College of Computing and Informatics

2020/2021 First Semester

Course Code	DS540
Course Name	Advanced Python for Data Science
CRN	14044
Assignment type	Project
Module	All Modules
Assignment Marks	10

Student ID	G200007615
Student Name	Abdulaziz M. Alqumayzi

Exam Performance Analysis

Abdulaziz Mohammed Alqumayzi
G200007615@seu.edu.sa,
Saudi Electronic University,
Riyadh, Saudi Arabia

Enad Saud Alotaibi
G200002614@seu.edu.sa,
Saudi Electronic University,
Riyadh, Saudi Arabia

1. Introduction

The student performance in exam is an important goal of the educational institution and it can improved by understanding the factors responsible for affecting it. Thus, we use a dataset of United States high school student's performance in exam to get insights into the performance of the students in subjects of Math, Reading and Writing. Furthermore, using (5) different attributes we managed to predict the scores of the students in those subjects using Random Forest as a regressor with Root Mean Square Error of approx. 15 for all three subjects.

2. Body section

2.1 Data

The provided data consists of scores of high school students from the United States in three different subjects (Math, Reading and Writing) and different other attributes. We will be analyzing whether the subject scores can attributed to these factors or not. The dataset consists of sample of 1001 students and has following (8) attributes:

Attribute	Type	Values
Gender	Object	female, male
Race/ethnicity	Object	group A, group B, group C, group D, group E
Parental level of education	Object	bachelor's degree, some college, "master's degree, associate's degree, high school, some high school
Lunch	Object	standard, free/reduced
Test preparation course	Object	none, completed
Math score	int64	
Reading score	int64	
Writing score	int64	

2.2 Methods

Initially, exploratory data analysis performed on dataset by first performing a descriptive analysis and then perform the visualization of the data. Subsequently, Random Forest Regression was used as the machine learning model. In order to provide the data to model, the object types were One Hot Encoded, and then the data was divided in training and testing data set with a ratio of 80:20. After that, in order to optimally select the parameters Grid Search was performed. Moreover, in order to evaluate the model, Root Mean Squared Error (RMSE) was selected as the primary evaluator, although Mean Absolute Error was also calculated.

2.3 Analysis

Data types were firstly investigated after that duplicates data. Data types are correct for each feature, but there were duplicates and were removed. Missing values were dropped. In figure 1 shows the scores are near each other, math is a little bit lower than the others.

Reading score clearly better than the other score and math has most outliers. In table 1, confirms the boxplot figure 1 which is the score of the reading subject is the best among the others. In figure 6 histogram shows the distribution of the scores. In figure 2, the average score of all subjects per gender. The two charts show that males are better than females in math subjects. In contrast, females are better than males in reading and writing subjects. In figure 3, the scores of subjects per parent level of education. The chart shows scores tend to be higher for students of a master's degree parents, then those parents having bachelor's degree, the least are those parents having a high school. In figure 4, scores per race chart show that students from groups A, D and E tend to have scored more than groups B and C. In figure 6, the charts show students who prepare for the test clearly get a better score than those who did not prepare for the exams. In figure 7, correlation plot shows that there is no strong correlation between features. Thus, our model will not be perfect for the machine learning model. In table 2, the RMSE of the scores in the random forest regression model is 15.9 for math, 16.1 for reading, and 16.3 for writing. These results are high which is not a good model. For example, if we predict that a student will get 75 scores in math subject. The RMSE tells us the score will be between 59.1 and 90.5.

2.4 Results

Task1:

```
#loading the dataset
import pandas as pd
df_data = pd.read_csv("StudentsPerformance.csv")
df_data.head()
#Get summary of dataset
df_data.info()
```

```
df_data.shape
```

The dataset consists of sample of 1001 students and has (8) attributes as shown below

```
RangeIndex: 1001 entries, 0 to 1000
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1001 non-null   object
 3   lunch                        1001 non-null   object
 4   test preparation course      1001 non-null   object
 5   math score                   1001 non-null   int64 
 6   reading score                1001 non-null   int64 
 7   writing score                1001 non-null   int64 

dtypes: int64(3), object(5)
```

The data consisted of two samples containing null values and one pair of duplicate sample, which removed while cleaning the data.

```
#Check missing values
df_data.isnull().sum()

#dropping the missing values
df_data = df_data.dropna()
df_data.isnull().sum()
df_data.shape

#dropping the duplicates
df_data.duplicated().sum()
df_data = df_data.drop_duplicates()
df_data.duplicated().sum()
df_data.shape
```

Task 2 & Task 3:

Performing exploratory data analysis (EDA) and then perform respective visualizations. First, we comparing the results of students in different subjects, we found that student scored highest marks in Reading as evident in Figure 1 and Table 1. Moreover, it observed that male students scored more marks in math while females outperformed the males in Reading and Writing. The trend is visible in Figure 2.

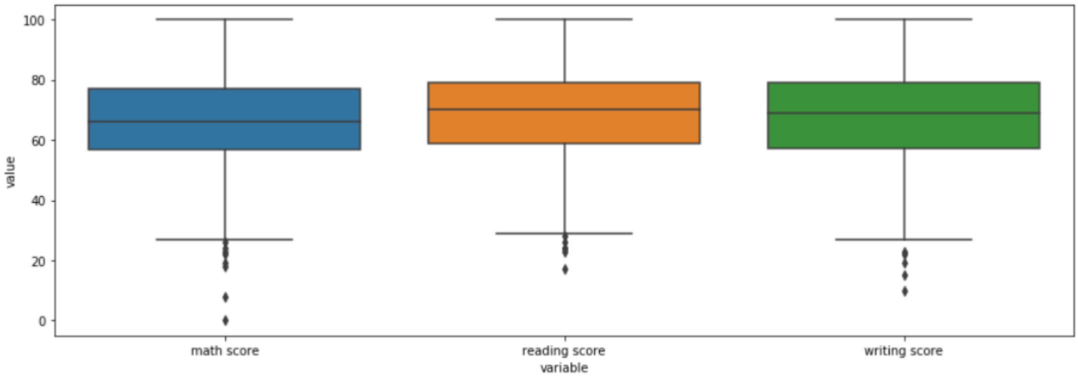
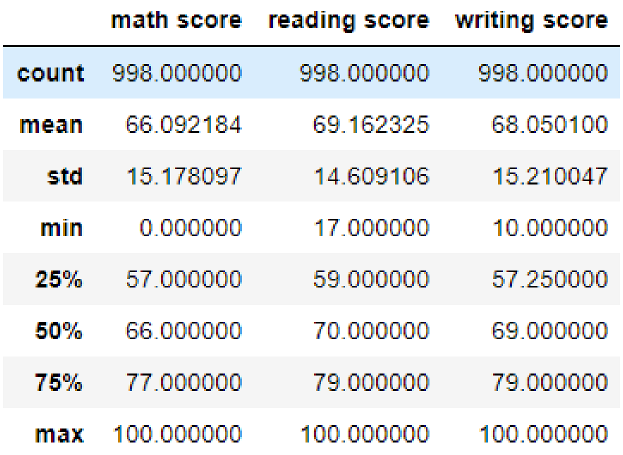
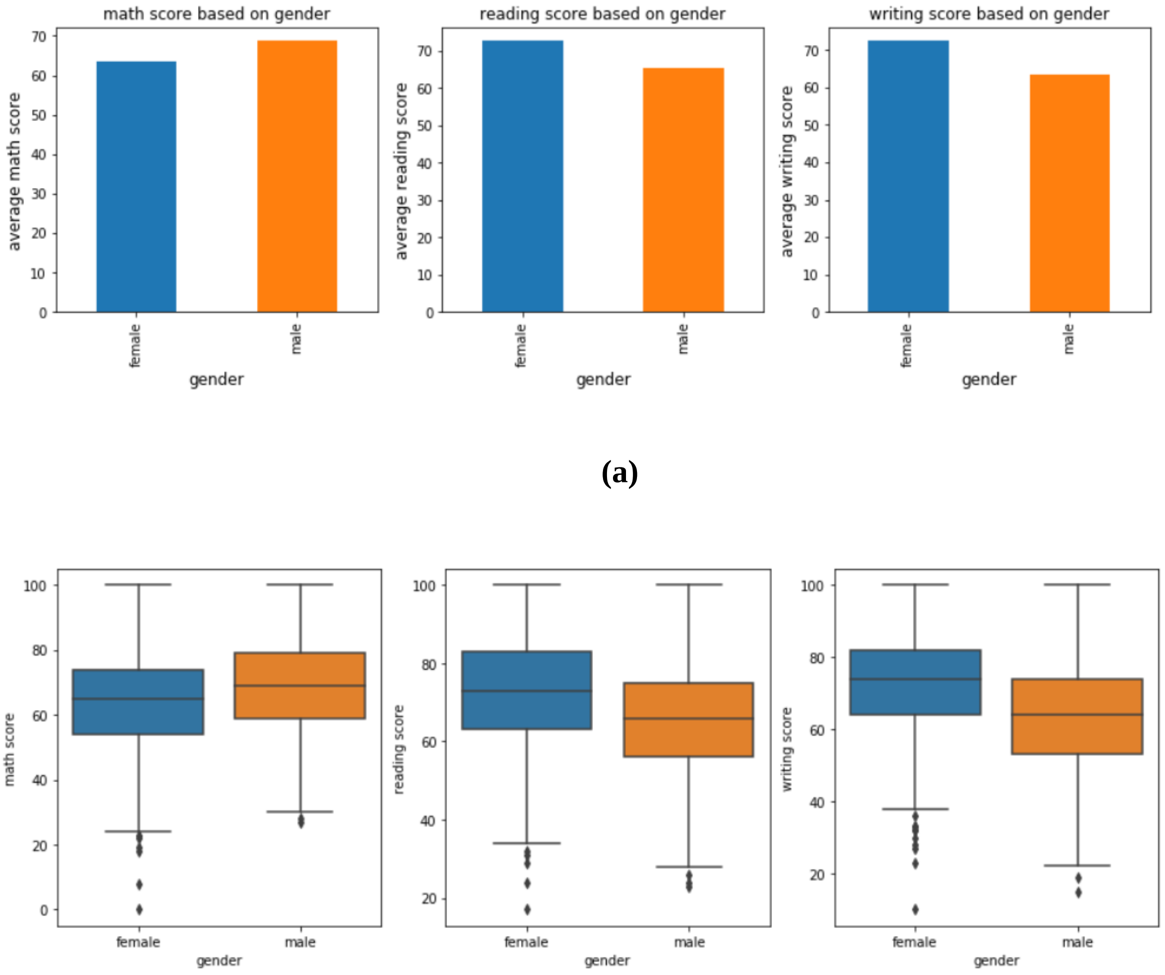


Figure 1. Box plot for subject marks.

Table 1. Descriptive stats of Subject Scores



	math score	reading score	writing score
count	998.000000	998.000000	998.000000
mean	66.092184	69.162325	68.050100
std	15.178097	14.609106	15.210047
min	0.000000	17.000000	10.000000
25%	57.000000	59.000000	57.250000
50%	66.000000	70.000000	69.000000
75%	77.000000	79.000000	79.000000
max	100.000000	100.000000	100.000000



(a)

(b)

Figure 2. (a) Subject Average Score Vs Gender, (b) Gender Wise Box Plot of Subject Scores

Moreover, the students whose parents were highly educated were less likely to score low-grade marks in the all the subjects. The trend is evident in the Figure 3.

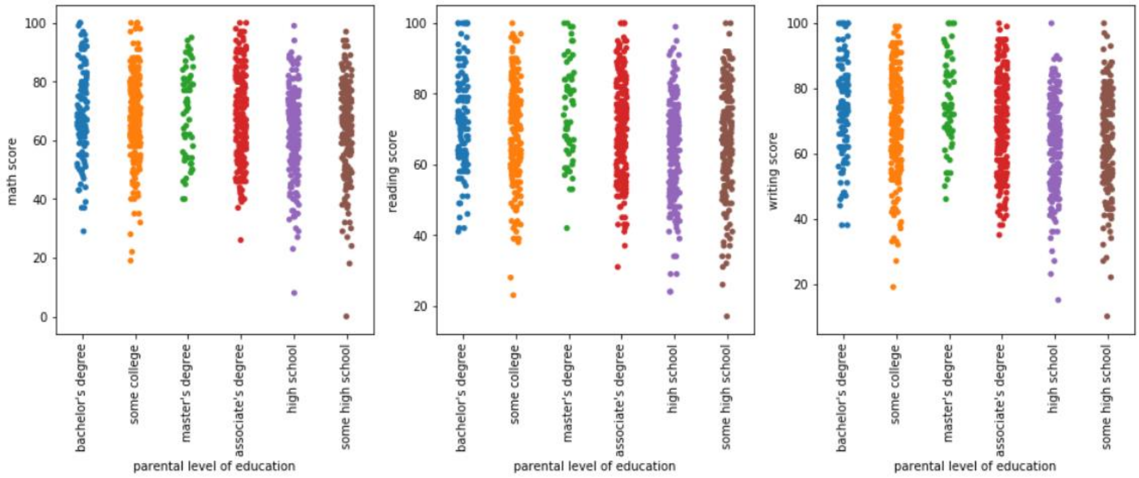


Figure 3. Subject scores plotted against Parent level of education

The scores were also dependent on the race of the students. For example, Group E students were mostly scoring high marks compared too Group B, although the trend was less evident in Reading scores. This can be visualized in Figure 4.

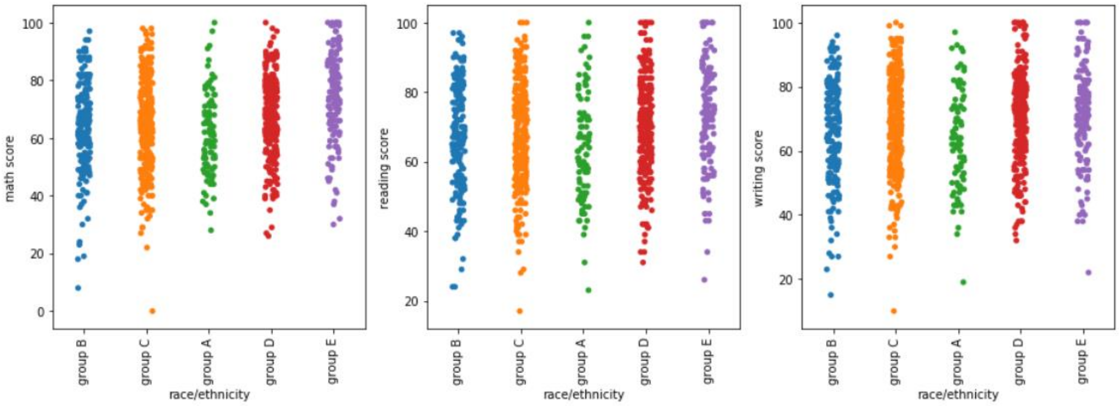
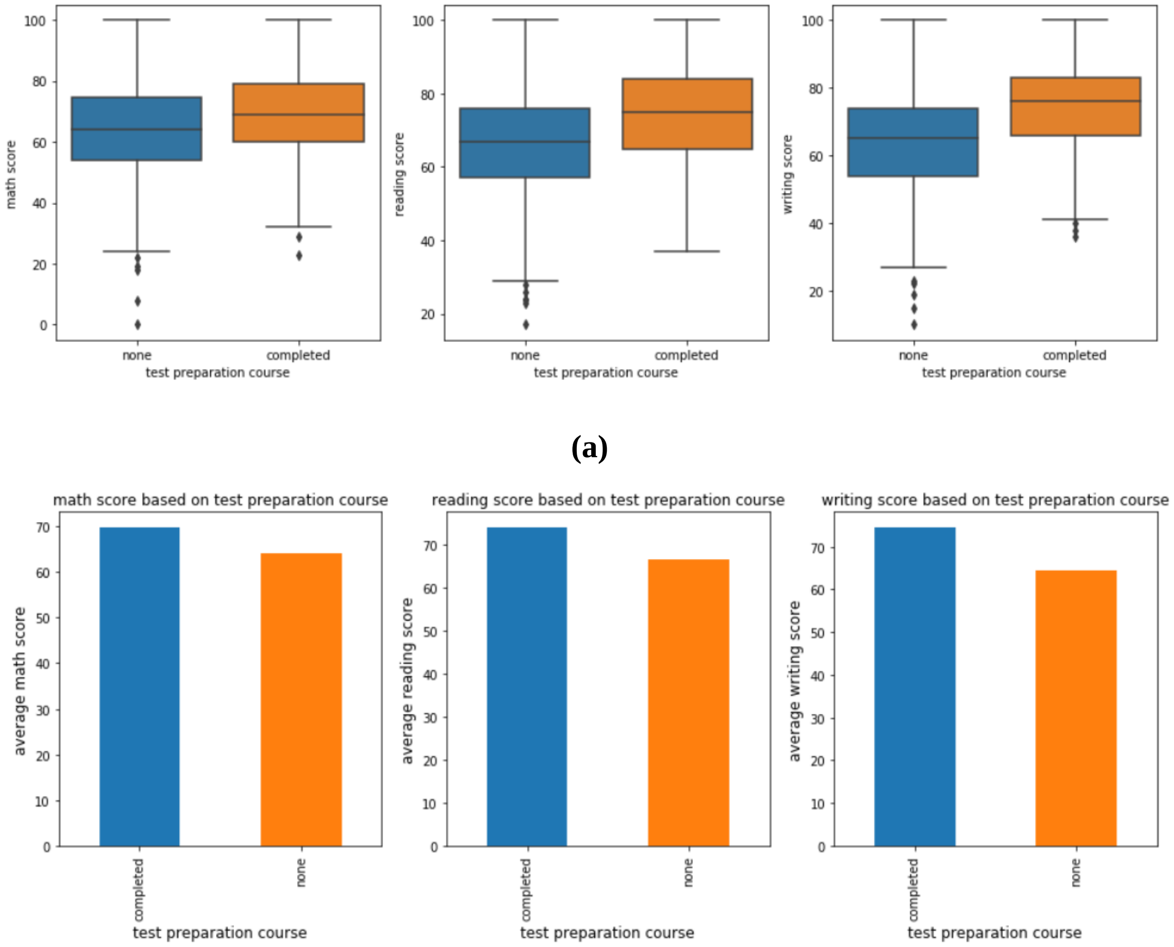


Figure 4. Subject scores plotted against Race.

In addition, the Test Preparation Course also proved to be effective in increasing the subject scores of the students as evident in Figure 5.

After that an analysis of the distributions of the subject score was performed and it was evident that the distribution was skewed more towards right for writing while for Math it was skewed towards left Figure 6.



(a)

(b)

Figure 5. (a) Subject Average Score Vs Test Preparation Course, (b) Test Preparation Course Box Plot of Subject Scores

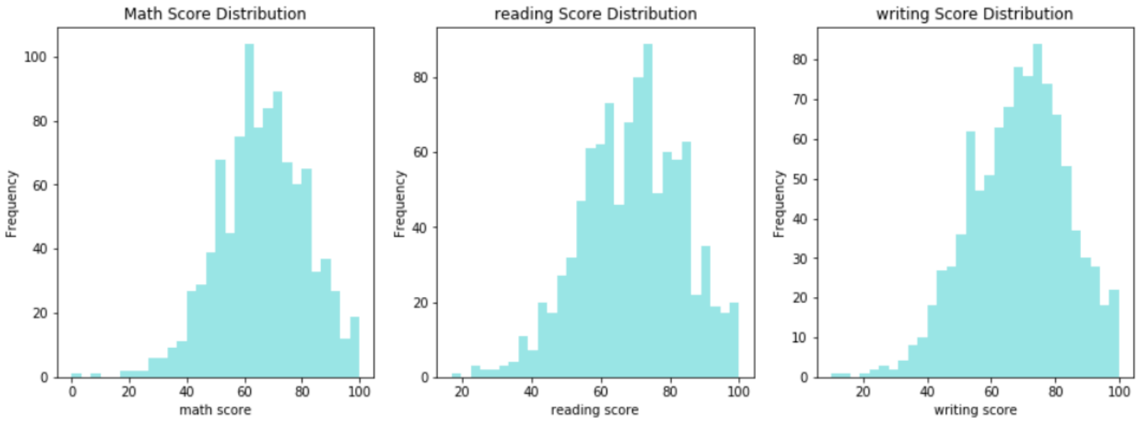


Figure 6. Distribution plots for Subject Scores

The covariance analysis shown that the scores among the three subjects were highly correlated, which means that's students who scored high in one subject were likely to score high in other two subjects Figure 7.

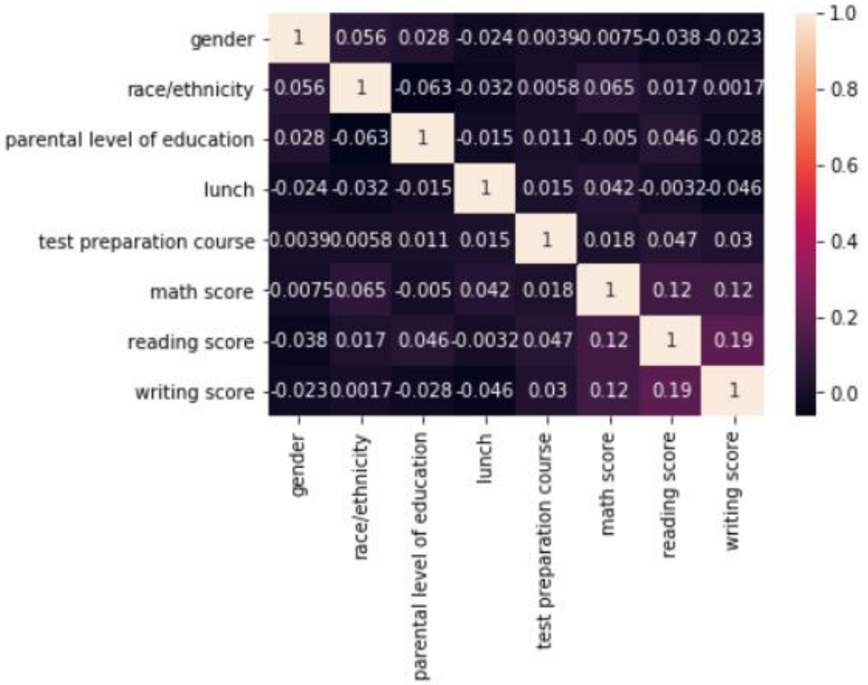


Figure 7. Correlation Plot

Task 4:

In our project, the Random Forest Regression Model used to predict the subject scores since it showed good results in preliminary analysis. After that, based on empirical evidence, best parameters to model the relationship by Random Forest selected, as they give the ideal results in minimizing the Root Mean Squared Error (RMSE) on test data. The results of the model on test data are given in Table 2.

Table 2. Model Performance on Test Data

Subject	**RMSE**
Math	15.9
Reading	16.1
Writing	16.3

3. Conclusion

In conclusion, as it is evident from the results that the progress of the students in different subjects is dependent on different factors and can be predicted using machine learning models. In future, it recommended adding other factors like student’s class performance, health attributes, attendance etc. in order to make the model more robust and more reliable.

References

- Deitel, P. J., & Deitel, H., (2020), Intro to python for computer science and data science: Learning to program with AI, Big Data and The Cloud

- Matplotlib_pyplot_bar,URl:https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.bar.html

- Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

- Grus, J. (2015). Data science from scratch: first principles with Python. First edition. Sebastopol, CA: O'Reilly.

- McKinney, W. (2013). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media