www.statsoft.com/Textbook/Principal-Components-Factor-Analysis [Accessed 17 Sep. 2014].

Stolcke, A., Kajarekar, S. and Ferrer, L. (2008) 'Nonparametric feature normalization for SVM-based speaker verification', in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pp. 1577–1580. doi: 10.1109/ICASSP.2008.4517925.

Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press.

Tufte, E. (2010). *Visual explanations*, Graphics Press

Tufte, E. R. (1990) *Envisioning information*. Graphics Press.

Ware, C. (2004) *Information Visualization: Perception for Design*. Morgan Kaufmann.

Watson, G. L., Sanders-Lawson, E. R. and McNeal, L. (2012) 'Understanding Parental Involvement in American Public Education', *International Journal of Humanities and Social Science*, 2(19). Available at: http://www.educateforexcellence.co/files/firstArticle.pdf (Accessed: 22 April 2014).

Wickham, H. (2009) *Ggplot2 elegant graphics for data analysis*. Dordrecht; New York: Springer. Available at: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=511468 (Accessed: 13 June 2014).

Van Wijk, J. J. (2005) 'The value of visualization', in *IEEE Visualization, 2005. VIS 05*, pp. 79–86. doi: 10.1109/VISUAL.2005.1532781.

Wilder, S. (2014) 'Effects of parental involvement on academic achievement: a meta-synthesis', *Educational Review*, 66(3), pp. 377–397. doi: 10.1080/00131911.2013.780009.

Wong, P. C. and Thomas, J. (2004) 'Visual Analytics', *IEEE Computer Graphics and Applications*, 24(5), pp. 20–21. doi: 10.1109/MCG.2004.39.

Wood, J., Badawood, D., Dykes, J. and Slingsby, A. (2011) 'BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers', *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp. 2384–2391. doi: 10.1109/TVCG.2011.174.

Yeomans, K. A. and Golder, P. A. (1982) 'The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors', *The Statistician*, 31(3), pp. 222–223. doi: 10.2307/2987988.

Yi, J. S., Kang, Y., Stasko, J. T. and Jacko, J. A. (2008) 'Understanding and Characterizing Insights: How Do People Gain Insights Using Information Visualization?', in *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*. New York, NY, USA: ACM (BELIV '08), pp. 4:1–4:6. doi: 10.1145/1377966.1377971.

Appendices

Appendix A

PROJECT PROPOSAL

Working Project Title: Investigating educational performance over London Boroughs through quantitative and visual analysis methods

Supervisor: Dr Cagatay Turkay

Introduction

Education is considered as a road to success. It is viewed as one of the most important things in life because without it, one may not be able to contribute positively to the world or even earn money. People oftentimes misinterpret the absence of formal education as an outright lack of knowledge. In a paper exploring the relationship between education and economic growth, Obradovic (2009) argues that "*education represents one of the primary components of human capital formation which is an important factor in modelling the endogenous production functions*". She goes on to state that for developed countries to achieve sustainable growth rates, human capital is of essential importance. For this reason, there are overwhelming studies that are concerned with forces that affect educational output from student. The majority of the findings suggest that the students' family background has a significant effect on a their performance in school, whereas school features have least effects (Brooks-Gunn and Duncan, 1997;Watson, Sanders-Lawson and McNeal, 2012).

In order to perform analysis on the data, the data type needs to be understood. Datasets that are categorical makes it a challenge for visualization. Data and visual analysis methods that usually work well with continuous datasets do not usually work well with categorical dimensions (Bendix, Kosara and Hauser, 2005). Hence, datasets that are heterogeneous in nature are challenging both for computational and visual analysis.

Researchers of information visualization(Matsui, Yamanouchi and Sunahara, 2011; Wood et al., 2011; Azeezullah et al., 2012) have used visualization techniques or created visualization tools to aid analyst, engineers, programmers etc in finding underlying patterns in large data sets. Although social scientist are constantly researching on factors that influence student performance, no research has been carried out on how visualization techniques can help the beneficiaries to find out the degree to which this factors influence student GCSE results based on spatially ordered variables.

Aims and Objectives

Overall Aims: The aim of this project is to investigate educational performance over London boroughs through the use of quantitative and visual analysis methods, taking into account the spatial relations between the variables and finding if a pattern exists over a period using quantitative data gathered by the London Datastore.

The main research question is: Does a pattern exist in the relationship between London borough profile and educational attainment?

Objectives: The objectives of the proposed project are

1. Categorize data set for data analysis.
2. Find the key main factors that may influence student grades.
3. Present and communicate findings using visualization techniques.
4. Investigate if a pattern exist in the relationship between the London borough characteristic and student grades over a period.
5. Complete the final report.

Project Products and Beneficiaries

The intended project will contribute to the already existing knowledge by providing the beneficiaries with an underlying understanding of the correlations between the two datasets, as well as possibly discovering a pattern, through the use of statistical and visualization techniques.

The intended beneficiaries of this project are policy makers, who sets plans pursued by the government; Local education authorities, who are the local councils in England and Wales responsible for the implementation of

educational laws in a system; social scientists in the educational field, who will want to adapt this technique in discovering more patterns; and Parents, who together with the other beneficiaries, would need this information to be able to make more informed decisions regarding education, like providing facilities and environment that will support learning and subsequently enhance students' academic achievements.

I have shown evidence by the reviewing literature that there exist a research gap in exploring whether the London borough variables can influence student GCSE grades. To achieve the objectives of this project, I will use the data from two datasets mentioned later. The proposal will be organized as follows. First, a review of the relevant literature to ground my understanding of the subject matter. Second, I will provide a detailed description of the methods and tools to be used for analysis and evaluation. Third, the project plan, project risk and a discussion of the ethical issue that may arise will be discussed. Finally, the references and ethic checklist will be provided.

Critical Context

The motivation behind this project is to contribute to the already existing knowledge comprising the support of the education sector's business processes with information technology. For the purpose of this project, the emphasis will be on providing social scientist with an underlying understanding of the influence of neighbourhood factors on student performance through the use of a different data analysis tool and visual analysis techniques. In the context of this research, the emphasis will be on the extent to which London borough characteristics influence GCSE results of students.

As with nearly all organisations, there is a need to collect and store huge amounts of raw data that can be used to provide useful information. In many research experiment, the collection of data are usually analysed in many different ways depending on the data type - to find relationships or patterns, however the important issue is how the results are communicated to the right bodies in order to find it useful. Previous work have found that research participants given a task in a visual form performed better than participants given the same task in plain text (Dickson, DeSanctis and McBride, 1986). Also, it

is found that decision makers memory work load is significantly lower when using visual interfaces, no matter how complex the task is, in comparison to text interfaces (Speier and Morris, 2003).

In a paper about the value of visualization, van Wijk (2005) stated that "*visualization of data makes it possible for researchers, analysts, engineers, and the lay audience to obtain insight in these data in an efficient and effective way…which enables us to detect interesting features and patterns in a short time*". He goes on to provide examples of the techniques in which visualizations can be used, one of which includes exploration (where users do not know what is in the data but want to gain more insight) and presentation (where information or findings has to be communicated to others). He also stated that researchers consider exploration as the major reason for visualization, where as presentation is not that complex. However, from his own experience, both these techniques are equally important.

To make informed decisions, we need to identify and visualize the knowledge contained in large datasets, through the process of filtering and analysing the data (Holtz et al., 2011). "*Visual analytics is the science of analytical reasoning assisted by interactive visual interfaces, which has already been applied and found to be effective in social sciences such as management, finance, marketing and organizational behaviour to aid in decision making*"(Savikhin, Maciejewski and Ebert, 2008). It focuses on processing managing heterogenous and dynamic volumes of information through visual representations and interaction techniques(D. A. Keim et al., 2006).

In the United Kingdom, individuals are expected to spend a minimum of eighteen years in school (primary and secondary). Although a majority of this time is spent within the confines of a school environment, it is very likely that internal and external factors would significantly influence students' educational performance. There is lots of research done on the factors influencing student performance.

In a paper concerned about the impact of environment on engineering student's academic performance in Malaysia, Holtz et al. (2011) selected environment, teaching methodology, teaching aid, student attitude and

lecturers involvement as parameters and collected data with the use of questionnaires. They found that environmental factors had more impact on student performance which they communicated via the use of the presentation visualization technique.

In another paper focusing on the performance of pupils and schools in east London, and on the importance of ethnicity, school characteristic and social background in influencing educational achievement - using the Pupil Level Annual School Survey (PLASC) data source of 2003. The data set contained 17,891 pupils at key stage 4 (i.e. GCSE). The analysis was only based across seven east London boroughs. It was found that social background was the important factor influencing educational attainment, which was also communicated via the use of presentation visualization technique as well as plain text(Hamnett, Ramsden and Butler, 2007).

Similarly, a study was carried out to examine the factors that influence students mainly in rural communities in the United States of America through the use of regression analysis. Results showed that parents' educational achievement greatly influences their child's performance (Rainey and Murova, 2004). Other studies have shown that a combination of some factors have a strong influence in a child's educational achievement, especially at the primary school level through the use of data envelopment analysis (Mancebon and Molinero, 2000;Carron and Chau, 1996).

A review of these literatures reveals that there is little evidence of the application of visual analytics techniques by social scientist in analysing the spatial relations of the factors that influence student performance over a period to enhance decision making. Also, there was no evidence of using R[1] to analyse the data gathered. In a book on the "Introduction to R", R was stated to be more efficient than other data analysis tool, because it is a programming environment within which statistical analysis and visualization is conducted (Clark, 2014) which is in line with the purpose of the proposed project stated earlier.

[1] R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible (http://www.r-project.org/about.html)

This project seeks to provide a relevant evidence of the applicability of the statistical analysis and visualization tool for analysing the degree and possibly finding a pattern in the factors that influence student performance. By providing a means of visualizing the broad range of factors associated with different London boroughs.

Methods and Tools for Analysis and Evaluation

In order to achieve the objectives of the project, thus answering the research question, Oates(2006) suggest that a research strategy approach should be adopted. In this case the experiment approach will be used. This approach is defined by oates as a "*strategy that investigates cause and effect relationships, seeking to prove or disprove a causal link between a factor and an observed outcome…*" I will experimentally examine the extent to which student grades over a period, are influenced by neighbourhood characteristics, as well as evaluate two hypotheses. *Hypothesis 1* states that there exist a significant pattern between student performance and factors influencing them over time. The other hypothesis, *Hypothesis 2,* states that there exist no significant pattern between student performance and factors influencing them over time.

Statistical analysis will be performed on two publicised quantitative datasets. Public datasets would be used because they contain data that are usually accurate and more importantly, datasets that contained the variables relevant for this research was already published online. The links to these datasets are http://data.london.gov.uk/datastore/package/london-borough-profiles and http://data.london.gov.uk/datastore/package/gcse-results-location-pupil-residence-borough and they comprise mainly of continuous data.

The independent variables will be the indicators in the London borough profile dataset; where as the dependent variables will be the test score for both genders in the GCSE result by location of pupil dataset. Although they are of the continuous data type, they will be treated as a categorical data type. An example is shown below

London borough profile (LBP)

	Population		
Area name	Proportion of population **aged 0-15**, 2013	Proportion of population of **working-age**, 2013	Proportion of population aged **65 and over**, 2013
city of london	7.9	77.5	14.5
barking and Dagenham	26.1	63.8	10.1
Barnet	20.9	65.5	13.6

GCSE Result by Location of pupil residence

	Percentage	
Area	All Pupils at the End of KS4 Achieving 5+ A* - G	All Pupils at the End of KS4 Achieving 5+ A* - C
City of London	75	100
Barking and Dagenham	59.3	94.3
Barnet	73.1	94.6

To determine the key indicators of each factor that may influence student test scores, factor analysis will be applied to the indicators of each London borough factor. Factor analysis "*consists of a number of statistical techniques, the aim of which is to simplify complex sets of data*."(Kline, 1994). He goes on to state that in social sciences, factor analysis is usually applied to correlations between variables…

Finding correlations between the LBP variables will be a very complex activity, seeing as there are a lot of values for each variable. Rather than having to look at the values of each variable to understand the correlations, I could understand them in terms of a number of factor loadings ("*a correlation of a variable with a factor*" (kline, 1994)). Therefore, factor analysis will be able to simplify, as well as illustrate the key factors relevant to each London borough.

Once the key factors are selected, Kendall tau rank correlation coefficient will be used to find the correlation between these key factors and student GCSE result for each related year. This statistical method is used to measure association between two measured quantities of categorical data type (George, 1948). A significance test will be used to test whether these two variables are statistically dependent or independent.

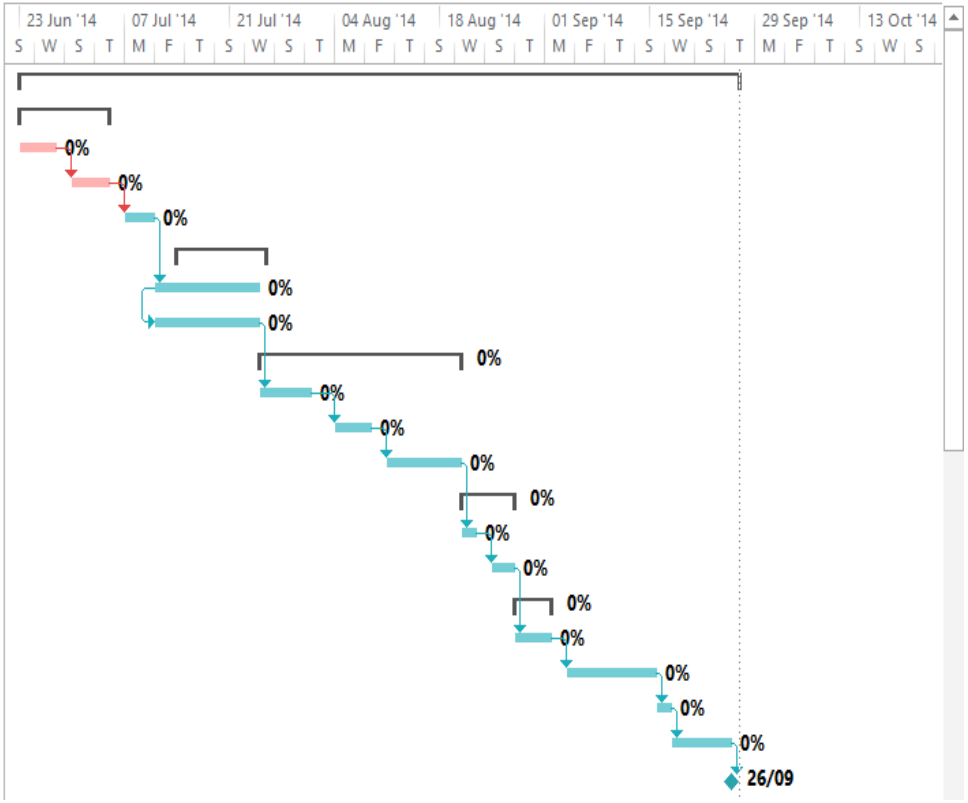
The tool which is going to be used for this project for performing the analysis and visualizing the patterns is R using the S[2] programming language, for reasons already stated. To communicate the findings and patterns, ggplot2 package will be used. It is a data visualization package in R that provides elegant graphics for data analysis (Wickham, 2009). Scatter plot matrix will then be used to present the results. A scatter plot is a type of graphic that is provided in the ggplot2 package. A scatter plot matrix is a great way to roughly determine if you have a linear correlation between multiple ordered categorical data(Galili, 2010). If a pattern exists in the findings, *Hypothesis 2* will be rejected.

[2] S is a statistical programming language developed primarily by John Chambers and (in earlier versions) Rick Becker and Allan Wilks of Bell Laboratories. The aim of the language, as expressed by John Chambers, is "to turn ideas into software, quickly and faithfully.(http://en.wikipedia.org/wiki/S_(programming_language)

PLAN OF WORK

The following figure shows the task I would need to accomplish to complete the proposed project.

Task Name	Start	Finish
Skeletal Proposed Project Plan	**Mon 23/06/14**	**Fri 26/09/14**
Literature Survey	**Mon 23/06/14**	**Fri 04/07/14**
Literature Search	Mon 23/06/14	Fri 27/06/14
Literature Review	Mon 30/06/14	Fri 04/07/14
Work on Report	Mon 07/07/14	Thu 10/07/14
Data Collection	**Mon 14/07/14**	**Fri 25/07/14**
Arrange Datasets	Fri 11/07/14	Thu 24/07/14
Clean Datasets (Error correction)	Fri 11/07/14	Thu 24/07/14
Data Analysis	**Fri 25/07/14**	**Wed 20/08/14**
Learn R Programming Language	Fri 25/07/14	Thu 31/07/14
Perform statistical analysis	Mon 04/08/14	Fri 08/08/14
Develop Visual designs to display patterns	Mon 11/08/14	Wed 20/08/14
Discussion	**Thu 21/08/14**	**Wed 27/08/14**
Discuss Findings with the supervisor	Thu 21/08/14	Fri 22/08/14
Draft Discussion Report	Mon 25/08/14	Wed 27/08/14
Evaluation and recommendation	**Thu 28/08/14**	**Mon 01/09/14**
Revise, Draft evaluation results	Thu 28/08/14	Mon 01/09/14
Draft final project report	Thu 04/09/14	Mon 15/09/14
Get feedback from supervisor	Tue 16/09/14	Wed 17/09/14
Make corrections	Thu 18/09/14	Thu 25/09/14
Submit Final project	Fri 26/09/14	Fri 26/09/14



RISKS

The table below identifies the risks that could occur and how they will be mitigated or recovered from, with likelihood (1) indicating low and (3) high, whereas the impact (1) indicating very low and (5) indicating very high.

NO.	RISKS	RISK TRIGGERS	LIKELIHOOD(1-3)	IMPACT(1-5)	MITIGATE	CONTINGENCY PLAN
1	Inability to meet project deadline	(i) Missing preliminary milestones	2	5	Mitigate: (i) Carry out a feasibility study to determine likeliness of the project completion within specified timeframe. (ii)Keeping the supervisor up to date with the progress of project. (iii)Increase project working hours	Provide a valid reason for not meeting the deadline to the programmes office for a request of an extension.
2	Loss of stored project work	(i) Corrupted files. (ii)Missing files. (iii) Hard Drive crashing.	2	5	Mitigate: Having various storage locations.	(i) Having an alternative machine to use. (ii) Resume work from the last saved work. (iii) Increase project working hours to meet up with the project deadline
3	Inadequate attention to detail	High frequency of errors	2	3	Mitigate :(i) Having checkpoints every week or after a task to go over my work. (ii) Having someone else/supervisor go through my work.	
4	Insufficient time to learn how to use the R software	Spending more time than allocated for this task	3	5	Mitigate: Delegating enough time to learn how to use R.	
5	Change of requirements	(i)Change of project aim and objectives. (ii)Discovering study during literature search	2	5	Mitigate: Making sure I carried out an extensive research before finalising my research topic	Consult with my supervisor and find a new project topic or a different direction as soon as possible to meet up with the project deadline

		that covers what I propose doing				
6	Project been beyond my technical ability	Struggling with a straight-forward implementation of a component in a new programming language	2	4	Mitigate: Choosing a tool that uses a programming language I am familiar with.	Find a tutor familiar with the programming language to put me through. But making sure I follow due procedures by getting consent from my supervisor or the project officers.

Ethical, Professional & Legal Issues

According to Oates (2006), "*researchers must treat everyone involved in your research, whether directly or indirectly fairly and with honesty*", in other words, they should be ethical researchers. She goes on to state the rights of the participants taking part in a research. These are "*right not to participate…to withdraw…to give informed consent...to anonymity…to confidentiality*". A researcher must also make sure that all legal issues regarding the topic are handled or be aware of the legal issues in their respective country.

For this proposed project, I do not foresee any ethical issues coming up. Due to not having any participants involved with my research. Also, no professional & legal issues because the data and tool to be used for my research is free to the public.

Refer to Appendix A for the Research ethics Checklist.

References

Azeezullah, I., Pambudi, F., Shyy, T., Azeezullah, I., Ward, N., Hunter, J. and Stimson, R. J. (2012) 'Statistical analysis and visualization services for Spatially Integrated Social Science datasets', in *2012 IEEE 8th International Conference on E-Science (e-Science)*, pp. 1–8. doi: 10.1109/eScience.2012.6404421.

Becker, G. S. (1974) *Essays in the economics of crime and punishment*. New York, NY: National Bureau of Economic Research.

Bendix, F., Kosara, R. and Hauser, H. (2005) 'Parallel sets: visual analysis of categorical data', in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pp. 133–140. doi: 10.1109/INFVIS.2005.1532139.

Bretaña, G. (1965) *Firearms. Act 1965*. Her majesty's stationery office.

Card, S. K., Mackinlay, J. D. and Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.

Cattell, R. B. (1966) 'The Scree Test For The Number Of Factors', *Multivariate Behavioral Research*, 1(2), pp. 245–276. doi: 10.1207/s15327906mbr0102_10.

Davies, T. (2010) 'Open data, democracy and public sector reform', *A look at open government data use from data. gov. uk*. Available at: http://www.opendataimpacts.net/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf (Accessed: 8 September 2014).

Dickson, G. W., DeSanctis, G. and McBride, D. J. (1986) 'Understanding the Effectiveness of Computer Graphics for Decision Support: A Cumulative Experimental Approach', *Commun. ACM*, 29(1), pp. 40–47. doi: 10.1145/5465.5469.

Educational Act (2002) *Education Act 2002*. Available at: http://www.legislation.gov.uk/ukpga/2002/32/section/82 (Accessed: 4 August 2014).

Elliott, C. and Ellingworth, D. (1996) 'The relationship between unemployment and crime: A cross-sectional analysis employing the British Crime Survey 1992', *International Journal of Manpower*, 17(6/7), pp. 81–88. doi: 10.1108/01437729610149358.

Few, S. (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press.

Fox, J. and Andersen, R. (2005) 'Using the R statistical computing environment to teach social statistics courses', *Department of Sociology, McMaster University*. Available at: http://socserv.socsci.mcmaster.ca/jfox/Teaching-with-R.pdf (Accessed: 13 September 2014).

Fry, B. (2003) *Visualizing data*. O'Reilly. Available at: https://www.dawsonera.com/abstract/9780596519308 (Accessed: 7 August 2014).

Galili, T. (2010) *Correlation scatter-plot matrix for ordered-categorical data*, *R-statistics blog*. Available at: http://www.r-statistics.com/2010/04/correlation-scatter-plot-matrix-for-ordered-categorical-data/ (Accessed: 13 June 2014).

George, M. K. (1948) *Rank correlation methods*. Oxford, England: Griffin.

Gershon, N. and Eick, S. G. (1995) 'Visualization's new tack: making sense of information', *IEEE Spectrum*, 32(11), pp. 38–40, 42, 44–7, 55–6. doi: 10.1109/6.469330.

Gershon, N., Eick, S. G. and Card, S. (1998) 'Information Visualization', *interactions*, 5(2), pp. 9–15. doi: 10.1145/274430.274432.

Gonzalez, V. and Kobsa, A. (2003) 'Benefits of information visualization systems for administrative data analysts', in *Seventh International Conference on Information Visualization, 2003. IV 2003. Proceedings*, pp. 331–336. doi: 10.1109/IV.2003.1217999.

Guttman, L. (1954) 'Some necessary conditions for common-factor analysis', *Psychometrika*, 19(2), pp. 149–161. doi: 10.1007/BF02289162.

Hale, C. and Sabbagh, D. (1991) 'Testing the Relationship between Unemployment and Crime: A Methodological Comment and Empirical Analysis Using Time Series Data from England and Wales', *Journal of Research in Crime and Delinquency*, 28(4), pp. 400–417. doi: 10.1177/0022427891028004002.

Hall, W., Shadbolt, N., Tiropanis, T., O'Hara, K. and Davies, T. (2012) 'Open data and charities'. Available at: http://eprints.soton.ac.uk/341346/ (Accessed: 8 September 2014).

Hamnett, C., Ramsden, M. and Butler, T. (2007) 'Social Background, Ethnicity, School Composition and Educational Attainment in East London', *Urban Studies*, 44(7), pp. 1255–1280. doi: 10.1080/00420980701302395.

Holtz, S., Valle, G., Howard, J. and Morreale, P. (2011) 'Visualization and pattern identification in large scale time series data', in *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 129–130. doi: 10.1109/LDAV.2011.6092333.

Hutchins, E. L., Hollan, J. D. and Norman, D. A. (1985) 'Direct manipulation Interfaces', *Human-Computer Interaction*, 1(4), p. 311.

Ihaka, R. and Gentleman, R. (1996) 'R: A Language for Data Analysis and Graphics', *Journal of Computational and Graphical Statistics*, 5(3), pp. 299–314. doi: 10.2307/1390807.

Kaiser, H. F. (1960) 'The application of electronic computers to factor analysis', *Educational and Psychological Measurement*, 20, pp. 141–151. doi: 10.1177/001316446002000116.

Kaiser, H. F. (1961) 'A Note on Guttman's Lower Bound for the Number of Common Factors1', *British Journal of Statistical Psychology*, 14(1), pp. 1–2. doi: 10.1111/j.2044-8317.1961.tb00061.x.

Keim, D. A., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006) 'Challenges in Visual Data Analysis', in *Tenth International Conference on Information Visualization, 2006. IV 2006*, pp. 9–16. doi: 10.1109/IV.2006.31.

Keim, D. ., Mansmann, F., Schneidewind, J. and Ziegler, H. (2006) 'Challenges in Visual Data Analysis', in *Tenth International Conference on Information Visualization, 2006. IV 2006*, pp. 9–16. doi: 10.1109/IV.2006.31.

Kline, P. (1994) *An Easy Guide to Factor Analysis*. Psychology Press.

Machin, S., Marie, O. and Vujić, S. (2011) 'The Crime Reducing Effect of Education*', *The Economic Journal*, 121(552), pp. 463–484. doi: 10.1111/j.1468-0297.2011.02430.x.

Machin, S., Marie, O. and Vujic, S. (2012) 'Youth Crime and Education Expansion', *German Economic Review*, 13(4), pp. 366–384. doi: 10.1111/%28ISSN%291468-0475.

Mancebon, M. J. and Molinero, C. M. (2000) 'Performance in primary schools', *Journal of the Operational Research Society*, 51(7), pp. 843–854. doi: 10.1057/palgrave.jors.2600980.

Mankin, E. (2010) *PCA How To*, *Scribd*. Available at: http://www.scribd.com/doc/39718072/PCA-How-To-1 (Accessed: 16 September 2014).

Matsui, K., Yamanouchi, M. and Sunahara, H. (2011) 'A Proposal of Framework for Information Visualization in Developing of Web Application', in *2011 IEEE/IPSJ 11th International Symposium on Applications and the Internet (SAINT)*, pp. 457–462. doi: 10.1109/SAINT.2011.85.

May, R., Hanrahan, P., Keim, D. ., Shneiderman, B. and Card, S. (2010) 'The state of visual analytics: Views on what visual analytics is and where it is going', in *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 257–259. doi: 10.1109/VAST.2010.5649078.

Muenchen, R. A. (2014) 'The Popularity of Data Analysis Software', *r4stats.com*. Available at: http://r4stats.com/articles/popularity/ (Accessed: 13 September 2014).

Norman, D. A. (1986) 'Cognitive engineering', *User centered system design*, pp. 31–61.

North, C. (2006) 'Toward measuring visualization insight', *IEEE Computer Graphics and Applications*, 26(3), pp. 6–9. doi: 10.1109/MCG.2006.70.

Oates, B. J. (2006) *Researching Information Systems and Computing*. London: SAGE.

Obradovic, S. (2009) 'Education and Economic Growth', *Lex ET Scientia International Journal*, 16, p. 377.

Office of National Statistics (2013) *Chapter 1: Property Crime - Overview*, *Office for National Statistics*. Available at: http://www.ons.gov.uk/ons/rel/crime-stats/crime-statistics/focus-on-property-crime--2011-12/rpt-chapter-1-overview.html#tab-Overview-of-property-crime (Accessed: 1 August 2014).

Paternoster, R. and Bushway, S. D. (2001) 'Theoretical and Empirical Work on the Relationship Between Unemployment and Crime', *Journal of Quantitative Criminology*, 17(4), pp. 391–407. doi: 10.1023/A:1012593805457.

Peugh, J. L. and Enders, C. K. (2004) 'Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement', *Review of Educational Research*, 74(4), pp. 525–556.

Rainey, D. V. and Murova, O. (2004) 'Factors Influencing Education Achievement', *Applied Economics*, 36(21), pp. 2397–2404.

Raphael, S. and Winter-Ebmer, R. (2001) 'Identifying the Effect of Unemployment on Crime', *Journal of Law and Economics*, 44(1), pp. 259–283. doi: 10.1086/jle.2001.44.issue-1.

Robertson, G. G., Card, S. K. and Mackinlay, J. D. (1993) 'Information Visualization Using 3D Interactive Animation', *Commun. ACM*, 36(4), pp. 57–71. doi: 10.1145/255950.153577.

Sabahat, U. (2012) 'Factors Influencing Students' Academic Performance at Higher Secondary Level: Teachers' Perception', *Communication & Mass Media Complete*, 12(9), pp. p524–548.

Salkind, N. J. and Rasmussen, K. (2007) *Encyclopedia of Measurement and Statistics*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. Available at: http://0-srmo.sagepub.com.wam.city.ac.uk/view/encyclopedia-of-measurement-and-statistics/n431.xml (Accessed: 12 September 2014).

Savikhin, A., Maciejewski, R. and Ebert, D. S. (2008) 'Applied visual analytics for economic decision-making', in *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08*, pp. 107–114. doi: 10.1109/VAST.2008.4677363.

Sexual Offences Act 2003 (2003). Available at: http://www.legislation.gov.uk/ukpga/2003/42/contents (Accessed: 20 August 2014).

Shlens, J. (2014) 'A tutorial on principal component analysis', *arXiv preprint arXiv:1404.1100*. Available at: http://arxiv.org/abs/1404.1100 (Accessed: 16 September 2014).

Speier, C. and Morris, M. G. (2003) 'The Influence of Query Interface Design on Decision-Making Performance', *MIS Quarterly*, 27(3), pp. 397–423.

Stolcke, A., Kajarekar, S. and Ferrer, L. (2008) 'Nonparametric feature normalization for SVM-based speaker verification', in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pp. 1577–1580. doi: 10.1109/ICASSP.2008.4517925.

Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press.

Tufte, E. R. (1990) *Envisioning information*. Graphics Press.

Ware, C. (2004) *Information Visualization: Perception for Design*. Morgan Kaufmann.

Watson, G. L., Sanders-Lawson, E. R. and McNeal, L. (2012) 'Understanding Parental Involvement in American Public Education', *International Journal of Humanities and Social Science*, 2(19). Available at: http://www.educateforexcellence.co/files/firstArticle.pdf (Accessed: 22 April 2014).

Wickham, H. (2009) *Ggplot2 elegant graphics for data analysis*. Dordrecht; New York: Springer. Available at: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=511468 (Accessed: 13 June 2014).

Van Wijk, J. J. (2005) 'The value of visualization', in *IEEE Visualization, 2005. VIS 05*, pp. 79–86. doi: 10.1109/VISUAL.2005.1532781.

Wilder, S. (2014) 'Effects of parental involvement on academic achievement: a meta-synthesis', *Educational Review*, 66(3), pp. 377–397. doi: 10.1080/00131911.2013.780009.

Wong, P. C. and Thomas, J. (2004) 'Visual Analytics', *IEEE Computer Graphics and Applications*, 24(5), pp. 20–21. doi: 10.1109/MCG.2004.39.

Wood, J., Badawood, D., Dykes, J. and Slingsby, A. (2011) 'BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers', *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp. 2384–2391. doi: 10.1109/TVCG.2011.174.

Yeomans, K. A. and Golder, P. A. (1982) 'The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors', *The Statistician*, 31(3), pp. 222–223. doi: 10.2307/2987988.

Yi, J. S., Kang, Y., Stasko, J. T. and Jacko, J. A. (2008) 'Understanding and Characterizing Insights: How Do People Gain Insights Using Information Visualization?', in *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization*. New York, NY, USA: ACM (BELIV '08), pp. 4:1–4:6. doi: 10.1145/1377966.1377971.

Research Ethics Checklist

School of Informatics BSc MSc/MA Projects

If the answer to any of the following questions (1 – 3) is NO, your project needs to be modified.

1.	Does your project pose only minimal and predictable risk to you (the student)?	**Yes**
2.	Does your project pose only minimal and predictable risk to other people affected by or participating in the project?	**Yes**
3.	Is your project supervised by a member of academic staff of the School of Informatics or another individual approved by the module leaders?	**Yes**

If the answer to either of the following questions (4 – 5) is YES, you MUST apply to the University Research Ethics Committee for approval. (You should seek advice about this from your project supervisor at an early stage.)

4.	Does your project involve animals?	**No**
5.	Does your project involve pregnant women or women in labour?	**No**

If the answer to the following question (6) is YES, you MUST complete the remainder of this form (7 – 19). If the answer is NO, you are finished.

6.	Does your project involve human participants? For example, as interviewees, respondents to a questionnaire or participants in evaluation or testing?	**No**

APPENDIX B: R Programming Commands used in Performing Data Analysis in RStudio

B1: Commands Used in Performing the Standardization process For Both Data Sets.

```
#Standardised crime sorted data
crimeStandard08 <- as.data.frame(scale(crimeSorted[1:32,3:10]))
crimeStandard09 <- as.data.frame(scale(crimeSorted[33:64,3:10]))
crimeStandard10 <- as.data.frame(scale(crimeSorted[65:96,3:10]))
crimeStandard11 <- as.data.frame(scale(crimeSorted[97:128,3:10]))
crimeStandard12 <- as.data.frame(scale(crimeSorted[129:160,3:10]))
crimeStandard13 <- as.data.frame(scale(crimeSorted[161:192,3:10]))

#Wrote eigenvalues to csv
write.csv(crimeStandard08,file="crimeStandard08.csv")
write.csv(crimeStandard09,file="crimeStandard09.csv")
write.csv(crimeStandard10,file="crimeStandard10.csv")
write.csv(crimeStandard11,file="crimeStandard11.csv")
write.csv(crimeStandard12,file="crimeStandard12.csv")
write.csv(crimeStandard13,file="crimeStandard13.csv")
```

```
#Standardised GCSE sorted data
GcseStandard08 <- as.data.frame(scale(GcseSorted[1:32,3:4]))
GcseStandard09 <- as.data.frame(scale(GcseSorted[33:64,3:4]))
GcseStandard10 <- as.data.frame(scale(GcseSorted[65:96,3:4]))
GcseStandard11 <- as.data.frame(scale(GcseSorted[97:128,3:4]))
GcseStandard12 <- as.data.frame(scale(GcseSorted[129:160,3:4]))
GcseStandard13 <- as.data.frame(scale(GcseSorted[161:192,3:4]))

#Wrote eigenvalues to csv
write.csv(GcseStandard08,file="GcseStandard08.csv")
write.csv(GcseStandard09,file="GcseStandard09.csv")
write.csv(GcseStandard10,file="GcseStandard10.csv")
write.csv(GcseStandard11,file="GcseStandard11.csv")
write.csv(GcseStandard12,file="GcseStandard12.csv")
write.csv(GcseStandard13,file="GcseStandard13.csv")
```

B2: Commands used for Performing Pearson's Correlation and Plotting the Scatter Diagrams for Both Datasets

```
#Function to show the correlation between crime rates variables
in descending order
mosthighlycorrelated <- function(mydataframe,numtoreport)
{
  # found the correlations
  cormatrix <- cor(mydataframe)
  # set the correlations on the diagonal or lower triangle to
zero,
  # so they will not be reported as the highest ones:
  diag(cormatrix) <- 0
  cormatrix[lower.tri(cormatrix)] <- 0
  # flattened the matrix into a dataframe for easy sorting
  fm <- as.data.frame(as.table(cormatrix))
  # assigned human-friendly names
  names(fm) <- c("First.Variable",
```

```
"Second.Variable","Correlation")
  # sorted and printed the top n correlations

head(fm[order(abs(fm$Correlation),decreasing=T),],n=numtoreport
)
}

#called the function for year 2008
mostHighlyCorrelatedCrime08<-
mosthighlycorrelated(crimeSorted[1:32,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime08,file="mostHighlyCorrelate
dCrime08.csv")
#Loaded gclus package to color code the scatterplot Matrix
library(gclus)
#plotted a scatterplot matrix of crime sorted file for the year
2008
crimeMatrix08<-abs(cor(crimeSorted[1:32,3:10]))#Got
Correlations Values
matrixColor08<-dmat.color(crimeMatrix08)#Got Color
matrixOrdered08<-order.single(crimeMatrix08)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[1:32,3:10], matrixOrdered08,
panel.colors=matrixColor08, gap=.5,
       main="Variables Ordered and Colored by Correlation 2008"
)#Plotted the Colored Scatterplot

#called the function for year 2009
mostHighlyCorrelatedCrime09<-
mosthighlycorrelated(crimeSorted[33:64,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime09,file="mostHighlyCorrelate
dCrime09.csv")
#plotted a scatterplot matrix of crime sorted file for the year
2009
crimeMatrix09<-abs(cor(crimeSorted[33:64,3:10]))#Got
Correlations Values
matrixColor09<-dmat.color(crimeMatrix09)#Got Color
matrixOrdered09<-order.single(crimeMatrix09)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[33:64,3:10], matrixOrdered09,
panel.colors=matrixColor09, gap=.5,
       main="Variables Ordered and Colored by Correlation 2009"
)#Plotted the Colored Scatterplot

#called the function for year 2010
mostHighlyCorrelatedCrime10<-
mosthighlycorrelated(crimeSorted[65:96,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime10,file="mostHighlyCorrelate
dCrime10.csv")
#plotted a scatterplot matrix of crime sorted file for the year
2009
crimeMatrix10<-abs(cor(crimeSorted[65:96,3:10]))#Got
Correlations Values
matrixColor10<-dmat.color(crimeMatrix10)#Got Color
matrixOrdered10<-order.single(crimeMatrix10)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[65:96,3:10], matrixOrdered10,
panel.colors=matrixColor10, gap=.5,
```

```
       main="Variables Ordered and Colored by Correlation 2010"
)#Plotted the Colored Scatterplot


#called the function for year 2011
mostHighlyCorrelatedCrime11<-
mosthighlycorrelated(crimeSorted[97:128,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime11,file="mostHighlyCorrelate
dCrime11.csv")
#plotted a scatterplot matrix of crime sorted file for the year
2009
crimeMatrix11<-abs(cor(crimeSorted[97:128,3:10]))#Got
Correlations Values
matrixColor11<-dmat.color(crimeMatrix11)#Got Color
matrixOrdered11<-order.single(crimeMatrix11)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[97:128,3:10], matrixOrdered11,
panel.colors=matrixColor11, gap=.5,
       main="Variables Ordered and Colored by Correlation 2011"
)#Plotted the Colored Scatterplot


#called the function for year 2012
mostHighlyCorrelatedCrime12<-
mosthighlycorrelated(crimeSorted[129:160,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime12,file="mostHighlyCorrelate
dCrime12.csv")
#plotted a scatterplot matrix of crime sorted file for the year
2009
crimeMatrix12<-abs(cor(crimeSorted[129:160,3:10]))#Got
Correlations Values
matrixColor12<-dmat.color(crimeMatrix12)#Got Color
matrixOrdered12<-order.single(crimeMatrix12)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[129:160,3:10], matrixOrdered12,
panel.colors=matrixColor12, gap=.5,
       main="Variables Ordered and Colored by Correlation 2012"
)#Plotted the Colored Scatterplot

#called the function for year 2013
mostHighlyCorrelatedCrime13<-
mosthighlycorrelated(crimeSorted[161:192,3:10], 8)
#Wrote to a csv file
write.csv(mostHighlyCorrelatedCrime13,file="mostHighlyCorrelate
dCrime13.csv")
#plotted a scatterplot matrix of crime sorted file for the year
2009
crimeMatrix13<-abs(cor(crimeSorted[161:192,3:10]))#Got
Correlations Values
matrixColor13<-dmat.color(crimeMatrix13)#Got Color
matrixOrdered13<-order.single(crimeMatrix13)# reordered
variables so those with highest correlation
# are closest to the diagonal
cpairs(crimeSorted[161:192,3:10], matrixOrdered13,
panel.colors=matrixColor13, gap=.5,
       main="Variables Ordered and Colored by Correlation 2013"
)#Plotted the Colored Scatterplot
```

```
#Found the correlation between the GCSE variables for the year
2008
cor.test(GcseSorted[1:32,3], GcseSorted[1:32,4])

#Found the correlation between the GCSE variables for the year
2009
cor.test(GcseSorted[33:64,3], GcseSorted[33:64,4])

#Found the correlation between the GCSE variables for the year
2010
cor.test(GcseSorted[65:96,3], GcseSorted[65:96,4])

#Found the correlation between the GCSE variables for the year
2011
cor.test(GcseSorted[97:128,3], GcseSorted[97:128,4])

#Found the correlation between the GCSE variables for the year
2012
cor.test(GcseSorted[129:160,3], GcseSorted[129:160,4])

#Found the correlation between the GCSE variables for the year
2013
cor.test(GcseSorted[161:192,3], GcseSorted[161:192,4])


#Created a Line gragh to ciew the difference in correlation
over time
# Defined the Gcsecorrelation vector with the correlation
values raised to three decimal places for each year
GcseCorrelation <- c(0.582, 0.662, 0.803, 0.728, 0.642, 0.470 )

# Calculated range from 0 to max value of GcseCorrelation
> g_range <- range(0.1,GcseCorrelation )

# Graphed GcseCorrelation using red points overlayed by a line
plot(GcseCorrelation, type="o", col="red",axes=FALSE,
ann=FALSE)

# Made x axis using yearlabels
axis(1, at=1:6, lab=c(2008,2009,2010,2011,2012, 2013))

# Made y axis with horizontal labels that displayed ticks at
# every 0.1 mark.
axis(2,las=1,at=seq(0, 0.9, by = 0.1))

# Created a title with a black, bold/italic font
title(main="Gcse Correlation By Time", col.main="black",
font.main=4)

# Created a box around plot
box()

# Labelled the x and y axes with dark green text
title(xlab="years", col.lab=rgb(0,0.5,0))
title(ylab="CorrelatedValues", col.lab=rgb(0,0.5,0))

#Plotted a scatterplot matrix for each time period
#scatterplot for the year 2008
plot(GcseSorted[1:32,3], GcseSorted[1:32,4],main="GCSE
Scatterplot 2008",,xlab="LevelTwo",ylab="LevelOne")

#scatterplot for the year 2009
plot(GcseSorted[33:64,3], GcseSorted[33:64,4],main="GCSE
```

```
Scatterplot 2009",xlab="LevelTwo",ylab="LevelOne")

#scatterplot for the year 2010
plot(GcseSorted[65:96,3], GcseSorted[65:96,4],main="GCSE
Scatterplot 2010",xlab="LevelTwo",ylab="LevelOne")

#scatterplot for the year 2011
plot(GcseSorted[97:128,3], GcseSorted[97:128,4],main="GCSE
Scatterplot 2011",xlab="LevelTwo",ylab="LevelOne")

#scatterplot for the year 2012
plot(GcseSorted[129:160,3], GcseSorted[129:160,4],main="GCSE
Scatterplot 2012",xlab="LevelTwo",ylab="LevelOne")

#scatterplot for the year 2013
plot(GcseSorted[161:192,3], GcseSorted[161:192,4],main="GCSE
Scatterplot 2013",xlab="LevelTwo",ylab="LevelOne")
```

B3: Commands Used In Performing PCA (Calculating The Eigenvalues, The Factor Loadings, And The PCA Scores)

```
#called prcomp function
crime.pca08 <- prcomp(crimeStandard08)
crime.pca09 <- prcomp(crimeStandard09)
crime.pca10 <- prcomp(crimeStandard10)
crime.pca11 <- prcomp(crimeStandard11)
crime.pca12 <- prcomp(crimeStandard12)
crime.pca13 <- prcomp(crimeStandard13)

#Summarized the PC of each variable
summary(crime.pca08)
summary(crime.pca09)
summary(crime.pca10)
summary(crime.pca11)
summary(crime.pca12)
summary(crime.pca13)

#Screeplot of the PC by year
screeplot(crime.pca08, xlab="Components")
screeplot(crime.pca09, xlab="Components")
screeplot(crime.pca10, xlab="Components")
screeplot(crime.pca11, xlab="Components")
screeplot(crime.pca12, xlab="Components")
screeplot(crime.pca13, xlab="Components")

#Found the eigenvalues to enable retain the necessary PC
eigenValues08<-crime.pca08$sdev^2
eigenValues09<-crime.pca09$sdev^2
eigenValues10<-crime.pca10$sdev^2
eigenValues11<-crime.pca11$sdev^2
eigenValues12<-crime.pca12$sdev^2
eigenValues13<-crime.pca13$sdev^2

#Wrote eigenvalues to csv
write.csv(eigenValues08,file="eigenValues08.csv")
write.csv(eigenValues09,file="eigenValues09.csv")
write.csv(eigenValues10,file="eigenValues10.csv")
write.csv(eigenValues11,file="eigenValues11.csv")
write.csv(eigenValues12,file="eigenValues12.csv")
```

```
write.csv(eigenValues13,file="eigenValues13.csv")

#loadings for the principal component
loadings08<-crime.pca08$rotation[,1]
loadings09<-crime.pca09$rotation[,1]
loadings10<-crime.pca10$rotation[,1]
loadings11<-crime.pca12$rotation[,1]
loadings12<-crime.pca12$rotation[,1]
loadings13<-crime.pca13$rotation[,1]
#wrote to a csv file
write.csv(loadings08, file = "loadings08.csv")
write.csv(loadings09, file = "loadings09.csv")
write.csv(loadings10, file = "loadings10.csv")
write.csv(loadings11, file = "loadings11.csv")
write.csv(loadings12, file = "loadings12.csv")
write.csv(loadings13, file = "loadings13.csv")

#Drew a dotchart of the loadings for each year
xlabs="variable loadings"
dotloadings08<-crime.pca08$rotation
sorted.loadings08<-dotloadings08[order(dotloadings08[,1]),1]
dotchart(sorted.loadings08,main="Loadings Plot for PC1
2008",xlab=xlabs,cex=1.0,col="black",bg="black" )

dotloadings09<-crime.pca09$rotation
sorted.loadings09<-dotloadings09[order(dotloadings09[,1]),1]
dotchart(sorted.loadings09,main="Loadings Plot for PC1
2009",xlab=xlabs,cex=1.0,col="black",bg="black" )

dotloadings10<-crime.pca10$rotation
sorted.loadings10<-dotloadings10[order(dotloadings10[,1]),1]
dotchart(sorted.loadings10,main="Loadings Plot for PC1
2010",xlab=xlabs,cex=1.0,col="black",bg="black" )

dotloadings11<-crime.pca11$rotation
sorted.loadings11<-dotloadings11[order(dotloadings11[,1]),1]
dotchart(sorted.loadings11,main="Loadings Plot for PC1
2011",xlab=xlabs,cex=1.0,col="black",bg="black" )

dotloadings12<-crime.pca12$rotation
sorted.loadings12<-dotloadings12[order(dotloadings12[,1]),1]
dotchart(sorted.loadings12,main="Loadings Plot for PC1
2012",xlab=xlabs,cex=1.0,col="black",bg="black" )

dotloadings13<-crime.pca13$rotation
sorted.loadings13<-dotloadings13[order(dotloadings13[,1]),1]
dotchart(sorted.loadings13,main="Loadings Plot for PC1
2013",xlab=xlabs,cex=1.0,col="black",bg="black" )

#Calculated the values of the first principal component
calcpc <- function(variables,loadings)
{
  # found the number of samples in the data set
  as.data.frame(variables)
  numsamples <- nrow(variables)
  # made a vector to store the component
  pc <- numeric(numsamples)
  # found the number of variables
  numvariables <- length(variables)
  # calculated the value of the component for each sample
  for (i in 1:numsamples)
  {
    valuei <- 0
```

```
    for (j in 1:numvariables)
    {
      valueij <- variables[i,j]
      loadingj <- loadings[j]
      valuei <- valuei + (valueij * loadingj)
    }
    pc[i] <- valuei
  }
  return(pc)
}

#Called the calcpc function for the first PC
loadings08PC1<-calcpc(crimeStandard08,
crime.pca08$rotation[,1])
loadings09PC1<-calcpc(crimeStandard09,
crime.pca09$rotation[,1])
loadings10PC1<-calcpc(crimeStandard10,
crime.pca10$rotation[,1])
loadings11PC1<-calcpc(crimeStandard11,
crime.pca11$rotation[,1])
loadings12PC1<-calcpc(crimeStandard12,
crime.pca12$rotation[,1])
loadings13PC1<-calcpc(crimeStandard13,
crime.pca13$rotation[,1])

#wrote to a CSV file
write.csv(loadings08PC1, file = "loadings08PC1.csv")
write.csv(loadings09PC1, file = "loadings09PC1.csv")
write.csv(loadings10PC1, file = "loadings10PC1.csv")
write.csv(loadings11PC1, file = "loadings11PC1.csv")
write.csv(loadings12PC1, file = "loadings12PC1.csv")
write.csv(loadings13PC1, file = "loadings13PC1.csv")
```

```
#called prcomp function
Gcse.pca08 <- prcomp(GcseStandard08)
Gcse.pca09 <- prcomp(GcseStandard09)
Gcse.pca10 <- prcomp(GcseStandard10)
Gcse.pca11 <- prcomp(GcseStandard11)
Gcse.pca12 <- prcomp(GcseStandard12)
Gcse.pca13 <- prcomp(GcseStandard13)

#Summarized the PC of each variable
summary(Gcse.pca08)
summary(Gcse.pca09)
summary(Gcse.pca10)
summary(Gcse.pca11)
summary(Gcse.pca12)
summary(Gcse.pca13)

#Screeplot of the PC by year
screeplot(Gcse.pca08, xlab="Components")
screeplot(Gcse.pca09, xlab="Components")
screeplot(Gcse.pca10, xlab="Components")
screeplot(Gcse.pca11, xlab="Components")
screeplot(Gcse.pca12, xlab="Components")
screeplot(Gcse.pca13, xlab="Components")

#Found the eigenvalues to enable retain the neccessary PC
G.eigenValues08<-Gcse.pca08$sdev^2
G.eigenValues09<-Gcse.pca09$sdev^2
G.eigenValues10<-Gcse.pca10$sdev^2
```

```
G.eigenValues11<-Gcse.pca11$sdev^2
G.eigenValues12<-Gcse.pca12$sdev^2
G.eigenValues13<-Gcse.pca13$sdev^2

#Wrote eigenvalues to csv
write.csv(G.eigenValues08,file="G.eigenValues08.csv")
write.csv(G.eigenValues09,file="G.eigenValues09.csv")
write.csv(G.eigenValues10,file="G.eigenValues10.csv")
write.csv(G.eigenValues11,file="G.eigenValues11.csv")
write.csv(G.eigenValues12,file="G.eigenValues12.csv")
write.csv(G.eigenValues13,file="G.eigenValues13.csv")

#loadings for the principal component
G.loadings08<-Gcse.pca08$rotation[,1]
G.loadings09<-Gcse.pca09$rotation[,1]
G.loadings10<-Gcse.pca10$rotation[,1]
G.loadings11<-Gcse.pca12$rotation[,1]
G.loadings12<-Gcse.pca12$rotation[,1]
G.loadings13<-Gcse.pca13$rotation[,1]

#wrote to a csv file
write.csv(G.loadings08, file = "Gloadings08.csv")
write.csv(G.loadings09, file = "Gloadings09.csv")
write.csv(G.loadings10, file = "Gloadings10.csv")
write.csv(G.loadings11, file = "Gloadings11.csv")
write.csv(G.loadings12, file = "Gloadings12.csv")
write.csv(G.loadings13, file = "Gloadings13.csv")

#Drew a dotchart of the loadings for each year
xlabs="variable loadings"
Gdotloadings08<-Gcse.pca08$rotation
Gsorted.loadings08<-Gdotloadings08[order(Gdotloadings08[,1]),1]
dotchart(Gsorted.loadings08,main="Loadings Plot for GCSE PC1
2008",xlab=xlabs,cex=1.0,col="black",bg="black" )

Gdotloadings09<-Gcse.pca09$rotation
Gsorted.loadings09<-Gdotloadings09[order(Gdotloadings09[,1]),1]
dotchart(Gsorted.loadings09,main="Loadings Plot for GCSE PC1
2009",xlab=xlabs,cex=1.0,col="black",bg="black" )

Gdotloadings10<-Gcse.pca10$rotation
Gsorted.loadings10<-Gdotloadings10[order(Gdotloadings10[,1]),1]
dotchart(Gsorted.loadings10,main="Loadings Plot for GCSE PC1
2010",xlab=xlabs,cex=1.0,col="black",bg="black" )

Gdotloadings11<-Gcse.pca11$rotation
Gsorted.loadings11<-Gdotloadings11[order(Gdotloadings11[,1]),1]
dotchart(Gsorted.loadings11,main="Loadings Plot for GCSE PC1
2011",xlab=xlabs,cex=1.0,col="black",bg="black" )

Gdotloadings12<-Gcse.pca12$rotation
Gsorted.loadings12<-Gdotloadings12[order(Gdotloadings12[,1]),1]
dotchart(Gsorted.loadings12,main="Loadings Plot for GCSE PC1
2012",xlab=xlabs,cex=1.0,col="black",bg="black" )

Gdotloadings13<-Gcse.pca13$rotation
Gsorted.loadings13<-Gdotloadings13[order(Gdotloadings13[,1]),1]
dotchart(Gsorted.loadings13,main="Loadings Plot for GCSE PC1
2013",xlab=xlabs,cex=1.0,col="black",bg="black" )

#Calculated the values of the first principal component
calcpc <- function(variables,loadings)
{
```

```
  # found the number of samples in the data set
  as.data.frame(variables)
  numsamples <- nrow(variables)
  # make a vector to store the component
  pc <- numeric(numsamples)
  # find the number of variables
  numvariables <- length(variables)
  # calculate the value of the component for each sample
  for (i in 1:numsamples)
  {
    valuei <- 0
    for (j in 1:numvariables)
    {
      valueij <- variables[i,j]
      loadingj <- loadings[j]
      valuei <- valuei + (valueij * loadingj)
    }
    pc[i] <- valuei
  }
  return(pc)
}

#Called the calcpc function for the first PC
Gloadings08PC1<-calcpc(GcseStandard08, Gcse.pca08$rotation[,1])
Gloadings09PC1<-calcpc(GcseStandard09, Gcse.pca09$rotation[,1])
Gloadings10PC1<-calcpc(GcseStandard10, Gcse.pca10$rotation[,1])
Gloadings11PC1<-calcpc(GcseStandard11, Gcse.pca11$rotation[,1])
Gloadings12PC1<-calcpc(GcseStandard12, Gcse.pca12$rotation[,1])
Gloadings13PC1<-calcpc(GcseStandard13, Gcse.pca13$rotation[,1])

#wrote to a CSV file
write.csv(Gloadings08PC1, file = "Gloadings08PC1.csv")
write.csv(Gloadings09PC1, file = "Gloadings09PC1.csv")
write.csv(Gloadings10PC1, file = "Gloadings10PC1.csv")
write.csv(Gloadings11PC1, file = "Gloadings11PC1.csv")
write.csv(Gloadings12PC1, file = "Gloadings12PC1.csv")
write.csv(Gloadings13PC1, file = "Gloadings13PC1.csv")
```

APPENDIX C: DETAILED DESCRIPTIONS OF THE DATA SET

C1: Detailed description of the raw GRLPR data set

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Area	All Pupils at the End of KS4	All Pupils at the End of KS4 Achieving 5+ A*-C	All Pupils at the End of KS4 Achieving 5+ A*-G	All Pupils at the End of KS4 Achieving 5+ A*-C Including English and Mathematics	All Pupils at the End of KS4 Achieving 5+ A*-G Including English and Mathematics	All Pupils at the End of KS4 With Any Parrer	All Pupils at the End of KS4 With No Parrer	Average GCSE and Equivalent Point Score Per Pupil at the End of KS4	Average Capped GCSE and Equivalent Point Score Per Pupil at the End of KS4	All Boy Pupils at the End of KS4	Boy Pupils at the End of KS4 Achieving 5+ A*-C	Boy Pupils at the End of KS4 Achieving 5+ A*-G	Boy Pupils at the End of KS4 Achieving 5+ A*-C Including English and Mathematics	Boy Pupils at the End of KS4 Achieving 5+ A*-G Including English and Mathematics	Boy Pupils at the End of KS4 With Any Parrer	Boy Pupils at the End of KS4 With No Parrer	Average GCSE and Equivalent Point Score Per Boy Pupil at the End of KS4
	Count	rcentage	rcentage	rcentage	rcentage	rcentage	rcentage	Score	Score	Count	rcentage	rcentage	rcentage	rcentage	rcentage	rcentage	Score
City of London	20	75	100	60	100	100	0	417.4	339.5	10	x	100	60	100	100	0	400.6
Barking and Dagenham	2,229	59.3	94.3	41.4	93.2	98.7	1.3	371.9	299.3	1,071	55.1	92.5	37.7	91.3	98.3	1.7	357.9
Barnet	3,130	73.1	94.6	61.7	93.9	98.9	1.1	418.6	336.6	1,572	70.2	94	59.5	93.2	98.6	1.4	406.1
Bexley	2,935	64.2	93.5	51	93.2	98.9	1.1	391	311.8	1,479	60.8	91.2	46.9	90.9	98.5	1.5	379
Brent	2,868	67.6	94.1	53.6	93.1	98.6	1.4	399	318	1,423	62.1	92.5	47.4	91	98.2	1.8	379.7
Bromley	3,287	75.2	94.4	62.7	93.6	98.6	1.4	424.9	330.4	1,677	72.6	93.2	60.1	92.4	98.4	1.6	403.1
Camden	1,183	60.9	93	43.8	91.5	98	2	365.7	302.2	567	54.5	90.7	39.7	88.7	97.7	2.3	346.9
Croydon	3,980	64.8	92.2	49.5	91.2	98.4	1.6	397.1	310.4	2,011	60.8	89.6	44.8	88.2	97.8	2.2	376.1
Ealing	3,004	69.3	96	54	95.1	99.4	0.6	426.9	327.9	1,502	64.3	95.1	49.3	94.1	99.4	0.6	407.5
Enfield	3,670	60.3	92.2	48.9	90.8	97.4	2.6	365.7	302.5	1,891	55.5	90.6	45.6	89.2	97.1	2.9	347.2
Greenwich	2,557	60.3	92.8	44.8	91.1	98.8	1.2	383.4	301.3	1,283	55.3	90.5	40.8	88.5	98.6	1.4	359.5
Hackney	1,941	57.7	91.7	41.5	90.2	98.3	1.7	357.1	294	892	53	88.6	36.5	86.7	97.8	2.2	337.3
Hammersmith and Fulham	991	63.2	91.9	43.5	90.7	98.7	1.3	377.7	304.5	510	59.8	90.6	40.2	88.6	98.6	1.4	366.9
Haringey	2,313	60.8	90.5	44.2	88.6	96.6	3.4	361.1	296.7	1,224	56.5	88.2	40.6	85.9	95.6	4.4	340.3
Harrow	2,477	74.1	95.5	62.9	94.7	98.7	1.3	417.8	335.1	1,252	68.6	94.3	57.7	93.2	98.6	1.4	396.1
Havering	2,997	66.2	94.7	53.9	93.4	98.8	1.2	405	314.9	1,520	63	93.8	50.5	92.4	98.6	1.4	391.6
Hillingdon	3,101	65.6	92.6	49.3	91.2	98.7	1.3	397	310.1	1,609	61.5	91.5	46.2	89.7	98.4	1.6	382
Hounslow	2,473	66.3	92.8	49.9	91.6	98.8	1.2	400.6	314.1	1,226	60.4	90.8	44.4	89.3	98	2	383.4
Islington	1,417	57.6	90.2	41.4	89	97.4	2.6	356	292.5	710	56.9	89.2	38	87.9	97	3	356.8
Kensington and Chelsea	518	67.6	94.4	51	93.6	98.5	1.5	390.7	315.9	265	x	94.3	46	93.2	98.1	1.9	374.7
Kingston upon Thames	1,424	72.3	93.4	60.4	92.6	98.4	1.6	424.6	331.6	719	67.5	91	56.5	89.8	97.5	2.5	396.6
Lambeth	2,203	62.3	90.9	45.6	89.7	98.9	1.1	382.3	300	1,076	57.1	87.9	40.5	86.3	98.6	1.4	360.4
Lewisham	2,665	61.1	91.7	45.9	90.4	98.4	1.6	376	301.3	1,306	56.3	89.5	42.3	88.4	98.2	1.8	349.9
Merton	1,732	67.4	90.8	53.3	89.9	97.7	2.3	386.1	309.7	939	63.7	89.4	49.1	88.2	97.3	2.7	372.8
Newham	3,243	58	94	45.7	92.1	99	1	363.4	300.7	1,611	51	93.2	39.4	91.3	99.1	0.9	340.9
Redbridge	3,196	72.7	95.7	62.7	94.8	98.5	1.5	418.4	330.8	1,617	67.6	94.4	56.8	93.3	98.3	1.7	401
Richmond upon Thames	1,169	73.1	92.5	64.5	91.8	98.5	1.5	426.2	332.3	533	69	93.1	58.7	92.5	99.2	0.8	403.4
Southwark	2,287	59.3	88.8	45.5	87.7	97.6	2.4	363.6	291.5	1,159	54.1	85.5	39.8	84.6	97	3	338.1
Sutton	2,209	72.6	95.3	57.5	94.1	99.1	0.9	417.9	329.9	1,129	71.4	94	54.2	92.6	98.9	1.1	415.3
Tower Hamlets	2,057	59.9	94.5	42.5	93.4	98.8	1.2	382.1	303.2	1,050	56.8	93.9	39.2	92.8	98.4	1.6	373
Waltham Forest	2,693	62.3	94.6	47.6	93.9	98.4	1.6	377.4	306.5	1,321	56.6	93.9	42.7	93	98.2	1.8	355.3
Wandsworth	1,622	62.9	90.7	48.3	89.3	98	2	377.7	304.2	797	60.5	88	44.4	86.8	97.2	2.8	369
Westminster	944	63.7	91.1	46.5	90.4	98.1	1.9	370.2	303.6	488	60.5	90.2	42.8	89.1	98	2	354.5
North East	32,059	66.5	91.8	45	89.7	98	2	404.8	305.9	16,418	62.2	90.1	41.6	87.7	97.5	2.5	384.4
North West	87,067	65.5	92.2	47.5	90.7	98	2	390.4	306.8	44,446	61.4	90.5	43.8	88.9	97.6	2.4	373.8
Yorkshire and The Humber	63,382	62.2	91.2	44.5	89.5	97.8	2.2	388	298.7	32,233	58	89.4	40.7	87.5	97.5	2.5	369
East Midlands	53,312	63.2	92.6	47.1	90.7	98.4	1.6	391	305.1	27,266	58.9	91.2	43.5	89.2	98.1	1.9	375.1
West Midlands	67,115	64.2	92.5	46.2	90.9	98.4	1.6	389.2	305.7	34,297	59.2	90.6	41.6	88.8	98.1	1.9	377
East of England	65,794	64.7	92.9	50.2	91.7	98.4	1.6	391.2	309.9	33,472	60.3	91.3	46	89.8	98.1	1.9	373
London	74,535	65.2	93.2	51	92.1	98.5	1.5	391.6	312.2	37,439	60.9	91.6	46.8	90.3	98.1	1.9	373.9
South East	92,114	66.1	93.3	51.8	92.2	98.6	1.4	399.5	313.5	47,004	62.4	92	48.1	90.7	98.3	1.7	383.5
South West	57,154	63.6	93	49.2	91.2	98.6	1.4	387.9	308.6	29,206	58.8	91.4	44.9	89.5	98.3	1.8	369.1

C2: Part of the filtered Dataset combined. Originally have 192 rows

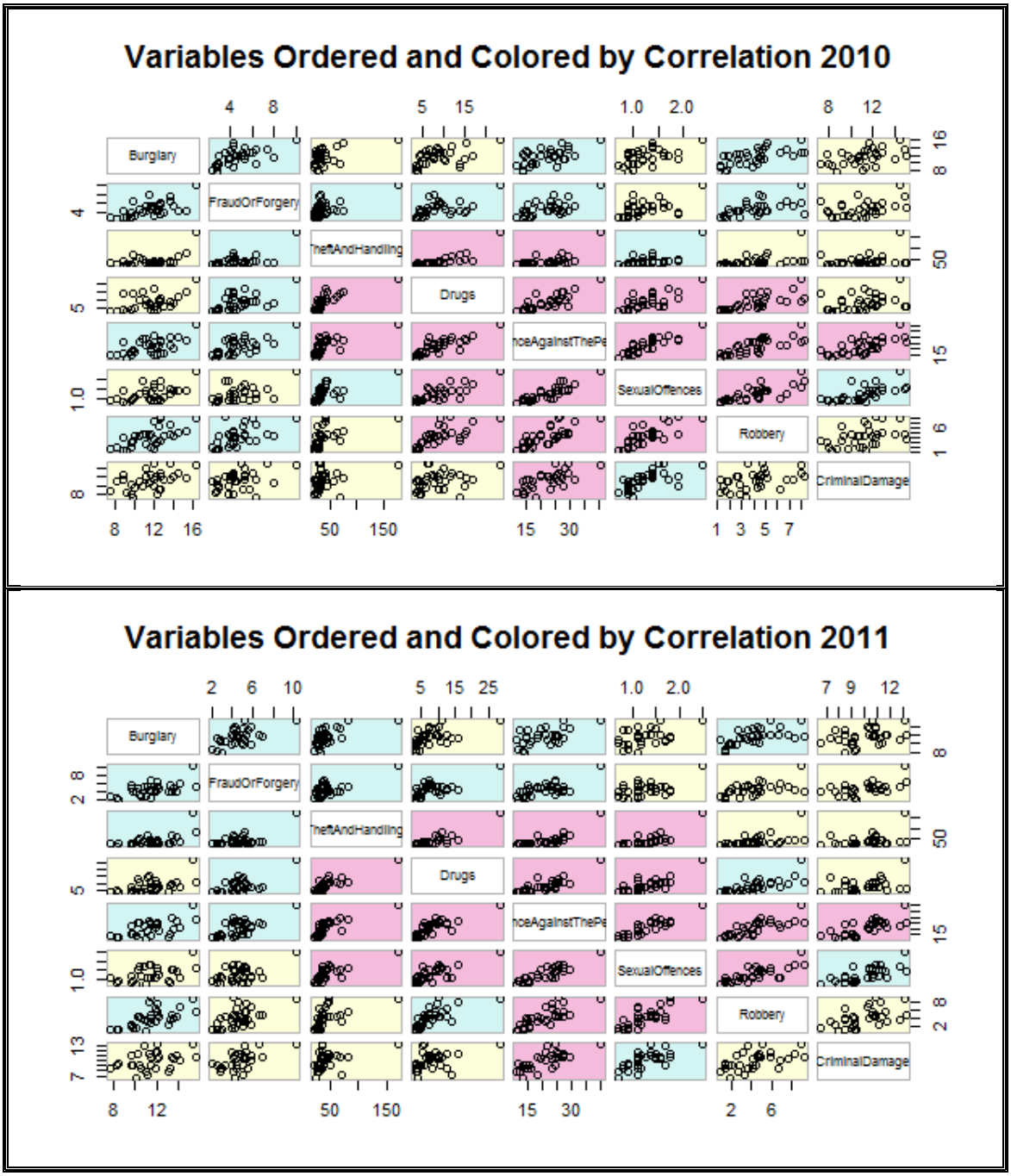
Area	Year	ViolenceAgainstThePerson	SexualOffences	Robbery	Burglary	TheftAndHandling	FraudOrForgery	CriminalDamage	Drugs	LevelTwo	LevelOne
BarkingAn	2008	29.3	1.7	3.7	9.2	36.5	7.1	18.9	7.8	41.4	93.2
Barnet	2008	15.6	0.8	2.9	12.2	30.4	2.9	10	3.9	61.7	93.9
Bexley	2008	14.4	0.7	1.7	9.7	24.9	3.2	15.3	4.2	51	93.2
Brent	2008	20.9	1.2	7.7	12.1	31	4	10.3	11	53.6	93.1
Bromley	2008	16.7	0.8	2.6	11	28.8	3.7	15.4	3.5	62.7	93.6
Camden	2008	27.4	1.1	6.3	18.3	81.5	3.5	14.1	13.2	43.8	91.5
Croydon	2008	18.5	1.1	4.9	10.7	27.1	4.7	13.7	7.2	49.5	91.2
Ealing	2008	24.2	1	4.9	13.3	34.4	4	12.4	11.1	54	95.1
Enfield	2008	14.5	0.8	4.4	12.5	29.9	3	12.5	8.1	48.9	90.8
Greenwicl	2008	30.7	1.4	5.5	14.5	43	5.3	19.1	8.5	44.8	91.1
Hackney	2008	31.4	1.7	5.8	12.3	52.8	4	13.6	20.1	41.5	90.2
Hammersı	2008	28.2	1.4	4.9	14.9	53.2	3.4	12.8	11.2	43.5	90.7
Haringey	2008	22.7	1.3	6.3	15.7	46.5	5.5	15.6	9.7	44.2	88.6
Harrow	2008	11.4	0.7	2.3	10.1	22.7	2.7	7.8	3.8	62.9	94.7
Havering	2008	13.6	0.6	1.4	9.7	27.3	3.9	12.7	5	53.9	93.4
Hillingdor	2008	22.9	1	3.1	13.1	33.5	4.2	16.4	7.5	49.3	91.2
Hounslow	2008	22.9	1.1	3.6	10.8	34.2	4.6	14.8	8	49.9	91.6
Islington	2008	28.4	1.4	6.4	17.7	72	2.5	15.5	14.6	41.4	89
Kensingto	2008	19.4	0.9	4	10.9	77.1	4	9.5	16.7	51	93.6
KingstonU	2008	16.7	1	1.8	6.9	33.7	2.4	11.5	4.4	60.4	92.6
Lambeth	2008	26.9	1.4	8.5	13	43.6	2.8	14.1	13.2	45.6	89.7
Lewisham	2008	32.1	1.4	6.4	12.1	35.8	5.5	15.2	8.8	45.9	90.4

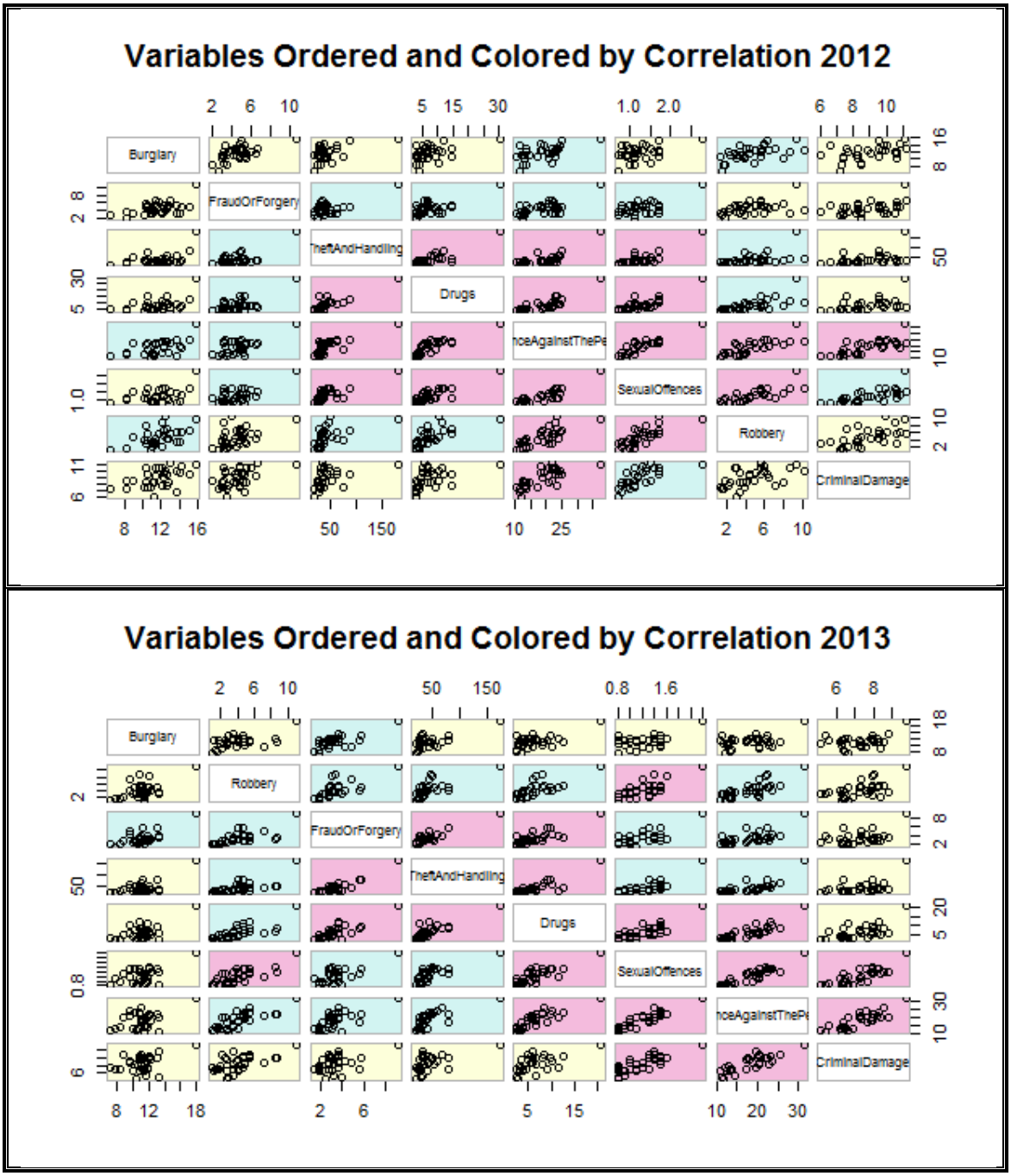
C3: Combined PCA scores for both variables.

Area	Year	CReducedVariables	GReducedVariables
BarkingAndDagenham	2008	-1.221464311	-0.42252186
Barnet	2008	2.15697366	1.849219676
Bexley	2008	2.386813599	0.532522075
Brent	2008	-0.114345172	0.755148158
Bromley	2008	1.742917046	1.84060847
Camden	2008	-1.948938095	-0.796298921
Croydon	2008	0.685265739	-0.337336534
Ealing	2008	-0.006293566	1.515574649
Enfield	2008	1.403331678	-0.541153378
Greenwich	2008	-1.755795367	-0.840941777
Hackney	2008	-2.179902048	-1.493522976
HammersmithAndFulham	2008	-1.016592074	-1.114397242
Haringey	2008	-1.348030042	-1.801423265
Harrow	2008	3.185558086	2.256853365
Havering	2008	2.580601469	0.89308823
Hillingdon	2008	0.304072136	-0.357233282
Hounslow	2008	0.409344568	-0.153416438
Islington	2008	-2.170282321	-1.935851147
KensingtonAndChelsea	2008	0.088082973	0.676648674
KingstonUponThames	2008	2.602971812	1.251479364
Lambeth	2008	-1.329741896	-1.265797878

APPENDIX D: SCATTER PLOT MATRICES

D1: Scatter plot matrices for years 2010, 2011, 2012, and 2013.




Variables Ordered and Colored by Correlation 2012
Burglary
FraudOrForgery
TheftAndHandling
Drugs
SexualOffences
Robbery
CriminalDamage
Variables Ordered and Colored by Correlation 2013
Burglary
Robbery
FraudOrForgery
TheftAndHandling
Drugs
SexualOffences
CriminalDamage

D2: Scree plot of GRLPR PCA values