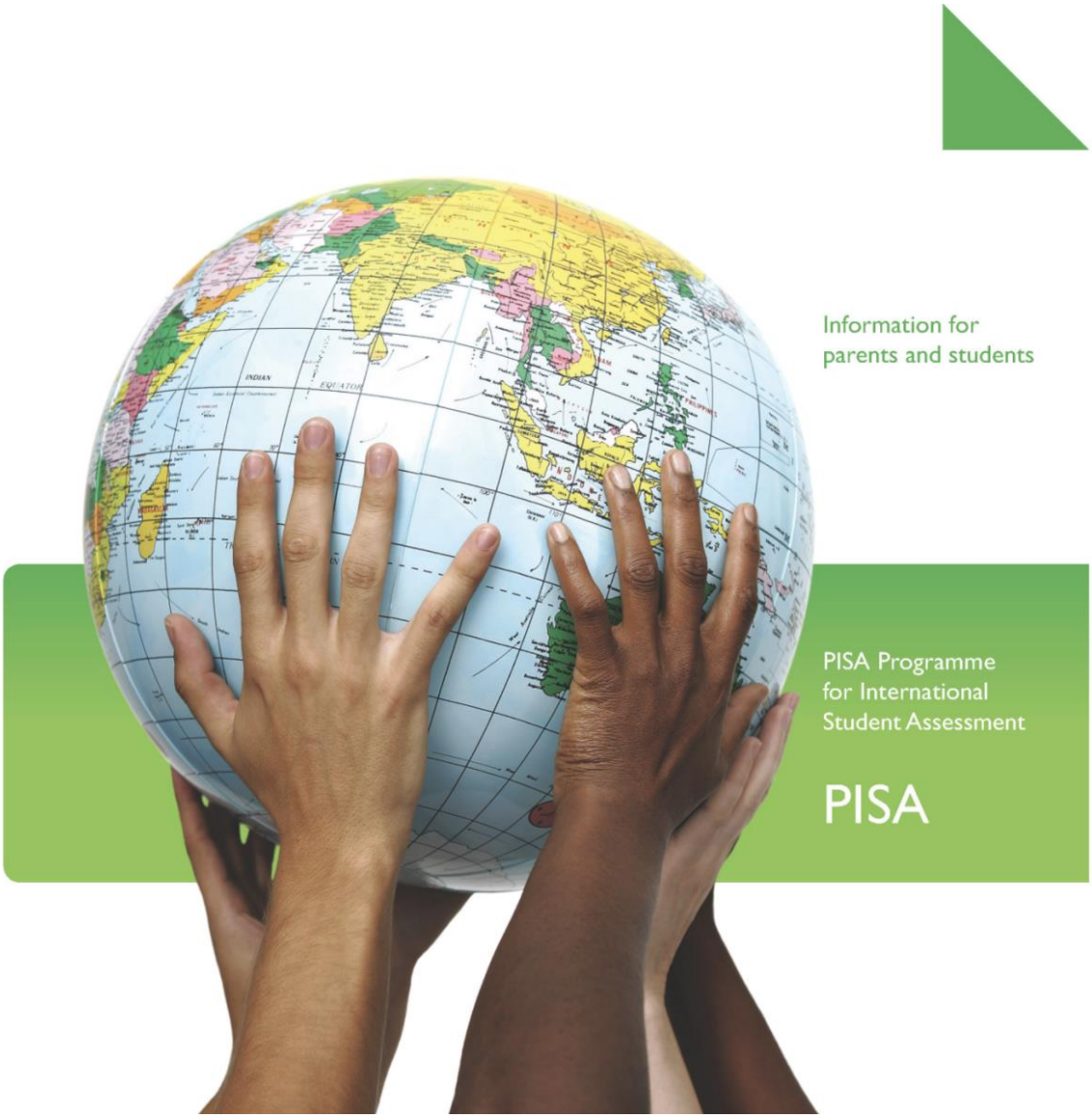


DATA1002 Project Stage 2 Group 8

470000946 Xiaobing Wang

470021332 Shiyu HU

490443505 Sizhuang Kang

Student Disclaimer

The work comprising this report is substantially our own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work. I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure. I understand that failure to comply with the University of Sydney Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under chapter 8 of the University of Sydney By-Law 1999 (as amended).

1. INTRODUCTION

The Programme for International Student Assessment (PISA) provides valuable information for various factors that may influence students' capabilities. Data analytics of PISA enable people in educational systems to draw insights by comparing with others and fine-tune educational policies. This project aims to apply data analysis skills on a PISA dataset to investigate on how occupational status of parent and some characteristics of teachers may correlate with students' capabilities in different subjects.

A picture says more than a thousand words. This saying implies the power of visualisation. In this regard, python is a powerful tool to learn. In this project, several packages like NumPy, Pandas, scipy, Matplotlib are found convenient and helpful for data analysis. Pandas package is used to read, group and process data; and Matplotlib is used for visualisation of quantitative data. In particular, scatter plots are drawn in subplots to compare and evaluate whether occupational status index of mother is higher than that of father, for performance scores of students in different indicators.

2. OBJECTIVES AND TARGET AUDIENCES

OBJECTIVES:

1. Investigate whether occupational status of mother influence students' score more than that of father?

2. Investigate whether family has better potential in influencing students' score more than teacher?

3. Provide insights for stakeholders in high school education, and possibly the general public who have interest in the performance of young students' ability in reading, mathematics and science across the world.

4. Enhance learning in data processing using Pandas and Matplotlib packages, for students with an interest in learning IT approaches to data analysis

This data analysis project can provide important insights for the government, education administrative departments and academic researchers on the regional and national education development level, and help the public understand if there is any trend in high school education. Moreover, this project could demonstrate data processing and visualisation skills to help students develop an interest in learning IT approaches to data analysis.

3. OVERVIEW OF DATA

3.1 DOMAIN KNOWLEDGE OF DATA:

PISA assessments are mandatory under the Australian Education Act, 2013("Important Notices - De La Salle College", 2018). These assessments established a profile of what 15-year-old students can do and how they perform as learners. Quantitative analysis of these assessments data could enhance our understanding about students of such age.

The Data set used is PISA2015-forStage2.csv and is collected by National Centre for Educational Statistics (NCES) in the United States in Fall 2015("PISA", 2018). This source is trustworthy and is prestigious in being objective and useful to reveal the relationship between scores and various factors such as family influence and teacher-student ratio.

3.2 USEFUL SKILLS FOR STATISTICS SUMMARY

This data set has been cleaned and assessed young students' average performances in science, reading, and mathematics at 31 countries, recorded every 3 years from year 2000 to 2015. The data set contains 31 observations of 9 variables.

In data analysis, it is always useful to get information about the statistics of your data. In Pandas, that can be quickly done using df.shape and df.head() after loading the data using pandas package.

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

df = pd.read_csv("PISA2015-forStage2.csv", skiprows = 2)
df.shape
```

```
(32, 9)
```

And the first 5 lines of the data are shown below to provide a glimpse of the data formats and structure:

	Jurisdiction	Region	Averages for PISA mathematics scale: overall mathematics, age 15 years	Averages for PISA reading scale: overall reading, age 15 years	Averages for PISA science scale: overall science, age 15 years	Averages for index mother occupational status, age 15 years	Averages for index father occupational status, age 15 years	Averages for student-teacher ratio, age 15 years	Averages for index proportion of all teachers isced level 5a master, age 15 years
0	Australia	Oceania	494	503	510	52	46	13	0.13
1	Austria	Europe	497	485	495	45	45	12	0.58
2	Belgium	Europe	507	499	502	47	46	9	0.38
3	Canada	Americas	516	527	528	53	48	16	0.17
4	Chile	Americas	423	459	447	40	41	21	0.09

As shown, it is easily observed that each row would have 9 fields (columns) of values.

- "Jurisdiction", refers to country the student is in.
- "Region", refers to the larger geographic area the student is in
- 3rd to 5th column refers to averages score in mathematics or reading or science.
- 6th and 7th column refers to occupation status index of mother and father
- 8th column refers to student-teacher ratio
- 9th column refers to proportion of all teachers isced level 5a master

3.3 RESEARCH QUESTIONS

- Research question 1:
  whether mother's occupational status influence score more than that of father?

Family education is an important part of the process of education. In the family, different characteristics such as family structure, family culture could have tremendous influences on the academic scores of students. In this report, the research is narrowed down and focus on the occupational status of parent.

Research is done by comparing the occupational status index of mother with that of father to gauge the maternal and paternal influences as students' PISA scores in 3 subjects across 31 countries. For the same occupational status, if the student with a mother of such an occupational status has a higher score than the score of another student with a father of such an occupational status, it would be taken as mother's occupational status influence more than father.

- Research question 2: whether family better potential in influencing more than teacher?

Research on influence of teacher ratio/ master proportion on the PISA scores of 3 subjects across 31 locations

Besides family education, teacher also plays an important role in helping the students perform better in academics. Differences in characteristics of the teachers, such as teacher's availability, patience, teaching skills and literacy level, may contribute to variances in students' learning capabilities. In this report, the research is narrowed down and focus on the student-teacher ratio and proportion of all teachers isced level 5a master as important characteristic of teachers that interest our project.

Research is done by comparing how the percentile of the occupational status of parents influence students' scores, with how the percentile of teacher characteristics influence students' scores, to gauge whether family occupational status differences influences more than teacher characteristics in contributing to students' PISA scores in 3 subjects across 31 countries. For the same percentile, if the student with such percentile of parents' occupational status has a higher score than the score of another student with such percentile of teacher characteristics, it would be taken as family has a better potential of helping students score better.

4. METHODOLOGY OF ANALYSIS

4.1 Select data and extract all columns and rename with simpler terms for the convenience of use

```
# extract all columns and  rename with simpler terms for the convenience of use
df.columns = ['LOCATION', 'REGION', 'MATHEMATICS', 'READING', 'SCIENCE', 'MOTHER', 'FATHER', 'TEACHER_RATIO', 'TEACHER_DEGREE']
```

4.2 Use filtering, grouping to construct list of scores for different indicators

```
indicators = ['MATHEMATICS', 'READING', 'SCIENCE']
titles = ["PISA – Maths", "PISA – Reading", "PISA – Science"] # titles of 3 subplots
for i in range(len(indicators)):
    # extract column with the target indicator
    y = df[indicators[i]]
    x1 = df["MOTHER"]
    x2 = df["FATHER"]
```

4.3 Prepare the data for 3 subplots, by plotting scatter plot of score of 3 subjects against maternal/paternal influence for different locations. Adjust bar plots and set parameters

```
ax = fig.add_subplot(3, 1, i + 1) # draw 3 subplots in 1 columns

# draw subplot and set parameters
# markers can be 'o', 'v', '^', '<', '>', '8', 's', 'p', '*', 'h', 'H', 'D', 'd', 'P', 'X'.
ax.scatter(x1, y, label = "Mother", marker='o')
ax.scatter(x2, y, label = "Father",  marker='x')

ax.set_xlabel("Mother/Father Occupational Status", fontsize = 15)
ax.set_ylabel("Score", fontsize = 15)
ax.set_title(titles[i], fontsize = 15)
ax.legend(loc = 2)
plt.axhline(np.mean(y), color = "orange")
```

4.4 Encode location as annotation on points

```
# put country annotations
for ind, txt in enumerate(country):
    ax.annotate(txt, (x1[ind], y[ind]))
for ind, txt in enumerate(country):
    ax.annotate(txt, (x2[ind], y[ind]))
```

4.5 Display p-value to see if there is a difference for occupational status for the same score

```
    (t_stats, p_value) = stats.ttest_ind(x1/y, x2/y)
    t_list.append(t_stats)
    p_list.append(p_value)

fig.suptitle("Performance vs Mother/Father Influence performance by different inidcators", fontsize = 18, y = 0.915)
fig.text(0.15, 0.64, ("t-statistic: " + str(t_list[0])), fontsize = 14)
fig.text(0.15, 0.63, ("p-value: " + str(p_list[0])), fontsize = 14)
fig.text(0.15, 0.37, ("t-statistic: " + str(t_list[1])), fontsize = 14)
fig.text(0.15, 0.36, ("p-value: " + str(p_list[1])), fontsize = 14)
fig.text(0.15, 0.11, ("t-statistic: " + str(t_list[2])), fontsize = 14)
fig.text(0.15, 0.10, ("p-value: " + str(p_list[2])), fontsize = 14)
```

5. RESULTS OF ANALYSIS

5.1 RESEARCH QUESTION 1: THE MATERNAL AND PATERNAL INFLUENCES ON THE PISA SCORES OF 3 SUBJECTS ACROSS 31 LOCATIONS

5.1.1 Findings: Distribution of scores in Reading, Math and Science at different maternal and paternal influences

There is a total of 3 subplots. 3 Scatterplots of score against maternal/paternal influence are drawn.

We encode locations as annotations for points on graph, y-position denotes for score for 1 subject (be it reading, mathematics or science), x-position denotes for maternal/paternal influence using the occupational status index

INTERPRETATION:

In a general trend for all indicators, it is found that mothers tend to have lower occupational status, as denoted by more orange points than blue points at lower occupational status ranges.

At the range of lower occupational status, more orange points are observed. Father (the orange points) seems to have higher influence than mother (the blue points). This kind of higher influence is especially true for scores in Mathematics and Science but not that apparent for score in Reading.

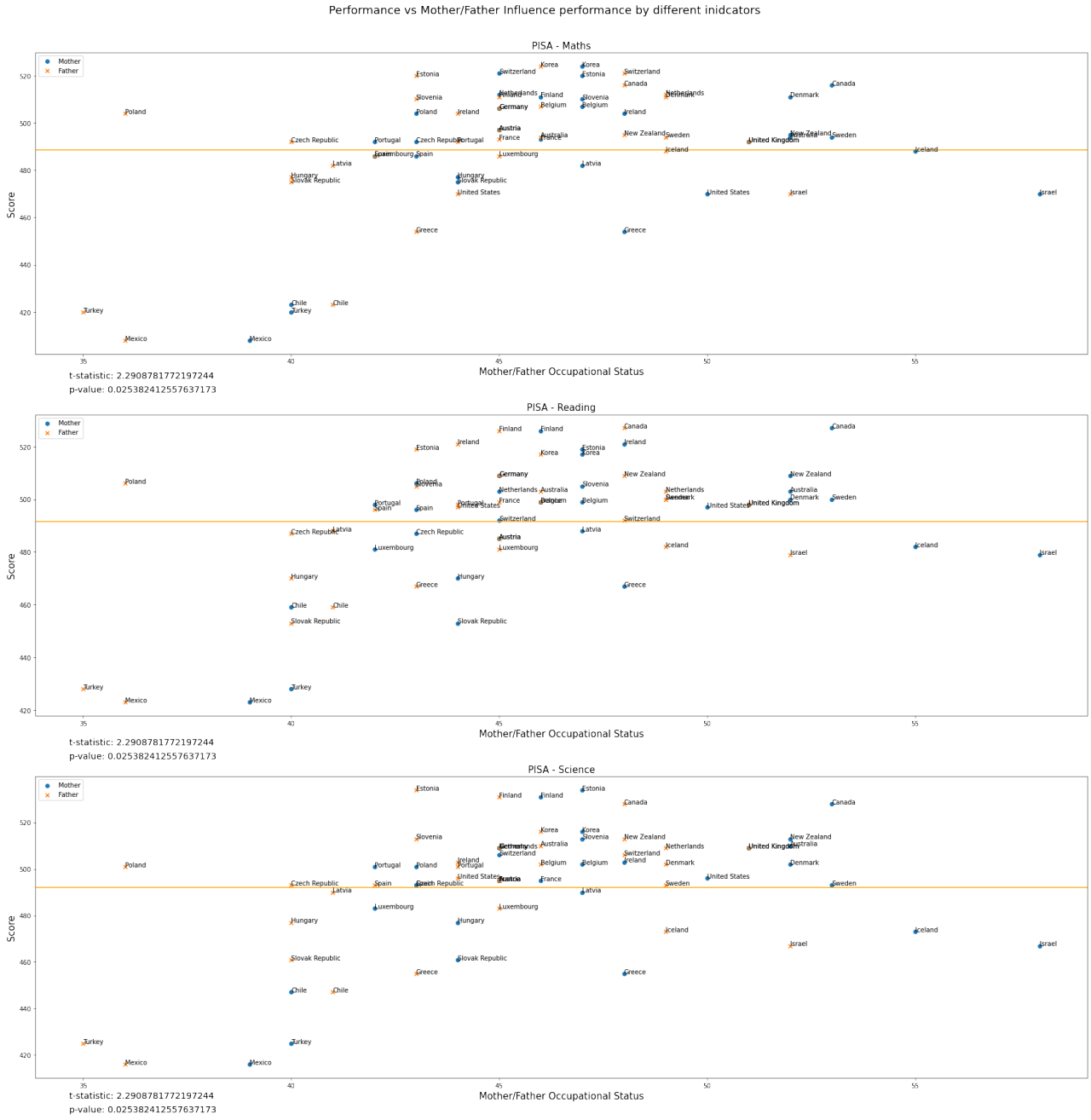
For instance, students with mother occupational status as 40 will have higher score than students with father occupational status as 40, as shown by the fact that mother point for Hungary is well above father point for Chile.

At middle range of occupational status, mother and father seems to have comparable influence, as shown by the fluctuating points in orange and blue.

At higher occupational status, there is a trend that they are the mothers, as shown by the prevalent blue points in the range of high occupational status, but few orange points in that range. Due to lack of points representing mothers, the influence is hard to find.

As shown by the p-values, which are all less than 0.05, the differences are actually not clear. This is maybe due to outliers or randomness of the dataset. Also, most of the parents are within middle range of occupational status, and within this range, the influence is quite comparable.

VISUALISATION

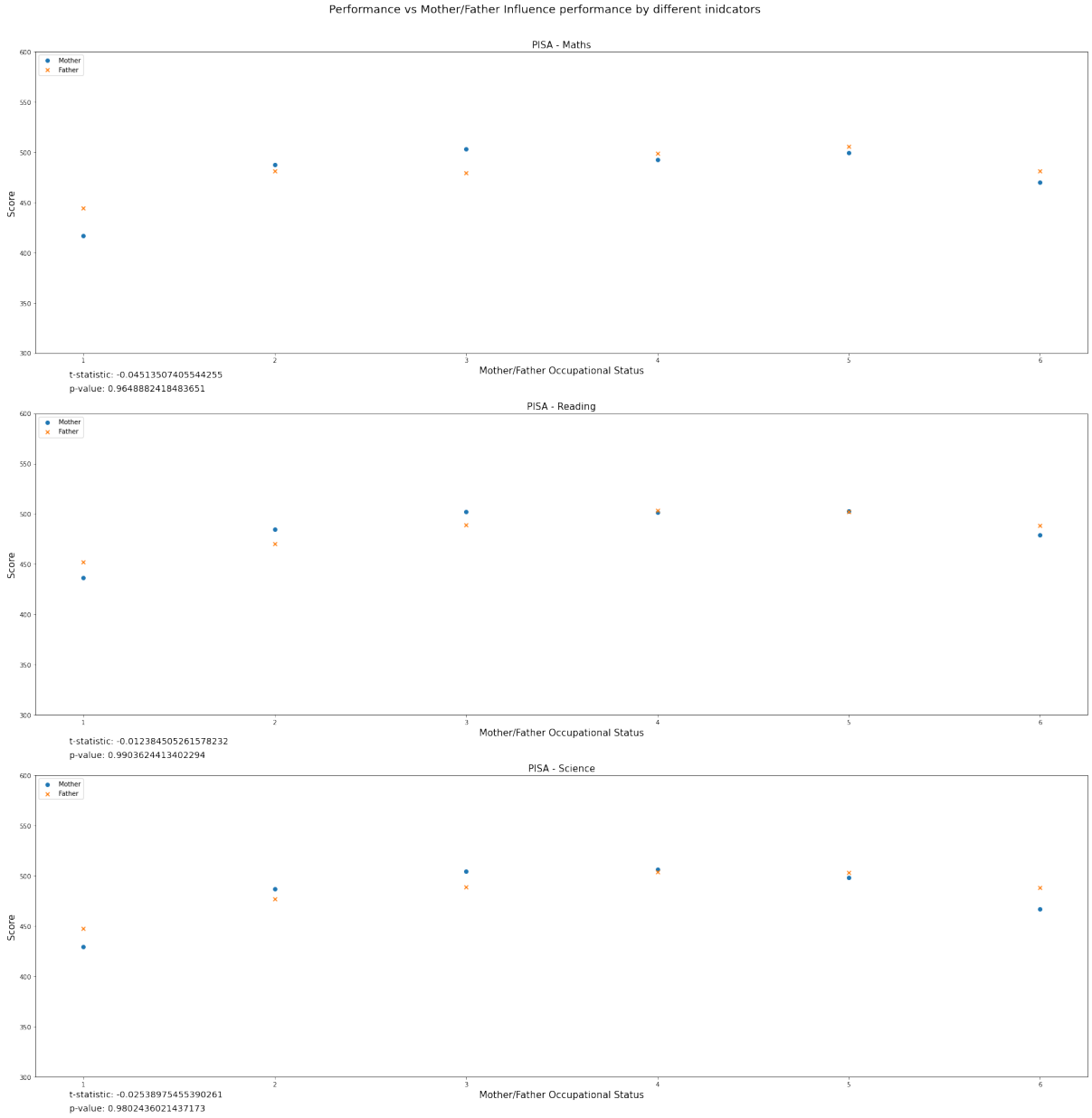


BINNING FOR CLEARED TREND

Binning is used to investigate further for a clearer trend. Scores are grouped by bins of occupational status. We use 6 bins for the convenience of visualisation.

```python
K = 6
def Kbins(data, K):
    result = []
    min_val = min(data)
    max_val = max(data)
    width = max_val - min_val
    for val in data:
        normalized_score =(val-min_val) * (K-1) / width + 1
        result.append(round(normalized_score))
    return result
df['MOTHER_BIN'] = Kbins(df['MOTHER'], 6)
df['FATHER_BIN'] =  Kbins(df['FATHER'], 6)
mother = df.groupby('MOTHER_BIN').mean()
father = df.groupby('FATHER_BIN').mean()
# y1 = mother[indicators[i]]
# y2 = father[indicators[i]]
```

VISUALISATION



INTERPRETATION:

As shown by the p-values, the differences in influence is clear, but the correlation is apparently not linear.

It is visualised that father has a higher influence when it comes occupational status of the parent being quite low or quite how. And then for the occupational status in bin 2 to 3, mother has a higher influence. For occupational status in bin 4 and 5, both parents have comparable influence.

These observations are true for all 3 indicators.

5.2 DISTRIBUTION OF SCORES IN READING, MATH AND SCIENCE AT DIFFERENT FAMILY/TEACHER INFLUENCE PERCENTILE

Normalized scores are calculated for family/teacher influence using sigmoid function of sum of z scores of factors such as teacher-student ratio and teacher-degree-proportion, as shown by the code below:

Since the values in the occupational status are in a small range, we decided to use the Z-score for the influential factors, which is standardised score assuming normal distribution

```
def z_score(df):
    df.columns = [x + "_zscore" for x in df.columns.tolist()]
    return (df - df.mean())/df.std(ddof=0)
factors =  df.iloc[:, [5, 6, 7, 8]].astype(float)
z_scores = z_score(factors)
print("🙊 ---------- the z_scores ---------- 🙊")
z_scores
```

```
🙊 ---------- the z_scores ---------- 🙊
```

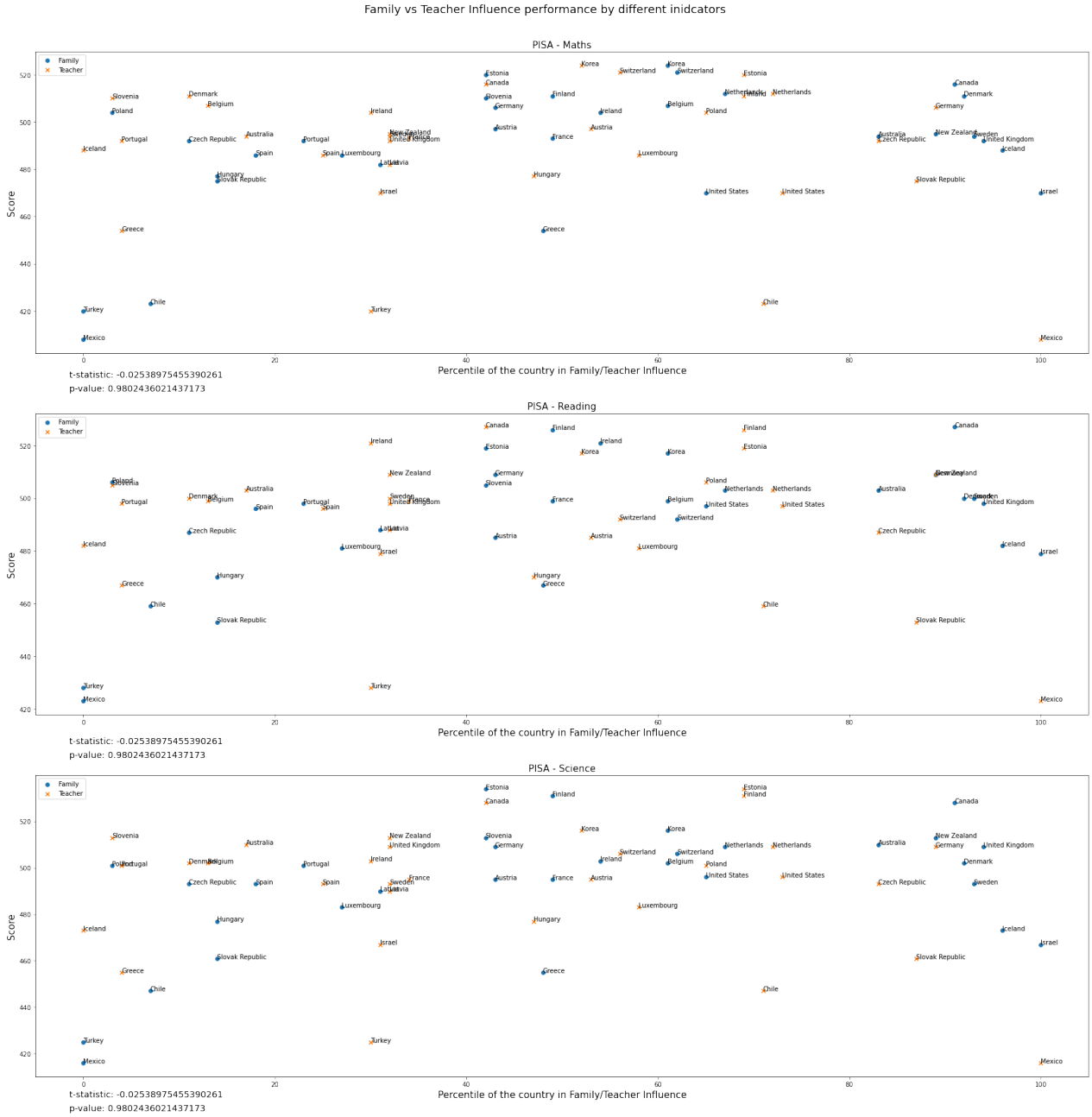
	MOTHER_zscore	FATHER_zscore	TEACHER_RATIO_zscore	TEACHER_DEGREE_zscore
0	1.140374	0.404520	-0.080951	-0.944520
1	-0.407770	0.164804	-0.339995	0.595807
2	0.034557	0.404520	-1.117125	-0.088783
3	1.361538	0.883951	0.696179	-0.807602

```
import math

def sigmoid(x):
  return 1 / (1 + math.exp(-x))

def normalize(data):
    result = []
    min_val = min(data)
    max_val = max(data)
    for val in data:
        normalized_score =(val-min_val) * 100 / (max_val - min_val)
        result.append(round(normalized_score))
    return result

# compute sum for family factors and teacher factors, and then normalize the sums
family = (z_scores['MOTHER_zscore'] + z_scores['FATHER_zscore']).apply(sigmoid)
df['FAMILY'] = normalize(family)
teacher = (z_scores['TEACHER_RATIO_zscore'] + z_scores['TEACHER_DEGREE_zscore']).apply(sigmoid)
df['TEACHER'] = normalize(teacher)
```

Three Scatterplots of score against family/teacher influence are drawn. We encode locations as annotations for points on graph, y-position denotes for score for 1 subject (be it reading, maths or science), x-position denotes for percentile of family influence using the occupational status index, or the percentile of teacher influence using sum of normalized scores.

As such, there is a total of 3 subplots, each contains a scatter plot of score against percentile of family influence and a scatter plot of score against percentile of teacher influence.

VISUALISATION



INTERPRETATION:

In a general trend for all indicators, it is found that the influence is apparent (as shown by the p-values) but is not linear.

At ranges of lower percentiles, more orange points (such as countries in Iceland and Slovenia) are observed than blue points. Students with teacher influence in lower percentile could still have high scores, but students having family influence in lower percentile rarely have high scores. This is true for all indicators. This enlighten us that teachers with average ratio or degree proportions could still cultivate 15-year old students with high capabilities.

At ranges of middle percentiles and high percentiles, family and teacher tend to have comparable influence.

6. EVALUATION OF CHARTS

6.1 DESCRIPTION OF ENCODING

In all three visualisation, we encode locations as annotations for points on graph, y-position denotes for score for 1 subject (be it reading, mathematics or science), x-position denotes for maternal/paternal influence using the occupational status index, or family/teacher influence using normalized score of sum of z-scores

6.2 EXPLANATION OF ENCODING DESIGN CHOICE

Generally, the quality of the chart is excellent to a large extent. Despite it seems messy at first sight, but we could observe a complex trend at different phases using scatter plots (which are easy to understand for the audience, and yet useful to draw complex insights even when the correlation is not that straightforward).

Scatter plot is chosen as it is the most commonly used useful chart to display the correlation between two variables. The audience could easily see if there is a trend between X and Y, for instance, as X increases, we can easily observe if Y increases.

P-values are calculated and displayed explicitly to see if the differences in parent occupation, or difference in family/teacher influences matter. As such, the audience could easily comprehend if these differences contribute to differences in scores or just that differences in scores are random variations.

The scores are chosen as the Y-variable as they are the target dependent variables. Subplots are chosen as such parallel comparisons allow us to use cross-reference and compare to see that if there is a trend only in certain indicators or that if there is a trend that affects all indicators.

Normalised scores are used instead of introducing multiple X-variables. Related variables are converted to z-scores and summed up assuming comparable weightages. Sigmoid function is used to simplify problems in logistics regression. These processing allows the use of scatter plots using only pairs of variables. And the final visualisation would be easy for the audience to understand, and to dig for a possible trend.

However, there is a limitation that for scatter plots, correlations are greatly affected by outliers, especially that for this dataset, noise may confuse the audience and hinder their ability to find a straightforward conclusion. At first sight, people could be intimidated by the observation that the trend is not linear, and it is complex for our audience to draw an easy and straightforward conclusion. As such, binned charts are drawn and maybe useful for audience who have an inclination of ‘easy-to-understand’ charts for complex correlations.

6.3 EVALUATION OF CHARTS

Apparently, these charts are effective in illustrating a complex interaction between 4 factors, namely scores, indicators, score/percentile of influence, and location. Not only appealing to the eyes (as shown by the binned plot), we are able to draw complex insights from these charts that are very hard to find through quantitative analysis (such that, the correlation is not linear, but vary in difference ranges).
Moreover, we are able to find a trend at different ranges of X-variables easily. For instance, father with lower occupational status index is found to have better potential in having students of high

scores than mother with lower occupational status index, teacher is found to have higher potential than family despite being at lower percentiles. These observations are easy to observe and easy to understand by comparing whether there are more points of one color than points of the other color, in certain ranges of X-variables.

6.4 FURTHER QUESTIONS FOR FUTURE RESEARCH

Hopefully, this project provides insights for educational development and better utilisation of educational resources for students with different family and teacher characteristics.

Despite that the trend is not linear, but the influence of differences in family/teacher is apparent. To investigate further, more techniques could be used to remove possible outliers. More importantly, machine learning skills such as nearest neighbors or polynomial regression models could be used to see if there is a trend in certain ranges and predict student score using complex models in machine learning. Such predictions could be promising to understand and even predict student's performances when the accuracy is good enough.

7. CONCLUSION

Through data processing and visualisation, important insights are drawn for people interested in the educational field. The students score in different indicators is found to have non-linear correlations with factors such as father occupation, mother occupation, teacher characteristics. Father with lower occupational status index is found to have better potential in having students of high scores than mother with lower occupational status index, and this is especially true for mathematics and science subjects. Teacher at lower percentiles is found to have higher potential of having students with high scores than family at lower percentiles, and this true for all indicators. This also enlighten the educators that teacher with average literacy could still cultivate students with high capabilities.

Besides providing insights for people interested in high-school education, data processing skills such as using pandas package to process data, and matplotlib to visualise data are demonstrated to help future students in data analysis develop more tools. Hopefully the visualisation codes provided and encoding designs are well explained to help people interested in IT approaches enrich their tools kit and and sharpen their skills in data analysis.

8. REFERENCE

1. PISA. Program for International Student Assessment (PISA) - Overview. (2018). Retrieved from https://nces.ed.gov/surveys/pisa/

2. Important Notices - De La Salle College. (2018). Retrieved from https://www.delasalle.vic.edu.au/2015/important-notices-15/