

Data Analysis

Final Project

Elaine Pei-Ling Chong

[Company Name]

Table of Contents

[1 Introduction 3](#_Toc501143355)

[2 Methodology 3](#_Toc501143356)

[3 Dataset 3](#_Toc501143357)

[4 Data Analysis 7](#_Toc501143358)

[4.1 Descriptive statistics 7](#_Toc501143359)

[4.2 Simple Linear Regression 8](#_Toc501143360)

[4.3 Multiple Linear Regression 11](#_Toc501143361)

[4.4 K-means Cluster Analysis 13](#_Toc501143362)

[4.5 Principal Component Analysis 14](#_Toc501143363)

[5 Conclusion 15](#_Toc501143364)

[6 References 16](#_Toc501143365)

# Introduction

The main purpose for this report is to carry out different analysis on with different purpose on the given dataset such as to predict student’s final exam grade based on their performance on previous tests such as lab test, Christmas test and Easter test.

# Methodology

I conducted different analyses to find out the best solution to predict the exam grade which included statistical analysis, simple and multiple linear regression analysis, k-mean cluster analysis and principal component analysis.

The first part is identification and treatment of outliers, generate visualisations and carry out the statistical analysis. The second part of the report is to determine the test which affects the exam grade the most by carrying out simple linear regression and data visualisation.

# Dataset

The dataset used in this analysis is given by lecture and can found in Moodle (<https://2018-moodle.dkit.ie/pluginfile.php/415016/mod_resource/content/1/FinalProjectData1718.csv)>

The data file contains multiple columns which describe student's result. To use the data inside the CSV file, I have read the file and convert into python object.

1. studentid: The Id of student
2. Lab 1: The result of lab 1
3. Christmas Test: The result of Christmas test
4. Lab 2: The result of lab 2
5. Easter Test: The result of Easter test
6. Lab3: The result of lab 3
7. parttimejob: 1 means have part time job; 0 means do not have part-time job
8. Exam Grade: The result of exam grade

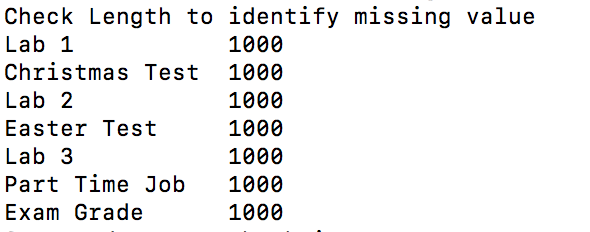
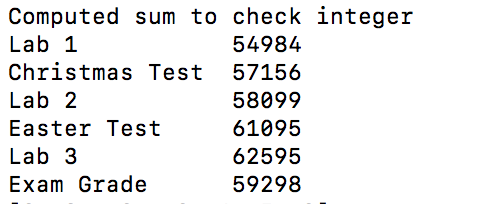
Identification and handling of missing data

Missing Data is data that is not present due to whatever reason, often marked as "NA" (not available). If data is missing and no treatment is performed, it will have a significant impact on the statistics result.

Treatments for Missing Data:

1. Delete the sample of missing data value
2. Impute the value
3. Remove the variable and find another variable which has similar aspects

In the dataset, the number of every variable's length is checked to determine if there any missing data and perform a straightforward statistic function (sum()) to make sure every data in each variable is an integer. The result shows that all the variables have the same length which is 1000 and the function sum carried out successfully.

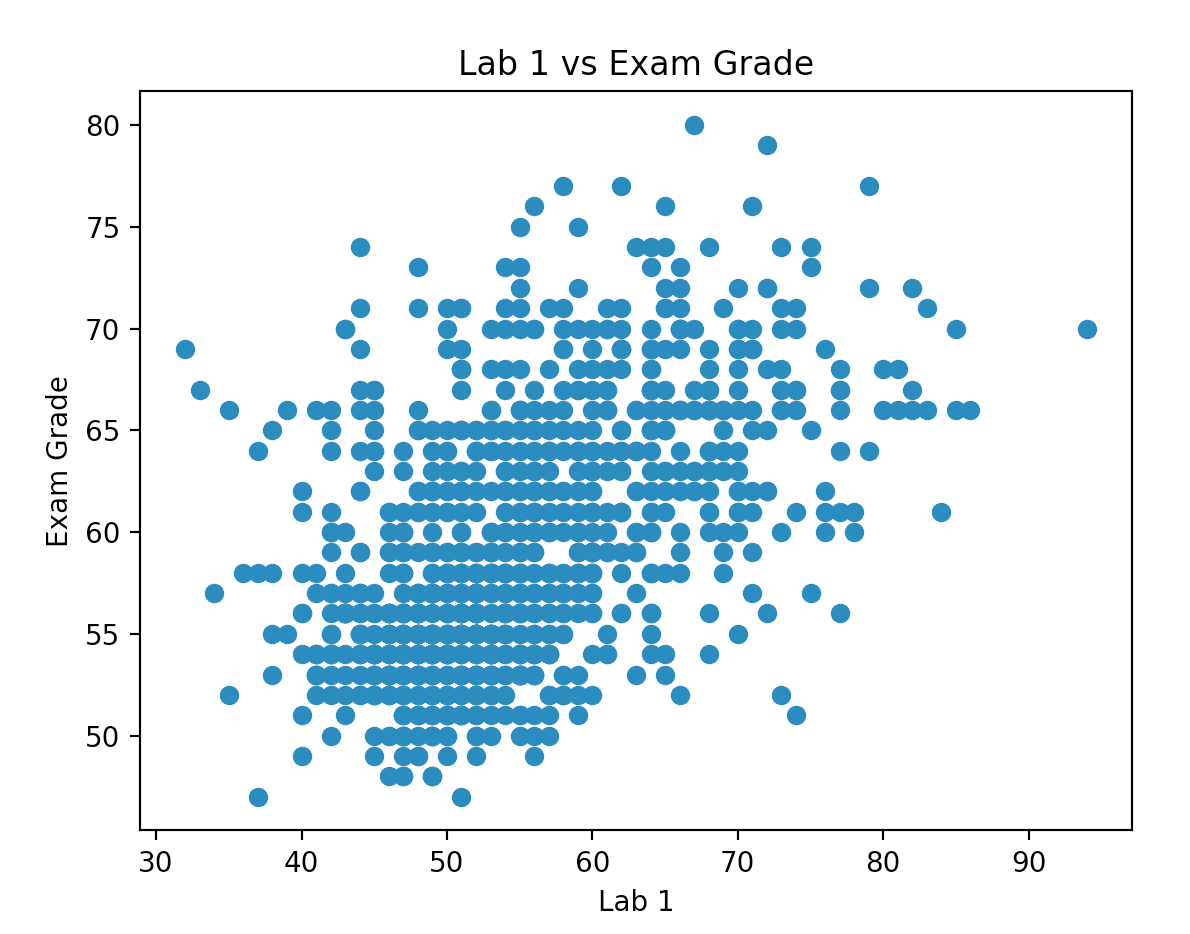
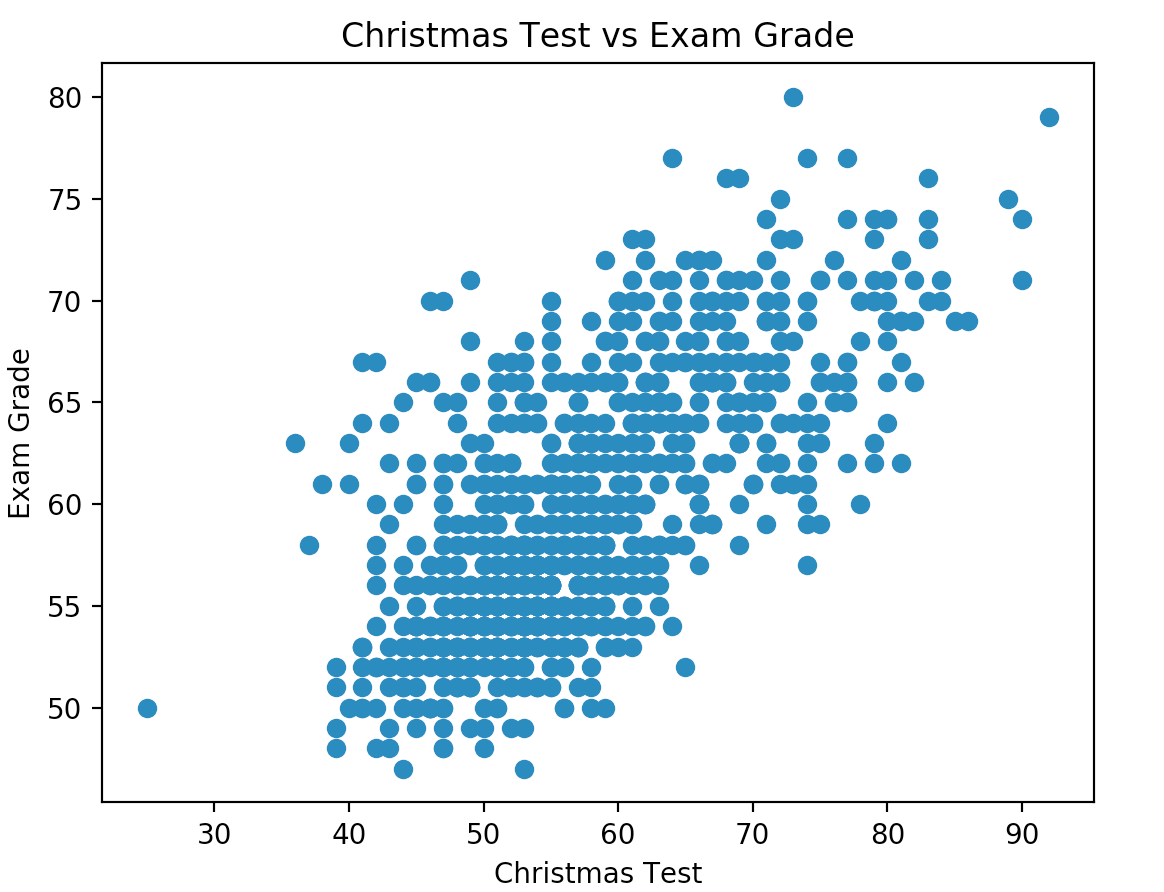
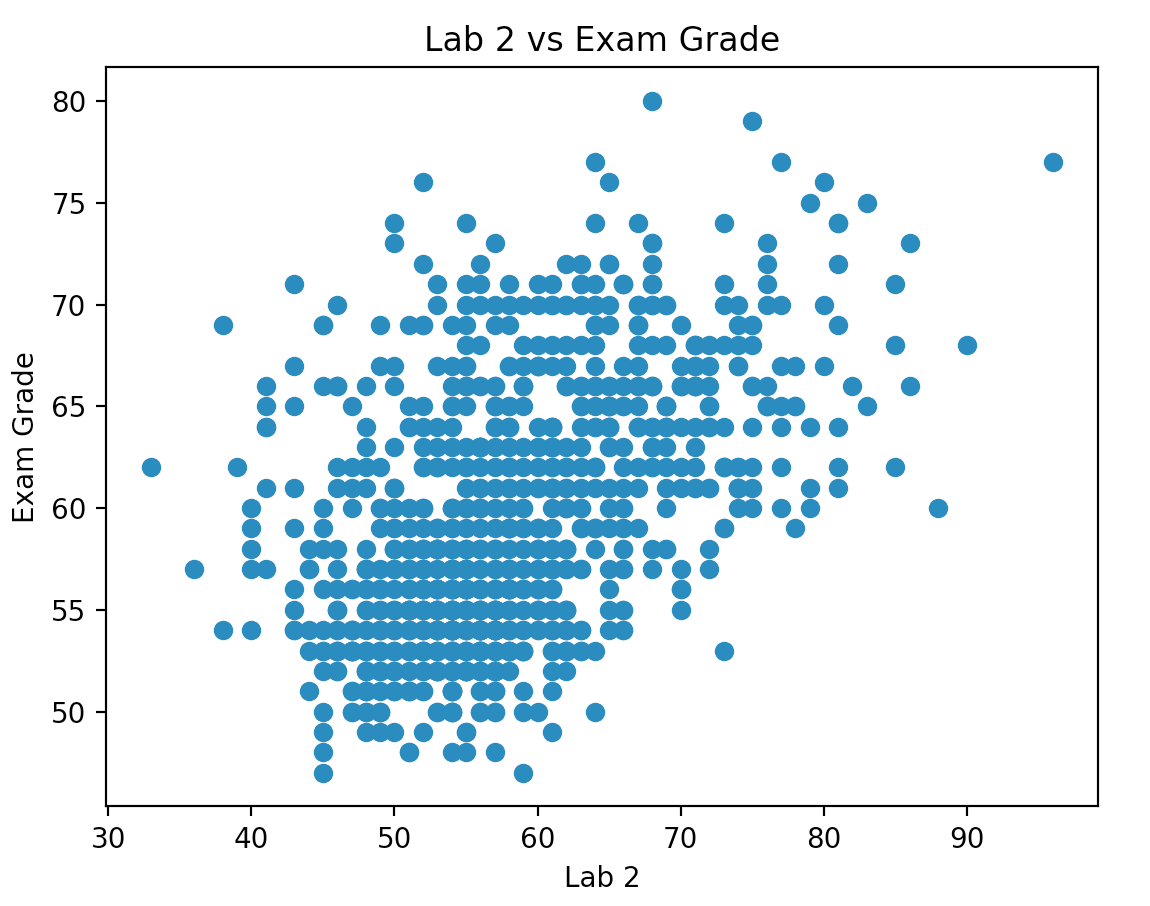
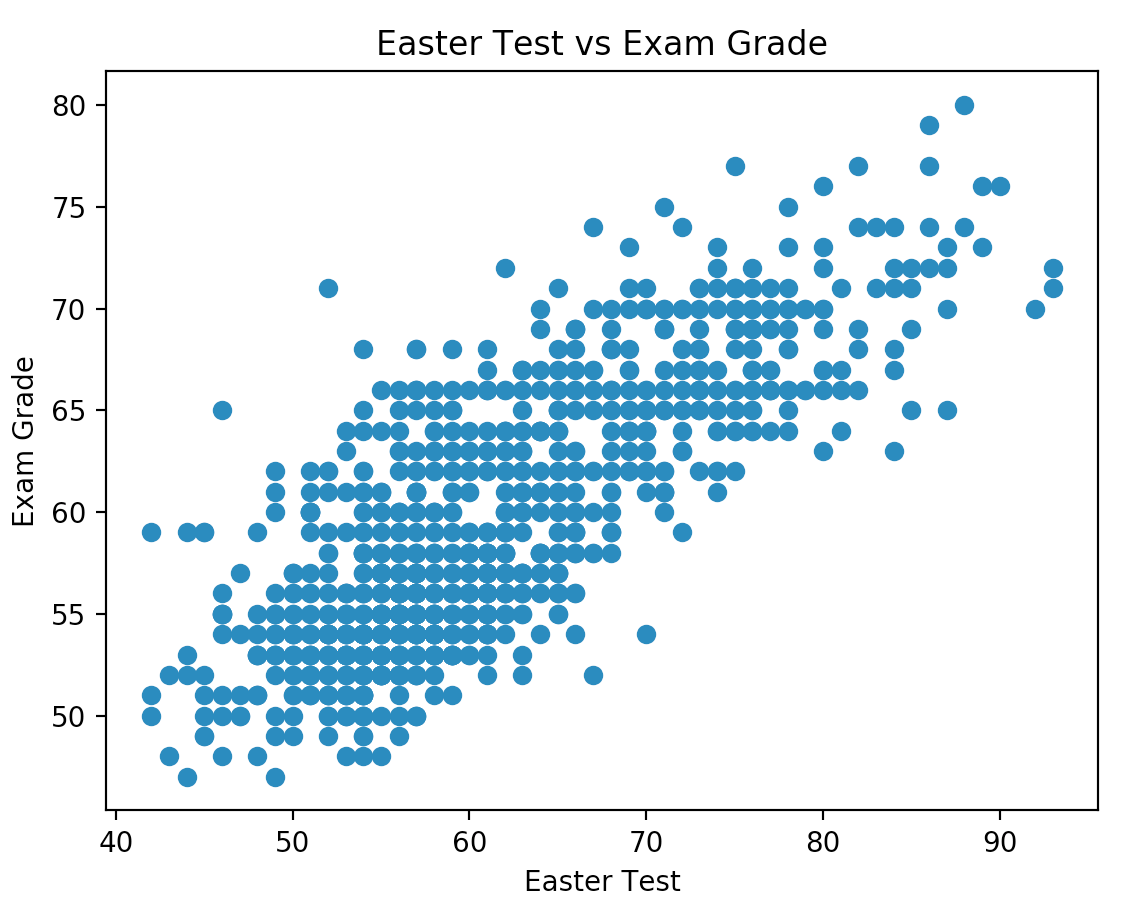
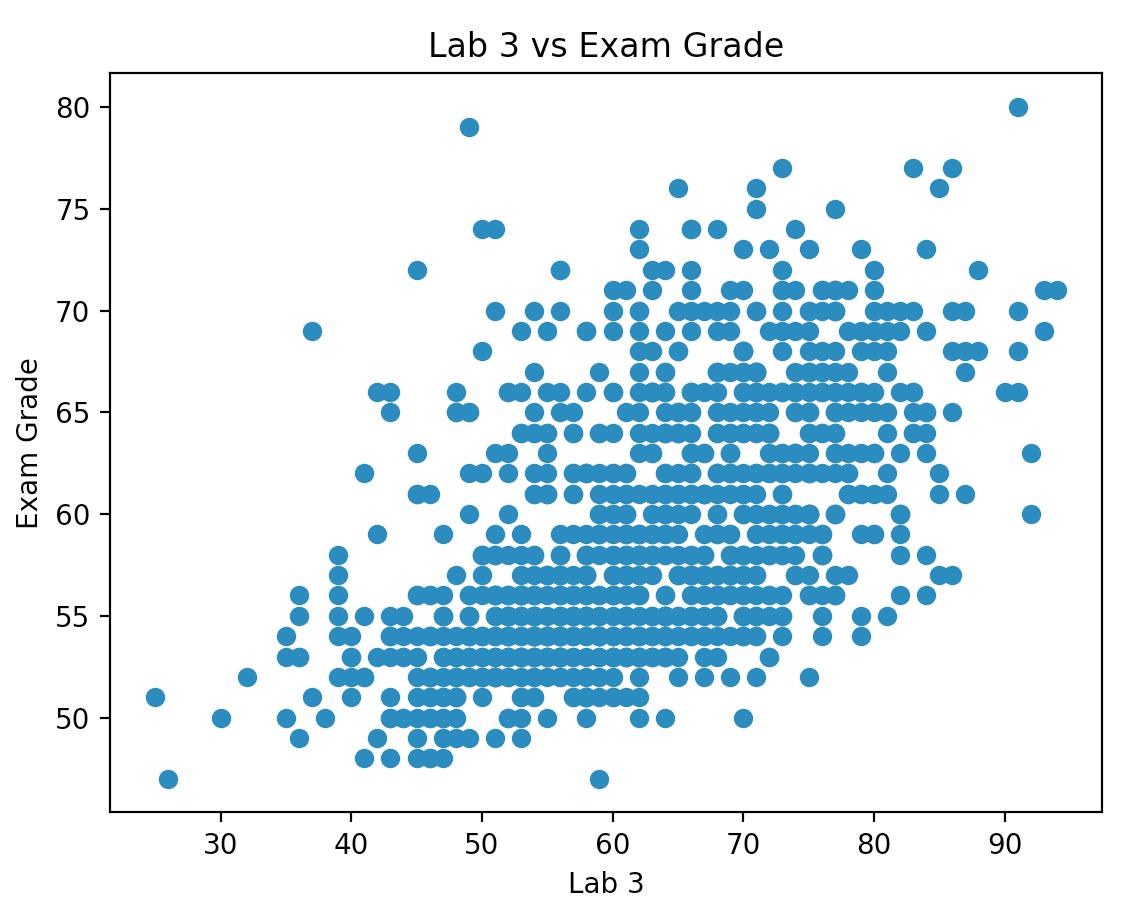
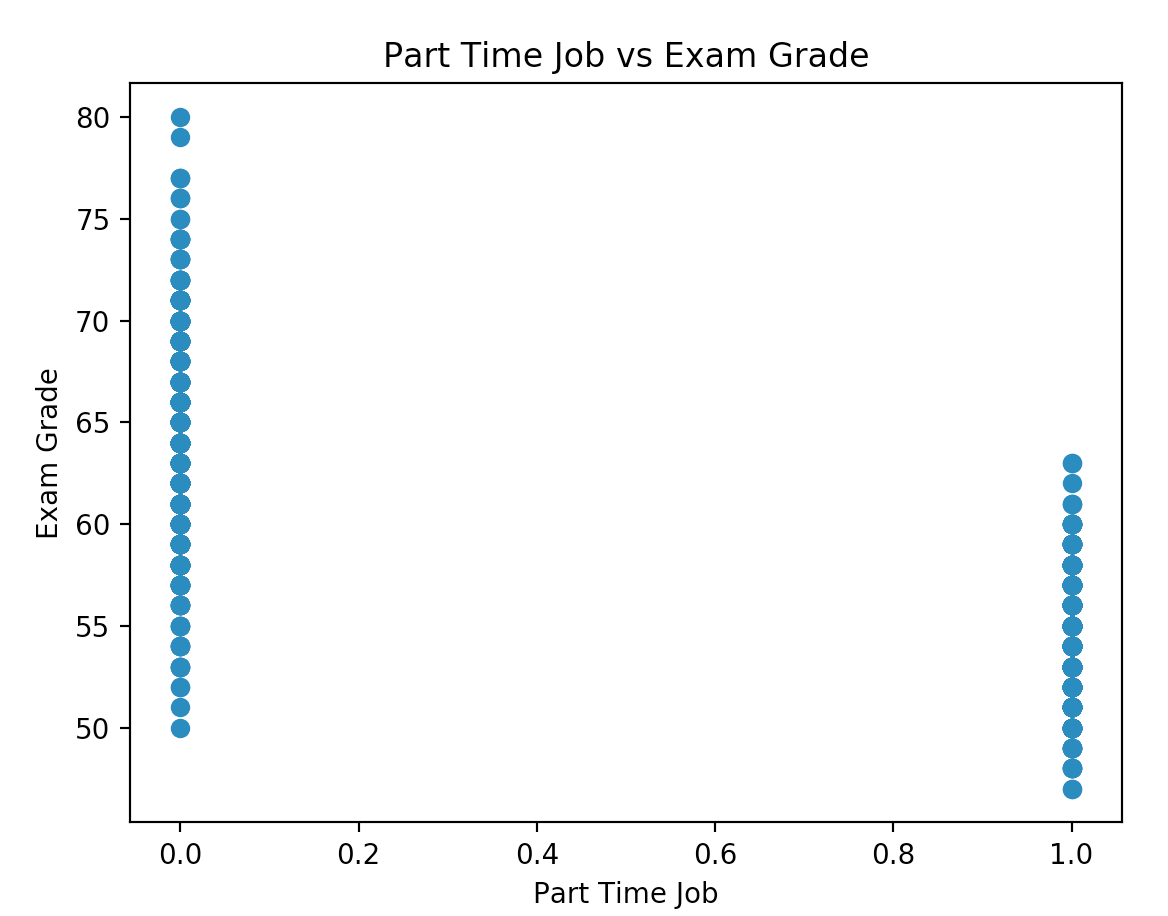
Treatment to outliers

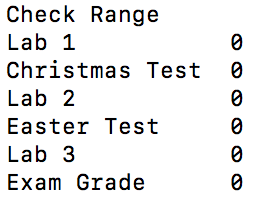
All the descriptive statistics are sensitive to outliers such as mean, standard deviation, variance and correlation.

An outlier is a value that notably differs from other values in a data set. One of the ways to detect outliers visually is to plot a scatter plot.

Another way to identify outliers is using interquartile range (IQR), an outlier is identified when the value is greater than the number of (Q3+ 1.5 \* IQR) or the value is lesser than then number of (Q1 - 1.5 \* IQR). If an outlier is greater than (Q3 + 3 \* IQR) or less than (Q1 – 3\*IQR), it is considered as an extreme outlier.

As students' result might vary from 0 to 100 thus removing outliers indicated by using interquartile range (IQR) is not appropriate for this dataset. To determine outliers which is bigger than the 100 and smaller than the value 0, I have plotted the scatter plot for every variable versus exam to check there any outliers. Other than that, I computed all continues variables to check is the data for student's test is from the range of 0 to 100. When the checkRange() function is called, it will return 0 when all the data is inside the range and returns one if one of the data is out of the range. From the result, none of the variables has outliers higher than 100 and smaller than 0 and for the categorised variable such as part-time job, there is no value other than 0 and 1.



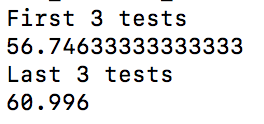
# Data Analysis

## Descriptive statistics

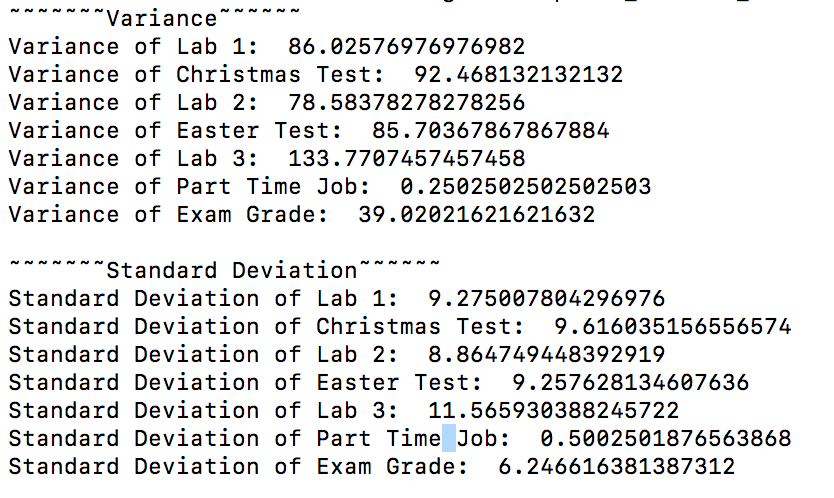
For categorised variables, frequency count (mode) is calculated. For continuous variables, mean, median, standard deviation and variance are computed.

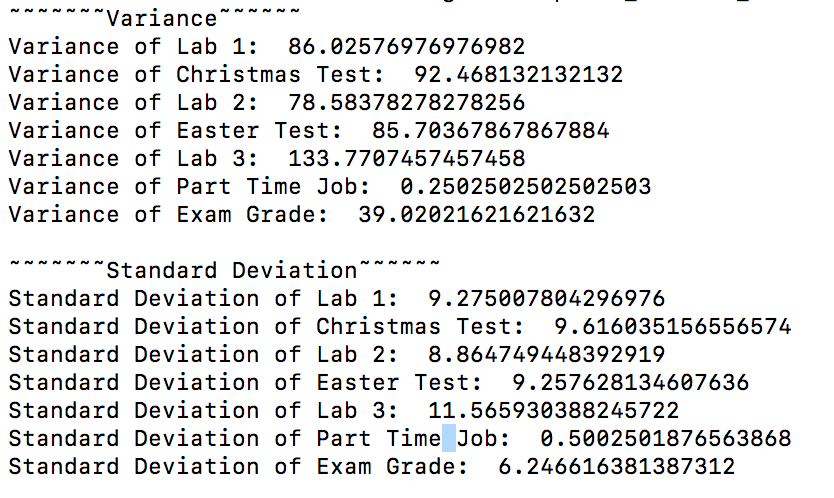
|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Mean | Median | Mode |
| Lab 1 | 54.984 | 51.0 | 52 |
| Christmas Test | 57.156 | 53.0 | 53 |
| Lab 2 | 58.099 | 54.5 | 54 |
| Easter Test | 61.095 | 57.0 | 57 |
| Lab 3 | 62.595 | 58.0 | 64 |
| Part Time Job |  |  | 0,1 |
| Exam Grade | 59.298 | 56.0 | 54 |

Mean for



From the result obtained, when at the end of the year, students tend to have a better performance compares to the beginning of the year as the mean for first 3 test is 56.75 and the mean for last three tests is 60.99. The mode for the part-time job is 0 and 1 which means the number of students with a part-time job is the same as the number of students without a part-time job.

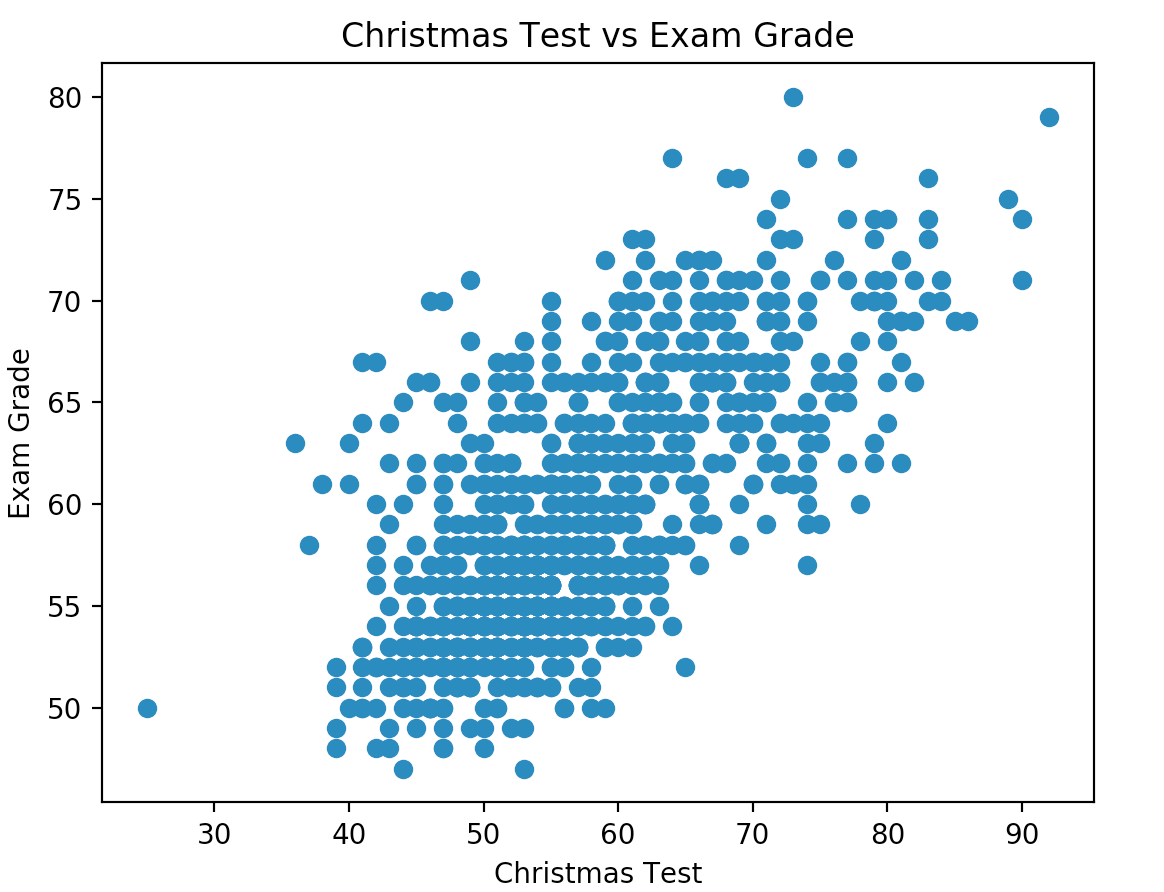


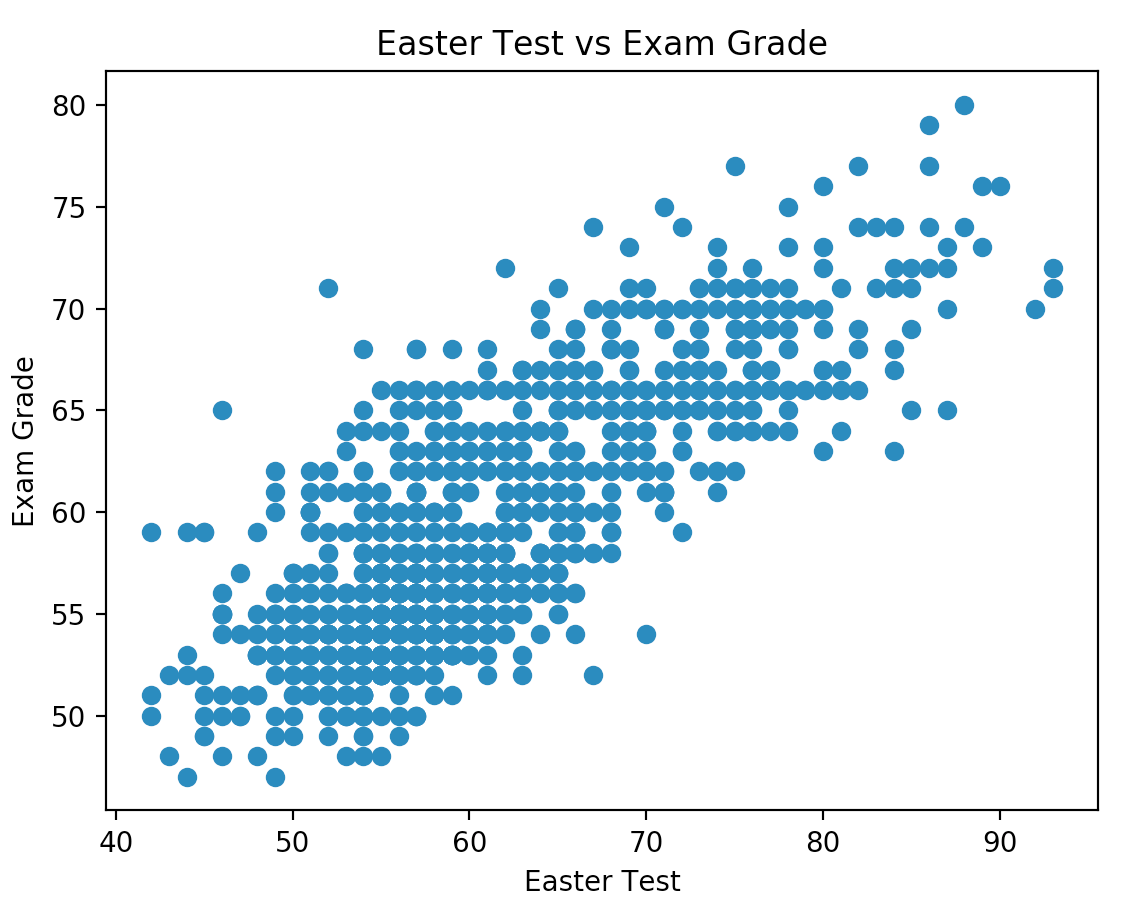


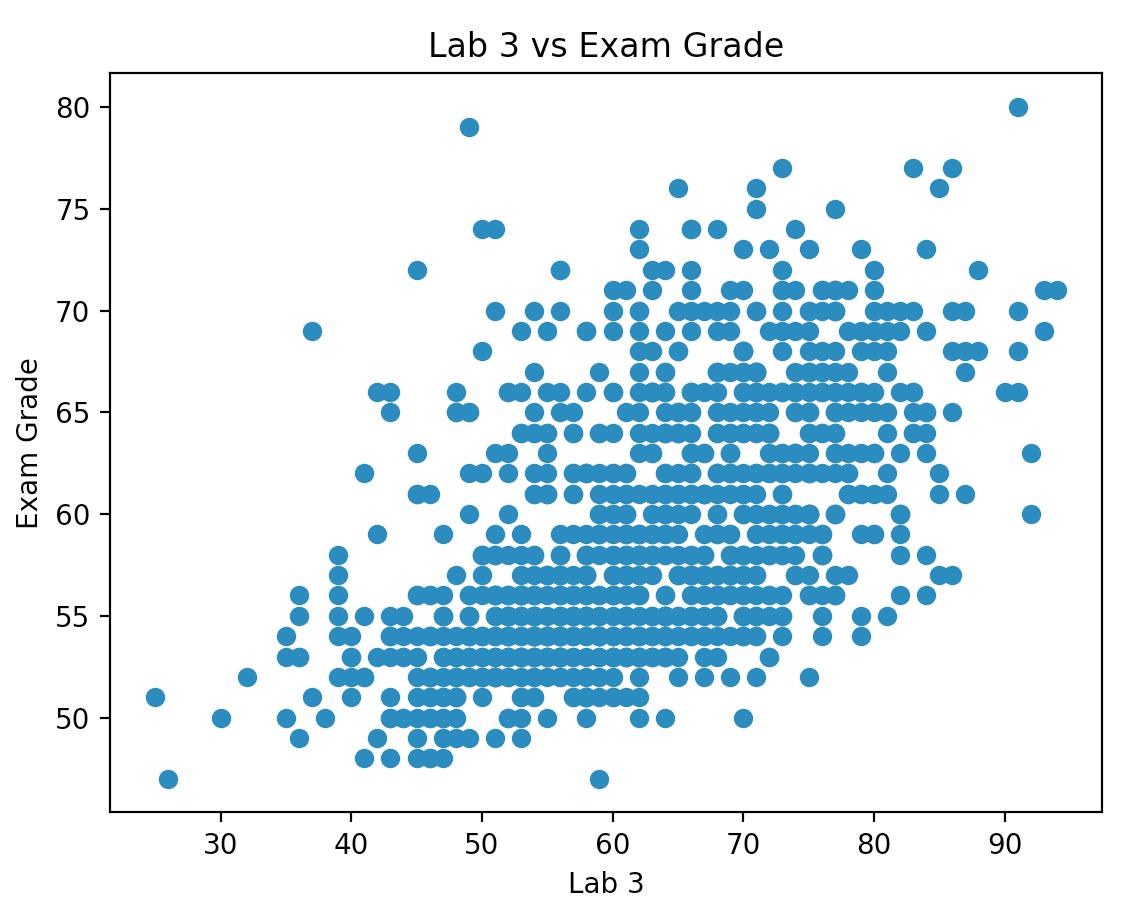
Variance and the standard deviation is calculated to measure the spread of the data around the mean. A high value of standard deviation explains that data is more spread away from mean and a low value of standard deviation explains that more data is around the mean. The standard deviation for exam grade is 6.25 which tells that the students' exam grade is relatively close to each other.

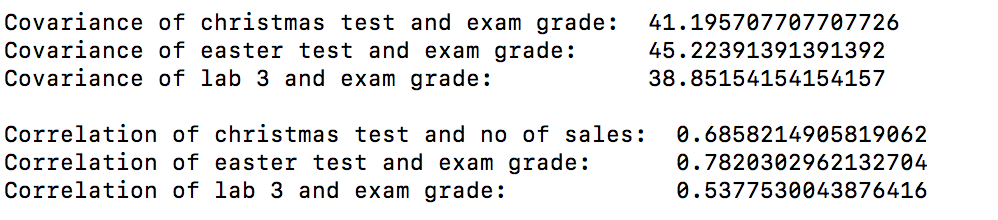
## Simple Linear Regression

In scatter plot, if the bivariate variables have a relationship, the data points are linear. When x increases and y also increase, it is a positive linear correlation. When x increase and y decreases, it is a negative linear correlation. In order to predict the exam grade based on 1 variable, I have chosen 3 variables where all the data points very dense together and have a linear correlation. There are Christmas test, Easter test and lab test 3. Correlation between those variables and exam grade were computed and the result shown below. Correlation coefficient, r has a value from -1 to 1, if r is a negative value, the relationship between 2 variables has a negative linear correlation. If the r is a positive value, the relationship between 2 variables has a positive linear correlation. R will be closer to 1 or -1 when both variables have a very strong relationship and the relationship is weaker when the r is closer to 0. From the values obtained, the correlation for Easter test and exam grade is 0.782 which is highest and shows both variables have a stronger relationship.



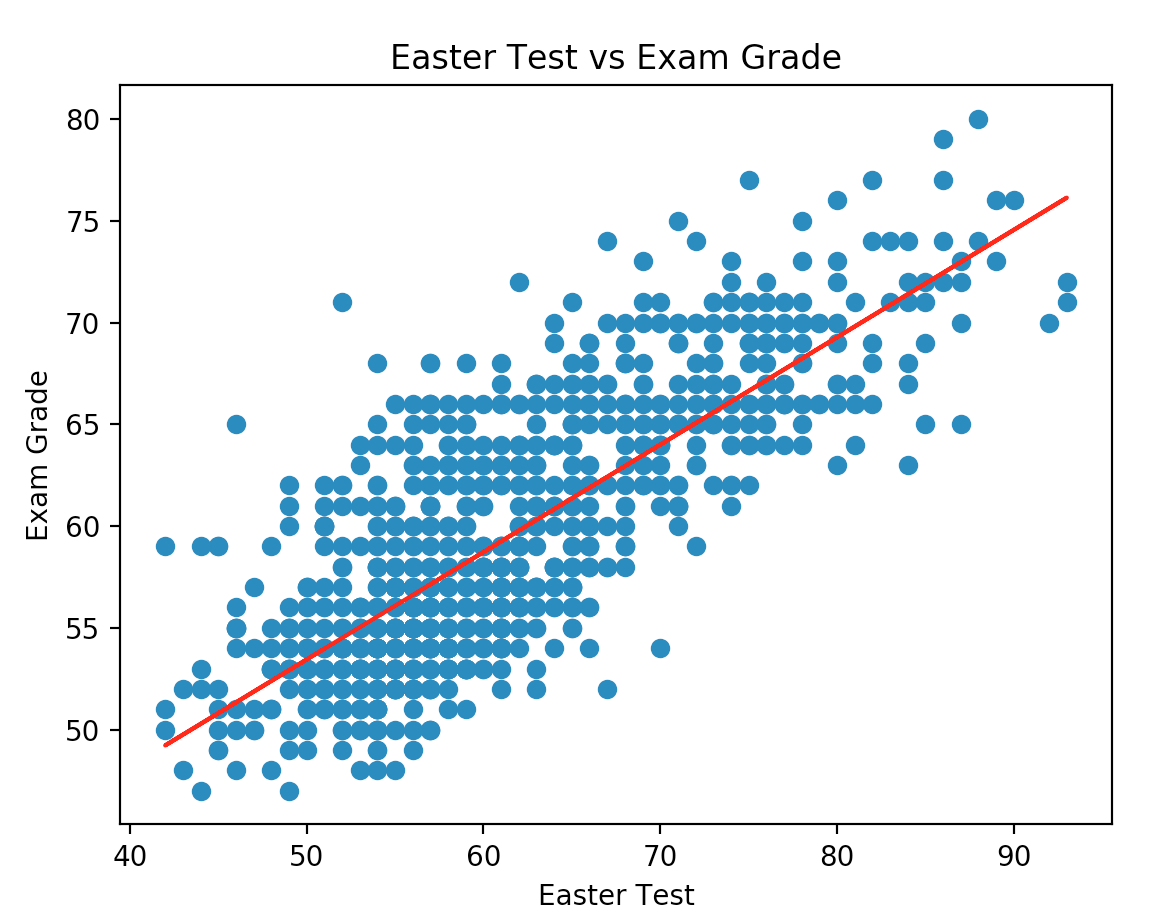


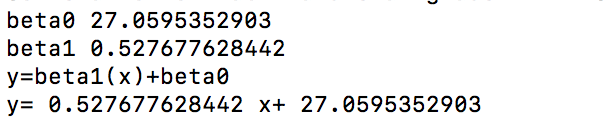




|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | Y | Beta0 | Beta1 | Equation | R-squared |
| Christmas Test | Exam Grade | 33.83 | 0.45 | Y=0.45\*x+33.83 | 0.4704 |
| Easter Test | Exam Grade | 27.06 | 0.53 | Y=0.53\*x+27.06 | 0.6116 |
| Lab 3 | Exam Grade | 41.12 | 0.29 | Y=0.29\*x+41.12 | 0.2892 |

I have generated least squares regression to obtain the regression line. It means hypothesising the linear relationship between the Easter test and exam grade and choose the line that minimises the errors between the model and data. The regression line can be used to predict the values for future use and the outliers will have a significant impact on it. I have generated 3 regression line for 3 variables I chose and the r-squared is calculated. From the result, as the higher the r-squared value, the better fit for the data, thus the best simple linear regression line is y=0.53\*x+27.06 for bivariate data (Easter test and the value) because this line has the value for r-squared =0.6116 which is the highest value. The purpose of the equation is used for estimation.





## Multiple Linear Regression

Multiple linear regression(MLR) is built to determine there any relationship between the independent variables in a model. There more independents variables in a data set, the more it can be able to respond to the reality. I computed a multiple linear regression model for all the variables and calculated the P-value to determine whether the variables should be the model. The result is shown below:

Dependent Variable: Exam Grade

Method: gradient descent stochastic

Sample: 5

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Coefficients | Standard Error | P-value |
| Constant | 3.321 | 0.28673894 | 0.0 |
| Lab 1 | 0.106 | 0.00102660 | 0.0 |
| Christmas Test | 0.248 | 0.00158275 | 0.0 |
| Lab 2 | 0.136 | 0.00100204 | 0.0 |
| Easter Test | 0.324 | 0.00141047 | 0.0 |
| Lab 3 | 0.130 | 0.00124660 | 0.0 |
| Part-time job | 0.410 | 0.23058994 | 0.07480 |
|  |  |  |  |
| Multiple R-squared = 0.98792045309 | | | |

The following is the linear equation for this regression model

Exam Grade= 3.321+ 0.106 x Lab 1 + 0.248 x Christmas Test + 0.136 x Lab 2 + 0.324 x Easter Test + 0.130 x Lab 3 + 0.410 x Part time job

From the results, although the p-value for the constant, lab 1, Christmas test, lab 2, Easter test and lab 3 have the p-value =0, the actual value is nearly zero but not zero. While most of the coefficients have very small p-value, the coefficient for "part\_time\_job" has p-value = 0.0748 which is slightly higher than others. As a result, the "part-time job" is removed from the model and I have created another MLR model.

Dependent Variable: Exam Grade

Method: gradient descent stochastic

Sample: 5

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Coefficients | Standard Error | P-value |
| Constant | 2.020 | 0.04627938 | 0.0 |
| Lab 1 | 0.110 | 0.00039005 | 0.0 |
| Christmas Test | 0.238 | 0.00171365 | 0.0 |
| Lab 2 | 0.140 | 0.00066773 | 0.0 |
| Easter Test | 0.345 | 0.00202805 | 0.0 |
| Lab 3 | 0.134 | 0.00037446 | 0.0 |
| Multiple R-squared = 0.995891091449 | | | |

Exam Grade= 2.020+ 0.110 x Lab 1 + 0.238 x Christmas Test + 0.140 x Lab 2 + 0.345 x Easter Test + 0.134 x Lab 3

R-squared, r2 is a proportion (explained variation / total variation) used evaluate the regression model. It is always a value between 0 and 1. When the r2 =1, all the data points lie exactly on the regression line, when the r2 closer to 1, it means that the regression line explains most of the variability, when the r2=0, it means that the regression line is horizontal and explains none of the variability. The r2 value for both MLR models is computed, after the variable, part-time job is removed, the r2 value for model 2 increased from 0.9879 to 0.9959 which indicates 99.56% of the variation in exam grade is being accounted for in the model. I have generated the linear regression with stochastic gradient descent to reduce the time to generate the equation for the line. The equation to estimate the exam grade is

Exam Grade= 2.020+ 0.110 x Lab 1 + 0.238 x Christmas Test + 0.140 x Lab 2 + 0.345 x Easter Test + 0.134 x Lab 3

Consider the coefficients of the model as representing (all-else-being-equal), the impact of each factor is estimated.

All else being equal, each additional point in lab test 1 corresponds to extra 0.11 points in exam grade.

All else being equal, each additional point in Christmas test corresponds to extra 0.238 points in exam grade.

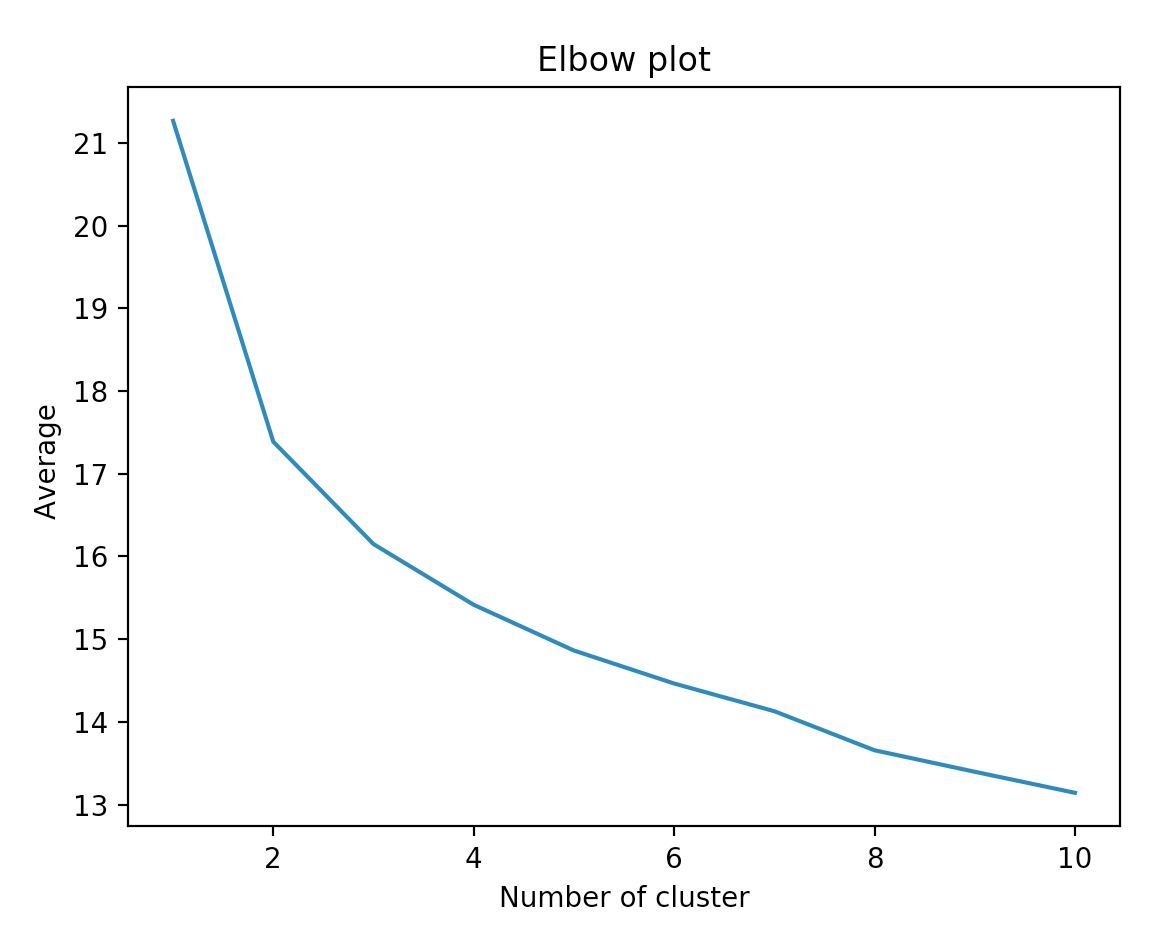
All else being equal, each additional point in lab test 2 corresponds to extra 0.140 points in exam grade.

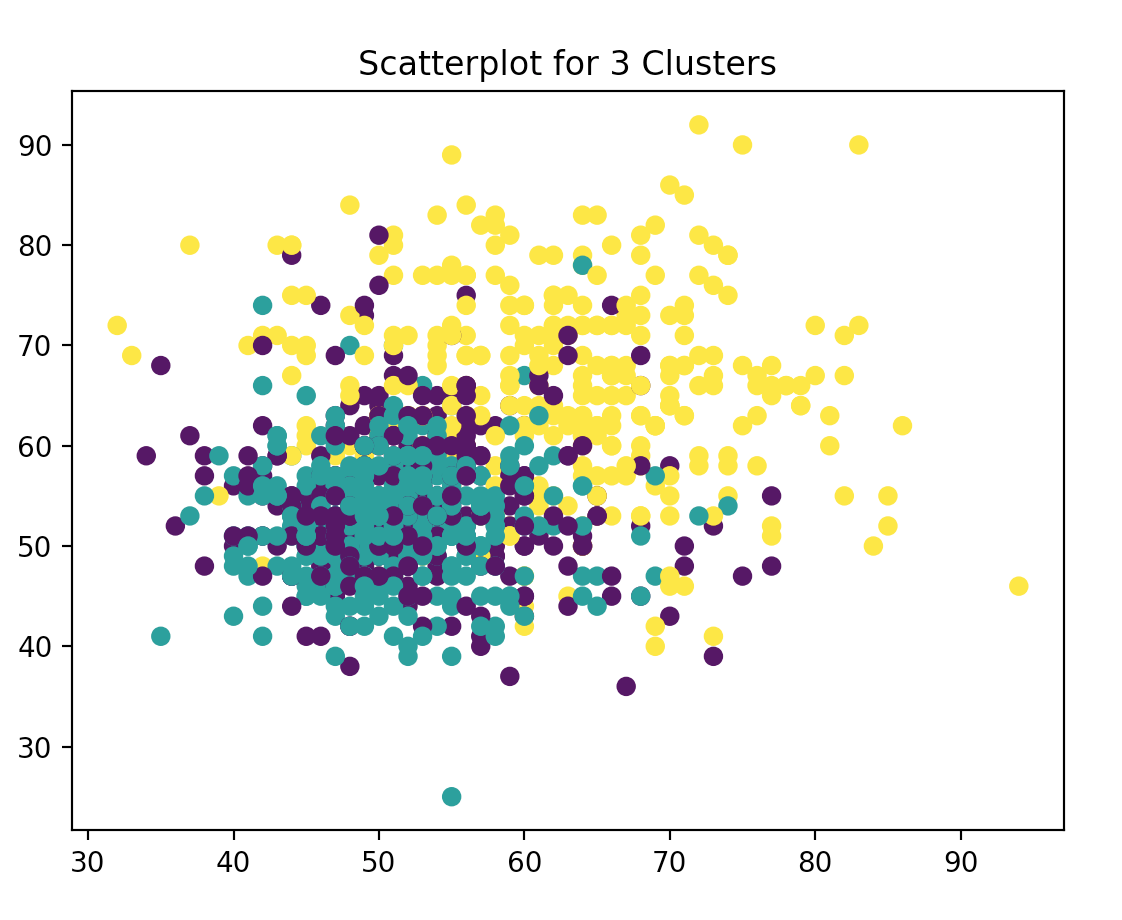
All else being equal, each additional point in Easter test corresponds to extra 0.345 points in exam grade.

All else being equal, each additional point in lab test 3 corresponds to extra 0.134 points in exam grade.

## K-means Cluster Analysis

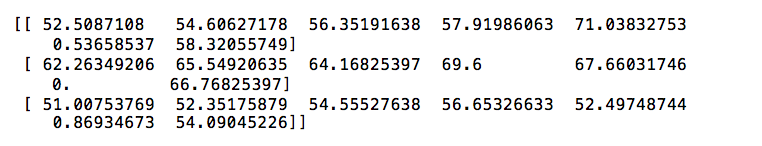
Cluster analysis is a method to classify similar objects into a different group with similar attributes. There have different ways to carry out cluster analysis and k-means cluster analysis is chosen. In order to specify the number of clusters in the analysis, elbow plot is plotted. From the elbow plot, it states that the number of clusters for this data set is 3.





The mean for each variables in each cluster:

[lab1, christmas\_test, lab2, easter\_test, lab3, part\_time\_job, exam\_grade]

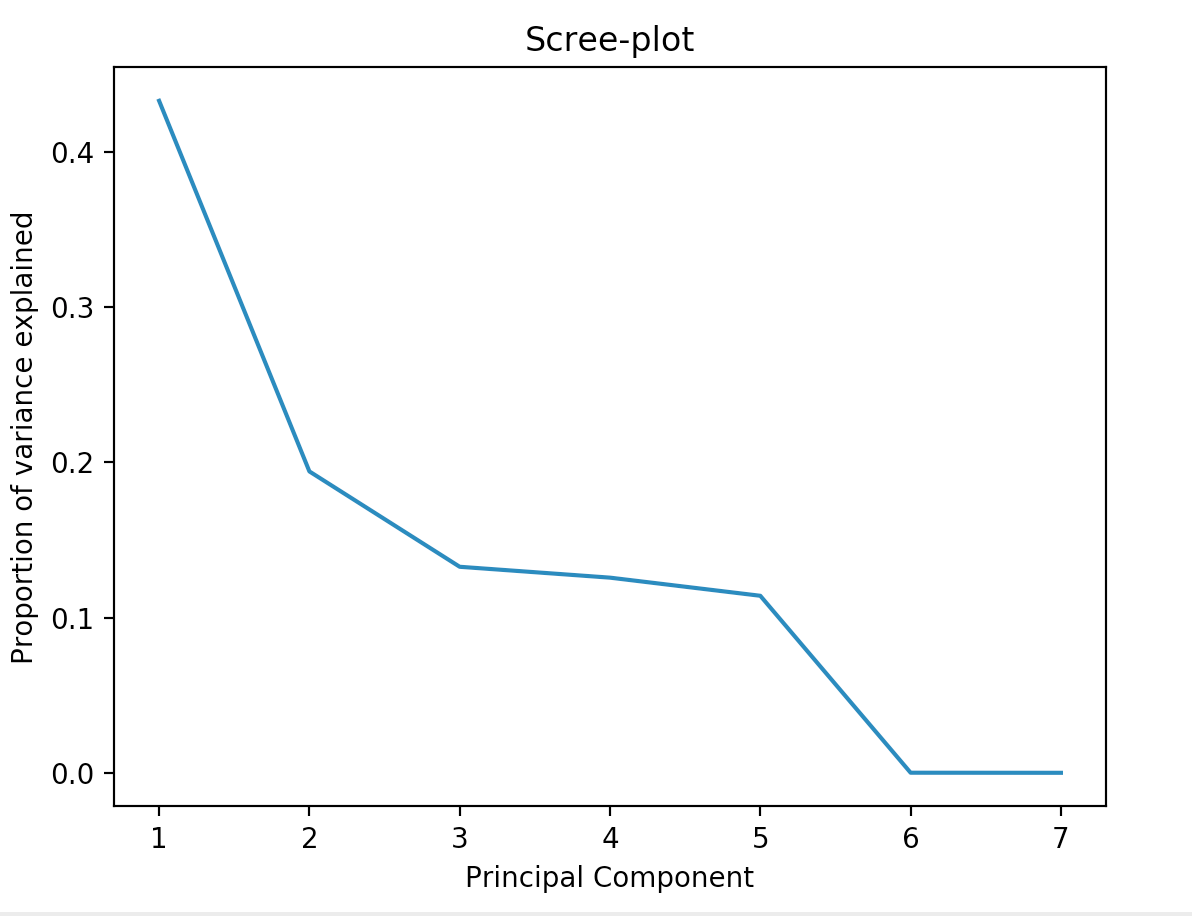


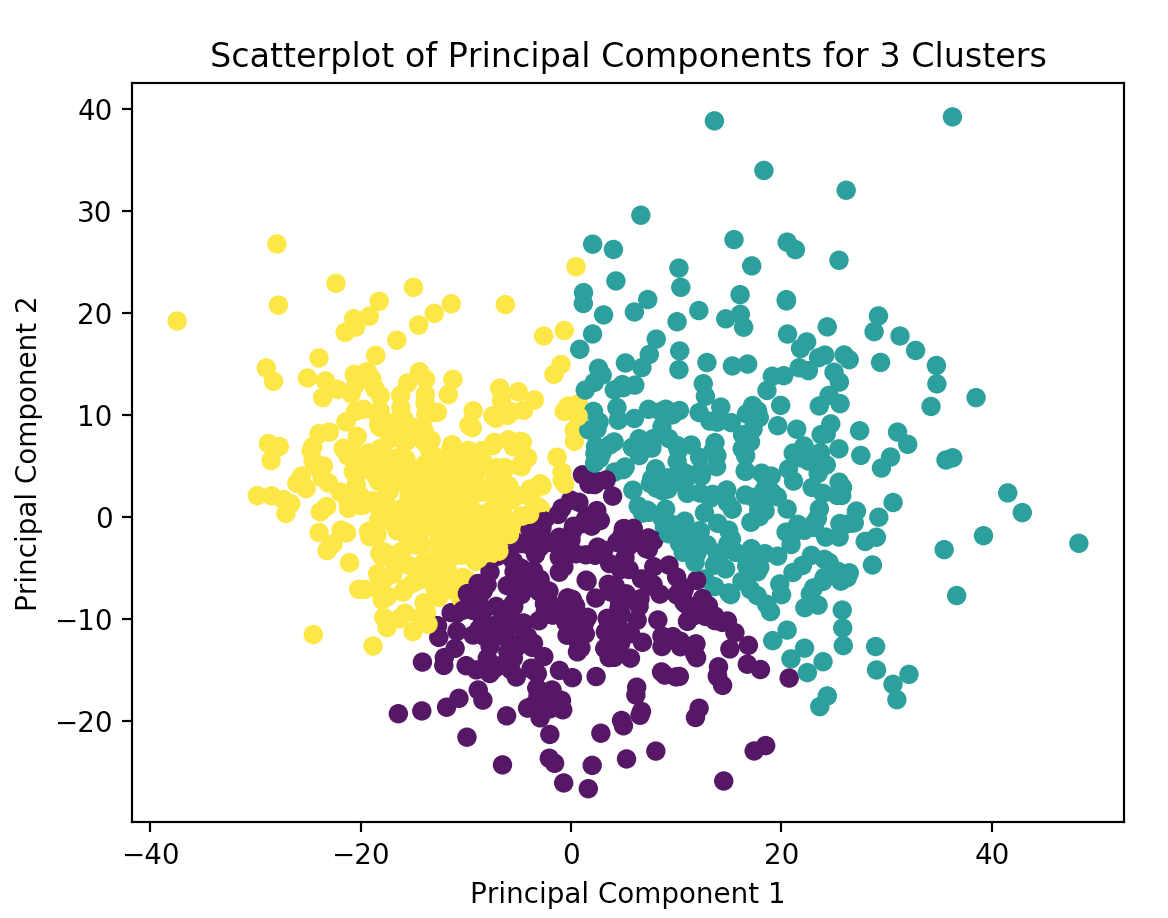
From the 3 clusters, the result shows the mean for cluster 2 for every test is the highest compares to cluster 1 and cluster 3 except for lab test 3, which is slightly lower than cluster 1. The average for the variable, part-time job, is 0 which means that every student in cluster 2 does not have a part-time job.

As a result, it can be concluded that the cluster analysis supports the hypothesis that having a part-time job has a negative impact on student performance.

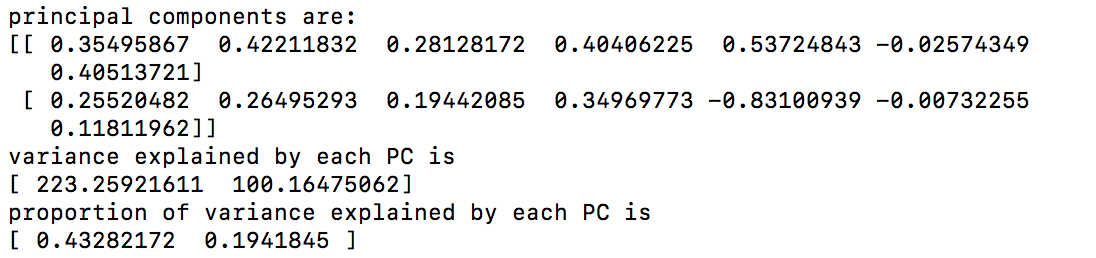
## Principal Component Analysis

The principal component analysis is an approach to reduce the dimensionality by transforming the original data set from linear correlation variables into non-linear correlation variables, known as principal components, that use new variable to represent the data in smaller dimension and keep most of the characteristics. From the scree-plot, it shows that the first principal component and second principal component explained more than 60% of the dataset so I have generated 2 principal components for the analysis. The result is shown below.





|  |  |  |
| --- | --- | --- |
| Variables | PC1 | PC2 |
| Lab 1 | 0.35 | 0.26 |
| Christmas Test | 0.43 | 0.26 |
| Lab 2 | 0.28 | 0.19 |
| Easter Test | 0.40 | 0.35 |
| Lab 3 | 0.54 | -0.83 |
| Part Time Job | -0.03 | -0.01 |
| Exam Grade | 0.41 | 0.12 |



# Conclusion

This report shows most of the analysis needed to carry out on the dataset for different purpose, such as to predict the final exam grade based on students’ previous performance. The best simple linear regression model to predict end-of-module exam grade is using Easter test result where the equation is y=0.53\*x+27.06. For multiple linear regression model, based on the p-value generated, a student with or without a part-time job should not be considered one of the factors that affect the performance on final exam grade. From the cluster analysis, it shows that student with a part-time job does have a negative impact on student's performance.

# References

Frost, J. (2013). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*. [online] Blog.minitab.com. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [Accessed 12 Dec. 2017].

onlinecourses. (n.d.). *http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit*. [online] Available at: https://onlinecourses.science.psu.edu/stat501/node/255 [Accessed 15 Dec. 2017].