**Probability Theory and Statistics**

Logo

Description automatically generated

ALY6010, WINTER 2022

**Week -5**

Module-5- R Practice Output

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Prof. Tom Beur

Date: 03/27/2022

**Introduction:**

Purchasing a fuel-efficient vehicle is constantly on the agenda; people like to purchase passenger vehicles to extend the range of their fuel tanks. Diesel passenger vehicle engines have a higher efficiency of 41%, while the average is closer to 30%. Petrol engines, on the other hand, have an energy efficiency of around 37%, although the average performance in a petrol automobile is closer to 20%. Therefore, diesel automobiles are in more demand than gasoline ones. According to the survey, turbodiesel provides a 25% increase in miles per gallon.

We used the diesel variant in this dataset to do correlation and regression analysis. There are 123 rows and 8 columns in the dataset, while i. Make denotes the car's manufacturer, model denotes the brand's variations, and type denotes the car's kind, such as passenger, SUV, light commercial, and so on. Cyl refers to the cylinders of a car's engine in order to assess its performance.Engine L denotes the engine's capacity, whereas the fuel tank denotes the size of the fuel tank. The fuel consumption per 100 kilometers is measured in liters per 100 kilometers, while the distance covered by the automobile is measured in kilometers.

**Task 1: Importing dataset:** In this task imported the dataset by using read.csv() in R while setting it as a working directory, it consists of the data about the performance if the diesel cars in Australia

Text

Description automatically generated with low confidence

**Table

Description automatically generated with medium confidence**

**Scatterplot Matrix:**

Plot to compare the variables by analyzing the correlation between Range and other variables.

Diagram, engineering drawing

Description automatically generated

**Task 2: Str()** In this task provide the structure of the data in the given dataset.

**Text

Description automatically generated**

**Task 3: Descriptive analysis:** In this task, it describes the descriptive statistics whereas, mean, median, standard deviation, minimum and maximum, range of the variables.

**Calendar

Description automatically generated with medium confidence**

**Task 4: Understanding Dataset:** In this task understanding the dataset by applying few functions in R

**A screenshot of a computer

Description automatically generated with low confidence**

**Table

Description automatically generated**

**Data Cleaning:**

In this task before implementing the regression model data need to be clean and executable which will help in getting a better understanding of the data by analyzing the datatype and other measures. Cleaning the data is the most important part which is also known as pre-processing of the data into readable format.

**Task 5: Converting datatype from int to numeric**

**1:** Range variable conversion from character datatype to numeric

**Table

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Preprocessing of data for correlation**

In this, we have preprocessed the data by converting datatype and dropin few columns as per the requirement of the correlation matrix.

**Data after converting datatype**

**A screenshot of a computer

Description automatically generated**

**A picture containing calendar

Description automatically generated**

**Table

Description automatically generated**

**Task6: Correlation Matrix, cor():** In this exercise, create a correlation matrix that shows the correlation coefficients between the variables in the diesel car performance dataset.

Positive numbers like 1 indicate a linear correlation between two variables, 0 shows no correlation, and -1 indicates a negative correlation between the variables in the supplied correlation table.

The below matrix shows the correlation among the two variables.

Diagonally, 1 indicates the correlation between two variables present in the matrix. Likewise, -0.11 indicates the correlation between Engine.L and Range km, which reflects that they are weakly negatively correlated. Increasing the capacity of the engine will negatively affect the performance of the car.

**Table

Description automatically generated with medium confidence**

In the below table, The correlation between the consumption and fuel tank per liter are indicating a positive result and is strongly correlated, because more consumption per liter depends on the fuel tank capacity.

**Table

Description automatically generated with medium confidence**

**Task7: rcorr() with p-value:** In this task below table show the number of observation in the correlation test which is n= 123. Where, as p values indicates the 63% percentage strong chances of accepting the alternative hypothesis in correlation range km and cyl. On the other hand, 24% low chances of acceptance the alternative hypothesis.

**A picture containing text, receipt

Description automatically generated**

**Task8: Histogram of the correlation matrix:** In this task representation of the histogram with the scatter plot which indicates the line and provided the correlation between each variable.

Chart, scatter chart

Description automatically generated

**Task9: Heat map for the correlation matrix:** In this task created the heat map which indicates the correlation with color and the performance for the fuel efficiency of the car.

Chart

Description automatically generated

**Task10: Correlation Models:** In this task, Pearson correlation two continuous variables indicates the linear relationship which is 4%, whereas Spearman indicates the monotonic relationship between two variables which is 9%. However, we look at the Kendall model which is also positive monotonic with 7%.

**Graphical user interface

Description automatically generated with low confidence**

**Task11:Regression matrix:**

In this task build the regression model between the dependent variable which is “Range.km” and the independent variable which are “FUEL.TANK.L”, “CYL”, “ENGINE.L”, “CONS..L.100km” which signifies the R square is 0.96 and p-value is 0.00 which indicates there is no effect of the null hypothesis as the coefficients are equal to zero.

**Table

Description automatically generated**Linear Regression indicates the Model Fit where the R square is 96% which indicates variance between independent variables and dependent variables, the higher the variance shows the higher effectiveness in predicting the model. However, adjusted R squared helps in predicting the variety of numbers to compare the descriptive power of the regression model.

**Graphical user interface, text, application, email

Description automatically generated**

**Summary:**

In conclusion, the key takeaway from this R practice is understanding about the correlation matrix and the regression matrix by implementing the relationship between independent and dependent variables.

**Reference:**

**[1]Exploring, cleaning, and analysing data in R - YouTube**

A replay of a non-technical livestream that walked through how to explore, clean, and analyse data in R, using the 'starwars' dataset that is built into the ...

[**https://www.youtube.com/watch?v=Ap1Q2fkqO\_I**](https://www.youtube.com/watch?v=Ap1Q2fkqO_I)

**[2]Correlations and Covariance in R with Example - YouTube**

Correlations and Covariance in R with Example: Learn how to calculate Pearson's correlation, Spearman's rank correlation, Kendall's rank correlation, and cov...

[**https://www.youtube.com/watch?v=XaNKst8ODEQ**](https://www.youtube.com/watch?v=XaNKst8ODEQ)

**[3]Correlation Matrix in R - YouTube**

This tutorial demonstrates how to create a correlation matrix in R. In addition, I show how to write (i.e., export) and APA style table containing a correlat...

[**https://www.youtube.com/watch?v=MfN8KGQU3MI**](https://www.youtube.com/watch?v=MfN8KGQU3MI)

**[4]Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software**

[**http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software**](http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software)

**[5]Linear Regression from Scratch in R**

[**https://datascienceplus.com/linear-regression-from-scratch-in-r/**](https://datascienceplus.com/linear-regression-from-scratch-in-r/)

**[6]Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software**

[**http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software**](http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software)

**[7]Deleting all apostrophes in a data frame in R**

Christopher YeeChristopher Yee 51522 gold badges55 silver badges1212 bronze badges et al.

[**https://stackoverflow.com/questions/35901778/deleting-all-apostrophes-in-a-data-frame-in-r**](https://stackoverflow.com/questions/35901778/deleting-all-apostrophes-in-a-data-frame-in-r)

[8][**https://cran.r-project.org/web/packages/jtools/vignettes/summ.html**](https://cran.r-project.org/web/packages/jtools/vignettes/summ.html)

**[9]Car Driving Distance Range Dataset**Mahimkar

[**https://www.kaggle.com/datasets/adityamahimkar/car-driving-distance-range-dataset**](https://www.kaggle.com/datasets/adityamahimkar/car-driving-distance-range-dataset)

**Appendix:**

Text

Description automatically generated

Text

Description automatically generated