# 1. Introduction

The Programme for International Student Assessment (PISA) provides a wide range of valuable information for comprehensive and reliable indicator of students’ capabilities. Data analytics of PISA enable Australia to consider the performance of our educational systems by comparing with others and fine-tune educational policies.

# 2. Objectives

## 2.1 Research questions

* Research question 1: whether mother influence more than father?

Research on maternal and paternal influences on the pisa scores in 3 subjects across 31 locations

* Research question 2: whether family influence more than teacher-ratio?

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

* Back-up research question: whether developed countries outperform developing countries in scores？

Back up research: Research on the discrepancy of pisa scores on 3 subjects by comparing developed countries with developing countries

## 2.2 Objectives and Audiences

* Provide insights for stakeholders in …, and possibly the general public who have interest in …
* Enhance learning in data manipulation for students in data analysis

# 3. Overview of Data

## 3.1 Domain Knowledge of data:

Data set PISA-2015 assessed ... The data set contains 31 observations of 9 variables…

表格

描述已自动生成

## 3.2 Origin of the data used:

PISA-data.csv was collected by National Centre for Educational Statistics (NCES) in the United States in Fall 2015("PISA”, 2018), which should be objective and useful to reveal the relationship between scores and various factors such as family influence and teacher-student ratio.

## 3.3 Summary Statistics of a database

# 4. Methodology of Analysis

4.1 Select data and obtain unique values for assessment indicators and assessment locations

4.2 Use filtering, grouping to stored scores in a dictionary with location as key, maternal/paternal influence, and scores as value

4.3 Construct list of scores and then a new data frame.

4.4 Prepare the data for 3 subplots, by plotting scatter plot of score of 3 subjects against maternal/paternal influence for different locations, encode location as different points, plot 2 best fit line on the same subplot for maternal/paternal influence, teacher ratio

4.5 Adjust bar plots by adding modifications and labels. Display p-value.

# 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

3 Scatterplots of score against maternal/paternal influence.

Encode locations as points on graph,

x denotes for score for 1 subject (be it reading, maths or science),

y-position denotes for maternal/paternal influence.

As such, there is a total of 3 subplots.

Plot best fit line for maternal/paternal influence.

## 5.2 Distribution of scores in Reading, Math and Science at different teacher ratios

3 Scatterplots of score against teacher-student-ratio, master-degree-ratio influence.

Encode locations as points on graph,

x denotes for score for 1 subject (be it reading, maths or science),

y-position denotes for maternal/paternal influence.

As such, there is a total of 3 subplots. Plot best fit line for teacher influence.

## 5.2 Distribution of scores in Reading, Math and Science at different locations

Findings: Discrepancy distributions of scores between low-performance countries and high-performance countries in various subjects

3 Barplots of scores against locations,

x denotes for score for 1 subject (be it reading, maths or science),

y-index denotes for location.

As such, there is a total of 3 subplots.

Label the benchmark line for average score.

图表, 条形图

描述已自动生成

6. Evaluation of Charts

6.1 Description of encoding

Describe encoding, style of chart, scale;

图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

6.2 Explanation of encoding design choice

Justify choice of encoding , evaluate the quality of the chart (its effectiveness of displaying how the relationship of some are impacted by the other attributes), indicate why you chose the summaries, and why you summarized in a particular way, link to audience’s aim and ease of understanding.

6.3 Evaluation of charts

6.4 Further questions for future research

7. Conclusion

8. Reference

# Code of Matplotlib

### Quick Start

*# import*

import pandas as pd

import matplotlib.pyplot as plt

from scipy import stats

import numpy as np

*# Prepare data*

df = pd.read\_csv("PISA2015-forStage2.csv", skiprows = 2)

df.head()

categories = list(df.columns[1:])

alice\_data = df.values[0, 1:]

bob\_data = df.values[1, 1:]

### subplots

plt.subplots(nrows=1, ncols=2, figsize=(20, 8), dpi=80)

**# Example**  
*# First Subplot*  
plt.subplot(1, 2, 1)  
plt.plot(x, y, color='green')  
  
*# Second Subplot*  
plt.subplot(1, 2, 2)  
plt.plot(x, y, color='steelblue')

# Possibel code to use

df = pd.read\_csv("PISA-data.csv")

df.head()

*# extract all columns*

df = df.iloc[:, [0, 1, 2]]

*# get list of variable categories of indicator and location*

locations = list(df['"LOCATION"']. unique())

**#…**

**for** location **in** locations:

d[location] = df[df['"LOCATION"'] == location].groupby("SUBJECT")["Value"].mean().to\_dict()

**#…**

*# create a new data frame with 4 columns: indicator, location, boy/girl averages*

location\_value = pd.DataFrame({"LOCATION": locs, {"INDICATOR": indis, "SCORE": score, "MOTHER": mother\_index, "FATHER": father\_index, "RATIO": ratio}})

*# Prepare the data for 3 subplots of score against influence/location*

…

*# Create figure*

plt.figure(figsize=(10, 6), dpi=80)

ax = fig.add\_subplot(3, 1, i ) *# draw 3 subplots in 1 columns*

*# draw subplot and set parameters*

ax.bar(x0, y1, width = 0.5, label = "Boy")

ax.set\_xlabel("Location", fontsize = 15)

ax.set\_ylabel("Value", fontsize = 15)

ax.set\_title(“PISA”, fontsize = 15)

ax.legend()

*# bar chart with xticks modified*

bar\_width = 0.35

xticks\_index = range(0, len(categories))

alice\_index = [x - bar\_width/2 for x in xticks\_index]

bob\_index = [x + bar\_width/2 for x in xticks\_index]

plt.xticks(xticks\_index, categories)

plt.bar(mother\_index, mother\_data, bar\_width, label='Mother')

plt.bar(father\_index, father\_data, bar\_width, label='Father')

*# save and display figure*

plt.savefig(‘fig.pdf’, dpi = 80)

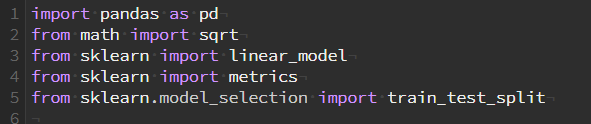
plt.show()

…

# Stage 3

We’ve made two predictive models with python(scikit-learn) and Excel respectively. We also tried many different models. However, not every one of them was suitable for our data.

## Scikit-learn (Linear Regression model)



We supply X and y in order to build the model, set the test\_size to be 0.1 (randomly select 10% of the data from the data set for the testing) In this case, we want to use the happiness ranks and score to predict the GDP score for a country. We will obtain the model ‘regr’ and put the sample data inside the model.

Code

The results of this prediction are displayed.

Screenshot

This predict model shows that…

### Evaluation:

It is efficient to use../Its effectiveness is undermined by…

## -Excel (Linear Regression model):

“Analysis ToolPack” is used by selecting the data needed, and it will produce a predictive model automatically. There is a small problem when we used this way to create a predictive model. If there is a missing value in that column, the function will not work, which means we need to clean the data again and delete those rows with any missing value.

