PISA Scores Data Analysis - INDUSTRIAL APPLICATIONS OF STATISTICS

Emre Albayrak[a]

[a]*Necmettin Erbakan University, , Konya, Türkiye*

Abstract

This paper addresses the OECD PISA dataset spanning from 2000 to 2018, assessing the performance of 15-year-old students in reading, mathematics, and science across various countries. The dataset provides mean scores categorized by country, indicator (reading, mathematical, or scientific), subject (boys, girls, total), and year of measurement.

Keywords: Data Analysis, R Programming, PISA Scores

1. Introduction

In our exploration of international student performance, we leverage R programming and sophisticated data analysis techniques on the OECD PISA dataset (2000-2018). Our goal is to interpret and report nuances within 15-year-old students' scores in reading, mathematics, and science.

Main Objectives:

1. **Exploratory Data Analysis (EDA):** Conduct thorough EDA, unraveling patterns and trends in performance scores across countries and years.
2. **Comparative Analysis:** Analyze variations based on gender and subjects.
3. **Temporal Trends and Correlations:** Explore temporal trends and correlations between socio-economic factors, policies, and performance.
4. **Predictive Modeling:** Develop predictive models to forecast future trends.

2. Understanding the Dataset

2.1. Columns

index	locations	indicators	subjects	time	value
0	AUS	PISAMATH	BOY	2003	527.00
1	AUS	PISAREAD	GIRL	2008	519.00
2	CRI	PISASCIENCE	BOY	2010	485.00
3	CRI	PISASCIENCE	BOY	2005	497.00
4	LTU	PISAMATH	GIRL	2009	505.00
5	LTU	PISASCIENCE	GIRL	2018	503.00

Figure 1: Means of columns

Before analysis, understand each column:

1. **Locations:** Country alpha-3 codes. OAVG indicates an average across all OECD countries.
2. **Indicators:** Reading (PISAREAD), Math (PISAMATH), or Science (PISASCIENCE).
3. **Subjects:** BOY (boys), GIRL (girls), or TOT (total).
4. **Time:** Year of measurement (2000-2018).
5. **Value:** Mean score.

3. Getting Started to Analysis

3.1. Part 1: Read & Understand To Data

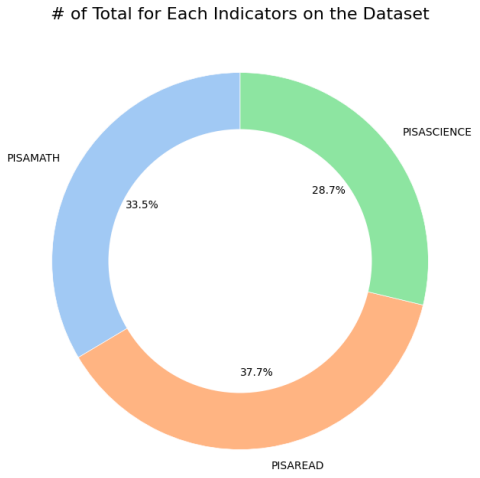

Figure 2: Distribution of each indicator in the total dataset (Pie Chart)

We are analyzing the percentage of each indicator in the total data set with a pie chart. A pie chart is a circular

graph that shows the relative sizes of different categories by dividing a circle into sectors. The angle of each sector is proportional to the percentage of the category in the data set. Pie charts are useful for comparing the parts of a whole and highlighting the differences among them.

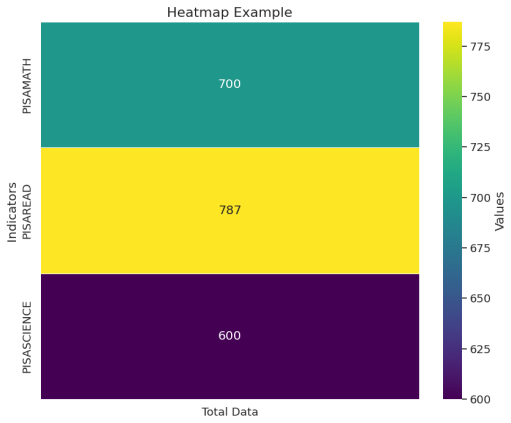

Figure 3: Distribution of each indicator in the total dataset (Heat Map)

Visualizing the same graph with a heat map.

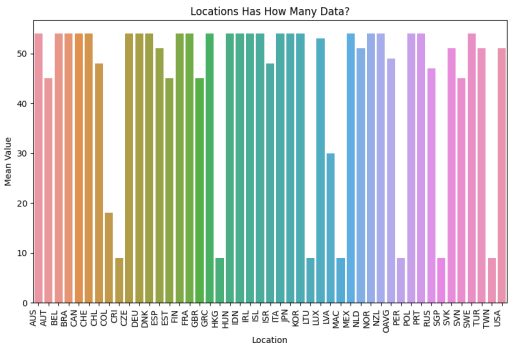

Figure 4: Number of Data Entries by Location

These graphs provide us with many clues about the data set, such as the distribution of the data sets, their impact on the average, and the analysis topic. This way, they help us to have a rough idea before starting the analyses. A paragraph is a group of sentences that are related to a common theme or topic. A long paragraph usually has more than five sentences and provides enough details and examples to support the main idea. An informational tone is a way of writing that conveys facts, data, or instructions in a clear and objective manner.

3.2. *Part 2: Overview of Average Values*

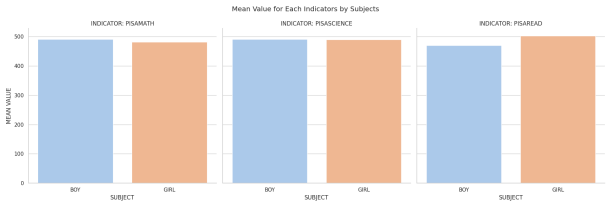

Figure 5: Mean Values for Each Indicator by Subject

We visualized the distribution of each subject in the total dataset for every indicator using a bar chart, illustrating the respective quantities.

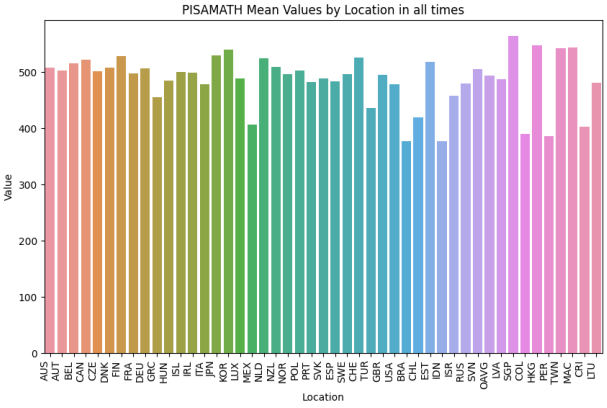

Figure 6: Mean Values for PISA Math by Location

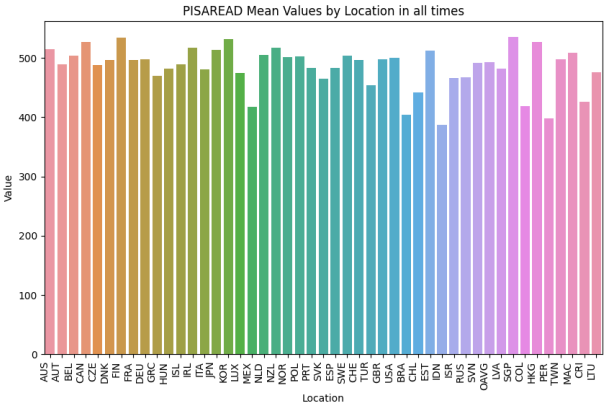

Figure 7: Mean Values for PISA Read by Location

These graphs show the average values for each country in the dataset for each indicator.

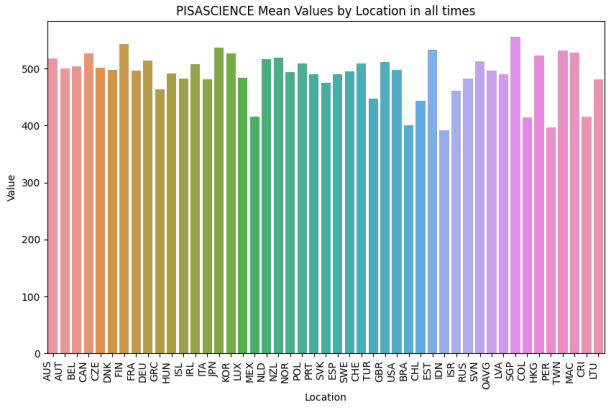

Figure 8: Mean Values for PISA Science by Location

*3.3. **Part 3: Value Change Over Time***

In this section, we will examine how different countries have changed their performance on various indicators over time. These indicators reflect the status of countries in areas such as mathematics, science and reading. This analysis reveals differences and similarities between countries and is useful for observing global trends. Our aim in this section is to graphically show the changes in indicator values over time and to interpret these changes.

Figure 9: PISA Math - Part 1

Figure 10: PISA Math - Part 2

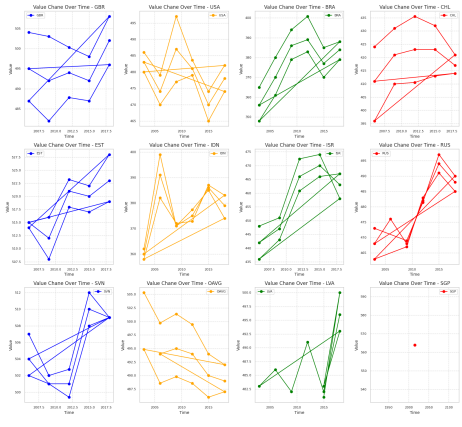

Figure 11: PISA Math - Part 3

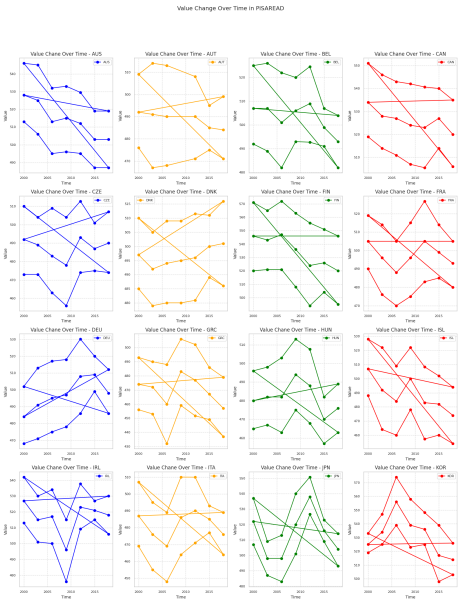

Figure 12: PISA Read - Part 1

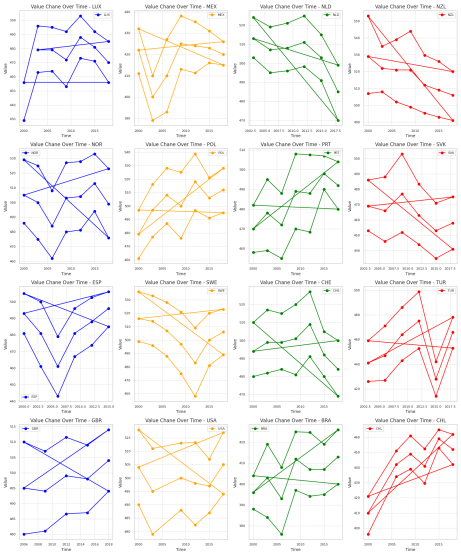

Figure 13: PISA Read - Part 2

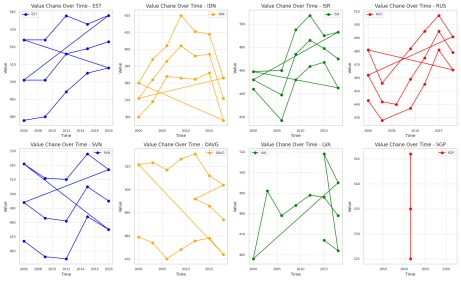

Figure 14: PISA Read - Part 3

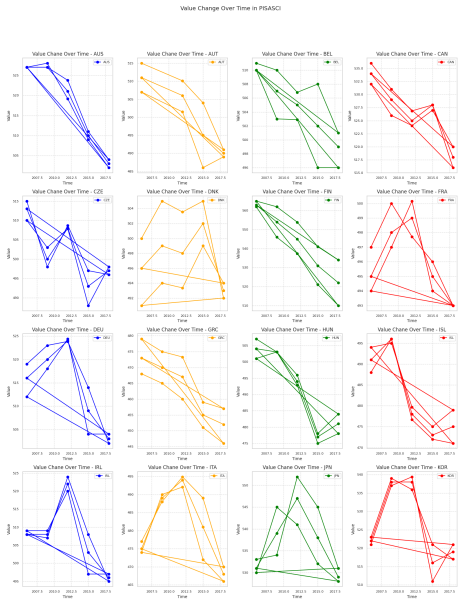

Figure 15: PISA Science - Part 1

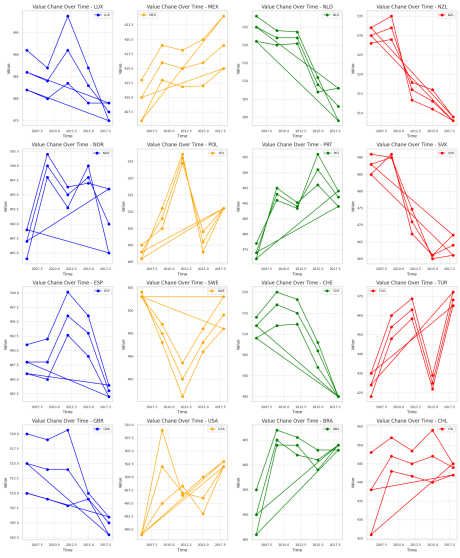

Figure 16: PISA Science - Part 2

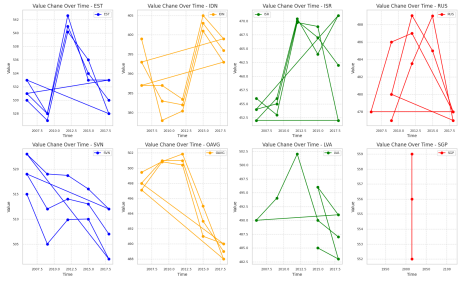

Figure 17: PISA Science - Part 3

*3.4. **Part 4: Linear Regression***

In this last section, we created a Linear Regression model using historical data for the ITS location. This model helps us predict how the ITS location will perform on each indicator between 2019-2029. You can see the prediction results in the chart. The notation of actual data and predicted data is indicated in the chart.

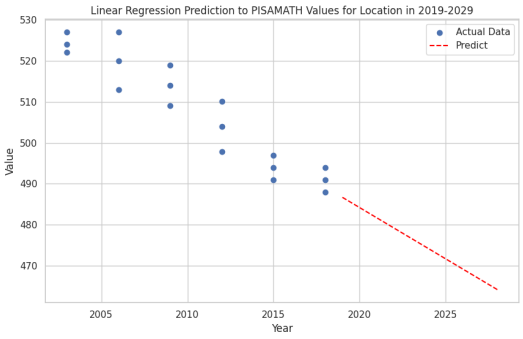

Figure 18: Regression Values for PISA Math

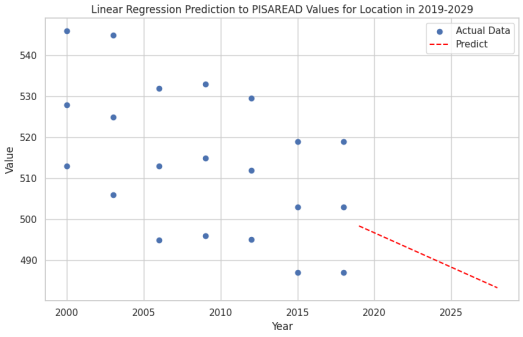

Figure 19: Regression Values for PISA Read

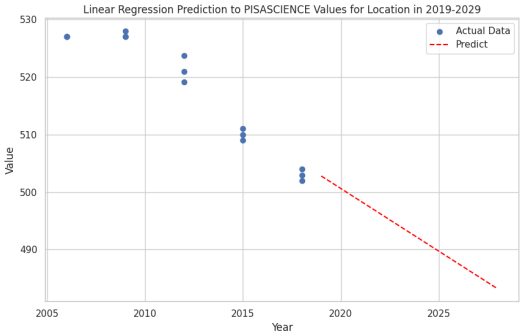

Figure 20: Regression Values for PISA Science

4. References

https://www.kaggle.com/code/tarikemre/pisa-scores-analysis

https://www.kaggle.com/datasets/thedevastator/pisa-performance-scores-by-country