BC2402 ANALYTICS I:  
VISUAL & PREDICTIVE TECHNIQUES

Machine Learning Proof of Concept (POC)   
for The Economist Intelligence Unit  
Text

Description automatically generated

AY21/22 SEM 1 | SEMINAR 2, TEAM 7  
Instructor: Prof Vivek Choudhary

| Matric No. | Group Member Name |
| --- | --- |
| U2010230J | Gabriella Angelina Lim |
| U2022593D | Justyn Phoa Zairen |
| U2021729K | Leow Ken Hing Bryan |
| U2022791E | Rodmond Tan Le Meng |

**Table of Contents**

[**EXECUTIVE SUMMARY**](#_heading=h.gjdgxs)

[Problem Statement](#_heading=h.ibqvs5ngum6o)

[Our Solution and results](#_heading=h.lqn0ijg19y1a)

[Future Recommendations](#_heading=h.qxdz6l9u9vkm)

[**1. Introduction**](#_heading=h.w2a9nh1l6ljc) **1**

[**1.1. Business Problem**](#_heading=h.azqw9i29r1d7) **1**

[**1.2. Significance of Business Problem**](#_heading=h.u9a9ug95ld1q) **1**

[**1.3. Justification of Business Problem**](#_heading=h.ewx4kzagogzs) **1**

[1.3.1 Diversity](#_heading=h.swzzn84kh42z) 2

[1.3.2 Predictability](#_heading=h.fs75rusrekeq) 2

[1.3.3 Causality](#_heading=h.sh1qh8m6ov0v) 3

[**2. Opportunity for Machine Learning (ML)**](#_heading=h.raz4dp14wslm) **4**

[**2.1. How ML Differs from Non-ML Forecasting**](#_heading=h.lfhrskt8jszb) **4**

[**2.2. Case Studies in ML-based Forecasting and Intelligences**](#_heading=h.4cc73kng0g8o) **4**

[2.2.1.1 Comparing statistical vs ML approach](#_heading=h.vs14564ug07u) 5

[2.2.1.2 Model and Results](#_heading=h.dbjv8dabr783) 5

[2.2.2.1 ML vs Non-ML methods](#_heading=h.1l1mpqjztqaa) 6

[2.2.2.2 Model and Results](#_heading=h.9kl1zfedfu7r) 6

[**2.3. Specific Value of ML to the EIU**](#_heading=h.odo04o5xv8io) **6**

[2.3.1 Quantitative indicators](#_heading=h.1nz3kelz62tu) 6

[2.3.2 Qualitative indicator**s**](#_heading=h.zablikmi6qn3)6

[**3. Application of Machine Learning**](#_heading=h.ft7wfioqvw1r) **7**

[**3.1 Data Shortlisting**](#_heading=h.tqoueukcekq) **7**

[3.1.1 Metric Category](#_heading=h.wzapnircg3uz) 7

[3.1.2 Least Amount of Missing Data](#_heading=h.boxk9hi583td) 7

[3.1.3 Making Exceptions for Metrics of Interest](#_heading=h.6vbqb5a18h1f) 7

[**3.2 Handling Missing Values**](#_heading=h.wclcwfm0cw7u) **8**

[3.2.1 Interpolating PISA Scores](#_heading=h.wyxoj1neylsx) 8

[**3.3 PISA Scores Regression Model using Education Metrics**](#_heading=h.v5q4y2a4uhjs) **9**

[3.3.1 Exploratory Analysis](#_heading=h.nn9ljc3r8388) 9

[3.3.2 Data Preparation of Predictor Variables](#_heading=h.v5qgjutgbr55) 9

[3.3.3 Linear Regression Modelling of PISA scores](#_heading=h.ljym7nzctel4) 9

[3.3.4 CART Modelling of PISA scores](#_heading=h.2x9bqqs39s7p) 10

[**3.4 GDP Growth Regression Model using Education Metrics**](#_heading=h.wlm67y230404) **10**

[3.4.1 GDP and PISA data preparation](#_heading=h.8ytcxzisqvsp) 10

[3.4.2 Continuous CART Model](#_heading=h.gd71r2dmyu5q) 10

[**3.5 GDP Growth Regression Model using Education & Economic Metrics**](#_heading=h.kjk5mqqjg18j) **10**

[3.5.1 Data Preparation](#_heading=h.n0pftwhwwu5o) 11

[3.5.2 Continuous CART Model](#_heading=h.2y6e1fsplkkk) 11

[3.5.3 Insights](#_heading=h.5cbj52lsp2gd) 11

[**3.6 GDP Regression Model using Education & Economic Metrics**](#_heading=h.xti7ohvelnft) **11**

[3.6.1 Continuous CART Model](#_heading=h.9fy2thdgy3lg) 11

[3.6.2 Linear Regression Model](#_heading=h.y1lpp6v0354g) 11

[3.6.3 Insights](#_heading=h.ew0voq41hxvc) 12

[**3.7 Data Preparation for Variables (% Change)**](#_heading=h.ch5b1ne0fls) **12**

[**3.8 GDP Growth Regression Model using % Change Inputs**](#_heading=h.gjlfwihybywe) **12**

[3.8.1 Continuous CART Model with Economic Metrics](#_heading=h.obepl6sigjba) 12

[3.8.2 Continuous CART Model with Education and Economic Metrics](#_heading=h.li612i99ccmy) 12

[3.8.3 Linear Regression Model with Education and Economic Metrics](#_heading=h.bny9k7ttnqsn) 12

[3.8.4 Insights](#_heading=h.solas4tpdpzs) 13

[**3.9 GDP Growth Classification Model using % Change Inputs**](#_heading=h.r2m3ympp3uqo) **13**

[3.9.1 Data Preparation](#_heading=h.ajtmb4h276fu) 13

[3.9.2 Categorical CART Model with Education and Economic Metrics](#_heading=h.82yrxfdtm4sv) 13

[3.9.3 Insights](#_heading=h.fmvdnygn8uu9) 13

[3.9.3.1 Undersampling](#_heading=h.12nmsi8q17a8) 13

[3.9.3.2 Oversampling](#_heading=h.vu6iqnjo7br4) 14

[**3.10 GDP Growth Classification Model using % Change Inputs (SMOTE)**](#_heading=h.kztjut8asf7g) **14**

[3.10.1 Categorical CART Model with Economic Metrics (SMOTE)](#_heading=h.wypt06klfqz9) 14

[3.10.2 Categorical CART Model with Education and Economic Metrics (SMOTE)](#_heading=h.hjzxa5q671ic) 14

[**3.11 Model Insights and Conclusions**](#_heading=h.mmdfngbdp6ht) **14**

[**4. Pilot Studies and Future Recommendations**](#_heading=h.wxlxs656hizn) **15**

[**4.1 Business Opportunity Understanding**](#_heading=h.s8vbfusag6b6) **15**

[**4.2 Data Mining and Collection**](#_heading=h.u98el5x8jabh) **15**

[4.2.1 Importance of data collection](#_heading=h.s18rtvgt1j4) 15

[4.2.2 Considerations for data collection](#_heading=h.iz9t9o1v5z9q) 15

[4.2.3 Data Storage](#_heading=h.xviyo2rm8eo) 15

[4.2.4 Challenges](#_heading=h.42wljlbwplnt) 16

[4.2.5 Relevance to Our Solution](#_heading=h.hvzj4rdls33d) 16

[**4.3 Data Cleaning and Preparation**](#_heading=h.gd85i3vb6f6k) **16**

[4.3.1 Importance of data cleaning](#_heading=h.68axdvltocv3) 16

[4.3.2 Process of cleaning](#_heading=h.7jpxtzilj6wd) 16

[4.3.3 Data Processing](#_heading=h.v4bxtme2vp5s) 16

[4.3.4 Challenges](#_heading=h.ezg03hd4b0km) 17

[4.3.5 Relevance to Our Solution](#_heading=h.vn04g729k6ns) 17

[**4.4 Data Exploration**](#_heading=h.w296nvprzttc) **17**

[4.4.1 Introduction](#_heading=h.1och65m0qv7w) 17

[4.4.2 Importance of EA](#_heading=h.sckfel1g49tv) 17

[4.4.3 Tools for EA](#_heading=h.43f81mdkh8t) 17

[4.4.4 Challenges](#_heading=h.eb8c7b6xkh2z) 18

[4.4.5 Relevance to Our Solution](#_heading=h.2iebccbvhtod) 18

[**4.5 Machine Learning Modelling**](#_heading=h.x605rotye7fz) **18**

[4.5.1 Introduction](#_heading=h.a2xp5muu9vug) 18

[4.5.2 Generating ML Models](#_heading=h.pm00iobferno) 18

[4.5.3 Model Evaluation](#_heading=h.8vf3biyvif11) 18

[4.5.4 Challenges](#_heading=h.getk9ufri9vi) 19

[4.5.5 Relevance to Our Solution](#_heading=h.c70ash3th03y) 19

[**5. Conclusion**](#_heading=h.l40t7pw2002s) **19**

[**6. References**](#_heading=h.i3w7fon4xhl9) **20**

[**Appendix A: Evaluating Model Fitness**](#_heading=h.b40j3brd2i83) **23**

[Evaluating Model Fitness: “Explainability” or R2](#_heading=h.5xti6ltbz8bc) 23

[Evaluating Model Fitness: “Error Rating” or RMSE](#_heading=h.v96z86hpue9g) 23

[Train Test Split](#_heading=h.74c24xslw3bk) 24

[**Appendix B: Exploratory Analysis**](#_heading=h.giatg4mcjtwl) **26**

[**Appendix C: Technical Details of Linear Regression**](#_heading=h.3auq47f26ss4) **44**

[Handling Missing Values](#_heading=h.svvoaahaq722) 44

[Pitfalls in Linear Regression](#_heading=h.22tnbuj0emv5) 44

[Required Assumptions in Linear Regression](#_heading=h.e7psdjjhq429) 44

[Influential Outliers](#_heading=h.kiqrt51jc92f) 45

[Checking for Required Assumptions & Influential Outliers](#_heading=h.1a6zn0vqetmi) 45

[Multicollinearity or “related variables”](#_heading=h.75kvfxkka0) 48

[Reducing collinearity using VIF and AIC](#_heading=h.4mx0qz2wolbw) 48

[Backwards Elimination](#_heading=h.85r6tmjcy9nb) 49

[**Appendix D: Technical Details of CART**](#_heading=h.8lvam8nqpdzf) **50**

[Technical Introduction to CART](#_heading=h.13zdime5gat7) 50

[Developing the CART](#_heading=h.krd00uek3ui5) 50

[Phase 1: Growing the (decision) tree to the maximum](#_heading=h.7crad2s9dqzh) 50

[Phase 2: Pruning the (decision) tree to the minimum](#_heading=h.eqywtomkne8n) 51

[**Appendix E: Oversampling Techniques**](#_heading=h.v9kw24lb1wmw) **52**

[**Appendix F: Pilot Study - Data Science Life Cycle**](#_heading=h.i1to2yjy9t1z) **54**

[Data Science Problems](#_heading=h.pb5krlyol17i) 54

[Data Normalisation](#_heading=h.j74xayd6af4p) 54

[Introduction](#_heading=h.mdzez1dlskvr) 54

[Methods of Normalisation](#_heading=h.wh44tjwrz28z) 54

[Exploratory Data Analysis Techniques](#_heading=h.j4d2kncxm8in) 56

[Classification and Dimension Reduction](#_heading=h.j61bdw5a7ls6) 56

[Univariate Visualisation](#_heading=h.gg4x70zh5vpy) 57

[Bivariate/Multivariate Visualisation](#_heading=h.7xoy1tv5tfnq) 58

[Machine Learning Modelling](#_heading=h.gcap15s1d128) 60

[ML Model Performance Metrics](#_heading=h.cc080rj3tje5) 61

# EXECUTIVE SUMMARY

### Problem Statement

Predicting economic recessions poses a challenge for economists as it is difficult to forecast business cycles, even for experts with decades of professional experience like the EIU. In this report, we present the ways in which we found current forecasting methods and results inaccurate and the significant implications this could cause for EIU, such as a loss in consumer’s trust. This motivated us to look into ways to improve prediction models and its accuracy for future assessment.

### Our Solution and results

ML models are able to efficiently manage large volumes of data, deliver greater accuracy, correlate data from various economic sources and respond quickly to dynamic changes, all without human assistance. Our team explores possible uses of machine learning using linear regression and Classification & Regression Tree (CART) models, including indicators of education as one of our influencing variables to predict whether a country’s GDP growth will be positive or negative, as a measure of economic activity. Our model has an overall accuracy of 57%, and managed to predict negative GDP growth with 42% accuracy, and positive GDP growth with over 70% accuracy. This is about 17% more accurate compared to EIU’s past predictions which only uses non-machine learning models and do not include education as a metric.

### Future Recommendations

In conclusion, we see the supplementation of machine learning models and education assessment to EIU’s current models as greatly beneficial to EIU’s forecasting capabilities. It is crucial to note that it is an extremely basic ML model. In spite of this, we were able to achieve positive results, which further goes to show the immense potential the inculcation of machine learning has to offer to EIU’s forecasts. With the use of more complex and advanced models, we are confident that it will yield significantly more reliable and accurate results .This has been proven in several academic literature. Therefore we recommend EIU to conduct pilot studies in addition to our research, and start harnessing the power of machine learning to add value to its flagship product of country forecasts.

# 1. Introduction

## 1.1. Business Problem

This report seeks to introduce new techniques to improve the accuracy of forecasts provided by the Economist Intelligence Unit’s (“EIU’s”) Country Report on economic growth, which is measured as the percentage change in a country’s Gross Domestic Product (“GDP”). We achieve this through the additional inclusion of educational metrics, and application of machine learning techniques.

## 1.2. Significance of Business Problem

Forecasts are crucial for all economic and business activities in implementing policy changes, as well as in planning operating activities (Elliott & Timmermann, 2008). The possibility of a prediction being inaccurate creates a serious dilemma for stakeholders and poses a threat to EIU’s credibility.

In addition, by showing that education can predict a country’s economic growth, we further solidify the narrative of the importance of education – not just for human capital development, but also for the country’s economic well-being in the form of an economic investment.

## 1.3. Justification of Business Problem

The EIU’s forecast accuracy for GDP growth comes into question when referencing Argentina’s 2015 Country Report. Their optimistic forecasts provide for an increasing trend in real GDP growth, as illustrated in Figure 1. Actual GDP growth, however, fell short of these expectations by a considerable amount, showcasing a fluctuating trend that alternates between positive and negative economic growth.

~~~~

*Figure 1: Forecast GDP Growth of Argentina from 2015-2019.*

*Source: EIU Argentina Country Report (2015, pp. 7)*

~~~~

*Figure 2: Actual GDP Growth of Argentina from 2015-2019.*

*Source: The World Bank Database*

This **huge discrepancy** between actual and predicted GDP growth is an area of concern worth exploring. Inaccurate forecasts, regardless whether they are underestimates or overestimates, incur additional costs for stakeholders (Ketchell, 2016).

In gauging economic growth, the EIU uses the following indicators to justify their forecasts:

* Industrial production and exports
* Consumer confidence, wages
* Government macroeconomic policy
* Economic conditions in neighbouring countries.

These indicators have their limitations when compared with education. We categorise them into the following three reasons:

### 1.3.1 **Diversity**

The current indicators, while strongly related to a country’s economic growth, are cyclical (Khan, 2011) and can change with varying economic outlooks over time. This is because they are largely influenced by policymakers, which in turn cause consumers and producers to react in certain ways.

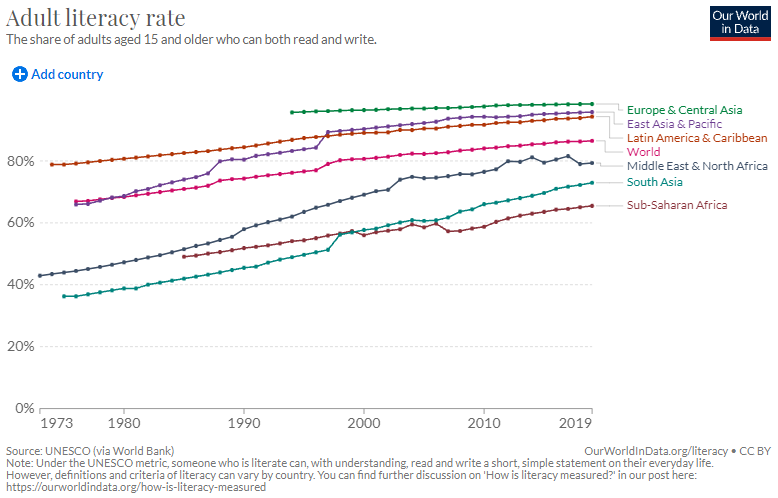
Education tackles this issue from a fundamentally different perspective, allowing us to observe the changes in the quality and productivity of labour (Chevalier & Harmon, 2004). This change is systemic, and much more permanent than the cyclical changes in policy stance.

### 1.3.2 Predictability

In making the predictions based on the selected indicators, the EIU makes huge assumptions about the government’s macroeconomic policies and the direction they will take the country, for example assuming that the Argentine government will address legal and regulatory concerns and persist their expansionary fiscal policies.

These are very risky assumptions to make since they are grounded largely in the human opinions of the analysts, who may have their own biases and blindsides (Hirshleifer, 2020) and may not come to fruition due to a multitude of factors on the policymakers’ ends.

In contrast, education rates are much more predictable in their trend. As an example, observing data from the world bank, we find that adult literacy rates consistently increase over the years, barring dips in certain periods.



*Figure 3: Adult Literacy Rate over Time, showing a consistent trend.*

*Source: Max Roser and Esteban Ortiz-Ospina (2016), Our World in Data.*

Owing to these predictable trends, we posit that our predictions for education and its projected rate of growth will be safer to make compared to those the EIU currently makes.

### 1.3.3 Causality

Lastly, there is a clear empirical correlation between education and economic growth (Marquez-Ramos & Mourelle, 2019). This relationship goes beyond correlation – there is a **causal relationship** between education and a country’s economic development. Specifically, Becker (1962) and Schultz (1963) propose education as an investment in human capital in its ability to raise productivity.

# 2. Opportunity for Machine Learning (ML)

## 2.1. How ML Differs from Non-ML Forecasting

Machine learning makes use of past data and statistics to “learn” and build a model for prediction. The models created using machine learning range from the simple uni/bivariate models, to more complicated ones that utilise dozens of variables across numerous databases. However, the core of machine learning is consistent – being able to efficiently handle big data to generate predictions beyond the capabilities of any human being manually sifting through the same dataset for insights.

Non-machine learning techniques of prediction are not as capable in finding the more in-depth statistical insights compared to machine learning, especially in the cases with large datasets. This is due to their generally interpretation-oriented nature. One thing which differentiates non-machine learning from machine learning is that they are programmed to perform a task, instead of learning to perform the task. This characteristic makes them more inflexible when dealing with intensive data. Moreover, non-machine learning techniques involve training a model based on historic training data and evaluating the resulting model against incoming data. (Bhatia, R., 2018) This may not be as feasible in the long run as the environment is in constant change.

Despite being more capable in processing huge amounts of data, machine learning may not be as effective as non-machine learning models in cases where variables are difficult to quantify. We will look through different case studies below to further explore the differences between the two types of forecasting.

## 2.2. Case Studies in ML-based Forecasting and Intelligences

**2.2.1 Forecasting Economic Recessions (Cicceri et al., 2020)**

Predicting economic recessions continues to pose a challenge for economists (Drautzburg, 2019) as it is difficult to forecast business cycles. For example, right before the 2001 recession, the median forecaster in the Survey of Professional Forecasters expected a 2.5% growth for U.S. GDP. Before the Great Recession, economists were also expecting GDP to grow 2.2% over the next year, and in both instances these predictions turned out to be terribly wrong. This is because recessions (as a result of pseudo-random business cycle fluctuations) are seen as “episodic and non-linear”. In this case study, Cicceri et al. implements several ML techniques to improve forecasting accuracy, using a Linear Regression Model and a Nonlinear Autoregressive with exogenous variables (NARX) model.

#### 2.2.1.1 Comparing statistical vs ML approach

Predictive statistical models are often unable to predict the outcome of a recession in a prompt and efficient manner, as it is often the case that they are released too late, and offer too many uncertainties for their economic forecasts – notwithstanding the fact that they are often subject to many revisions. For example, the National Bureau of Economic Research’s (US) methodology requires 2 consecutive quarters with negative growth to announce the beginning of a recession.

ML models are able to manage large volumes of data, and provide better guarantees in terms of accuracy even when faced with small volumes of observations and a large number of variables. ML models are able to correlate data from various economic sources and respond quickly to dynamic changes, without human assistance.

#### 2.2.1.2 Model and Results

Using various Italian economic predictors from 1995 to 2019 (taken from ISTAT database), the paper used 7 regression estimation techniques to predict GDP growth and recession periods. The results indicate with a great majority that ML models outperformed the traditional statistical models, having a higher degree of explainability and accuracy (see Appendix A). In particular, the NARX model was able to correctly forecast not only the 3 recessions that occurred, but also the strength of these events during that time period.

**2.2.2 Forecasting Business to Business Sales (Rohaan et al., 2022)**

In a business-to-business environment, forecasting future demand is crucial to the production and supply process, and most companies are using traditional forecasting methods based on past sales and cyclical forecasting. Rohaan et al. posits there exists a valuable opportunity to solve this business problem through ML and Natural Language Processing (NLP) via the form of processing Requests for Quotations (RFQs). RFQs are requests for a quote for spare parts (usually by email) and do not necessarily result in a sale. The case study is carried out at a large after-sales service and maintenance provider, which receives a large volume of RFQs, yet only 17% of which becomes an actual sale.

#### 2.2.2.1 ML vs Non-ML methods

Currently the way of processing these requests is manually through the service and maintenance provider, who then checks through the emails one by one. When the number of RFQs gets large, there were numerous instances of significant delays in replying the queries which likely affected sales negatively. Through the use of machine learning, particularly through text mining, a model was used to predict the likelihood of a sale from the RFQs which allows sorting according to priority.

#### 2.2.2.2 Model and Results

The ML approach (supervised ML and NLP) in the paper was shown to have an 155.3% increase in performance as compared to manual handling, allowing the organisation in question to generate more sales by prioritising the processing of RFQs with higher sales potential.

## 2.3. Specific Value of ML to the EIU

#### 2.3.1 Quantitative indicators

Quantitative indicators such as economic growth, inflation, and exchange rates, for example, will be able to benefit from the predictions generated by regression, decision-tree, and neural network models. These predictions can either be numeric or categorical, serving as a second opinion to corroborate or question an analyst’s manual predictions. State-of-the-art models today include gradient boosting models such as XGBoost (Natekin, et al., 2013), and advanced time series models such as ARIMA specialise in providing time-dependent predictions (Bontempi, et al., 2013), potentially assisting the EIU greatly in preparing its quarterly forecasts.

#### 2.3.2 Qualitative indicators

On the other hand, qualitative indicators such as political stability and international relations will be able to benefit from NLP techniques by scraping text data from social media and news outlets to aggregate sentiments from journalists, citizens, and the like. Image recognition techniques can also be applied to geotagged images to provide analysts fresh perspectives on the country’s situation.

In this paper, we focus on value-adding to EIU’s economic growth forecasts using two specific machine learning models: linear regression models and Classification and Regression Tree (CART) models.

# 3. Application of Machine Learning

## 3.1 Data Shortlisting

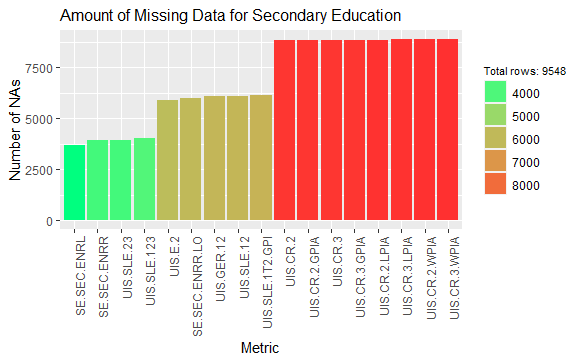
With over 4000 education-related metrics in the World Bank Database (Teixeira, P.N., 2014), our first step was to shortlist a few that we could concentrate on. We performed this initial selection based on the following criteria:

### 3.1.1 Metric Category

We made sure to select a variety of aggregated, nation-wide metrics from each category in order to paint a more holistic picture of a country’s educational landscape (for more details, refer to the Data Dictionary).

### 3.1.2 Least Amount of Missing Data

Having a large amount of data is important for machine learning (Help Net Security, 2019). It is hence important that we also select metrics that would give us the most amount of raw data to work with.



*Figure 4: missing data for each metric in the Secondary Education category, showing which metrics (in green) have the least missing data. Source: RSCRIPT\_DATA\_CLEANING.R*

For each metric category (see 3.1.1), we visualised the amount of missing data for each metric, and selected the best from each category.

### 3.1.3 Making Exceptions for Metrics of Interest

However, such a method of preliminary data shortlisting is flawed since certain metrics will inherently have less data points due to their difficulty of collection, such as test scores (educational outcomes) compared to, say, enrolment rate (educational inputs). Thus, we decided to also shortlist assessment scores from the Programme for International Student Assessment (“PISA”), a standardised test format representative of a given country’s educational outcomes in terms of math, science and reading. Other metrics such as completion rate and percentage of STEM graduates were also shortlisted at this stage for their posited impact on economic growth.

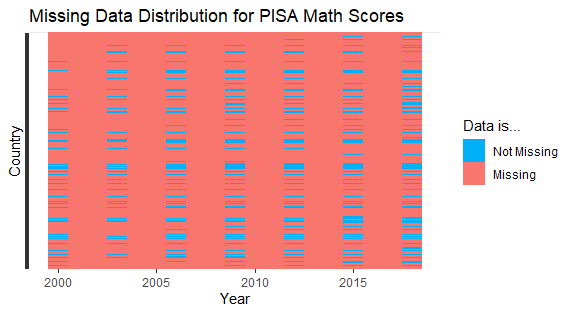
## 3.2 Handling Missing Values

Despite our best efforts to ensure data quality and volume in Section 3.1, missing values will still be inevitable in real-world datasets. There are three different ways to handle them (Chew, 2020).

* Using a model with automatic missing value handling e.g. CART.
  + Most other models, however, will ignore rows with missing data, creating issues with data volume (see 3.1.2). For these models, we will need to handle the missing values manually.
* If missing-at-random, replace with the mean of the dataset
* If missing systematically, replace with the mean of the relevant subgroup or with another model.

### 3.2.1 Interpolating PISA Scores

There is a substantial difference between the number of missing PISA values and that of the other indicators. Values are missing systematically across time and space since PISA is administered every 3 years, and not every country participates every year.



*Figure 5: distribution of missing PISA Math scores, showing systematically missing data. Source: RSCRIPT\_DATA\_CLEANING.R*

This distribution made it prohibitively difficult to replace them with the mean of the relevant subgroup, since most years had no data at all, and using the country’s mean would remove the data’s time-series value. We hence first used interpolation to infer the intermediate PISA scores for countries with at least two data points, while keeping the original data points as they are. One caveat to note here is that interpolation can only work between the first and last year that the PISA test was administered – if a given country has PISA scores for 2000 and 2009, using interpolation would not be able to give us PISA scores for the years before 2000 or after 2009.

Hence, interpolation supplements our linear regression by providing higher data volume, but does not replace regression entirely as the latter is still needed to build a generalised model for extrapolation (Poncela., 2019).

## 3.3 PISA Scores Regression Model using Education Metrics

To complete the filling of missing values, we make use of supplementary models to fill in the missing PISA values before 2000 (when PISA was first administered) and after 2018 (when PISA was last administered).

### 3.3.1 Exploratory Analysis

The results of our EA are outlined in Appendix B.

### 3.3.2 Data Preparation of Predictor Variables

In order to account for population size differences, we derived government education expenditure per capita. We also lagged expenditure by 3 years to account for the delayed effects of education spending on (education quality and hence) education outcomes. We also observed that expenditure per capita has a skewed relationship with PISA scores, thus we included log(expenditure) to improve its linear relation to the response variable, satisfying assumption #1 of linear regression (see Appendix C).

### 3.3.3 Linear Regression Modelling of PISA scores

Linear Regression is an ML model whose aim is to find a mathematical equation for a numeric variable Y as a function of X variables, predicting Y when only the Xs are known (Prabhakaran S., 2017).

Using the data in Section 3.3.1, we generated a linear regression model that would predict PISA scores based on other education metrics. One problem we faced was the high percentage of missing values for predictor variables as outlined in Section 3.1.2, especially for variables we shortlisted for qualitative reasons. While it is standard advice to drop features having more than 60% missing values (Swalin., A., 2018), given the large size of our original dataset, we set an arbitrary threshold of 80% since 20% would still give us a reasonable amount of data to work with. We performed a train test split (see Appendix A) and generated a model on the testset. After dropping those features, we dropped variables that did not contribute meaningfully to the model via backward elimination, then dropped variables that were multicollinear to give us a final model with an error rating of 24-28. Refer to Appendix C for a technical discussion of this variable selection process, and Appendix A for an appreciation for model fitness evaluation.

### 3.3.4 CART Modelling of PISA scores

A Classification and Regression Tree, or CART, is another machine learning model that we used in our predictions. It automatically handles missing values (see 3.2.1), and does not require any assumptions on the data to work (Ojha, 2017). Using CART, there was no need to omit any columns or rows, and fitting this model on our data gave us an error rating of 15-21. CART’s performance is significantly better than that of linear regression, even in its worst case versus the latter’s best case. We hence proceed to use CARTl to populate PISA scores. See Appendix D for a technical discussion on CART.

## 3.4 GDP Growth Regression Model using Education Metrics

Since we have now obtained a relatively populated dataset from the modelling in Section 3.3.3, we can proceed to predict GDP growth using education metrics.

### 3.4.1 GDP and PISA data preparation

Since our objective is to use previous years’ data to predict the current/next year’s growth, we lagged the GDP data by 1 year such that the data retrieved in 2019 would correspond to the GDP growth in 2020. In this sense, it would have much more value to the EIU as well.

Additionally, since PISA is administered on youths who are between 15-16 years old, the effects of these educational outcomes would not manifest itself until several years later, hence we chose to lag our PISA variables by 3 years as well, to account for this delay. While it can be argued that this assumption of delayed impact should not be generalised to 3 years (since the average age at which adults start work in each country is different), it is also true that PISA scores over the years is highly correlated with itself, thus lagging it by 3 years can be said to be generally representative of the delayed effect (although in reality might take longer to manifest). Another factor to take into account here is that lagging such variables will decrease the size of our dataset, which would undermine the robustness and goodness-of-fit of our model.

### 3.4.2 Continuous CART Model

We fitted our dataset onto a continuous CART model, which yielded disappointing results, having an error rating of 5 (see Appendix A). While this might seem small, do keep in mind that this represents GDP growth annually, and an error margin of 5% would make the model more or less unusable.

## 3.5 GDP Growth Regression Model using Education & Economic Metrics

From 3.4.2, it is clear that using education alone to predict GDP growth would not be enough to yield accurate results. Formal education is merely one aspect of human capital, which itself is one aspect of GDP growth (Swalin., A., 2018). Other aspects of human capital (e.g. life expectancy) and metrics measuring macroeconomic policies (e.g. government expenditure, trade), as discussed in Section 1.3, were not accounted for in this initial model.

Hence, our methodology shifts from showing how education metrics alone can accurately predict GDP growth, to how education metrics can ***value-add*** to existing ways of predicting GDP growth in terms of enhancing explainability and accuracy.

### 3.5.1 Data Preparation

We used the data previously cleaned in Section 3.4.1, and added on several key economic indicators that have been discussed in previous literature to have high correlation to economic growth and low collinearity, namely: life expectancy at birth, value of trade as a percentage of GDP, gross capital formation as a percentage of GDP, and total government expenditure as a percentage of GDP.

### 3.5.2 Continuous CART Model

We fitted our dataset onto a continuous CART model, which performed better than the model in Section 3.4.2, having an error rating of 3 (RMSE - refer to Appendix A). While this is a significant improvement from previous predictions, we believe that an error margin of 3% would still be too large for the model to have any real value-add to the EIU.

### 3.5.3 Insights

Perhaps a reason why our model is not performing well is due to the fact that GDP growth is represented in percent change, year on year, whereas our economic and education variables are in absolute values. Thus, predicting GDP, and absolute value, instead of GDP growth would seem to yield more promising results.

## 3.6 GDP Regression Model using Education & Economic Metrics

### 3.6.1 Continuous CART Model

Using a continuous CART Model, we tried to predict the GDP of a country given its education and economic metrics, but the model performed badly and had an error rating in the tens of billions (RMSE - Appendix A).

### 3.6.2 Linear Regression Model

This model performed worse than CART at an error rating of 100 billion (RMSE - Appendix A).

### 3.6.3 Insights

Circling back to Section 3.5.3, since using economic and education metrics to predict GDP did not yield any significant results (as shown in the above models), we can flip our perspectives and explore if calculating percentage change of the variables year on year would serve as better predictors for GDP growth.

## 3.7 Data Preparation for Variables (% Change)

To find the percentage change of any variable for a given year, we would compute the difference between the metric value in that given year and its value in the previous year, then represent it as a percentage of the base value, which is the value in the previous year. E.g. Given the PISA score for math of a given country in 2008 is 350, and the PISA score for math of that same country in 2009 is 385, then the percentage change for PISA math score for 2009 would be (385-350)/350 = 10%.

We used this methodology for all the education and economic metrics highlighted in Section 3.5.1, and we carried out the lagging of specific variables that was previously outlined in Section 3.4.1. From this data processing, we obtained the relevant data that we will be using to build the models in the subsequent sections.

## 3.8 GDP Growth Regression Model using % Change Inputs

### 3.8.1 Continuous CART Model with Economic Metrics

We fitted a continuous CART model on our dataset to predict GDP growth using the percentage change of our economic variables over the years. However, this model performed badly and predicted the same value for GDP growth for all data points.

### 3.8.2 Continuous CART Model with Education and Economic Metrics

Ee explored if adding education metrics into the mix would improve the model, by predicting GDP growth using the percentage change of both education and economic variables over the years. However, this model performed just as badly as the previous model, predicting the same value for GDP growth for all data points.

### 3.8.3 Linear Regression Model with Education and Economic Metrics

We then investigated if using a linear regression model instead of a continuous CART model would yield better results: whether it was a problem specific to the model we used, or if it was a problem relating to the data itself. The linear regression model performed badly as well, hinting that we might have to circle back to investigate the relationship of the response variable to the predictor variables.

### 3.8.4 Insights

As seen from the above models, even using percentage change of metrics as inputs did not help much with model performance. It might be the case where the basic models such as CART and Linear Regression are unable to capture the high-level complexity of the relationship between GDP growth and the economic and education inputs used.

We postulate that if we are able to reduce the complexity of the problem at hand, then the model might be able to perform better. Instead of predicting the actual values of GDP growth (E.g. +1.5% or -0.7%), we can reframe and generalise this regression problem such that it becomes a classification problem: whether GDP growth is positive or negative. By making the problem “easier”, the model might yield better results.

## 3.9 GDP Growth Classification Model using % Change Inputs

### 3.9.1 Data Preparation

Each GDP growth value would either be labelled as “positive” or “negative”, essentially transforming it from a continuous variable into a categorical variable.

### 3.9.2 Categorical CART Model with Education and Economic Metrics

We used a CART model to predict if GDP growth would be positive or negative based on the percentage change of education and economic metrics the year before. The model performed badly, making 100% positive predictions and was fundamentally unable to distinguish the difference between positive and negative samples.

### 3.9.3 Insights

Considering that the above model had a strong tendency to make positive predictions, we can postulate that maybe it was because the majority of GDP growth values in the data set are positive, which might lead to the model having this particular bias (If 90% of the data is positive, and if the model predicts positive only, then it would be correct 90% of the time). There are 2 ways to correct this data imbalance:

#### 3.9.3.1 Undersampling

Undersampling keeps all the data in the minority class and decreases the size of the majority class by deleting data in the majority class. However, one drawback of this technique is that potentially important information may be lost in the process, especially if the ratio of majority to minority is high - e.g. 10:1, meaning that we would have to drop 90% of the data in the majority class.

#### 3.9.3.2 Oversampling

Oversampling keeps all the data in the majority class and creates artificial events. One such way is to randomly duplicate sample data from the minority class, meaning that there will be multiple copies of the same instance in the rebalanced data. This however, introduces the problem of over-specificity, whereby models are overexposed to the same samples, which diminishes its generalisation capabilities (ability to be applied to “new” data).

One particular way to tackle this limitation is to use Synthetic Minority Oversampling Technique (SMOTE, refer to Appendix E), where artificial events are synthesised from existing examples. We will be using this technique to tackle the problem of imbalanced data classes that was highlighted above.

## 3.10 GDP Growth Classification Model using % Change Inputs (SMOTE)

We used 75% of the original dataset to obtain a “training” dataset, on which we would then use SMOTE to rebalance the data classifications. The model would then be trained on this dataset and its performance will be determined by testing the accuracy of the model’s predictions on the remaining 25% of the original dataset.

### 3.10.1 Categorical CART Model with Economic Metrics (SMOTE)

This model performed much better than all previous models, being able to actually distinguish between positive and negative classes to a certain degree. The model had an overall accuracy of 52%, with it being able to predict a negative GDP growth correctly 33% of the time, and predicting positive GDP growth 70% of the time.

### 3.10.2 Categorical CART Model with Education and Economic Metrics (SMOTE)

This model performed the best so far, having an overall accuracy of 57%, with it being able to predict a negative GDP growth correctly 42% of the time, and predicting positive GDP growth 70% of the time, which is a 27% increase in accuracy (due to the inclusion of education metrics as inputs).

## 3.11 Model Insights and Conclusions

We postulate that basic models such as linear regression as well as decision trees might not be able to capture the full complexity of the relationships between socio-economic factors and GDP, but by reducing this complexity by generalising the problem (transforming it into classification) and solving data issues (imbalanced data and data processing), we can see that our model has performed decently well, considering that predicting recessions have been a problem that has been confounding economists for a long time.

Further studies and recommendations to the EIU regarding the usage of our model or how they can go about implementing ML to solve similar problems will be outlined in the following section.

# 4. Pilot Studies and Future Recommendations

This section will outline the process of how the EIU can leverage our solution to implement ML, detailing each stage of the pilot study, along with any challenges that the EIU might face along the way.

## 4.1 Business Opportunity Understanding

As discussed in Sections 1 and 2, it is important that the EIU is aware of and thoroughly understands the capabilities of ML and how it functions. It is also equally as important that the EIU identifies certain areas in which they believe that ML can value-add to their operations (similar to Section 2.3).

These 2 questions are a necessary first step to categorise the problem at hand, in terms of data science problems (Appendix F), so that the EIU can narrow down the correct data to acquire.

## 4.2 Data Mining and Collection

### 4.2.1 Importance of data collection

It is without question that data is integral for ML. Models are only as good as the data from which they are built, so data collection and data mining is definitely a high priority.

### 4.2.2 Considerations for data collection

When it comes to data collection, there are several considerations, mainly: accuracy, variety, reliability, utility, and cost.

The most important consideration here is utility: how useful the data will be. In this sense, it would be pointless for the EIU to perform data collection on “junk” or irrelevant data. Rather, it would be best for the EIU to leverage on its decades of professional experience in each industry to identify certain key variables that might be relevant to the problem at hand (similar to section 3.1).

In terms of variety, the selection of variables should be as diverse as possible to prevent overexposure of the model to certain factors. Accuracy and reliability of the collected data is crucial, since having inaccurate inputs to the model will surely result in inaccurate outputs. This will depend on the data collection methods, which are directly related to costs. Hence, the EIU needs to balance this trade-off between costs and accuracy of data.

### 4.2.3 Data Storage

In a large corporation like the EIU, there is no doubt that it already has its internal database that stores relevant historical data. However, with the additional process of data collection, EIU might need to expand its data storage. One might think that we only need enough storage to contain the datasets, but in fact, throughout each stage of the data processing pipeline (cleaning, exploring, etc), there will be data byproducts that will take up storage as well. Thus it is important to consider these factors when thinking about data storage.

### 4.2.4 Challenges

With the EIU being such a large corporation, it goes without saying that the EIU will have the means to collect data for current and future years. The problem manifests itself when the EIU needs to refer to historical data that cannot be found in its internal database. As a result, the EIU will need to obtain data from third parties, which might pose problems pertaining to accuracy and cost - if obtained from a private vendor.

### 4.2.5 Relevance to Our Solution

In the case of using education metrics to predict GDP growth, we faced the problem of a high percentage of missing values when collecting data (Section 3.1). Thus, we would recommend that the EIU uses its own variables, or source more extensively for reliable data to enhance the model.

## 4.3 Data Cleaning and Preparation

### 4.3.1 Importance of data cleaning

As mentioned in 4.2.2, accurate data is necessary for accurate models. There is a need for thorough data cleaning to weed out erroneous and inconsistent data, which can throw off the entire model, and these might often manifest themselves as missing data or unwanted influential outliers (refer to Appendix C). Wrong data can drive a business to wrong decisions and poor analysis, which will cost the organisation greatly.

### 4.3.2 Process of cleaning

While much of data cleaning can be done by software, it must be closely monitored and inconsistencies reviewed, making it quite a labour-intensive and expensive process. (E.g. removal of null values, outliers) Hence, it is even more important that the data collection process (Section 4.1) is done properly, to reduce the need for data cleaning.

### 4.3.3 Data Processing

The next step of data preparation is to process the data (a.k.a. data wrangling). This can be done by updating the format or entries of the data such that the dataset is well-defined, and more easily understood. This is especially so if the raw data is unstructured (e.g. surveys, website logs, etc), and should be tabulated and formatted before any analysis. Data normalisation (refer to Appendix F) should also be done at this step if necessary, after which the data will be stored in the database, or channeled into a business intelligence tool such that further processing and analysis can take place.

### 4.3.4 Challenges

As mentioned in Section 4.3.2, the biggest challenge of data cleaning and preparation is that it is a highly labour-intensive process that requires a significant amount of time. While there is software available that can help to speed up the process, ultimately, this stage requires a great attention to detail in order to ensure that the subsequent stages can proceed smoothly. One way to reduce the amount of cleaning needed is to have a robust data collection system, such that erroneous and inconsistent data are few and far in between.

### 4.3.5 Relevance to Our Solution

In the context of the datasets that we used for the above models, the only cleaning we did was to exclude samples with missing data. Whereas, more processing was done to lag the variables or to enhance the linearity of the associations between variables. In this case, we would recommend to the EIU to ensure that any new variables that it introduces is properly cleaned, and to explore more in terms of data processing and normalisation.

## 4.4 Data Exploration

### 4.4.1 Introduction

Exploratory analysis (EA) is the process of analyzing and looking through the data sets that have been collected. In doing so, the EIU will be able to summarise the main characteristics of the data, and get a general feel for the data through plotting and visualising the data. Through EA, the EIU can also further identify certain outlying data points which will have to be further investigated.

### 4.4.2 Importance of EA

The main purpose of data exploration is to verify assumptions, identify obvious errors, better understand patterns and trends within the data, as well as discover any interesting relations between variables. (E.g. Section 3.5.3 where we discovered that government expenditure on education seemed to have a logarithmic association with PISA scores). EA provides the context needed to develop an appropriate ML model and interpret the results accordingly.

### 4.4.3 Tools for EA

EA involves a lot of plotting and visualisation, which can be done using software like R (like we have done) or Python, which is more popular nowadays. In terms of the visualisation techniques and statistical functions that can be used for EA, refer to Appendix F.

### 4.4.4 Challenges

The most common challenge in EA is choosing how to represent the data; should it be a pie chart, or a column graph, a scatter plot, etc. Furthermore, data exploration becomes exponentially more difficult as the number of variables increases, as the number of relationships increase as well. One approach that our team took when faced with such an issue was to perform some qualitative analysis beforehand, to clarify any existing assumptions regarding the relationships of variables (e.g. the highly linear association of PISA scores in math, reading, and science) and then verify such relationships through the data visualisation. However, it is also important to note that preexisting assumptions can still be wrong, hence it is still crucial to investigate the data and reaffirm or discover new relationships between variables.

### 4.4.5 Relevance to Our Solution

WIth regards to EA, the EIU has rich experience and expertise in specific industries, which will allow them to perform detailed qualitative analysis of patterns, as well as relationships, which will enhance the value of the EA stage. The EIU can add on to the EA that we have done, which might uncover new trends and relationships that laymen might miss out on.

## 4.5 Machine Learning Modelling

### 4.5.1 Introduction

Modelling is used to find behaviours and patterns in data. However, it is important to re-emphasise at this stage that the model is only as good as the data given to it, thus the EIU must make sure sufficient time and effort is spent in the previous stages outlined above.

### 4.5.2 Generating ML Models

Depending on the problem type (Section 4.1), the EIU will have to select the appropriate type of model to implement. Once we have done so, we can then choose from a variety of ML algorithms that can tackle the problem, and tune their parameters to fit the problem at hand. Several relevant examples are included in Appendix F.

### 4.5.3 Model Evaluation

Similar to how we compare the “goodness-of-fit”, accuracy and explainability of the models that we have generated in Section 3, the EIU must also choose certain performance metrics (Appendix F) to determine the “best model”. Model evaluation will have to be done iteratively based on model accuracy and relevance, and the EIU has to ensure that models have a correct balance between specificity and generalisability.

### 4.5.4 Challenges

Some challenges that the EIU might face is the lack of significant results or findings, similar to the first few models we generated in Section 3.3. This might stem from improper data cleaning, or collecting the wrong data, or having flawed assumptions about certain relationships. In any case, we would advise the EIU to circle back to the data preparation and EA stage to sort out any data inconsistencies or perhaps explore more options in terms of variables and ML algorithms.

### 4.5.5 Relevance to Our Solution

The model that we have generated in Section 3.6 can predict whether a country will have positive or negative GDP growth the next year, based on the economic and education variables that we have outlined. Hence, the EIU can use our model as a baseline reference in its predictions, and combined with its expertise in the industry, the EIU will definitely be able to improve the accuracy of its predictions. One point to note here is that CART is still a very basic model with low complexity, and that might explain its mediocre performance. We would recommend the EIU look into more advanced models and algorithms (Appendix F), which would have a greater likelihood of being able to capture the high-level complexity of the relationship between GDP and other predictor variables, as shown in the example of the case study in Section 2.2.1.

# 5. Conclusion

In this report, we addressed the issue of the inaccuracy of traditional statistical prediction methods that the EIU currently employs, and explored the potential of using Machine Learning models to enhance the EIU’s forecasting capabilities. In predicting GDP growth, our model has an overall accuracy of 57%, managing to predict negative growth with 42% accuracy and positive growth with 70% accuracy. We are confident that using more complex models in this context will yield significantly better results for the EIU, and we recommend the EIU to harness the power of machine learning to add value to its product.

# 6. References

1. *A. Chevalier, C. Harmon et al. (2004). Does Education Raise Productivity, or Just Reflect It? The Economic Journal, 114(499).* [*https://www.jstor.org/stable/3590169*](https://www.jstor.org/stable/3590169)
2. *Allison., P. (2012) When Can You Safely Ignore Multicollinearity?* [*https://statisticalhorizons.com/multicollinearity*](https://statisticalhorizons.com/multicollinearity)
3. *Andrea Bassanini and Stefano Scarpetta - OECD. (n.d.). Retrieved November 6, 2021, from https://www.oecd.org/economy/growth/18450995.pdf.*
4. *Bhatia, R. (2018). How Do Machine Learning Algorithms Differ From Traditional Algorithms.* [*https://analyticsindiamag.com/how-do-machine-learning-algorithms-differ-from-traditional-algorithms/*](https://analyticsindiamag.com/how-do-machine-learning-algorithms-differ-from-traditional-algorithms/)
5. *Boehmke, B., & Greenwell, B. (2020). Decision Trees. In Hands-On Machine Learning With R (p. 7). Github.* [*https://bradleyboehmke.github.io/HOML/DT.html*](https://bradleyboehmke.github.io/HOML/DT.html)
6. *Bontempi, Gianluca & Ben Taieb, Souhaib & Le Borgne, Yann-Aël. (2013). Machine Learning Strategies for Time Series Forecasting.* [*https://www.researchgate.net/publication/236941795\_Machine\_Learning\_Strategies\_for\_Time\_Series\_Forecasting*](https://www.researchgate.net/publication/236941795_Machine_Learning_Strategies_for_Time_Series_Forecasting)
7. *By: IBM Cloud Education. (n.d.). What is exploratory data analysis? IBM. Retrieved November 3, 2021, from* [*https://www.ibm.com/cloud/learn/exploratory-data-analysis#:~:text=Why%20is%20exploratory%20data%20analysis,interesting%20relations%20among%20the%20variables*](https://www.ibm.com/cloud/learn/exploratory-data-analysis#:~:text=Why%20is%20exploratory%20data%20analysis,interesting%20relations%20among%20the%20variables)
8. *Chew C.H. (2020). Artificial Intelligence, Analytics and Data Science, Volume 1: Core Concepts and Models. Cengage, August 2020*
9. *Considerations for collecting data. Ebrary. (n.d.). Retrieved November 3, 2021, from https://ebrary.net/1291/education/considerations\_for\_collecting\_data.*
10. *Dey, V. (2021, October 7). Understanding the importance of data cleaning and normalization. Analytics India Magazine. Retrieved November 3, 2021, from* [*https://analyticsindiamag.com/understanding-the-importance-of-data-cleaning-and-normalization/*](https://analyticsindiamag.com/understanding-the-importance-of-data-cleaning-and-normalization/)
11. *Education statistics - all indicators. DataBank. (n.d.). Retrieved November 6, 2021, from* [*https://databank.worldbank.org/source/education-statistics-%5E-all-indicators*](https://databank.worldbank.org/source/education-statistics-%5E-all-indicators)
12. *Elliott, G., & Timmermann, A. (2008). Economic forecasting. Journal of Economic Literature, 46(1), 3–56.* [*https://doi.org/10.1257/jel.46.1.3*](https://doi.org/10.1257/jel.46.1.3)
13. *Exploratory data analysis and its importance to your business. upGrad blog. (2021, October 20). Retrieved November 3, 2021, from* [*https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/*](https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/)
14. *Help Net Security., et al. (2019) Volume and quality of training data are the largest barriers to applying machine learning.* [*https://www.helpnetsecurity.com/2019/05/28/applying-machine-learning/*](https://www.helpnetsecurity.com/2019/05/28/applying-machine-learning/)
15. *Hirshleifer, D. (2020). Behavioural Biases of Analysts and Investors. National Bureau of Economic Research (2).* [*https://www.nber.org/reporter-2020-02/behavioral-biases-analysts-and-investors*](https://www.nber.org/reporter-2020-02/behavioral-biases-analysts-and-investors)
16. *Ketchell, M. (2016) Economic Forecasting: Why It Matters and Why It’s So Often Wrong.* [*https://theconversation.com/economic-forecasting-why-it-matters-and-why-its-so-often-wrong-53829*](https://theconversation.com/economic-forecasting-why-it-matters-and-why-its-so-often-wrong-53829)
17. *Khan, N.S. (2011). Cyclical Behaviour of Macroeconomic Policies and Capital Flows: A Study of Asian Countries. The Bangladesh Development Studies, 34(2), pp. 47-85.* [*https://www.jstor.org/stable/23339843*](https://www.jstor.org/stable/23339843)
18. *Lounge, C.-learning. (2021, October 1). Data Science Project Life cycle. Medium. Retrieved November 3, 2021, from* [*https://medium.com/co-learning-lounge/complete-data-science-project-life-cycle-9eae6e4ed4c9*](https://medium.com/co-learning-lounge/complete-data-science-project-life-cycle-9eae6e4ed4c9)
19. *Marquez-Ramos, L. and Mourelle, E. (2019), Education and Economic Growth: An Empirical Analysis of Nonlinearities, Applied Economic Analysis, 27(79), pp. 21-45.* [*https://doi.org/10.1108/AEA-06-2019-0005*](https://doi.org/10.1108/AEA-06-2019-0005)
20. *Natekin, Alexey & Knoll, Alois. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics.* [*https://www.researchgate.net/publication/259653472\_Gradient\_Boosting\_Machines\_A\_Tutorial*](https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial)
21. *Normalization. Codecademy. (n.d.). Retrieved November 3, 2021, from* [*https://www.codecademy.com/articles/normalization*](https://www.codecademy.com/articles/normalization)
22. *Poncela., J. (2019). Brief example about data interpolation. Amaral Lab.* [*https://amaral.northwestern.edu/blog/brief-example-about-data-interpolation*](https://amaral.northwestern.edu/blog/brief-example-about-data-interpolation)
23. *Prabhakaran S. (2017). Complete Introduction to Linear Regression in R.* [*https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/*](https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/)
24. *Rohaan, D., Topan, E., & Groothuis-Oudshoorn, C. G. M. (2022). Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. Expert Systems with Applications, 188, 115925.* [*https://doi.org/10.1016/j.eswa.2021.115925*](https://doi.org/10.1016/j.eswa.2021.115925)
25. *Romer, P. M. (1990). Human Capital and Growth: Theory and Evidence. Carnegie-Rochester Conference Series on Public Policy, Vol. 32.* [*https://www.nber.org/papers/w3173*](https://www.nber.org/papers/w3173)
26. *Schultz, T. (1963). Economic Value of Education. Columbia University Press, New York.* [*https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1400922*](https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1400922)
27. *Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological Methods, 14(4), 323-348.* [*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927982/*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927982/)
28. *Swalin., A. (2018) How to Handle Missing Data. towardsdatascience.* [*https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4?gi=5bea8ac9f51b*](https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4?gi=5bea8ac9f51b)
29. *Teixeira, P.N. (2014). Gary Becker’s Early Work on Human Capital – Collaborations and Distinctiveness. IZA J Labor Econ 3(12).* [*https://doi.org/10.1186/s40172-014-0012-2*](https://doi.org/10.1186/s40172-014-0012-2)

# Appendix A: Evaluating Model Fitness

This Appendix serves as supplementation for Sections 3.3.2, 3.5.4, and 3.6, where we discuss the concept of model fitness, specifically with regard to the terms “explainability” and “accuracy” as used in this paper.

## Evaluating Model Fitness: “Explainability” or R2

The explainability of a model refers to whether the output of the model can be explained in a way that is understandable to human beings. R2 or R squared is a common identifier which shows how well a linear regression model fits the data. It is calculated by taking the sum of squares due to the regression divided by the total sum of squares.

When we use machine learning to formulate a model, at times we are able to get a good model fit and hence a low R2 when we choose to rigorously train the model with numerous variables and data. However, this may come at the cost of explainability as it would in turn be harder to explain the reasoning and implications behind the model.

## Evaluating Model Fitness: “Error Rating” or RMSE

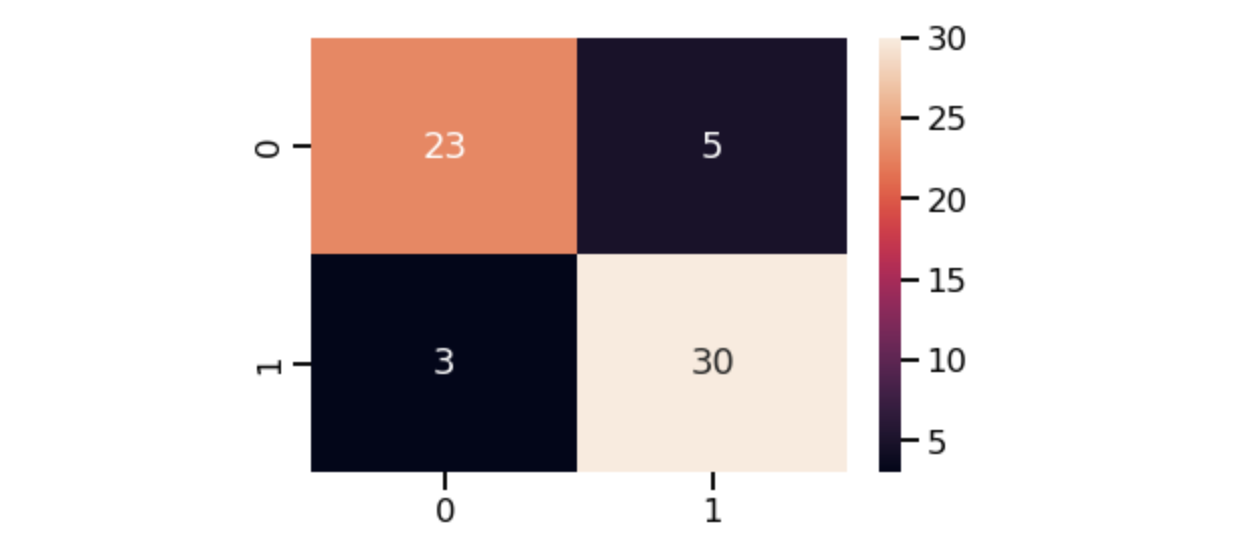
Model accuracy and Root Mean Squared Error (RMSE) both assess how good the predictions made are. However, model accuracy looks to assess classification prediction, with models such as logistic regression or CART, while RMSE assess numerical continuous variables and their deviation from the actual data.

RMSE is calculated by taking the square root of the mean of the squared residuals, where each residual is the deviation of the predicted value compared to the actual value. The figure below shows the how residuals are found



*Retrieved from* [*Hatari Labs*](https://hatarilabs.com/)

Model accuracy is calculated using the number of correct predictions divided by the total number of predictions. This can be illustrated by a confusion matrix, with an example shown in the figure below. The bottom axis shows the predictions while the left shows the true values. Through this we can find the true positive and true negative values which will enable us to find the model accuracy.



*Retrieved from* [*Medium*](http://medium.com/@dtuk81/)

## 

## Train Test Split

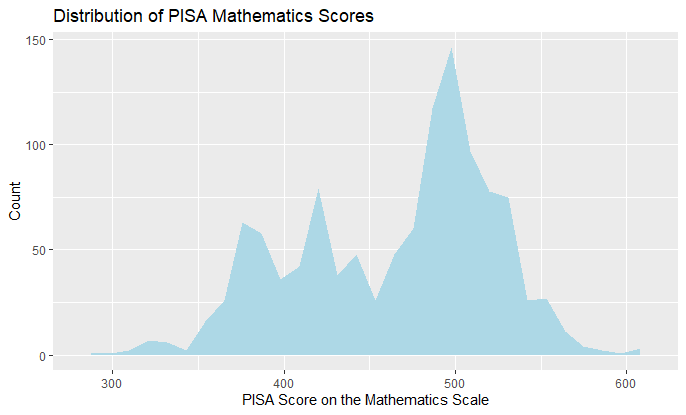
In the field of statistical analysis and regression modeling, to test the accuracy of a model, the data will be split randomly into two subsets. A train set and a test set. The train set would usually contain the majority of the data (about 70 percent) and is used to train the regression model. The test set will then be used to test the capabilities of the model. The assumption here is that the train and test set are split randomly so that the two sets will be independent of each other.

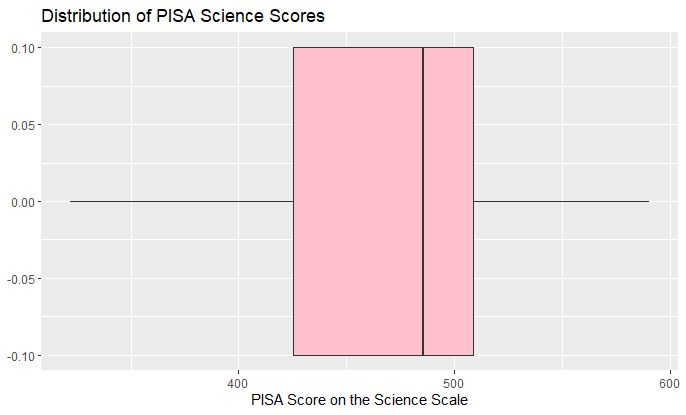
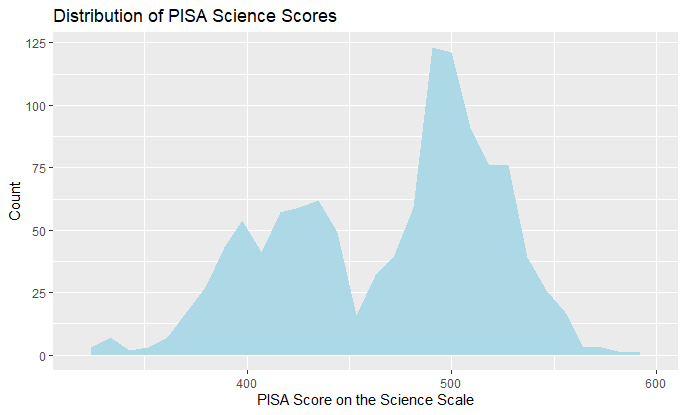
# Appendix B: Exploratory Analysis

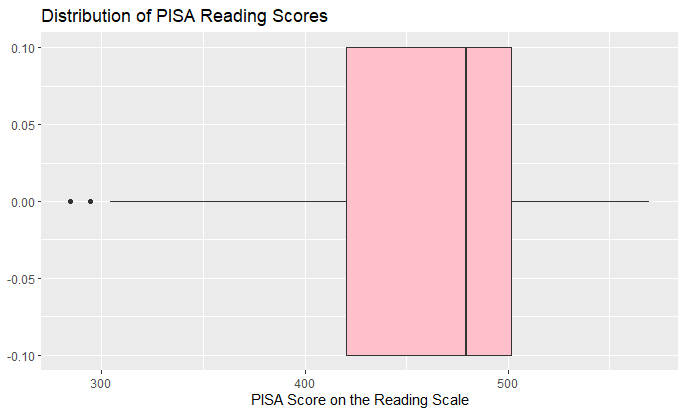
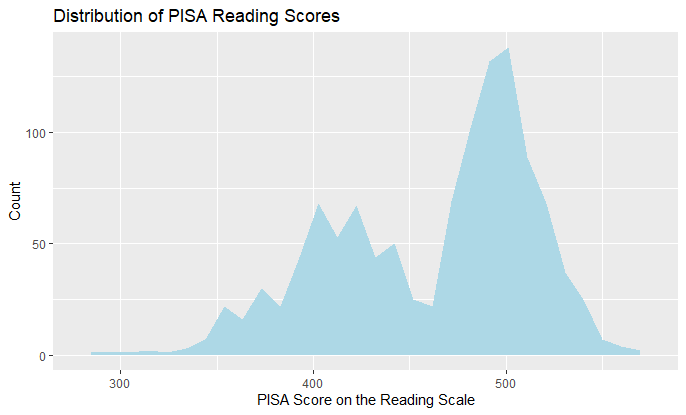
Comments on these plots can be found in the Machine Learning R script.

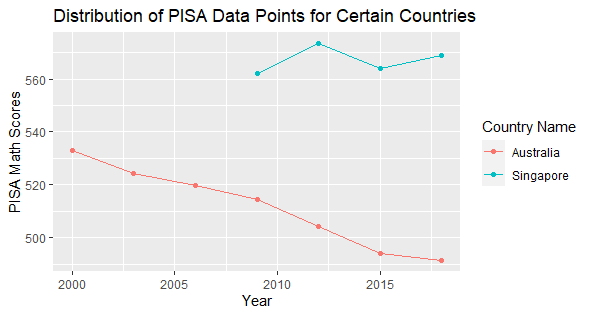
## Univariate Plots

### PISA Scores

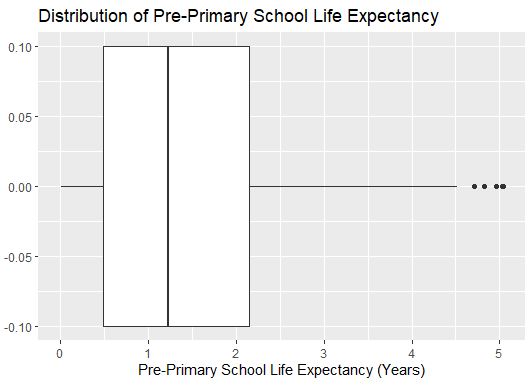
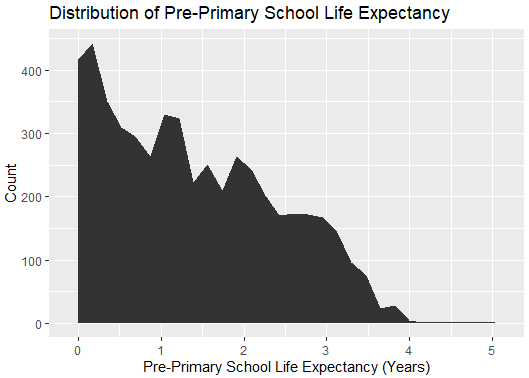


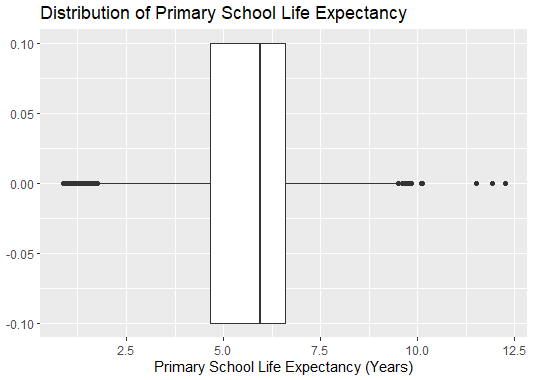
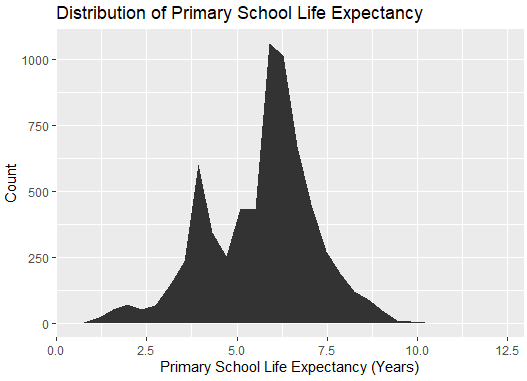


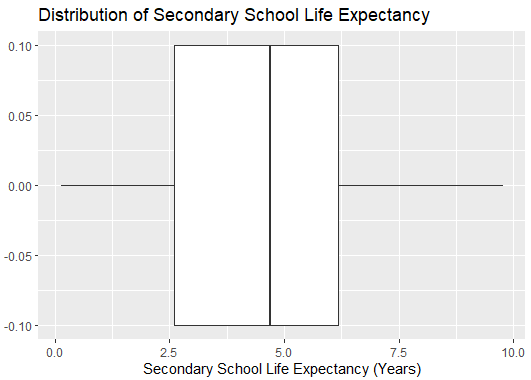
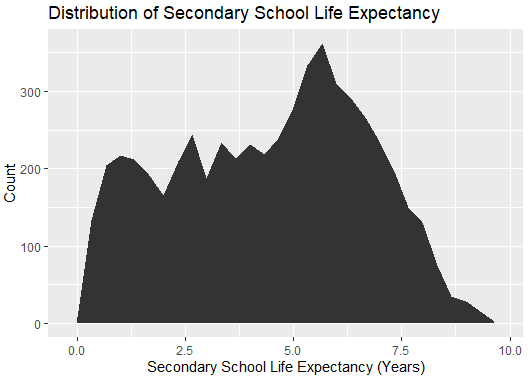


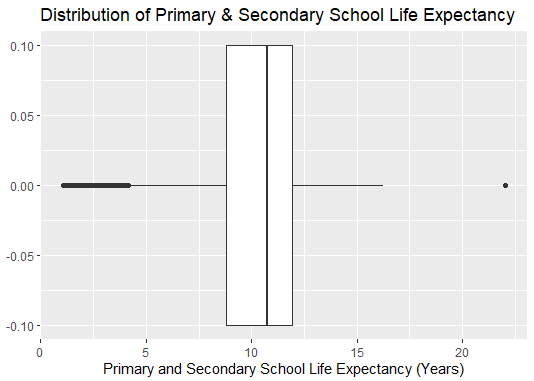
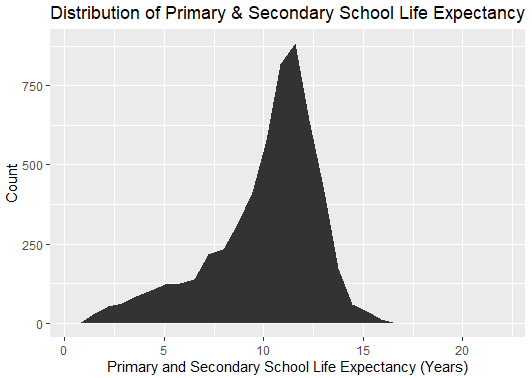


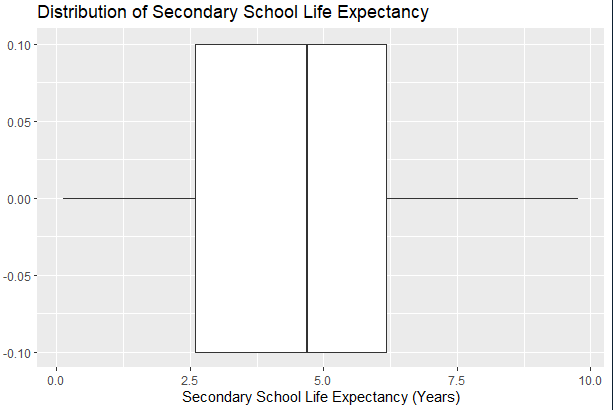
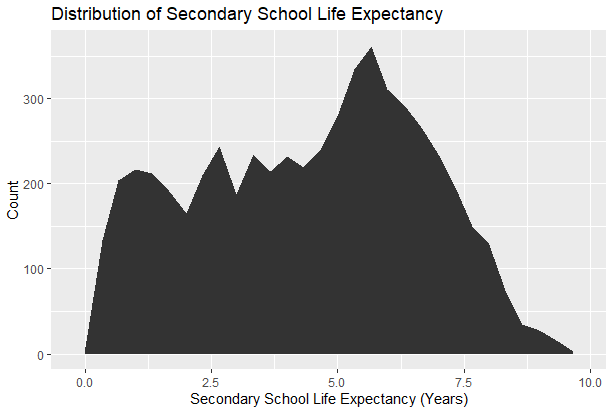
### School Life Expectancy



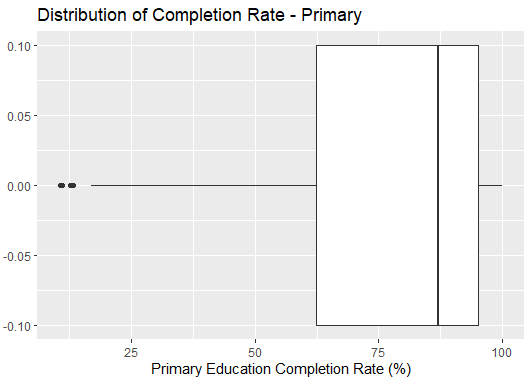
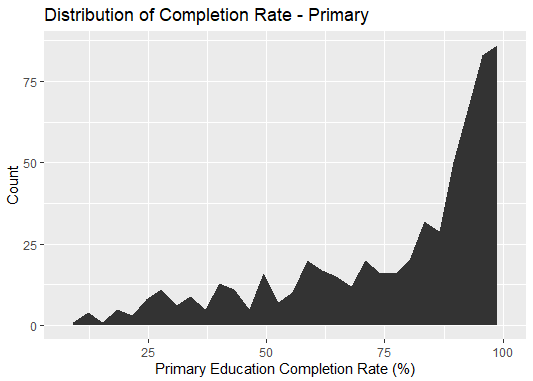


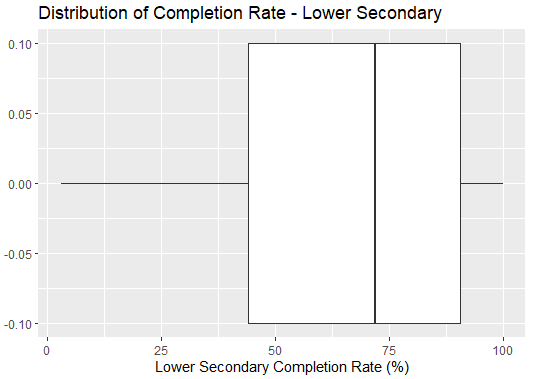
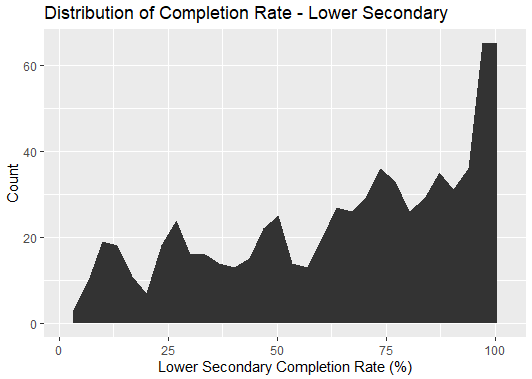


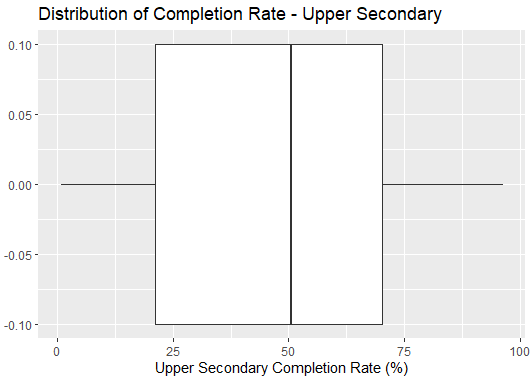
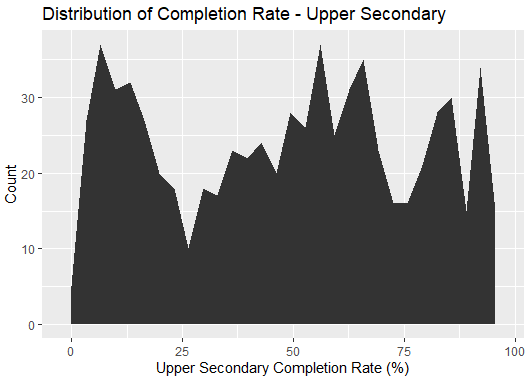




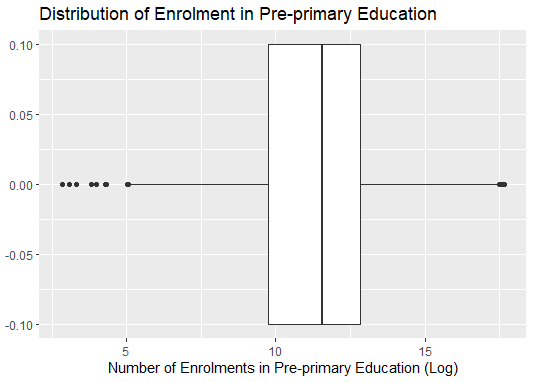
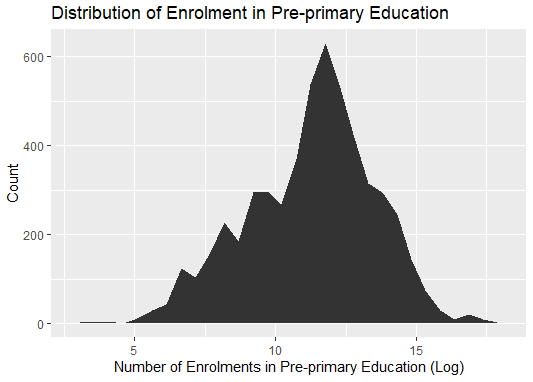
### School Completion Rate

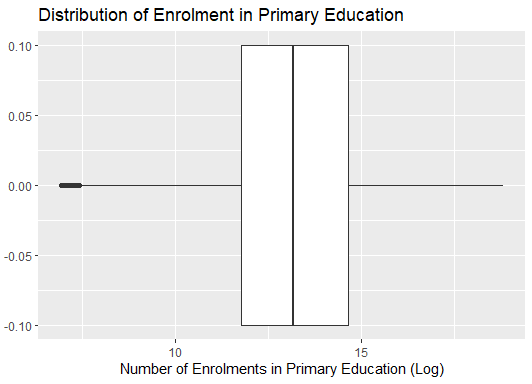
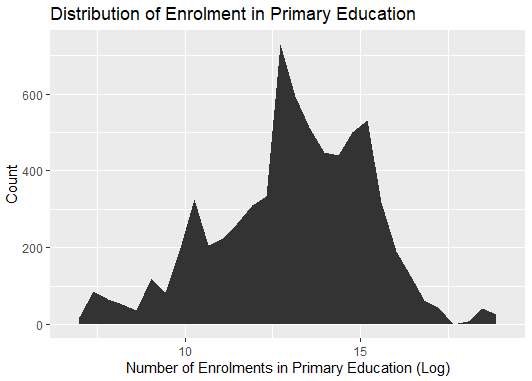


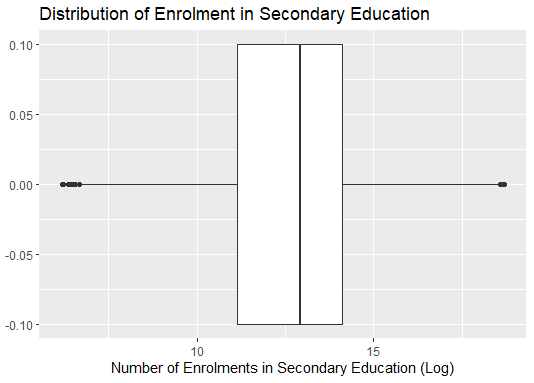
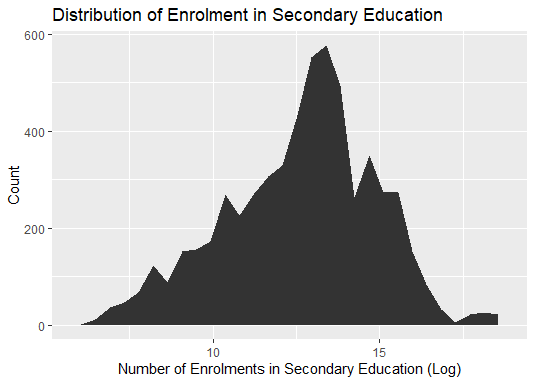


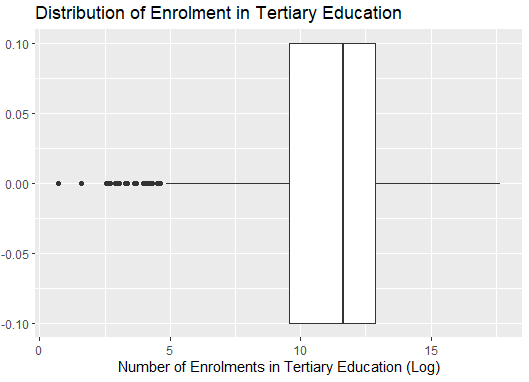
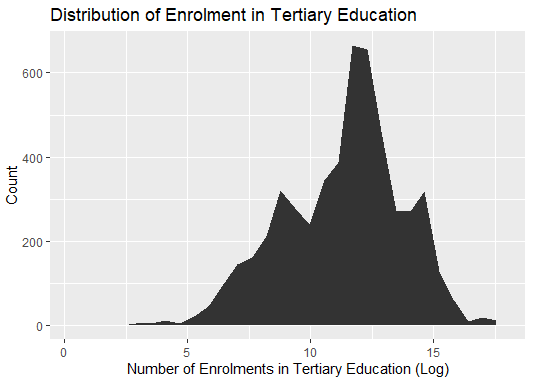


### Number of Enrolments

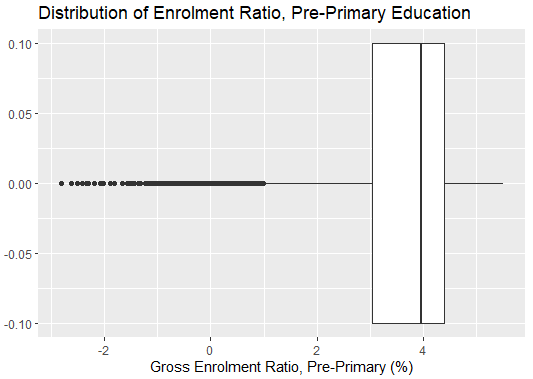
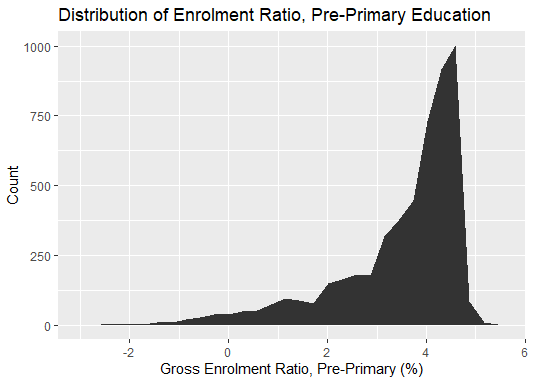


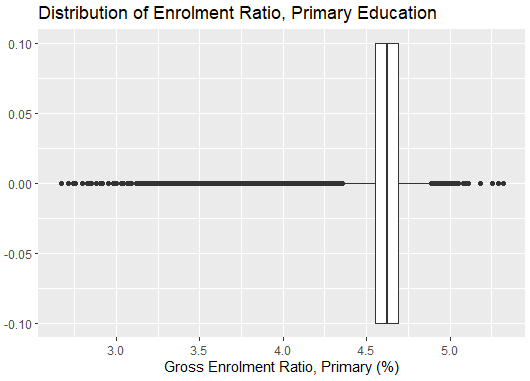
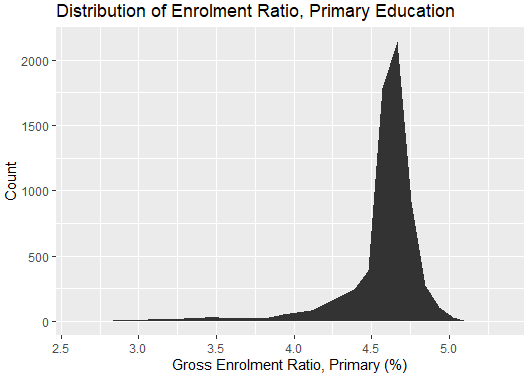


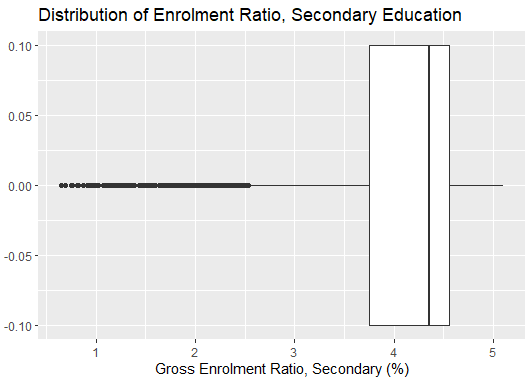
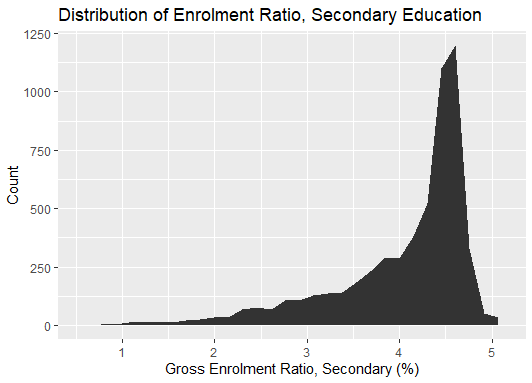




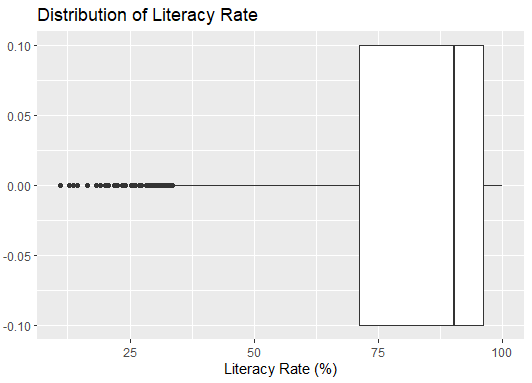
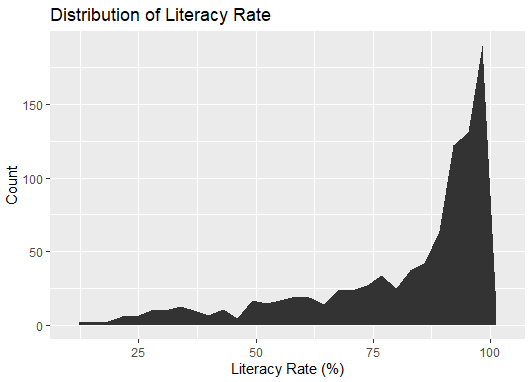
### Enrolment Ratio

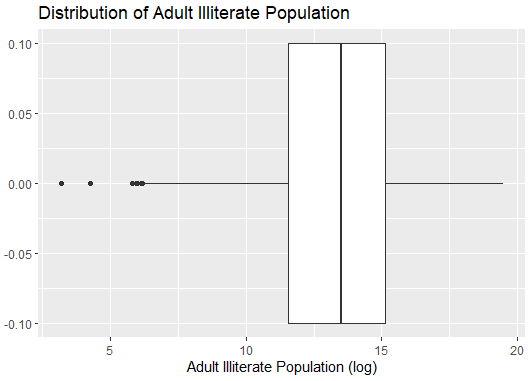




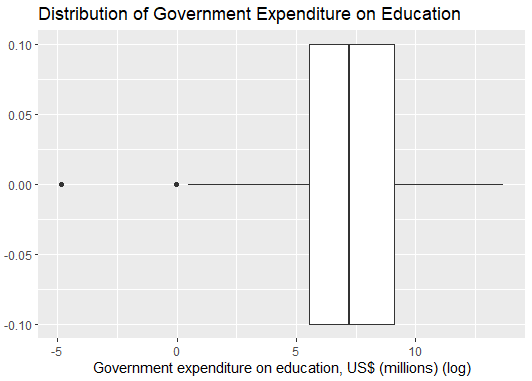
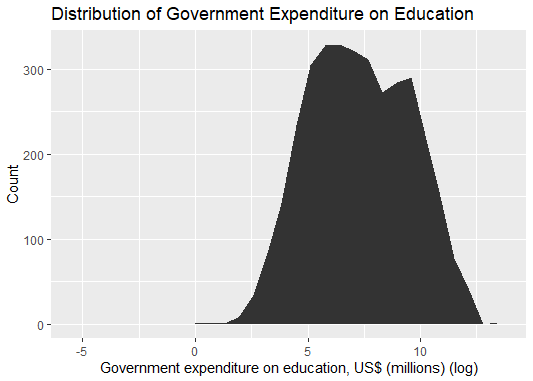


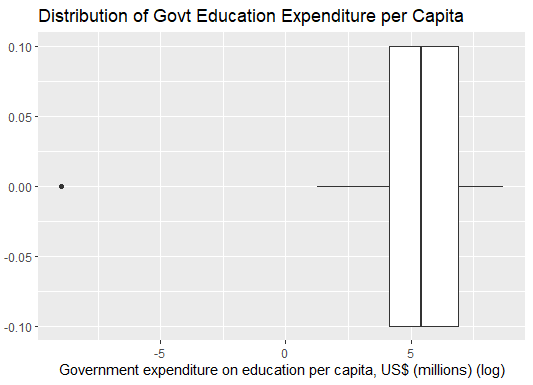
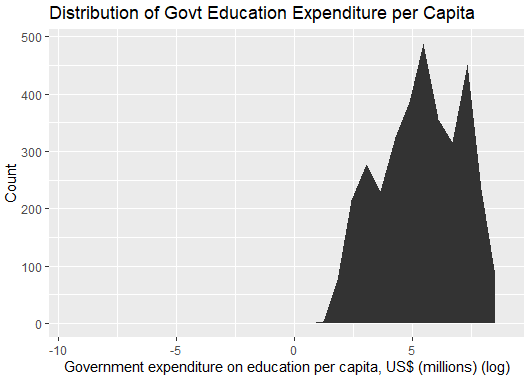
### Literacy Rate





### Government Expenditure





### 

### 

### 

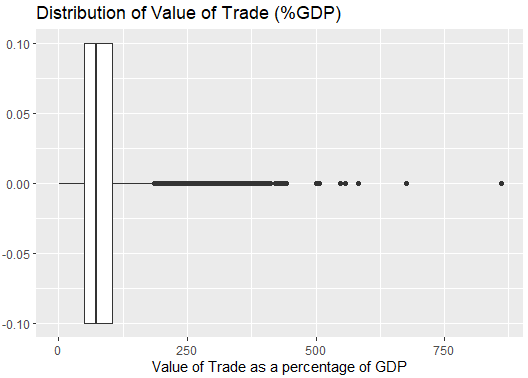
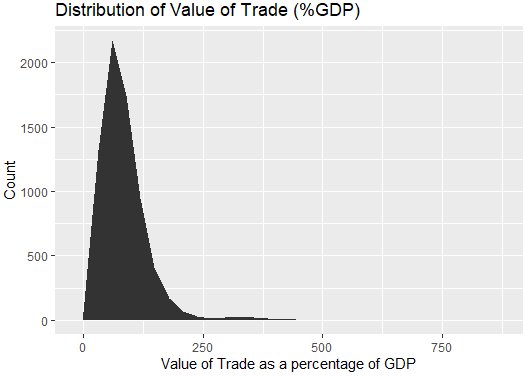
### 

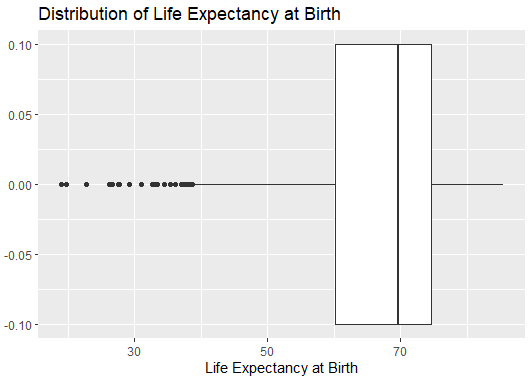
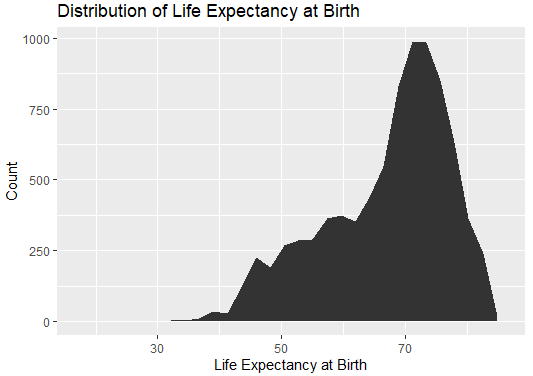
### 

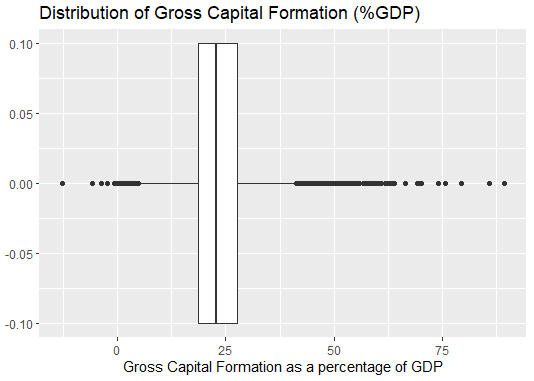
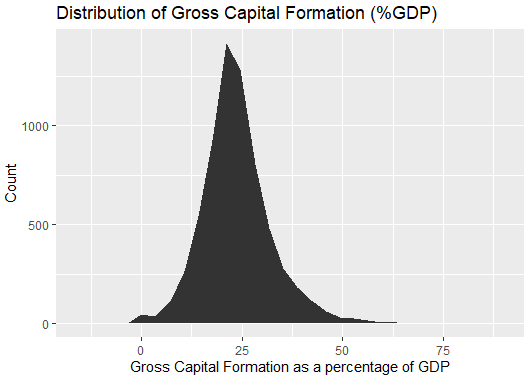
### 

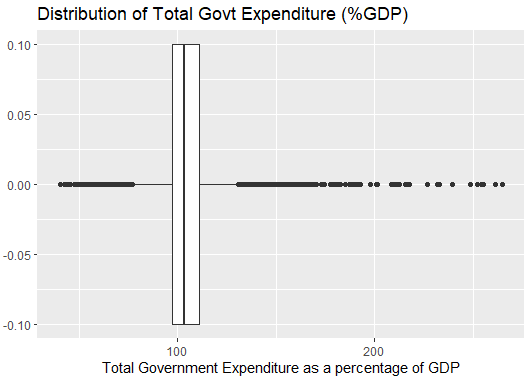
### 

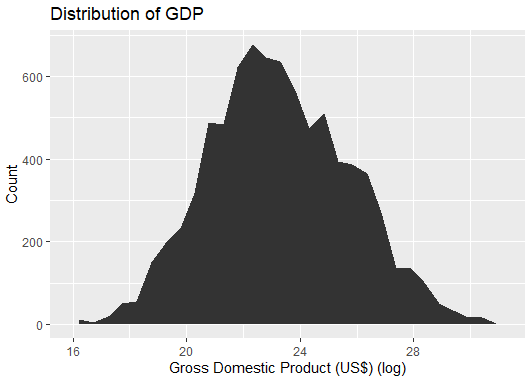
### Economic Indicators

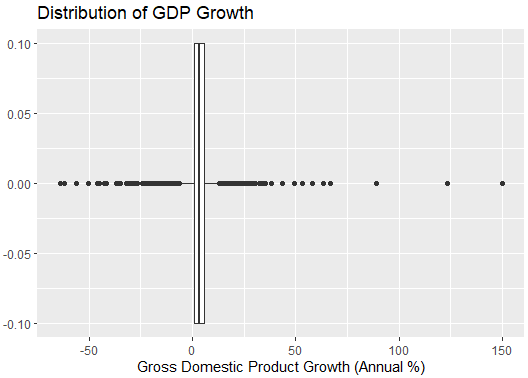
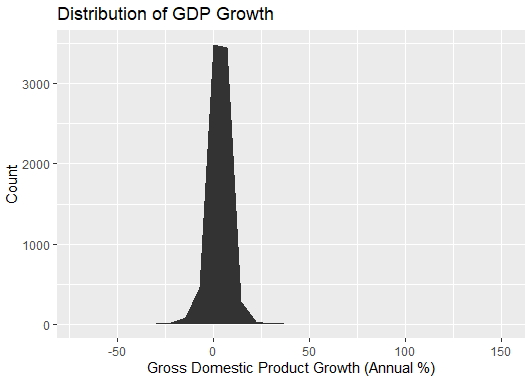


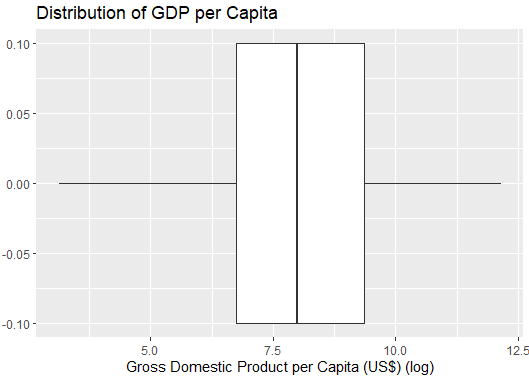
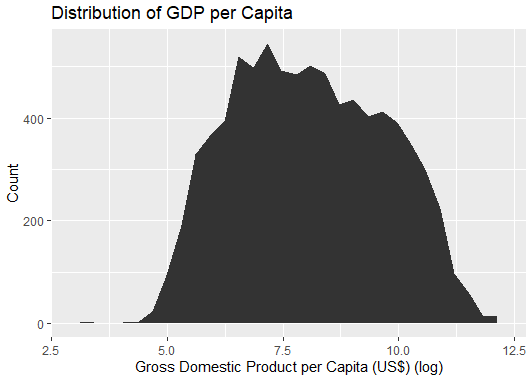


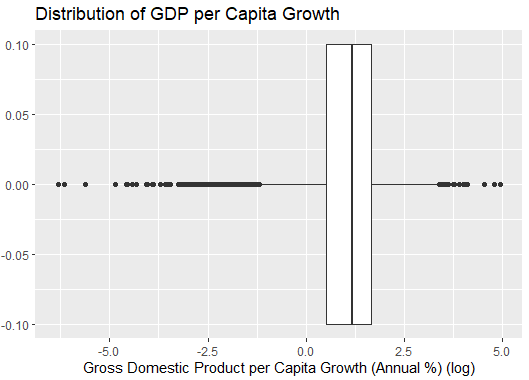
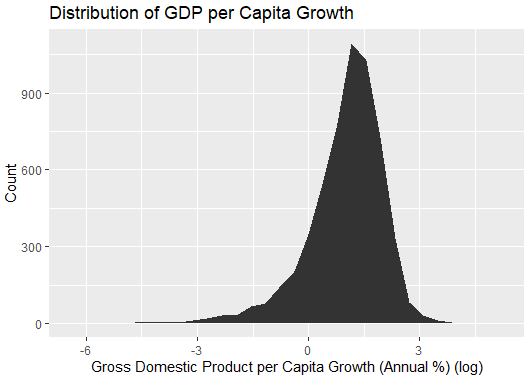




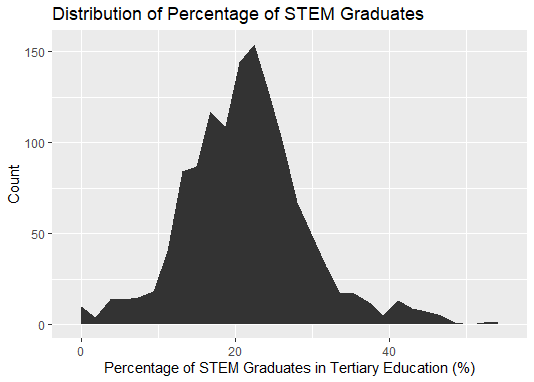


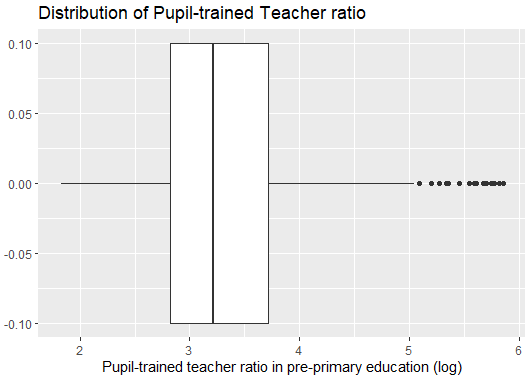
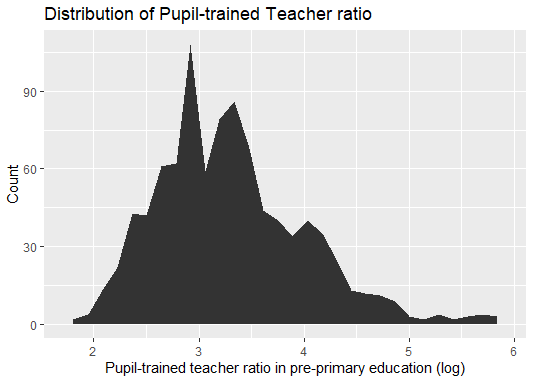






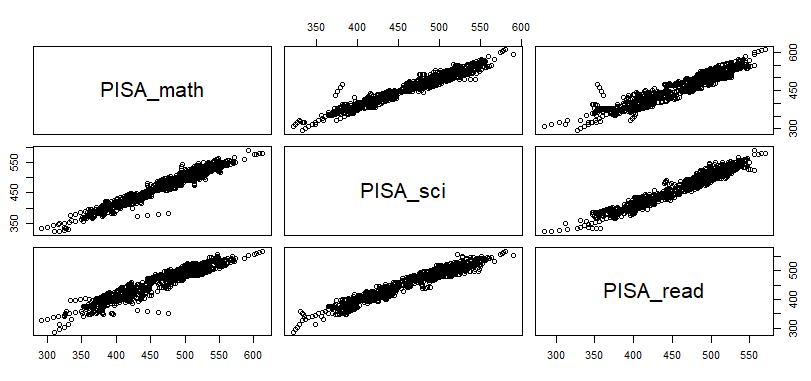
### Others

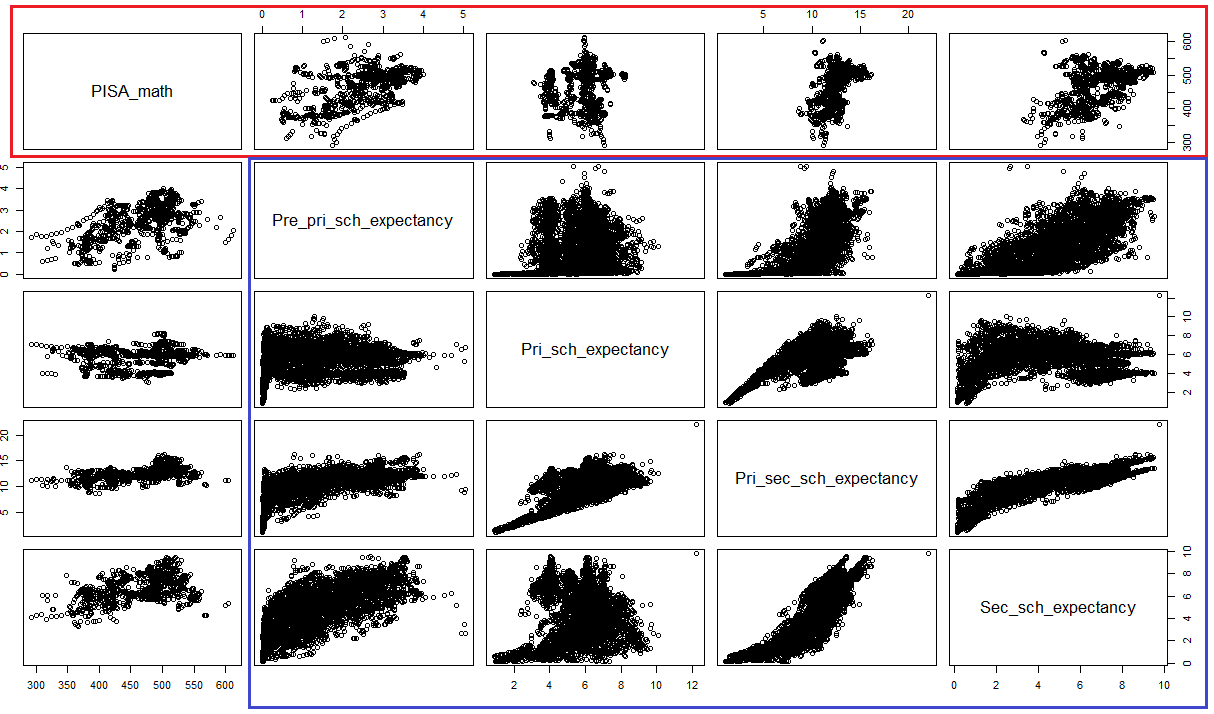




## Bivariate Plots

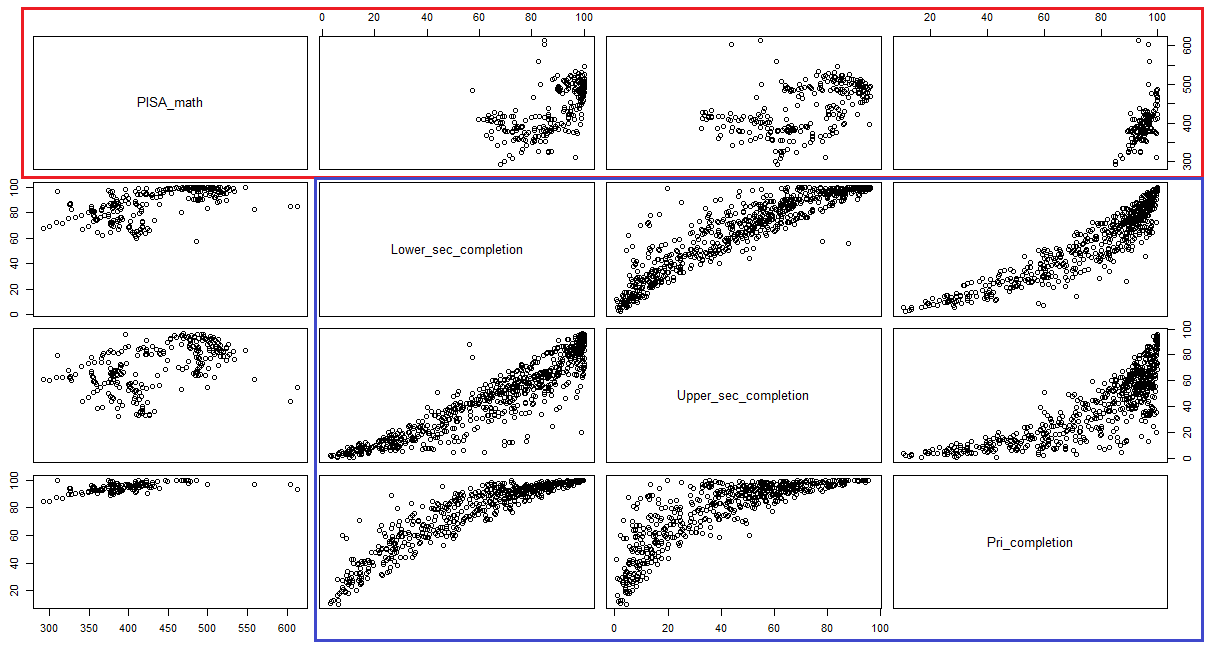
### PISA as response variable





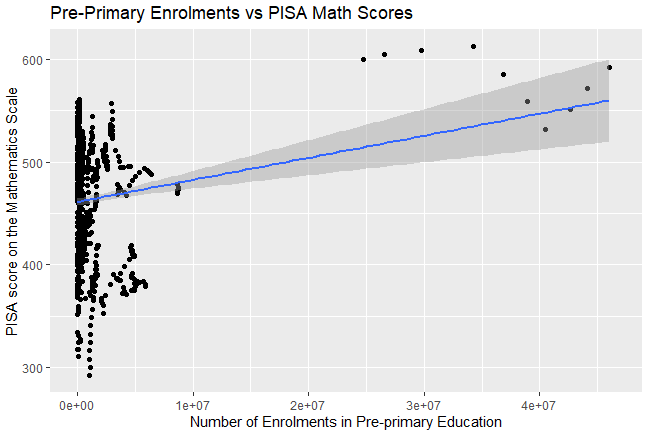
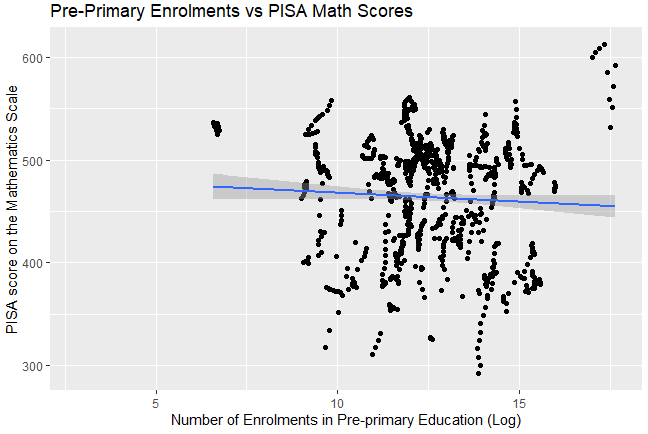
Outlined in red: Moderate linear association observed between PISA math scores and pre-primary school life expectancy, as well as secondary school expectancy. Satisfies Assumption 2 in Appendix C.

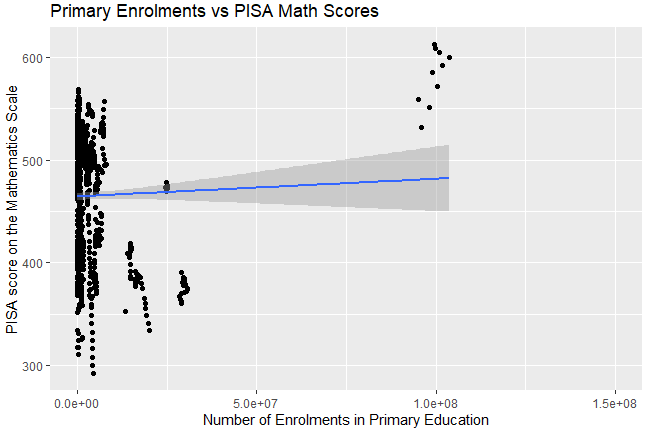
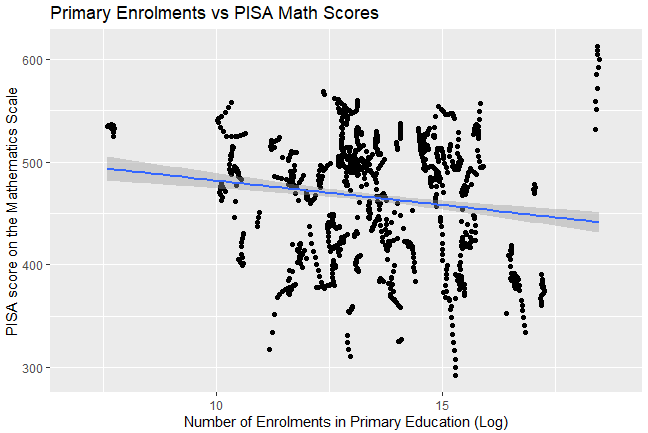
Outlined in blue: The strong linear associations observed between some of the variables may cause issues later with regard to multicollinearity as outlined in blue.

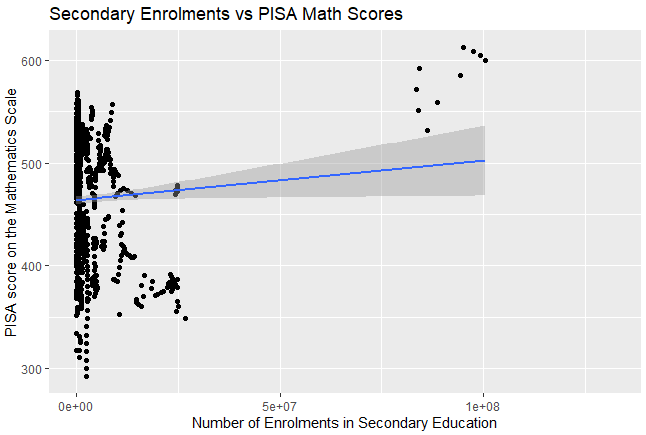
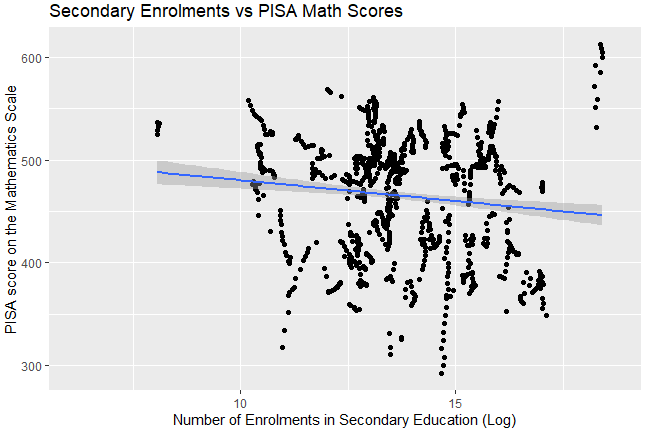


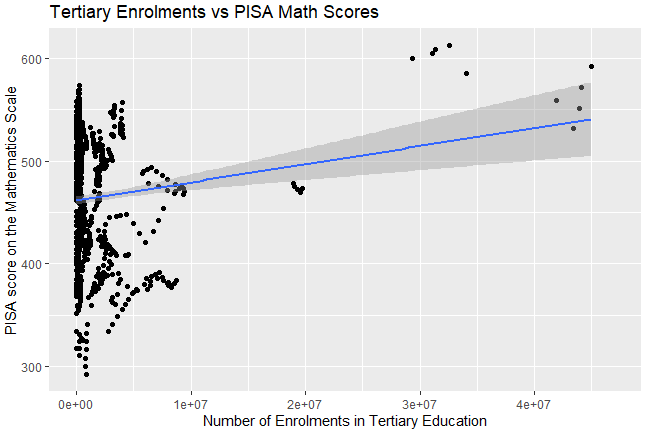
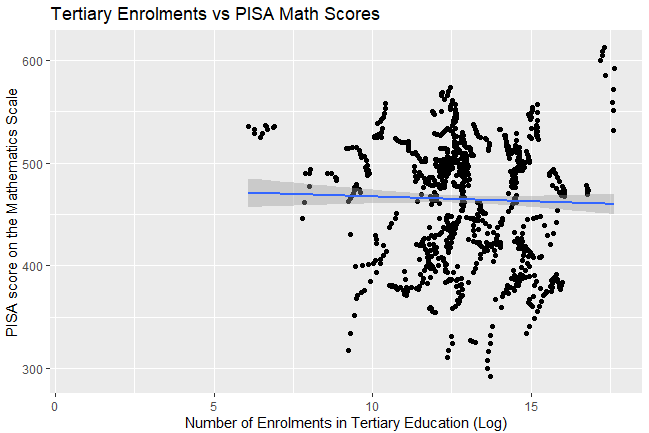
Moderately strong linear association observed between PISA math scores and school completion rates as outlined in red.

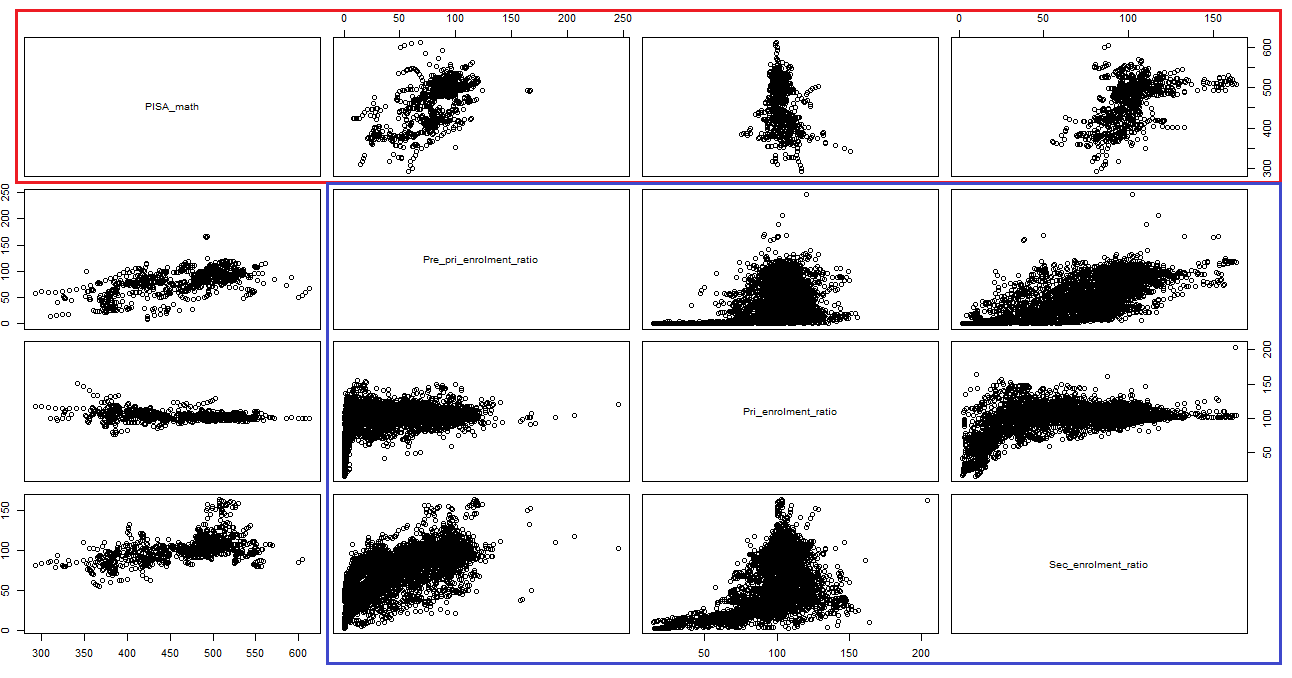
Collinearity issues between the school completion rates may arise due to strong linear and logarithmic associations observed between some of the variables as outlined in blue.





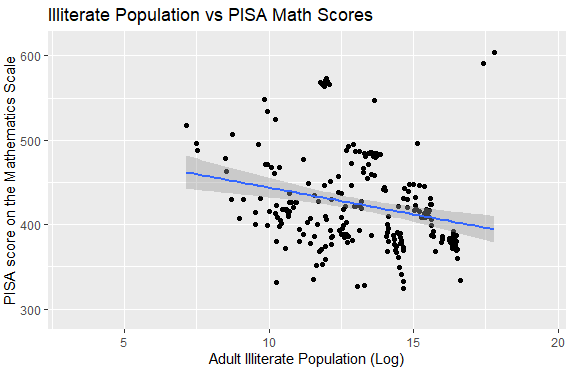
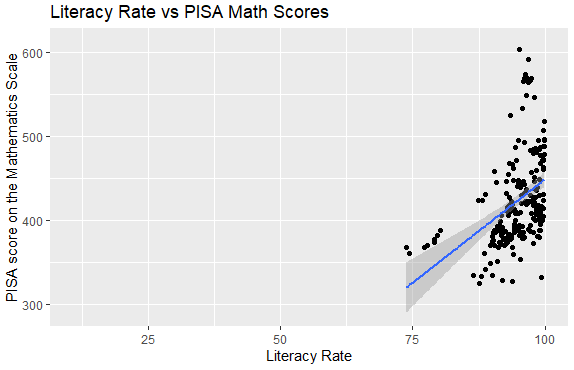




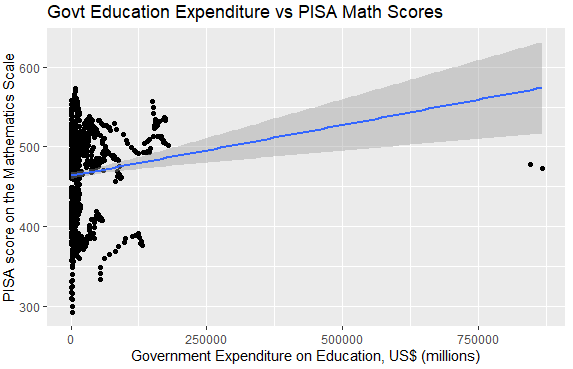
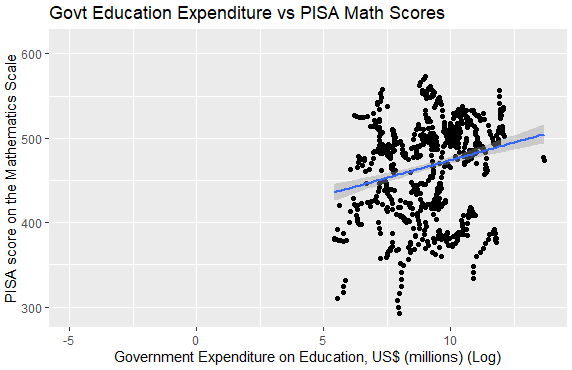


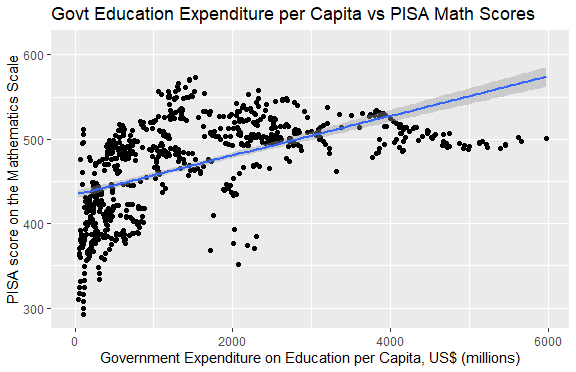
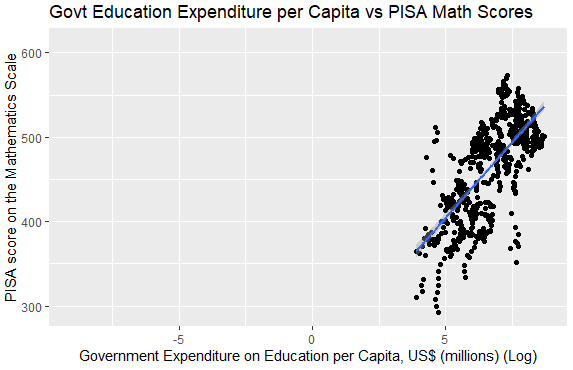
Weak to moderate linear association observed between PISA math scores and pre-primary enrolment ratio as well as secondary enrolment ratio (outlined in red).

Collinearity issues between the school life expectancies may arise due to strong linear associations observed between pre-primary and secondary enrolment ratio as well as a logarithmic association between secondary and primary enrolment ratio (outlined in blue).

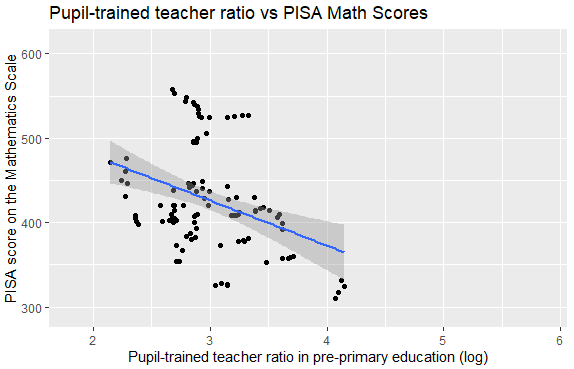
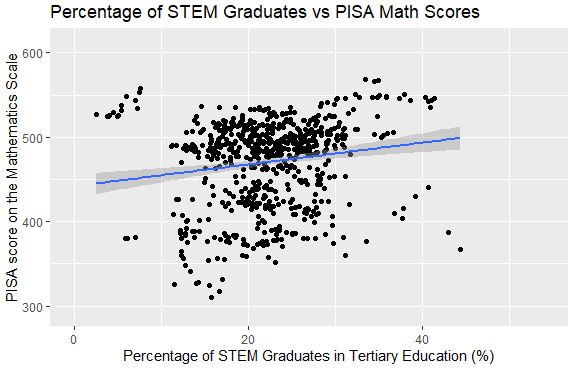


Moderately strong linear association observed between the log(illiterate population) against PISA math scores.

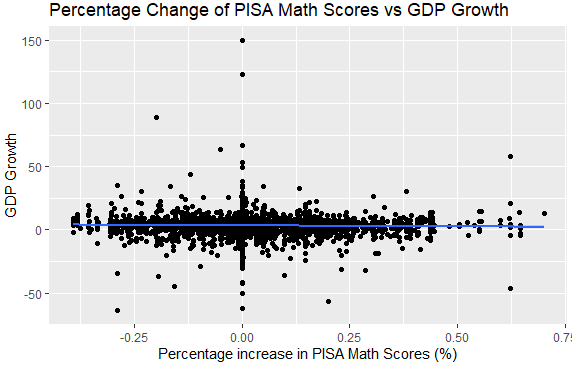
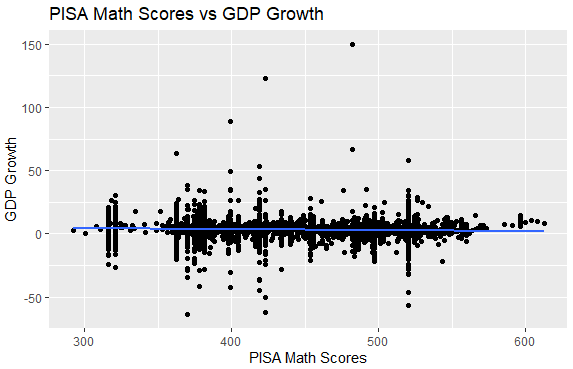


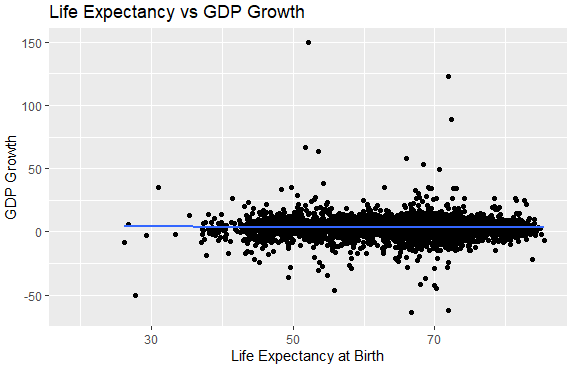
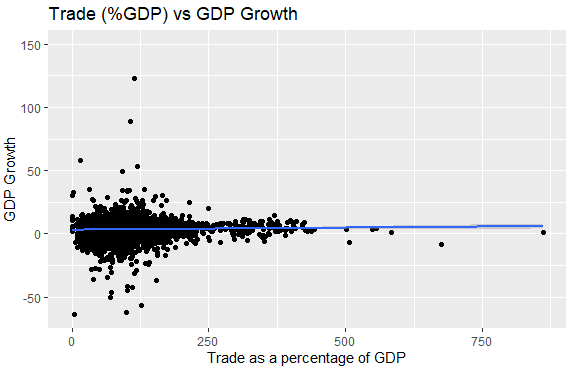


Strong logarithmic association observed between Government Education Expenditure per Capita against PISA math scores, which translated into a strong linear association between log(Govt Education Expenditure per Capita) against PISA.



### GDP Growth as response variable





# 

# Appendix C: Technical Details of Linear Regression

This Appendix serves as supplementation for Section 3.3.3, where we introduced the Linear Regression model. This model was subsequently used in Sections 3.6.2 and 3.8.3. We delve into its technical details here.

## Handling Missing Values

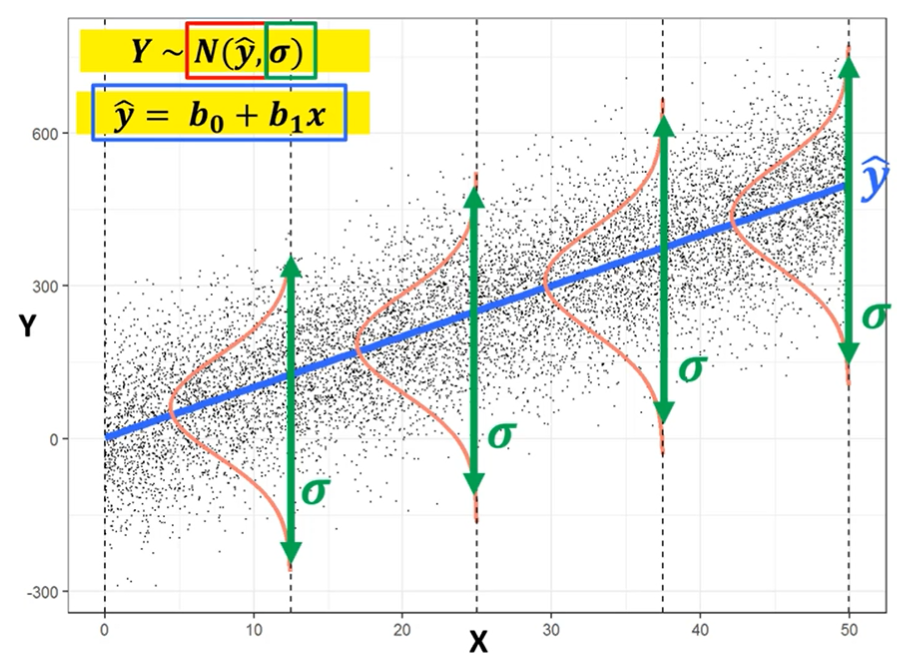
Linear regression will take all the values of the dependent variable (y-axis) and plot them against the independent variables (x-axis). However, when there are missing values, in some cases the model will count them as 0 values, which will be included as part of the model. This would bring the mean value and gradient of the plotted points to be closer to 0 then intended. Hence we need to either make sure that the missing values are not included in the regression model, or insert a reasonable value into the missing data.

## Pitfalls in Linear Regression

### Required Assumptions in Linear Regression

1. Linear association between Y and Xs
2. Errors have a normal distribution with mean 0
3. Errors are independent of X and have constant standard deviation

Visually, these four assumptions are summarised into the following graph. The straight blue line represents assumption #1, the red normal distribution curves represent assumption #2, and the green standard deviation range represents assumption #3.

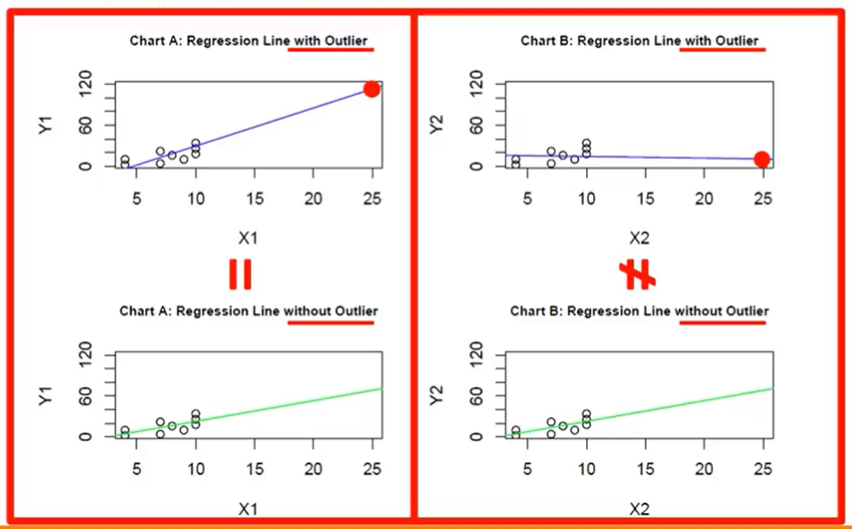
**

*Retrieved from BC2406 Unit 6 Slides.*

### Influential Outliers

An outlier is a data point that falls an abnormally large distance away from the other values in a dataset.

An outlier becomes influential if its exclusion does not significantly change the plotted regression line. In the below example, both Charts have outliers, but only Chart B’s one is influential since its presence has an effect on the regression line unlike that of Chart A.



*Retrieved from BC2406 Unit 6 Slides.*

### Checking for Required Assumptions & Influential Outliers

By using the plot() function provided in R, we are able to produce 4 diagnostic plots as follows:

| **Residuals vs Fitted**  To test:   * Assumption 1: linear association between Y and Xs. * Assumption 2: errors have a normal distribution with mean 0.   **Ideal:**   * **Straight line**   Examples:      Clear curvature in Case 2, where non-linear relationship is not explained by the model.  **Resolution:**   * **Try non-linear variables e.g. y = a+bx+cx^2 for u-shapes. Other options are using higher polynomial terms such as x^3 or using a spline** | **Q-Q Plot**  To test:   * Assumption 2: errors have a normal distribution with mean 0.   **Ideal:**   * **Data points fall on the dotted line**   Examples:    *Some problems at the tail ends, but not serious*      Residuals significantly deviate from the straight line in Case 2, violating normal distribution assumption.    **Resolution:**   * **Possibly taking a log of the Y variable e.g., log(Y) = a + bx** * **Other methods include:**   + **using different family in GLM**   + **non-parametric method using bootstrapping** |
| --- | --- |
| **Scale-Location**  To test:   * Assumption 3: errors are independent of X and have constant standard deviation.   Ideal:   * Points are distributed with the same spread in each vertical slice.   Examples:      Residuals begin to spread wider along the x-axis in Case 2, violating constant standard deviation of error assumption since error now depends on X.  **Resolution:**   * **Use robust standard error** | **Cook’s Distance**  To test:   * Assumption 3: errors are independent of X and have constant standard deviation.   Ideal:   * Points are distributed with the same spread in each vertical slice and do not fall beyond the dotted line   Examples:        The outlier in Case 2 is influential as it lies outside the dotted line.  **Resolution:**   * **Remove those outliers** |

*Retrieved from BC2406 Unit 6 Slides, BC2406 Seminar 2 Slides*

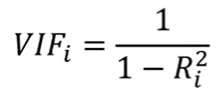
### Multicollinearity or “related variables”

Multicollinearity occurs when an X variable can be expressed statistically well as a linear combination of some other X variables. A lot of information about that X variable is already contained in the other X variables, e.g. length of left arm and right arm.

The multicollinear X can still be used to predict Y, but will result in unstable model coefficients. Hence coefficients cannot be interpreted as the impact of the variable on the response.

### Reducing collinearity using VIF and AIC

Mathematically, given a model M, the VIF of the ith X variable, Xi, is:



where Ri2 is the R2 statistic in the linear regression with Xi as the outcome variable (Y) on all the other X variables in the Model M.

R provides a function, vif(), in the ‘car’ library, to check for multicollinearity. There is no consensus on the cut-off. Some research papers conclude multicollinearity if VIF > 5 while others are more strict and use VIF > 10.

When dealing with multicollinearity, variables should be removed one at a time. This is because the collinearity of a variable can be caused by “overlaps” with the Y variable, with other X variables, or both. In the centre case, removing X2 will also largely reduce the collinearity of X3 and X1.

### 

*Retrieved from BC2406 Seminar 2 Slides*

Collinearity can also be safely ignored in three conditions (Allison., 2012):

1. The variables with high VIFs are control variables, and the variables of interest do not have high VIFs.
2. The high VIFs are caused by the inclusion of powers or products of other variables.
3. The variables with high VIFs are indicator (dummy) variables that represent a categorical variable with three or more categories.

### Backwards Elimination

R provides the step() function to perform backwards elimination by automatically performing variable selection for us. It uses AIC, a metric for information loss where a lower AIC is more desirable. All variables whose removal would reduce AIC are candidates for removal, although this removal should still be done one at a time similar to variable selection for multicollinearity.

# Appendix D: Technical Details of CART

This Appendix serves as supplementation for Section 3.3.1, where we introduced the Classification and Regression Tree model. This model was subsequently used in many Sections (3.3.4, 3.4.2, 3.5.2, 3.6.1, 3.8.1, 3.8.2, 3.9.2, 3.10.1, 3.10.2). Here, we delve into the technical details behind this machine learning model.

## Technical Introduction to CART

The output to CART is a decision tree. The following is a table of important decision tree-related terminologies to serve as a baseline for your understanding of subsequent sections in this Appendix:

| Non-terminal Nodes | Points in the decision tree at which a split occurs. |
| --- | --- |
| Split | A point at which divergence in the decision tree occurs. The variable that determines this split is called the splitting variable. The variable must be above or below (inclusive/exclusive) a specific value as defined by this split. |
| Terminal Nodes | A node where no splitting occurs. Predictions at terminal nodes are based on majority. |
| Node Purity | The proportion of the majority at the terminal node. This directly affects model prediction accuracy. |

## Developing the CART

### Phase 1: Growing the (decision) tree to the maximum

At each node, the CART algorithm will consider and test all x variables and all possible values in that x variable to determine the best splitting variable and binary split, which will ideally result in the purest possible child nodes, on average. The best splits are then used to generate 2 child nodes, and the x variables with the highest associations with the y variable will be repeatedly splitted first to generate the splitting rules.

This process continues for each child node, until a predetermined stopping criteria is met. These criteria include all terminal nodes being completely pure and the number of observations in each terminal node reaching the pre-specified minimum. This methodology is technically called recursive partitioning (Strobl. et al, 2009).

The fully grown tree will likely overfit the train set used and the resulting model may be overly complex and difficult to interpret, leading to poor test set performance (Boehmke & Greenwell, 2020). This is solved via Phase 2, where the tree is pruned to its optimal level.

### Phase 2: Pruning the (decision) tree to the minimum

Before we discuss Phase 2, here is a summary of other important terminologies used in this section:

| CP | Prune triggers at which size of tree changes  (if CP = inf, it is root node) |
| --- | --- |
| nsplit | Number of splits (number of terminal nodes-1) |
| rel error | Train set MSE |
| xerror | Average cross-validation testset MSE |
| xstd | Standard deviation of cross-validation test set error (since we have 10 separate values for cross-validation error) |

In Phase 2, we prune to simplify the overfitted tree by applying the optimal complexity parameter (cp). As all trees whose 10-fold Cross Validation (CV) error is below the CV error cap are statistically significant in terms of error, the optimal tree will be the simplest tree with the highest CV error right within the CV error cap (Kumar., n.d.). The CV error cap is determined by adding a 1 standard error (xstd) to the minimum CV error. The pruned optimal tree will be the most stable tree that still performs well.

# Appendix E: Oversampling Techniques

## Synthetic Minority Oversampling Technique ([SMOTE](https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5))

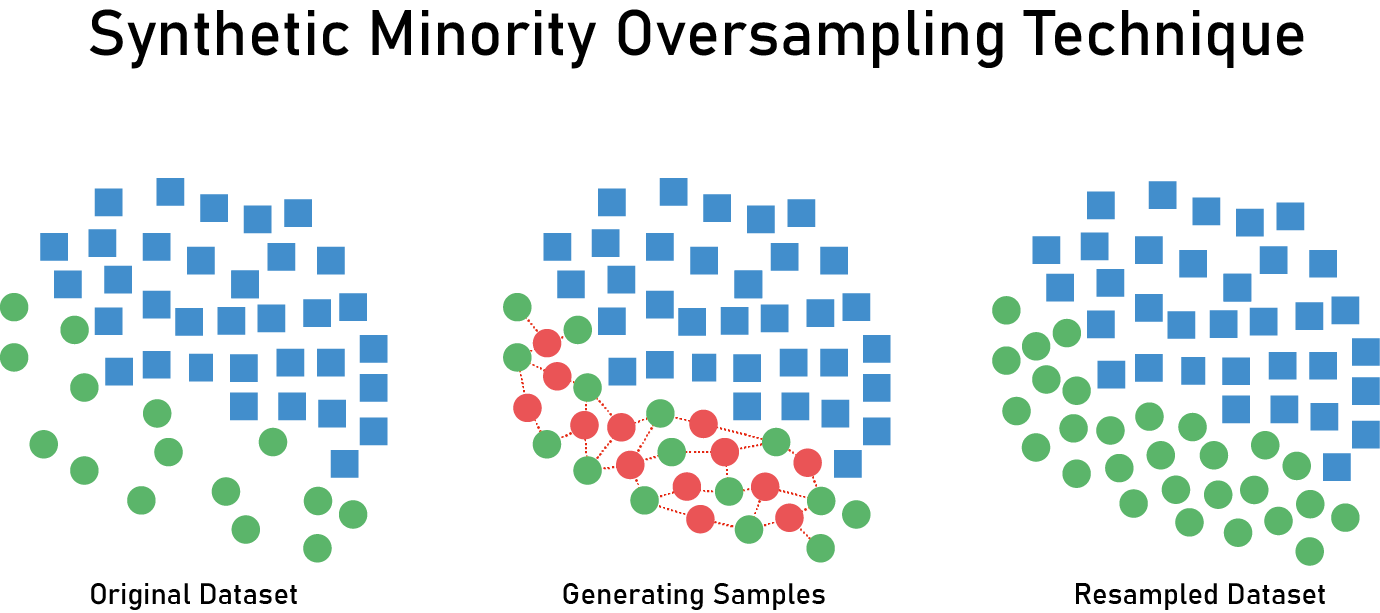
### Introduction

The problem with imbalanced classification is that there are too few samples of the minority class. Hence the model is unable to effectively learn the decision boundary, a.k.a. when to predict otherwise.

One way to overcome this problem is to oversample the data in the minority class. This can be done by duplicating samples from the minority class, but this does not provide any additional information to the model, and can cause the model to be prone to overfitting. SMOTE solves this overfitting problem by creating artificial data points from the minority class.

### Process

SMOTE works by using a *k-*nearest neighbour algorithm to create synthetic data. It starts by choosing a random data point from the minority class, then synthetic data will be created between the selected data point and a randomly selected *k*-nearest neighbour, illustrated in the diagram below:



*Retrieved from* [*Medium*](https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2)

### Types of SMOTE

* Default SMOTE (the technique that we used in Section 3.10)
* SMOTE-NC
  + Creates synthetic data for mixed (continuous and categorical features)
* Borderline-SMOTE
  + Makes synthetic data along the decision boundary between the two classes
* Borderline-SMOTE SVM
  + Uses SVM algorithm instead of *k*-nearest neighbours
* Adaptive Synthetic Sampling
  + Creates more synthetic data in regions where density of minority of examples are low

# Appendix F: Pilot Study - Data Science Life Cycle

## Data Science Problems

Fundamentally, there are 5 questions that ML can answer:

1. How much or how many? (Regression)
2. Is it A or B? (Classification)
3. Is the data organised? (Clustering)
4. Is this weird? (Anomaly detection)
5. What should I do next? (Reinforcement Learning)

By classifying the business problem at hand into one of these five subcategories, we can then determine which type of model and algorithm can be applied to solve the problem.

## Data Normalisation

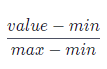
### Introduction

Normalisation is the process of standardising and organising data, which includes creating tables and establishing relationships, thereby eliminating redundancy and inconsistency, ensuring logical data storage.

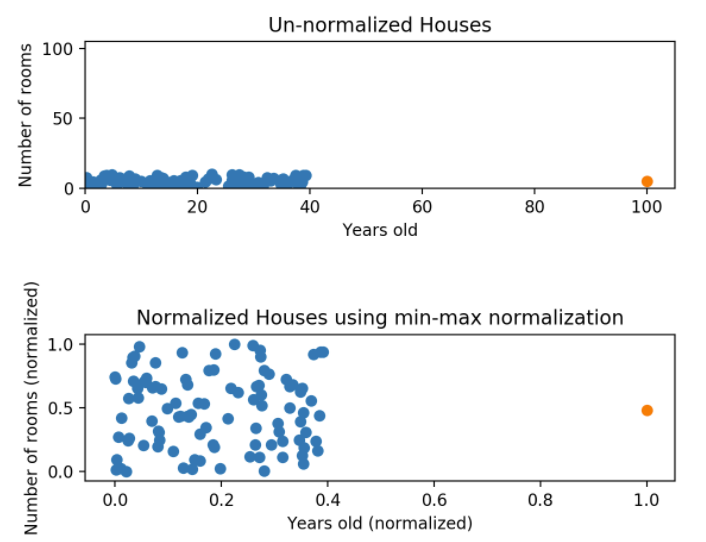
Data normalization transforms the multiscaled data all to the same scale. After normalization, all variables have a similar weightage on the model, hence improving the stability and performance of the learning algorithm. Normalization gives equal importance to each variable so that no single variable drives the model performance.

### Methods of Normalisation

1. Min-Max normalisation:



It converts features from their original range into a standard range (usually between 0-1). One thing to note is that Min-Max normalisation is highly influenced by maximum and minimum values in the dataset, hence there is a need to be careful of outliers.



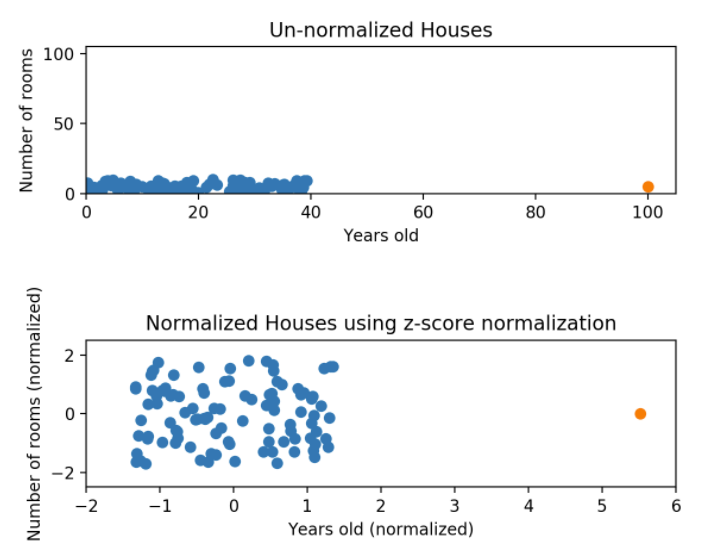
*Retrieved from* [*CodeAcademy*](https://www.codecademy.com/articles/normalization)

1. Z-score normalisation

Z-score normalisation avoids the issue of outliers in the data, as compared to Min-Max normalisation. The formula is as follows:



If a value is equal to the mean of all the values of the variable, then it will be normalised to 0. Else if it is above the mean, it will be positive and if it is below the mean, it will be negative.



*Retrieved from* [*CodeAcademy*](https://www.codecademy.com/articles/normalization)

However, one point to note here is that Z-score normalisation does not produce data with the exact same scale as the original data.

## Exploratory Data Analysis Techniques

### Classification and Dimension Reduction

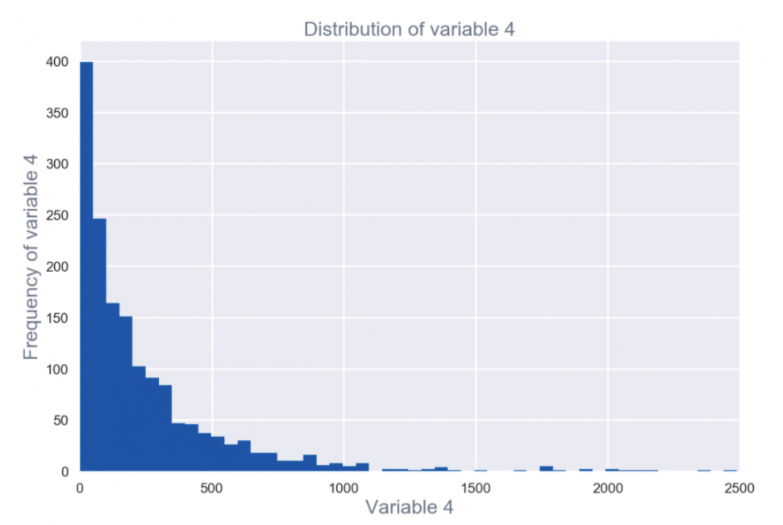
Clustering is used to group different datasets together based on certain common parameters, which allows us to identify groupings in the data. If the data is multi-dimensional, dimensionality reduction techniques like Principal Component Analysis (Refer to [this link](https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad)) or Linear Discriminant Analysis (Refer to [this link](https://machinelearningmastery.com/linear-discriminant-analysis-for-dimensionality-reduction-in-python/)) can be used to provide a projection of a training dataset which we can then work with.

### 

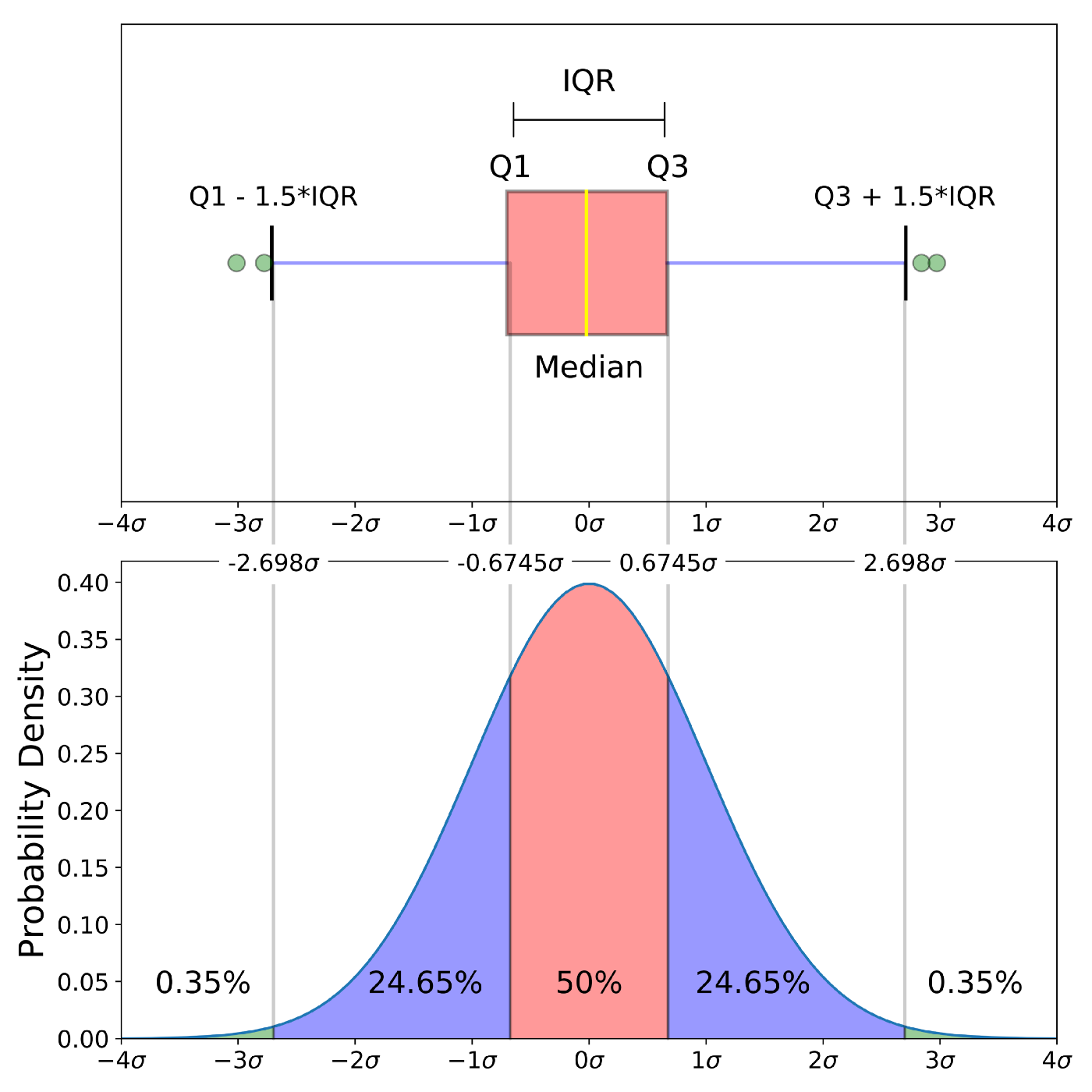
*Retrieved from* [*Upgrad*](https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/)

### Univariate Visualisation

Univariate visualisations provide a summary statistic of each feature in the raw dataset, depending on what type of plot you choose (bar chart, box plot, etc)



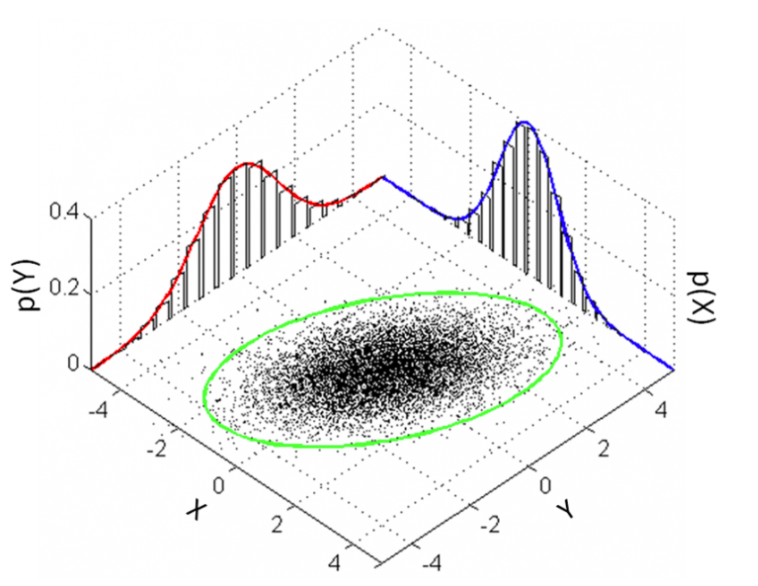
*Retrieved from* [*Upgrad*](https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/)



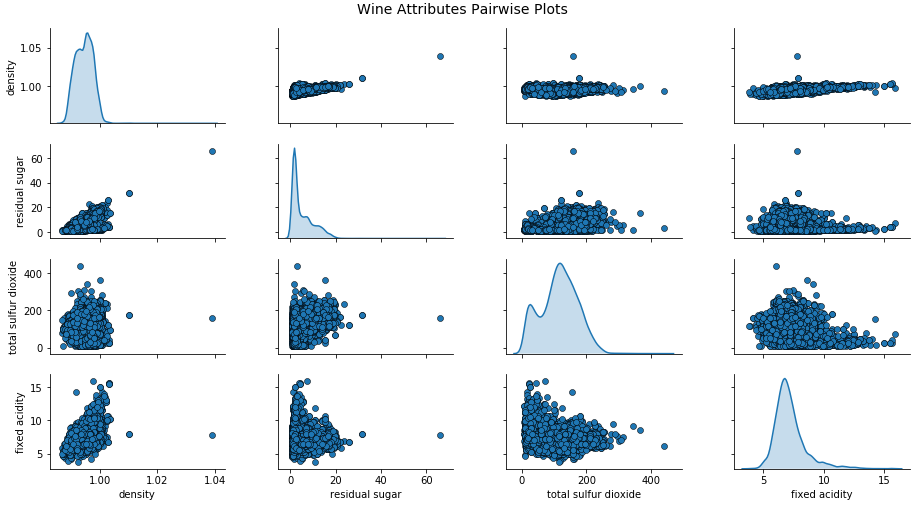
*Retrieved from* [*Towards DataScience*](https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51)

### Bivariate/Multivariate Visualisation

Bivariate plots allow you to visualise and assess the relationship between variables in the dataset. Depending on the type of variable (discrete or continuous), appropriate graphs should be chosen to represent the relationship accordingly.



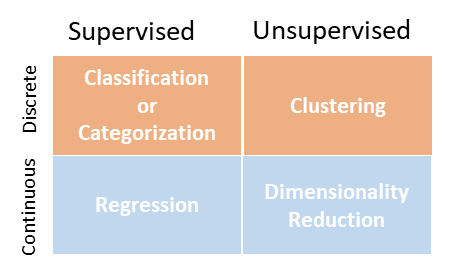
*Retrieved from* [*Upgrad*](https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/)



*Retrieved from* [*Towards DataScience*](https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57)

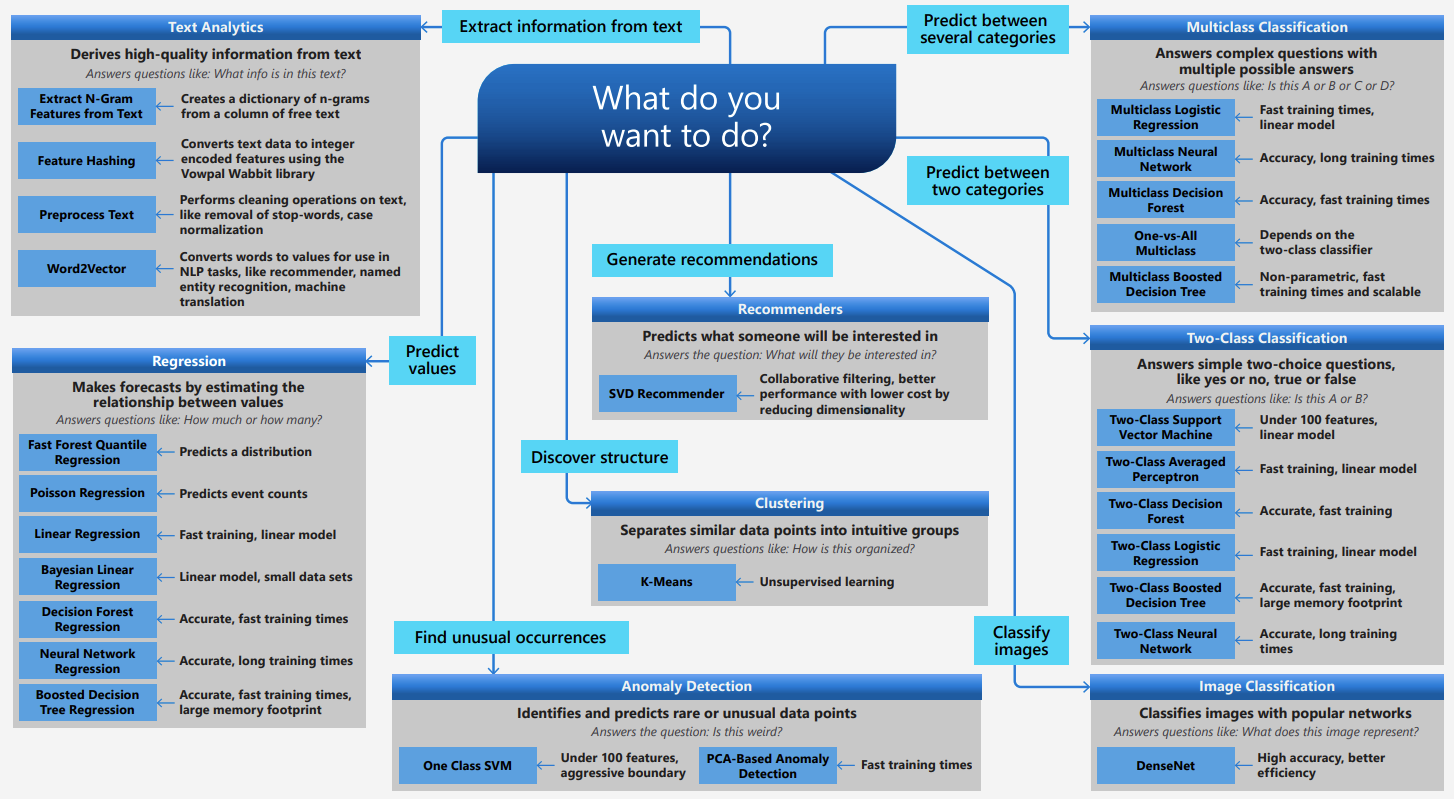
## Machine Learning Modelling

Selecting a ML model depends on the problem at hand, as discussed above in the Data Science problems section.



*Retrieved from* [*Upgrad*](https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/)

For each of these types of problems, there are many different algorithms that can be applied to the data, as shown in the flowchart below.



*Retrieved from* [*Microsoft Azure*](http://aka.ms/mlcheatsheet)

## ML Model Performance Metrics

Performance metrics are necessary to evaluate ML models, so that you can choose the best model that best suits your problem. For each type of problem, there are several metrics that can be used, as shown in the table below:

| [**Regression**](https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/) | [**Classification/**](https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/)  [**Anomaly Detection**](https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/) | [**Clustering**](https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/) | [**Reinforcement Learning**](https://analyticsindiamag.com/metrics-for-reinforcement-learning/) |
| --- | --- | --- | --- |
| Mean Squared Error | Accuracy | Silhouette Coefficient | Dispersion across Time |
| Mean Absolute Error | F-score | Dunn’s Index | Short-term Risk across Time |
| Root Mean Squared Error | Precision |  | Long-term Risk across Time |
| Root Mean Squared Log Error | Recall |  | Dispersion across Runs |
| R Square (refer to Appendix A) | Confusion Matrix |  | Risk across Runs |
| Adjusted R Square | Sensitivity, Specificity |  | Dispersion across Fixed-Policy Rollouts |
|  | ROC & AUC |  | Risk across Fixed-Policy Rollouts |