Yiling Chen

Hw#1 Write-up

Feb. 17,2021

**Question 1**

The goal for this problem is to analyze the College dataset base on given instructions. This dataset contains 777 observations and 13 variables. 12 of the 13 variables are continuous variable that show detail information of each observation university along with one other categorical variable that determines whether the observation university is private or not.

Graphical user interface, text, application

Description automatically generated

My first step is to set the work directory, load the College dataset from ISLR, and preprocess the dataset by looking at the summary and produce a scatterplot using only continuous variables, which are all variables except the first column.

A close - up of a document

Description automatically generated with low confidence

From the summary function, I can get a brief understanding of the dataset. I am able to know how many schools are private, the minimum, maximum, mean and median for the number of acceptances, the graduation rate, etc. Then I use the pairs function to produce scatter plots that shows the relationship between each variable.

A picture containing grate

Description automatically generated

My next step is to create a new column called “Elite”, which include only if the university accepted the top 10 percent students are greater than 50%. Calculate the number of Elite schools as well as to see how many elite schools are private.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Hence, I first created if the top10perc is greater than 50, then insert yes in the Elite variable, vice versa for less than 50. Next, I use the table function to look at how many colleges are elite. As well as produce another table to compare how many colleges are elite and private. The result I got back is there are 78 elite colleges and out of those 78 elite colleges, 67 are private college.

Finally, I want to know whether if those colleges that are elite result in a higher graduation rate compare to those colleges that are not elite. I created two subset groups which differentiate the elite colleges and non-elite colleges. Then I calculate the mean graduation rate for each of these two subgroups. The result shows that Elite colleges has an average rate of 83.38% graduation rate whereas non-elite colleges have a 63.46% graduation rate.

Graphical user interface, text, application

Description automatically generated

In conclusion, throughout looking and analyzing the College dataset, I can conclude that when more than 50% of the top 10 percent students (in their high school) enroll in those colleges, the likelihood of graduation is higher than those that are not top 10percent students (in their high school. Also, the colleges that has a higher graduation rate are most likely to be a private school. Even though, conclusion can successfully be drawn but the reason for these conclusions can be further analyzed. For example, the reason of higher graduation rate in private and elite colleges could be the student skill, or small classroom settings help students grade in multiple factors (which the dataset doesn’t have data on). It would be interesting if we can get those data of different colleges see if those factors are affecting the student’s graduation rate.

**Question 2**

Graphical user interface, text, application

Description automatically generated

The goal for this problem is to analyze the Auto dataset in ISLR base on given instructions. This dataset contains 392 observations and 9 variables including gas mileage, horsepower and other information. The origin variable contains three different numeric categories (1 American, 2 European, 3 Japanese).

My first step is to check if there is any missing value, the result is no missing value contain in the dataset. I then check how many variables are numerical and how many are factors. The result I got back is only the “name” variable is factor and the rest is numeric. Next, I use the summary function to get a brief understanding of the Auto dataset.

A close - up of a document

Description automatically generated with low confidence

Next, after defining a new dataset as continuous, which contains all the variables except for the name variable. I use the sapply function to define that I want to see the mean and standard deviation of all the variables.

A picture containing text

Description automatically generated

Then, base on the instruction, I remove the 5th through 55th observation and use the sapply function again to get the range, mean, and standard deviation of all the continuous variables.

Table

Description automatically generated with medium confidence

2e)In the full Auto dataset, I would change the origin variable from numerical into their original name. Using American, European, or Japanese instead of “1,2,3” helps us to identify the identity of this variable because R automatically consider “1,2,3” as numerical variable but the origin is actually a categorical variable.

Table

Description automatically generated

Next, I use scatter plot and corrplot to explore the relationships between the variables graphically. But I excluded the origin and name variable because those are not numerical. From the scatter plots we can see some trend, for example, in general, the smaller the horsepower, the smaller the weight of the vehicle is. However, we can see there is barely any trend in the year vs. acceleration.

Graphical user interface, text, application

Description automatically generated

Since the scatter plot is hard to tell the trends in detail, I then generated a corr plot as shown below. From here, we can see the relationships between different variables more closely. For example, there is a very high correlation between displacement and cylinders, displacement and weights, displacement and cylinders. Also, there is a very weak relationship between weight and mpg, displacement and mpg, etc.

Diagram

Description automatically generatedA picture containing table

Description automatically generated

Finally, I created a new category base on each car’s mpg on low, med, and high. After inserting this into a new variable and save as a new dataset, I exported the dataset into r.data file.

Graphical user interface, text, application, chat or text message

Description automatically generated

In conclusion, throughout analyzing the Auto dataset, I saw some interesting relationships between the variables. I saw that there are high correlation between variables like displacement and cylinders, displacement and weights, displacement and cylinders, as well as low correlation between variables like weight and mpg, displacement and mpg, year and acceleration. I have also change the origin from a numerical category values into word category so that R wouldn’t treat origin variable as a continuous variable anymore.