DATA1002
Project 2

This report reflects the contribution of the members of Group 7:
Jonathan Then 490605077
Benjamin Teo 490580833
Susan Sim 440398956
Mingzhe Li 490347812

PROJECT PART 1

Executive Summary

The role of education plays an imperative role in determining future interests within society, influenced by indicative attributes. In the era of continuous globalisation, it is imperative to continue to redevelop and improve our education system, thus requiring constant analysis of the factors that contribute to the performance of the student scores. Utilizing 'Educational Achievement Dataset' provided by the Organisation for Economic Co-operation and Development, this report aims to draw a relationship between the attributes that affects student performances in Maths across 32 countries. The first section of the report will provide the initial and outcome analysis of the dataset. The second section of the report will illustrate the methodology undertaken to achieve the analysis illustrated in tables and figures. The final section will explain the methodology undertaken to develop a predictive model, created to forecast future educational scores. By exploring the methodology taken to yield the desired results demonstrates how student's education scores are determined or influenced by multiple attributes. This will draw understanding on the student performance and the attributes that affecting it, serving as a potential guide in educational reform.

Domain Situation

To ensure a student succeeds in developing society, requires the development of a strong foundational skill set in mathematics, reading and science. To allow this development requires teachers to obtain the required qualifications to assist with the student's learning development and furthermore ensure enough time is allocated to each individual. The 'Educational Achievement Dataset' obtained from the 'Organisation for Economic Co-operation and Development' emcompasses the 15 year old students scores in Reading, Mathematics and Science across 32 countries. Alongside the student scores data, are the percentages of teachers that have obtained the ISCED level 5a qualifications and the ratio of students per classroom. The overarching dataset is separated by 32 countries, which provides a reflection on the level of education obtained by the students in comparison with neighbouring countries.

Aim

By extracting information from the data, this report aims to illustrate the hypothesis of establishing a correlation behind the student scores and its dependency on teacher aid attributes; the amount of teacher aid and focus time catered to the student. To do so, we aim to find:

1. Firstly, a relationship between the students Mathematics Scores and Student-Teacher Ratio. In finding this initial analysis will allow further analysis to be predicted. In establishing this relationship allows a clear understanding that the student scores correlates with the amount of attention given to students in a classroom, thus affecting the student's learning experience.
2. Secondly, a relationship between the Science Scores, Student-Teacher and Teacher's Qualification. Stemming from the first analysis, we aim to further narrow down attributes that contribute to the nature of the student scores. In doing so, this allows a more comprehensive understanding on the specific factors that contribute to performance, serving as a guide to future educational reforms.
3. Lastly, a relationship between Mathematics Scores, Reading Scores, Student-Teacher Ratio and Teacher's Qualification.

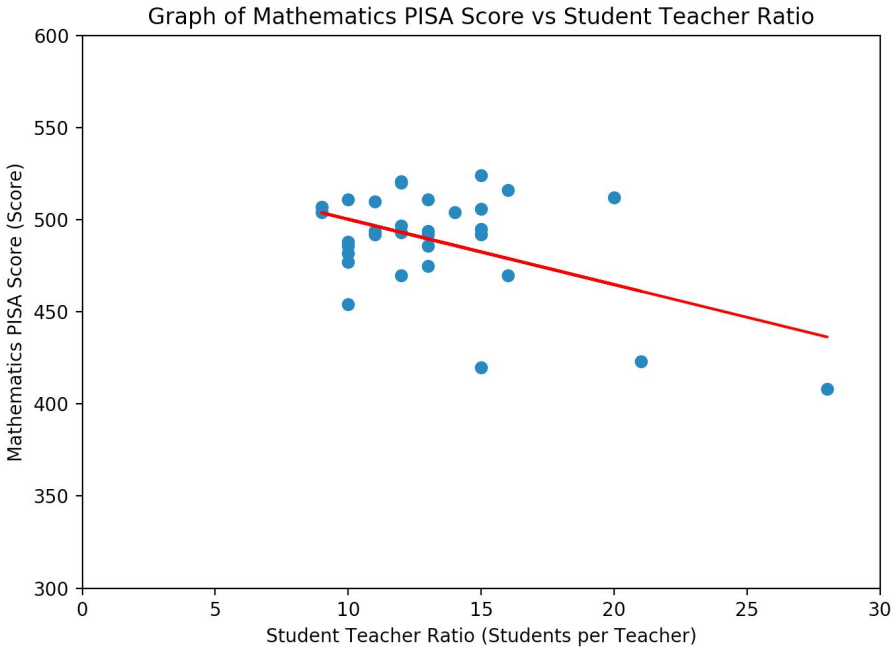
Origin of the Data

The 'Educational achievement and other factors, in PISA 2015' (downloaded on Oct 18 2019, from https://pisadataexplorer.oecd.org/ide/idepisa/

Analysis Results:

Research Question 1:

Figure 1:



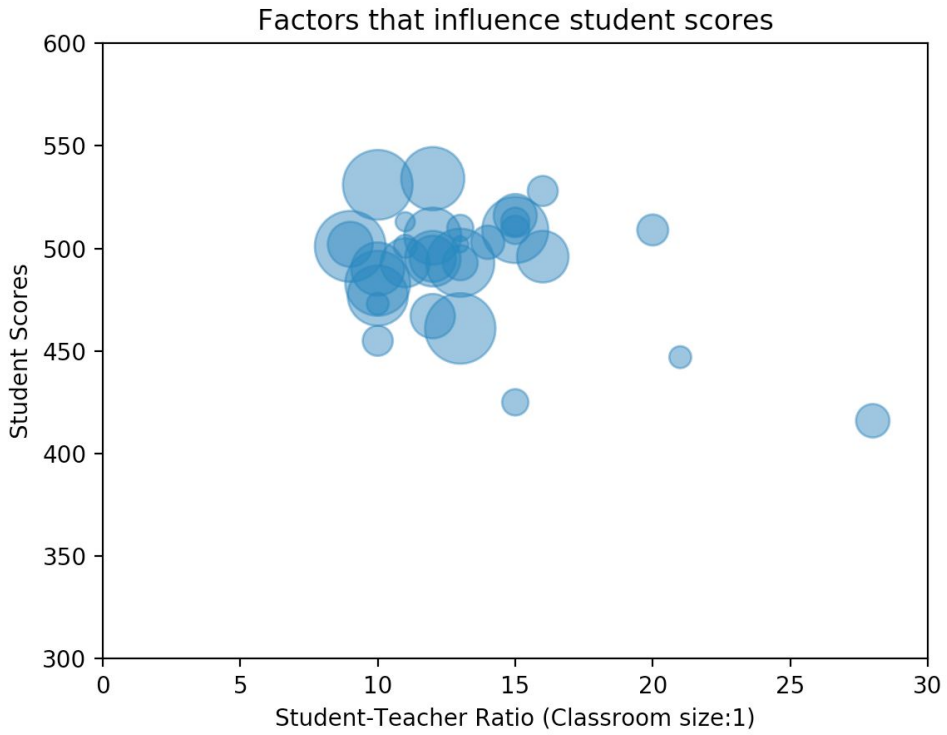
To obtain information about the domain, a scatter plot was chosen to demonstrate the linear regression used to illustrate the relationship between the Student-Teacher Ratio and the Mathematic Scores illustrated in a scatter plot, reflected in Figure 1. The red regression line reflects a declining slope, illustrating the relationship of both attributes; as the Student-Teacher ratio increases simultaneously, there is a decrease in the mathematics scores.

The domain situation reflects a sound relationship ultimately allowing the desired results Research Question 1 to be achieved. The situation reflects a logical explanation on how the Student-Teacher Ratio attribute is a factor that influences the student's Mathematics Scores:

- As teachers cater to a larger classroom of students, less attention is given to each student. This is indicative that when a teacher has a larger classroom size, their responsibility is shared amongst the students therefore it is inevitable for some students to be neglected reflected in their lower obtained scores suffer.
- While there are outliers illustrating extreme results of student scoring lower despite a smaller classroom size or student scoring higher despite having a larger classroom size, there is a general consensus representative of the general trend as reflected in point 1. It is important to note that the outliers are not representative of the general trend and it is assumed that external factors or limiting information may have skewed the outliers, which is not accounted for in the analysis.

Research Question 2:

Figure 2:

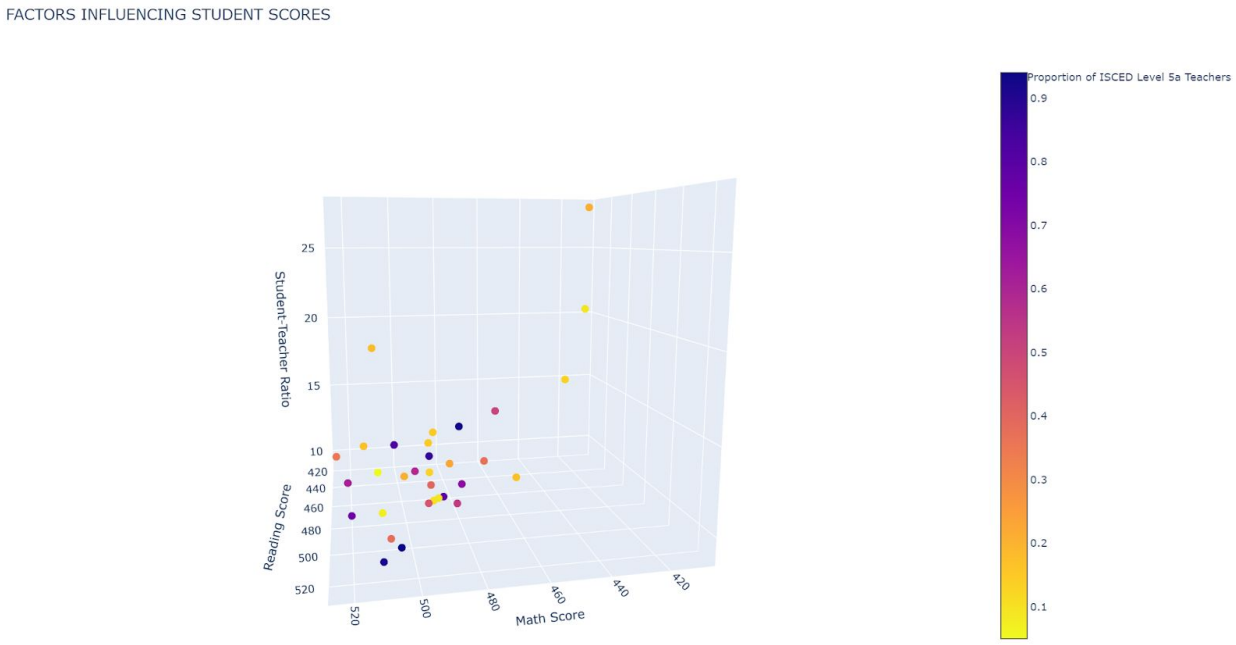


A bubble chart was chosen to demonstrate a three way relationship between the Science Scores, Student-Teacher Ratio and the Teacher's Qualification, as represented in Figure 2; aiming to further explore how educational resources contribute to the learning experience of students, as reflected in their marks. The bubble's size is indicative of the amount of teachers obtaining the Level 5 OECD qualification. The larger the bubbles size represent a higher proportion of teachers obtaining the OECD Level 5 certificate per country, in contrast to the smaller bubbles.

The domain situation reflects a sound relationship ultimately allowing the desired results of Research Question 2 to be achieved.The situation reflects a logical explanation on how the Student-Teacher Ratio and Teacher Qualification attributes contribute to the student's Science Scores:

- In conjunction with a lower Student-Teacher ratio, teachers that obtained the OECD Level 5 qualification contributed to the performance of the students.
- Outliers were present illustrating where a higher proportion of teacher qualifications correlating with a lower science mark, and a lower student-teacher ratio. It is important to note that the outliers are not representative of the general trend and it is assumed that external factors or limiting information may have skewed the outliers, which is not accounted for in the analysis.

Research Question 3:
Figure 3:



A 4-D analysis is performed to further give a visual idea about how both the reading and math scores correlate with the Student-Teacher Ratio and Teacher Qualifications. Looking at the 4-D representation of the data, it is discernible that students with higher scores have a lower Student-Teacher Ratio, and a varying mix of Teacher Qualifications. There are several ways to interpret this data set:

- Student-Teacher Ratio is the only factor in student scores, because there are many different teacher qualifications within that range.
- The student scores are higher because of a low Student-Teacher Ratio, and although there are many different Teacher Qualifications, but a majority of the data points still score above 0.5, which is a significant contributing factor to high student scores as well.

In conclusion, the weight of the student scores are affected by a higher weighted Student-Teacher Ratio and a less weighted Teacher Qualifications.

Recommendation:
With the rise of globalisation increasing the demand for skilled workers, education plays a vital role in developing the skill set of today's students. it is clear from the study conducted that the scale of the classroom plays an important role in contributing to the capacity of the individual to understand knowledge and learning experience. This will allow teachers to properly divide their time across a smaller group of students thus allows solid amount of focus allocated for each student. It is furthermore imperative for teachers to receive a credible qualification as it plays an influence in shaping the student's knowledge and worldviews. The analysis revealed in Research Question 2, reveals the credibility of the Teacher's Qualification, playing a contribution

in the student's learning experience. The more credible the teacher's skillsets are the higher the students will score, in contrast with a lower qualification obtained. This influence is further emphasised in the analysis conducted in Research Question 3; where the Teacher Qualification plays a role in contributing to the scores of the students.

PROJECT PART 2

This section aims to provide a methodology undertaken to achieve the analysis represented in forms of charts and graphs represented in Project Part 1. Python code snippets are attached to illustrate the steps taken to generate meaningful outputs. The main steps taken to provide the analysis are first, cleaning the dataset and second, generating graphs as reflected below. To properly reflect the methodology and communication of both research questions, the second step is divided into two sections.

1. **Cleaning the Dataset:**

```
import csv

csv_reader = csv.reader(open('PISA2015-clean.csv'))
output_file = open('PISA2015-cleaned.csv', 'w', newline='')
csv_writer = csv.writer(output_file)

is_first_line = True
is_sec_line = True
is_third_line = True

for row in csv_reader:
    if is_first_line:
        is_first_line = False
        csv_writer.writerow(['Country', 'Math Score', 'Reading Score',
            'Science Score', 'Index Mother Occupation Status', 'Index Father Occupation Status',
            'Student-Teacher Ratio', 'Proportion of ISCED Level 5a Teachers'])
        continue
    if is_sec_line:
        is_sec_line = False
        continue
    if is_third_line:
        is_third_line = False
        continue
    else:
```

Step One:
The first stage before performing any analysis on the data is to clean the data. The first thing that was done was to delete the first few empty lines in the table that was given. To do this in the python code, was to set the first three lines to true. Therefore when reaching the ‘if’ statements it will set each line that was ‘true’ and set it to ‘false’.

Step Two:
To capture the rest of the rows of data, an ‘else’ statement was used in order to process the dataset without incorporating the header.

```
    else:
        country = row[0]
        math = row[1]
        reading = row[2]
        science = row[3]
        mother = row[4]
        father = row[5]
        stratio = row[6]
        teacher = row[7]
        csv_writer.writerow([country, math, reading, science, mother, father, stratio, teacher])
```

Step Three:

To analyse the dataset, requires assigning values to the required variables. This was done by indexing to determine the correct corresponding columns. For example, row[0] was assigned as country, row[1] was assigned as math, etc.

Step Four:

Finally, we created a new table by using the 'csv writer' to construct a new table with the relevant columns (country, math, reading, science, mother, father, Student-Teacher ratio, teacher qualifications) and rows of data connected with each column.

2. **Generating Graphs:**

Research Question 1 (Two Variables):

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

from sklearn.linear_model import LinearRegression

df = pd.read_csv("PISA2015-cleaned.csv")

X = df["Student-Teacher Ratio"].values.reshape(-1, 1)
Y = df["Math Score"].values.reshape(-1, 1)
linear_regressor = LinearRegression()  # create object for the class
linear_regressor.fit(X, Y)  # perform linear regression
Y_pred = linear_regressor.predict(X)  # make predictions

plt.scatter(X, Y)
plt.plot(X, Y_pred, color='red')
plt.ylim(300, 600)
plt.xlim(0, 30)
plt.title("Graph of Mathematics PISA Score vs Student Teacher Ratio")
plt.xlabel("Student Teacher Ratio (Students per Teacher)")
```

```
plt.ylabel("Mathematics PISA Score (Score)")
plt.show()
```

Step One: To generate graphs and relate data columns more manageable, it was important to use several tools to achieve this. Three toolkits were used pandas, numpy, and matplotlib.pyplot. Pandas is used for quick and flexible managing of relational and labeled data. Next numpy is used for array processing, and since a table was used, numpy was a valuable tool to manage table arrays easier. Finally, pyplot was used to plot graphs based on the data table.

Step Two: To read the data from the csv file, “PISA2015-cleaned.csv” into the python program and setting it to variable df.

Step Three: Set the ‘x’ and ‘y’ labels; x-axis is the ‘Student-Teacher Ratio’ and y-axis is the ‘Math Scores’.

Step Four: To include a linear regression into the graph, it was crucial to assign and label the correct attributes to the ‘x’ and ‘y’ axis, which is ‘student-teacher ratio’ and ‘math scores’.

Step Five: To get a better understanding of how the data relates to each other, required an increase in the range of the ‘x’ and ‘y’ axes; y-axis range is [300, 600] and x-axis range is [0, 30], using the ‘ylim’ or ‘xlim’ function provided by pyplot.

Step Six: Labeled the title of the graph, and titles of the ’y’ and ‘x’ axes.

Step Seven: plt.show() is the code used to display the graph.

Communication of Results:

In Figure 1, the relationship between the student ratio and mathematics scores indicate the performance of students is dependent on the Student-Teacher Ratio ranging between [10,15]. There are several noticeable outliers, specifically with scores around the 400’s range, where the Student-Teacher Ratios are greater than 20. In order to see the overall general trend, required the extension of the x-axis ranging from [0,30] and the y-axis ranging from [300,600]. This extension provides a clearer explanation, allowing the domain to properly include all ranges of scores and student teacher ratio. The outlier presents a data point with a lower student-ratio accompanied with a low math score indicating that there is a possibility that external factors could have singularly attributed to the value of the dataset. When conducting a regression analysis on the dataset, it is evident that the regression line has a negative slope, implying that the Student Teacher Ratio is negatively correlated with Math Scores.

Research Question 2 (Three Variables):

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

from sklearn.linear_model import LinearRegression

df = pd.read_csv("PISA2015-cleaned.csv")

x = df["Student-Teacher Ratio"]
y = df["Science Score"]
z = df["Proportion of ISCED Level 5a Teachers"]
plt.title("Factors that influence student scores")
plt.xlabel("Student-Teacher Ratio (Classroom size:1)")
plt.ylabel("Student Scores")
plt.xlim(0,30)
plt.ylim(300,600)
plt.scatter(x,y,s=z*1000, alpha=0.5)
plt.show()
```

Step One: Most of the steps to generate the graph for question 2 was similar to the steps used to generate the graph for question 1, with a few differences. Again, to generate graphs and relate data columns more manageable, it was important to use several tools to achieve this. Three toolkits were used pandas, numpy, and matplotlib.pyplot. Pandas is used for quick and flexible managing of relational and labeled data. Next numpy is used for array processing, and since a table was used, numpy was a valuable tool to manage table arrays easier. Finally, pyplot was used to plot graphs based on the data table.

Step Two: We had to read the data from our csv file, “PISA2015-cleaned.csv” into the python program and setting it to variable df.

Step Three: For this step, we had to add an extra variable, ‘z’ to get a bubble graph. We set the ‘x’, ‘y’, and ‘z’ labels; x-axis is the ‘Student-Teacher Ratio’, y-axis is the ‘Math Scores’, and z-axis is the ‘Teacher Qualification’.

Step Four: We decided to include a linear regression into the graph, to show the trendline, and in order to show the right regression line, we had to run it against the ‘x’, ‘y’, and ‘z’ data, which is ‘Student-Teacher Ratio’, ‘Science Scores’, and ‘Teacher Qualifications’. Finally, a linear predictor was coded to run a predicted linear regression line based on the ‘x’, ‘y’, ‘z’ data.

Step Five: Beautifying the graph is important especially for anyone that wants to understand what is being done with the data. Therefore we ran lines of code to label the axes. To get a better understanding of how the data relates to each other, we had to increase the range of the ‘x’ and ‘y’ axes; y-axis range is [300, 600] and x-axis range is [0, 30]. To increase the range we had to use the ‘ylim’ or ‘xlim’ function provided by pyplot. However, it was unnecessary for us to change anything with the ‘z’ variable since it is represented as a bubble.

Step Six: Finally, we coded in the title of the graph, and titles of the ’y’ and ‘x’ axes, but it is important to note that the ‘z’ variable which is represented as a bubble is ‘Teacher Qualifications’.

Step Seven: plt.show() is the code used to display the graph.

Methodology Explained:
A further analysis stemming from Research Question 1 is conducted, through the addition of Teacher Qualification attribute, assigned to the ‘z’ variable. To properly illustrate the three way relationship; the Science Scores, represented assigned to the y-axis, Teacher Qualification assigned to the z-axis and the Student-Teacher Ratio assigned to the x-axis. The range for the x-axis is extended from 0 to 50, and the y-axis from 300 to 600. This allows feasible observation on the range for students who perform better at science represented in the increase of the x-axis, correlating with the size of the bubble represented by the z variable. The range of students that obtained higher scores in science ranging from [450,540], has a classroom size between [8,18] correlating with the larger bubble size which indicates the credibility of the teachers.

In comparison to the first graph, the second graph presenting a three way relationship x-y-z variables give a better picture about student scores. The one dimensionality of the x-y graph tells one story, which is that Student-Teacher Ratio is the only factor contributing to student scores. However, with the added ‘z’ variable, it becomes more apparent how teacher qualifications also play a major role in student scores. In the case of the x-y graph, it was considered that the datapoint at (28, 420) was an outlier, but by factoring in the ‘z’ variable, it is evident that the high Student-Teacher Ratio and the lower teacher qualification coefficient are major factors that contribute to lower student scores. Therefore, a reasonable conclusion is that student scores are not only a result of low Student-Teacher Ratios but also attributable to low teacher qualifications.

Research Question 3 (Four Variables):

```
import pandas as pd
import plotly
import plotly.graph_objs as go
```

```
#Read PISA2015-cleaned data from csv
data = pd.read_csv("PISA2015-cleaned.csv")

#Set marker properties
markercolor = data['Proportion of ISCED Level 5a Teachers']

#Make Plotly figure
fig1 = go.Scatter3d(x=data['Math Score'],
                    y=data['Reading Score'],
                    z=data['Student-Teacher Ratio'],
                    marker=dict(color=markercolor,
                                opacity=1,
                                reversescale=True,
                                colorscale='Blues',
                                size=5, autocolorscale=True,
                                showscale=True),
                    line=dict (width=0.02),
                    mode='markers',
                    showlegend=True,
                    name="Proportion of ISCED Level 5a Teachers",
                    textposition="top right")
#Make Plot.ly Layout
mylayout = go.Layout(title="FACTORS INFLUENCING STUDENT SCORES",
                    scene=dict(xaxis=dict(title="Math Score"),
                    yaxis=dict(title="Reading Score"),
                    zaxis=dict(title="Student-Teacher Ratio")))

#Plot and save html
plotly.offline.plot({"data": [fig1],
                     "layout": mylayout},
                     auto_open=True,
                     filename=("4DPlot.html"))
```

Step One: A different technique is used to graph 4 variables. In this case, two libraries were imported, pandas and plotly. Pandas was used as a tool to simplify managing relational and labeled data. This time, numpy was not used; instead, plotly was used to generate the graph based on the data set.

Step Two: We had to read the data from our csv file, “PISA2-15-cleaned.csv” into the python program and set the variable ‘Proportion of ISCED Level 5a Teachers’ as the data point that will include color..

Step Three: The next step is defining the variables to generate a graph. The function go.Scatter3d is used to generate a 3 dimensional graph with three axes, ‘x’, ‘y’, and ‘z’. The math score is represented by the x-axis, reading score represented by the y-axis, and student-teacher ratio represented by the z-axis. The next few lines of data was to set the visual representation of the graph, using a color-scale set to blue. The graph will also include a legend to let the user know what the data points in the graph represent.

Step Four: Next, the axes on the graph were labeled accordingly, with the x-axis being labeled Math Score, y-axis labeled as Reading Score, and z-axis labeled as Student-Teacher Ratio.

Step Five: The final step was to generate the graph and save it into an html format. The html format allows the user to rotate the graph on 3-axes to see how the data points are being represented. Since the computer screen is a 2-D screen, it is difficult to see a 3-D graph and therefore, the ability to rotate the graph is a useful tool for allowing the user to see the data points more clearly in a 3-D environment.

Methodology Explained:
Representing a 3-D graph is difficult using a 2-D space (computer screen) but a 3-D graph can give more visual characteristics to the data. In the 3-D graph with 4 variables, users can clearly see the data points in a 3-D space and how the different factors affect each other. The range for the x-axis (math score) is between [400, 540] and the same goes for the y-axis (reading score) with a range of [400, 540]. For the z-axis, which is representational of student-teacher ratio, the range is between [10, 30]. The data points, which are representative of teacher qualifications is plotted according to the three linking factors, math scores, reading scores and student-teacher ratios. However, the quantitative aspect of these data points is represented in a color scale, ranging from light yellow to deep blue. Each color on the color scale is represented with a number, ranging from [0,1].

The visual representation of a 3-D graph is in many ways more accurate than a 2-D graph because it can show information (data points) on a 3 dimensional scale. In this case, 4 variables are represented in the 3-D graph, and the data shows a somewhat weak but discernible linear pattern. Where the 2-D graphs fail to show a pattern in the data, the 3-D graph excels, and shows a weak but rather obvious linear pattern to the data. Therefore, arguably, a linear regression can be applied to the data set in the 3-D graph, and can be modeled to much greater accuracy versus modeling it in the 2-D graph.

Reflection:
In the earlier stages of the analysis, excel was initially used to perform the analysis on the dataset. While its utilisation allowed filtering and sorting tools (pivot tables, vlookup, match, index) functions, limitations were encountered while performing the simple regression analysis on multiple variables. Therefore Python programming was chosen due to its re-usability, ease of replication nature and limited time consumption, specifically when analysing the second research question. Additionally, Python's functionality to perform predictive modeling was met with ease, specifically through the feasible integration of software packages such as pandas, numpy, pyplot, and sklearn. In contrast to using excel, there are limited functions to allow predictive modeling thus proving to be more challenging.

Prior to demonstrating a relationship for the research questions, required the selection of attributes to display the relationship. The initial impression towards the teacher qualification attribute held no significance to the dataset in comparison to the student-teacher ratio or parental occupation status, it held more influence to the student scores. After much reflection, it was decided that to truly understand student scores, the student-teacher ratio and teacher qualification coefficient variables held the most significance as it was two factors that was part of the education system.

Upon establishing our research questions, it was difficult to select which school subjects would be the most useful in understanding the education quality in the different countries. At first, reading and math was chosen, because these two represented quantitative and literary skills that the education system are assessed on. After deliberation, math and science was chosen, because reading skills can be honed and strengthened without the help of a teacher, but math and science skills required good instruction and discipline that can only be attributed to a professional educator. Further, math and science skills are highly correlated due to its interconnectedness. Therefore, math and science were the two chosen subjects to test if these scores were attributable to the teacher-student ratios or teacher qualification variables.

PROJECT PART 3

This section aims to provide the methodology undertaken to achieve the production of predictive models for Research Questions 1 & 2. Linear Regression and K-nearest neighbor technique were selected as methods to predict future analysis based on our clear data. The first section of the report consists of the overarching method used to conduct the predictive analysis.

Predictive Modelling Steps:

Step One: Import the required libraries needed in the prediction modelling. The modules are: pandas, math and sklearn.

Step Two: Read the data from our cleaned csv file, “PISA2-15-cleaned.csv” into the python program through the use of pandas.

Step Three: The next step is to sieve out the variables that we want to be used in building of the predictive model. With the data in our DataFrame, we want to slice it to input (independent) variables and target (dependent) variables. The value 0.1 means we want to allocate 10% of data for testing. This function will sample records from our data randomly and the seed will make the random process to behave the same across multiple runs. Once that is all done, we use the LinearRegression object and fit our data to build the model.

Step Four: Next, we create a sample that we would like to predict from the given variables. In the first scenario, we used a sample size of Student Teacher Ratio of 1 to predict the Mathematics Score. In the second scenario, we used a sample size of Student-Teacher Ratio of 50 and Proportion of ISCED 5a Master of 0.1 to predict Science Scores.

Step Five: The final step was to see how accurate our prediction model is. We used the Root Mean Squared Error (RMSE) and R-Squared Score as measures to find out the accuracy of the prediction. The lower the RMSE, the greater the accuracy of the model.

Linear Regression

Research Question 1:

```
import pandas as pd
from math import sqrt
from sklearn import linear_model
from sklearn import metrics
from sklearn.model_selection import train_test_split

df = pd.read_csv('PISA2015-cleaned.csv')

X = df.values[:, 6].reshape(-1, 1)      # slice dataFrame for Student Teacher Ratio
y = df.values[:, 1].reshape(-1, 1)      # slice dataFrame for Mathematics PISA Scores

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
regr = linear_model.LinearRegression().fit(X_train, y_train)
```

```
# Let's create one sample and predict the Mathematics PISA Score
sample = [1]        # A sample - Student Teacher Ratio of 1
print('----- Sample case -----')
for column, value in zip(list(df)[6], sample):
    print(column + ': ' + str(value))
sample_pred = regr.predict([sample])
print('Predicted Mathematics Score:', int(sample_pred))
print('-----------------------')

# The coefficients
print('Coefficients:')
print(regr.coef_)
# Use the model to predict y from X_test
y_pred = regr.predict(X_test)
# Root mean squared error
mse = metrics.mean_squared_error(y_test, y_pred)
print('Root mean squared error (RMSE):', sqrt(mse))
# R-squared score: 1 is perfect prediction
print('R-squared score:', metrics.r2_score(y_test, y_pred))
```

Output:

```
$ python regline.py
----- Sample case -----
S: 1
Predicted Mathematics Score: 529
-----------------------
Coefficients:
[[-3.40049597]]
Root mean squared error (RMSE): 16.037639117523444
R-squared score: -1.0064816652475335
```

Research Question 2:

```python
import pandas as pd
from math import sqrt
from sklearn import linear_model
from sklearn import metrics
from sklearn.model_selection import train_test_split

df = pd.read_csv('PISA2015-cleaned.csv')

X = df.values[:, 6:8]      # slice dataFrame for Student Teacher Ratio and Proportion of ISCED 5a Master Teachers
y = df.values[:, 3]     # slice dataFrame for Science PISA Scores

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
regr = linear_model.LinearRegression().fit(X_train, y_train)

# Let's create one sample and predict the Science Score
sample = [50, 0.1]        # A sample - Student Teacher Ratio of 50 and Proportion of ISCED 5a Master Teachers of 0.1
```

```python
print('----- Sample case -----')
for column, value in zip(list(df)[6:8], sample):
    print(column + ': ' + str(value))
sample_pred = regr.predict([sample])
print('Predicted Science Scores:', int(sample_pred))
print('-----------------------')

# The coefficients
print('Coefficients:')
print(regr.coef_)
# Use the model to predict y from X_test
y_pred = regr.predict(X_test)
# Root mean squared error
mse = metrics.mean_squared_error(y_test, y_pred)
print('Root mean squared error (RMSE):', sqrt(mse))
# R-squared score: 1 is perfect prediction
print('R-squared score:', metrics.r2_score(y_test, y_pred))
```

Output:

```
$ python regline2.py
----- Sample case -----
Student-Teacher Ratio: 50
Proportion of ISCED Level 5a Teachers: 0.1
Predicted Science Scores: 401
-----------------------
Coefficients:
[-2.40487515  3.01183267]
Root mean squared error (RMSE): 21.80007004513536
R-squared score: -0.7396222520166846
```

Linear Regression as a Predictive Model:

The first tool used to conduct a predictive model for Research Question 1 and 2, is the Linear Regression model, create a foundation in developing a simplistic understanding of the trends in the data set. Linear regression was used to conduct a predictive analysis on how student-teacher ratios can affect student's mathematical scores and furthermore how student-teacher ratio and teacher's qualification can affect the student's science scores. This method allowed simplicity when implementing and deciphering due to the generated straight line prediction compared to other more modelling methods. While, linear regression does not take into consideration outliers, which may be helpful, but in the scope of our research, the nature of the outliers play a crucial role in determining influential contributions therefore it is imperative for it to be considered.

Limitations:

Limitations encountered when modeling the data set is indicated by the data closely clustered instead of having some sort of linearity, creating the assumption that data points located outside the cluster to be considered outliers. This was encountered in our dataset with the R-squared generating a -0.7 value. This immediately raises concerns on the incomplete collection process of the dataset, given there are 195 countries in the world, the dataset of 32 countries is not enough to generate a valid prediction. In being inclusive of all countries, allows a succinct trend when using regression analysis, therefore the method was considered weak because it did not perform a reliable data based prediction.

K-Nearest Neighbor Regression

Research Question 1:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
neigh = neighbors.KNeighborsRegressor(n_neighbors=4).fit(X_train, y_train)

# Let's create one sample and predict the Mathematics PISA Score
sample = [1]        # A sample - Student Teacher Ratio of 1
print('----- Sample case -----')
for column, value in zip(list(df)[6], sample):
    print(column + ': ' + str(value))
sample_pred = neigh.predict([sample])
print('Predicted Mathematics Score:', int(sample_pred))
print('-----------------------')

# Use the model to predict X_test
y_pred = neigh.predict(X_test)
# Root mean squared error
mse = metrics.mean_squared_error(y_test, y_pred)
print('Root mean squared error (RMSE):', sqrt(mse))
# R-squared score: 1 is perfect prediction
print('R-squared score:', metrics.r2_score(y_test, y_pred))
```

Output:

```
$ python kline.py
----- Sample case -----
S: 1
Predicted Mathematics Score: 495
-----------------------
Root mean squared error (RMSE): 12.987975207860538
R-squared score: -0.3159434422233056
```

Research Question 2:

```
neigh = neighbors.KNeighborsRegressor(n_neighbors=28).fit(X_train, y_train)

# Let's create one sample and predict the Science Score
sample = [50, 0.1]        # A sample - Student Teacher Ratio of 50 and Proportion of ISCED 5a Master Teachers of 0.1
print('----- Sample case -----')
for column, value in zip(list(df)[6:8], sample):
    print(column + ': ' + str(value))
sample_pred = neigh.predict([sample])
print('Predicted Science Scores:', int(sample_pred))
print('-----------------------')

# Use the model to predict X_test
y_pred = neigh.predict(X_test)
# Root mean squared error
mse = metrics.mean_squared_error(y_test, y_pred)
print('Root mean squared error (RMSE):', sqrt(mse))
# R-squared score: 1 is perfect prediction
print('R-squared score:', metrics.r2_score(y_test, y_pred))
```

Output:

```
$ python kline2.py
----- Sample case -----
Student-Teacher Ratio: 50
Proportion of ISCED Level 5a Teachers: 0.1
Predicted Science Scores: 489
-----------------------
Root mean squared error (RMSE): 25.237014240101807
R-squared score: -1.3313910327342993
```

K-Nearest Neighbor as a Predictive Model:

The second tool used to conduct a predictive model for Research Question 1 and 2, is the K-Nearest Neighbour, due to the unreliability of the Linear Regression method. The k-nearest neighbor technique was used due to the curved model generated, thus allowing an accurate prediction for the attributes that affect the student's maths and science scores. When using linear regression, the regression line gravitated towards zero which in the nature of the data set, it is improbable for students to get zero scores for an entire country even if they have extremely high student-teacher ratios or low teacher qualifications. This is where the k-nearest neighbor method was reliable based on its curve towards a more reasonable prediction.

Limitations:

However, some limitations with K-Nearest Neighbor technique was encountered. The K-nearest neighbor technique did not follow logical conventions when modeling student scores against Student-teacher ratios. N-neighbor is a variable that determines how much data is used to predict the model. For example, a small n-neighbor factor limits the data to a small subset, which means that a prediction based on it would use the data within that subset and skew the model based on that bias. However, if a high n-neighbor factor, for a data set that was unique like the one being used, the n-neighbor would predict a model and draw a conclusion based on the majority of the data which again, will include some sort of bias. This was illustrated when assigning 1 to the n-neighbour which received an output of 420, however when increasing the n-neighbour to 28. the output generated was 480. This demonstration reveals the skewed nature of our dataset. With these limiting factors, it was again concluded that perhaps a different predictive model could be used to model the data set because of its uniqueness.