NILE UNIVERSITY

Descriptive analysis on Egypt High School Grades (Thanaweya Amma) using Multi Linear Regression models and visuals.

A Math203 Project Report

By

Mohamed ElGemeie (202000206)
Mohamed Ismail (202003000)
Hesham Reda (19105550)
Ibrahim Moustafa (202000938)
Ahmed Kamal Ahmed (211000202)

Submitted in partial fulfillment of the requirements

for Math-203 Project

Date

4/19/2022

ABSTRACT

In this paper we will be Analyzing a Dataset about El- Thanaweya El-Amma students which contains more than 100k records of their high school grades, and using some analysis techniques to answer a couple of questions. Some of these like Multi Linear Regression Model, and scatterplots. We started by cleaning and organizing the data to facilitate the process of analysis, Furthermore, use the techniques to end up with insight to answer our proposed questions. Three questions were asked: what is the hardest subject in Adaby? Who does better in chemistry and Physics, Alm Aloom or Alm Ryada? Who has the most passing/failing students? For the first question We found that the hardest subject was Arabic by using a regression model to draw a theory and scatter plots to prove it visually. As for the second Question after comparing these different groups, we concluded that Alm Ryada does better in Physics, and Chemistry, However, the difference between the two groups in chemistry was very slim. For the third, Alm Ryada is the best group regarding passing, while Adaby has the most failure rates.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

SECTION I: Introduction

This paper reviews a descriptive analysis on 2021 Thanaweya Amma students' grades containing 100K records to answer a set of questions, no previous papers or articles were made that relate to this field regarding Thanaweya Amma Students. Many were made on some bad effect of Thanaweya Amma's English Test. The original aim was to investigate the data and find any insights and so the questions were based on how far we got with our analysis techniques. To fully grasp the contents of this paper, you must know what Thanaweya Amma is, how does it work and what Regression Models are.

Thanaweya Amma is the last year of High School for Egyptian students, their scores this year determine which college and university they will be able to enter. Thanaweya Amma students are usually split into two groups, Adaby and Almy. Adaby students take 7 subjects: Arabic, English, mandatory second language like French or German, History, Geography, Psychology and Philosophy. Almy students are split into two more group Almy Aloom and Almy Ryada. Both of which have Arabic, English, second Language, Chemistry and Physics. Only Almy Aloom students take Geology and Biology, while Almy Ryada students take both Pure and Applied Mathematics. Both groups – Almy and Adaby – have other extra subjects such as Statistics and Religion.

The Regression model is a function that represents the connection/relation between a response, dependent, or target variable and one or more independent variables. And Hopefully by the end of reading this paper you will have a better understanding to what a regression model is and how it is applied to see certain patterns in huge amounts of data.

Descriptive graphs and visuals were used to visualizes the data and explain any correlations we found. These graphs compare between one dependent variable and one independent variable like scatterplots and boxplots.

SECTION II: Background and Literature Review

Regression model

In data analytics, the regression model is used to predict certain outcomes from a set of data. It is one of the most widely used predictive analytics methods, and serves as the foundation for many others machine learning models. it was one of the first to be utilized in business, and has been applied in a wide range of applications from predicting a client’s purchase behavior to calculating the risk of a severe weather event. The main goal of these models is to find whether X has a positive or negative correlation with variable Y.

History

It has a long and storied history. The first regression model was developed by Sir Francis Galton in the late 19th century. He plotted different size measurement of daughter and mother peas to find a correlation. In his study, he observed that sons do not propagate toward their fathers' heights but instead approach the mean of the population.

Multi linear regression model.

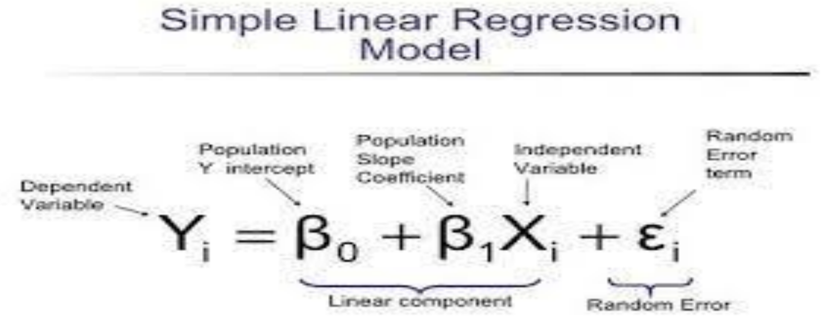
Multilinear regression models are mainly used to compare between the effect of many independent variables and one dependent variable to see which of the independent variables influences the most in terms of the other independent variables. Example being: what affects your mental state more, your job or your marital life? What are important factors for plant growth humidity, temperature or region? As more independent variables are added, more $\mathbf{X_i}$ variables are added with their coefficients in the below equation.

Equations

The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable we want to predict), X is the independent variable which is plotted on the X axis and is the feature we use for the prediction, b is the slope of the line and a is the y-intercept which represent the value of y when all features X are zero.

$$y = a + b(x)$$

Other variations have a margin of error variable:



Figures

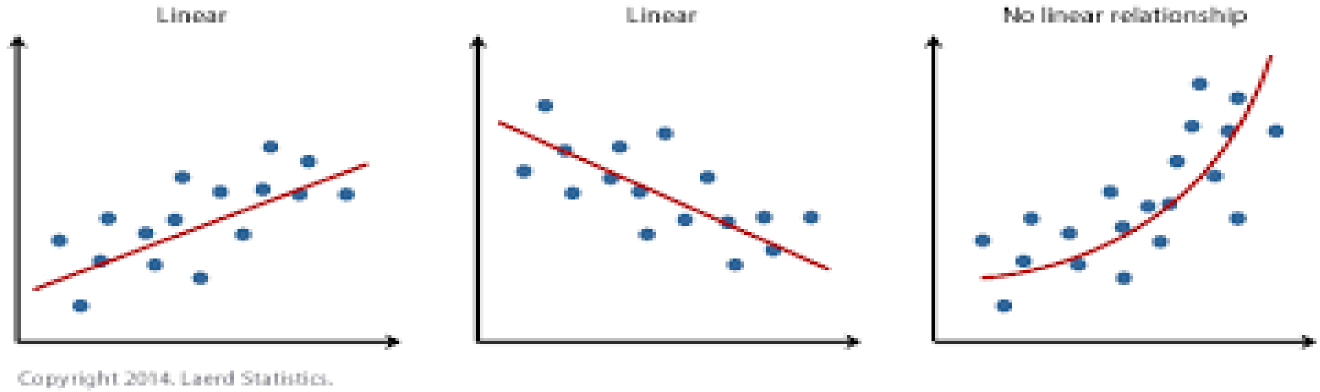


Figure 1: different variations of relation between the variable X and y, being positive, negative, and a non-linear relationship.

SECTION III: METHODOLOGY

Our dataset was retrieved from
https://www.kaggle.com/datasets/raamyy/egypt-high-school-grades-sanwya-amma

We programmed the analysis on Jupiter notebook using python and python data manipulation packages. In this project, we have gone through three stages to do analysis.
Stages:

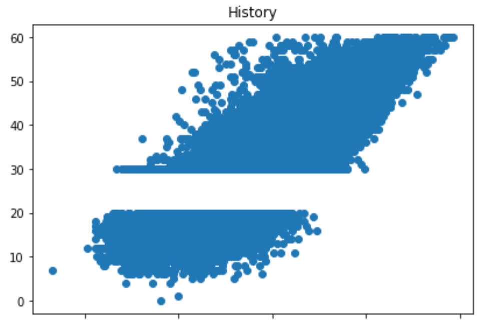
- setup
    - we import python libraries like pandas for data manipulation, statsmodels.api for the regression model, and matplotlib to visualize the data.
    - we import our dataset using pandas' data reading functions
- cleaning the data
    - Dropping unnecessary columns in the dataset.
    - Removing non-numerical (string) values in data set that cannot be used in analysis.
    - Creating dummy variables to store categorical data using a binary format, as regression models can only handle numerical data.
    - Separate different groups of the students with their subjects like Adaby, Alm Aloom and Alm Ryada.

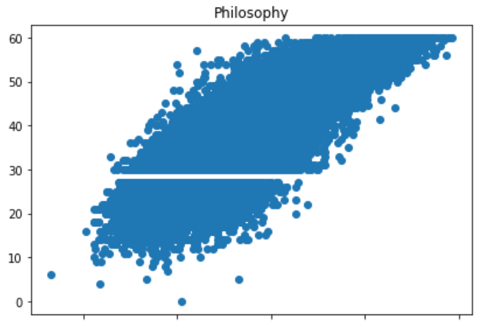
- Analysis
    - Use Matplotlib to visualize and find insights.
    - Use regression models to prove that insight statically.

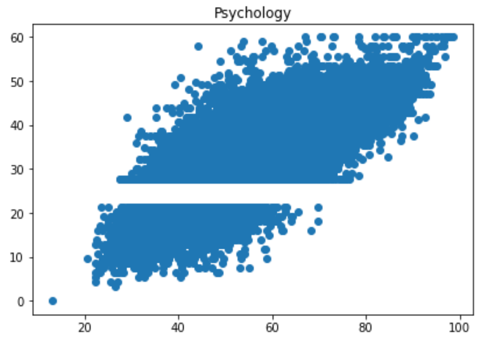
SECTION IV: RESULTS

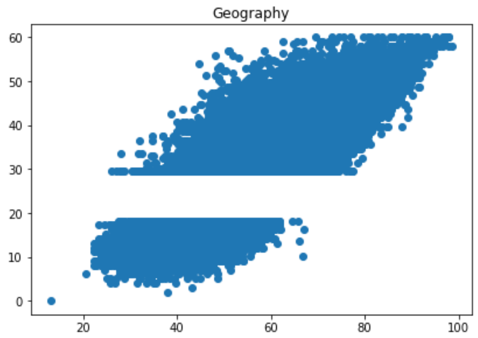
Q1) What is the hardest subject in the Adaby group?

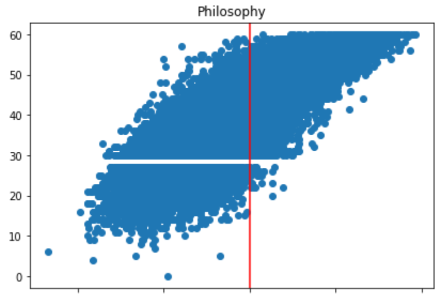
We first compared all the subjects to the total grade of the students, and the goal was to find which subject has high grades regardless of the total grade increase which would conclude that, this subject is easy, and a hard subject would be one that is directly proportional with the total grade meaning that as the total grade increases: the subject's score increases.
Using scatter plots on the following:

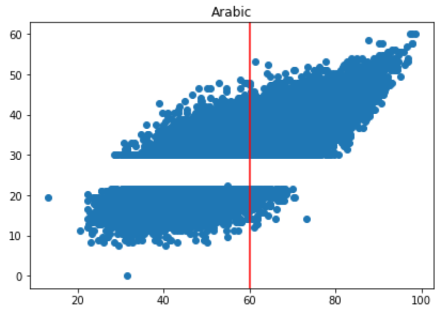






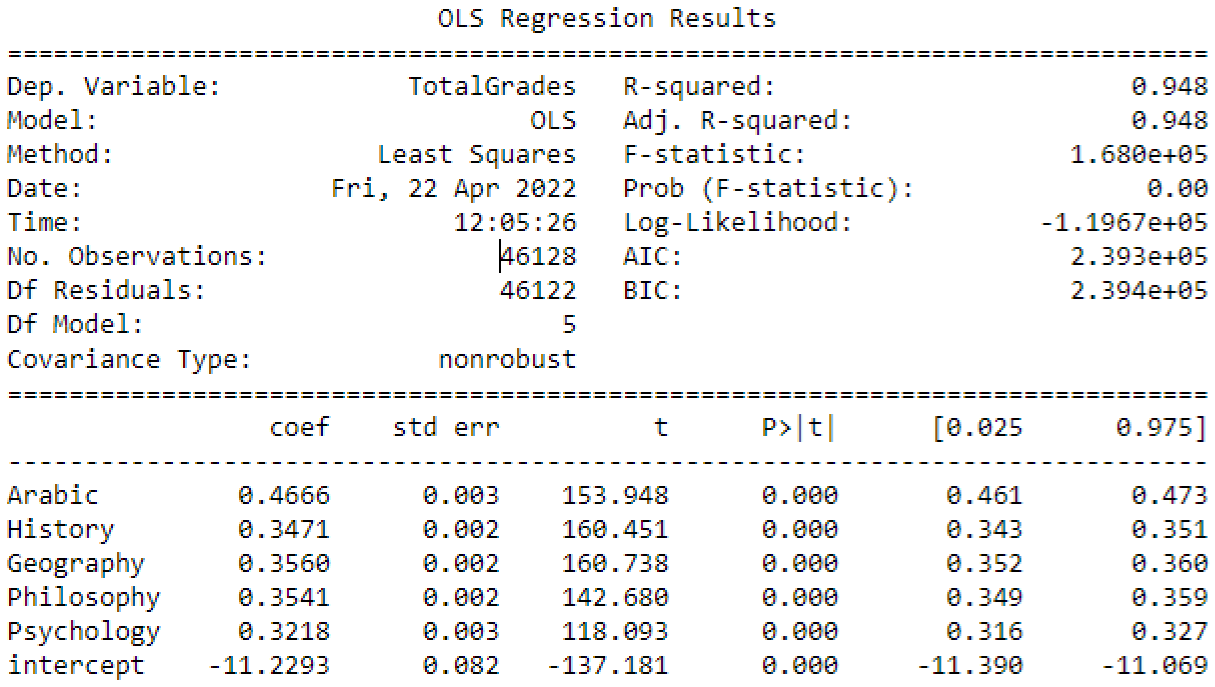






As shown, the Empty Part of the Plots is String values that were deleted from the dataset, and rows that contained one Null values, so please ignore them. a lot of students who score more than 60 in the total grades seem to achieve very high scores that subjects above which implies that high good achieving students find these subjects easy. But in Arabic there seems to be direct positive relation between Arabic and the total grade and there isn't a lot of students the score high Arabic scores after passing 60 in the total grades for the Arabic table which implies that Arabic is harder than other subjects.

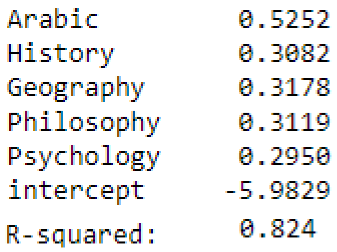
We also used a multilinear regression model to prove this finding statistically, and the intercept is used to center calculate the coefficients:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            TotalGrades   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.948
Method:                 Least Squares   F-statistic:                 1.680e+05
Date:                Fri, 22 Apr 2022   Prob (F-statistic):               0.00
Time:                        12:05:26   Log-Likelihood:            -1.1967e+05
No. Observations:               46128   AIC:                         2.393e+05
Df Residuals:                   46122   BIC:                         2.394e+05
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Arabic         0.4666      0.003    153.948      0.000       0.461       0.473
History        0.3471      0.002    160.451      0.000       0.343       0.351
Geography      0.3560      0.002    160.738      0.000       0.352       0.360
Philosophy     0.3541      0.002    142.680      0.000       0.349       0.359
Psychology     0.3218      0.003    118.093      0.000       0.316       0.327
intercept    -11.2293      0.082   -137.181      0.000     -11.390     -11.069
```

These are the coefficients of each subject that affect the total grades, and Arabic has 46% on the total grade more than other subjects, while philosophy has a positive correlation of 45%. This statistic has an R squared value of 0.948 which measures how well does this regression model predict the Total grade, being very high (close to 1) indicates that this statistic can be evaluated with 95% certainty.

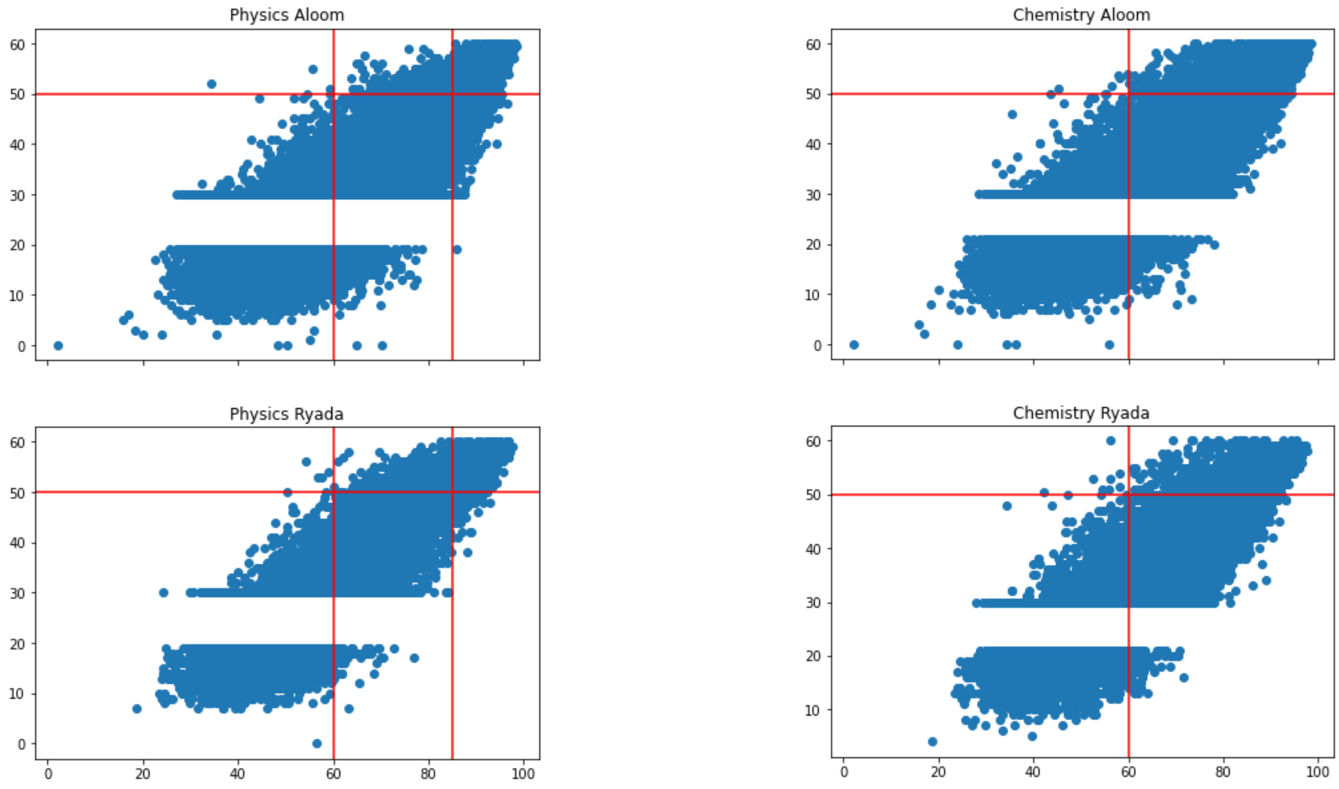
After decreasing the data to only the students that score more than 60 on the total grades and using a regression model, we also found that the percentage increased, and the R squared value didn't decrease much. This proves the insight that Arabic is the hardest subject.



```
Arabic         0.5252
History        0.3082
Geography      0.3178
Philosophy     0.3119
Psychology     0.2950
intercept     -5.9829
R-squared:     0.824
```

Q2) Who Does better in Chemistry and Physics: Alm Aloom or Alm Ryada?

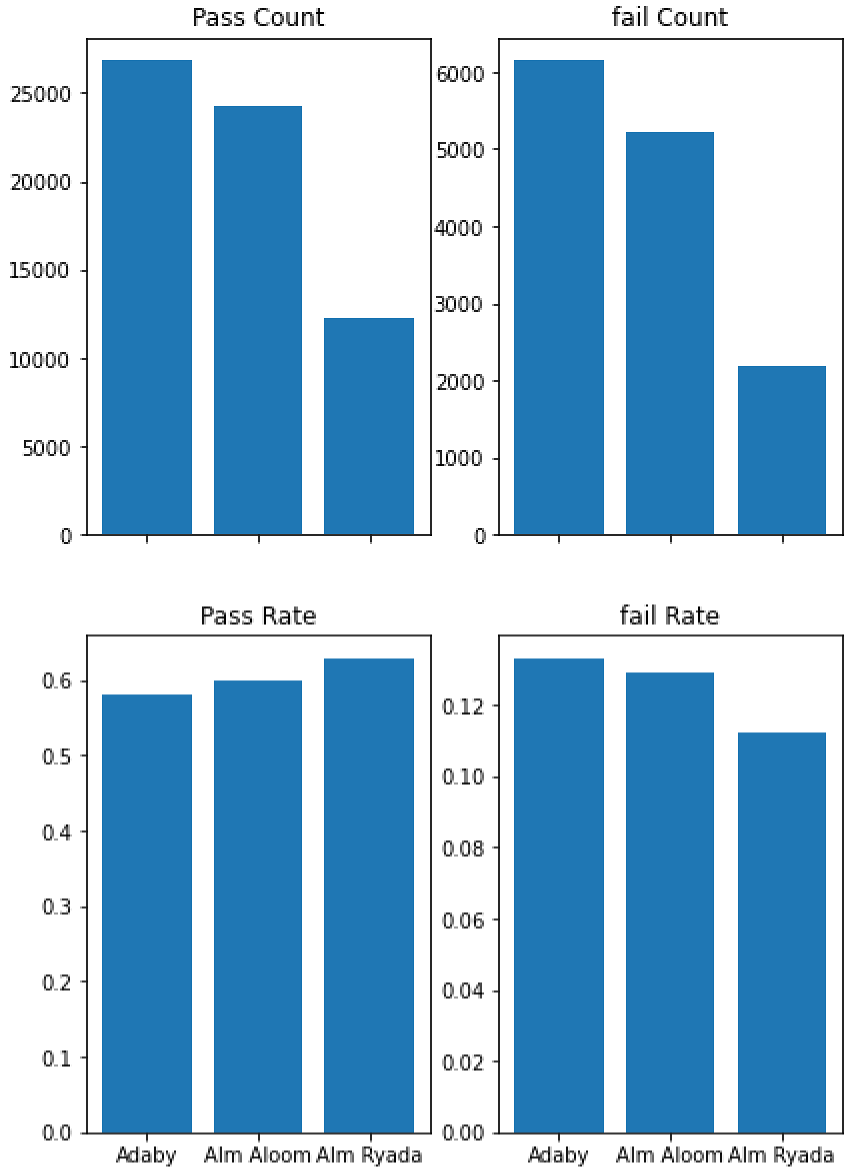
Using the same techniques as above:



Alm Ryada seems to do better in chemistry than Alm Aloom, as the red lines show that students who score higher than 60% and have a chemistry score higher than 50 are mostly Alm Ryada but the difference is very slim.
The area blue area marked by the red lines show the portion of students who get less than 85% and more than 50 in Physics which is covered by more blue dots in the Alm Ryada group, showing that Alm Ryada did better in Physics than Alm Aloom

Q3) Who has the most passing/failing students

Groups	Pass Count	Fail Count	Pass Rate	Fail Rate
Adaby	26810	6142	0.5812	0.1332
Alm Aloom	24242	5210	0.5998	0.1289
Alm Ryada	12205	2177	0.6287	0.1121



SECTION V: CONCLUSION

This was an explanatory analysis meaning that we formulated our questions based on our findings. To answer our previously stated questions.

1st statistic

The hardest subject in Adaby is Arabic with a regression model R-squared value of 94% that indicates the credibility of this statistic.

2nd statistic

There doesn’t seem to be much variance in both groups but there is a small portion which is better achieved by Alm Ryada. We couldn’t use a regression model in this question as it was difficult to group the data by subject.

3rd statistic

Adaby has the highest rate of failure, while more than half of the students in Alm Ryada seem to pass the most with a rate of 0.63 making it the best group to apply to statistically speaking. Alm Aloom seems to be unbiased in terms of passing and failing.

REFERENCES

[1] Ibrahim, M. (2019). The washback effect of the thanaweya amma English test: Drawbacks and solutions. In *English language teaching research in the Middle East and North Africa* (pp. 73-91). Palgrave Macmillan, Cham.

[2] Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, *9*(3).