Data Mining and Regression

Introduction

In a paper published in 1936, Alan Turing, who is considered to be the father of theoretical computer science and artificial intelligence, introduced the idea of a universal machine capable of performing computations like our modern-day computers. This idea at that time was considered bizarre and unlikely. However, within a century many computers were able to carry out such tasks. This was made possible by the abundant data currently present in our society. Therefore, data are considered to be one of the most valuable assets for humans in this digital age. Learning how to use and analyze them can be an important skill that can help solve many questions from different fields of study. For e.g.: In economics, data analysis can help us answer questions about optimal behavior, in physics data analysis can help in visualizing different physical phenomena like the recent picture of Blackhole that is hard to observe with our human eyes, etc. Utilizing the available data and using data analysis techniques can help us find trends and patterns hidden in datasets that might help some businesses understand problems facing their organization or create models to determine whether, natural disasters, etc. There is an entire field called Data Science that teaches people tools on how to work with data and analyze them. This field has been around for a long time and in recent years the tools to practice Data Science has gone through a major boost because of all the technological advancement. In this project, I have attempted to learn about one of the subsets of Data Science called Data Mining. Data Mining is the process of discovering different structures and patterns in large and complex data sets (Hand and Adams, 2014). This process can be used for any kind of data as long as the data is meaningful for its target application. There are different techniques associated with Data Mining. Some of the popular techniques are Classification, Clustering, Regression, Prediction, Sequential Pattern, etc. (Han and Kamber, 2011). Therefore, in this final project, I had a chance to dig deep into two of the techniques, Regression, and Classification, and apply them to different datasets.

Note: I have used response variable and dependent variable interchangeably as they mean the same thing. I have also used independent variable and explanatory variable interchangeably for the same reason.

Methods

I used three different datasets for this project: iris (dataset within R), penguins, and Happiness Score (from Kaggle). Iris dataset was used for learning the data mining techniques: Regression and Classification, whereas the other two datasets were used to apply those newly learned techniques. Penguin dataset contained penguins categorized into three different species of penguins based on their geographical and physical features such as location, Culmen length, Culmen depth, Flipper length, Body mass, sex, etc. The Happiness Score dataset contained the rank of various countries based on its happiness score which depended upon different factors such as levels of GDP, life expectancy, corruption, etc. The penguin dataset was used for classifying different categories whereas the Happiness Score dataset was used for analyzing the relationship between the dependent variable (Happiness Score) with other independent variables. For these analyses, I used two different data mining techniques which are given as:

Regression

It is a statistical technique used to determine the relationship between one variable (dependent) with other variables (independent). There usually exists a functional relationship between two or more variables in most physical processes which might be too complicated for us to grasp (Zou and Silverman, 2003). Some of these relationships can be approximated by some simple mathematical functions like polynomial or linear which helps us learn more about the underlying effects produced by changing different variables in the physical processes (Draper and Smith, 1998). Regression helps us find these relationships between variables. Furthermore, they can also help us predict how those variables might change in the future by investigating the relationship. There are mainly three types of regression: Linear, Multiple Linear and Non-Linear. In this project, I learned about Linear Regression and used it for analyzing the Happiness Score dataset.

Linear Regression: It is a model of regression that describes the relationship between two more variables using straight line. There are two types of Linear Regression.

a) Simple Linear Regression: It is a model of regression that describes the relationship between one independent variable and one dependent variable using a straight line. The straight line can be represented by this main equation

$$y = mx + b$$

where y is the response variable, x is the explanatory variable, b is the intercept and m is the slope of the line. To find this equation of the straight line, we need to calculate the mean and standard deviation (SD) of both dependent and independent variable and their correlation (r) because the slope of the line and the y-intercept are calculated as:

$$Slope(m) = r * \frac{SD\ of\ dependent\ variable}{SD\ of\ independent\ variable}$$

$$Y - intercept(b) = y - mx$$

Solving the slope and y-intercept using the data from our Happiness Score dataset and plugging the value in our main equation gives us the equation of the straight line or a line of best fit that represents our dataset. Calculating a straight line using the above formulas can also be called as a line of least square regression which means the slope of the regression line predicts the change independent variable when the dependent variable changes by one unit. In R, all of this calculation is done by the linear model function lm.

In the Happiness Score dataset, Simple Linear Regression is used to show the relationship between happiness score of countries and other variables like country's GDP, Life Expectancy and Corruption. Since the Simple Linear Regression can only describe relationship between two variables, *Fig 1*, *Fig 2* and *Fig 5* show the relationship between the happiness score and other variables separately. For this process, the dataset was downloaded from Kaggle. The downloaded dataset was cleaned, and the response variable was checked for its normality. After checking its normality, regression was done with the help of different in-built R functions.

b) Multiple Linear Regression: It is a model of regression that describes the relationship between three or more variables using a straight line. In this regression, one dependent variable is modeled as a function with several independent variables with their corresponding coefficient. The equation takes the form of:

$$y = m_1x_1 + m_2x_2 + \cdots + m_nx_n + b$$

where m_1, m_2…m_n are the regression coefficient(slopes) and x_1,…x_n are the explanatory variables and b is the intercept.

The process of carrying out multiple linear regression was similar to Simple Linear Regression. The only difference between the two types of Linear regression was that we used multiple independent variables for multiple linear regression and just one independent variable for simple linear regression.

Classification

Classification is a process of finding a model that describes and differentiates different categorical class labels. It determines where a certain class/observation belongs by training the program and creating a model using a portion of the dataset with known categories. The training model is then used to classify the other portion of the dataset or other similar observations. One of the examples of a classification problem is classifying different types of plants with the help of other variables such as leaves length, color, plant's length, etc. For this project, I used the penguins' dataset to perform the classification. The dataset was downloaded from Kaggle and it contained around 344 penguins divided into three different species based on their different physical and geographical features such as their location, body mass, culmen length, sex, etc. For this project, I decided to classify the penguins based only on their culmen length, culmen depth, flipper length, and body mass which were my independent variables. I only chose four physical features/independent variables because they were numeric while the rest were non-numeric, so because of the project's time limit, I thought dealing with numeric variables would be easier than non-numeric.

In the classification process, the penguin dataset was separated into two datasets: training data and testing data. 70% of the penguin dataset was used as training data and 30% was used as testing data. 70% of data was used by the computer to learn the connection between the independent/explanatory variables with the dependent/response variables. Training data was used by the program to learn the classification and create a model that was tested on the testing data. For this classification, I used two different techniques: Decision Tree and Random Forest.

1) Decision Tree: It is a type of classifier that maps the possible outcomes of a series of related choices. The map contains a root node, internal nodes, leaf nodes, and branches. The root node is at the top of the map as shown in *Fig 4*. This node is used for dividing the whole data into two sets by applying certain conditions related to the independent variable. For eg: If the Flipper length < 207, the data belongs to one category, and if it's not then it belongs to a different category. These categories form internal nodes. The internal nodes then apply other conditions related to other variables to separate the dataset into different leaf nodes. To determine which conditions to use, different algorithms such as GINI, CRAT, Chi-Squared, etc. are used. These algorithms check the impurity of different conditions. The impurity refers to the chance of being incorrect if you randomly assign a category/label to an observation/example in the data set. It is calculated by subtracting the total probability, usually 1, to the probability of correctly categorizing an observation using the given interval. The condition with the lowest impurity creates the interval node and the other condition are tested again for their impurity for the next internal nodes. In this way, by checking the impurity of different conditions and splitting data into different categories, a decision tree is formed. This is all done using a function rpart() in R.

   In this project, the boxes (nodes) consist of different species name of penguins along with their probability of occurrence and the total percentage of penguins classified as that species. The lines (*Fig 4*) represent branches which consist of different conditions that help in the categorization of different species

2) Random Forest: It is a similar, but improved classification technique as the decision tree. In this classifier numerous decision trees are created. This is done by creating numerous bootstrap datasets. A bootstrap dataset contains data created by using the provided training data and randomizing or repeating some of the rows in that data. Each bootstrap datasets forms its own decision tree by following the procedure described in the decision tree section. All the decision trees predict a categorization when data is supplied, and the categorization predicted by the higher number of the decision tree is supplied as the final

categorization. However, while creating bootstrap datasets some of the data from our original training data won't be utilized because of the repetition of data in those datasets. These missing data are put together to create a new dataset called Out-Of-Bag (OOB) dataset. This OOB dataset is then tested using the decision trees made earlier and categorized in the same way as other training data. All of this process is done by using the randomForest() function from R.

After applying the Random Forest model, a graph (*Fig 4*) was created that represents how the program improved over time by showing the error rate vs the number of decision trees. Also, *Table 2* and *Table 3* shows us how the algorithm performed in categorizing different dataset using training data and testing data respectively.

Results/Discussion

For Happiness Score dataset (Regression)

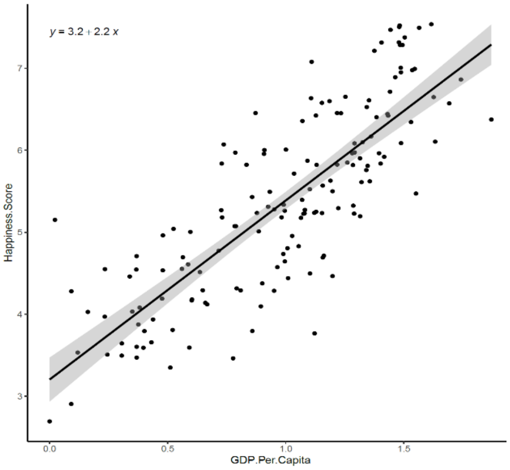

Fig 1: GDP Per Capita vs Happiness Score

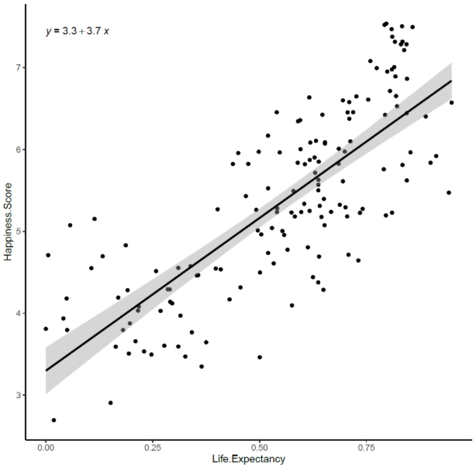

Fig 2: Life Expectancy vs Happiness Score

Fig 1 and Fig 2 were created using Happiness Score dataset. The graphs were formed after performing linear regression on the Happiness Score (dependent variable) and other variables. Fig 1 shows the relationship between two variables: Happiness Score and GDP Per Capita. The line of best fit has the equation $y = 3.2 + 2.2x$ where 2.2 is the slope and 3.2 is the intercept of that line. From the line of the best fit, we can infer that increasing GDP increases the Happiness Score. Similarly, for figure 2, the line of best fit has the equation $y = 3.3 + 3.7x$ which means increasing Life Expectancy also increases the Happiness Score. However, the slopes of the two equations are different, therefore comparing the two slopes we can infer that increasing the Life Expectancy increases the Happiness Score more than the GDP as the slope for line of best fit in figure 2 is higher.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.990451	0.130162	22.97484	3.80E-51
GDP.Per.Capita	1.241823	0.221566	5.604767	9.64E-08
Life.Expectancy	1.645305	0.38357	4.289451	3.18E-05
Corruption	1.897321	0.513723	3.693275	0.000309

Table 1: The multiple regression coefficient table

Table 1 was created using the Happiness Score dataset. The t-value evaluates if there is any significant connection between the independent and the dependent variable. It measures how many standard deviations our coefficient is far away from 0. We want the t value to be far away from 0 (higher or lower) to reject the null hypothesis and claim there exists some relation between dependent and independent variables. Since the GDP.Per.Capita has the highest t-value which shows that GDP.Per.Capita is more significant to Happiness Score than any other independent variable. The Estimate row in Table 1 is the respective coefficient of the variables. Since the p-value, in the 5th column has a value less than 0.05 (significance level), it indicates that the relationship between the dependent and independent variable is not by chance. Also, our R2 found to be 0.7125. Since it measures the linear relationship between our independent variable and our dependent variable, this means that almost 71% of the variance found in the dependent variable can be explained by the independent variable. Finally the equation for our multiple regression becomes:

$$y = 1.24x_1 + 1.64x_2 + 1.9x_n + 2.99$$

Where x1, x2, and x3 are GDP.Per.Capita, Life.Expectancy and Corruption respectively.

For Penguins dataset (Decision Tress and Random Forest)

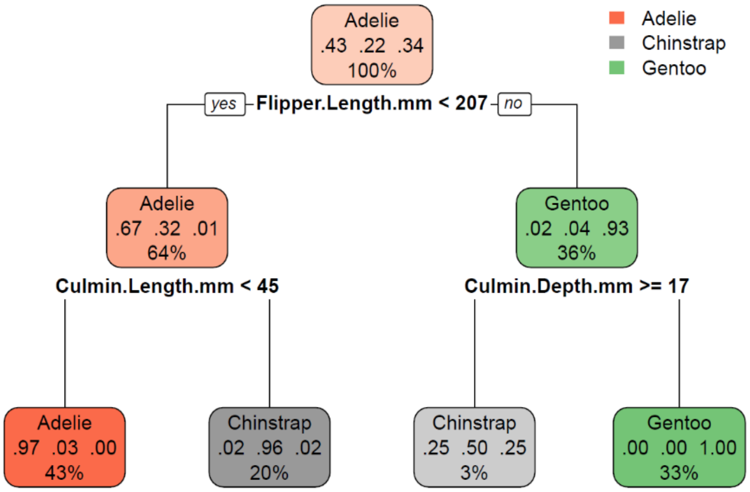

Fig 3: Decision tree to categorize different species of penguins

Fig 3 represents the model created by the program using the training data. Different nodes (boxes) represent different penguin species. Firstly, the whole training dataset was categorized into three different species and 43 percent of them were Adelie, 22% were Chinstrap and 34% of them were Gentoo. This is shown in the root node (top box). For the right-hand side of the decision tree, the classification starts with categorizing the penguins with flipper length. The penguins with flipper length greater than 207 mm had the probability of 0.93 to be from a species Gentoo whereas the probability of the same penguin to be from the species Chinstrap or Adelie is 0.04 or 0.02 respectively. Now, a different variable(Culmen Depth) is used and the penguins with Culmen depth more than or equal to 17 had a probability of 0.25 to be Adelie, 0.50 to be Chinstrap and 0.25 to be Gentoo whereas if the penguins has Culmen depth less than or equal to 17 it has the probability of 1 to be Gentoo. In this way, the whole decision tree is formed in R. The boxes in the figure contain the species name, their probability, and the percentage of the dataset that were classified as the species from that box. The decision tree also shows the most important variable. In this case, Culmen length, Culmen depth, and Flipper length are important because they are used to form the decision tree whereas other variables like Body Mass that was present in the dataset was not an important variable as it was not used for categorization.

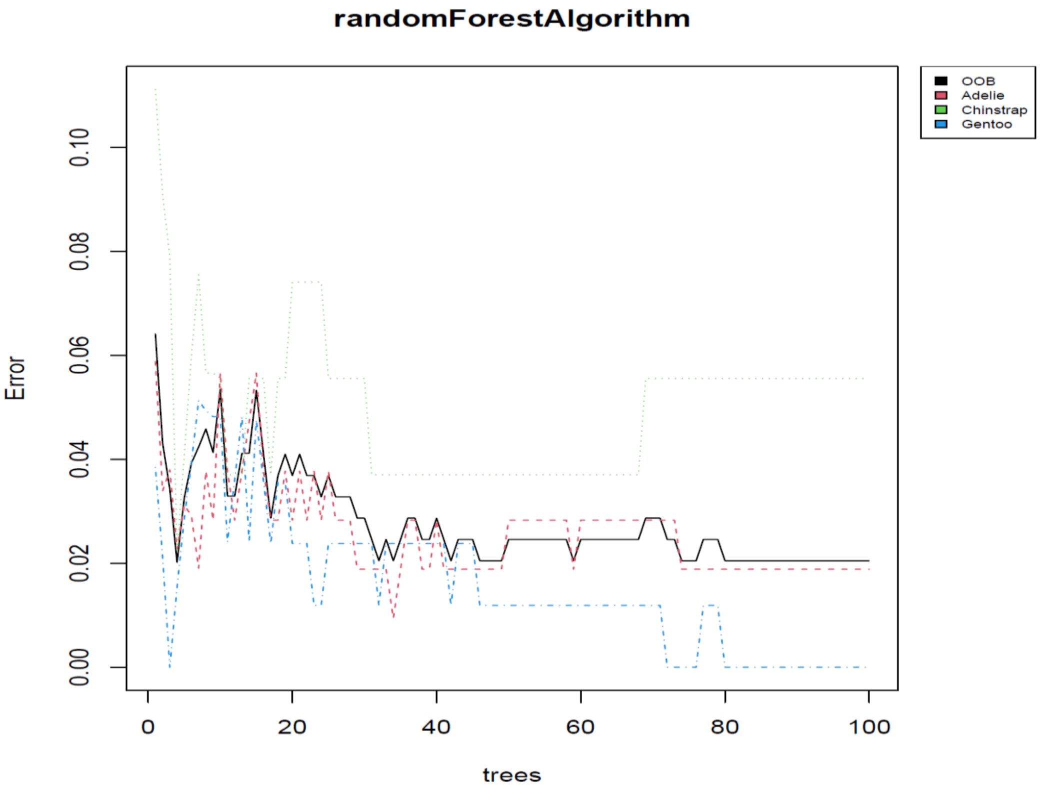

Fig 4: Error vs random trees graph for my random forest

Fig 4 shows how the numerous decision trees formed for random forest classification performed in terms of categorization of different penguin species. The x-axis represents the total number of decision trees formed (i.e 100) in this classification. The y-axis represents the error made by each decision tree. Different colors represent different species as shown in the index at the top right corner. The black line in the graph shows the error made by 100 decision trees while classifying the Out-Of-Bag (OOB) dataset. It is clear from the graph that as the number of decision trees increased the error rate decreased which as well is considered learning in the computer language. Also, the decision trees were more efficient in predicting the Gentoo species than other species.

	Adelie	Chinstrap	Gentoo	Class Error
Adelie	104	2	0	0.018868
Chinstrap	3	51	0	0.055556
Gentoo	0	0	84	0

Table 2: Result of using random forest algorithm in training data.

	Adelie	Chinstrap	Gentoo
Adelie	43	1	0
Chinstrap	2	13	0
Gentoo	0	0	39

Table 3: Result of using random forest Algorithm in the testing data

Table 2 and 3 show the result of the prediction of random forest classifier in training data and testing data respectively. For training data as shown in table 2, 104 Adelie were correctly categorized whereas 3 Chinstrap were wrongly categorized as Adelie. 51 Chinstrap were correctly categorized whereas 2 Adelie were wrongly categorized as Chinstrap. All the individuals in Gentoo were correctly classified. The class error represents the error made by the algorithm to correctly categorize different species. Similarly, for table 3, 43 Adelie were correctly categorized, 2 Chinstrap were wrongly categorized as Adelie, and so on. The algorithm performed well with the testing data.

Conclusion

The main aim of this project was to learn new techniques such as Regression and Classification surrounding Data Mining and apply those techniques in real datasets. After completing this project, I feel like I have achieved that aim as I have a good understanding of the underlying concept regarding Regression, Decision Tree, Random Forest which I have described in the methods section. Using the decision tree and random forest I was able to train my algorithm to predict different species of penguins using certain variables. For the happiness score dataset, I was able to use simple linear regression and multiple linear regression to find the relationship between different variables in my dataset and get the equation for the straight line representing the regression.

I will use these newly learned skills in different machine learning projects in the future. As linear regression is fundamental to a lot of machine learning algorithms, I think this project brought me one step closer to my aim of having a good grasp of the machine learning algorithms.

Reflecting on the Review

The Review was very helpful as my reviewer pointed out many errors made in this research paper. Almost all the comments were helpful, and I have made changes accordingly. Some of the important errors pointed out to me were writing introduction of data mining in the method section, missing out the equation for multiple linear regression, unclear decision tree figure, clearly explain some of the methods, explain data set clearly, improve the clarity on figure 9 and 10 and use more statistical approach to show results. I added the introduction of data mining to the Introduction section of this paper. I added the multiple regression equation and corrected the image for the figure of the decision tree. I have also added some more descriptions on how the Decision tree and Random forest work and expanded more on multiple linear regression. I moved Figures 9 and 10 to the appendix in order to explain the result of multiple linear regression statistically. After adding the statistical explanation, I felt more confident about my paper because statistical techniques were missing from my paper. One positive and helpful comment provided by my reviewer was that the equation in my simple linear regression was very helpful to understand the regression method. This encouraged me to add some more equations to my regression section for more clarity. I was also pointed out that I didn’t explain the Class Error in my caption for table 2. It was a great find, however, I had explained what it was in the paragraph describing the two tables, so I didn’t think it was necessary. Nevertheless, it was helpful to read all my other captions and make changes if necessary.

Appendix

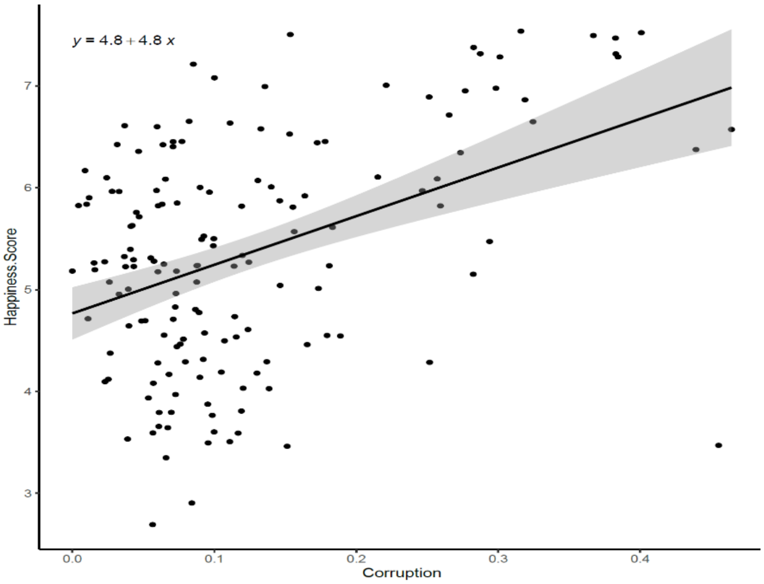

Figure 5: Country's Happiness Score vs Country's Corruption. $y = 4.8 + 4.8x$ *is the equation of line of best fit*

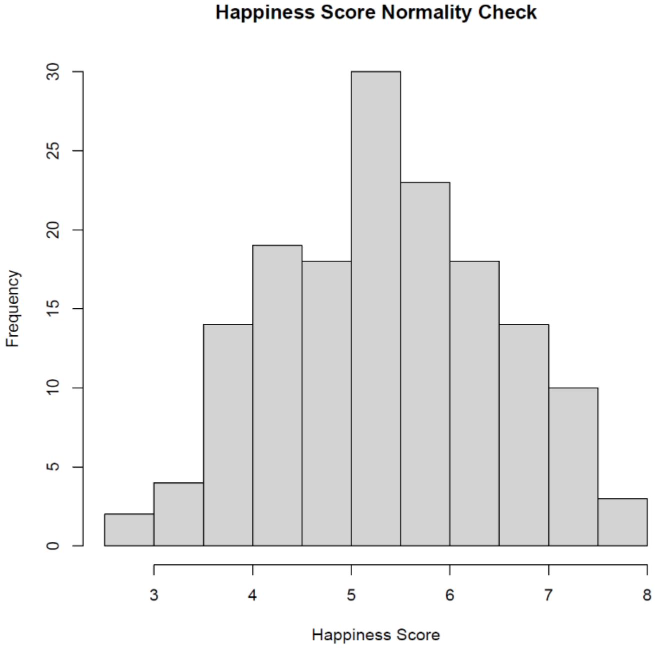

Figure 6: Checking the normality of the happiness score of countries. It is normally disributed.

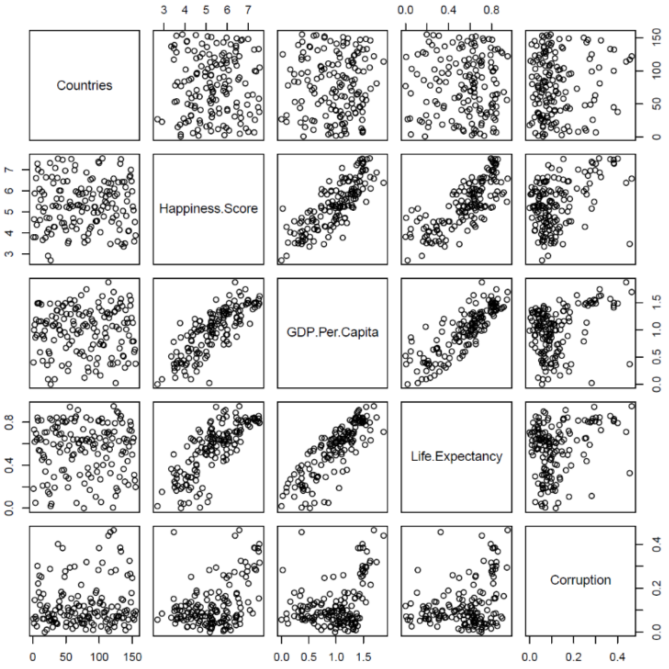

Figure 7: Plotting the multiple regression plot containing all the variables from the dataset. This graph shows the relation between different variables. For eg: The box in the 2nd row 3rd column graph shows the relation between happiness score and GDP. That graph is similar to the graph from Fig 1 that looked at the relationship between those two variables. Similarly, most of the boxes also represents relationships.

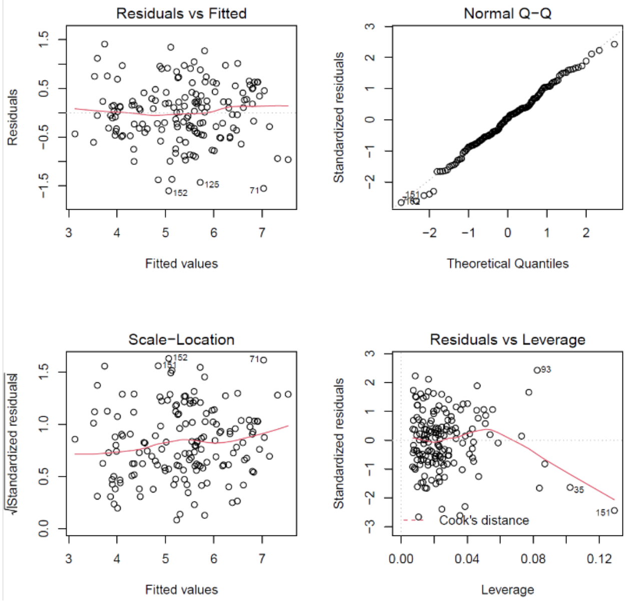

Figure 8:The top left is the graph of residual of the linear model vs its fitted values, top right graph is the graph of quantile distribution of the residuals, bottom left graph is the graph of standardized residuals vs its fitted values and the bottom right graph shows us the Cook's distance of the residual.

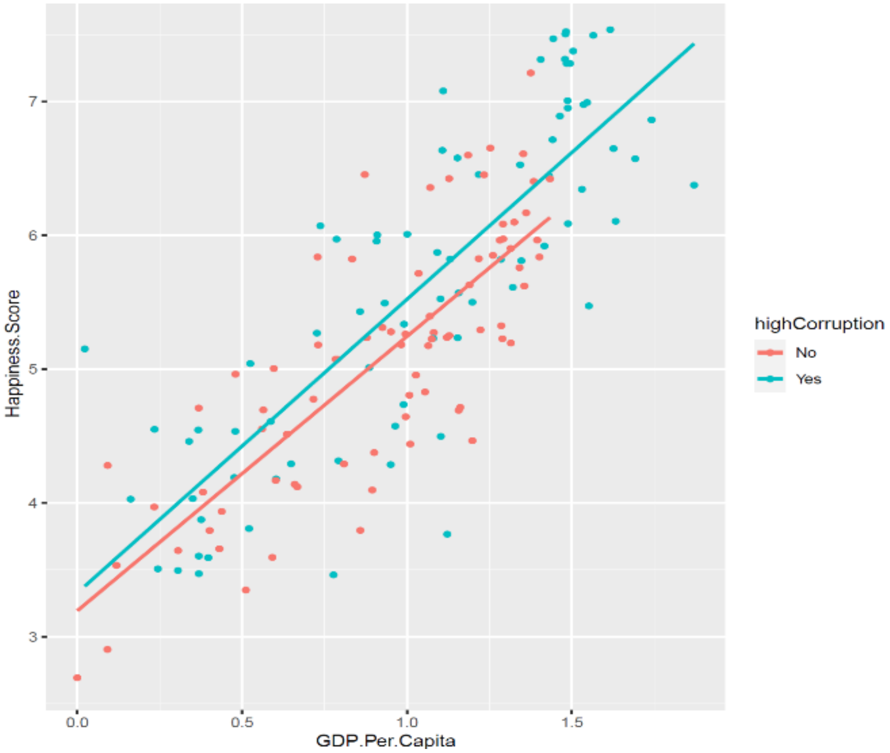

Fig 9: Graph comparing the GDP with Happiness Score along with the rate of corruption

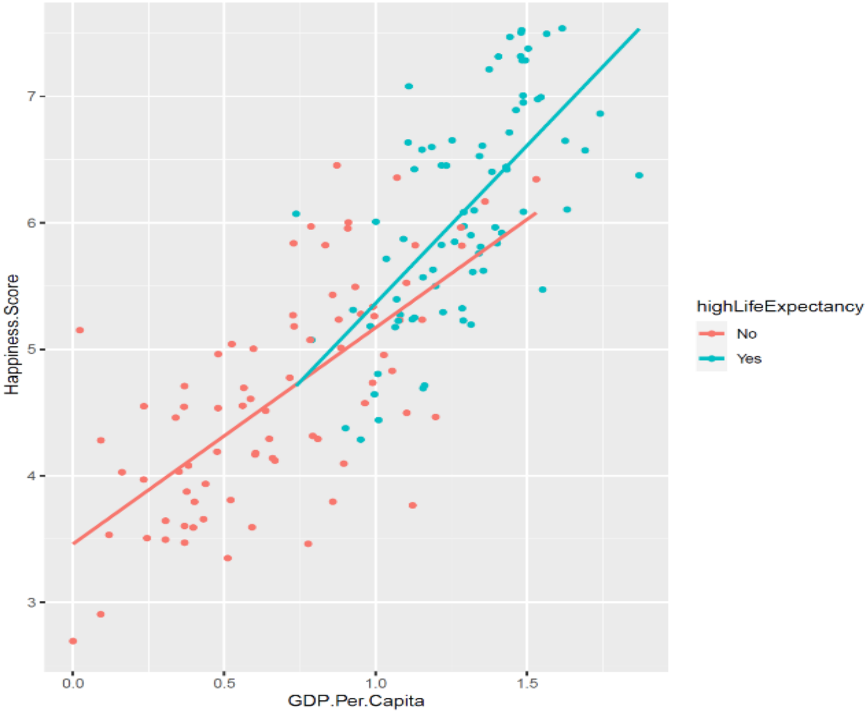

Fig 10: Graph comparing the GDP with Happiness Score along with the higher and lower Life Expectancy

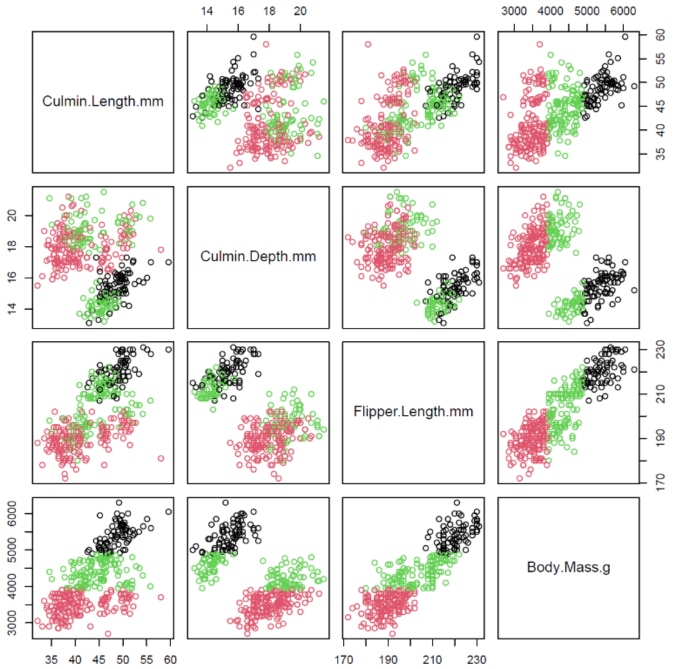

Fig 11: This graph shows us the clustering of the penguin dataset using different variables

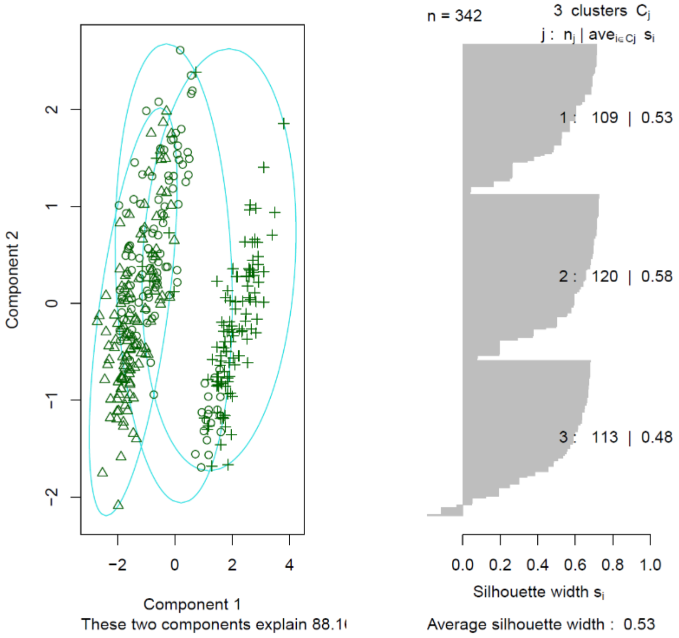

Fig 12: Figure showing the clustering of three different penguin species.

Reference

Hand, D. J., & Adams, N. M. (2014). Data Mining. *Wiley StatsRef: Statistics Reference Online*, 1-7.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, *227*(3), 617-628.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.