NATIONAL COLLEGE OF IRELAND

Higher Diploma in Science in Data Analytics

2013/2014

An Exploratory Study into the Association between School Expenditure Levels per Student and School Performance in High Stake Examinations

Dissertation

Louise Blake

13110535

louise.blake@student.ncirl.ie

Abstract

The principal objective of this project was to present the results and findings of an exploratory study, investigating the association between school expenditure levels per student head and overall school performance in high stakes examinations. The topic of students and schools performance is by no means uncharted territory. Annual reports are created and published by Government Departments, both in Ireland and abroad. An Open dataset sourced from the Department of Education[1] UK, was explored for student performance and expenditure levels, to see what association existed. The Department of Education, provides a number of performance tables of interest. The focus of this exploratory study is for the years 2010 to 2013. The data explorative study includes variables on performance, percentage of pupils achieving 5+ A*-C or equivalents in both English and mathematics for General Certificate of Secondary Education (GCSEs) and the school expenditures per pupil from the Department of Education's Consistent Funding Reports. "The consistent financial reporting framework (CFR) is a standard framework into which schools should code their income and expenditure to enable production of simple, standardised reports for governors and local authorities." (Department of Education, 2014).

Within a PISA report[2] published in 2009, the OEDC[3] referred to the expenditure per student stating "Across OECD countries, expenditure per student explains 9% of the variation in PISA mean reading performance between countries" (OECD, 2009). The same report stated that it examined the data for the UK and found that students results were "not statistically significantly different from the OECD average" (OECD, 2009). PISA, is a Programme for International Student Assessment, it "is a worldwide study of 15-year-old school pupils' scholastic performance on mathematics, science, and reading" and compiled by the Organisation for Economic Co-operation and Development (OECD).

Data mining methods were used in order to analyse patterns in the data, while inference statics were used in order to assess the relationships between the variables considered. The results presented make use of visualisation, charts and graphs. The data analysis was conducted using R, MS Excel and Tableau analytics software.

[1] http://www.education.gov.uk/
[2] 'Viewing The United Kingdom School System Through The Prism Of Pisa', 2009
[3] OEDC Organisation for Economic Co-operation and Development

Acknowledgements

I would like to acknowledge and thank the following for their assistance; Ioana Ghergulescu, project supervisor, to Michael Bradford, course director and Jonathan Lambert, mathematics support and the staff from the School of Computing, National College of Ireland.

My sincere gratitude to my parents who provided enormous emotional and financial support throughout this year. Without them I would not have been able to complete this course.

Declaration

I declare that the material contained in this dissertation is the result of my own work and that due acknowledgement has been given in the bibliography and references to all sources be they printed, electronic or personal.

Signed:

_____
Louise Blake

Date: 28th of May, 2014

Table of Contents

List of Figures

List of Tables

List of Appendices Figures and Tables

Appendix 4

Abbreviations

	Description
CFR	Consistent financial reporting framework
EDA	Exploratory Data Analysis
ELT	Extraction, Loading and Transformation
FSM	Free School Meals Eligibility
GCSE	General Certificate of Secondary Education
GDP	Gross Domestic Product
KS4	Key Stage 4
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment

Chapter 1

1 Introduction

This project was compiled for the Graduate Diploma in Science in Data Analytics programme, at the National College of Ireland. The purpose of this project was to analyse and investigate the data sourced to see what association between school expenditure Levels per Student and overall School Performance in High Stakes Examinations.

1.1 Background

The initial research for this project stemmed from an interest in performance and the metrics utilised for measuring performance. A valuable source of material for this topic is education and there are numerous articles and resources available on the web which reference information on Education and its needs. From a World Bank source,

> "Education is one of the most powerful instruments for reducing poverty and inequality and lays a foundation for sustained economic growth. The World Bank compiles data on education inputs, participation, efficiency, and outcomes. Data on education are compiled by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics from official responses to surveys and from reports provided by education authorities in each country" (The World Bank, 2014).

In seeking an 'open data source' relating to Education a number of countries were looked at to see what was available to the public. The investigation viewed data from Worldwide organisations to specific countries and government organisations. The UK Education system was the open data source of choice. Performance was not the only variable to be used in this study. Its association with expenditure and other variables were explored.

1.2 Motivation

The initial incentive for this project originates from different data sources relating to both school performance statistics and expenditure per student data available from the UK Government sourced website. The motivation was consolidated by a report compiled by the Organisation of Economic Co-operation and Development (OEDC). Within the PISA report[4] published in 2009, the OEDC[5] referred to the expenditure per student stating "Across OECD countries, expenditure per student explains 9% of the variation in PISA mean reading performance between countries" (OECD, 2009). The same report stated that it examined the data for the UK and found that students results were "not statistically significantly different from the OECD average" (OECD, 2009). PISA, is a Programme for International Student Assessment, it "is a worldwide study of 15-year-old school pupils' scholastic performance on mathematics, science, and reading" and compiled by the Organisation for Economic Co-operation and Development (OECD).

1.3 Aims

The aim of the project was to compile an exploratory study, investigating the Association between School Expenditure Levels per Student and overall School Performance Attained in High Stake Examinations.

One of the primary reasons for this dissertation was investigate the associations relating to Performance and expenditure. To give a wider picture to this dissertation a number of other available social indications are included; Frees School Meals eligibility, English as First Language, Disadvantage percentages along with knows discrete variables Gender, Location (London Non London), School Type, Free School Meals Band.

---

[4] 'Viewing The United Kingdom School System Through The Prism Of Pisa', 2009

5 OEDC Organisation for Economic Co-operation and Development

The Department of Education gathers data on English schools and students from a number of bodies which govern and have close links to the inspection systems. The data is compiled by the Department for Education and Skills (DfES), the Office for Standards in Education (OFSTED), and at the local level, local education bodies and school governing board, for instance the National School League Tables and the National Pupil Database. These pre-existing reports which might be of interest to parents, to assist school choice. They are also of significance to any organisation or individuals interested in education. The data analysis presented in this dissertation is unconnected to previous research by the Department of Education.

The objective is rather to investigate the data with an unbiased analysis approach that is free of any presumptions, to potentially reveal correlations that have not yet been discovered. In the conclusion, the research results are evaluated and visualized.

1.3.1 Scope

The scope of the project was to develop an exploratory study into the Association between School Expenditure Levels per Student and overall School Performance in High Stake Examinations. The system focused on the deliverables required by the client, who would utilise the study in future planning strategies and provide the information to other related stakeholders.
The data utilised in this system was based on specific tables, provided by the Department of Education and relates to student performance and expenditure per student. It was acknowledged that the information in the tables only provided part of the picture of each school and its student's achievements. As the study includes expenditure, the data acquired has details of the Department of Education's Consistent Funding Reports. "The consistent financial reporting framework (CFR) is a standard framework into which schools should code their income and expenditure to enable production of simple, standardised reports for governors and local authorities" (Department of Education, 2014). The data provided was utilised to provide data analysis, using Data Mining principles and to visualise the data for the client. The data incorporates both numeric and categorical variables relating to, Expenditure per student, Performance Attained, Gender, School Type, Free School Meals (FSM) band and London-Non London location, Disadvantaged percent, FSM Eligibility and English as a First Language.
The subject of student performance and expenditure has varying opinions on the use of performance indicators (PI's). Performance indicators in general can be used to made a complex situation less subjective. The questioned of how valid performance indicators are as an approach and should they be used for making decisions in education is not one reviewed at in this dissertation. Within the scope of this dissertation, performance is an indicator however it is not focusing on individual students but on schools so the purpose of performance is in the context of a factor and it association to other factors.

The scope of the dissertation was to develop an exploratory study into the Association between School Expenditure Levels per Student and overall School Performance in High Stake Examinations. The system focused on the deliverables required by the client, who would utilise the study in future planning strategies and provide the information to other related stakeholders.

The data utilised in this system was based on specific tables, published by the Department of Education and relates to Key Stage 4 Achievement and Attainment and Expenditure per student. It was acknowledged that the information in the tables are only part of the wider picture of each school and its student' achievements. As the study includes expenditure, the data acquired has details of the Department of Education's Consistent Funding Reports. "The consistent financial reporting framework (CFR) is a standard framework into which schools should code their income and expenditure to enable production of simple, standardised reports for governors and local authorities" (Department of Education, 2014).The provided data was utilised to provide data analysis and to visualise the data for the client. The data incorporates categorical variables relating to gender, school type, Free School Meals band and London-Non London location.

1.3.2 Research Questions

In this dissertation the Data Mining analysis will use methods and algorithms to answer a number of questions relating to the multiple variants in the study.

A. What associations (or interactions) has student Expenditure Levels Per Pupil with Performance Attained?
B. What relationships (or interactions) has student Performance Attainment(%) at Key Stage 4(KS4) to the other variables; Free School Meals Eligibility, Disability, English as First Language?
C. What relationships (or interactions) has student Expenditure Levels Per Pupil to the other variables; Free School Meals Eligibility, Disability, English as First Language?
D. Are there significant effects on the dependent variables
E. Are there patterns and trends within the data?
F. What relationships has the categorical variables, Gender, School type, Free School Meal Band and location (London verses Non London) with Performance Attainment and Expenditure levels?
G. What associations or relationships in the data can a decision tree determine?

1.4 Solution Overview

The solution for this study was to investigate the data, using Data Mining and Analysis process. The process commenced with the initial data and a general overview and understanding of the data is attained. The requirements from the clients perspective were compiled. Once agreed work on a new system and deliverables commenced. The agreed process would use the principles of Data mining techniques and analysis, which would incorporate explorative and predicting models.

Preparation of the data was required, followed by the Data exploration phase. Different visualization techniques and statistical measures were applied, to investigate on single data attributes and discover correlations between different data attributes. Furthermore the quality of the datasets and its attributes is evaluated. After the revision and detailing of the research questions. This provided information for the Predictive Modeling and what was the most appropriate model to use with the data supplied.

The next step of the process concentrated on the development of an appropriate data analysis model, to approach the revised research question. Under consideration of the desired structure of the analysis result, diverse statistical methods are selected and combined. The evaluation stage of the project introduces the interpretation of the results and the visualisation of the findings. Throughout each analysis stage an overview of the analysis was complied both from an results and interpretative position. The final stage of the project is to conclude and summarise the study. To assess the knowledge value to the information and results obtained

The diagram in figure 1 below outlines the Data Mining and Analysis solution process path taken in this study.

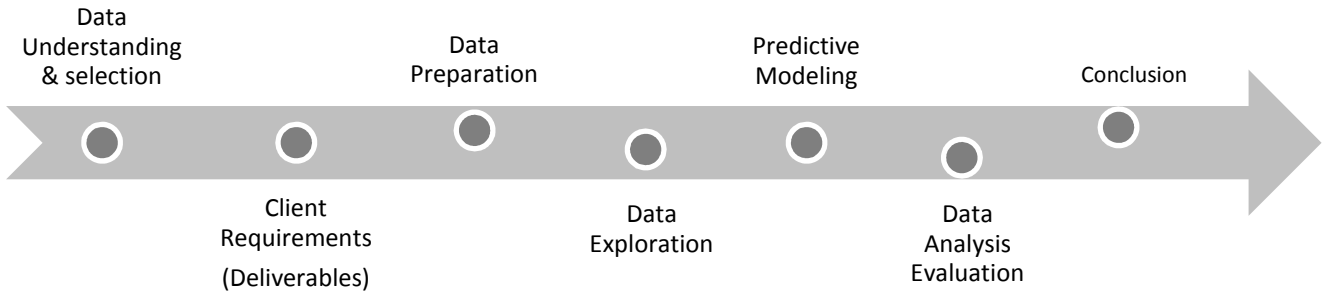


Figure 1 Outline of Data Mining and Analysis Solution Path

1.5 Structure

The main structure of the research is divided into five parts.

Chapter 1 is the background which explains the motivation, aims and scope of the research. The solution overview and research questions are incorporated in Chapter 1. In Chapter 2 the focuses is on the literature reviewed in the course of presenting the dissertation. This includes an overview and details on the Data Mining models incorporated in this study.

Following on from this in Chapter 3, are the system and datasets, required for the Data Mining analysis where exploratory and predictive models is discussed in chapter 4. Consecutively, within chapter 4 the methods, results and interpretations are discussed. In Chapter 5 the analysis conclusion and further recommendations.
.

Chapter 2

2 Related Work

2.1 Literary Review

From reading through various reports, it was decided to include other target variables and not just student performance attained and the expenditure per student. This decision came from reading a number of literary reviews and in particular a report relating to the UK Education data sourced from both the Department of Education and on an international level the OECD[6].

According to this report "To make meaningful comparisons between academies, it is important to consider the percentage of children eligible for Free School Meals, the type of academy (including whether it is a primary or secondary academy) and whether it is in London or not. This is because all these factors will affect how much an academy spends" (Department of Education, 2013)

Schools are inspected in England and standards are set at the national government level. There are a number of bodies which govern and have close links to the inspection systems. These include; the Department for Education and Skills (DfES) and the Office for Standards in Education (OFSTED), and at the local level, local education bodies and school governing board. Reports are issued, summarising key points, for example a report issued by the Department of Education as part of their Statistical First Release (SFR) presented information on the income and expenditure in academies in England. This report highlights that "considerable progress has been made in aligning the income and expenditure categorised for Benchmarking return with those of the Consistent Financial Reporting (CFR) for maintained schools (financial year 2011-12)" (Department of Education, 2013)

An OECD, PISA report[7] published in 2009 looked into research into education and the opportunities in learning, it stated "opportunities in learning lies in the distribution of resources across students and schools" (OECD, 2009). In the same report referring to the UK school system "characterised by an equitable distribution of educational resources, the quality or quantity of school resources would not be related to a school's average socio-economic background as all schools would enjoy similar resources. Therefore, if there is a positive relationship between the socio-economic background of students and schools and the quantity or quality of resources, this signals that more advantaged schools enjoy more or better resources. A negative relationship implies that more or better resources are devoted to disadvantaged schools. No relationship implies that resources are distributed similarly among schools attended by socio-economically advantaged and disadvantaged students" (OECD, 2009).

The diagram is below, figure 2, is a over view of the key findings in a 2001 study compiled by the Department of Education (UK) called "The relationship between capital investment and pupil performance: An analysis by the United Kingdom" (DfES Publications, 2001).

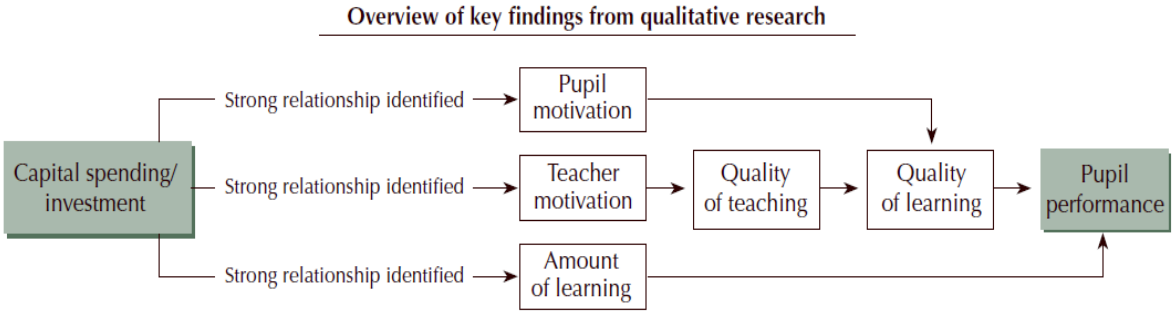


Figure 2 Overview of Key findings from qualitative research by DfES

[6] OEDC Organisation for Economic Co-operation and Development
[7] 'Viewing The United Kingdom School System Through The Prism Of Pisa' , 2009

There are subsequent reports since the above 2001 report. A report that would have a relevance to this dissertation titled 'Resources and Attainment at Key Stage 4, Estimates from a Dynamic Methodology' (Geoff Pugh et al, 2008). In this report, the research looked at data sources on Achievement and Attainment tables, the Consistent Financial Reporting and the School Census over a five year period 2002-03 to 2006-07. The dependant variables were the Average Points Score and Level 2 (the percentage of pupils achieving 5+ A*-C grades at GCSE). This dissertation compared to this study includes more recent data which is subsequent to the 2008 study. It does not use levels in the Performance attained or points scores. The expenditure values include additional Pupil Premium which was not available at the time of the earlier study. The earlier study included absenteeism and incorporated Excellence in Cities (EiC) initiative which are not included in this dissertation.

The 2001 report, "The relationship between capital investment and pupil performance: An analysis by the United Kingdom" has being referenced in other studies internationally. In a report compiled in New Zealand Government in 2013 reviewed New Zealand's expenditure per school student. At the time New Zealand's student Attainment level was below the OECD average for both annual public and private education. The purpose of this report was to use data from English schools in a quantitative analysis to provided "evidence of a positive and statistically significant relationship between capital investment and pupil and performance" (Ministry of Education of New Zealand, 2013).Expenditure per student was acknowledged by the New Zealand Government as "of particular importance when the socio-economic status of the student's family or the socio-economic mix of the school community is low" (Ministry of Education of New Zealand, 2013). According to this document resources at a elevated level were needed "to overcome barriers to learning associated with access to reference material and resources, information. and communication technologies(ICT), and other opportunities linked to cultural capital and educational achievement" (Ministry of Education of New Zealand, 2013). They acknowledged that to achieve a high quality education, with the level of resources for each student and at a cost that can balances expenditure is a difficult task. Hence they looked at comparing education at an international level and used the UK report.

In another report, an independent evaluation commissioned by the Department for Education, presented the findings of an evaluation study on the Pupil Premium. This Pupil Premium introduced in April 2011, and provides schools with additional funding. It is based on the numbers of pupils eligible for free school means (FSM) (Department of Education, 2013). This information is very important to this dissertation as the higher schools expenditure levels should correlate with the higher eligibility for FSM[8] in a schools. The report investigated how the Pupil Premium funds were spent by the schools and other related interests, such as perception of the Pupil Premiums impact on support. This report study took a selection of schools and surveyed them on a number of criteria so as to define disadvantage, the scope entailed more than FSM[8]. In the context of this dissertation, the report highlighted the need to include additional factors, such as the percentage of FSM Eligibility and the percentage of disadvantage values available in the datasets.

In the course of researching performance in education and the use of Predictive models a recent project by Kin Fun Li et al, 2013 called "Predicting Student Academic Performance" was looked at. This investigation was "to identify the factors that serve as good indicators of whether a student will drop out or fail the program" (Kin Fun Li, 2013). The education program in question relating to engineering courses. This project is targeted towards improving the

[8] FSM Free School Meals

numbers of students staying on to complete their chosen course. So although it used predictive models the project did not targeted expenditure.

As expenditure is part of this dissertation it seems prudent to assess the GDP rates for the UK between 2009 and 2013 were reviewed. During the four years the growth levels, sourced from the World Bank are show in Figure 3 and 4 (The World Bank, 2014). This shows an increase since 2009. For the same period only two years of public expenditure per pupil as a % of GDP per capita. at Secondary school level was available, the details are shown in figure 2-3.

The definition of "Public expenditure (current and capital) includes government spending on educational institutions (both public and private), education administration as well as subsidies for private entities (students/households and other privates entities)" (The World Bank, 2014).

Country name	2009	2010	2011	2012
United Kingdom	-5.2	1.7	1.1	0.3

Figure 3 UK GDP growth (annual %) Years 2009-2013

Country name	2009	2010	2011	2012	2013
United Kingdom	30.5	33.6			

Figure 4 Expenditure per student, secondary (% of GDP per capita)

2.2 Data Mining

Due to the large amounts of data that is openly available Data mining and analysis has become more and more important. There are numerous opportunities to investigate the data and discover new knowledge relating to it. The reason these opportunities are more readily available is due to the developments of existing methods and algorithms models. Technology can facilitate in the capture and storage of large amounts of data. It is Data Mining that helps to discover the underlying trends and patterns along with outliers and anomalies in the data (Rapid Miner, 2014).

The principles of Data mining and the associated techniques, methods and algorithms were used in this study. The challenge was to further the existing knowledge relating to the data and find potentially useful information from data (Rapid Miner, 2014).

2.3 Data Mining Applications

As part of the architecture of this study there was a requirement to extract and store the data in a warehouse repository and a need for a Database management system to consolidate and query the data. There were two chosen Database management systems; MySQL and MS Access. The initial intention was to utilise MySQL and create handlers to link with the chosen reporting applications. During the study MySQL was only utilised at the initial stages and MS Access was the Database Management system of choice.

There are numerous software applications' available to provide reporting tools and algorithm models. In this study the chosen applications includes R, MS Excel, Weka and Tableau. R was chosen for a number of reasons. The initial motivation was due to the availability of packages which are designed for Data Mining activities. The second reason was due to the DBI package in R which could provide a uniform, client side interface to different database

management systems, such as MySQL and Oracle, and the third advantage to using this application, the availability and knowledge resources to a novice commencing a Data Mining study. The use of MS excel was noted at an early stage, as the Data sourced was extracted in a CSV format. MS Excel was utilised at various stages of the project. Initial data viewing and data quality checks were processed in MS Excel. Weka was introduced to the project due to the Predictive algorithms' available such as Classifications using Decision Trees and the option to create Clustering.

The following technology Tableau were investigated and utilised for some data mining processes. Tableau is business intelligence software. It creates dashboards which allows the visualize of data along with interactive, sharable dashboards. "Tableau Desktop is a powerful data discovery and exploration application" (Tableau Software, 2014). During the visualisation process the data was processed some of the charts and graphs are from this application. Google Fusion - initially chosen to create geo graphic map of the data. Initial tests showed it would import data with postal codes and show it on a map. It was not utilised for the final visualisation of the results.

2.4 Data Mining algorithms Models

Once the data was extracted an essential part of the study was to devise a strategy on what data mining methods would be suitable for this data. There were many Data Mining methods available. As part of this study there is a requirement to gain an understanding of the functionality of the Data Mining algorithms. The research commenced with an exploratory investigation which is the descriptive statistics processing activities. This takes into account many elements such as, average, standard deviation, variance, distribution and range of the data variable. The second set of models applied are Data Mining and Predictive processing. These are also referred to as advanced or inferential statistics, generating models, inferences and predictions to evaluate the data. It can account for the random and uncertainty factors of the observations. The inferential statistics includes; hypothesis testing, estimations, predictions, decision trees, clustering, association descriptions (correlation) and Regression analysis (Aerd Statistics, 2014).

Both Descriptive and Predictive Analytic techniques were used in this study.

2.4.1 Data Mining - Descriptive and Inferential Analysis

> "Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data" (Aerd Statistics, 2014).

Descriptive analysis describes the main characteristics of a collection of data. The purpose of descriptive analysis is to look at the data, inspect the data for size and structure followed by basic statistics and various charts like pie charts and histograms.

In Figure 5 Overview of Descriptive Models the Descriptive models used in the exploratory analysis are broken down into Individual (univariate) variables and bivariate variables followed by the type of data the attribute comprises of; Categorical and Numeric.

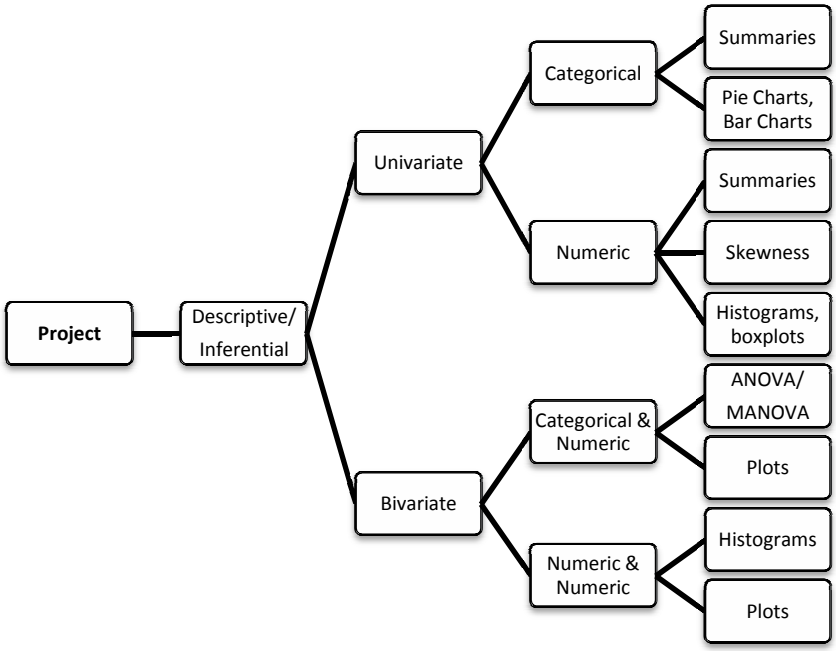


Figure 5 Overview of Descriptive Models

The categorical (nominal) data are shown as factors, and by exploring the data using a number of R functions, it was possible to find out what the levels of the Factors were. The frequency was calculated and plots created, bar charts and pie charts.

For numeric variables the summarize data comprises of a series of statistics; the mean, median and first 25% and the third 75% quartiles and the extreme values maximum and minimum (the Range). This gives the initial distribution of the variables values. In the case of the factors (nominal) frequency are shown as levels.

2.4.1.1 Summary Statistics - Descriptions

Mean: is the average, measure of central tendency,
Median: is the central value, measure of central tendency,
Range: is the highest (maximum) and lowest values (minimum),
Variance: is a measured of how spread out a distribution is, average of the squared differences from the mean,
Standard Deviation: used to measure how spread out the numbers are square roots of variance,
3rd Quartile-1st Quartile: By observing the difference between the mean and median as well as the inter-quartile range (3rd Quartile-1st Quartile) we can get an idea of the shewness of the distribution and also its spread.

2.4.1.2 Normality and Symmetry - Tests

"Since a number of the most common statistical tests rely on the normality of a sample or population, it is often useful to test whether the underlying distribution is normal, or at least symmetric. This can be done via the following approaches:

- Review the distribution graphically (via histograms, boxplots, QQ plots)
- Analyze the skewness and kurtosis
- Employ statistical tests (esp. Chi-square, Kolmogorov-Smironov, Shapiro-Wilk)

If data is not symmetric, sometimes it is useful to make a transformation whereby the transformed data is symmetric and so can be analyzed more easily" (Zaiontz, 2014).

Within this project the data was tested for Normality and Symmetry using a number of tests and graphs.

2.4.1.3 Histogram, Skewness and Kurtosis

Histogram

The histogram is used to show the distributions of independent and dependent variables.

"The histogram is a standard type of graphic used to summarise univariate data where the range of values in the data set is divided into regions and a bar (usually vertical) is plotted in each of these regions with height proportional to the frequency of observations in that region. In some cases the proportion of data points in each region is shown instead of counts" (GM-RAM Limited, 2010).

"The shape of the histogram is determined by the width and number of regions that divided up the data. A histogram provides an indication, the following features of a set of data: the general shape, symmetry or skewness of data and modality (uni-, bi- or multi-modal). There are some situations where a different type of graph would be preferable but histograms are useful for describing the general features of the distribution of a set of data" (GM-RAM Limited, 2010).

Skewness

Skewness is an indicator used in distribution analysis. It is one of the checks that can be used to test for symmetry and normality. The following is a summary of how the results can be Interpreted (Tompkins Cortland Community College, 2012).

Skewness > 0	Right skewed distribution,
Skewness < 0	Left skewed distribution,
Skewness = 0	mean = median, distribution is symmetrical around the mean.
If skeweness = 0	Perfectly symmetrical
If less than-1 or greater than 1	Highly skewed
If between -1 and -0.5 / between 0.5 and 1	Moderately skewed
If between -0.5 and 0.5	Approximately skewed

Table 1 Skewness Table

Kurtosis

Kurtosis is an indicator used in distribution analysis. It indicated a sign of flattening or "peakedness" of a distribution.

Interpretation: "

- Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values.
- Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.
- Kurtosis = 3 - Mesokurtic distribution - normal distribution for example" (Tompkins Cortland Community College, 2012).

2.4.1.4 Density Plots & Kernel Density plot

Density Plots are used to examine the distribution of specific variables in the data. By overlaying the Histogram with parametric or non-parametric Kernel density plots we could ask the question; Are the variables normally distributed or is another distribution suitable?

"This kind of data inspection is very important as it may identify possible errors in the data sample, or help to locate values that are so awkward that they may only be errors, or at least we would be better off by disregarding them in the posterior analysis" (Torgo, 2011)

2.4.1.5 Shapiro-Wilks normality test

A Shapiro-Wilks normality test was carried out, as a supplement to the graphical assessment of normality a normality tests, The Shapiro-Wilkes test is based on the correlation between the data and the corresponding normal scores.

2.4.1.6 QQ - Plot

The QQ-Plot is a more precise check for graphically combining two probability distributions, "which plots the variable values against the theoretical quintiles of a formal distribution" (Torgo, 2011).

A Q–Q plot is based on quintiles and it is difficult to determine which point in the QQ plot determines a given quintile. It is possible in most cases to observe if there are several lower and upper values of the distribution that don't observer the assumptions of the normal distribution. Using a paired samples it is possible to see if the probability plot correlation coefficient is between the paired sample quartiles. A Q–Q plot is generally a more powerful approach, than the common technique of comparing histograms of the two samples.

2.4.1.7 Box plot

This provides a visual inspection of the data. This plot gives a summarisation of some of the key properties of the variable distribution. It shows an inner box with the limits for the 1st and 3rd quartiles of the variable. The box has a line which represents the median value of the variable. The box plot gives information on the central value and spread of the variable, and also outliers.

2.4.1.8 Outlier Inspection

Outliers are extreme scores that are more than two standard deviations above or below the mean. Outliers are values that are much bigger or smaller than the rest of the data. In a box plot these are represented by a dot at either end of the plot. In order to be an outlier, the data value must be; larger than Q3 by at least 1.5 times the interquartile (IQR) range and smaller than Q1 by at least 1.5 times the IQR. In a linear regression an Outlier will move the line towards it - no hard & fast rule as to what to do. It is important to identify potential outliers as they can skew findings which in turn can skew decisions and programming.

Strategy for identifying outliers the dataset

- Visualise using boxplots and scatter plots
- Look for residuals
- Determine the approach to take with the outliers; discard or keep in.

2.4.1.9 The Mann Whitney test

The Mann Whitney test is an alternative test to the t-test. It is an Inferential Statistics test and a non-parametric test which is used in order to overcome the underlying assumption of normality in parametric tests. It is used to test whether two Independent samples are from the same or identical distribution. It is also called Mann-Whitney-Wilcoxin test or Wilcoxin Rank-Sum test. The test assumes continuous variables without ties. This test does not assume that the difference between the samples is normally distributed or that the variances of the two populations are equal (Two sample t-test does). The t-test should not be used if the distribution of the scores are skewed and the sample size is small (< 30). The Wilcoxin Sign test assumes a continuous variable, it does not deal with values that happen to fall at the Null Hypothesis median. It is not a particularly powerful test. It makes no distribution assumptions - therefore useful. The advantage is the two samples need not have the same number of observations. (Explorable.com, 2014). The test statistic for the Wilcoxon test is T.

Strategy for testing the assumptions

- the two samples under consideration are random, and are independent of each other, as are the observations within each sample.
- the observations are numeric or ordinal (arranged in ranks).

2.4.2 ANOVA/ MANOVA Test on Variants

This is an inferential statistic which refers to something about the population. It has a test statistic used to test whether a hypothesis is significant or not for the distribution of the data. The purpose of this test is to check the suitability of the multiple regression model using the F-test in the ANOVA table. If there is a significant F-value indicates a linear relationship between Y and at least one of the X's.
ANOVA and regression analysis give a dependent variable that is a numerical variable, while hierarchical optimal discriminate analysis gives a dependent variable that is a class variable.

Strategy for testing the assumptions
There are four basic assumption used in ANOVA.

1. the expected values of the errors are zero
2. the variances of all errors are equal to each other
3. the errors are independent
4. they are normally distributed

MANOVA is the Multivariate analysis of variance or multiple analysis of variance a statistical test procedure for comparing multivariate (population) means of several groups. It differs from the ANOVA test in that, "it uses the variance-covariance between variables in testing the statistical significance of the mean differences" (Wikipedia, 2014). It has a number of uses when there are two or more dependent variables. The same assumptions apply to MANOVA as the ANOVA Tests, listed above.

In this dissertation it was used to assist in answering a number of question relation to the multiple variants in the study.

- Are there significant effects on the dependent variables when changes are made to independent variable(s)?
- What are the interactions among the dependent variable?
- What are the interactions among the independent variable?

From the analysis the individual p-values provided for each dependent variable indicating whether differences and interactions are statistically significant. (Stevens, 2007)
In the MANOVA test "the information is expressed in terms of *sums of squares and cross products".* (Stevens, 2007) . In MANOVA the group mean squares is looked at between and within the groups. The between group mean squares is the hypothesis mean squares and the within group mean squares is called the error mean squares which is the F statistic and can be shown to be the product of two ratios.

2.4.3 Data Mining - Predictive analytic techniques

The predictive models reviewed and included in this study are, Correlation and Regression and Classification Decision Trees. The target variables included both categorical or nominal values hence some algorithms would work better than others with this data.

The figure 6 is an overview of Predictive Analytical Techniques. It shows the Regression model and Classification models.

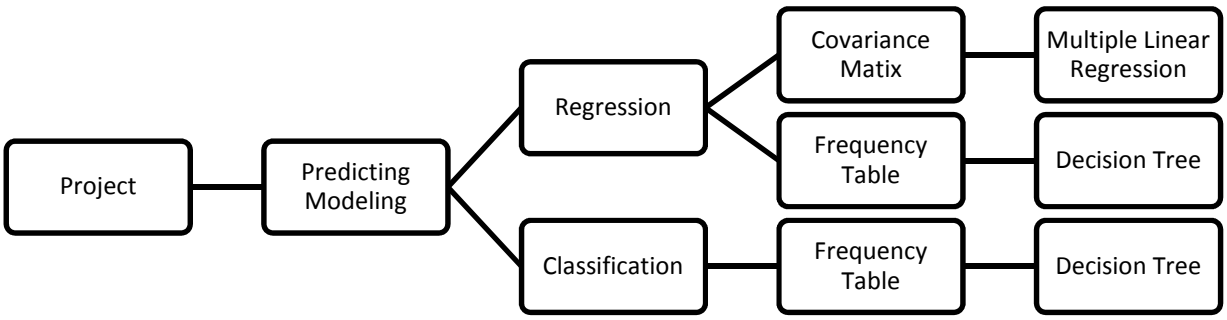


Figure 6 Overview of Predictive models

2.4.3.1 Regression Models

A simple linear regression is a way to describe a relationship between two attributes through an equation of a straight line, called line of best fit, that most closely models this relationship. "The origin of the name "e;linear"e; comes from the fact that the set of solutions of such an equation forms a straight line in the plane" (Livephysics.com, 2014).
Line of best fit (trend line)
This is a line on a scatter plot which can be drawn near the points. to more clearly show the trend between two sets of data. If the line of best fit that rises quickly from left to right is referred to as a positive correlation. If the line of best fit that falls down quickly from left to the right is referred to as a negative correlation. "Strong positive and negative correlations have data points very close to the line of best fit. Weak positive and negative correlations have data points that are not clustered near or on the line of best fit. Data points that are not close to the line of best fit are called outliers" (Livephysics.com, 2014).

Strategy for testing the assumptions of linear regression:
Quantitative models always rest on assumptions about the way the world works, and regression models are no exception. There are four principal assumptions which justify the use of linear regression models for purposes of prediction:

- linearity of the relationship between dependent and independent variables,
- independence of the errors (no serial correlation),
- homoscedasticity (constant variance) of the errors,
- normality of the error distribution.

2.4.3.2 Multivariate Regression

"Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables - also called the predictors" (Explorable.com, 2014). "By multiple regression, we mean models with just one dependent and two or more independent (exploratory) variables. The variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent (exploratory) variables" (Explorable.com, 2014).
It is possible to check the suitability of the multiple regression model by the F-test in the ANOVA table. If there is a significant F-value indicates a linear relationship between Y and at least one of the X's.

Checking model in terms of predictability

> "Once a multiple regression equation has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R2). R2 always lies between 0 and 1" (Explorable.com, 2014).

The closer R^2 is to 1, the better is the model and its prediction. The R^2 is the coefficient of determination.

2.4.3.3 Regression Tree

> "The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. Trees used for regression and trees used for classification have some similarities-but also some differences, such as the procedure used to determine where to split" (Wikipedia, 2014).

ANOVA and regression analysis provides a dependent variable that is a numerical variable.

2.4.3.4 Decision Trees

A widely used technique in data mining are Decision trees. If a decision tree is created well, it can be easy to use, and an advantage is it can be understood by non-experts. "The basic idea behind decision trees is a so-called divide-and-conquer approach. In each step the dataset is divided into different parts while each part should better represent one of the possible classes. The final result will be a tree structure where each inner node represents a test for the value of a particular attribute and each leaf is representing the decision for a particular class" (Rapid Miner, 2014).

When constructing a decision tree the first step is to select one of the attributes for the root node. The algorithm utilises a process of training data, testing data and evaluating the performance of the results.

Chapter 3

3 Systems and Datasets

3.1 Design and Architecture

The following section describes the system architecture and provides representation of the system, showing the system components, how they interacted with each other, and the extraction of the data from the website. The components included the creation and management a MySQL Database. The initial requirements proposed the use of connectors where possible to get data out of the MySQL database into the various Reports Generating tools to generate results. This proposal was adjusted and an MS Access database was created using the Warehouse Repository created in MySQL.

This architecture for the system was chosen because of the following reasons.

The data comprised of data from two 'major' data tables and two 'minor' tables. It required a 'mashup' so as to data could be accurately matched. This is server-based mashups which provided a data source with the Target variables, aggregated data sources all which were used in the Reporting Tools.

In addition both MySQL and MS Access have add-in options which, can be used to integrate with the selected Report generating and Data analysis tools, such as R, Tableau and MS Excel. This is a recommended future requirement for a more automated process.

The process to establishing the Warehouse Repository and the MS Access are described in more detail in the Appendix 3 - Systems and Datasets.

An overview of the system architecture is illustrated in figure 7 below.

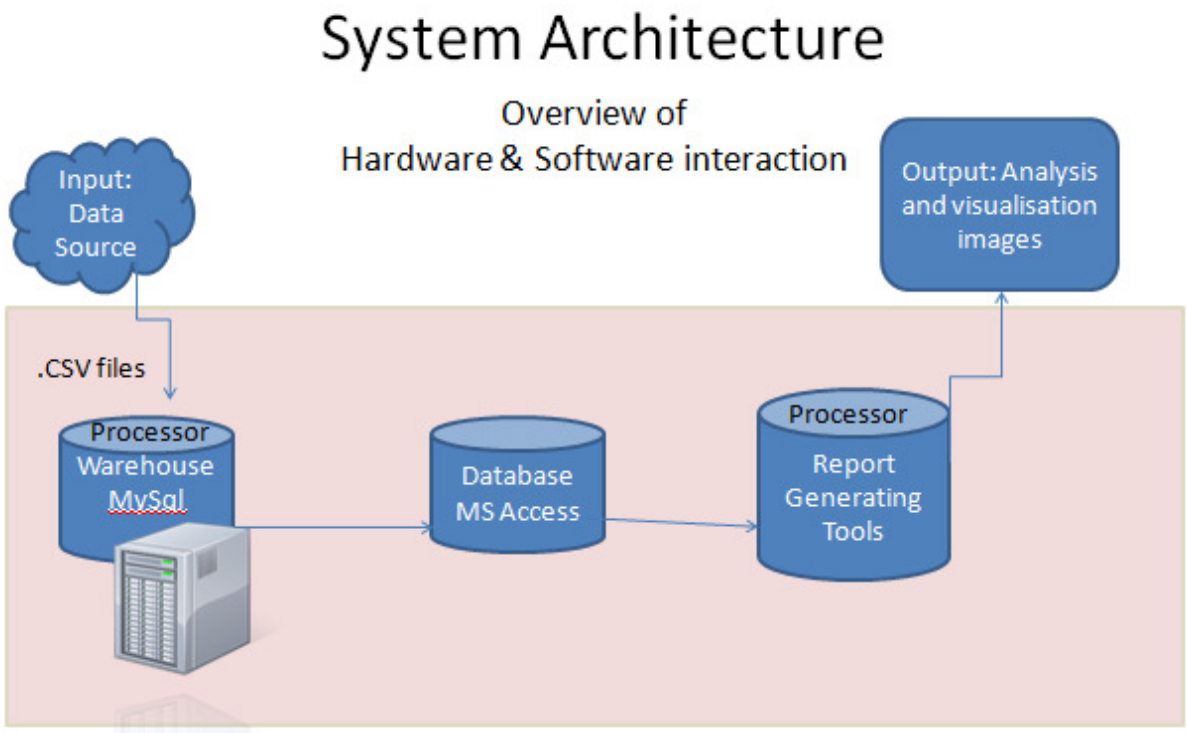


Figure 7 Overview of System Architecture

3.2 Implementation

The implementation of the study follow a number of pre-defined key steps. The figure 8 shows the flow of Data during implementation process. The system architecture was in the previous section and was implemented as described.

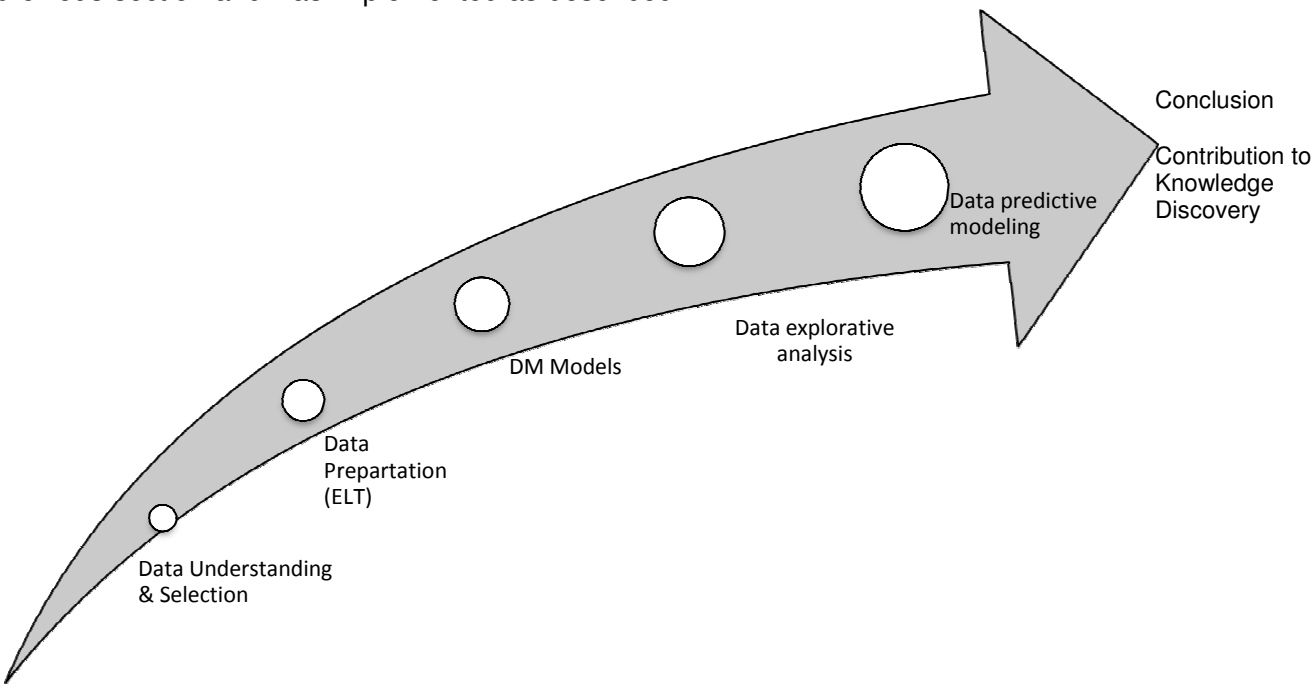


Figure 8 Flow of Data during implementation process

Following the set up of the data repository the data preparation step commenced.

Data understanding and selection

Gaining an understanding of the Domain (the education sector) and an overview of the data was one of the first steps in the implementation process. Following on from this the data attributes required in the study was determined. This stage of the implementation also explored the data types within the database. The data comprised of both numeric and categorical data. It was decided at commencement a full in depth analysis of all available data was not feasible, therefore specific variables were targeted. In order to exclude data from the analysis, the initial specification of the research area was determined at this project stage.
In figure 9 highlights the target variables and their data types.

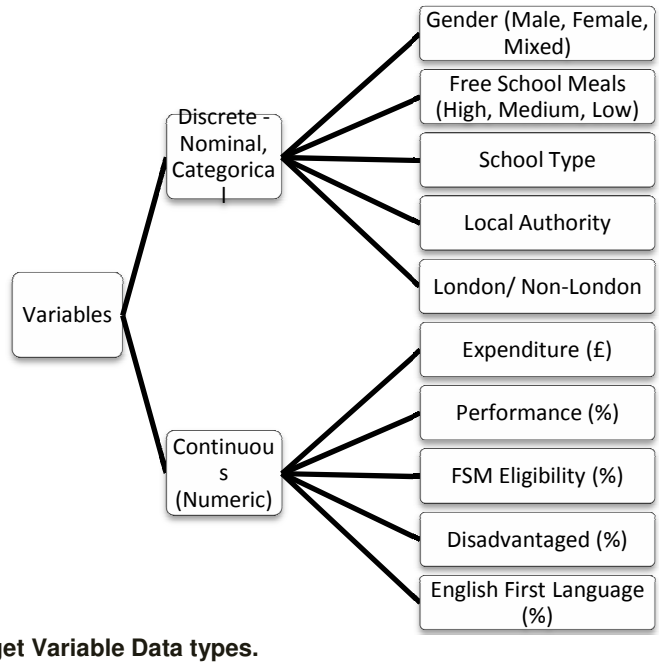


Figure 9 Target Variable Data types.

3.2.1 Data preparation

Following the decision on the target variables the next stage of implementation was to prepare the data. In the data preparation stage the implementation stage applied was an Extraction, Loading and Transformation (ELT) process. In Figure 10 below, provides a summary of the process.

3.2.1.1 Extraction

The data was sourced from the Department for Education website web site. As it consisted of two main datasets, the first for KS4 Attainment Results and the other main dataset for Consistent Financial Reporting (CFR) data. The other datasets required composed of data relation to regions and school types. Each file was saved as a CSV ("Comma Separated Value") file format. Data manipulation and joining was required, this was initially set up in an MySQL database, used as a repository and then a MS Access database was used for processing , joining data tables and to create a consolidated database for use in analysis. The use of MS Access was not the initial chosen data processing tool. It was the stand by option, to enable processing to continue. The connections between the analysis tools (R) and MySQL were not established, hence the need for MS Access.

3.2.1.2 Transformation

Transformation and manipulation of the datasets source was required. This included some cleanup of the dataset. Prior to using the data sourced, it had being processed and the datasets were published as individual datasets. Hence the data needed to be joined together, this was carried out in MS Access. The datasets had a considerable amount of information, not all was targeted for the analysis and visualisation process. The transformation process determined what variables would be kept. Thirty three variables in total were kept for possible exploration and analysis.

3.2.1.3 Loading

The datasets sources were also loaded into MS Access, created a 'mashup' table of data, rechecked for data consistency and aggregation analysis. The aggregated data was exported to excel. The data was also imported into R for various methods and Tableau.

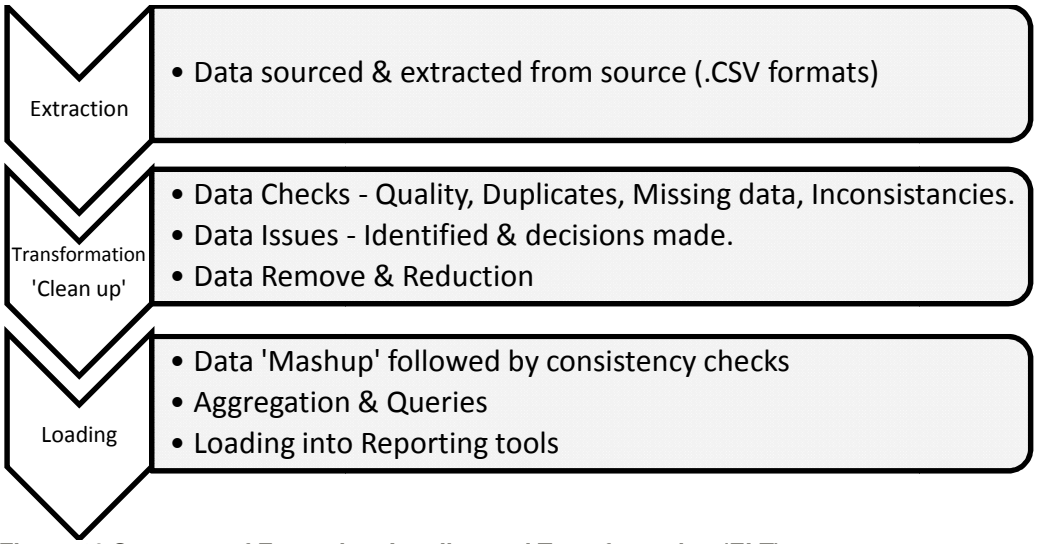


Figure 10 Summary of Extraction, Loading and Transformation (ELT) process

More in depth details on the steps use in the ELT Process are supplied in Appendix 2 - Systems and Datasets.

3.2.2 Data Mining Models

Following Data preparation the next stage of implementation was to determine what Data mining activities would be appropriate. The flow diagram in figure 11 provides a summary of the implementation activities for both data mining exploratory analysis and predictive data analysis.

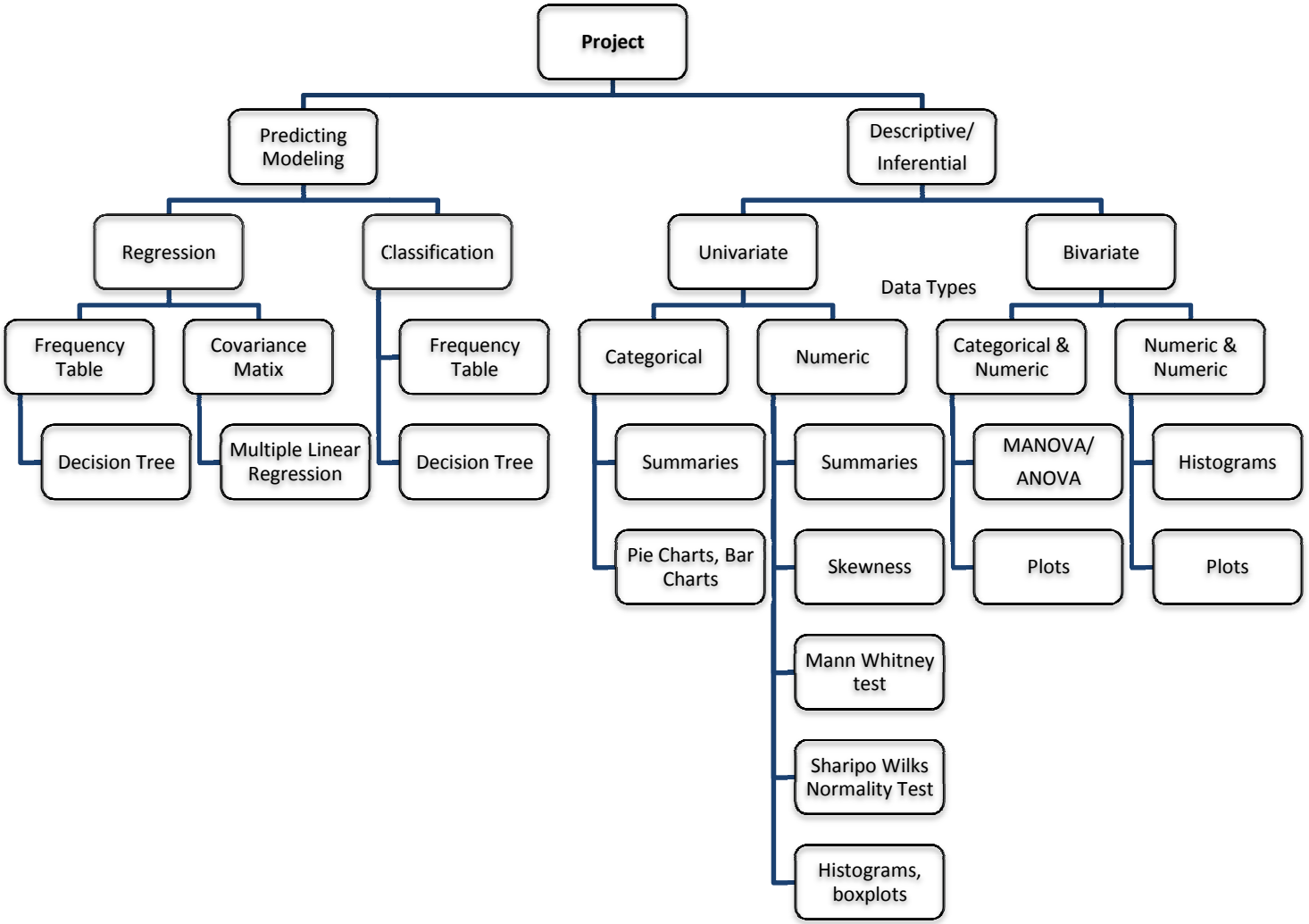


Figure 11 Data Mining Analysis Schema

3.3 Requirements

3.3.1 Project Restrictions

Restrictions were acknowledged at the commencement of the project which included the duration of the project confined to pre-agreed dates which had a commencement date February 2014 and a completion date May 2014. The availability of specialist people to complete all tasks was accepted as limited. The software resources were accepted as what, is provided by the college software, free downloads and students own resources.
The need for a budget to source additional software did not occur.

Legal and Copyright
The data provided was open source and it is acknowledged that the department of Education are the corporate author and supplier of the material. It is used under the Crown copyright Open Government Licence. (Department of Education, 2013).

3.3.2 Functional requirements

3.3.2.1 Data Requirements

The student performance on a subject by subject basis was not part of this study or a requirement of the system. It was acknowledged at the commencement of the project that on an annual basis results in schools can change in the future compared to historical results. There were other sources relating to schools performance which could have being sourced from organisations such as Ofsted and obtained from www.ofsted.gov.uk[9]. The details from Ofsted and any other post primary student performance were not included as part of this study, only the data tables sourced from the Department of Education (UK) as referred to in the requirements definition.

3.3.2.2 User requirements

The objectives for the system were to provide the requirements from the client's perspective, to agreed on a new system development and to provide agreed deliverables.

The required an exploratory study into the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations. The school expenditure levels per student head provided between the years 2009 to 2012 (four years) with the overall performance available for students at GCSE School curriculum; Key Stage 4 (KS4). These Tables gave information on the achievement of student in secondary schools in a percentage format. They also included details of the school; numbers of students, school type, gender, age range, local authority (LA) area in England as a whole. Only the data tables sourced from the Department of Education relating to England_KS4 Performance and Consistent Funding Reports (CFR) were utilised.

[9] www.ofsted.gov.uk

3.3.3 Requirements Specification

3.3.3.1 Output Requirements

The required analysis and visualisation of the database data, included the following:

Descriptive Analysis (Explorative)

Descriptive statistics, purpose to summarise the data with visuals (box plot; summary tables); Explore the categorisation of the schools, by involving a number of other variables; such the following target variables; school type, gender, Free School Meal band, location (London-non-London) and age.

Predictive analytics

As part of this dissertation, the associations within the data (correlation) were investigated. The predictive relationships explored within the data using Linear and Multiple regression analysis. A number of other predictive models such a Decision Models were explored; the relationship between student performance percentage, expenditure and other socio economic attributes available within the data. Aggregated data was used to reduce the data for specific visualisations where it was considered a reduction in the data is of benefit to the visualisation process. The creation of a map locating the top 5% and lowest 5% of performing schools using a suitable geospatial tool. Tableau was the geographical tool of choice, however the map analysis was limited in this dissertation, and further analysis is part of the future recommendation. Written reports were required, for Stakeholders and presentations to be create and delivered.

3.3.3.2 Non Functional Requirements

Performance/ Response Time

The volume of data in this data analysis did not require to be analysed at a significantly fast speed or with urgency. It is a noted in this project due to the system design, whereby there is a utilisation of a number of tools which are not fully integrated and automated the speed of performance and response time are not a critical element.

Availability requirement

The system did not need to be available for an end user, just the Data Analyst.

Recover requirement

In the event of hardware failure a current backup version of the project data was available at all times.

Robustness requirement

The system used a filtering process which was independent of the algorithms and sieved out the relevant from the irrelevant data requirements. During this process a number of important issues were addressed. These issues included, missing data, outliers and unwanted characters. These issued needed to be considered as they would affect the data analysis and needed to be controlled and managed during the study.

Security requirement

To gain access to the 'raw' data taken from the Department of Education's web site did not require any specific security rights. The data required was available for public use, as an Open Source database.

To access the system developed and designed required access to the device (laptop) which currently hosts the system. The system did not pose a high security requirement, as it was not accessible by the public or other users, that were likely to damage or corrupt the data.

Reliability requirement

There was a requirement to select appropriate tests that were valid and fit for purpose. The tests selected should be able to provide consistent and stable results. The data used in this study is compiled annually. The study was dependant on the accuracy and availability of data

produced and compiled by the Department of Education. There was a need for the extraction process be reliable, to ensure data that is not corrupted during the study.

Maintainability requirement

There was no need to maintain the system designed once it is created.

Portability requirement

The methods processed can be utilised by other users, along with access to the data and the applications utilised in this study. The application scripts in SQL and data queries in MS access generated during the study may not run properly if the Data metadata is amended by the providing source, the Department of Education.

Extendibility requirement

There was no future plan to extend the study, however it is possible that the methods and processes created could be re-used for future years data.

3.3.3.3 Resource Utilization Requirement

Hardware and Software

There was a need to provide a laptop, internet access, backup storage device. The installation of software was also a requirement.

Project Planning

There was a need to manage the project within the time specified, to succeed in producing successful deliverables. MindGenius software was installed and utilised at the requirements stage of the dissertation. This tool aids in visualising ideas and information flow during the project. It provided a Mind map and a Gantt Chart both of which are available in the Appendix section of the dissertation. (Appendix 1 and Appendix 2)

Learning - Knowledge

There was a requirement to consult with Lecturers and specialist person(s). Further resources include Library resources' and Online help tools.

3.3.4 Interface Requirements

This section was not applicable hence, no end user interface was designed. Future scope, would be to incorporate an end user interface where the user can request a particular analysis or graph result. However it is possible to use Tableau as an end user interface as it is designed with this feature. This is also a recommendation for future development.

3.3.5 System Evolution

The system can evolve by developing more integration and automation which, this would produce faster results with less reliance on specialist people. This would be benefit the end user and any future development to incorporate data from other student cohort years. The system was designed to focuses on students whose performance relates to the KS4 cohort and provide analysis and visuals relating to this data. By evolving the system could accommodate other Key stages in the student cycle such as KS3 or KS5. Another derivative to the system would be to incorporate future years of expenditure and performance for the same student cohort KS4.

3.3.6 System Evolution Constraint

A prominent system constraint to the system evolving, would be if the Department of Education amends the open data availability. In addition to this threat, the Department of Education's could in the future amend the attributes on the tables used in this dissertation.

3.4 Datasets

3.4.1 Datasets Source

The datasets sourced for this project, came from the Department for Education website[10], UK. The subject matter for this project was based on two school education datasets relating to 'Spend Per Pupil' and the 'KS4 Attainment results'. Each dataset comprised of data for schools within England, which are identified by the school name and a primary key called Unique Reference Number (URN). The URN was available in both datasets, hence making it a common primary key to enable joining of Tables.

The website also provided the metadata descriptions, for each of the above datasets. Two other datasets tables were required, to assist in the analysis and dataset processing. The two other datasets were the list of regions and local authority codes and the fourth dataset comprises of the school types descriptions. The main datasets each had a huge number of attributes. Specific attributes were identified as target variables of interests, while the others were exclude from the study this was carried out during the Data preparation implementation.

3.4.2 Entity Relationship Diagram

Below in figure 12 is an entity relationship diagram (ERD) showing the relationships between the tables and some of the attributes.

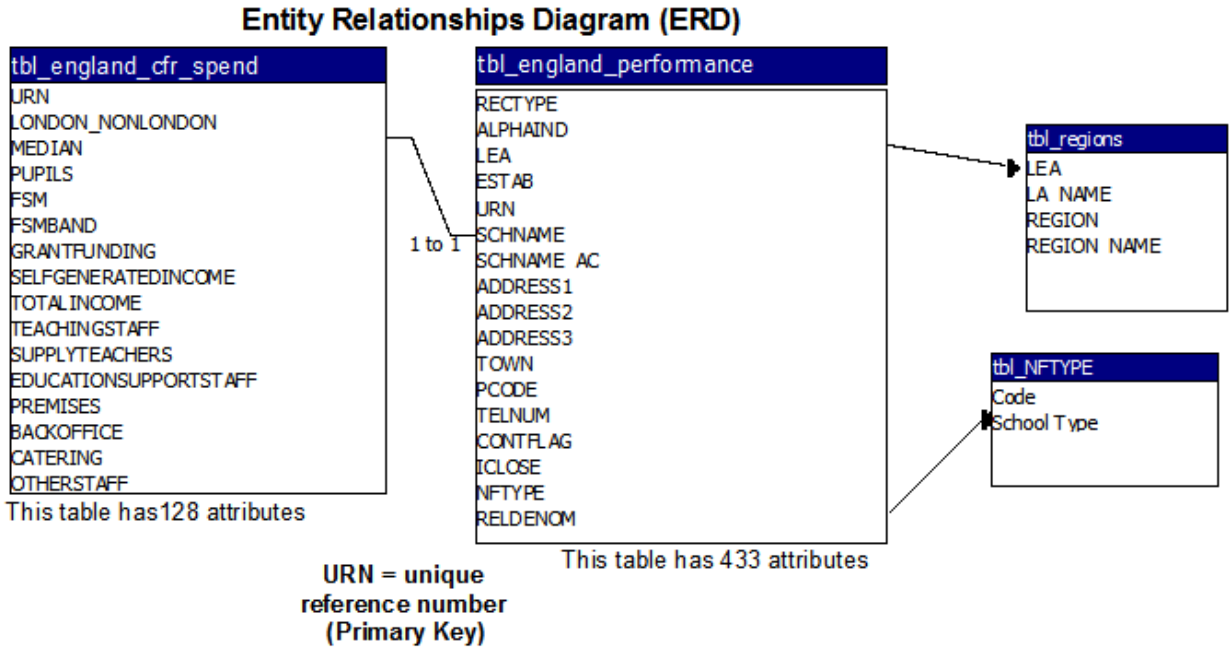


Figure 12 Entity Relationship Diagram (ERD)

[10] http://www.education.gov.uk/schools/performance/download_data.html

3.4.3 Description of Datasets

3.4.3.1 Expenditure Per Student

This Expenditure per Student dataset came from a report called the Consistent Financial Reporting (CFR) data. The data sourced from this report included the Expenditure figures supplied by the schools to the Department for Education. The original CFR data table, provided values for all schools, broken down into the individual CFR codes (17 Income, 31 Expenditure). Schools are responsible for their own CFR returns, many work closely with their local authority on their CFR returns, before it is submitted to the Department. In total there were 128 attributes and the original raw data was over 18114 rows. The data comprised of primary and Secondary School data. The dissertation focused on Secondary education. Overall, not all the attributes were targeted, hence not all rows details were required, and they were not part of the current study.

The attributes that were targeted in this project included:
URN - Unique Reference Number, which provided the join to KS4 Attainment results table.
London_Non-London - A category reference for London and non London schools.
Median - Type of school (secondary with KS4, secondary without KS4)
Pupil Numbers -The pupil numbers (full time equivalent) taken from the 2013 Annual Schools' Census. These numbers were used to calculate per pupil expenditure amounts.
School level FSM - This is the percentage of solely and dually registered pupils eligible for Free School Meals (FSM).
FSM Band -The three broad band's used to group pupils eligible for FSM are: Low: =<20.0%, Medium: 20.1%-35.0% and High: >35.0%.
Total Expenditure - Spend per pupil(£) available for the years 2009-10, 2010-11, 2011-12 & 2012-13.

3.4.3.2 KS4 Attainment Results

This table was one of the main tables utilised for the project. Following the data preparation not all of the 5208 records available in this dataset was used, the same applied to the attributes. The attributes comprised of the following targeted attributes:

RECTYPE -This attributes indicates record type, 1 = Mainstream is the type we are interested in.
URN - Unique Reference Number which will provide the link to the Spend Per Pupil
LEA - Local Authority Code, it is a numeric format, hence the need for the Regions table
SCHNAME - School name
TOWN - Town of where school is located
PCODE - Postal code, required to provide geographical location
NFTYPE - Type of school, abbreviations used, hence the need for School Type table
EGENDER - School gender of entry
TOTPUPS - Number of pupils on roll (all ages)
TPUP - Number of pupils at the end of Key Stage 4
KS2APS - Key Stage 2 Average Points Score of Key Stage 4 cohort
TFSMCLA -Number of disadvantaged pupils
PTFSMCLA -Percentage of pupils who are disadvantaged
TNOTFSMCLA - Number of non-disadvantaged pupils
PTNOTFSMCLA - Percentage of pupils who are not disadvantaged
AC5EM10 - Percentage in 2010 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
AC5EM11 - Percentage in 2011 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
AC5EM12 - Percentage in 2012 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
AC5EM13 - Percentage in 2013 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
NUMBOYS - Total boys on roll (including part-time pupils)
NUMGIRLS - Total girls on roll (including part-time pupils)
P15END4 - Percentage of pupils at the end of Key Stage 4 aged 15

P14END4 - Percentage of pupils at the end of Key Stage 4 aged 14 or under
TEALGRP1 - Number of Key Stage 4 pupils with English as their first language
PTEALGRP1 - Percentage of Key Stage 4 pupils with English as their first language

3.4.3.3 Regions and Local Authorities

This table was created from the contents of a web page relating to the Region and Local Authorities codes[11]. It joins the table called 'tbl_england_performance' using the LEA attribute. Table naming convention in SQL is tbl_regions. There were 153 regions available, of which 146 local authorities had schools which were included in the dissertation.

3.4.3.4 Type of school

This table is based on the Type of school, abbreviations[12]. The attribute NFTYPE, in the table called 'tbl_england_performance' is joined to the code attribute in the 'tbl_NFTYPE'. Table naming convention in SQL is tbl_NFTYPE. The abbreviations and descriptions of the schools types that were part of this dissertation are shown in the Table 2 below

Type of School (institution)

CY	Community school maintained by the local authority (LA). The LA is the admissions authority - it has main responsibility for deciding arrangements for admitting pupils.
FD	Foundation school maintained by the LA. Some may have a foundation (generally religious) which appoints some - but not most - of the governing body. The governing body is the admissions authority.
VA	Voluntary aided school maintained by the LA, with a foundation (generally religious) that appoints most of the governing body. The governing body is the admission authority.
VC	Voluntary controlled school maintained by the LA, with a foundation (generally religious) that appoints some - but not most - of the governing body. The LA is the admission authority.

Table 2 School Types

These datasets joined and a Consolidated dataset was created for use in R, Weka and Tableau.

3.4.4 Merged Data

Following the data preparation a consolidated table was created with the relevant Target attributes as well as relevant row. The was done by a data reduction process using Filters and queries. The final dataset used consisted of 33 attributes and 1447 records.

A number of subsets of the dataset were prepared, in addition to the initial consolidated dataset of 33 attributes. One subset that was created, contained the aggregated values taking the average over the four years for both the Expenditure Levels and Performance Attained attributes.

[11] http://www.education.gov.uk/schools/performance/laregcode.html
[12] http://www.education.gov.uk/schools/performance/ks4_abbr.html

Chapter 4

4 Testing and Evaluation

4.1 Preparation of Data

Preceding the data preparation, a review was carried out to discover out what issues were identified by the dataset source, the Department of Education. This review discovered that there were some pre-existing issues with the datasets, as identified by the Department of Education. Schools that are listed as Academies do not have to return Consistent Financial Returns(CFR), therefore they not available to be included in this study, although they it is acknowledged there would be student sitting GCSE's.
The specific issues noted by the Department of Education included the following, SUPP, indicated data had been suppressed, the reasons stated were if the number of pupils who entered or passed the qualification was five or fewer. NA, meant that the grade was not available for the GCSE qualification. No results, this was given if pupils did not achieve a result due to possible disqualified or not enough grades to be awarded an overall grade, ungraded or were absent also applied to this indicator. Data with these pre-conditions were excluded from the study during the Data Preparation implementation stage. (Department of Education, 2014)

As well as exclusion of some data, during the implementation of the Data Preparation a number of checks were completed immediately after extraction from the website source. These checks included the need to establish the quality of the data and applicability. The dataset was merged into one dataset and further verification checks were completed at this stage of the Dissertation study.

4.2 Data Verification

One of the first verification checks completed on the data, was to assess for record duplicates. This was an uncomplicated task as each record had a Unique Reference Number (URN) assigned. No duplicates needed to be removed. The Unique Reference Number also provided the relationship between the datasets, which was a requirement during the data merge process.
The next data verification check was to locate missing data from the target attributes or Null Values. It was established that some of the attributes extracted did have some missing data and a number attributes had unwanted characters and references, such as SUP (Supplementary). The records affected by these missing values were investigated. The decision was to delete these data records described from the study. It was determined to be the most effective option and the affect on the overall study by keeping them in the study would be greater on the analysis models. Overall numbers were small and there was a need for specific attributes, Performance Attained and Expenditure per pupil to their values grouped using the mean aggregation.

It was not just missing data that need to be verified, attributes considered in the initial choice of target attributes were found to be unsuitable due to inconsistent data. The attribute relating to Student Age Range had issues. The format of the data entry was not consistent, therefore this field was not used in the data analysis. It consisted of varying ranges, dates and text formats.

The in-depth details and processes involved are located in the supporting documentation Appendix 2 section.

4.3 Data Exploration

An exploratory study of the merged data was carried out on the merged dataset. Data attributes within this dataset contained two attributes which had four years of values each. These attributes were looked at individually, however the Average values were calculated and used in the Descriptive analysis, the Correlation and Regression, Regression and Decision Tree analysis. By aggregating the values the analysed disregarded the year of their original. Some analysis was carried out on the individual years during the discovery of trends and patterns and also as a verification check in some tests completed.

4.3.1 Descriptive and Inferential Analysis

One of the objectives of this dissertation was to explore the data and discover patterns and trends. The exploratory stage investigated the data using descriptive analysis, with the anticipation of discovering patterns and trends. The first step of the descriptive analysis was to look at the data; inspect the data for size and structure, review the attributes followed by basic statistics and the use of charts such as pie charts and histograms.
Using R as the as the chosen Analytical application tool for this process, the dimension and names of data were obtained respectively, using the R functions dim(), and names(). Other R functions used included, str(), attributes(), summary() and names() which returned the structure, attributes, summary and names of the data.

The objective of this task was to provided an easy and compact way to explore the datasets for information relating to the factors. A factor in R is an attribute generally categorical type. In the dataset it was noted that the factors showed a number of levels or categories values. There were 1447 observations and each observation, had information on the 33 attributes in the 'cleaned up' dataset. Of these 33 attributes, there were eleven numeric attributes; four attributes relating to Expenditure per pupil(£) and four attributes relating to Performance Attained(%), FSM[13].Eligibility(%), Disadvantaged(%) and English as a First Language(%). The attributes Expenditure per pupil(£) and Performance Attained(%) related to data collected from schools over four years, school years 2009/10 to 2012/13.

4.3.1.1 Exploring Individual Attributes

The purpose of this section was to explore, the individual attributes using summary options and to visualise the data using various tools available. The attributes comprised of both continuous and discrete data types. The exploration was divided into two divisions nominal attribute (discrete) attributes and numeric attributes.

Nominal Frequency Tables

To provide a breakdown of nominal data a number of frequency tables were created on the target variables. The four target nominal attributes are as follows, Gender, Type of Schools, Location (London Non-London), Free School Meal Band. In the frequency tables below (Table 3) the nominal attributes are listed with the categories within each attribute broken down and a count provided.

Location		Gender		Free School Meal Band (FSM)		School Type	
London	193	Boys	68	High	146	Voluntary Controlled School	39
Non-London	1254	Girls	87	Medium	372	Community School	753
		Mixed	1292	Low	929	Foundation School	332
						Voluntary Aided School	323

Table 3 Breakdown of Nominal attributes by Count

The nominal attributes were looked at in more depth than just a count. The frequency of the factors were calculated in R using with function table(), and then plotted as a bar chart with barplot() or a pie chart in MS Excel. The charts created are representing the nominal's are presented in the figures 13 to figures 15.

[13] FSM Free School Meals

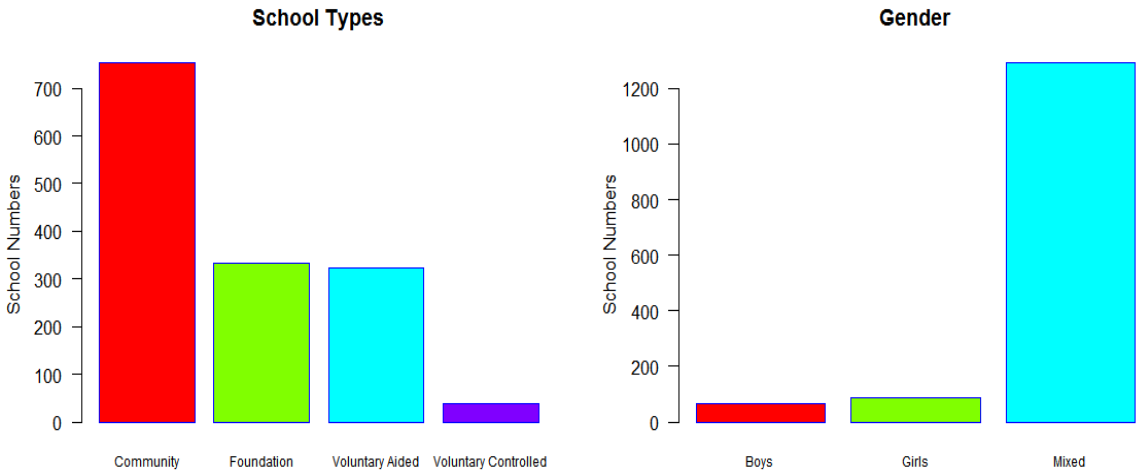


Figure 13 Two Bar Charts a) School Types and b) Chart of Gender

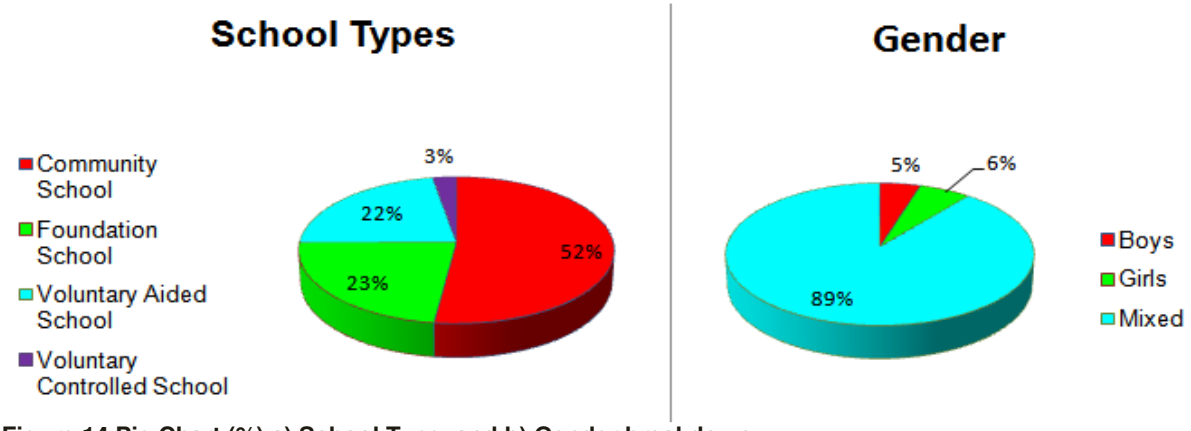


Figure 14 Pie Chart (%) a) School Type and b) Gender breakdown

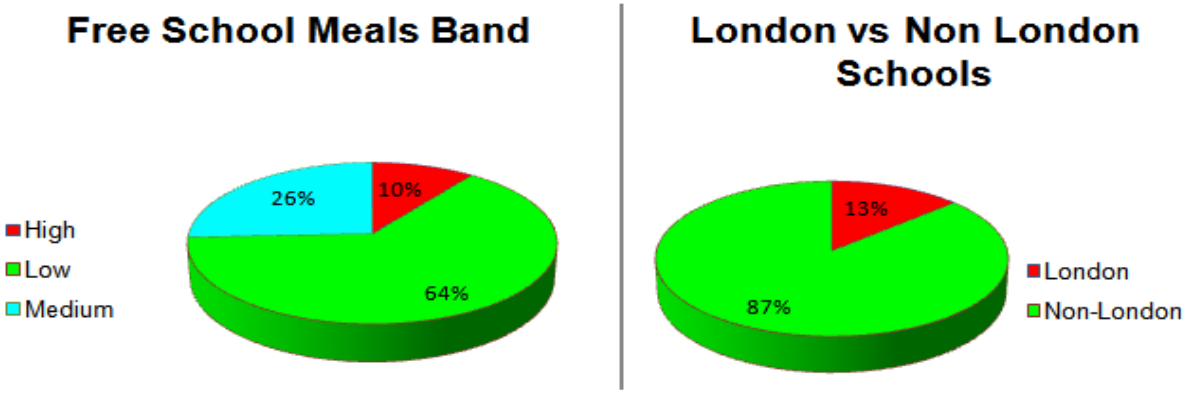


Figure 15 Pie Chart (%) a) Free School Meal Band & b) London vs. Non London

The analysis and visuals presented a number of results. The results presented that the majority of the schools are Community schools (52%) and the gender balance of the schools provided results whereby 89% of the schools were Mixed gender schools (boys and girls). Not surprising, the majority of the schools in the study are based outside of London with a high percentage of 87% outside of London. The Free School Meals eligibility band show the majority of the schools (64%) are on the Low band for Free School Meals eligibility.

There were ten numeric attributes, chosen as target attributes to contribute to the analysis. Out of the ten there are four years of values for both Performance Attained (%) & Expenditure per pupil (£). This is data is for the school years, 2009/10, 2010/11, 2011/12 and 2012/13. Firstly, the summaries of these attributes were performed individually and later a mean aggregate was created and used in the Predictive analysis models.

Numeric Summaries

The application tool R was used, to create the summaries for the numeric attributes. A series of statistics; the mean, median and first 25% and the third 75% quartiles and the extreme values maximum and minimum. By observing the difference between the mean and median as well as the inter-quartile range (3rd Quartile-1st Quartile) we can get an idea of the shewness of the distribution and also its spread. Additional tests were completed to tells us more about the variation of the dataset attributes. These numeric summaries are provided in Table 4

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Expenditure Per Student Mean (£)	4135	4982	5489	5766	6220	18160
Performance Attained Mean (%)	0.21	0.48	0.56	0.57	0.64	1.0
English First Language (%)	0.01	0.83	0.95	0.86	0.99	1.0
Disadvantaged pupils (%)	0.02	0.16	0.26	0.29	0.39	0.99
FSM Eligibility (%)	0.01	0.09	0.15	0.18	0.25	0.73

Table 4 Statistical Summary - Numeric attributes

The findings from the statistical values presented the following. The majority of the students in the study at KS4 stage, had an average of 86% of students with English as their first language. The Disadvantaged percentage in the schools was low with an average of 29% and too was the percentage of students eligible for Free School Meals (18%).

The statistical values from the individual years for the Performance attained(%) and Expenditure (£ Spend) per pupil, the tabulated results are in appendix 3. Over the four years the annual mean for both Expenditure per pupil and Performance Attained showed an increasing trend. In the case of Performance the average at 54% in the school year 2009/10 increasing to 59% by 2012/03. This was an increase of 5% in the overall performance in the percentage of students achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs. The Expenditure Per Student had also increased over the four year period from £5560 in 2009/10 to £5972 in 2012/13 an increase of £412 per student. The range average value for Expenditure Per Student had a minimum of £4135 and a maximum of £18160 over the four years. Throughout the four years the range between the minimum and maximum remained quite large.

The summary results were presented on two, bar charts prepared in R. The purpose of the graph is to visualise the trend of the averages over the four years. The trend in the attributes, Performance Attained (%) and Expenditure per pupil (Spend) are shown in the figure 16.

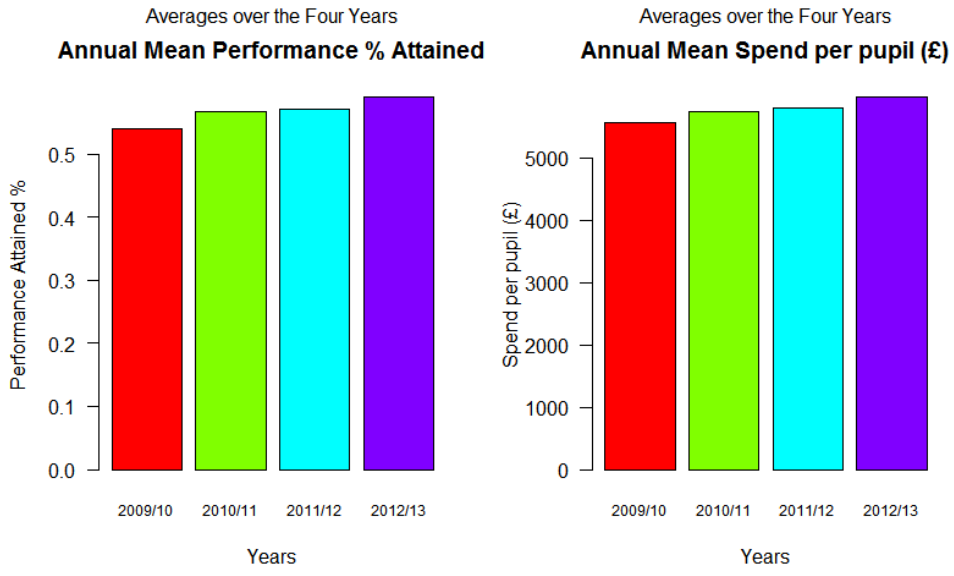


Figure 16 Bar Charts Averages over four years Performance & Expenditure

Patterns and Trends

There were a number of patterns and trends in the data that not as oblivious as the increasing trend in the Expenditure Levels and Performance Attained attributes. The trend over the four years, an increase in Expenditure levels would not have been an expectation result. The interesting element of the increase in expenditure levels was the discovery that Performance has also increased. However the association between the two attributes still needed to be established in the study.

There were other patterns, Girls schools out performed Boys and mixed schools. Another pattern presented that greater Expenditure levels occurred in London schools compared to Non London schools. Illustrated in Figures 17 the variable Frees School Meals band trend between the London Schools and the Non London schools.

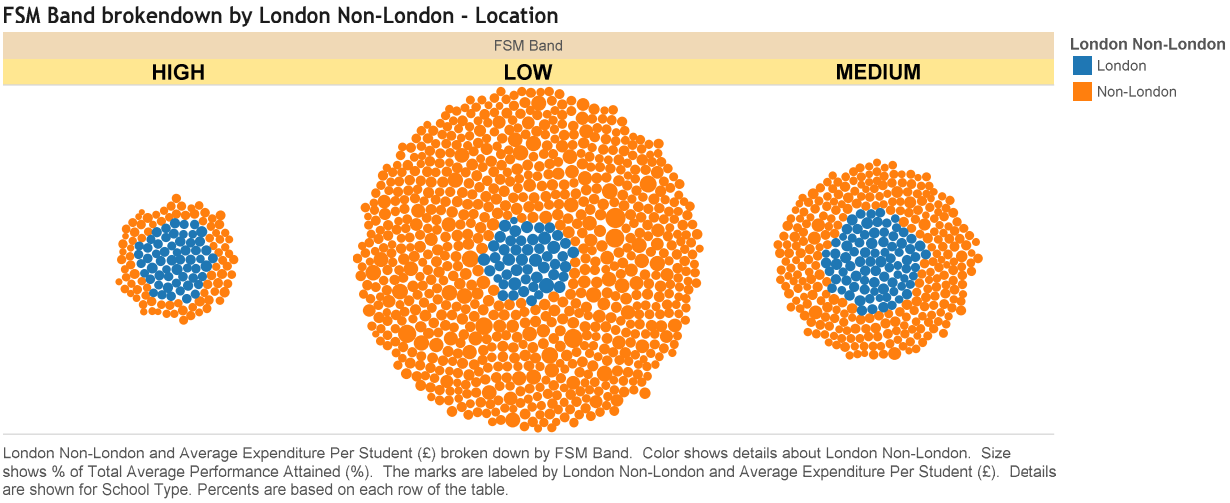


Figure 17 Trend of FSM and London Non-London Schools

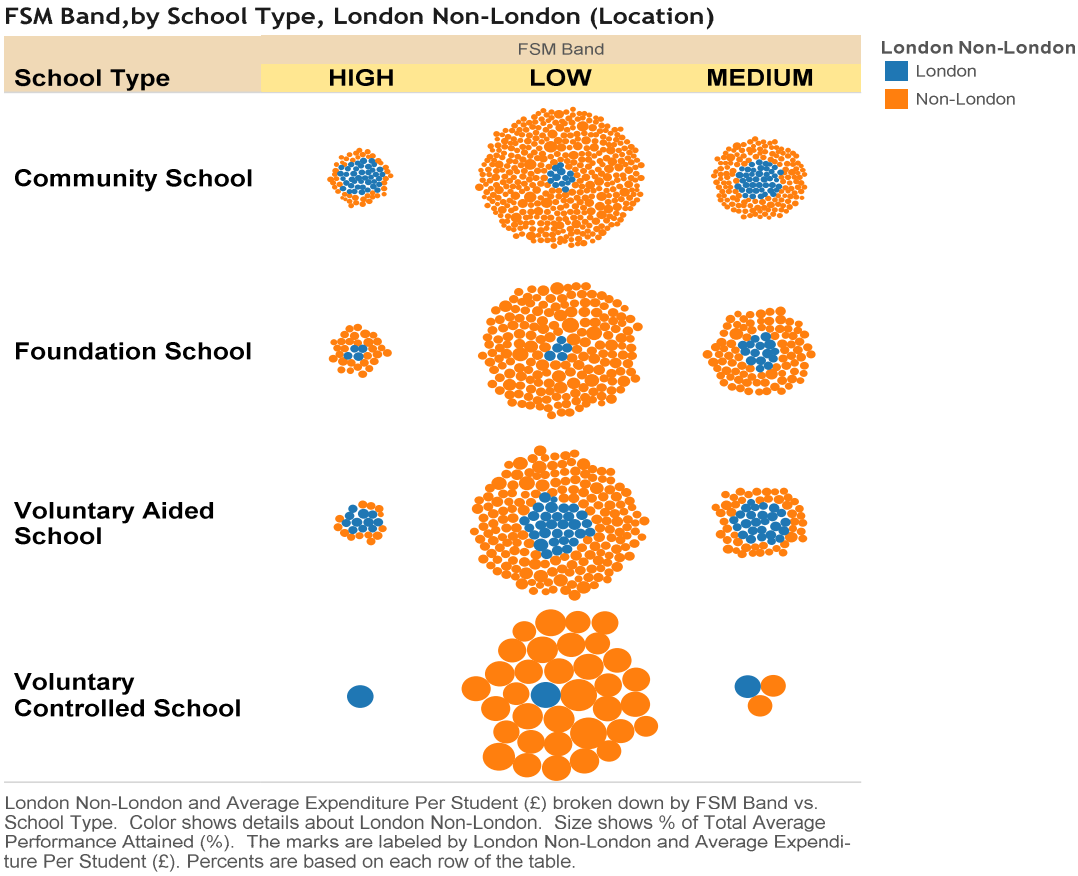


Figure 18 Trend -FSM Band, School Type and Location

In figure 18 the data was extended to include the schools types. The blue colour indicates the schools in London and the yellow Non London schools. Further variables were explored with FSM Band. In figures 19 and 20 Gender and Expenditure are explored in more detail.

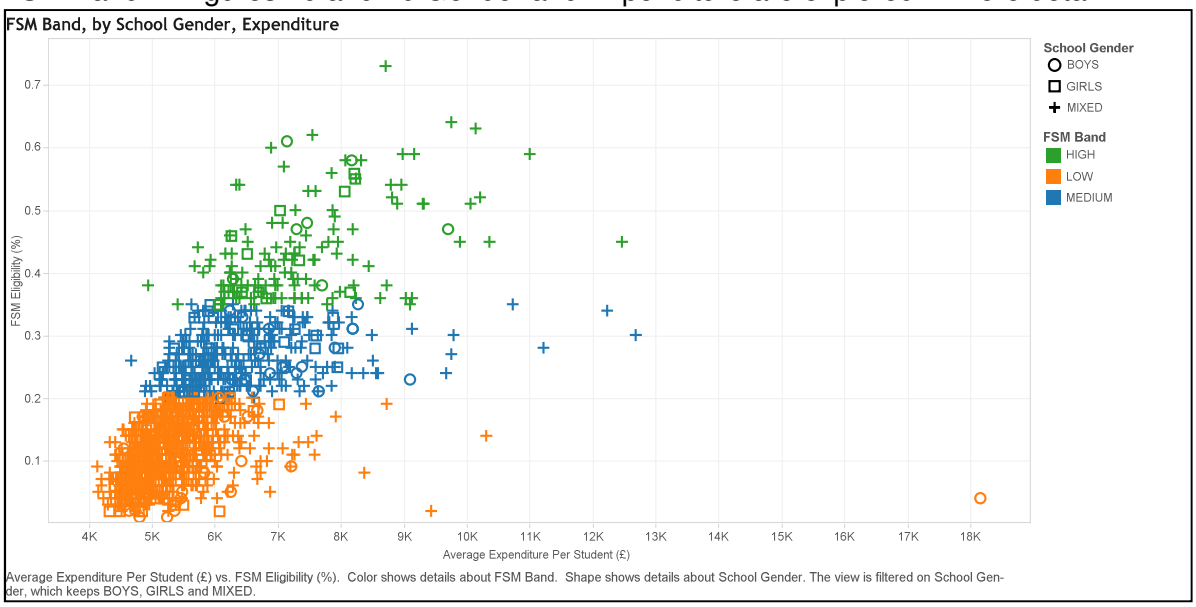


Figure 19 Trend Graph - FSM Band, Gender, Expenditure

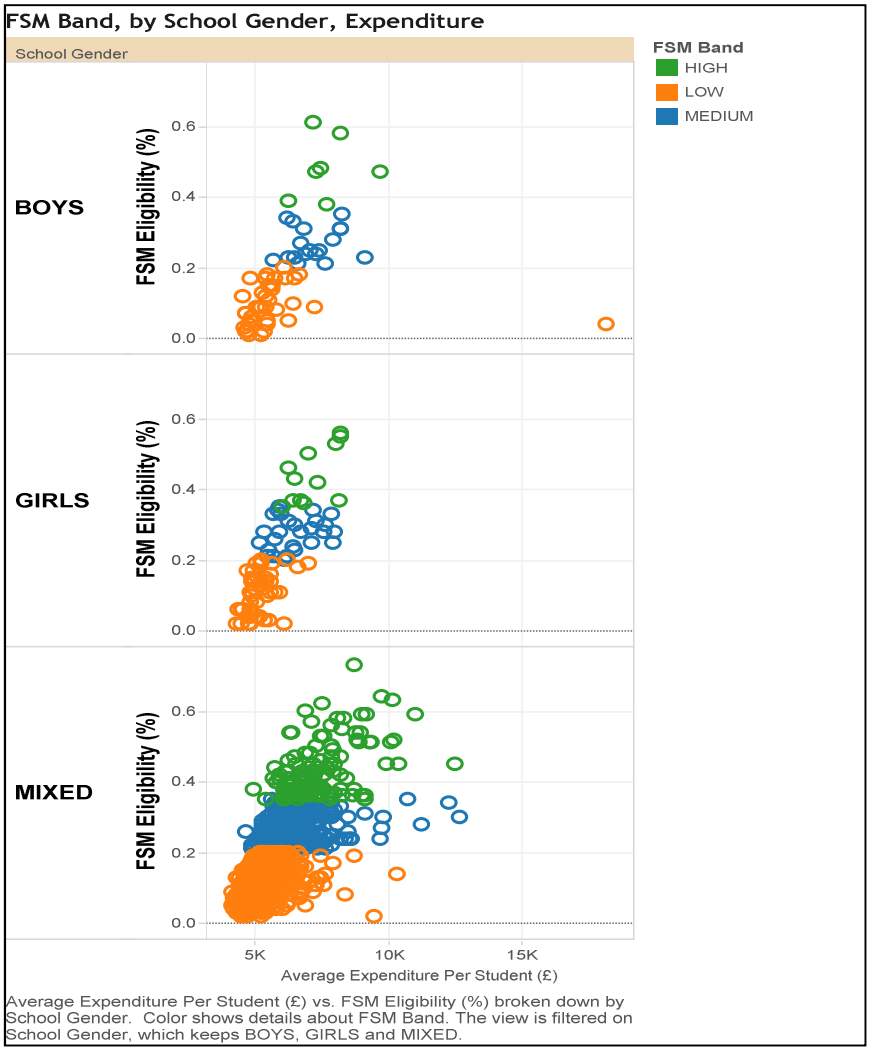


Figure 20 Trend -Average Expenditure, FSM Band and Gender

The graphs on this page focus on Average Expenditure levels per Student vs FSM Eligibility broken down by Gender

The Low Band for FSM is yellow and the funding for these schools is the lower left corner indication - low funding levels.

The graphs presented a distinct pattern in presenting the Low, Medium and High bands. As the FSM band increased so did the expenditure, this occurred regardless of Gender.

Normality and Symmetry

A number of the proposed tests required some assumptions to be upheld. If these alternative were not upheld alternatives within the suite of tests would be needed. The first assumption to be tested was to determine the normality and symmetry of the data attributes. To determine the assumption both visualisation graphs and statistic tests were completed, first on the uni-variant attributes and then the bi-variant attributes.

Histogram and Density plots

Histograms and density plots provided visualisation aids for the Normality and Symmetry analysis. The histogram, plots the frequency of the data attributes and the density the probability. In addition to the plots, both the Kernel and normal density plots were used. The histogram was used to show the distributions of independent and dependent variables. The purpose of overlaying the plots with the Normal parametric and non-parametric Kernel density plots was to determine whether the attributes are normally distributed.

The uni-variant attributes were plotted for the numeric attributes there are, English as first language(%), Disadvantaged pupils(%), Performance Attained 4-Year Mean(%), Expenditure 4-Year Mean(£), Free School Meal Eligibility(%), each of the four years for Performance Attained and Expenditure per student. The plots were carried out in R using hist() and density() functions. In addition to the to these functions there were other functions used, to enhance some of the plots. These additional functions included rug() and jitter() and used the library("car") in R.

An enriched version of the histogram according to Torgo "shows probabilities of each interval of value" In figure 21 shows the enriched histogram and density plot for the attribute mean Performance Attained(%). This chart shows the normal density curves in red and a kernal density estimate of the distribution of the variable in a broken yellow line. Near the x-axis the real values of the variable are shown, this allows for easy viewing of outliers. "This kind of data inspection is very important as it may identify possible errors in the data sample, or help to locate values that are so awkward that they may only be errors, or at least we would be better off by disregarding them in the posterior analysis" (Torgo, 2011)

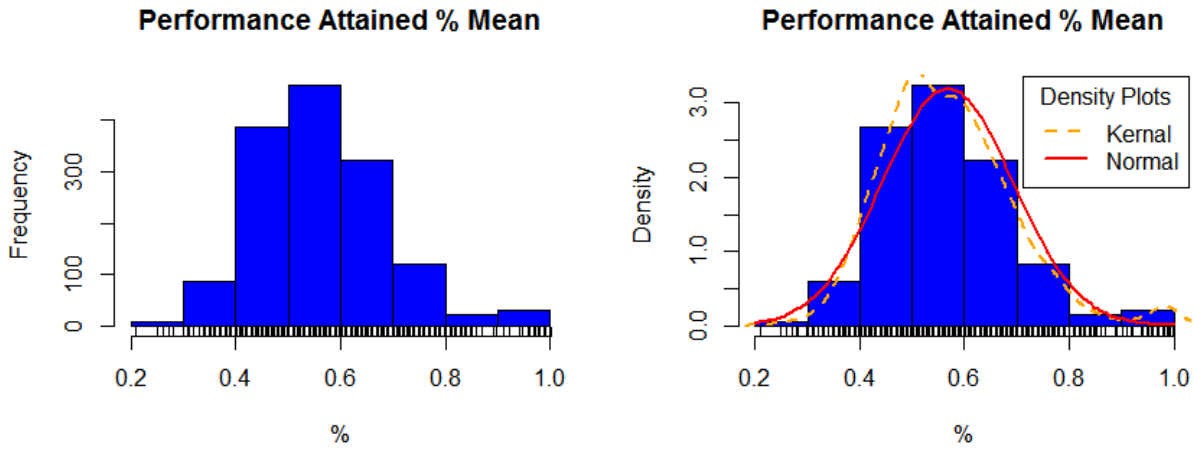


Figure 21 Enriched Histogram & Density Plot - Performance Attained % Mean

Overall, the results from the visualisation of the uni-variant attributes provided an indication that shape, symmetry and skewness for the individual data attributes was non-symmetrical and had some skewness. In Appendix 3 all the uni-variant attribute histograms and density plot are available. It was determined that further tests for normality would be carried out.

Shapiro-Wilks test

As part of this further investigation a number of normality tests were completed. One of the selected tests was the Shapiro-Wilks test for normality. The p-values results from this test, for each of the following attributes, English as first language(%), Disadvantaged pupils(%), Performance 4-Year Mean(%), Expenditure 4-Year Mean(£) and Free School Meal Eligibility(%) were low, with a small p-value < 0.05. The individual results are tabulated in Appendix 3. The p-value is small enough to reject the null hypothesis and accept the alternative hypothesis, that the data is not sampled from normally distributed data. The null hypothesis is that, the data are sampled from a normal distribution. Due to the small p-value, the deviations were "statistically significant", further tests were considered with the purpose of determining which Non parametric test should be applied.

H_0 = Null Hypothesis, sampled from a normally distributed data
Ha = Alternative hypothesis, not normally distributed sample

Boxplots

The Boxplot was used as another visualisation method for numeric attributes. The results for the attributes tested are presented in Figures 22 and 23, all of the presented boxplots showed the presence of outliers. Outliers are extreme values and in the box plots there are "values that are "far" from the middle of the distribution" (Statsoft Inc., 2013). The presence of outliers, may result in rejection of the normality distribution. Therefore, further inspections were carried out on the attributes. The boxplots below shows the five numeric attributes of interest in this study.

The attributes in this figure are as follows:

D = Disadvantaged (%)
E = English as a first Language (%)
P = Performance attained (%) 4 year Mean
F = Free School Meals Eligibility (%)

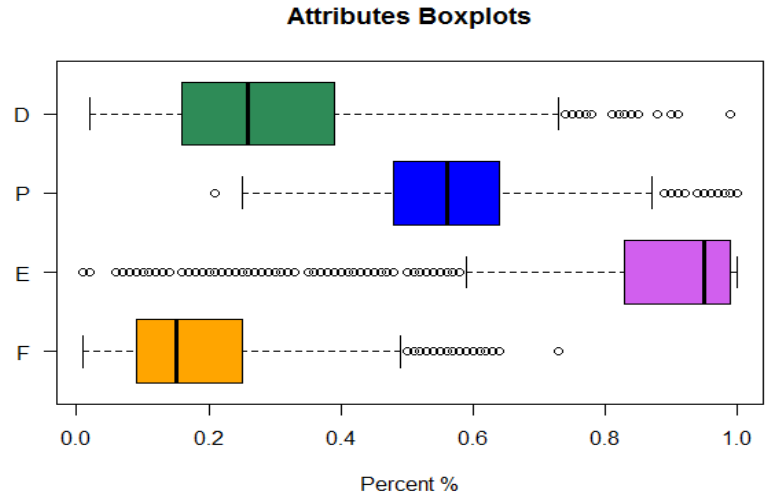


Figure 22 Boxplots Numeric Attributes

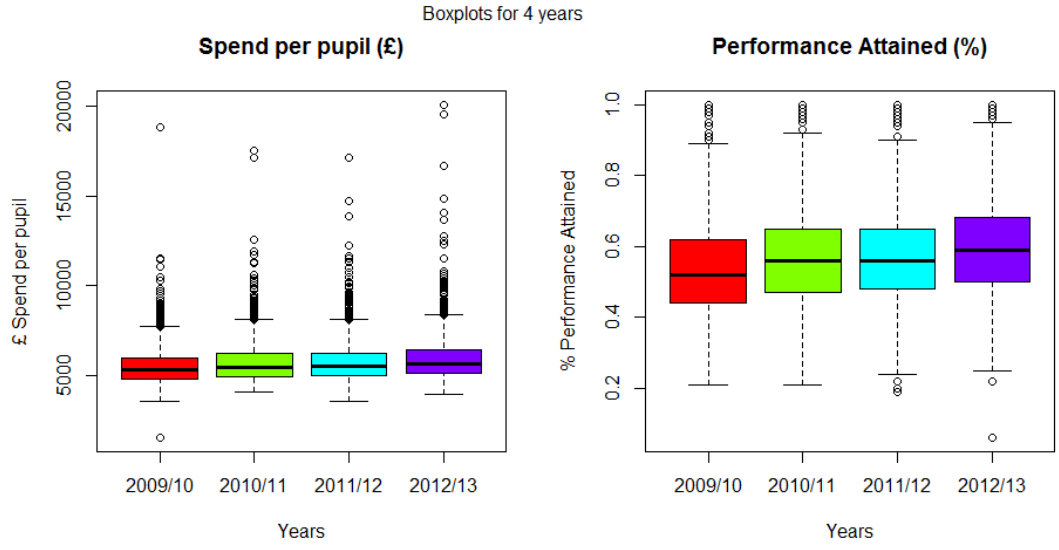


Figure 23 Boxplots - Performance and Expenditure 4 years

Further boxplots, histograms and density charts are shown in the supporting documentation Appendix 3.

QQ-Plots

Another powerful graphical check includes the QQ-Plot which is a more precise check "which plots the variable values against the theoretical quintiles of a formal distribution" (Torgo, 2011).

In figure 24, a QQ-plot for Average Performance Attained (%) can be observed, there are several lower and upper values of the distribution and that don't observe the assumptions of the normal distribution. The QQ-plot function has a default confidence interval of 95% of the normal distribution showed by the red broken lines. In figure 25 the QQ-plot for Average Expenditure Per Student shows several upper and lower values of the distribution and similar to the Average Performance Attained don't observe the normal distribution assumption. As both of these attributes are important in the study, the outliers were investigated further and decisions make on whether to exclude specific values to see if the distribution would improve.

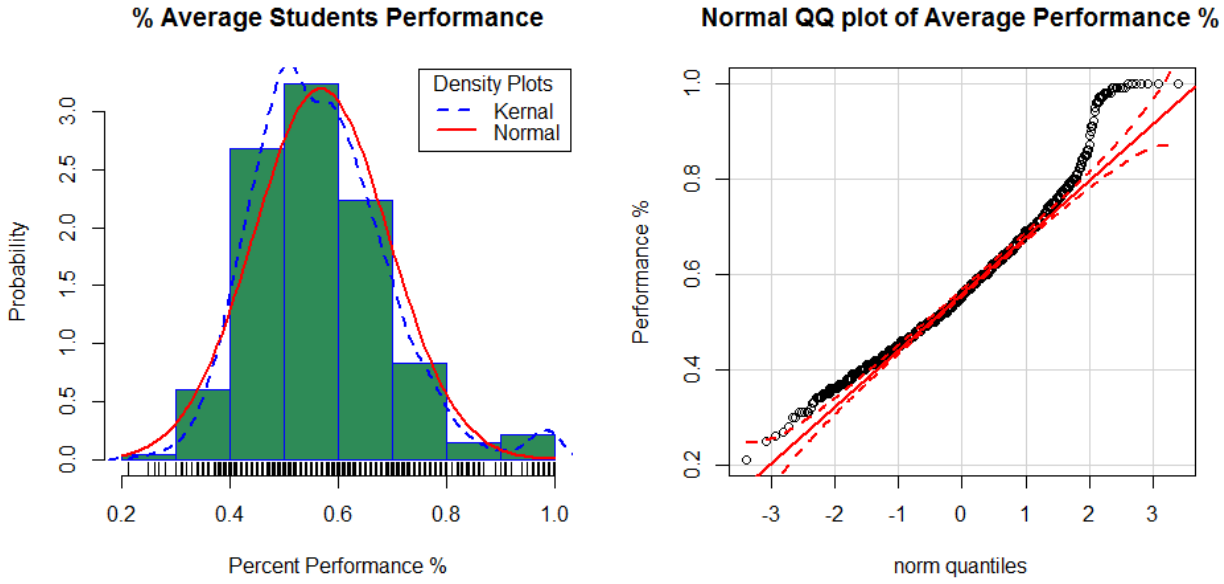


Figure 24 Density & QQ Plot for Average Performance Attained

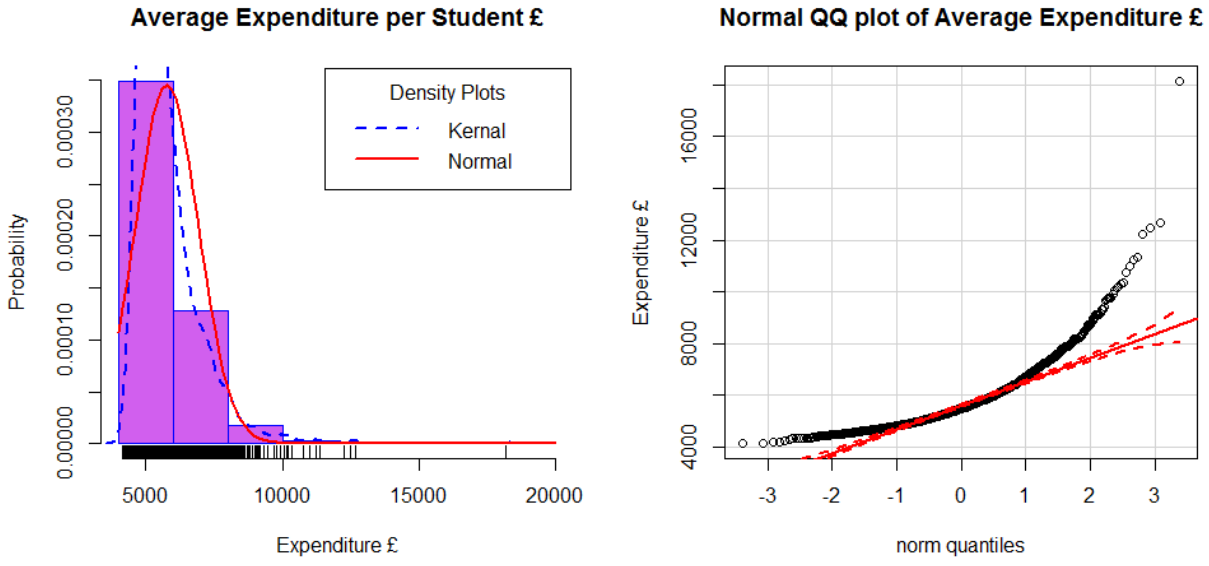


Figure 25 Density & QQ Plot for Average Expenditure per Student (£)

In addition to determining if an improvement in the distribution it was also necessary to determined whether to leave them included in future tests or to transform them.

A number of checks were carried out to review outliers in the data. The initial data check was performed in MS Excel. The data was sourced from a reliable and legitimate source, the Department of Education, UK. Having sourced the data from the Internet there was no facility to verify if any entries were input incorrectly. The attributes at this stage in the study were only looked at individually and not in the context of the overall data. They are from a legitimate source and the results are representative of the data available.

There are varying attitudes about the removal of data with outliers. Some, statisticians like Orr feel that the data is more likely to be representative of the population as a whole if outliers are not removed (Orr J.M. et al, 1991). Outliers can severely affect normality and homogeneity of variance, so it is important to decide on a method to detect the outliers.

The decision was made at this stage in the exploration not to exclude the values from the study as removal of an outlier record would mean eliminating a school from the study due to an outlier occurrence in an individual target attribute. A number of the schools, were subjected to an individual review to see if there was any evidence as to whether it might be an "indicative occurrence of a phenomenon that is qualitatively different than the typical pattern observed or expected in the sample" (Tompkins Cortland Community College, 2012). This review comprised of looking at schools which had High Expenditure and Low performance over the four years followed by a check with Outstad to see if that school had notable issues.

The findings in relation to this presented some interesting results. There were some schools closing down in 2013, another was a specialised agricultural establishment and some schools were in the process of amalgamations. A small sample of these are discussed in this dissertation to reflect on where the sample data originally came from. Wakmean school, closed in July 2013 with 18 students, it this was decided to close the school in 2011 and from then the number of students dropped. It is probable the high expenditure was required to keep the school running at a sufficient level even though the number of students had dropped due to expected closure. Brymore Academy an agricultural boarding school, its expenditure level was high and as a school that specialised in agricultural courses the numbers taking GCSE exams, were low. Schools in the lower quartile had Ofsted reports which reflected the lower than average performance. In one school Hameldon Community College, the Ofsted 2013 report showed a drop in points for Mathematics. In one year, there was a 28% decrease in percentage points. As this is not a Dissertation investigating individual schools only a few schools were looked at. The samples taken were on the high expenditure and low performance levels.

Mann Whitney Test

This is a test is an Inferential Statistics test and a non-parametric test which is used in order to overcome the underlying assumption of normality in parametric tests. This was used as an alternative to the t-test. It was used to compare two independent group samples means, where normality distribution can't be assumed.

The null and two-sided research hypotheses for the nonparametric test are stated as follows:

H_0: The two populations are equal versus

Ha: The two populations are not equal.

The results from this test on the paired numeric attributes presented a p-value is less than 0.05 in all the Mann Whitney tests. At a significance level of 0.05, we can conclude that the data y and x in our study are non identical populations.

4.3.1.2 Exploring Multivariate's

Correlation

As one of the research questions was to investigate the association Expenditure levels per student with Performance Attained, it was appropriate, to include a correlation test. Correlation refers to the degree to which two or more sets of data show a tendency to vary together. The correlation coefficient when positive has varying strengths between 0 and 1 and when negative can has varying strengths between 0 and -1.

A common measure of correlation is Pearson Product Moment correlation, however there are alternatives. Spearman Brown is an alternative and both tests were utilised in this dissertation, primarily to increase the reliability of the estimates produced. It was known in advance that specific assumptions for Pearson's correlation were required and the data did not uphold these requirements. Data tested with Pearson's correlation should uphold assumptions they are, that the data should be interval or ratio level and linearly related and bi-variant normally distributed. Spearman's test had no requirement of normality and is a non-parametric statistic. It is therefore a more appropriate test for this data.

The results of the Correlations using Spearman test indicated that there is a positive relationship between Disadvantaged Students(%) and FSM Eligibility(%) with a results of 0.95 for the correlation coefficient. The Average Expenditure per Student level(£) and Disadvantages Students(%) presented a positive correlation of 0.75. The correlation coefficient for the relationships between Average Expenditure level(£) and Average Performance Attained(%) at -0.53 indicated an intermediate negative correlation. The results do not imply causation.
A correlation significance matrix using rcov() was also prepared in both Spearman and Pearson test. The purpose was discover what significance levels were presented and how they compared in both tests. The tables, using the rcov() function indicated that Average Performance Attained(%) and English First Language(%) has a significance level. In Spearman test, the significance level result was 0.004 and in Pearson Coefficient significance table the same pair of variables had a result of 0.1694.

Using the Spearman test, the correlations were calculated for the four years between the attributes, Pupil Performance (%) Attainment and Total Spend per pupil (£). The purpose of this was to see what type of relationships the variables had prior to the aggregated values looked at earlier. Not unsurprisingly the relationships between the four years of Pupil Performance (%) Attainment values showed a strong positive relationship ranging from a result of 0.84 to 0.86. A similar result was discovers between the four years of Expenditure per Student (£) levels with a range of values between 0.7 and 0.83. The relationship between the two variables, Pupil Performance (%) Attainment and Total Spend per pupil (£), indicated a moderately weak negative relationship with a range from -0.39 to -0.55.

Scatter plots

The visualisation of the data was necessary, using scatter plots on the multivariate values. This option was used to explore the relationships between two categorical variables using a two-way table. The creation of paired scatter plots presented illustrates the associations of the correlation coefficient on one matrix. This was created in R using the pairs() function. The scatter plot presented all of the numeric attributes in a matrix which can be viewed in Figure 26. The visualisation of data in a pair matrix, provides an overview, but the process can be a somewhat subjective process, hence the need for individual scatter plots. The scatter plots chosen to view were determined by results in the Matrix in Figure 26.

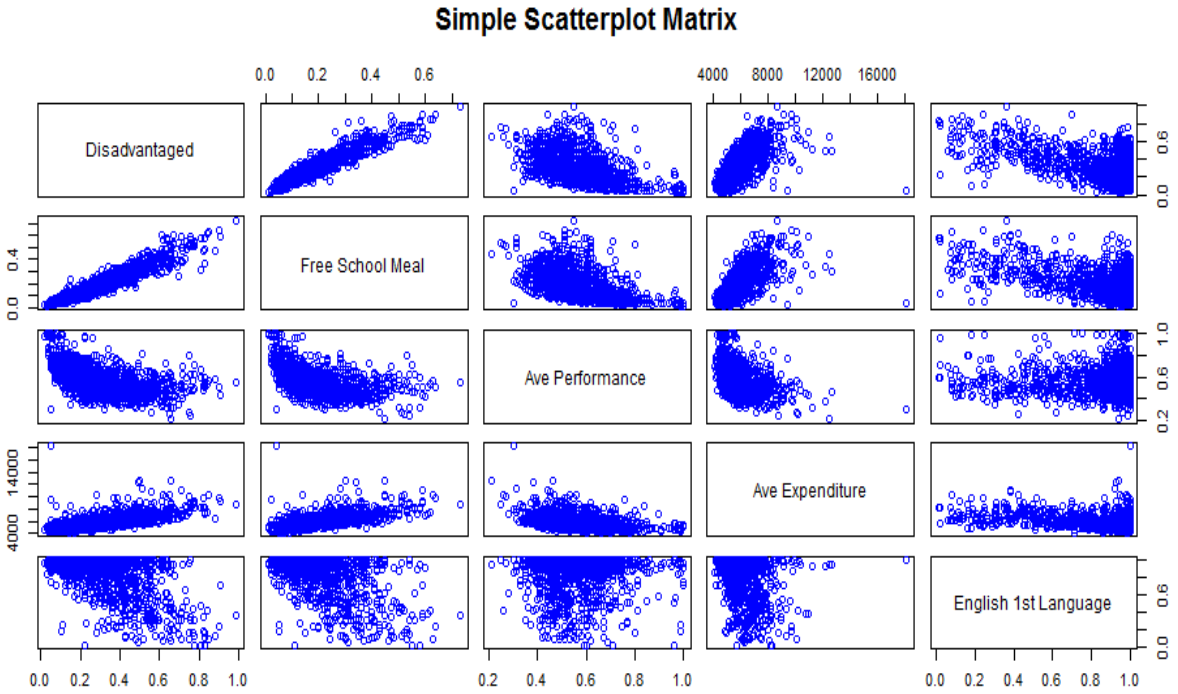


Figure 26 Pairs Scatter plot

The pairs scatterplot above in Figure 27 indicates results similar to the correlation coefficients. The scatterplots suggests that there was a positive relationship between Disadvantaged Students(%) and FSM Eligibility(%), however evidence of non-linearity from the plots was hard to determine. It was decided to create further charts and view the bi-variants using an enhanced scatterplot( ) function in the r, car package. This application offered many enhanced features, including fit lines, marginal box plots and conditioning on a factor. In figures 27 and 28 the variables of interest at this stage in the study were plotted using the enhanced method. The enhanced scatterplots provide boxplots in addition to the scatterplot. The boxplots for Disadvantaged Students(%) showed the median was closer to the Upper Quartile, the skewness was already investigated, its static result did show a moderate skewness when explored individually. From the boxplot for FSM Eligibility(%), the median indicated it was closer to the lower quartile which would suggest a positive skewness. It was noted that Pearson's test is sensitive to skewness so using the alternative, Spearman was more appropriate for reporting the relationships.

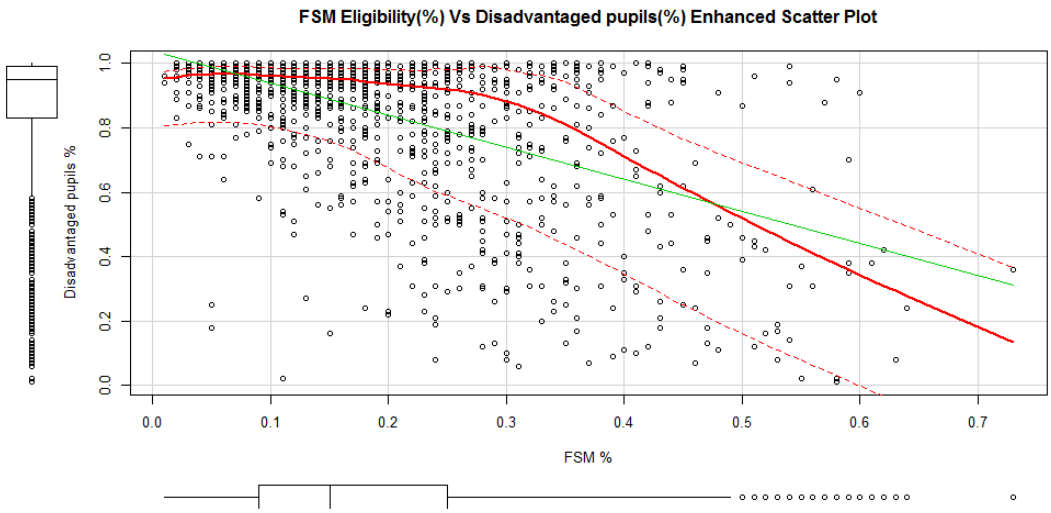


Figure 27 FSM Eligibility (%) v Disadvantaged (%) (Enhanced Scatterplot)

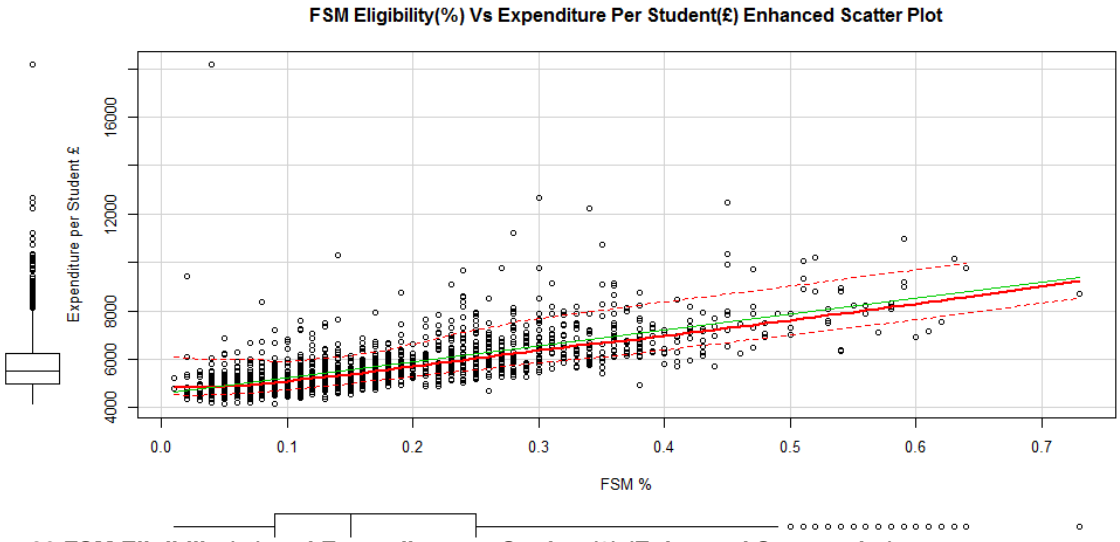


Figure 28 FSM Eligibility(%) and Expenditure pe Student(£) (Enhanced Scatter plot)

Boxplot analysis was also carried out the individual attributes, the results indicated skewness was present in the target attributes which included, Disadvantaged Students(%),FSM Eligibility(%), Average Performance Attained(%) and Average Expenditure levels(£), The skewness indicated a positive skewness, which was of a moderate to high level of skewness. In the attribute English as First Language(%), it also showed skewness which was highly negative.

Determine Relationships

After checking the distributions of individual variables, it was determined that the relationships would be explored using multivariate the analysis MANOVA model.
The initial task was to create scatter plots, to graph the relationship between two variables in a data set. The x and y axes are used for the values of the two variables and a symbol (circles in the figure 29 and 30). The graphs represented a combination of pair values in the data set. This type of graph is commonly used in many situations and can convey a lot of useful information.

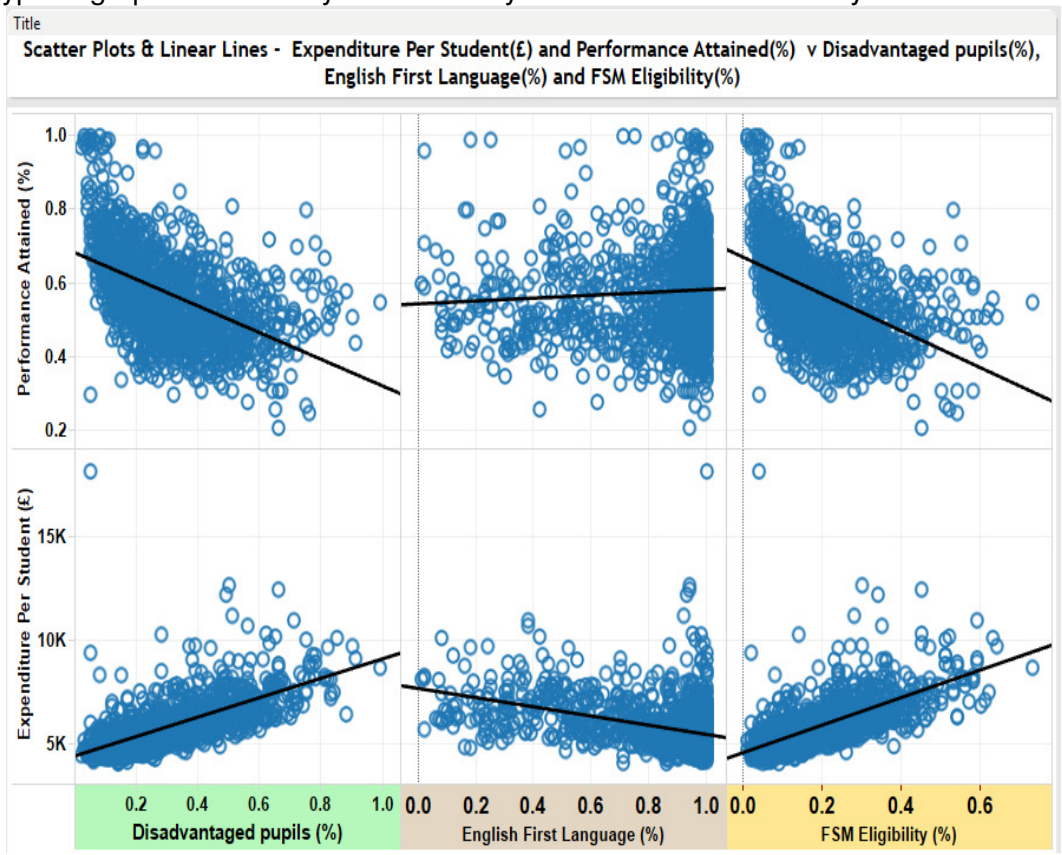


Figure 29 Scatter plot Multivariate Analysis

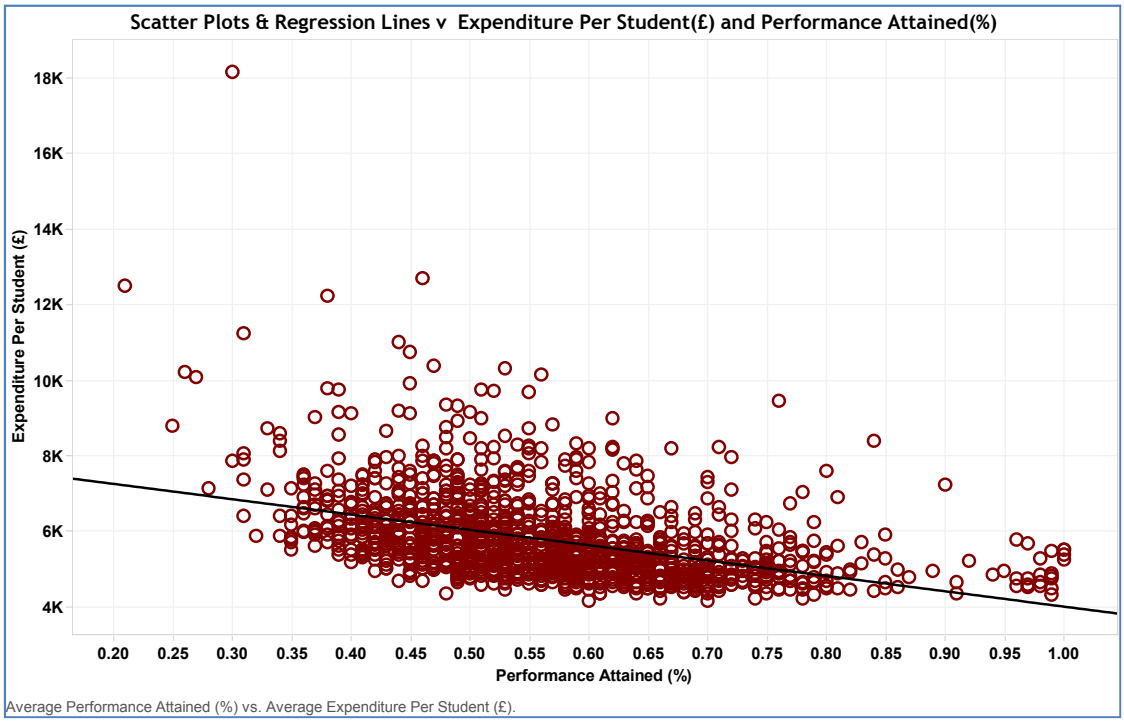


Figure 30 Scatter plot, Bi-variante Analysis Expenditure v Performance

The charts in figures 29 and 30 were created in Tableau a powerful visualisation software. This software was chosen to complement the tests and Models created in R. It also provided results which could be used as comparison technique. to evaluated the results by cross referring. The cross-check was carried out on the regression models were built in R with these provided on Tableau. When reviewing the results the data had similarities but were not identical in all cases. The complex process of analysing data using two different tools can provide varying results. This was also seen when similar tests were carried out in Excel.

Regression

A regression analysis was built using R. The R functions used were ln() and summary() to provide the output for the analysis. The information extracted from the results included the R-squared, which is the coefficient of determination of a linear regression model. The coefficient is the proportion of the variances of the fitted values and observed values of the dependent variable. The purpose of this test was to discover further information on the relationships within the data and if there were associations of note.

Two models were build using the variables, Performance Attained and Expenditure per Student as the independent attributes in the two models for multivariate analysis models. The models were created to find out more about the associations of the independent variables and the dependant variables. In each model built, the dependant variables were the same, Disadvantaged Students(%), FSM Eligibility(%) and English as First Language(%). Each model provided results output on the significant of the R-square scores in the model containing the variable In the first model using Performance Attained (Model1), the R-squared value was 37.5%. With a significant level for the hypothesis set at 0.05. then with the p-value <0.05 Null Hypothesis can be rejected and the Alternative hypothesis accepted. In the results of Model1, there were three significant variables, which included the intercept. All of the dependant variables had a p-value <0.05. Hence there was a significant relationship between the variables in the linear regression Model1 and Performance Attained. The R-squared is a good indicator of how well our model is fitted. The value was low at 37%, in Model1 but it was significant enough not to be neglected. So based on Model1, Performance levels had some association with the variables, but it was not possible to give the nature of it relationship.
The model was replicated with Expenditure levels (Model2) as the Independent variable. In this second model, the predicted intercept. FSM Eligibility and English as First Language all indicated significant levels where the p-value < 0.05. In the case of the Disadvantaged pupils the p-value was greater than 0.05 at 0.52. The R-squared valued provided a result which was

higher than the previous model at 48.2 %.As with Model 1, presence of a relationship was indicated for Expenditure levels and the dependant variable, but not the nature of the relationship. It is accepted that the data within this Dissertation investigates only limited number of variables that could influence the result. There are many socio-economic influences that could influence the relationships but not included in this dissertation. A more in depth, analysis and result output are provided in appendix 3.

Residuals Plots

In the regression models it was possible to observe the difference between the dependant variable and the fitted y values, using a residual plot. The residual plots were created using Res() in r and the results are presented in Figures 31 and 32.

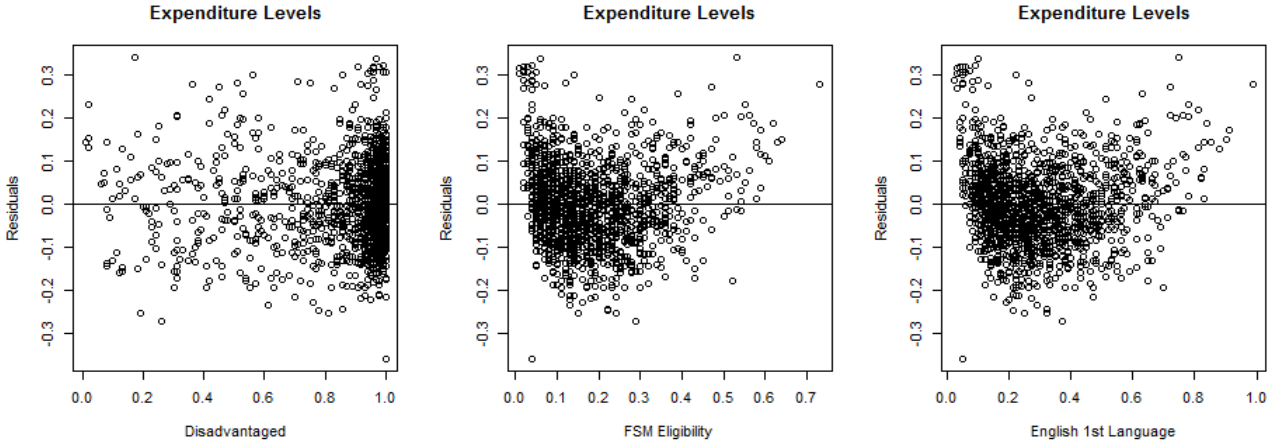


Figure 31 Residuals Regression plots Model1

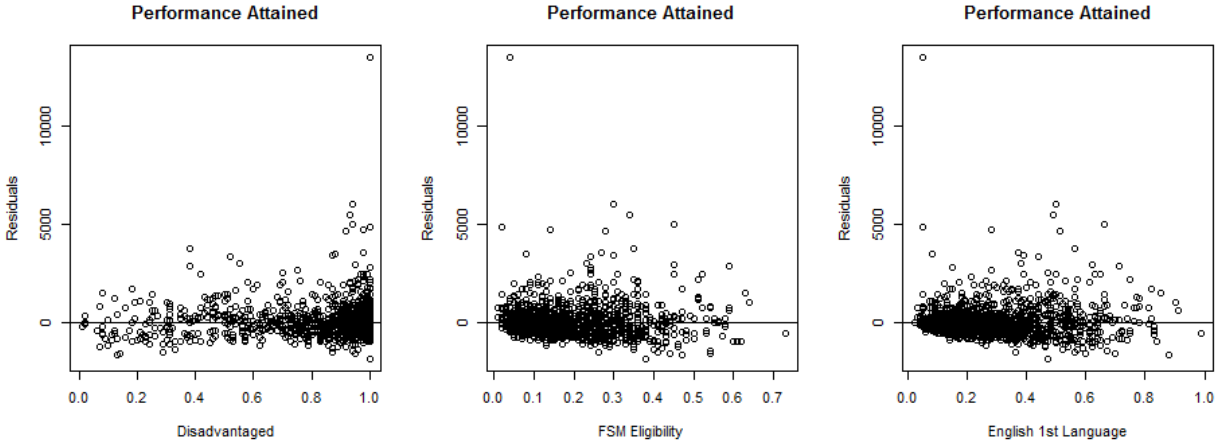


Figure 32 Residual Regression plots Model2

MANOVA

A model was also built to included Multiple Analysis of Variance. This was built in R, using aov(), lm() and summary. The MANOVA results tables confirmed that there are differences between the groups which were highlighted in the model summary. Residual Standard errors were small 0.097 in the Model1, however in Model2 the Residual Standard errors were 826. When the two models were tried in a test together, an error in the type of data was referenced to, hence the MANOVA models were not combined.

Regression and Decision Trees

Both a Regression Tree was created and a Decision Tree.

Regression Tree

The purpose of creating a Regression Tree was to see if there were stronger possible relationships to explore in the data, in particular with the nominal data within the Analysis. The Regression Tree in Figure 33, illustrated Expenditure levels at the root node level followed by the nominal attributes Gender, London Non London, Schools types and FSM Bands at node level. Patterns relating to these attributes indicated trends in Gender and London Non-London.

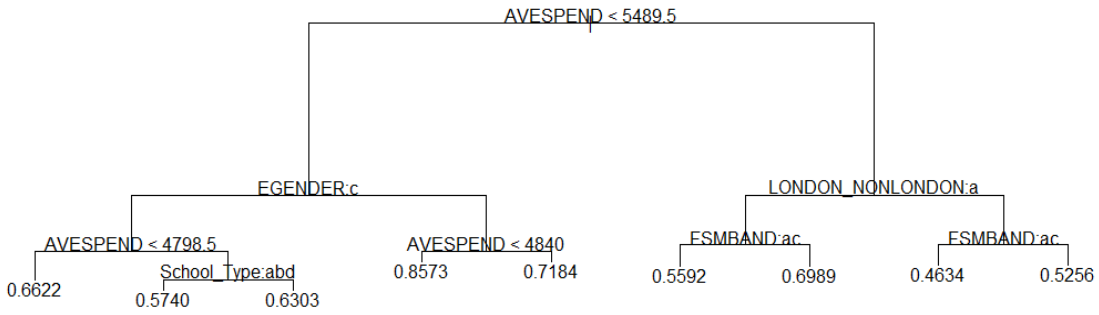


Figure 33 Regression Tree

Decision Trees

A number of Decision trees were created n both R and Weka.

In Figure 34 a small Decision Tree was created using only Gender and Free School Meals Band. The initial purpose of completing this test was to discover if FSM Bands could be further classified by other attributes. This process using Weka, was not investigate sufficiently to provide results of note. The decision tree in R in figure 35 was more informative.

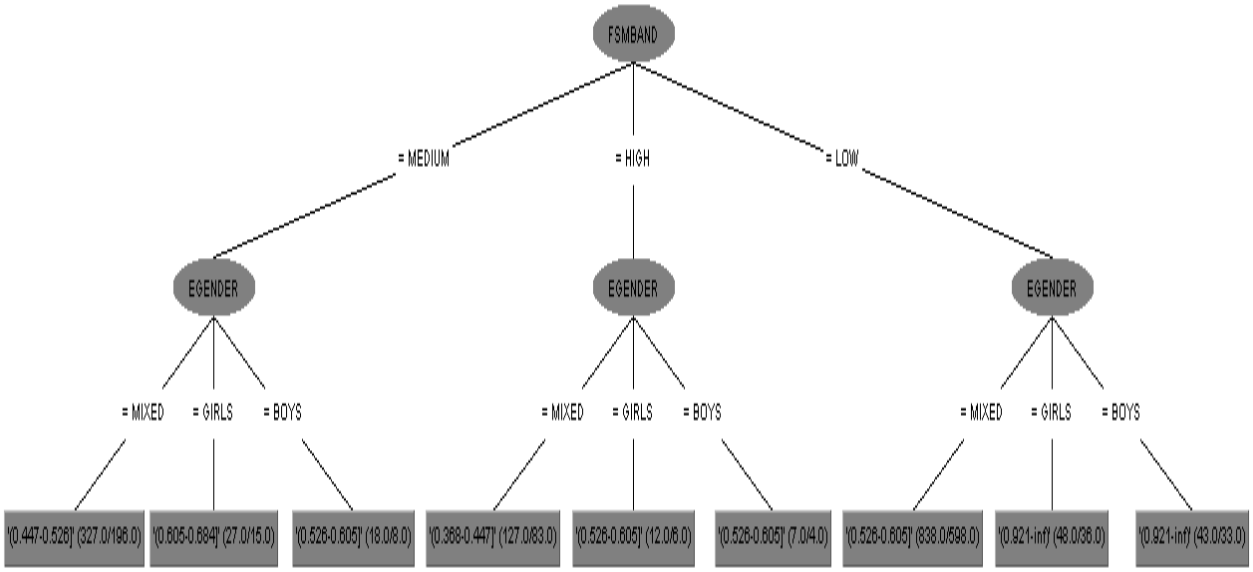


Figure 34 Decision Tree - Weka

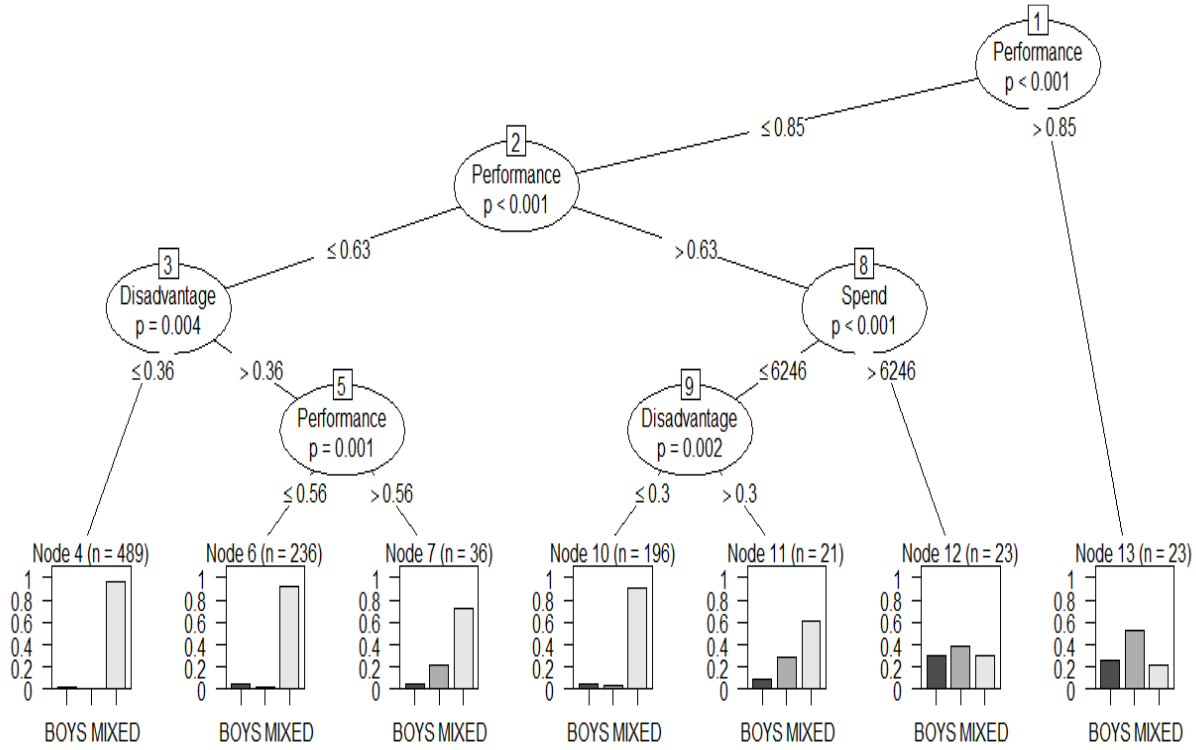


Figure 35 Decision Tree - R

The decision tree could be presented using the rules or a plot. The plot is generally easier to follow. The decision tree in figure 35 required more formatting so as to clearly determine the Gender details shown on the end node.

Interpretation

The Decision Tree created in R illustrated Performance dividing into two nodes the <=85% and the > 85%. In total there were 9 divides shown for Gender and the attributes Performance, Spend and Disadvantaged.

The right brand had just one end node, which was the breakdown of schools by gender.

The left brand provided more information and split Performance again into the <=63% (left branch) and the > 83% (right branch). The interesting finding in the Decision Tress is the split at this level between Left branch divide and the right branch divide. The Left divide from Performance <=63% goes to Disadvantaged while the >63% goes to Expenditure.

Continuing with the left node the split at this level for Disadvantaged, the left divide shows <= 36% with one node to the breakdown of schools by Gender. The right divide had another divide in Performance before the final end node.

Going back to the >63% performance the Expenditure levels divide into <=6246 and >6246. The left branch has another divide before the node ends, This node is Disadvantage with a divide of <=30% and >30%

Having completed the analysis to this stage the Dissertation results illustrate a number of items of interest. In the conclusion in Chapter 5, the items are summarised and further work recommended

Chapter 5

5 Conclusions

5.1 Overview

This Chapter concludes the Dissertation and how the tests completed in Chapter 4 relate to the research questions prepared at the beginning of the dissertation. At the commencement of the Dissertation the background and the motivation for the Dissertation were presented. The background and motivation assisted in compiling the aims and research questions. The solution was outline in Chapter 1 on how Data Mining and Analysis processes would be used to provide a solution to the Research in this Dissertation.

Research Questions

1. What associations (or interactions) has student Expenditure Levels Per Pupil with Performance Attained?
2. What relationships (or interactions) has student Performance Attainment(%) at Key Stage 4(KS4) to the other variables; Free School Meals Eligibility, Disability, English as First Language?
3. What relationships (or interactions) has student Expenditure Levels Per Pupil to the other variables; Free School Meals Eligibility, Disability, English as First Language?
4. Are there significant effects on the dependent variables
5. Are there patterns and trends within the data?
6. What relationships has the categorical variables, Gender, School type, Free School Meal Band and location (London verses Non London) with Performance Attainment and Expenditure levels?
7. What associations or relationships in the data can decision trees determine?

The process of exploring the data for associations, using Data Mining processes required extensive Data Preparation. This preparation of data was followed by the Analysis and interpretation of the results. The assessment of the results and with particular reference to the research questions been asked are of importance. The following are the findings:.

Associations between Expenditure levels per Pupil with Performance Attained.

The Dissertation motivation was due to an interest in Education and initially performance. Related literature provided interesting concepts about performance and what influences Students performance. A reports prepared by the OEDC had an interesting finding relating to Performance Attained and Expenditure levels, stating that a 9% variation in Performance could be explained by expenditure levels. This Dissertation does not quantify the relationship between the two variables. However the results do provide information in favour of an association. A relationship was present using the Data in the Dissertation and Data Mining algorithms and models. The result, from Spearman's Coefficient test provided an intermediate positive correlation value of 0.53, between Expenditure Levels and Performance Attained. This result does not however imply causation, that if more capital was spent on education the performance results would improve.

Performance Attained and Expenditure Levels Relationships with other variables

The Correlation and Regression models created indicated that there was a strong positive relationship between Average Expenditure levels and Disadvantages students, the value of the correlation was 0.75. There was a stronger relationship between Disadvantaged Students and FSM Eligibility with a strong correlation of 0.95 for the coefficient. This was not know prior to the Dissertation. The Correlation test also indicated that the relationships between English as a First Language and Performance (0.08) was very weak, as was its relationship with Expenditure Levels (-0.38). Performance and FSM Eligibility had a stronger relationship of 0.61

Effects of Significance

The level of significance was explored by two Regression models, to discovered the interactions of the coefficient responses of the dependant variable to the Expenditure Levels (independent).

The results indicated a significant level on the intercept and two of the dependant variables FSM Eligibility (%) and English as a First Language(%). The Disadvantaged Students (%) variable had a p-value > 0.05. The Performance Attained models indicated that the intercept and all three dependant variable had significant levels with p-values<0.05.

Patterns and Trends

Overall, there were a number of patterns and trends in the data, not as oblivious as the increasing trend in the Expenditure Levels and Performance Attained attributes. The patterns and trends illustrated increasing trends over the four years data available for Performance Attained and Expenditure levels. Overall there was a 5% increase in the four years for student Performance Attained and an increase of £412.

There Performance level of students illustrated a higher trend of Girls schools compared to Mixed and all boys schools. Expenditure levels per student was also notably higher in the London are compared to schools outside of London.

FSM has three bands levels, low, medium and high. When graphed with Gender and Expenditure Levels the pattern on the graphs presented a distinct pattern in presenting the Low, Medium and High bands. As the FSM band increased so did the expenditure, this occurred regardless of Gender.

Decision and Regression Trees

The Decision Tree created in R illustrated Performance dividing into two nodes the <=85% and the > 85%. In total there were 9 divides showing for Gender and the attributes Performance, Spend and Disadvantaged. The divide which was most interesting was where Performance split into <=63% and >63%. This showed that Disadvantages Schools of <=36% have lower performance levels.

5.2 Further development or research

There are a number of areas where further development and research could be applied to this project. By applying these options the Data mining process would increase in scope and be a more enhanced model. The key areas for development are, the decision tree, explore other attributes for Trends and pattern and a more automated data warehouse and data preparation process. The areas for research included the incorporation of more data relating to socio-economic influences.

The Decision Tree provided some interesting observations. A recommended to include FSM Eligibility as a variable of interest in an extended Decision Tree. In addition to this variable the FSM Band would also provide in option to extend the model. The root node in this Decision Tree used Gender, changing the root node in future models should be considered. There is a lot of potential to enhance the Decision Tree model and increase the scope. Future development lies in the investigation and use of Classification Models.

The Regression and Correlation models, did provide interesting findings. However due the limited data in available in this Dissertation relating to Socio Economic influences the results would not be conclusive. An option to progress this analysis further would required data from other sources. This would involved further research, to both source and determine what approach to take. If additional data were included, the Regression Tree could assist in the exploration of the relationships of the Nominal data. The factor plot could be used as well to create a better model fit.

The Data warehouse and preparation of the data could be developed into a more automated process and to develop the integration of the applications.

The patterns and trends could be expanded to find, more patterns and trends than lie undiscovered, using the Local Authority and the Regional information. The data was mapped on a map but it was not explored to give potential results. This is another area that would provide relevant information and for further development.

In relation to Robustness and reliability, if further data is included then further consideration relating to both robustness of the methods currently used in this Dissertation and their reliability would need to taken into account. Overall there is potential to increase the scope of this project to provide a more Robust set of models.

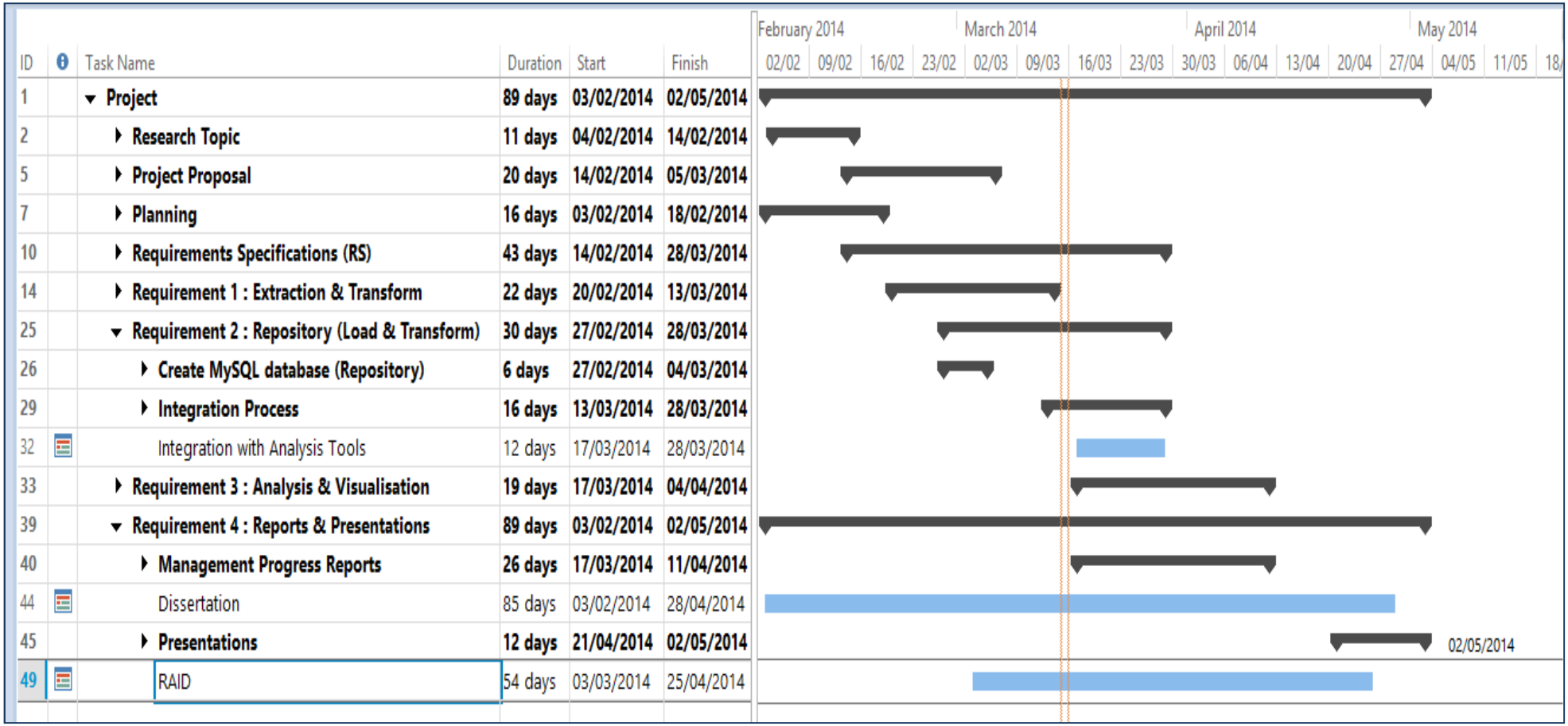
Bibliography

Aerd Statistics, 2014. *Descriptive and Inferential Statistics.* [Online]
Available at: https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php
[Accessed 5 May 2014].
Department of Education, 2013. *Evaluation of Pupil Premium Research Brief July 2013,* Manchester: Department of Education.
Department of Education, 2013. *Income and expenditure in academies in England: 2011/12 (Experimental Statistics).* [Online]
Available at: http://www.education.gov.uk/schools/performance/download/SFR24-2013.pdf
[Accessed 25 April 2014].
Department of Education, 2013. *Use of Crown copyright material.* [Online]
Available at: http://www.education.gov.uk/help/legalinformation/a005237/crown-copyright
[Accessed 15 February 2014].
Department of Education, 2014. *Consistent financial reporting.* [Online]
Available at:
http://www.education.gov.uk/schools/adminandfinance/financialmanagement/consistentreporting
[Accessed 28 February 2014].
DfES Publications, 2001. The Relationship Between Capital Investment and Pupil Performance: an Analysis by the United Kingdom. *PEB Exchange THE JOURNAL OF THE OECD PROGRAMME ON EDUCATIONAL BUILDING,* Oct, Volume 44, p. 8.
Explorable.com, 2014. *Mann-Whitney U-Test.* [Online]
Available at: https://explorable.com/mann-whitney-u-test
[Accessed 15 April 2014].
Explorable.com, 2014. *Multiple Regression Analysis.* [Online]
Available at: https://explorable.com/multiple-regression-analysis?gid=1586
[Accessed 5 April 2014].
Geoff Pugh et al, J. M. a. J. G., 2008. *Resources and Attainment at Key Stage 4, Estimates from a Dynamic Methodology,* Staffordshire University: Institute for Education Policy Research, Staffordshire University.
GESIS, 2014. *GESIS: European Values Study.* [Online]
Available at: http://www.gesis.org/en/services/data-analysis/survey-data/european-values-study/
[Accessed 16 February 2014].
GM-RAM Limited, 2010. *Summarising data using histograms.* [Online]
Available at: http://www.wekaleamstudios.co.uk/posts/summarising-data-using-histograms/
[Accessed 14 April 2014].
Information Standards Board, 2014. *About the ISB.* [Online]
Available at: http://www.education.gov.uk/escs-isb/about/whoweare/a0075244/stakeholders
[Accessed 24 February 2014].
Kabacoff, R. I., 2014. *Tree-Based Models.* [Online]
Available at: http://www.statmethods.net/advstats/cart.html
[Accessed 15 May 2014].
Kin Fun Li, D. R. &. F. S., 2013. *Predicting Student Academic Performance.* Victoria, Canada, IEEE Computer Society, p. 27.
Livephysics.com, 2014. *Calculate Linear Regression and Graph Scatter Plot and Line of Best Fit.* [Online]
Available at: http://www.livephysics.com/tools/mathematical-tools/calculate-linear-regression-graph-scatter-plot-line-fit/
[Accessed 19 April 2014].
Ministry of Education of New Zealand, 2013. *Annual expenditure per student.* [Online]
Available at: http://www.educationcounts.govt.nz/__data/assets/pdf_file/0008/143891/2013-Indicator-inID-2043.pdf
[Accessed 4 May 2014].
National Archives, 2014. *Open Government Licence for public sector information.* [Online]
Available at: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[Accessed 15 February 2014].

OECD, 2009. *Viewing The United Kingdon School System Through The Prism of PISA.* [Online]
Available at: http://www.oecd.org/pisa/46624007.pdf
[Accessed 16 February 2014].
O'Loughlin, E., 2012. *How to Edit a Basic Gantt Chart in Excel 2010.* [Online]
Available at: http:/www.eugeneoloughlin.com/2010/10/how-to-edit-basic-gantt-chart-in-excel.html
[Accessed 17 February 2014].
Orr J.M. et al, P. S. a. C. D., 1991. OUTLIER DETECTION AND TREATMENT IN I/O PSYCHOLOGY: A SURVEY OF RESEARCHER BELIEFS AND AN EMPIRICAL ILLUSTRATION. *Personal Psychology*, pp. 473-486.
Rapid Miner, 2014. *Rapid Miner Data Minining Use Case and Business Analysis Applications,* s.l.: Chapman & Hall/CRC.
RStudio, 2014. *R Studio Projects.* [Online]
Available at: http://www.rstudio.com/projects/
[Accessed 17th February 2014].
Stevens, J. P. a. W. D., 2007. *Applied multivariate Statistics Analysis.* 6 ed. New Jersey: Pearson Pretice Hall.
Tableau Software, 2014. *Tableau Business Intelligence.* [Online]
Available at: http://www.tableausoftware.com/business-intelligence
[Accessed 2 April 2014].
The World Bank, 2014. *Education Data.* [Online]
Available at: http://data.worldbank.org/topic/education
[Accessed 6 March 2014].
The World Bank, 2014. *Expenditure per student, secondary (% of GDP per capita).* [Online]
Available at: http://data.worldbank.org/indicator/SE.XPD.SECO.PC.ZS/countries
[Accessed 22 May 2014].
The World Bank, 2014. *GDP growth (annual %).* [Online]
Available at: http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG
[Accessed 22 May 2014].
Tompkins Cortland Community College, 2012. *Measures of Shape: Skewness and Kurtosis.* [Online]
Available at: http://www.tc3.edu/instruct/sbrown/stat/shape.htm#Kurtosis
[Accessed 21 May 2014].
Torgo, L., 2011. Predicting Algae Blooms. In: V. Kumar, ed. *Data Mning with R Learning with Case Studies.* London: CRC Press, pp. 45-47.
Wiki, 2014. *Programme for International Student Assessment.* [Online]
Available at: http://en.wikipedia.org/wiki/Programme_for_International_Student_Assessment
[Accessed 16 February 2014].
Wikipedia, 2014. *Decision tree learning.* [Online]
Available at: http://en.wikipedia.org/wiki/Decision_tree_learning
[Accessed 5 May 2014].
Wikipedia, 2014. *Multivariate analysis of variance.* [Online]
Available at: http://en.wikipedia.org/wiki/Multivariate_analysis_of_variance
[Accessed 22 May 2014].
Zaiontz, C., 2014. *Testing for Normality and Symmetry.* [Online]
Available at: http://www.real-statistics.com/tests-normality-and-symmetry/
[Accessed 22 May 2014].

Appendices

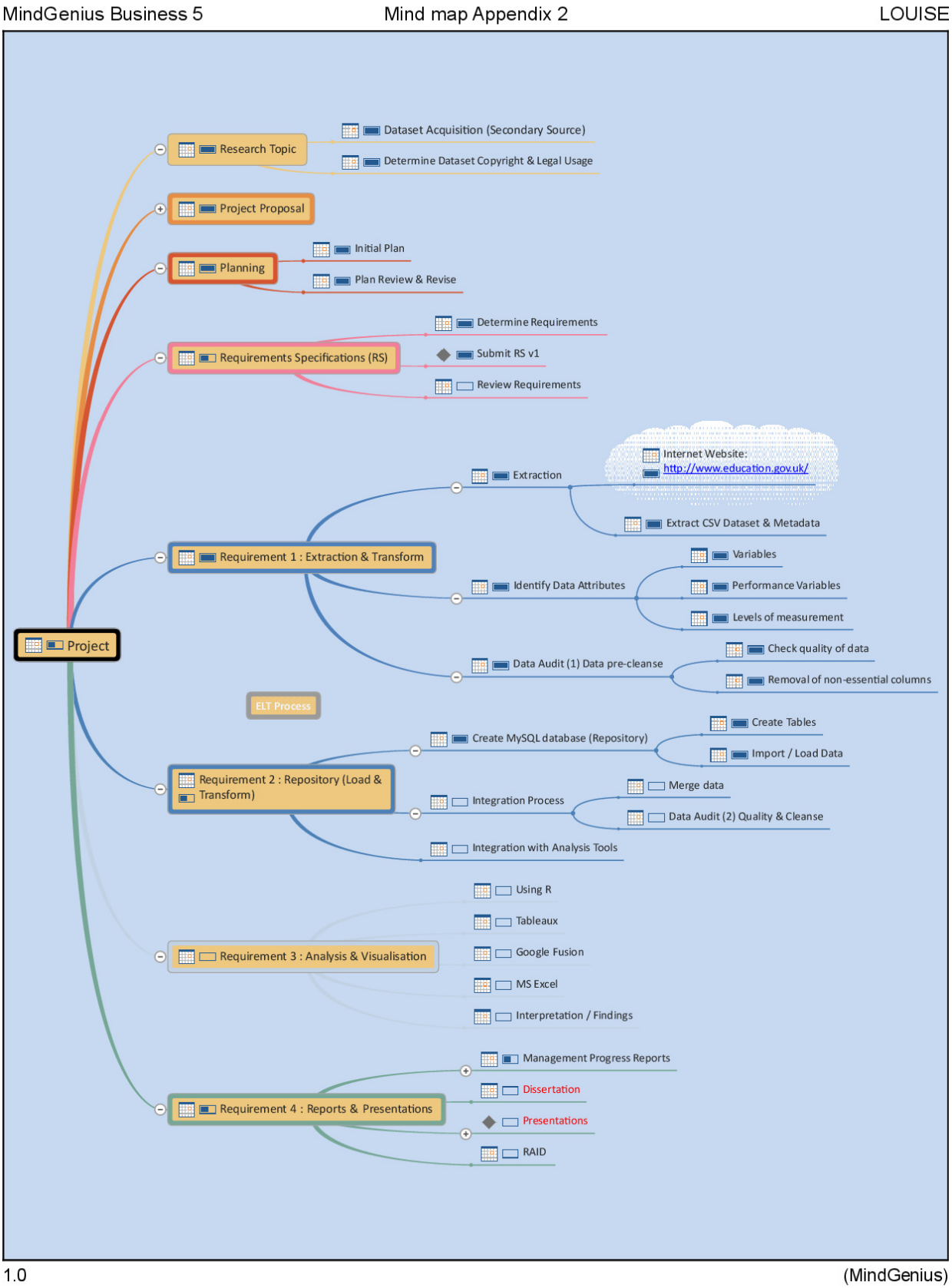
Appendix 1- 1 Final Project Plan - Gantt Chart



ID	Task Name	Duration	Start	Finish
1	**Project**	**89 days**	**03/02/2014**	**02/05/2014**
2	**Research Topic**	**11 days**	**04/02/2014**	**14/02/2014**
5	**Project Proposal**	**20 days**	**14/02/2014**	**05/03/2014**
7	**Planning**	**16 days**	**03/02/2014**	**18/02/2014**
10	**Requirements Specifications (RS)**	**43 days**	**14/02/2014**	**28/03/2014**
14	**Requirement 1 : Extraction & Transform**	**22 days**	**20/02/2014**	**13/03/2014**
25	**Requirement 2 : Repository (Load & Transform)**	**30 days**	**27/02/2014**	**28/03/2014**
26	**Create MySQL database (Repository)**	**6 days**	**27/02/2014**	**04/03/2014**
29	**Integration Process**	**16 days**	**13/03/2014**	**28/03/2014**
32	Integration with Analysis Tools	12 days	17/03/2014	28/03/2014
33	**Requirement 3 : Analysis & Visualisation**	**19 days**	**17/03/2014**	**04/04/2014**
39	**Requirement 4 : Reports & Presentations**	**89 days**	**03/02/2014**	**02/05/2014**
40	**Management Progress Reports**	**26 days**	**17/03/2014**	**11/04/2014**
44	Dissertation	85 days	03/02/2014	28/04/2014
45	**Presentations**	**12 days**	**21/04/2014**	**02/05/2014**
49	RAID	54 days	03/03/2014	25/04/2014

Created using MindGenius

Appendix 1- 2 Project Mind Map



Appendix 1- 3 Mashup Table

```
# 'data.frame':    1447 obs. of  34 variables:
# [1]  LA_NAME         : Factor w/ 146 levels "Barking and Dagenham",.. Local Authority
# [2] REGION_NAME     : Factor w/ 25 levels "East Midlands A\n",..:
# [3] SCHNAME         : Factor w/ 1429 levels "Abbeyfield School",..: School name
# [4] TOWN            : Factor w/ 524 levels "Abingdon","Accrington", Town
# [5] PCODE           : Factor w/ 1445 levels  Postal Code, required to provide geographical location
# [6] School_Type     : Factor w/ 4 levels "Community School",
# [7] TOTPUPS         : int  The pupil numbers (full time equivalent) are taken from the 2013 Annual Schools' Census.
#                      These numbers have been used to calculate per pupil expenditure amounts.
# [8]  TPUP           : int  Number of pupils at the end of Key Stage 4 / Number of pupils on roll (all ages)
# [9]  TFSMCLA        : int  Number of disadvantaged pupils
# [10]  PTFSMCLA      : num Percentage of pupils who are disadvantaged
# [11]  FSM           : Factor w/ 426 levels "1.2","1.4","  School level FSM - This is the percentage of solely and dually registered pupils eligible for Free School Meals (FSM)
# [12]  FSMBAND       : Factor w/ 3 levels "HIGH","LOW","MEDIUM" FSM Band, three broad band's used to group pupils eligible for FSM are:Low: =<20.0%, Medium:20.1%-35.0%,High: >35.0%
# [13]  EGENDER       : Factor w/ 3 levels "BOYS","GIRLS",   School gender of entry
# [14]  TNOTFSMCLA     : int Number of non-disadvantaged pupils
# [15]  PTNOTFSMCLA    : num Percentage of pupils who are not disadvantaged
# [16]  AC5EM10        : num Percentage in 2010 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
# [17]  AC5EM11        : num Percentage in 2011 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
# [18]  AC5EM12        : num Percentage in 2012 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
# [19]  AC5EM13        : num Percentage in 2013 (Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs)
# [20]  NUMBOYS        : int Total boys on roll (including part-time pupils)
# [21]  NUMGIRLS       : int Total girls on roll (including part-time pupils)
# [22]  P15END4        : num Percentage of pupils at the end of Key Stage 4 aged 15
# [23]  P14END4        : num Percentage of pupils at the end of Key Stage 4 aged 14 or under
# [24]  TEALGRP1       : int Number of Key Stage 4 pupils with English as their first language
# [25]  PTEALGRP1      : num Percentage of Key Stage 4 pupils with English as their first language
# [26]  LONDON_NONLONDON: Factor w/ 2 levels "London","Non-London": A categorical reference for London and non London schools.
# [27]  MEDIAN         : Factor w/ 1 level "Secondary with KS4": Type of school (secondary with KS4, secondary without KS4)
# [28]  T0910CAT5      : int Spend per pupil (£) available for the years 2009-10
# [29]  T1011CAT5      : int Spend per pupil (£) available for the years 2010-11
# [30]  T1112CAT5      : int Spend per pupil (£) available for the years 2011-12
# [31]  T1213CAT5      : int Spend per pupil (£) available for the years 2012-13
# [32]  URN            : int Unique Reference Number which will provide the link to the KS4 Attainment results.
# [33]  LEA            : int ...Local Authority Code, it is a numeric format, hence the need for the Regions table
# [34]  NFTYPE         : Factor w/ 4 levels "CY","FD","VA",. Type of school, abbreviation, hence the need for School Type table
```

Appendix 2 - Systems and Datasets

Design & Architecture - Creation & Setup of MySQL repository

Description: To create a Warehouse Repository for storing and querying the data.
Tools: The following tools were used Oracle SQL Developer, MySQL and MS Access.

Method:

i. Setup Database

The following are steps in the process, used to create the Warehouse Repository.

1. Software installation, Oracle SQL Developer and MySQL (MS Access already availalbe)
2. In MySQL, create a new database called ukeduc in MySql

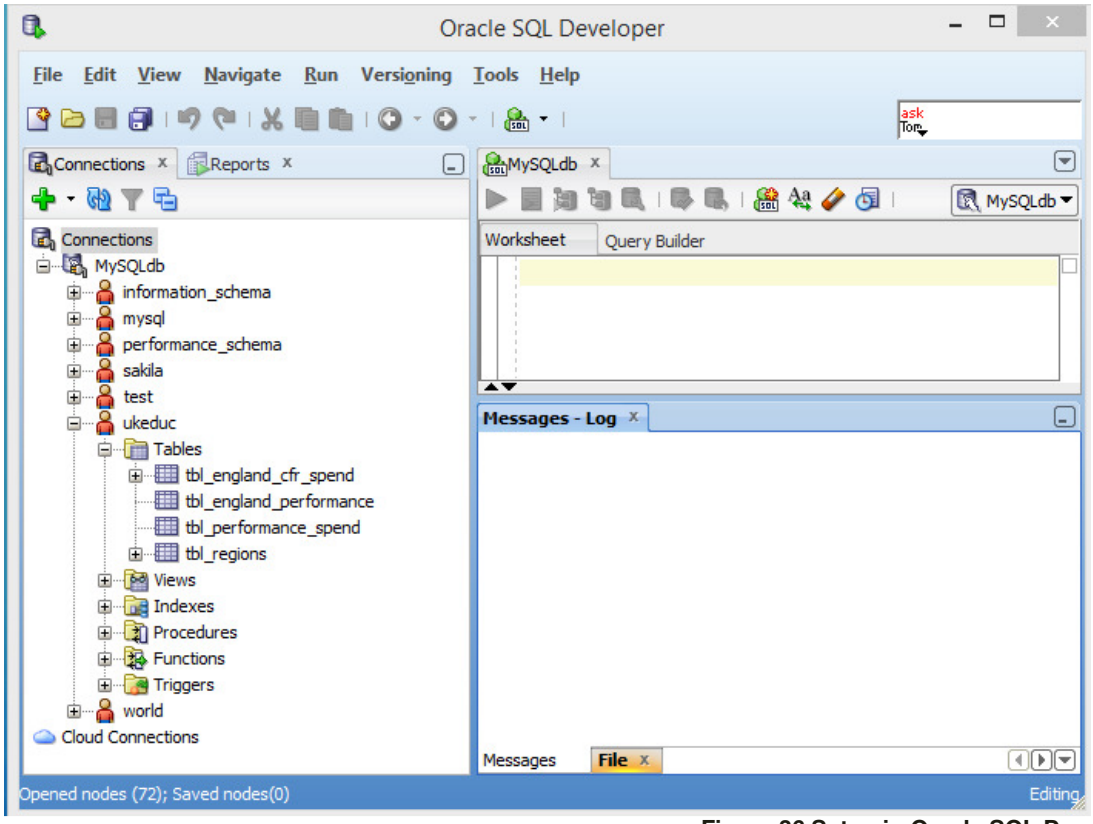


Figure 36 Setup in Oracle SQL Developer

The image Appendix 36 above illustrates the setup in Oracle SQL Developer

3. Create a connection between Oracle SQL Developer and MySQLdb.
4. The image left Appendix 2-2, illustrates the Configuration set up between Oracle SQL Developer & MySQLdb.
5. Created scripts a number of scripts to build tables.
6. Created bulk import scripts to import the datasets from CSV formats into the MySql.

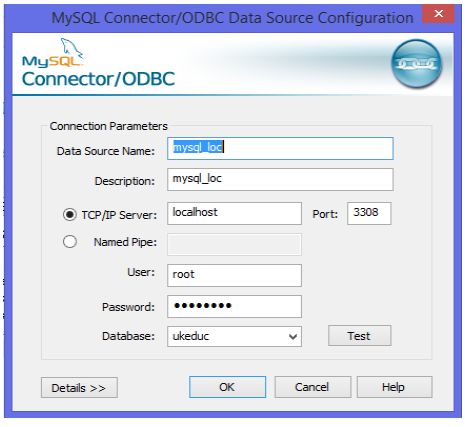


Figure 37 Configuration Setup Oracle SQL Developer & MySQL

7. This task required the metadata from the databases to create the required tables.
   - tbl_england_cfr_spend
   - tbl_england_performance
   - tbl_regions
   Appendix 2-3 illustrates the table configuration is the final MySQL

```
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sakila             |
| test               |
| ukeduc             |
| world              |
+--------------------+
7 rows in set (0.00 sec)

mysql> show tables;
+-------------------------+
| Tables_in_ukeduc        |
+-------------------------+
| tbl_england_cfr_spend   |
| tbl_england_performance |
| tbl_performance_spend   |
| tbl_regions             |
+-------------------------+
4 rows in set (0.00 sec)
```

Figure 38 Table Setup in SQL

ii. Preparation and creation of New database tables

1. Open workbook, the CSV file extracted from source (Department of Education website)
2. Copied first two rows of the CSV worksheet, opened a blank sheet.
3. Selected cell A1 in new sheet, selected Paste Special --> Transform option. This created a skeleton for the new table.
4. In figure 38 provides an illustration of the prepared sheet. It lists of all the fields in a column format.
5. With all the attributes from the sources file, in a column format it was possible to use this skeleton to create the script for the MySQL table.
6. Some amendments were required to the new sheet. Two rows were inserted above data pasted, a new column inserted and the size of the characters in the fields estimated.
7. A field size of 20 characters (VARCHAR(20)) was used for the majority of the attributes. It was necessary to increase this for some attributes. If the field is not large enough the data would be truncated when imported.
8. Below in figure-38, illustrates the additional columns (A,C,C,E) rows(1 and 2), commas, brackets, and new text. The sample data copied from row 2 of the original datasheet was removed.

	A	B	C	D	E	F	G	H	I	J
1	CREATE TABLE	tbl_england_cfr_spend								
2	(									
3		URN	VARCHAR(30)	COMMENT	Unique Reference Number					
4	,	LONDON_NON-LONDON	VARCHAR(20)	COMMENT	London/Non-London					
5	,	MEDIAN	VARCHAR(25)	COMMENT	Group to compare					
6	,	PUPILS	VARCHAR(20)	COMMENT	Number of pupils (FTE)					
7	,	FSM	VARCHAR(20)	COMMENT	Percentage of pupils eligible for Free School Meals (FSM)					
8	,	FSMBAND	VARCHAR(20)	COMMENT	Free school meals eligibility band					
9	,	GRANTFUNDING	VARCHAR(20)	COMMENT	Grant funding (£ per pupil)					
10	,	SELFGENERATEDINCOME	VARCHAR(20)	COMMENT	Self generated income (£ per pupil)					
11	,	TOTALINCOME	VARCHAR(20)	COMMENT	Total income (£ per pupil)					
12	,	TEACHINGSTAFF	VARCHAR(20)	COMMENT	Teaching staff (£ per pupil)					

Figure 39 Attributes - Transposed in MS Excel

9. Once complete, it was saved as a .CSV format
10. The .CSV file was opened in Notepad, where further editing was carried out.
11. The edited file was saved as a .SQL format, see next illustration Figure 40

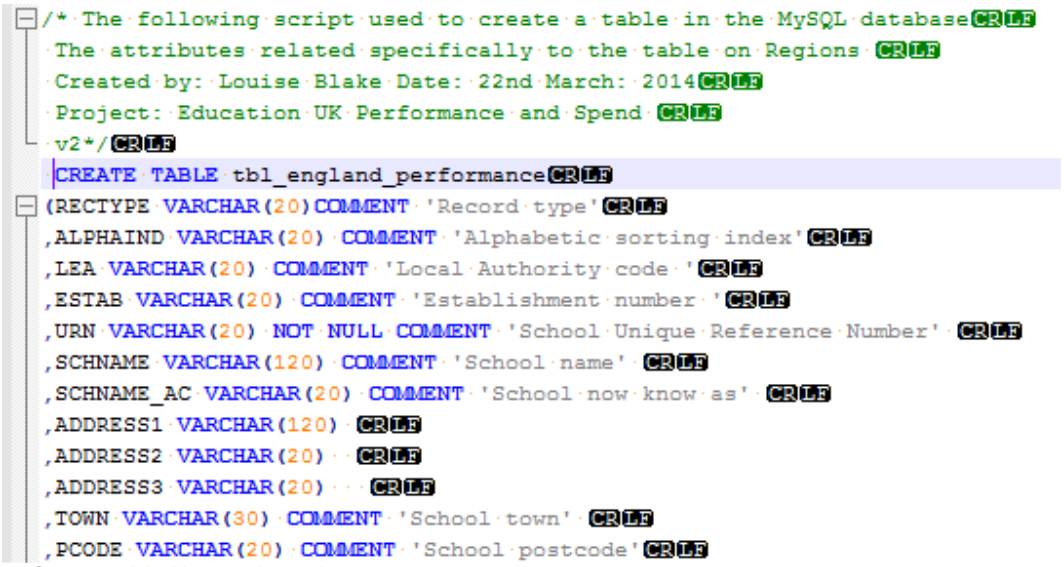


Figure 40 Create table Notepad version

12. Below is sample code used to create the Regions table.
13. This code was run in Oracle SQL Developer and created the new table.

```
CREATE TABLE tbl_regions (
 LEA VARCHAR(20) COMMENT 'Local Authority code'
,  LA_NAME VARCHAR(50) COMMENT 'Local Authority Name'
,  REGION VARCHAR(20) COMMENT 'Region Code '
,  REGION_NAME VARCHAR(50) COMMENT 'Region Name'
)
```

14. The above process was carried out for three tables.

Additional Tables set up in MS Access:
Two tables were created directly in MS Access. One of these was the table containing the School Type details. The reason it was not created with the rest of the tables is due to the decision to include the full description into the Mashed up table at a later stage in the process. The abbreviations for school type were already there, however for completeness and ease of use the full name was included. The second table is the list of Regions Geo codes which were sourced at a later stage and again this table was created in MS Access.
It is acknowledged that to keep all the data tables in the same place would have been a better option.

Recommendation: Future work recommendation, would be to ensure all data tables are kept in the one Warehouse Repository. This would create a more robust warehouse repository.

iii. Pre-processing Checks

Description: Prior to the batch import from CSV to MySQL, the data required additional pre-processing and checking.

Check for unwanted characters
The purpose of this check was to ensure a successful batch import. An example of undesirable characters are commas, they could split the data values when importing, hence they needed to be removed.

Method:

How to remove unwanted characters

1. A search was carry out for frequently know characters such as , ' % £ ^ * &.
2. The related columns identified, the unwanted characters were assessed as to whether removal or leave in.
3. Completed, a simple Find and Replace in MS Excel to manage this issue.

Findings in this dataset

In this dataset the Address fields, held a number of commas. This did cause some issues when importing into MySQL. Once all the commas were removed, the importing process completed successfully.

.

Deletion of Column or Rows

No columns were removed from the worksheets. During the Mashup process the required Target attributes were selected. Two to three rows at the end of each sheet needed to be removed as they contained partial summary information.

iv. Bulk Import Script

Description: The next part of the set of the Warehouse repository was to populated it with the data. This was done by using a batch import script, to take the extracted and pre-cleansed CSV dataset and upload it to the MySQL database.

Method: Using the script below as an example of what was run in Oracle SQL Developer.

Sample of code:

```
LOAD DATA LOCAL INFILE "C:/myscripts/List_Of_Regions_LocalAuthorities.csv"
REPLACE INTO TABLE tbl_regions FIELDS TERMINATED BY ',' OPTIONALLY
ENCLOSED BY "" LINES TERMINATED BY '\n'
```

Design & Architecture - MS Access database

Prior to setting up the MS Access database a number of trials scripts were created in R to

Sample of code:

```
library(RODBC)
conn <- odbcConnect(dsn="projsql", uid="root", pwd="password")
#sqlTables(channel)
queryResult <- sqlQuery(conn, "SELECT * FROM tbl_england_cfr_spend")
odbcClose(conn)
dim(queryResult)
```

Following several unsuccessful attempts and time constraints the alternative option to of importing pre-cleansed and 'mashed' data in a CSV format into R.

The following reasons determined the use of MS Access

- Import link to MySQL the warehouse repository available,
- Export options available from MS Access, also readily available.
- Flexible queries for pre-checking data for missing, nulls and other data inconsistencies
- Easy joining and aggregation queries.

The alternative of using more MySQL and R integration was deferred and by the end of the project remained unused.

Setting up of MS Access Database

1. A new MS Access database called 'ukschools' was created.
2. An ODBC driver was set up to allow a connection between the MS Access database and the MySQL database warehouse repository called 'ukeduc'
3. Once established an External Data link was established for each table.

Implementation - Data Cleanse and Verification Checks

Description: This was a more extensive cleanse and verification check process of the data. This process looked at the quality of the data under a number of criteria, duplicate record, missing data, unwanted characters and data consistency.

Tools: The application tools used in this process included MS Excel and MS Access.

Data Quality

i. Duplicate Records

Data verification was a requirement. One of the first verifications completed on the data was to assess for duplicates. This was an uncomplicated task as each record had a Unique Reference Number (URN) assigned. No duplicates needed to be removed. The Unique Reference Number also provided the relationship between the datasets, which was a requirement during the data merge process.

ii. Individual attributes check - unwanted characters and missing data

Further investigation into individual variables were completed. The results from this verification check, established that some of the attributes extracted did have some missing data and a number of attributes had unwanted characters and references such as SUP (Supplementary). The records affected by missing values were investigated. The decision was to delete these data records described from the study. It was determined to be the most effective option and the affect on the overall study by keeping them in would be greater on the analysis models. Overall numbers were small. Aggregation of the Performance Attained and Expenditure per pupil values was a reason for removing them from specific attributes.

iii. Missing Data

Specific columns were checked for missing values. In total, 25 entries with na reference in one of the percentage attributes relating to performance. The decision was to not include these records in the study.

iv. Inconsistent Data attributes

This check was carried out for each of the Target attributes. One of the attributes relating to Student Age Range had issues. The format of the data entry was not consistent, therefore this field was not used in the data analysis.

Implementation - Transformation

Description: In the transformation process data reduction occurred. This used the data tables joined and filtered using queries in MS Access. This process required a significant amount of queries set up in MS Access. The initial stage of the transformation setup was to join the tables and ensure the data available comprised of Secondary schools with KS4 level details.

Tools: MS Access was used as the application to create queries and transform the data.

i. Process - Joining of tables

When the tables were combined, the number of Secondary schools with pupils sitting the KS4 totalled 1476. The CFR table tells us the schools that have KS4 (attribute reference is called Medium). The total schools in this category is 1476, However following attributes data checks a small number of records were removed bringing the final data set to 1447.

ii. Process - Data Reduction

As part of this process in MS Access, the data was queried to reduce the types of schools in the study to the relevant Secondary with KS4. The query was created in MS Access using the attribute in the Expenditure table coded as RECTYPE with a value equal to 1. This attribute indicates record type and the value 1 = mainstream schools.

School Types	
	Count
Primary	15644
Secondary with KS4	1476
Secondary without KS4	123
Special	871

Figure 41 Table of School Types pre-cleanse numbers

The final dataset size had a total of 1447 record.

Appendix 3 - Testing and Evaluation

Test 1 Data Summaries

Description

Mean: is the average, measure of central tendency,
Median: is the central value, measure of central tendency,
Range: is the highest (maximum) and lowest values (minimum),
Variance: is a measured of how spread out a distribution is, average of the squared differences from the mean,
Standard Deviation: used to measure how spread out the numbers are square roots of variance,
3rd Quartile-1st Quartile: By observing the difference between the mean and median as well as the inter-quartile range (3rd Quartile-1st Quartile) we can get an idea of the shewness of the distribution and also its spread.

Method

R application, summary() function

Results

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
English First Language (%)	0.01	0.83	0.95	0.86	0.99	1.0
Disadvantaged pupils (%)	0.02	0.16	0.26	0.29	0.39	0.99
FSM Eligibility (%)	0.01	0.09	0.15	0.18	0.25	0.73

Abbreviations on the table include min. - minimum, qu. - quartile, max- maximum

Table 5 Summary of Numeric Attributes

Expenditure (£)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2009/10	1530	4802	5295	5560	5978	18809
2010/11	4104	4938	5462	5749	6213	17500
2011/12	3565	5007	5512	5797	6246	17147
2012/13	3941	5100	5651	5972	6406	20080
4 year average	4135	4982	5489	5766	6220	18160

Abbreviations on the table include min. - minimum, qu. - quartile, max- maximum

Table 6 Summary Expenditure per pupil (%) for each year

Performance (%)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2009-10	0.21	0.44	0.52	0.54	0.62	1
2010-11	0.21	0.47	0.56	0.57	0.65	1
2011-12	0.19	0.48	0.56	0.57	0.65	1
2012-13	0.06	0.5	0.59	0.59	0.68	1
4 year average	0.21	0.48	0.56	0.57	0.64	1.0

Abbreviations on the table include min. - minimum, qu. - quartile, max- maximum

Table 7 Summary Performance Attained per pupil (%) for each year

Interpretation

Over the four years the annual mean for both Expenditure per pupil and Performance Attained showed an increasing trend.
In the case of Performance the mean, 54% in the school year 2009/10 increasing to 59% by 2012/03. This was an increase of 5% in the overall performance attainment by students at GCSE level. The Expenditure also increased over the four year period from £5560 in 2009/10 to £5972 in 2012/13 an increase of £412 per student.

The following are graphical representations for the factors plotted against Target variables Average Performance Attained and Average Expenditure Levels.

In figure 42 the Factors plotted against the Average Performance

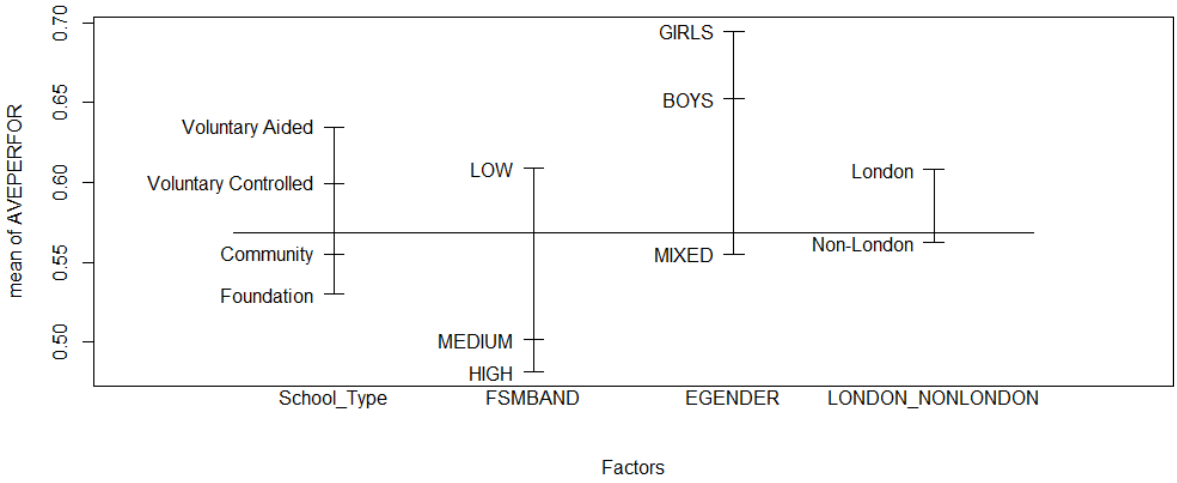


Figure 42 Factors plotted against the Average Performance

In figure 43 the factors plotted are against Average Expenditure

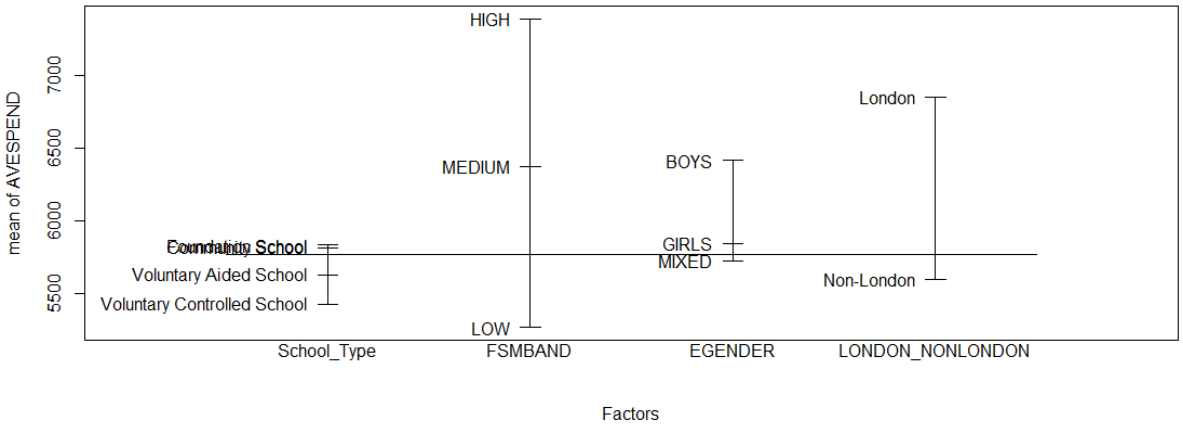


Figure 43 Factors plotted against Average Expenditure

Test 2 Data patterns - Categorical and Data

Description

The purpose of this analysis was to discover what trends and patterns already existed in the data. This analysis included categorical data which was already available in the data, Gender, School Type, FSM Band and London Non London,

Method: Tableau, was the application of choice to discover trends and patterns.

Results

The results show a series of graphs analysing the data.

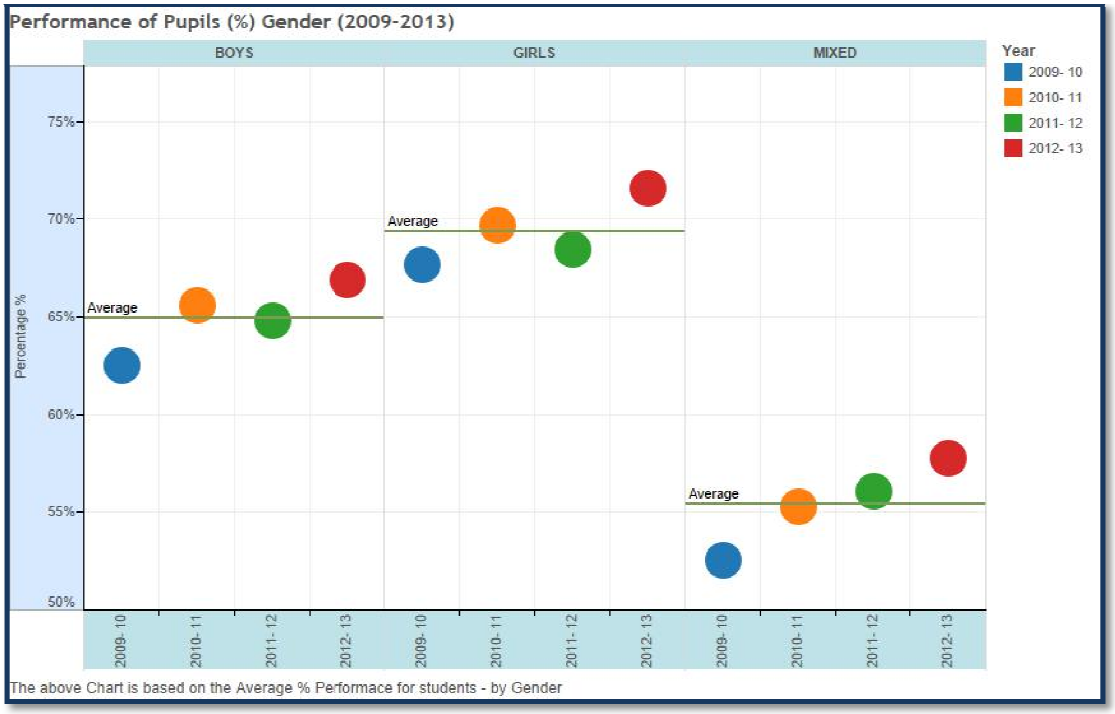


Figure 44 Performance by Year, School Gender

The above chart presents a breakdown of the Performance Attained by School Gender.

Interpretation

In figure 44, over the course of the four years of performance data available the all Girls schools outperformed the all Boys and mixed schools.

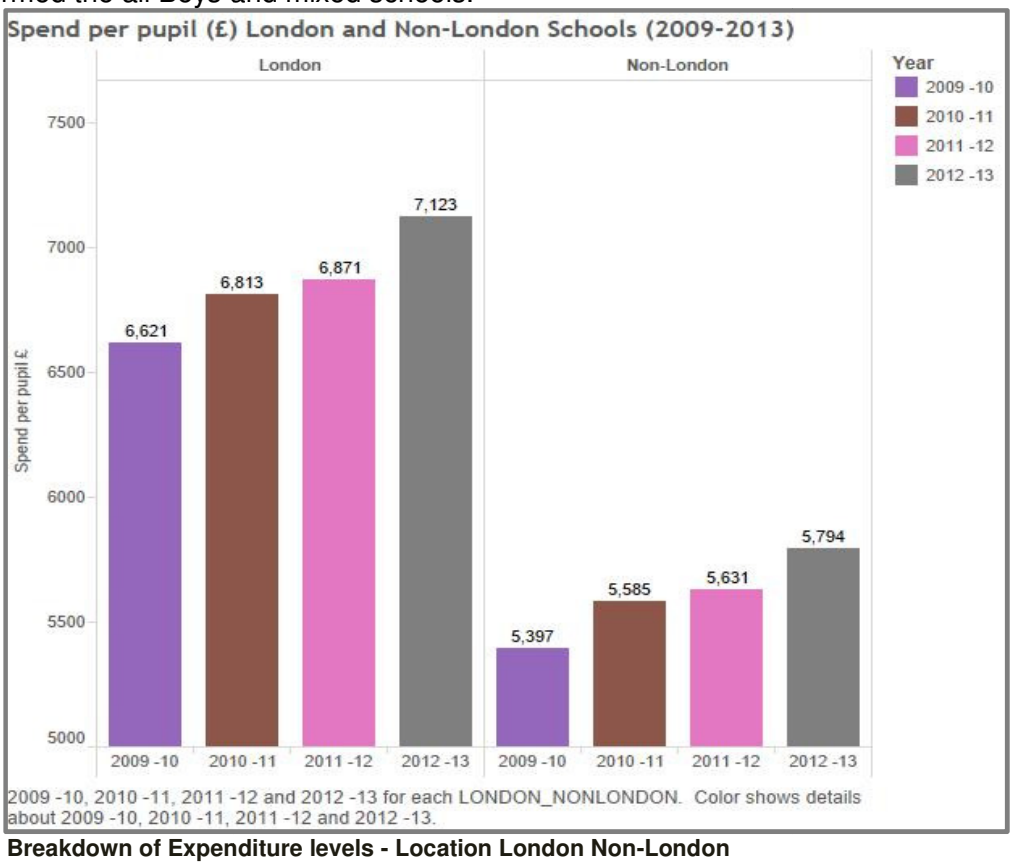


Figure 45 Breakdown of Expenditure levels - Location London Non-London

Interpretation
In figure 45, over the course of the four years expenditure levels for London schools was notably higher compared to Schools outside of London.

Tests for Normality and Symmetry

Test 3 Visualisation Histograms & Density plots
Description: The purpose of both the Histograms and density plots was to graph the data for visualisation. The frequency of the data attributes was plotted using the histogram. In addition to the density graphs, both the Kernel and normal density plots were used. Density plots can have negative values, although we have none in the data studied. The histogram was used to show the distributions of independent and dependent variables. The purpose of overlaying the plots with the Normal parametric and non-parametric Kernel density plots, assists in whether the attributes are normally distributed or is another distribution is suitable.

Method: R was used to create the Histograms using the hist() function, Density using the density() function, lines() function is used to add the Kernel density plot .The curve () function is used for the normal density plot with the mean and standard deviation equal to that of the data explored.

Results: The plots shown provided an indication about features of the data. This included the general shape, symmetry or skewness. The histograms and density plot below are from some of the individual attributes, Figures 46 to 52.

Interpretation
The charts below give an indication of non-symmetry and some skewness, hence further tests were required. In this scenario, it was decided that the attributes were not normally distributed therefore another distribution would be more suitable.

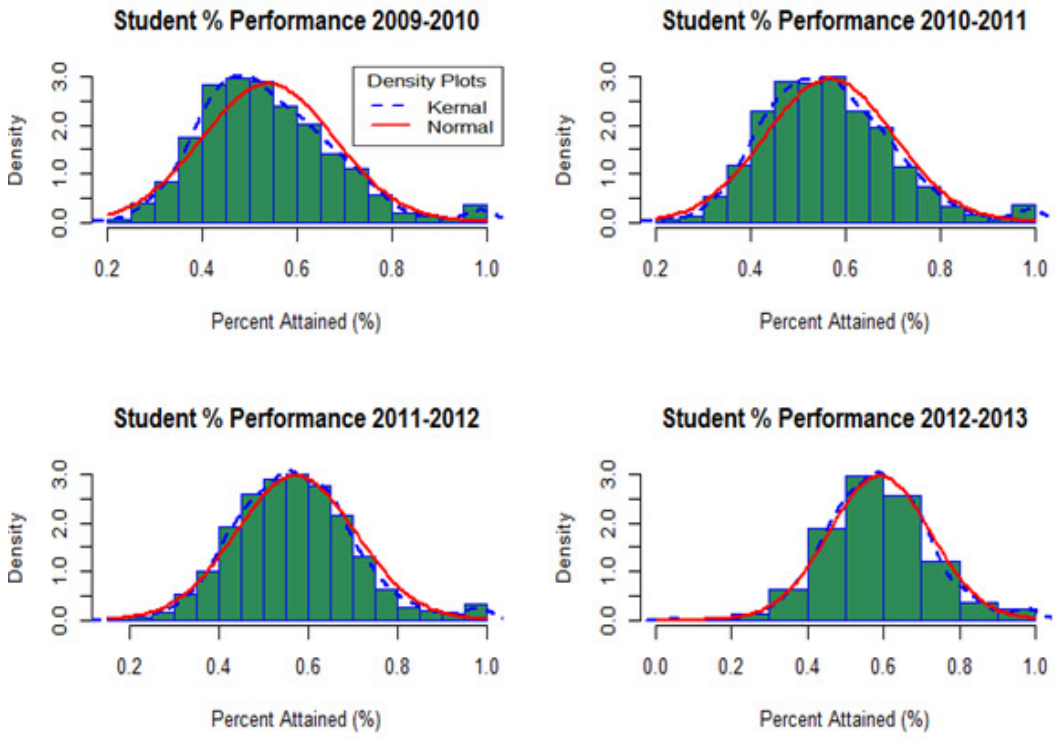


Figure 46 Density plots Performance % 4 years

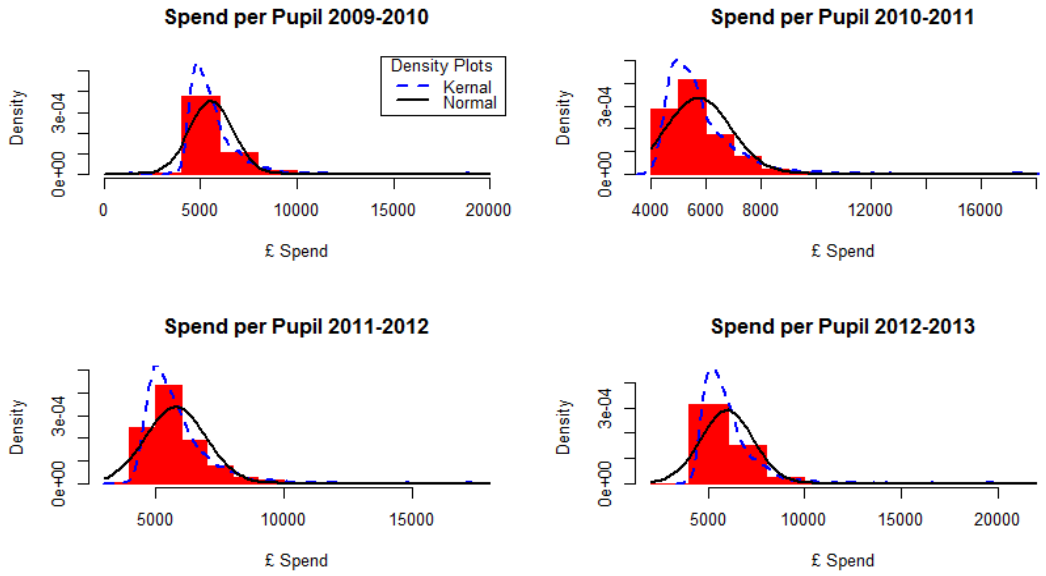


Figure 47 Density plots Expenditure £ 4 years

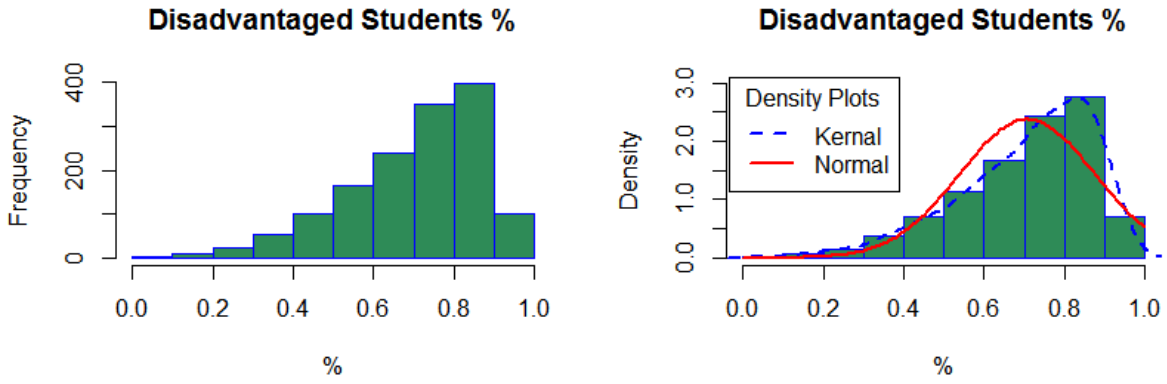


Figure 48 Histogram & Density Plots - Disadvantaged Students %

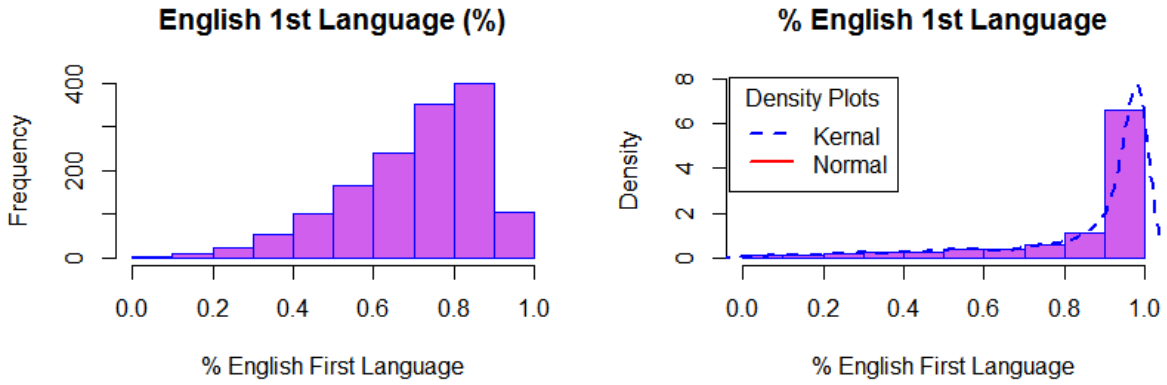


Figure 49 Histogram & Density Plots - English First Language %

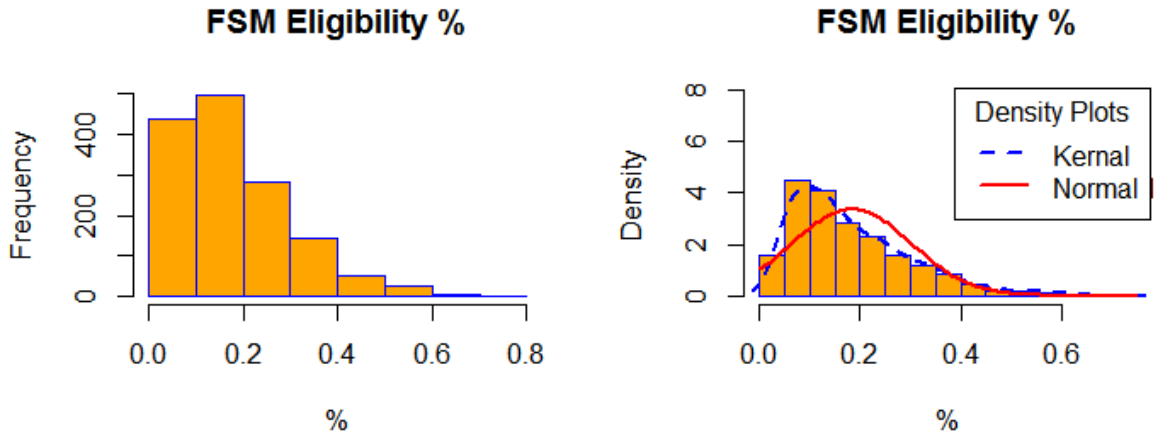


Figure 50 Histogram & Density Plots - FSM Eligibility %

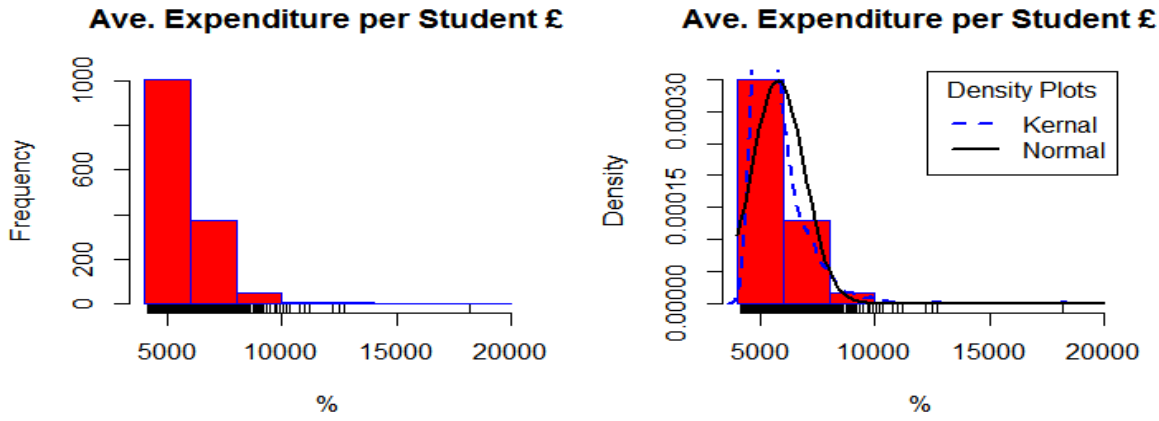


Figure 51 Histogram & Density Plots - Mean Expenditure per student £

Both Figures 51 and 52 are enriched plots using rug() and jitter() in R and library(car)

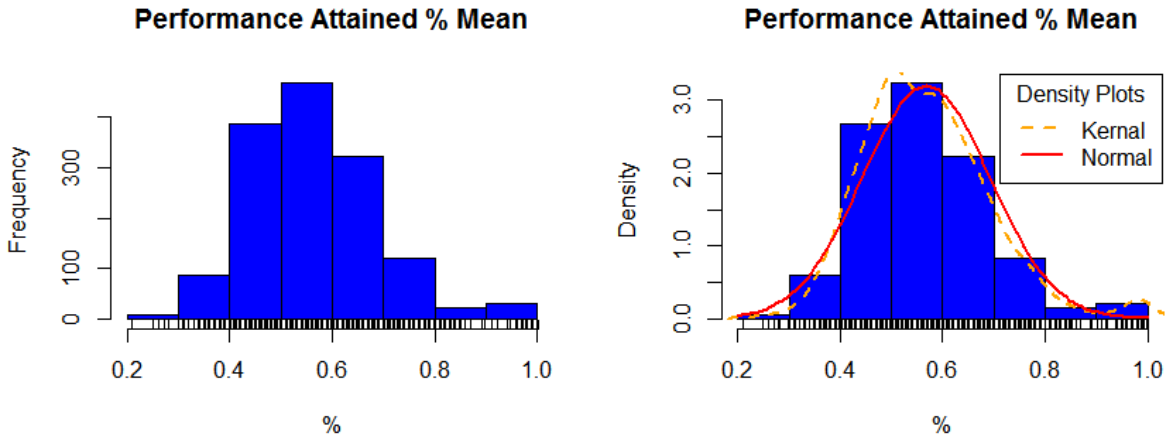


Figure 52 Histogram & Density Plots - Mean Performance Attained %

Test 4 Visualisation - boxplots

Description

The purpose of this check was to look and see if that were "no significant outliers" in the data that might have a negative influence on the results. This check was carried out using the box plot. Outliers are unusual, irregular observations; data points that do not appear to follow the characteristic distribution of the rest of the data. However defining an outlier is subjective, and a decision on identifying each of the outliers would require individual investigation.

Method In R the data was plotted using the boxplot() function.

Results

The following boxplots were carried out on the uni-variant numeric attributes. The boxplots shown below in Figure 53 are the individual years covering the four years between 2009/10 to 2012/13 for Performance Attained (£) and Expenditure per student (£) spend. All of the boxplots indicated outliers are present in each year and each attribute.

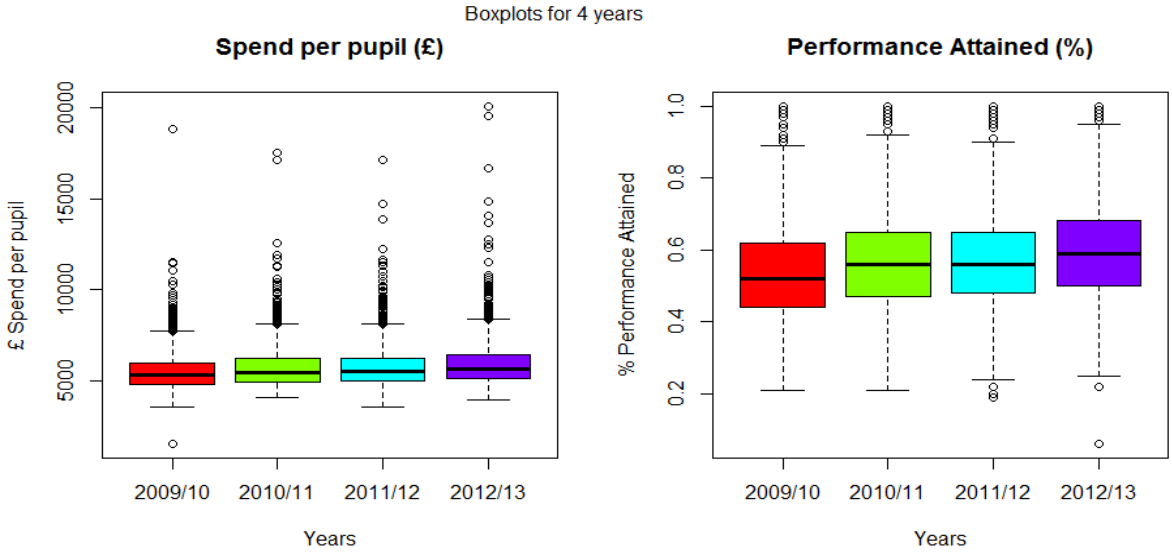


Figure 53 Boxplots for Expenditure and Performance over 4 years

In Figure 54 the boxplot is an enriched boxplot for the average of the four years Expenditure. As expected outliers were present.

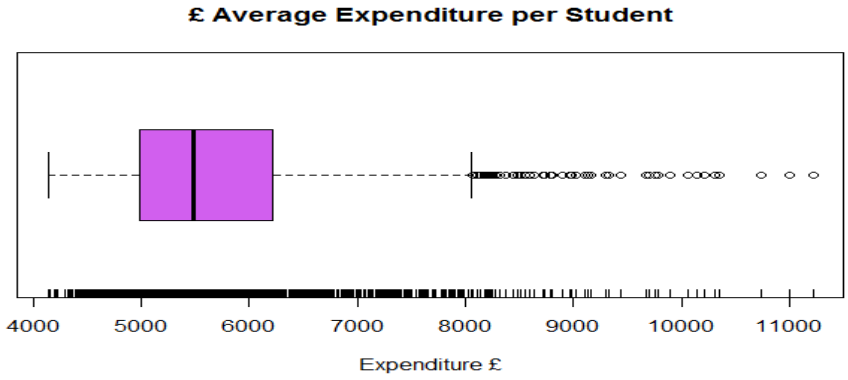


Figure 54 Boxplots of Average Expenditure (4 years mean)

Min.	1st Qu.	Median	Mean 3rd	3rd Qu.	Max.
4135	4982	5489	5766	6220	18160

Table 8 Summary of Average Expenditure (4 years mean)

Interpretation
Each of the target numeric attributes showed a number of outliers. In the boxplot for Average Expenditure and the summary table statistics, there was information on key properties available from the boxplot. The boxplots were a rich source of information. As an example of the information from the boxplots the following is a summarisation of the key properties of the attribute Average Expenditure. In this boxplot, it shows an inner purple box with the limits for the 1st and 3rd quartiles of the variable, which are approximately £4135 and £6220 for the variable. The box has a line which represents the median value of the variable, which is £5489. The small horizontal dash line to the right of the box are the largest observations "that is less than or equal to the 3rd quartile (plus 1.5 the inter quartile range). The small dashed line to the left of the box shows the smallest observations that are greater than or equal to the 1st quartile (minus 1.5 the inner quartile range). The circles right "represent observations that are extremely high compared to all the others, usually considered outliers" (Torgo, 2011). This box plot gives information on the central value and spread of the variable, and also outliers.

Test 5 Skewness and Kurtosis

Description
The purpose of the statistic test for skewness and kurtosis was to confirm and provide additional information on the histograms and density plots in Test 1.
Method R function skewness() and kurtosis(). and library(e1071)
Results

Attribute	Skewness	Kurtosis
English as first language	-1.98	3.21
Disadvantaged pupils	0.91	0.53
Performance 4-Year Mean	0.74	1.11
Expenditure 4-Year Mean	2.43	12.90
Free School Meal Eligibility	1.12	1.16

Table 9 Skewness & Kurtosis Results

Interpretation

	Skewness	
English as first language	-1.98	Left skewed, highly skewed
Disadvantaged pupils	0.91	Right skewed, moderately skewed
Performance 4-Year Mean	0.74	Right skewed, moderately skewed
Expenditure 4-Year Mean	2.43	Right skewed, highly skewed
Free School Meal Eligibility	1.12	Right skewed, highly skewed

	Kurtosis	
English as first language	3.21	Central peak, close to a Normal distribution (Mesokurtic)
Disadvantaged pupils	0.53	Central peak is lower and broader compared to a normal distribution (platykurtic)
Performance 4-Year Mean	1.11	Central peak is lower and broader compared to a normal distribution (platykurtic)
Expenditure 4-Year Mean	12.9	Central peak is higher and sharper compared to a normal distribution (leptokurtic)
Free School Meal Eligibility	1.16	Central peak is lower and broader compared to a normal distribution (platykurtic)

Bi-variant and Multivariate Analysis

Test 6 Shapiro-Wilks Normality test

Description
As a supplement to the graphical assessment of normality, a normality tests using the Shapiro-Wilks normality test was carried out. The Shapiro-Wilks test is based on the correlation between the data and the corresponding normal scores.

Method
The Shapiro-Wilks normality test was carried out in R using the shapiro.test() function performs normality test of a data set with hypothesis that it's normally distributed.

Results

Attribute	w	p-value
English as first language	0.6851	< 2.2e-16
Disadvantaged pupils	0.9372	< 2.2e-16
Performance 4-Year Mean	0.9716	3.124E-16
Expenditure 4-Year Mean	0.8281	< 2.2e-16
Free School Meal Eligibility	0.9131	< 2.2e-16

Table 10 Shapiro-Wilks Test Results

Shapiro-Wilks normality test Expenditure (£)				
Year	2009/10	2010/11	2011/12	2012/13
W	0.8419	0.8158	0.8263	0.7725
p-value	<2.20E-16	<2.20E-16	<2.20E-16	<2.20E-16

Table 11 Shapiro-Wilks Results Expenditure over 4 Years

Shapiro-Wilks normality test Performance (%)				
Year	2009/10	2010/11	2011/12	2012/13
W	0.9692	0.9764	0.9855	0.9909
p-value	<2.20E-16	=1.14E-14	=7.02E-11	=8.76E-08

Table 12 Shapiro-Wilks Test Results (b)

Interpretation
- Since the p-value is smaller than 0.05 for each of the attributes, English as first language(%), Disadvantaged pupils(%), Performance 4-Year Mean(%), Expenditure 4-Year Mean(£) and Free School Meal Eligibility(%), it's rejected that attributes are normally distributed.
- Since the p-value is smaller than 0.05, it's rejected that the Expenditure(£) attributes for each year are normally distributed.
- Since the p-value is smaller than 0.05, it's rejected that the Performance(%) attributes for each year are normally distributed.
- Overall the eight attributes relating to Expenditure per pupil(£) and Performance Attained(%), each have a p-value smaller than 0.05, it's rejected that the attributes are normally distributed.

Test 7 Regression

Description: In order to find out more about the relationship between out variables two regression models were built in r. This test was used to find out what significant levels the dependant and independent attributes to each other and the R-squared value.

Method

The functions used in are to build the model comprised of lm() and summary(). The information provided from the results of these functions related to the coefficient responses of the interaction of the dependant to the independent variable.

Results

The results are tabulated below relate to the regression model built for finding out more about our independent and dependant variables.

Model 1 Output

Model: Ave. Performance and Ave. Expenditure ~ Disadvantaged. pupils + FSM. Eligibility + English1st

Coefficients	(Intercept)	Disadvantaged	FSM Eligibility	English 1st Language
Ave. Performance	0.9184	-0.2301	-0.3558	-0.3052
Ave. Expenditure	4523.58	-81.35	3313.96	2405.26

Residuals

Min	1Q	Median	3Q	Max
-0.35877	-0.06457	-0.00763	0.05737	0.33816

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.91839	0.01681	54.644	< 2e-16	***
Disadvantaged.pupils	-0.23013	0.01487	-15.472	< 2e-16	***
FSM.Eligibility	-0.35579	0.06784	-5.245	1.80E-07	***
English1st	-0.30515	0.04915	-6.208	7.00E-10	***

Residual standard error: 0.09718 on 1437 degrees of freedom
Multiple R-squared: 0.3757, Adjusted R-squared: 0.3744
F-statistic: 288.2 on 3 and 1437 DF, p-value: < 2.2e-16

Table 13 Regression Model 1 Results

Model 2 output

Ave. Expenditure ~ Disadvantaged. pupils + FSM. Eligibility + English1st

Residuals

Min	1Q	Median	3Q	Max
-1883	-438.9	-133.4	240.2	13462.9

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4523.58	142.86	31.666	< 2e-16	***
Disadvantaged.pupils	-81.35	126.43	-0.643	0.52	
FSM.Eligibility	3313.96	576.62	5.747	1.11E-08	***
English1st	2405.26	417.79	5.757	1.04E-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 826 on 1437 degrees of freedom
Multiple R-squared: 0.4822, Adjusted R-squared: 0.4811
F-statistic: 446.1 on 3 and 1437 DF, p-value: < 2.2e-16

Table 14 Regression Model 2 output

Test 8 Residual Plots

Method: The residual plots were created using Res() in r for the two models created for the regression model.

Results

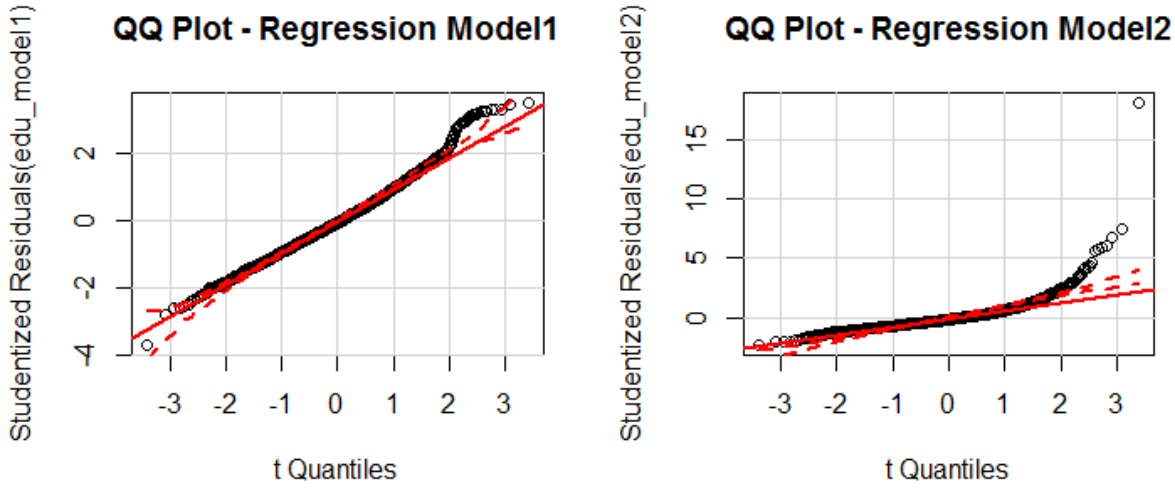


Figure 55 Residual QQ plot from Regression Models

Test 8 MANOVA

Description: MANOVA is the statistical test procedure for comparing multivariate (population) means of several groups. It differs from the ANOVA test in that, "it uses the variance-covariance between variables in testing the statistical significance of the mean differences" (Wikipedia, 2014).

Method

The functions were used to build the ANOVA models comprised of lm() and summary(). The information provided from the results of these functions related to the coefficient responses of the interaction of the dependant to the independent variable.

Results

The results are tabulated below relate to the regression model built for finding out more about our independent and dependant variables.

Analysis of Variance Table Response: Ave.Performance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Disadvantaged.pupils	1	0.0285	0.0285	3.0196	0.08248	.
FSM.Eligibility	1	7.7739	7.7739	823.1214	< 2.2e-16	***
English1st	1	0.364	0.364	38.5422	7.00E-10	***
Residuals	1437	13.5717	0.0094			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 15 ANOVA, Response to Ave Performance

Analysis of Variance Table Response: Ave.Expenditure

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Disadvantaged.pupils	1	320017065	320017065	469.008	< 2.2e-16	***
FSM.Eligibility	1	570448282	570448282	836.033	< 2.2e-16	***
English1st	1	22615033	22615033	33.144	1.05E-08	***
Residuals	1437	980504125	682327			

Table 16 ANOVA, Response to Expenditure

Interpretation

In the results of the first model Table 15 the response coefficient, p-value was less than 0.05 for two variables, FSM Eligibility and English First Language in the ANOVA test. At a significance level of 0.05, the two variables indicate a significant level. The p-value of the response from Disadvantage was 0.08 > 0.05. However if the significance was set at a higher significance level this would also be indicated as a significant response to the coefficient.
In the second model all three responses indicated a significance level. present as the p-values were less than 0.05.

Test 9 Mann Whitney Test

Description

The Mann Whitney test is a nonparametric tests of Group Differences. The null hypothesis is that the data of y and x are identical populations. To test the hypothesis, to compare the independent samples the wilcox.test() function is applied. With a significance level of 0.05 if p>0.05 we accept the Null hypothesis, if p<0.05 we reject the null hypothesis and accept the alternative that the tested attributes are non identical populations.

H0: The two populations are equal
H1: The two populations are not equal

Method R function wilcox.test(), wilcox.test(y, x) where y and x are numeric

Results

y	x	W	p-value
Ave. Performance (%)	Ave. Expenditure (£)	0	< 2.2e-16
Ave. Performance (%)	English First Language(%)	266324	< 2.2e-16
Ave. Performance (%)	Disadvantaged pupils (%)	1863860	< 2.2e-16
Ave. Performance (%)	FSM Eligibility (%)	2038133	< 2.2e-16
Ave. Expenditure (£)	English First Language	2076481	< 2.2e-16
Ave. Expenditure (£)	Disadvantaged pupils (%)	2076481	< 2.2e-16
Ave. Expenditure (£)	FSM Eligibility (%)	2076481	< 2.2e-16
English First Language (%)	FSM Eligibility (%)	2027064	< 2.2e-16
Disadvantaged pupils (%)	FSM Eligibility (%)	1468694	< 2.2e-16
English First Language (%)	Disadvantaged pupils (%)	1976241	< 2.2e-16

Table 17 Mann Wilcox Test results

Performance v Spend

	W	p-value
2009/10	0	< 2.2e-16
2010/11	0	< 2.2e-16
2011/12	0	< 2.2e-16
2012/13	0	< 2.2e-16

Table 18 Mann WIlcox Test results (b)

Interpretation
The p-value is less than 0.05 in all the Mann Whitney tests, then we can accept the alternative hypothesis Ha of statistical equality of the means of two groups. At a significance level of 0.05, we conclude that the data y and x in our study are non identical populations

Test 10 Visuals - Pairs Scatter plots

Description
There are many types of scatter plots in R. The simple scatter plot uses the function plot() and assists in visualising the relationship between two continuous variables. The scatter plot of the variables supports in checking for linearity. If the relationship of the two variables are not linear, the correlation coefficient should not be calculated. In the scatter plot chosen the axis the variables are on does not matter. Normally there is an independent (or explanatory) variable which is plotted on the x-axis (horizontally) and then the dependent (or response) variable is plotted on the y-axis (vertically).

Method
The pairs() function in R this produced plots that are helpful in exploring the data

Results

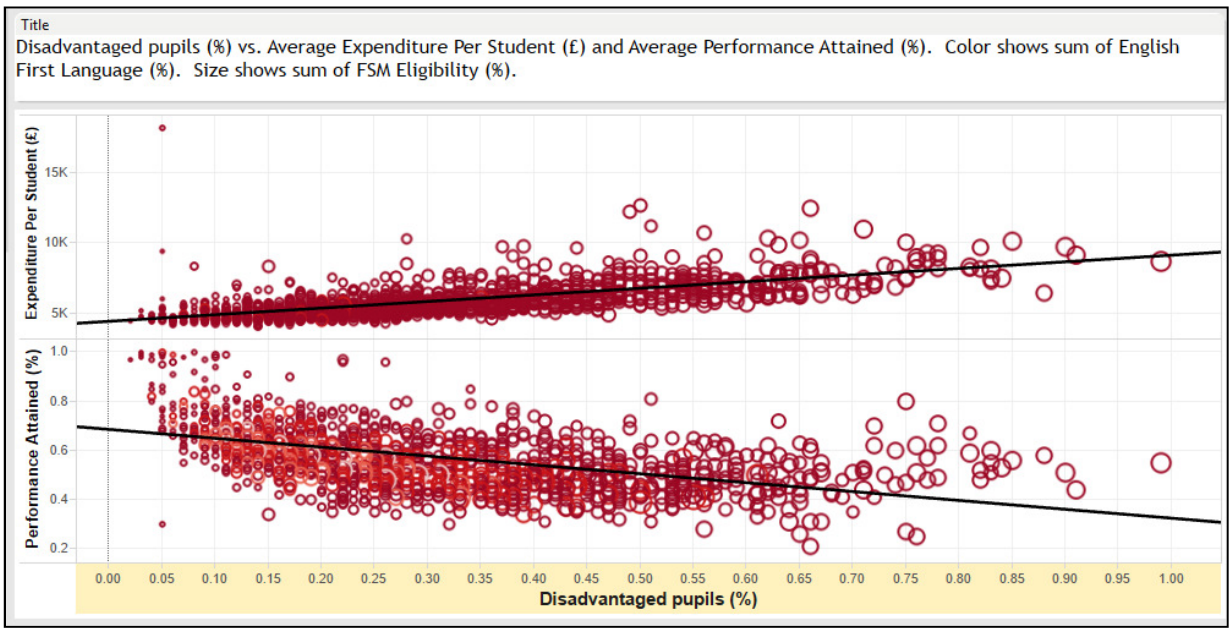


Figure 56 Two Scatterplots - Disadvantage vs Expenditure and Performance

The following are enhanced scatter plots showing bi-variant variables, this provides a closer inspection compared to the pairs plot matrix show in figure 56.

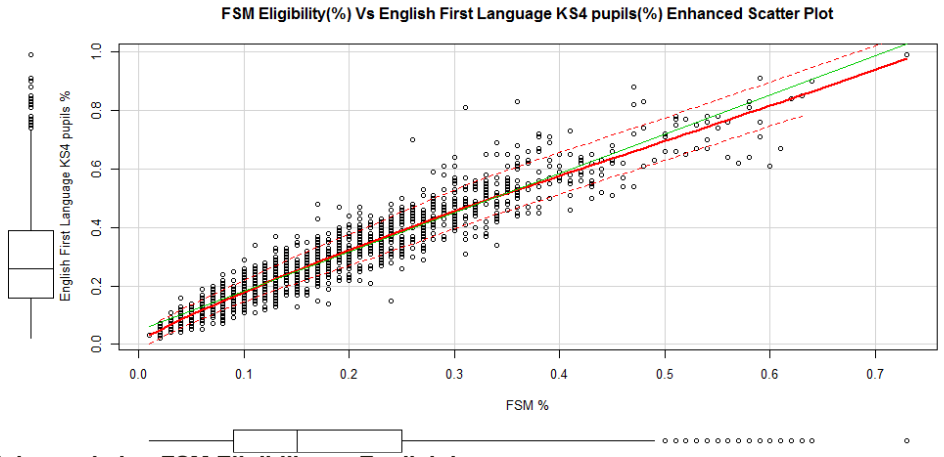


Figure 57 Enhanced plot, FSM Eligibility vs English language

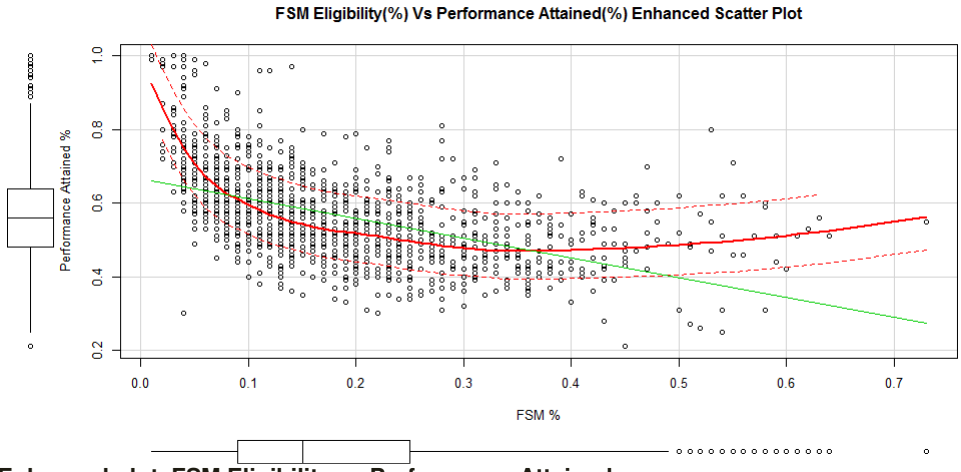


Figure 58 Enhanced plot, FSM Eligibility vs Performance Attained

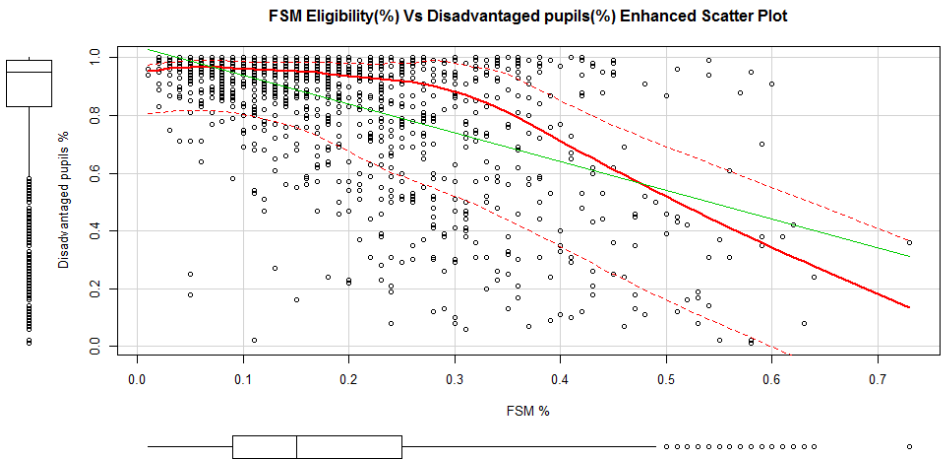


Figure 59 Enhanced plot, FSM Eligibility v Disadvantaged

Interpretation

Discussion in Dissertation

Test 11 Correlation & Covariance

Description

In Correlation the relationship between two or more variables is measured. When two things are correlated, it means that they vary together. The correlation can be either a positive or a negative correlation and the values range between the range -1 and 1. A common statistic for calculating bi-variate correlations is Pearson r. The closer the correlation is to 1 the stronger the relationship, between the two variables. If the coefficient is positive the variable 1 increases variable 2, if the coefficient is negative the reverse is a decrease. The variables move in the same direction.

Method cor() in r with methods Spearman and Pearson Correlation Coefficients

Spearman Test Results

	Disadvantaged pupils (%)	FSM Eligibility (%)	Ave Performance	Ave Expenditure	English First Language
Disadvantaged pupils (%)	1.00	0.95	-0.58	0.75	-0.44
FSM Eligibility (%)	0.95	1.00	-0.61	0.75	-0.43
Ave. Performance (%)	-0.58	-0.61	1.00	-0.53	0.08
Ave. Expenditure(£)	0.75	0.75	-0.53	1.00	-0.38
English First Language%	-0.44	-0.43	0.08	-0.38	1.00

Table 19 Spearman Test Results

Pearson's Correlation Coefficient

	Disadvantaged pupils (%)	FSM Eligibility (%)	Ave Performance	Ave Expenditure	English First Language
Disadvantaged pupils (%)	1.00	0.95	-0.51	0.69	-0.59
FSM Eligibility (%)	0.95	1.00	-0.52	0.68	-0.55
Ave. Performance (%)	-0.51	-0.52	1.00	-0.43	0.04
Ave. Expenditure (£)	0.69	0.68	-0.43	1.00	-0.41
English First Language%	-0.59	-0.55	0.04	-0.41	1.00

Table 20 Pearson's Correlation Coefficient Test Results

Interpretation

The correlation indicating a significance levels in both result tables using the rcov() function indicated that Average Performance(%) and English First Language(%) was at a significance level. Spearman significance level table the result was 0.004. In Pearson Coefficient significance table the same pair of variables has a result of 0.1694.

Results

Individual Years 2009/10 to 2012/13 Performance Attained and Expenditure per Student, using Spearman's Coefficient

	AC5EM10	AC5EM11	AC5EM12	AC5EM13	T0910CAT5	T1011CAT5	T1112CAT5	T1213CAT5
AC5EM10	1	0.83	0.7	0.72	-0.46	-0.51	-0.53	-0.55
AC5EM11	0.83	1	0.75	0.75	-0.42	-0.47	-0.49	-0.52
AC5EM12	0.7	0.75	1	0.75	-0.37	-0.39	-0.4	-0.44
AC5EM13	0.72	0.75	0.75	1	-0.38	-0.42	-0.43	-0.45
T0910CAT5	-0.46	-0.42	-0.37	-0.38	1	0.93	0.87	0.84
T1011CAT5	-0.51	-0.47	-0.39	-0.42	0.93	1	0.93	0.88
T1112CAT5	-0.53	-0.49	-0.4	-0.43	0.87	0.93	1	0.93
T1213CAT5	-0.55	-0.52	-0.44	-0.45	0.84	0.88	0.93	1

Table 21 Spearman Results table (4 years)

Interpretation

Spend vs Performance	Performance vs Performance	Spend vs Spend
Moderately weak negative correlations	Strong positive Correlations	Strong positive Correlations
-0.39 to -0.55	0.7 to 0.83	0.84 to .93

Test 12 Regression Analysis Tree and Decision Tree

Description

"Recursive partitioning is a fundamental tool in data mining. It helps us explore the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome." (Kabacoff, 2014)

Method: R application using rpart()

Results

An number of regression trees were created in order to see how the data would appear on the regression plots. In figure 60 is the regression trees with the most interest.

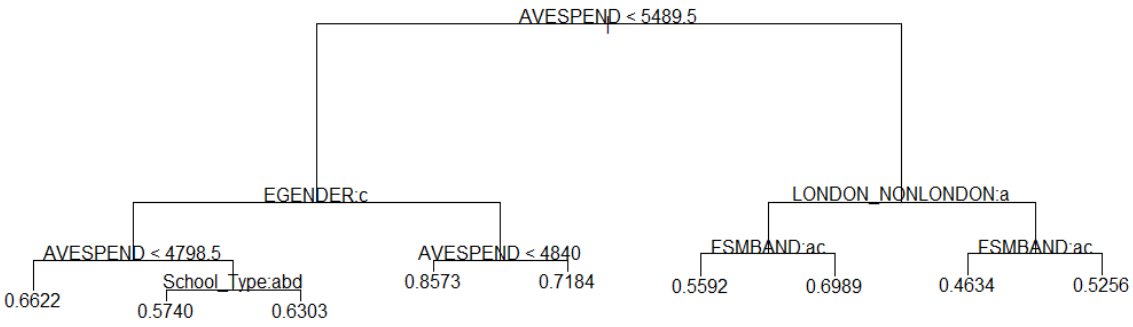


Figure 60 Regression Tree

Interpretation
This shows Average Spend with two branches, the left shows Gender having the most influence and the right the Location (London Non London).
Following the left hand side of the Regression Tree Gender branched off into two divisions for Average spend and the school types. On the right hand side of the branch the branches focus on gender.

Decision Tree
Method: Coding for this decision tree used the package Party in R and rpart. The tree was built using the r functions ctree() in package party
Before modeling, the data was split below into two subsets: training (70%) and test (30%).
The random seed is set to a value below to make the results reproducible, set.seed(1234)

```
ind <- sample(2, nrow(mydatasub), replace=TRUE, prob=c(0.7, 0.3))
trainData <- mydatasub[ind==1,]
testData <- mydatasub[ind==2,]
```

The party package was loaded to build a decision tree, and check the prediction result.
The target variable was set as Gender and all other variables were the independent variables.

```
myFormula <- Gender ~ Spend + Performance + Disadvantage
mydatasub_ctree <- ctree(myFormula, data=trainData)
```

The prediction was checked using the testing data. The rules were printed out and a tree plotted. A number of target variables were tested and it was decided that Gender would be the most interesting variable.

Result
The decision Tree plot show is a (Simple Style)

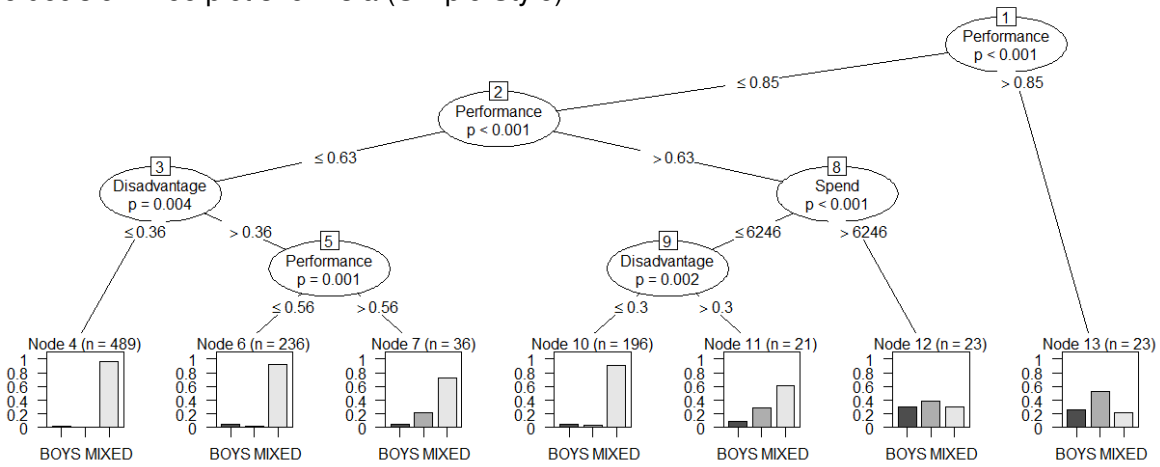


Figure 61 Decision Tree (R)

Interpretation
The decision tree shown could be presented using the rules or a plot. The plot is generally easier to follow. The decision tree in figure 61 required more formatting so as to clearly determine the Gender data.

Map of Geo Coding

Map based on Longitude and Latitude. Size shows Perfor1213 as an attribute.

Figure 62 Geo Coding Map 1 year

Appendix 4

Appendix 4- 1 Project Proposal

Project Proposal

An Exploratory Study into the Association between School Expenditure Levels per Student and overall School Performance in High Stake Examinations

Louise Blake
X13110535
Louise.Blake@student.ncirl.ie

20th of February, 2014

TABLE OF CONTENTS

1. Introduction

This project proposal is compiled for the Graduate Diploma in Science in Data Analytics programme, at the National College of Ireland. The purpose of this project will be to analyse and investigate a small to medium data to see if there is any association between school expenditure Levels per Student Head and overall School Performance in High Stakes Examinations.

2. Objectives and Contribution to the Knowledge

The principal objective of this project is to provide an exploratory study of the Association between School Expenditure Levels per Student Head and overall School Performance in High Stakes Examinations.

The topic of Students and Schools performance is by no means uncharted territory and this can be seen by the numbers of reports created by the numerous Government Department reports issued annually both in Ireland and abroad. In this project the performance and expenditure will be investigated to see if there is any association. By utilising statistics and initialling exploring the data it will give an approach to what to analyse, how to summarize their main characteristics, show patterns and use visualisation methods. This will advance my knowledge as to what is required when analysing data by utilising various techniques such as exploratory and descriptive analysis, building and evaluating predictive modeling.

There are a number of technical objectives; utilising languages and technologies will broaden my knowledge of their functionality and capabilities.

3. Background

With an interest in education and the measurement of performance, the search initially focused on the Irish education system. Firstly to see what information and datasets were available publically. This process illustrated that the Irish Department of Education[14] websites, currently have limited information available for public viewing and analysis. Alternatives viewed included www.schooldays.ie[15], again the availability was unsuccessful.

Following on from Irish options, alternative open data sources relating to Students Performance in Education included the GESIS European Values Study (GESIS, 2014). This site has a number of education related datasets but unfortunately not all in English. The UK Department of Education[16] website provides a number of performance tables of interest. Available are options to view an interactive map and reports on, school performance, characteristics, a range of socio-economic background indicators and funding reports. Earlier years from 1994 to 2009 limited access to pupil performance, proceeding years 2010 to 2012 have access to spend per pupil. The focus of this project proposal is on the years where both performance and school spend per pupil is available.

Subsequent research found a 2009 PISA report, 'Viewing The United Kingdom School System Through The Prism Of Pisa' that commented on the results examined by PISA for the UK were "not statistically significantly different from the OECD average." (OECD, 2009) PISA is the abbreviation for the Programme for International Student Assessment, it "is a worldwide study of 15-year-old school pupils' scholastic performance on mathematics, science, and reading" and compiled by the Organisation for Economic Co-operation and Development (OECD). According to the Wiki, the first PISA study performed in 2000 and then repeated every three years. The purpose of taking 15 year olds from schools worldwide is to improve education policies and outcomes. (Wiki, 2014)

Within the same PISA report, the OEDC referred to the UK expenditure per student stating:

[14] http://www.education.ie/en/
[15] http://www.schooldays.ie/articles/progression-to-college-data-on-schooldays.ie
[16] http://www.education.gov.uk/

"Only seven OECD countries spend more per student than the United Kingdom. While GDP per capita reflects the potential resources available for education in each country, it does not directly measure the financial resources actually invested in education. However, a comparison of countries' actual spending per student, on average, from the age of 6 up to the age of 15 puts the United Kingdom at an advantage, since only seven countries spend more than the United Kingdom on school education per average student. Across OECD countries, expenditure per student explains 9% of the variation in PISA mean reading performance between countries. " (OECD, 2009).

Data available, the GCSE's are generally taken by students aged between 14-16.This is the focus group for the project.

4. Technical Approach

Research into a topic for project proposal.
Investigate what research has being carried before.
Describe the framework for the project: duration of project, current resources, additional resources, costs if any apply.
Data source - Identify variables, measurements levels
Manipulation of the data source.
Explore the associations of the data
Investigate and research the tools and analysis
Utilise the determined analysis tools to be used and carry out tests
Visualise through graphs and charts
Evaluate analysis
Review the exploratory study results and highlight concerns.
Plan to develop an approach to meet the different audiences for presentation
Write up and finalise

5. Special resources required

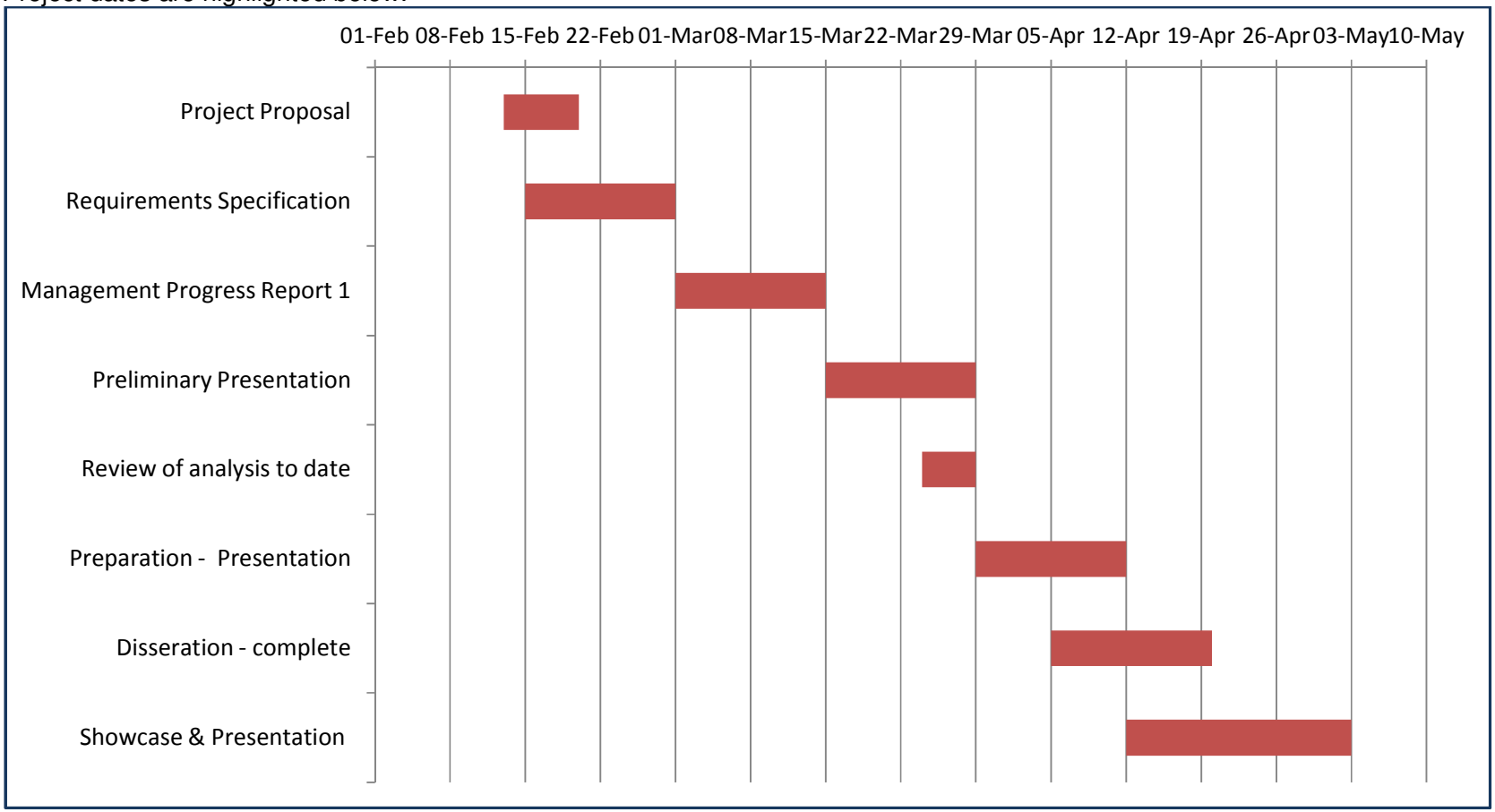
The statistical analysis tests will require further consultation with experienced members of staff. Additional reading on how to evaluate, test and analysis will be required determine what is appropriate for the data in question.

Research into the analysis and visualisation tools chosen will be required, along with the utilisation of application online Help options, as well as websites.

The software listed in the of the Technical details are available either on my own laptop or by utilising the college resources. The visualisation and analysis tools that will be utilised will include: R Studio, MS Excel, MS Access, Google Fusion, Python , again these resources are available.

6. Initial - Project Plan

The project is scheduled over a three month period; commending in February and concluding in May,2014.
Project dates are highlighted below:



Gantt chart using Microsoft Project with details on implementation steps and timelines.
(O'Loughlin, 2012)

7. Technical Details

MS Excel

Data extraction formats from the source are CSV and XML format. Initial processing of the data will be in MS Excel. It is possible it will be used for further analysis of the data but is too early to say at his point. There is an Add in option available to connect MySQL to Excel.

R

As well as being an open source, free software it is also a software environment for statistical computing and graphics. This application has a wealth of packages and supporting material for analysing data and graphing results these include packages such as ggplot2 and lattice:

Information from RStudio website:

- Ggplot2 is an enhanced data visualization package for R and it creates multi-layered graphics.
- Lattice graphics is a powerful data visualization system with an emphasis on data collected on several variables for each sampling unit. (RStudio, 2014)

MySQL / SQL/ MS Access

One of these options will be utilised in the process of Extraction, Transformation and Loading (ELT) of the data sourced from the website. The current preference is to utilise SQL.

Google Fusion Tables

As part of the visualisation of the results the schools are sub divide into regions and have postal codes so it possible provide a geographical element to the analysis. Some investigation is required as to whether it can accommodate the larger dataset. An initial test showed it could accommodate a small dataset.

Python

Python has packages which maybe capitalised on, with online supporting material which can be used to analyse and graph data results.

At this stage in the project the complete listing of tools, that will be utilised are not determined. It is possible alternatives will be sources in addition to the above. The preference will be to utilise the tools' available and become more proficient in what these options can do. Utilising to many newer tools environments

8. Datasets
The source data to be used is from the UK Department for Education.
http://www.education.gov.uk/schools/performance/download_data.html

Available on this website are a number of datasets, the 2 datasets which will be used for the purpose of this project are related as follows: KS4 Attainment Results and Spend per Pupil data

The metadata for the above datasets can be sourced at the following:
http://www.education.gov.uk/schools/performance/metadata.html
KS4 Attainment Results
Spend per Pupil data

9. Evaluation, Tests and Analysis
A number of methods and techniques will be explored for quantitative and visual data analysis. The results will be evaluated and presented, with regard to significance and quality.

Following an informal exploratory examination of the data, the main objectives are to explore the association between the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations. The tests expected to be used as part of the research will focus on certain variables, to see if they show interesting patterns. The variables have already being collected by the Department of Education (UK). However the data will require preparation prior to focusing on the main testing. The data quality has missing or omitted values and outliers need to determined.

Types of descriptive statistics which may be suitable;

- box plot;
- tables;
- model formulation;
- regression;

Aggregation of data
This will be to reduce the data for specific visualisations where a reduction in the data is benefit to the visualisation process.

Other evaluation, tests and analysis have yet to be defined.

10. Legal and Copyright Information

The copyright for this material was checked to see if it could be used for the purpose of this project. The UK Department for Education named as the corporate author and is acknowledged as the source of the material supplied.

"The Crown copyright-protected material featured on this website (other than Departmental or agency logos and visual media) may be reproduced free of charge in any format or medium, under the terms of the Open Government Licence[17]." (Department of Education, 2013)

This was verified with the Open Government Licence for public sector information to source the conditions for using the information under this licence. The following is an extract relating to the conditions:

> "Use of copyright and database right material expressly made available under this licence (the 'Information') indicates your acceptance of the terms and conditions below.
>
> The Licensor grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information subject to the conditions below.
>
> This licence does not affect your freedom under fair dealing or fair use or any other copyright or database right exceptions and limitations.
>
> You are free to:
>
> - copy, publish, distribute and transmit the Information;
> - adapt the Information;
> - exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.
>
> You must, where you do any of the above:
>
> - acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;
>
> If the Information Provider does not provide a specific attribution statement, or if you are using Information from several Information Providers and multiple attributions are not practical in your product or application, you may use the following: Contains public sector information licensed under the Open Government Licence v2.0.
>
> These are important conditions of this licence and if you fail to comply with them the rights granted to you under this licence, or any similar licence granted by the Licensor, will end automatically." (National Archives, 2014)

[17] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

11. Consultation with Specialisation Person(s)

Dr. Ioana Ghergulescu, Lecturer
Discussed the concept and the dataset sourced. The feedback from Ioana was this was an interesting idea and to look at predictive analysis.

Jonathan Lambert, Mathematics Development and Support Officer,
Discussed the project and the need for a title. The feedback suggested the title and to ensure to identify Variables, Levels of Measurement and Performance Variables.

LOUISE BLAKE 20/02/2014
Signature of student and date

Bibliography

Department of Education, 2013. *Use of Crown copyright material.* [Online]
Available at: http://www.education.gov.uk/help/legalinformation/a005237/crown-copyright
[Accessed 15 February 2014].
GESIS, 2014. *GESIS: European Values Study.* [Online]
Available at: http://www.gesis.org/en/services/data-analysis/survey-data/european-values-study/
[Accessed 16 February 2014].
National Archives, 2014. *Open Government Licence for public sector information.* [Online]
Available at: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[Accessed 15 February 2014].
OECD, 2009. *Viewing The United Kingdon School System Through The Prism of PISA.* [Online]
Available at: http://www.oecd.org/pisa/46624007.pdf
[Accessed 16 February 2014].
O'Loughlin, E., 2012. *How to Edit a Basic Gantt Chart in Excel 2010.* [Online]
Available at: http:/www.eugeneoloughlin.com/2010/10/how-to-edit-basic-gantt-chart-in-excel.html
[Accessed 17 February 2014].
RStudio, 2014. *R Studio Projects.* [Online]
Available at: http://www.rstudio.com/projects/
[Accessed 17th February 2014].
Tukey, J. W., n.d.
Wiki, 2014. *Programme for International Student Assessment.* [Online]
Available at: http://en.wikipedia.org/wiki/Programme_for_International_Student_Assessment
[Accessed 16 February 2014

Abbreviations

	Description
EDA	Exploratory Data Analysis
ETL	Extraction, Transformation and Loading
GCSE	General Certificate of Secondary Education
GDP	Gross Domestic Product
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment

Appendix 4- 2 Initial Requirements Specification (RS)

Requirements Specification (RS)

An Exploratory Study into the Association between School Expenditure Levels per Student and overall School Performance in High Stake Examinations

Louise Blake
x13110535
Louise.Blake@student.ncirl.ie

2nd March, 2014
Higher Diploma in Science in Data Analytics

Document Control
Revision History

Date	Version	Scope of Activity	Prepared	Reviewed	Approved
02/03/2014	1.0	Create Requirements Specification	LB	X	X

Distribution List

Name	Title	Version
Dr. Ioana Ghergulescu	Lecturer	1.0

Related Documents

Title	Comments
Use Case Model - User Prospective Of System	Included in section 4

1. Introduction

1.1 Purpose

The purpose of this document is to set out the requirements for the development of an Exploratory Study into the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations. The audience this document is intended for; includes the related stakeholders and the project team. This version (1.0) provides the general description of the system and to ensure there is adequate information central for the commencement of the project.

The requirements specification provides the agreed frame work in which the deliverables and communication are set out.

The precise nature of the requirements specification is to show the various stakeholders what features are described. The major stakeholder, who is referred to as the client is the Information Standards Board (ISB) and works with a wide range of stakeholders, including: the Department of Education; various sponsors and directors of Department for Education executive agencies and other Government departments and relevant organisations. (Information Standards Board, 2014).

This document will be reviewed and revised during the project lifecycle, by the system developer to ensure any changes are incorporated.

1.2 Project Scope

The scope of the project is to develop an exploratory study into the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations. The system focuses on the deliverables required by the client, who will utilise the study in future planning strategies and provide the information to other related stakeholders.

The data to be utilised in this system is based on specific tables, provided by the Department of Education (Data provider) relates to student performance and cost per student. It is acknowledged that the information in these Tables only provided part of the picture of each school and its student' achievements. As the study includes expenditure, the data acquired has details of the Department of Education's Consistent Funding Reports. "The consistent financial reporting framework (CFR) is a standard framework into which schools should code their income and expenditure to enable production of simple, standardised reports for governors and local authorities." (Department of Education, 2014). The provided data will be utilised by the Statistician to provide data analysis and to visualise the data for the client. Following this process reports and presentations will be prepared and presented.

1.3 Project Restrictions

Restrictions are acknowledged at the commencement of the project which incorporate the following:

Data

The student performance on a subject by subject basis is not part of this study or a requirement of the system. It is acknowledged that on an annual basis results in schools can change in the future compared to historical results. There are other sources relating to schools performance which can be sourced from organisations such as Ofsted and obtained from www.ofsted.gov.uk[18]. The details from Ofsted and other post primary student performance related sources of data are not included as part of this study, only the data tables sourced from the Department of Education (UK) as referred to in the requirements definition.

Time

The duration of project is confined to pre agreed dates which has a commencement date February 2014 and a completion date 5th May 2014.

Human Resource

The availability of specialist people to complete all tasks is limited.

[18] www.ofsted.gov.uk

Software Resources
The software resources are accepted as what, is provided by the college software, free downloads and students own resources.
Budget
There is a no budget to source additional software, however if specifically required it may be possible to sourced via the college.
Legal and Copyright
The data provided is open source and it is acknowledged that the department of Education are the corporate author and supplier of the material. It is used under the Crown copyright Open Government Licence. (Department of Education, 2013).

1.4 Definitions, Acronyms, and Abbreviations

Definitions

Stakeholder/ user	Individuals or organisations who have a vested interest in this project and need to be considered throughout the project.

The Actor Glossary contains a list of all the participating Actors (external users and systems) in the system.

Actor Glossary (alphabetical order)

Client (External)	The customer is the person representing the organisation who authorises the study, receives the progress reports, final report and presentations.
Data Provider	Provides the initial source of the 'raw' data utilised in the system.
Presenter:	The person or persons who creates and presents the required presentations, for the specific audiences.
Project Manager	Interacts with the Client and the other project team members.
Reports writer:	The person or persons who writes the report for the client and compiles the deliverables in a readable format.
SQL DB / Administrator:	The person or persons who will prepare the database data, manage access and security to the database data.
Statistician:	The person or persons who analyses the database data and produces findings for the reports.
System Designer:	The person or persons who designs the system to deliver the deliverables. (Use case author)

Acronyms and Abbreviations

CFR	Consistent Funding Reports
ESCS	Education, Skills and Children's Services in England
GCSE	General Certificate of Secondary Education
ISB	The Information Standards Board
KS4	Key Stage 4
LA	Local Authority

2. User Requirements Definition

The objectives for the system are to provide the requirements from the client's perspective, to agreed on a new system development and to provide agreed deliverables.

The client requires an exploratory study into the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations. The school expenditure levels per student head provided are from 2009 to 2012 (four years) with the overall performance available for students at GCSE School curriculum; Key Stage 4 (KS4). These Tables give information on the achievement of student in secondary schools in a percentage format. They also include details of the school; numbers of student, school type, gender, age range, local authority (LA) area in England as a whole. Only the data tables sourced from the Department of Education relating to England_KS4 Performance and Consistent Funding Reports (CFR) will be utilised.

3. Requirements Specification

3.1 Output requirements

The client requires analysis and visualisation of the database data, this includes the following:

- Descriptive statistics, to summarise the data.
    - Visuals: box plot; summary tables;
- Predictive analytics, explore the associations within the data (correlation) and modeling relationships within the data. Predictive models
    - Correlation
    - Regression analysis.
- Utilise a number of technical algorithms:
    - Cluster analysis – K-means
- Decision Models, to describe the relationship between student performance percentage and pupil funding spend.
- Explore the categorisation of the schools, by involving a number of other variables; such as school type, gender, product preferences and life stage
- Aggregation of data - Reduce the data for specific visualisations where, a reduction in the data is of benefit to the visualisation process.
    - A map locating the top 5% and lowest 5% of performing schools.
    - Geospatial predictive modeling
- Other evaluation, tests and analysis have yet to be defined.
- Provide visualisation of results where possible using, graphs, charts and maps,
- Create a report for stakeholders,
- Create and deliver presentations.

3.2 Database Data Description

Two tables are recognised as a requirement for this study, each table has a huge number of attributes. Specific attributes identified as relevant and the others can be exclude.

Below are the identified attributes of interest for table england_ks4
433 attributes
Over 5,300 records

Column	Metafile heading	Metafile description	Datatype
1	RECTYPE	Record type	1=mainstream school; 2=special school; 4=Local Authority; 5=National (all schools); 7=National (maintained schools)
3	LEA	Local Authority code (see separate list of Local Authorities and their codes)	VARCHAR
5	URN	School Unique Reference Number	NUMERIC PRIMARY KEY
6	SCHNAME	School name	VARCHAR
11	TOWN	School town	VARCHAR
12	PCODE	School postcode	VARCHAR
16	NFTYPE	School type	varchar - abbreviations in a list
19	EGENDER	School gender of entry	CATEGORICAL
23	AGERANGE	Age range	VARCHAR
25	TOTPUPS	Number of pupils on roll (all ages)	NUMERIC
26	TPUP	Number of pupils at the end of Key Stage 4	NUMERIC
27	KS2APS	Key Stage 2 Average Points Score of Key Stage 4 cohort	NUMERIC
34	TFSMCLA	Number of disadvantaged pupils	NUMERIC
35	PTFSMCLA	Percentage of pupils who are disadvantaged	PERCENTAGE
36	TNOTFSMCLA	Number of non-disadvantaged pupils	NUMERIC
37	PTNOTFSMCLA	Percentage of pupils who are not disadvantaged	PERCENTAGE
53	AC5EM09	Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs - 2009	PERCENTAGE
54	AC5EM10	Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs - 2010	PERCENTAGE
55	AC5EM11	Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs - 2011	PERCENTAGE
56	AC5EM12	Percentage of pupils achieving 5+ A*-C or equivalents including A*-C in both English and mathematics GCSEs - 2012	PERCENTAGE
128	NUMBOYS	Total boys on roll (including part-time pupils)	NUMERIC
129	NUMGIRLS	Total girls on roll (including part-time pupils)	NUMERIC
130	P15END4	Percentage of pupils at the end of Key Stage 4 aged 15	PERCENTAGE
131	P14END4	Percentage of pupils at the end of Key Stage 4 aged 14 or under	PERCENTAGE
132	TEALGRP1	Number of Key Stage 4 pupils with English as their first language	NUMERIC
133	PTEALGRP1	Percentage of Key Stage 4 pupils with English as their first language	PERCENTAGE

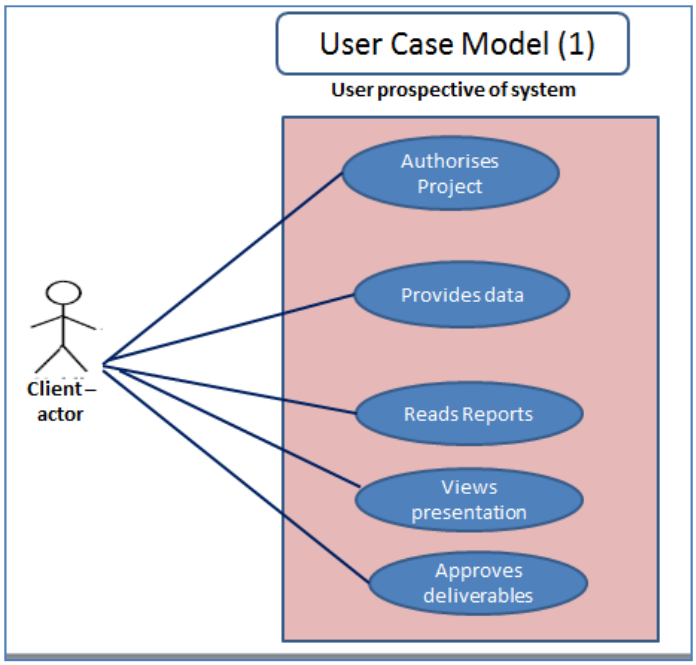
Attributes of interest in england_crf shown below.
There are 128 attributes in total
Over 18,114 rows

column no	Attribute	Description	Developer Comment
1	URN	Unique Reference Number	PK
2	LONDON_NON-LONDON	London/Non-London	categorical option
3	MEDIAN	Group to compare	categorical option
4	PUPILS	Number of pupils (FTE)	numeric
5	FSM	Percentage of pupils eligible for Free School Meals (FSM)	categorical option
6	FSMBAND	Free school meals eligibility band	categorical option
39	T0910CAT5	2009-10 Total Expenditure (£ per pupil)	Currency
40	T1011CAT5	2010-11 Total Expenditure (£ per pupil)	Currency
41	T1112CAT5	2011-12 Total Expenditure (£ per pupil)	Currency
42	T1213CAT5	2012-13 Total Expenditure (£ per pupil)	Currency

4. Functional requirements

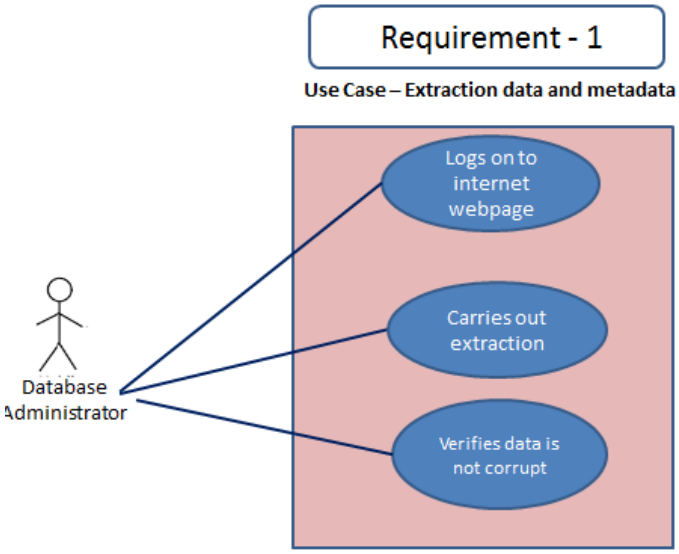
4.1 User case diagram - User Prospective Of System

This diagram represents the user case model from the prospective of the Client -Actor.



4.2 Requirement 1 - Extraction Data and Metadata

Scope
The Client Actor authorises the project, provides database data and metadata.
Description
The Use Case - Extraction Data and Metadata the initial requirement. Without this information the other requirements cannot take place.
Use case Diagram



Preconditions
The user needs to have access to the internet and to the related web pages for downloading the database data[19] and metadata[20] for extraction.
Trigger /event
Authorisation to go ahead with the project.

[19] http://www.education.gov.uk/schools/performance/download_data.html
[20] http://www.education.gov.uk/schools/performance/metadata.html

Flow Description

1. Download data in csv format,
2. Save to a location, accessible to the user,
3. Inspect data integrity following extraction, to ensure it is not compromised.

Post conditions

Database and metadata are extracted from the website the user case is complete.

Termination

The data extraction is not accepted if there is corruption of the data occurs and a new extraction is required.

4.3 Requirement 2 - Database Data Management

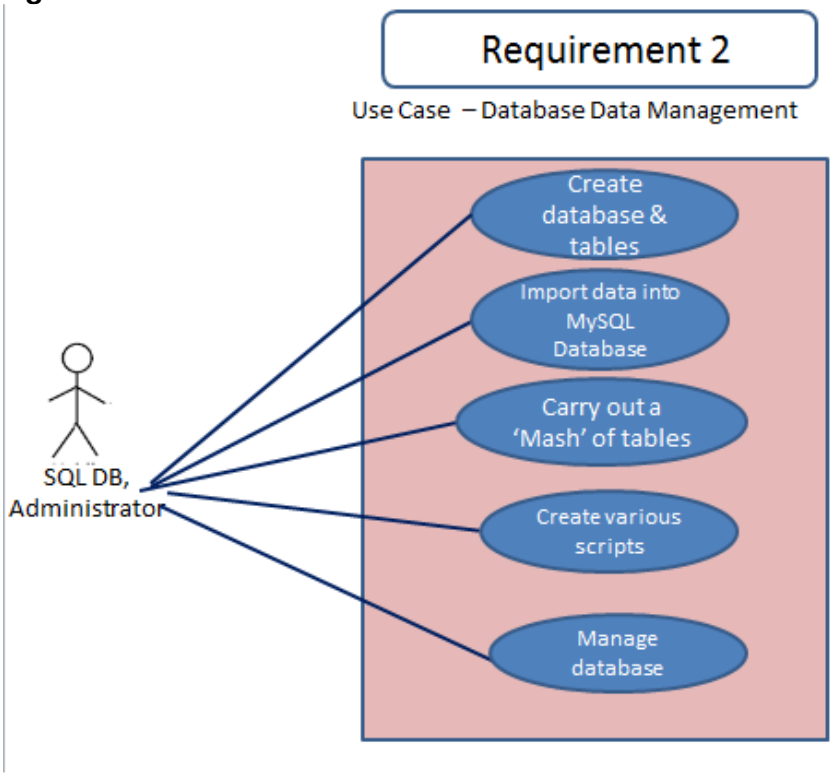
Scope

The Database Actor creates the MySQL database and tables, imports the database data and metadata provided. The MySQL Database serves as a repository which can allow for searches, analyse, visualisation, which is the next requirement.

Description

The requirement 2 -Database Data Management is where the data transfer to a repository a MySQL Database, it is cleansed, mash up and queried for quality.

Use case Diagram



Preconditions

The user needs to have access to the initial database data and metadata extracted in requirement 1.

Trigger /event

Commence when Database data and metadata is available.

Flow Description

1. MySQL database is created,
2. Tables created based on the metadata,
3. Data imported into the tables,
4. Data integrity is inspected following import, to ensure it is not compromised,
5. Cleanse the data,
6. Create scripts to allow for data analysis.

Post conditions

Once the data is available in the MySql database and scripts prepared the next requirement can commence.

Termination

Termination can occur is data is in accessible or corrupt.

4.4 Requirement 3 - Analysis of Data & Visualisation

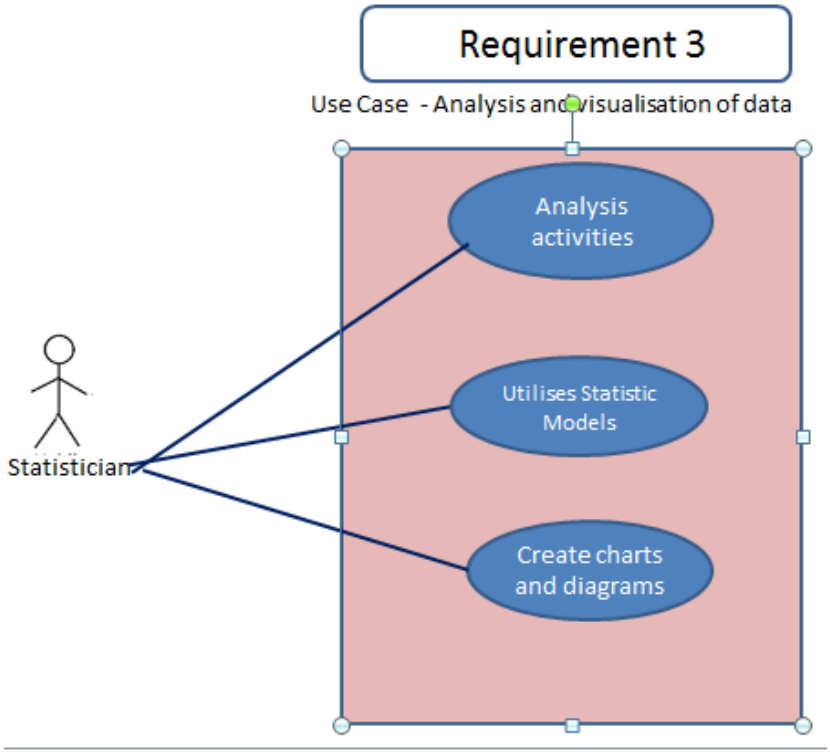
Scope

The Statistician Actor creates the MySQL database and tables, imports the database data and metadata provided.

Description

The requirement 3 -Analysis of Data and Visualisation is where the data is analysed and visualisation takes place.

Use case Diagram



Preconditions

The user needs to have access to the MySQL Database and the requirements specifications output.

Trigger /event

When the My SQL Database is available and populated with data.

Flow Description

1. Access database data,
2. Compile scripts to provide analysis as per specifications output,
3. Create visualisation chart, graphs, maps,
4. Compile findings for report

Post conditions

The outputs from this required is essential to complete the next requirement.

Termination

The analysis will be terminated only if project is terminated.

4.5 Requirement 4 - Report and Presentations

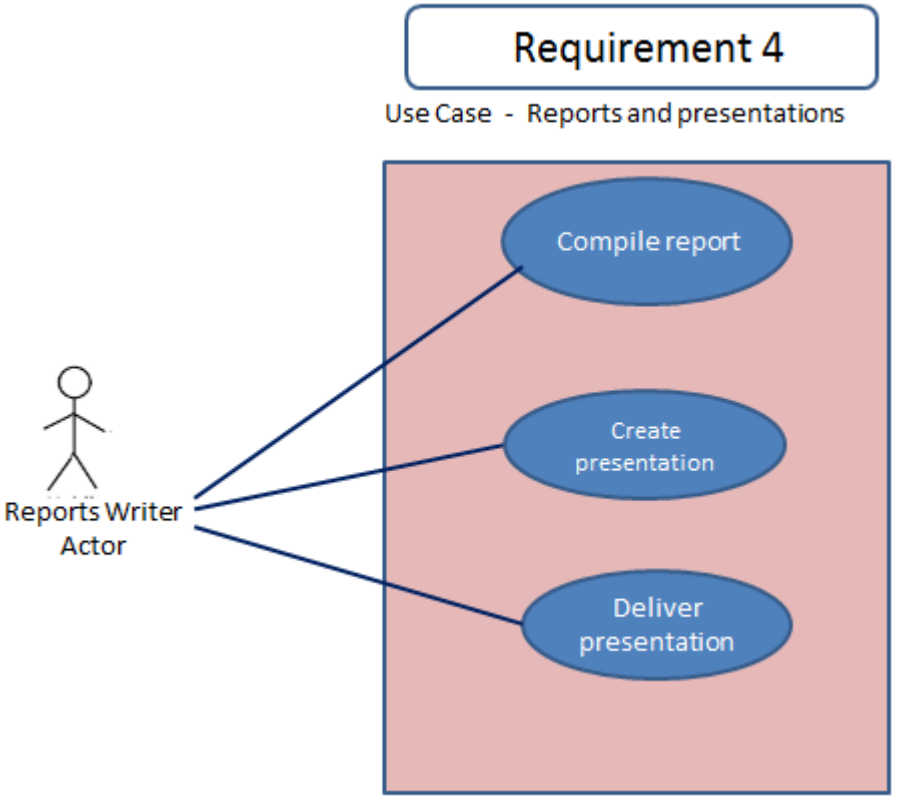
Scope

The Report Writer Actor creates the report and presentations.

Description

Deliver two presentations to the client; the target audience (a) a technical focus group and (b) non-technical group.

Use case Diagram



Preconditions

The Reports Writer user needs to have access to the analysis and visuals created in requirements specifications 3. They require all progress reports, change reports and communications.

Trigger /event

This Reports Writer can commence writing the report after requirement 2 has commenced. The report is prepared in conjunction with the other requirements.

Flow Description

1. Access database data,
2. Reviews all stages
3. Incorporated visualisation chart, graphs, maps,
4. Compile findings for report

Post conditions

The outputs from this required is essential to produce a report required to Client Actor.

Termination

The requirement will only be terminated if the project is cancelled.

4.6 Non Functional Requirements

4.6.1 Performance/ Response Time
The volume of data in this data analysis does not require to be analysed at a significantly fast speed or with urgency. It is a noted in this project due to the system design, whereby there is a utilisation of a number of tools which are not fully integrated and automated the speed of performance and response time are not a critical element.

4.6.2 Availability requirement
The system is not required to be available for the user.

4.6.3 Recover requirement
In the event of hardware failure a backup of project data will be required. It is expected that the project data and all scripts will be backed up on an external device (external hard drive) and a second on line back up kept in a cloud storage solution (Dropbox). In the event of hardware failure the software utilised for this project will require installation on an alternative laptop.

4.6.4 Robustness requirement
There is no requirement for robustness in this project. Refer to Recover requirement (3.2.3).

4.6.5 Security requirement.
To gain access to the 'raw' data taken from the Department of Education's web site does not require any specific security rights. The data required is available for public use, as an Open Source database.

To access the system developed and designed requires access to the device (laptop) which currently hosts the system. The system does not pose a high security requirement, as it is not accessible to the public or other users, that are likely to damage or corrupt the data. The treat of hardware failure

4.6.6 Reliability requirement
The data used in this study is compiled annually. It is dependant of the accuracy and availability of data produced and compiled by the Department of Education. The software and hardware components refer to the Recover Requirement (3.2.3).

4.6.7 Maintainability requirement
There is no need to maintain the system designed once it is created.

4.6.8 Portability requirement
The methods processed can be utilised by other users, along with access to the data and the applications utilised in this study. The application scripts generated during the study may not run properly if the Data metadata is amended by the providing client.

4.6.9 Extendibility requirement
Currently there is not a requirement to extend the study, however it is possible that the methods and processes created can be re-used for future years.

4.6.10 Reusability requirement
There is currently no requirement for reusability.

4.6.11 Resource utilization requirement
Hardware and Software

There is a need to provide a laptop, internet access, backup storage device.

The installation of software currently not installed.

Project Planning

There is a need to manage the project within the time specified, to succeed in producing successful deliverables. There a number of key deliverables which will require research prior to development.

Learning - Knowledge

There will be a requirement to consult with Lecturers and specialist person(s). Further resources include Library resources' and Online help tools .

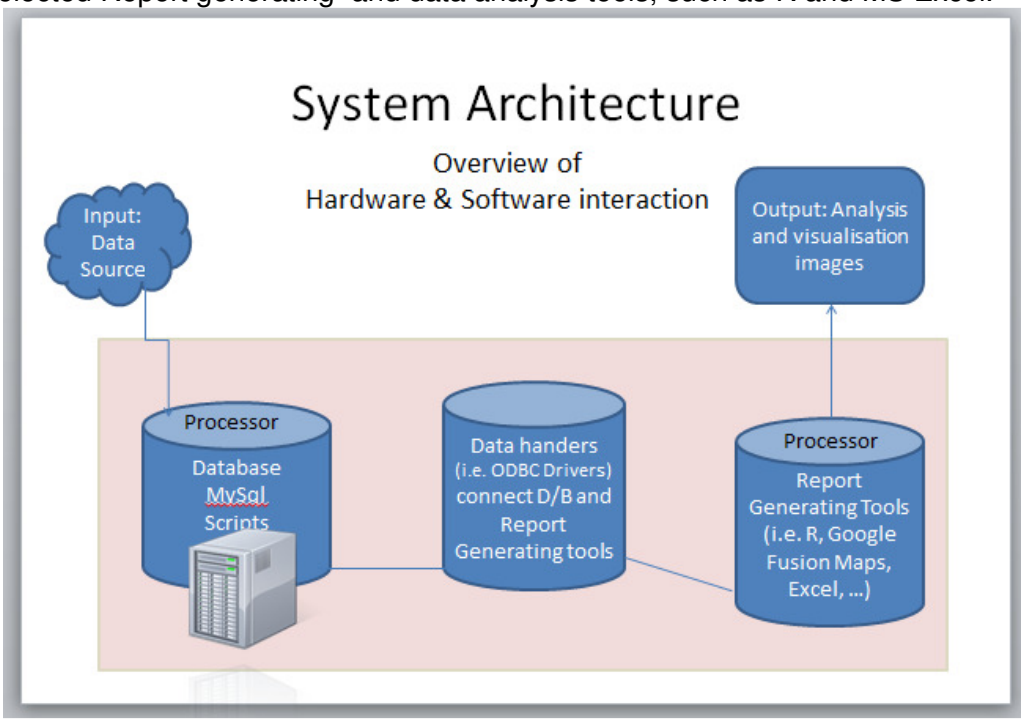
5. Interface Requirements

This section is not applicable as there will not be an end user interface designed. There is scope in the future to incorporate an end user interface where the user can request a particular analysis or graph result.

6. System Architecture

The following section describes system architecture and is a formal representation of the system, showing the system components, how they interact with each other,and the extraction of the data from the website. The components include the creation and management the MySQL Database, use of connectors where possible to get data out of the MySQL database into the various Reports Generating tools to generate results.
This architecture for the system was chosen because of the following reasons.

- The data is two sourced tables and requires a 'mashup' so as to data is accurately matched. This is server-based mashups which will analyze and reformat the data. Using an MySQL database will meet these requirements.
- In addition MySQL has add-in options which, can be used to integrated with the selected Report generating and data analysis tools, such as R and MS Excel.



7. System Evolution

The system can evolve by developing more integration and automation which, would produce faster results with less reliance on specialist people. This would be benefit the Client, as the organisation would have future options to incorporate data for other student cohort years. The system is designed to focuses on students whose performance relates to the KS4 cohort and provide analysis and visuals relating to this data. By evolving the system can accommodate other Key stages in the student cycle such as KS3 or KS5. Another derivative to the system is to incorporate future years of expenditure and performance for the same student cohort KS4.

7.1 System Evolution Constraint

A noted system constraint to the system evolution would be if the Department of Education amends the open data availability. In addition to this threat, the Department of Education's could in the future amend the attributes on the tables selected.

Bibliography

Department of Education, 2013. *Use of Crown copyright material.* [Online]
Available at: http://www.education.gov.uk/help/legalinformation/a005237/crown-copyright
[Accessed 15 February 2014].
Department of Education, 2014. *Consistent financial reporting.* [Online]
Available at:
http://www.education.gov.uk/schools/adminandfinance/financialmanagement/consistentreporting
[Accessed 28 February 2014].
Information Standards Board, 2014. *About the ISB.* [Online]
Available at: http://www.education.gov.uk/escs-isb/about/whoweare/a0075244/stakeholders
[Accessed 24 February 2014].

Appendix 4- 3 Management Progress Report 1

Management Progress Report 1

An Exploratory Study into the Association between School Expenditure Levels per Student Head and overall School Performance in High Stake Examinations

Report Period 02/02/2014 to 16/03/2014
Release: 1.0
Date: 16 March 2014
Author: Louise Blake

1. Report History

1.1 Document Location
This document is valid on the day of submission via Moodle 16/03/2014.

1.2 Revision History

Revision date	Author	Version	Summary of Changes	Changes marked
15/03/2014	Louise Blake	1.0	First Management Progress Report	Not applicable

1.3 Approvals
This document requires the following approvals:

Name	Title	Date of Issue	Version
Louise Blake	Author	15/03/2014	1.0

1.4 Distribution
This document has additionally been distributed to:

Name	Title	Date of Issue	Status
Dr. Ioana Ghergulescu	Lecturer	16/03/2014	submitted

2. Highlight Report from 02/02/2014 to 16/03/2014

3. Highlight Report Purpose
The purpose of this report is to provide details to the Project Board with a summary of the status of the project and is used to monitor progress. The Project manager uses the Highlight report to advise the Project Board of any potential problems or areas where the Project Board can help.

4. Achievements Completed during the Period
During the period referenced the following documents were prepared and submitted:

- Project Proposal,
- Requirement Specification v1.0.

The following requirements and accompanying tasks were complete:

- Requirement (1) Extraction & Transformation process,
- Requirement (2) Repository, Load and Transform.

5. Activities during the Period
Documents prepared and submitted

- Project Proposal
- Requirement Specification v1.0

ELT process
Requirement 1 : Extraction & Transform

- Extracted datasets and Metadata from Source (CSV format)
- Identify attributes, variables and levels of measurement
- Transform data so it fits operational needs

Data audit (1)

- Check quality of data extracted
- Pre - Cleanse of data for import
- Results are inspected to verify correctness

Requirement 2 : Repository (Load & Transform)

- Load data into the data warehouse
- MySQL database (Repository) created
- Tables created using Metadata
- Import of Data from CSV to MySQL
- Merge data

Data audit (2)

- Data Quality and Cleanse
- Detection and removal of anomalies

Documents (WIP[21])

- Dissertation
- RAID [22] Log

Requirement 2 : Repository (WIP)

- Integration Created ODBC driver for integration MySQL --> R

6. Budget Status
No change to the budget.

7. Schedule Status
See Appendix 1 for updated Gantt chart and in Appendix 2 a Mind map.
Reason for update is to include additional tasks and changes to the following milestone events, Preliminary Presentation and Management Progress Reports.

[21] WIP = Work In progress
[22] RAID = Risks, Assumptions, Issues and Dependencies

8. Planned Work for Next Period (to 29/03/2014)

Proceed with Requirement 3 : Analysis & Visualisation

Analysis Tool	Task
R	1. Statistical methods: Analyzing the data using the values of mean, standard deviation, range, or clustering algorithms.
R	2. Locate Outliers - determine to remove or leave.
R	3. Create box plots
R	4. Predictive analytics, explore the associations within the data (correlation) and modelling relationships within the data. Predictive models; Correlation, Regression analysis.
R	5. Utilise a number of technical algorithms: Cluster analysis – K-means
SQL	6. Create scripts to aggregate data for use in R

9. Project, Risks, Assumptions, Issues and Dependencies

In the RAID Log details are recorded and below are a summary of the contents of the Log.

9.1 Project issues

There are currently have 6 issues on our project issue log, 5 have been resolved and 1 is outstanding. This outstanding issue is relating to Integration with R. Progress with the Requirement (3) will not be hindered by this issue, as there is an alternative option until the issue is resolved.

9.2 Project Risks

To date the Risks log contains 5 items. Most of these project risks were identified during the requirements process. The main risk is time required to complete the project and the required milestones. This needs to be carefully monitored and a revised Gantt chart has been created to assist in the management of the project. The other risks relating to technology; a Hardware and data backup is in place, the required software is software installed and available. Currently there is no high or medium risk item logged to cause concern.

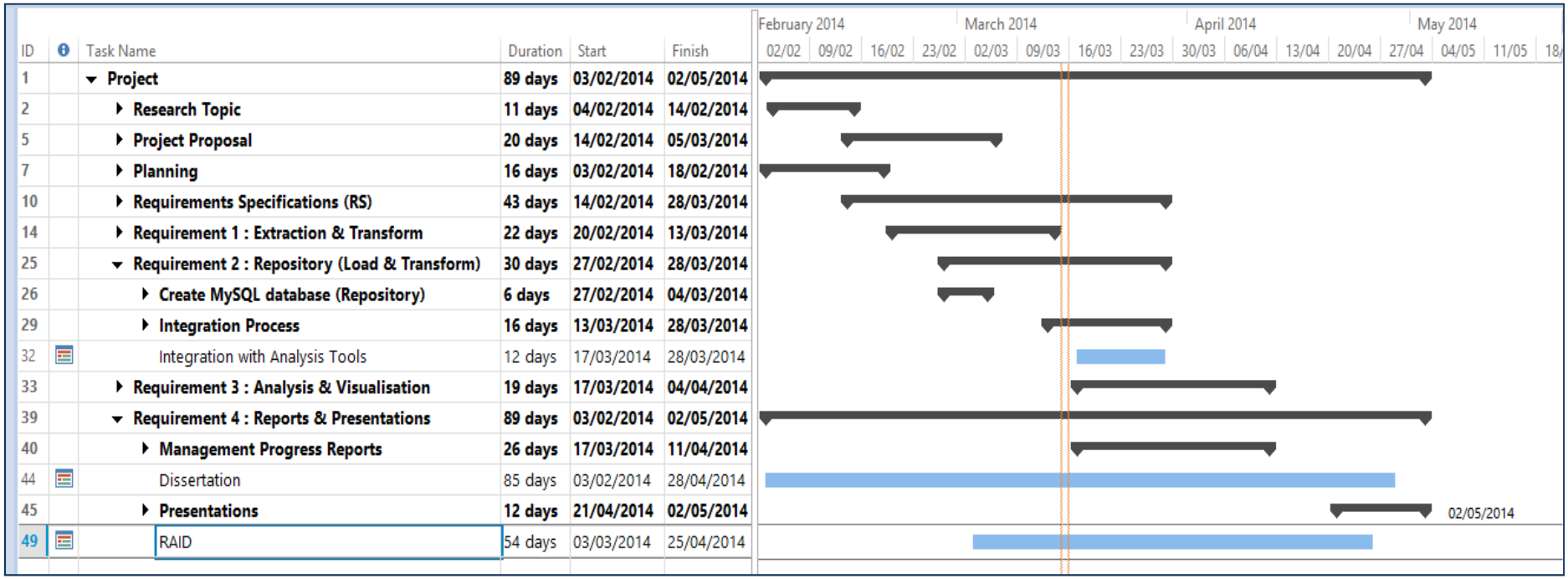
9.3 Project Assumptions

The assumptions logged during the planning stage have been revised to take into account the breakdown of the tasks needed in each of the requirements. The assumptions incorporated hardware, software, learning resources and the application of models to produce visualisation of the data.

9.4 Project Dependencies

To date a number of key project dependencies are complete allowing for the project to progress to the next dependency. The dependencies are listed on the RAID Log.

APPENDIX 1

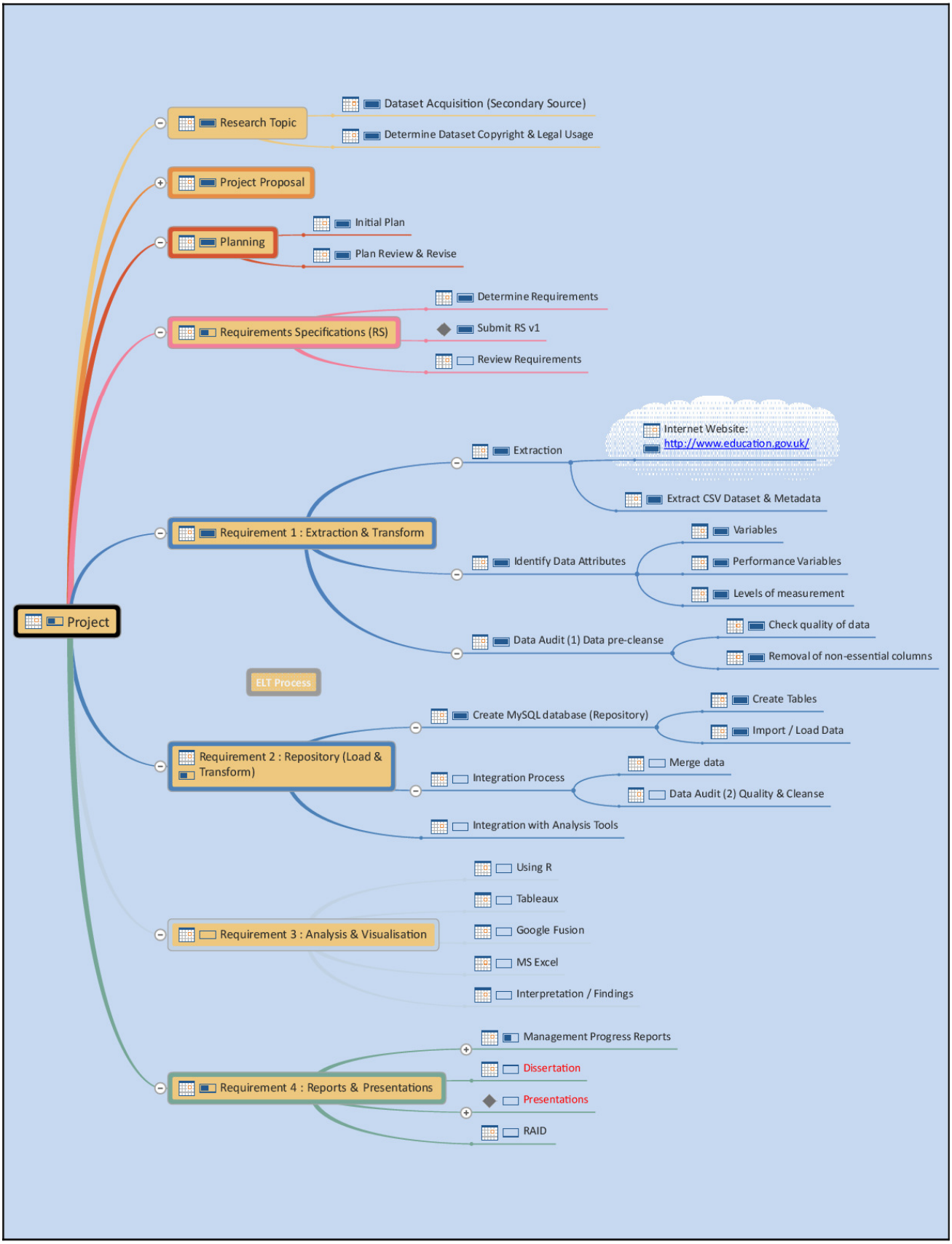


ID	Task Name	Duration	Start	Finish
1	**Project**	**89 days**	**03/02/2014**	**02/05/2014**
2	**Research Topic**	**11 days**	**04/02/2014**	**14/02/2014**
5	**Project Proposal**	**20 days**	**14/02/2014**	**05/03/2014**
7	**Planning**	**16 days**	**03/02/2014**	**18/02/2014**
10	**Requirements Specifications (RS)**	**43 days**	**14/02/2014**	**28/03/2014**
14	**Requirement 1 : Extraction & Transform**	**22 days**	**20/02/2014**	**13/03/2014**
25	**Requirement 2 : Repository (Load & Transform)**	**30 days**	**27/02/2014**	**28/03/2014**
26	**Create MySQL database (Repository)**	**6 days**	**27/02/2014**	**04/03/2014**
29	**Integration Process**	**16 days**	**13/03/2014**	**28/03/2014**
32	Integration with Analysis Tools	12 days	17/03/2014	28/03/2014
33	**Requirement 3 : Analysis & Visualisation**	**19 days**	**17/03/2014**	**04/04/2014**
39	**Requirement 4 : Reports & Presentations**	**89 days**	**03/02/2014**	**02/05/2014**
40	**Management Progress Reports**	**26 days**	**17/03/2014**	**11/04/2014**
44	Dissertation	85 days	03/02/2014	28/04/2014
45	**Presentations**	**12 days**	**21/04/2014**	**02/05/2014**
49	RAID	54 days	03/03/2014	25/04/2014

Created using MindGenius

APPENDIX 2

MindGenius Business 5 Mind map Appendix 2 LOUISE



1.0 (MindGenius)

Appendix 4- 4 Management Progress Report 2

Management Progress Report 2

An Exploratory Study into the Association between School Expenditure Levels per Student and Overall School Performance in High Stake Examinations

Report Period 16/03/2014 to 12/04/2014
Release: 1.0
Date: 12 April 2014
Author: Louise Blake

1. Report History

1.1 Document Location
This document is valid on the day of submission via Moodle 12/04/2014.

1.2 Revision History

Revision date	Author	Version	Summary of Changes	Changes marked
12/04/2014	Louise Blake	1.0	Second Management Progress Report	Not applicable

1.3 Approvals
This document requires the following approvals:

Name	Title	Date of Issue	Version
Louise Blake	Author	12/04/2014	1.0

1.4 Distribution
This document has additionally been distributed to:

Name	Title	Date of Issue	Status
Dr. Ioana Ghergulescu	Lecturer	12/04/2014	submitted

2. Highlight Report from 16/03/2014 to 12/04/2014

3. Highlight Report Purpose

The purpose of this report is to provide details to the Project Board with a summary of the status of the project and is used to monitor progress. The Project manager uses the Highlight report to advise the Project Board of any potential problems or areas where the Project Board can help.

5. Achievements Completed during the Period

The progress in the project during this period consists of:

- The commencement and progress on the data analysis and visualisation tasks.
- Writing up the details in the Dissertation relating to the background, dataset descriptions, issues with data and the ELT[23] processes to date.
- Installation and utilisation of Tableau Software.
- Attendance of a Tableau Showcase.

During this period, no documents were prepared and submitted to Moodle.

6. Budget Status

No change to the budget.

7. Schedule Status

The project is still on schedule.

Due to an extension in the project completion date there are changes to key milestone events dates; these are noted on an updated version of the project plan in the Appendix 1.

Reason for updating the schedule is to incorporate changes to the following milestone events, Management Progress Reports, Dissertation submission and Presentation Dates.

23 Extraction, Loading and Transformation

8. Activities during the Period

Status	Analysis Tool	Task	Comment
Complete	R	1. Statistical methods: Analyzing the data using the values of mean, standard deviation, ranges	Descriptive statistics - write up required (8. Planned Work)
Complete	R	2. Locate Outliers	Determined to leave in.
Incomplete	R	3. Create box plots	Reassessing attributes to use.
Part Complete	R	4. Predictive analytics, explore the associations within the data (correlation) and modelling relationships within the data. Predictive models; Correlation, Regression analysis.	To continue (8. Planned Work)
Incomplete	R	5. Utilise a number of technical algorithms: 6. Cluster analysis – K-means	To explore - (8. Planned Work)
Scripts created	SQL	7. Create scripts to aggregate data for use in R	ODBC driver connection - unsuccessfully (Details in RAID) Alternative utilised - MS Access for aggregation.
	MS Access	8. Connected to SQL Repository 9. Scripts create to - aggregation of Data	See RAID for change
	Tableau	10. Set up & sourced Education Licence 11. Learned how to use Tableau for visuals 12. Attended a Showcase on Tableau 27th March, 2014 13. Sourced specific geo codes for utilising in Map visualisation	Additional software for visualisation of Data and analysis. This is also going to be used instead of Google maps. (8. Planned Work)
WIP[24]		14. Dissertation - Writing in progress	
Updated		• RAID [25] Log	

[24] WIP = Work In progress
[25] RAID = Risks, Assumptions, Issues and Dependencies

9. Planned Work for Next Period (to 27/04/2014)

Continue with progress, Requirement 3: Analysis & Visualisation continue

Analysis Tool	Comment
R	Descriptive statistics - document and tabulated work complete to date Histogram and Scatter plots
R	Box plots - reassess attributes to use
R	Predictive analytics to continue Document work complete to date, example: - Regression analysis - Correlation - Mann Whitney Test / Two sample Test tests
Weka	Explore the use of Weka for Cluster analysis and Decision Tree options
MS Access	Create as required subsets from the data for use in Weka , R and Tableau.
Tableau	Document existing analysis and visuals created and suitable for inclusion.

10. Project, Risks, Assumptions, Issues and Dependencies

In the RAID Log details are recorded and below are a summary of the contents of the Log.

10.1 Project issues

There are currently have 8 issues on our project issue log 7 have been resolved and 1 is open. This open issue relates to a proposal to Integrate R with SQL. Progress is continuing with analysis using MS Access and Excel to aggregate data so the project is not hindered by this issue. The purpose of this connection was to give a more automated system, and will be worked on towards the later stage of the project. The reason for deferment is give sufficient time to the analysis and visualisation.

10.2 Project Risks

To date the Risks log contains 5 items. This has not had any inclusions since the project risks were identified during the requirements process. One of the primary concerns was Time, due to an extension in the project dates a revision in dates is required. This will still needs to be carefully monitored and a revised Gantt chart (Appendix 1) has been created to assist in the management of the project. The other risks relating to technology; a Hardware and data backup is in place, the required software is software installed and available. Currently there is no high or medium risk item logged to cause concern.

10.3 Project Assumptions

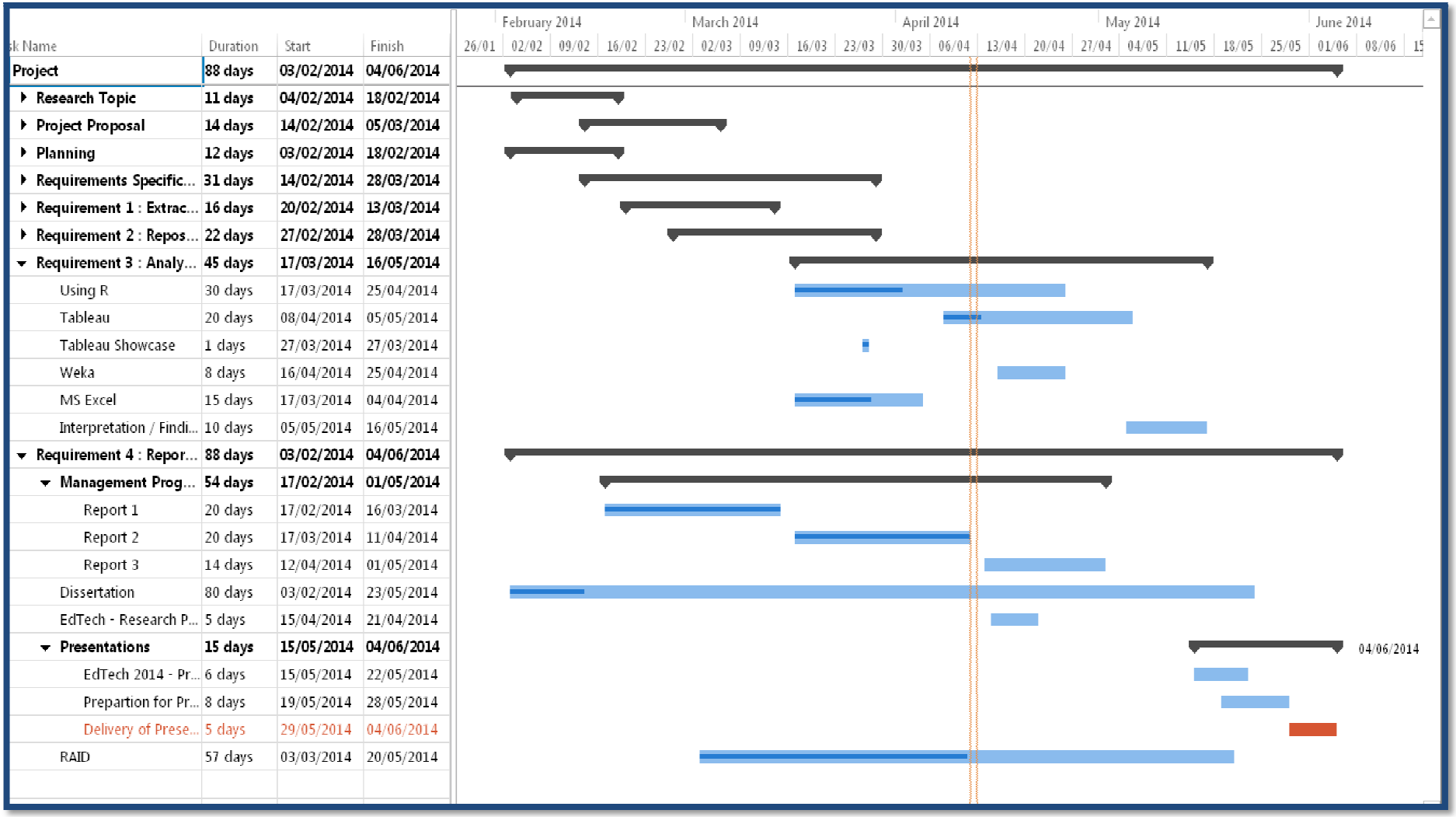
The assumptions logged during the planning stage have been revised to take into account the breakdown of the tasks needed in each of the requirements. The assumptions incorporated hardware, software, learning resources and the application of models to produce visualisation of the data.

10.4 Project Dependencies

A new addition to the log was added, due to the need to revise the Gantt chart. To update the chart required a new install of the software on an alternative laptop, due to the trial period had expired. No other dependency was affected during this period. The dependencies are listed on the RAID Log.

APPENDIX

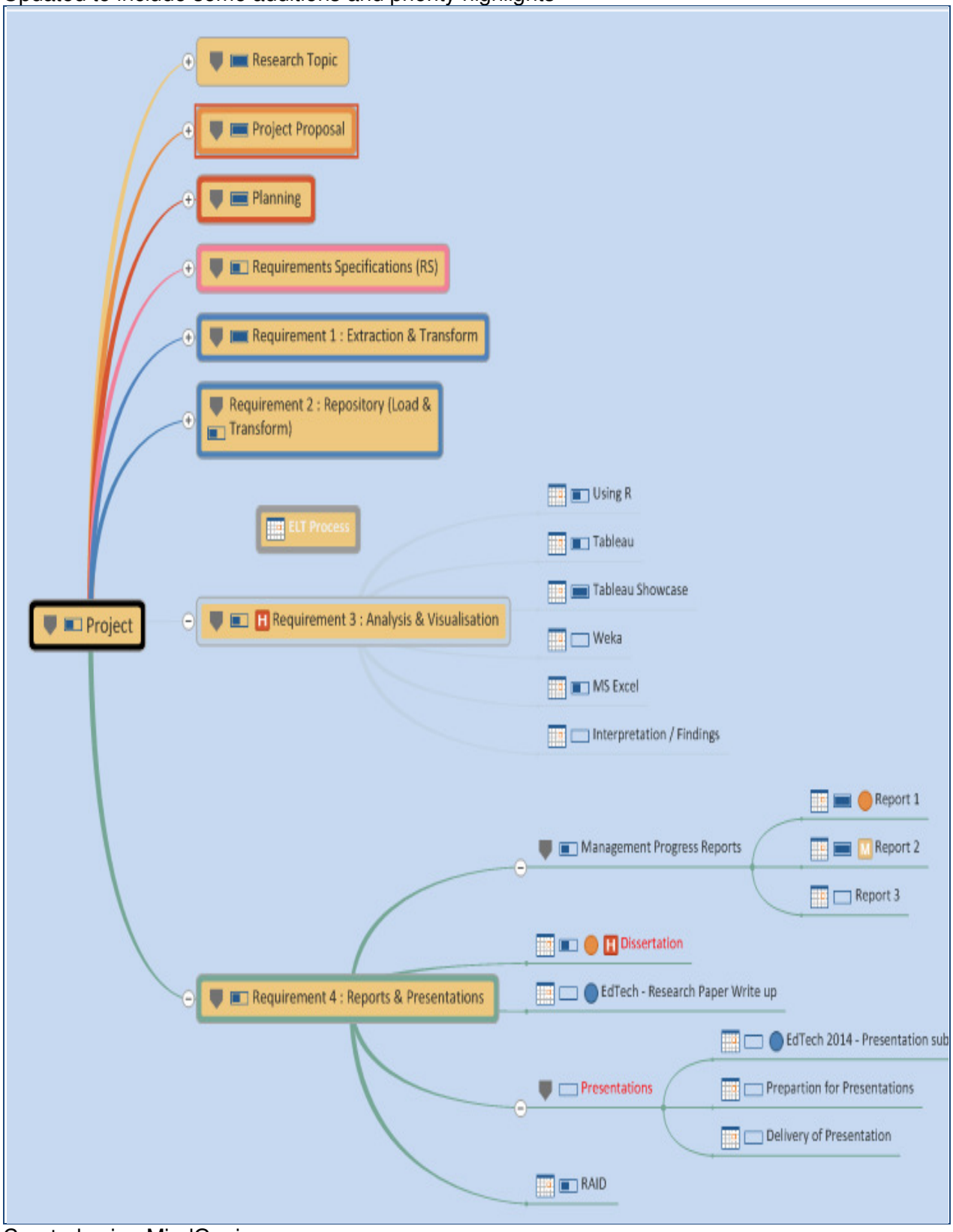
Plan schedule revised, due to extension in dates.



sk Name	Duration	Start	Finish
Project	**88 days**	**03/02/2014**	**04/06/2014**
Research Topic	**11 days**	**04/02/2014**	**18/02/2014**
Project Proposal	**14 days**	**14/02/2014**	**05/03/2014**
Planning	**12 days**	**03/02/2014**	**18/02/2014**
Requirements Specific...	**31 days**	**14/02/2014**	**28/03/2014**
Requirement 1 : Extrac...	**16 days**	**20/02/2014**	**13/03/2014**
Requirement 2 : Repos...	**22 days**	**27/02/2014**	**28/03/2014**
Requirement 3 : Analy...	**45 days**	**17/03/2014**	**16/05/2014**
Using R	30 days	17/03/2014	25/04/2014
Tableau	20 days	08/04/2014	05/05/2014
Tableau Showcase	1 days	27/03/2014	27/03/2014
Weka	8 days	16/04/2014	25/04/2014
MS Excel	15 days	17/03/2014	04/04/2014
Interpretation / Findi...	10 days	05/05/2014	16/05/2014
Requirement 4 : Repor...	**88 days**	**03/02/2014**	**04/06/2014**
Management Prog...	**54 days**	**17/02/2014**	**01/05/2014**
Report 1	20 days	17/02/2014	16/03/2014
Report 2	20 days	17/03/2014	11/04/2014
Report 3	14 days	12/04/2014	01/05/2014
Dissertation	80 days	03/02/2014	23/05/2014
EdTech - Research P...	5 days	15/04/2014	21/04/2014
Presentations	**15 days**	**15/05/2014**	**04/06/2014**
EdTech 2014 - Pr...	6 days	15/05/2014	22/05/2014
Prepartion for Pr...	8 days	19/05/2014	28/05/2014
Delivery of Prese...	5 days	29/05/2014	04/06/2014
RAID	57 days	03/03/2014	20/05/2014

Created using MindGenius

APPENDIX
Updated to include some additions and priority highlights



Created using MindGenius

Appendix 4- 5 Management Progress Report 3

Management Progress Report 3

An Exploratory Study into the Association between School Expenditure Levels per Student and Overall School Performance in High Stake Examinations

Report Period 12/04/2014 to 04/05/2014
Release: 1.0
Date: 04 April 2014
Author: Louise Blake

1. Report History

1.1 Document Location
This document is valid on the day of submission via Moodle 04/05/2014.

1.2 Revision History

Revision date	Author	Version	Summary of Changes	Changes marked
04/05/2014	Louise Blake	1.0	Third Management Progress Report	Not applicable

1.3 Approvals
This document requires the following approvals:

Name	Title	Date of Issue	Version
Louise Blake	Author	04/05/2014	1.0

1.4 Distribution
This document has additionally been distributed to:

Name	Title	Date of Issue	Status
Dr. Ioana Ghergulescu	Lecturer	04/05/2014	submitted

2. Highlight Report from 12/04/2014 to 04/05/2014

3. Highlight Report Purpose

The purpose of this report is to provide details to the Project Board with a summary of the status of the project and is used to monitor progress. The Project manager uses the Highlight report to advise the Project Board of any potential problems or areas where the Project Board can help.

This is the final management report before the submission of the Dissertation.

4. Achievements Completed during the Period

The progress in the project during this period consists of:

- Continued progress on the data analysis and visualisation tasks.
- Dissertation template headings received and existing write up reviewed and revised to bring in line with this documentation format.
- Commencement predictive models construction using Weka and R.

During this period, no documents were prepared and submitted to Moodle.

5. Budget Status

No change to the budget.

6. Schedule Status

The project is still on schedule.

There have been no amendments to the submission date. The key milestone events dates were not amended during this time with one exception the third Management Report submission date was revised, to the current date 04/05/2012. There is no updated version of the project plan in this report, as the extension does not effect on going work schedules.

7. Activities during the Period

Status	Analysis Tool	Task	Comment
In Progress	R	7. Document work complete to date, example:	Documentation prepared partly revised in the format of the template received for the Dissertation.
Complete	R	8. Descriptive statistics - Histogram and Scatter plots	
Complete	R	9. Attributes for Decision Tree model	Reassessed attributes to use for Decision Tree and Clustering
In progress	R	10. Predictive analytics, explore the associations within the data (correlation) and modelling relationships within the data. Predictive models; Correlation, Regression analysis.	
Incomplete	R / Weka	11. Utilise a number of technical algorithms: Decision Tree	Explored Decision Tree options using party package and rpart package and looked at Decision Tree options in Weka. Need to build on this in next phase.
Incomplete	Tableau	12. Sourced specific geo codes for utilising in Map visualisation	Reconsidering the merit of a map and initial proposal
WIP[26]		13. Dissertation - Writing in progress	
Updated		• RAID [27] Log	

[26] WIP = Work In progress
[27] RAID = Risks, Assumptions, Issues and Dependencies

8. Planned Work for Next Period (to 27/05/2014)

Continue with progress, Requirement 3 : Analysis & Visualisation continue
Commence and complete Requirement 4

Analysis Tool	Comment
R	Recheck existing explorative data analysis documentation
	Research required for interpretation and analysis using Weka
Weka /R	Decision Tree options (Machine Learning method) - Compare R and Weka results and visuals
Weka	Decide & Commence work on the use of Weka for Cluster analysis (Descriptive Data Model)
	Write up interpretations of Weka Analysis
Tableau	Document existing analysis and visuals created and suitable for inclusion.
Write up	Interpretation - review of work done to date Continue and Complete Write up of Dissertation Write up revise in the format of Template
Conclusion	To do
Proofing	Completed write up - needs to be proofed & ready by an independent person
Print/ CD	2 print outs plus CD of content.
Presentation	Prepare and practice (Schedule required & Time off work to be arranged)

9. Project, Risks, Assumptions, Issues and Dependencies

In the RAID Log details are recorded and below are a summary of the contents of the Log.

9.1 Project issues

There were currently have 8 issues on our project issue log 7 have been resolved and 1 open. This item is now closed. Open issue, related to a proposal to Integrate R with SQL. This item is deemed as unobtainable in the final stage of the project due to time. It does not affect the analysis or hinder the progress of the project. It is an item which would be desirable for a more automated system and could be worked on if the project were to develop to a commercial level. There are external commitments added to the project issues, and it is added as a high issue, as it could affect the project completion outcome.

9.2 Project Risks

To date the Risks log contains 6 items. There a change is status to one Risk during this reporting period. Due to the obtainment of work commencing in May, the Time Risk needs to be monitored for the completion of the project and presentation schedule. The existing Gantt chart in place is still on schedule. External commitments is added to the project risks, as a high risk as it could affect the project completion outcome. The other risks relating to technology; a Hardware and data backup is in place, the required software is software installed and available. Currently the high risk item relates to time, there are no other high risk items to cause concern.

9.3 Project Assumptions

The assumptions logged during the planning stage have been revised to take into account the breakdown of the tasks needed in each of the requirements. The assumptions incorporated hardware, software, learning resources and the application of models to produce visualisation of the data. An additional assumption is added; time off work requirement.

9.4 Project Dependencies

Addition dependency added to the log includes an external organisation, and attendance at a presentation will be dependant of co-operation of this organisation. No other dependency was affected during this period. The dependencies are listed on the RAID Log.

Appendix 4- 6 Entity Relationship Diagram