Data Analysis 2 Report

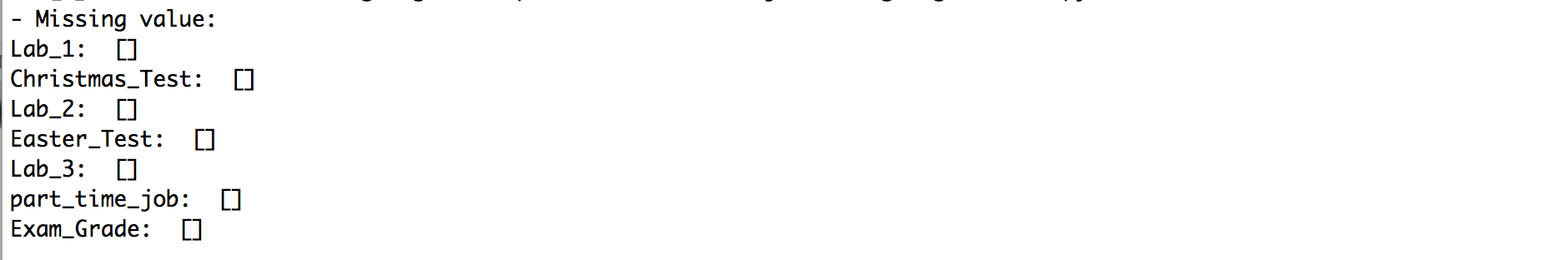
Analysis Results on Student Grades Data Set

# Introduction

In this report, I’ll be providing details on the analysis results of student grades data set from 5 aspects which are initial analysis, simple linear regression, multiple linear regression, K-means cluster analysis and principal components analysis respectively.

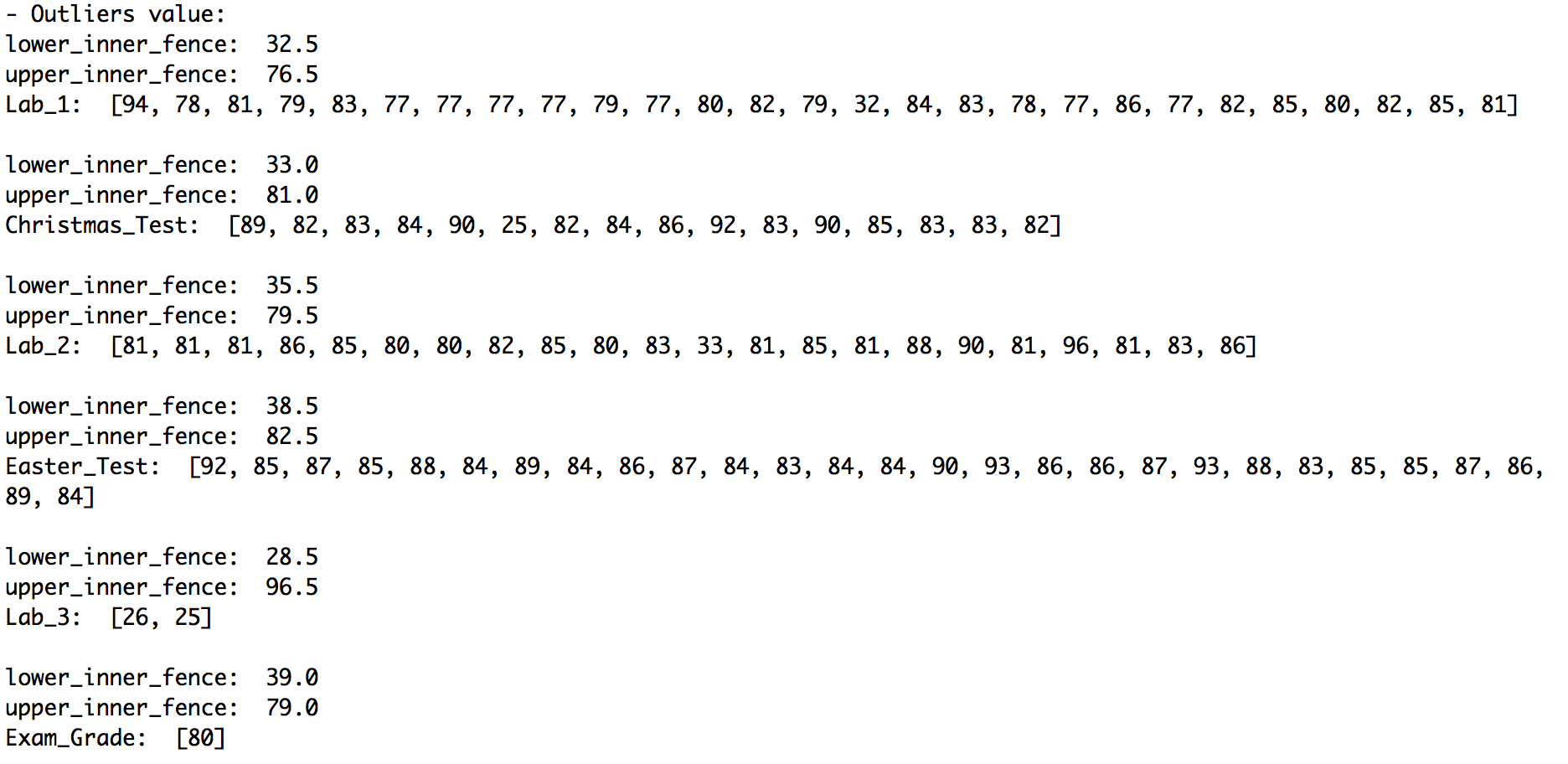
# Initial Analysis

Before carrying out any analysis, I checked the data set to see if there is any missing value by looping through each data point and adding the index of value equal to 0 or Nah. The output (P 1) shows there is no any missing value in the data set.



(P 1)

The output (P 2) shows the mild fence of each test grade’s data and the outliers exiting in each set. In terms of the treatment, I want to just keep all these outliers because the data set itself contains the CA grades and end-of-module exam grades and students’ performance may vary from time to time, so the existence of these outliers is reasonable and it could help with building the linear model predict the end-of-module exam grades.



(P 2)

The histograms (P 3) below show the distribution of grades in each CA element, students who have a part-time job and grades of exam grades. For most CA, the number of students with grades spanning from 40 to 50 are the largest except for Lab 3, the number students with score in the range within 50 and 60 seem to be greater than the numbers of students with in other ranges. For the distribution of students who have/don’t have a part-time job, they both occupy a half.



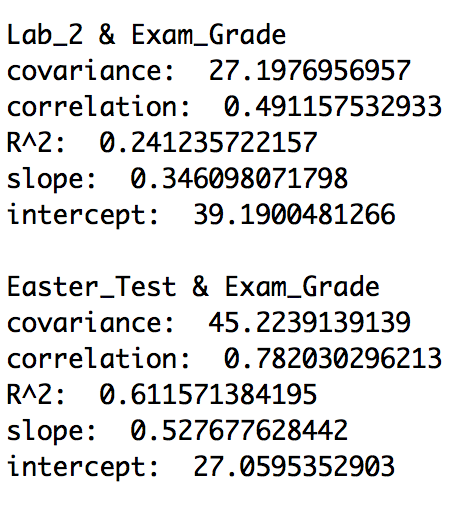
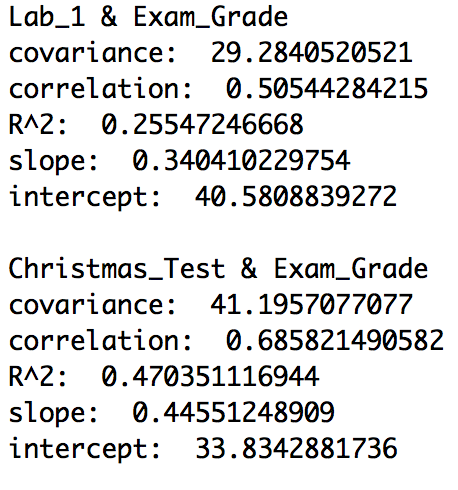
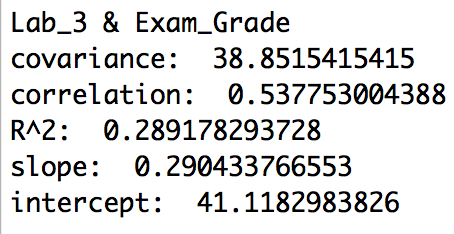






(P 3)

Then I calculated covariance and correlation of each CA grades against the end-of-module r exam grades as (P 4) shows. Covariance measures how each CA grades and end-of-module exam grades vary from their means and correlation indicates the relationship between each CA grades and end-of-module exam grades in statistical way. Both statistics helped me to find out the model to predict the end-of-module exam grades in next section.

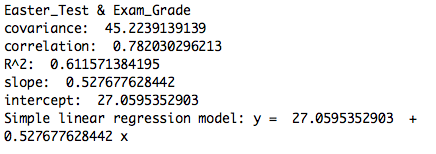
 

（P 4）

# Simple Linear Regression

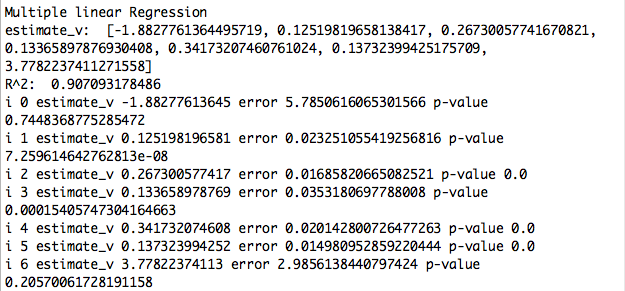
To find out the best simple linear model to predict the end-of-module exam grades, I took advantage of the covariance and correlations calculated in last section. Among the five bivariate data sets, Easter test grades and end-of-module exam grades seemed to have a stronger positive relationship than others as their correlation is 0.782 which is the highest.

Then I calculated the coefficients based on these two variables, slope and intercept are 0.528 and 25.06 respectively. What’s more, the determination of coefficient (R^2) indicates approximately 61.16% variance in end-of-module exam grades is predictable from Easter test grades. So, the best simple linear regression model predicting predict the end-of-module exam grades is y = 0.528 x + 25.06.



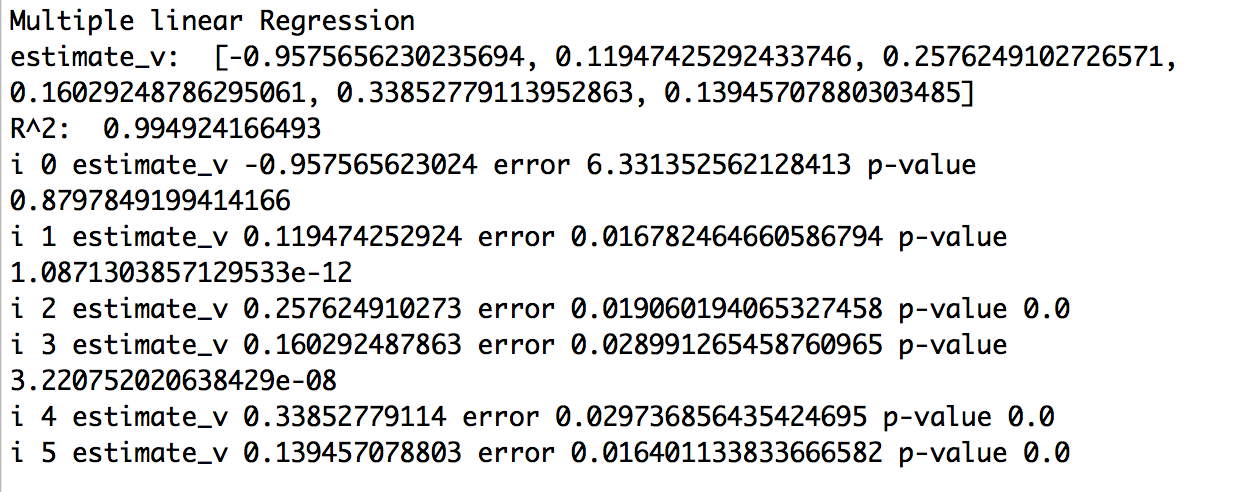
# Multiple Linear Regression

In order to get a more accurate model to predict the end-of-module exam grades, more variables need to be considered instead of one single variable. I put all CA grades as independent variables against end-of-module exam grades as independent variables through stochastic gradient descent to find the coefficients of the multiple linear regression model and then calculated p-values and R^2 to help me determine the usefulness of my model and the significance of each of the factors included.



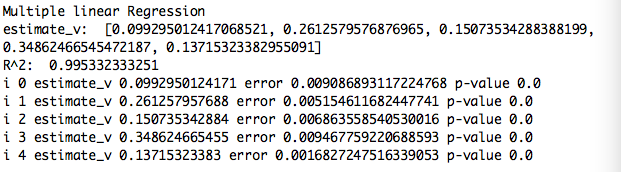
(P 5)

The output (P 5) shows the estimated coefficients [constant\_term, Lab\_1, Christmas\_test, Lab\_2, Easter\_test, Lab\_3, Part\_time\_job] based on the grades of all CA elements and whether or not a student has a part-time job. The determination of coefficient (R^2) implies 90.7% variance in end-of-module exam grades can be predicted from the selected variables, which means it’s a good model. Yet, the P-values of constant term and dummy variable of whether or not a student has a part-time job are very high and they are 0.745 and 0.206 respectively, suggesting the true coefficients of these two variables are not significantly different from 0 and they are much random rather meaningful to the model. So, I decided to take out these variables and ran it again.



(P 6)

The output (P 6) shows the coefficients after I removed dummy variable for part-time job, the P-value of constant term was still very high so that I decided to remove the constant term from independent variables.



(P 7)

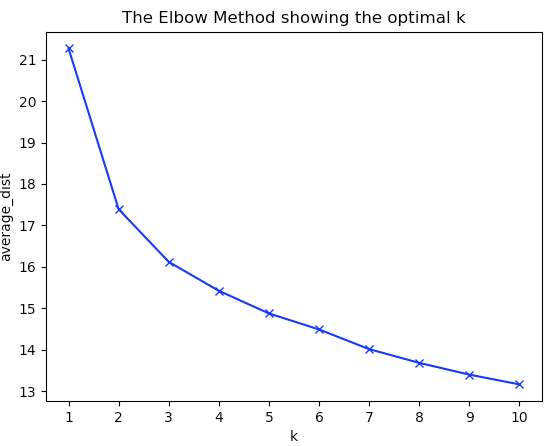
The output (P 7) shows the coefficients [Lab\_1, Christmas\_test, Lab\_2, Easter\_test, Lab\_3]

calculated after I took out the constant term and the dummy variable indicating whether or not a student has a part-time job. This time all P-values are 0 which implies the true coefficients are non-zero and meaningful. So, lab 1 grades, Christmas test grades, lab 2 grades, Easter test grades and lab 3 grades are the right features which should be used to produce the model predicting end-of-module exam grades. Furthermore, the determination of coefficient (R^2) shows up to 99.5% of variance in end-of-module exam grades is predictable from these variables.

To conclude, the multiple linear regression model is

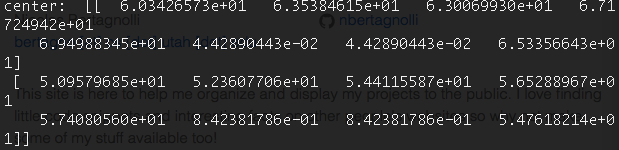
*End-of-module\_exam\_grade = 0.099 \* Lab1\_grade + 0.261 \* Christmas\_test\_grade + 0.151 \* Lab2\_grade + 0.349 \* Easter\_test\_grade + 0.137 \* Lab3\_grade* and this is probably the best model to predict end-of-module exam grades based on the give data set.

# K-Means Clusters Analysis

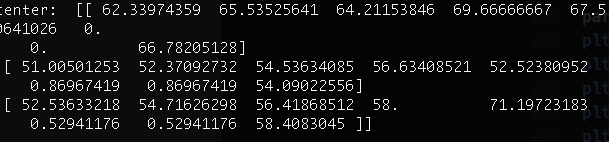
K-means clusters analysis helps to find the groups in the student grades data with the number of group called K. The student grades data was categorized by the features provided in in the data set. To get the output, I used Kmeans model from sklearn package to produce clusters model with number of clusters up to 10.

The elbow plot (P 8) shows the number of student grades clusters against the average distance between the cluster centroid. This one is not the typical case of elbow plot because it doesn’t start to flatten significantly at any points and it’s very ambiguous to say how many clusters are there in the data. But I still see that the line goes much smoother after K = 2 or K = 3. So, in my point of view, the optimal K can either be 2 or 3. (P 8)

In order to answer the Question 5 in project specification, I took the optimal K of 2. Then the python code splited the student grades of this year-long module into two clusters and printed out the centers of these two clusters. In the output (P 9), the second last element in each list represents whether or not a student have a part-time job and the last element represents the average end-of-module exam grades in that cluster. The second last element is very close to 0 in the firtst list, while that is close to 1 in second list, which means students in cluster 1 don’t have a part-time and they have a higher average end-of-module exam grade of 65.33 than students in cluster 2 with a part-time job have a lower average end-of-module exam grade of 54.76. The findings in K-means clusters analysis supports the hypothesis that having a part-time job can have a negative impact on a student’s performance.

 (P 9)

Besides K=3, I also wanted to see what the clusters K = 3 would tell me. So, I printed the center of each in 3 clusters (P 10). Genrally, the 3 clusters also supports my conclusion in last paragraph. The first cluster representing the grades of student without a part-time job has a higher average end-of-module exam grade, while the other 2 clusters are kind of a mix of students with/withou a part-time job. But these two clusters still have a lower average end-of-module exam grades compared to the firtst cluster.

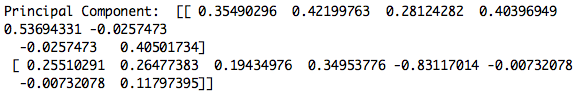


(P 10)

# Principal Component Analysis

Principal component analysis was used here to find the directions of maximum variation in the student grades data thereby reducing the dimensionality of data itself and viewing it in a more appropriate co-ordinate system and the principal components also indicates the direction of max variation in the data set.

By using PCA model from sklearn package, I got the first two principal component which is used to determine the direction or variation in data set as shown in (P 11). The first principal components is [ 0.35490296 0.42199763 0.28124282 0.40396949 0.53694331 -0.0257473 -0.0257473 0.40501734] and the second principal component is [ 0.25510291 0.26477383 0.19434976 0.34953776 -0.83117014 -0.00732078 -0.00732078 0.11797395].



(P 11)

The scree plot (P 12) shows that the first principal components explains about 40% of variation in the data and the second principal components explains approximately 20% of variation. And the scatter plot (P 13) shows the two clusters of student grades data of a year-long module in more appropriate axes based on the first and second components of the data.



(P 12) (P 13)