Shay Snyder

April 28, 2020

CSCI-4957 Data Analytics

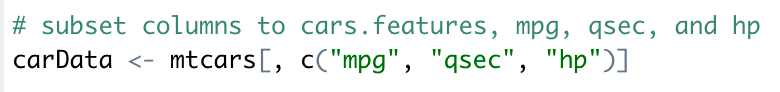
Husari

**Exercise 13: Outlier Detection**

*Task 1: Jay Leno Reincarnated*

As an expert in the antique car market, I need to determine which automobiles stand out among the rest. This knowledge enables me to better select my motor vehicle investments to get the most unique antique automobiles. Since I have been in the business for so long, I have decided to take a different approach to determine my next purchase. For this purchase, I will use the programming language R and the mtcars dataset which was extracted from the 1974 Motor Trend US magazine.

The 1974 Motor Trend dataset contains 32 observations ranging from your everyday Toyota Corolla to the Italian Ferrari. Since I am an elite antique collector, I am only interested in miles per gallon, quarter mile times, and horsepower. The data was subset to meet my desires



Now that the data is subset, I would like to determine which automobiles are the most unique (outliers) using my ‘spectacular’ data analytics skill. The scatter3d method in R was used to visualize the data. Let’s see if any standout…

A picture containing knife, table

Description automatically generated

A close up of a map

Description automatically generatedA picture containing umbrella, rain

Description automatically generated

As you can see from the plots (you can interact with them upon running my script), it is very difficult to visually identify any outliers. Let’s try some mathematical calculations. The mvoutlier package in R specified very unique cars from our data set. The results of the uni.plot() function are detailed below.

A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generated

From our results, we can see that four automobiles stood out from the others: Mercedes 230, Fiat 128, Maserati Boar, and the legendary Toyota Corolla. These are the four vehicles that I will acquire and collect in my iconic collection.

*Task 2: Wolf of Wall Street 2.0*

As a young data scientist at Amaya, I am tasked to predict whether various company obligators are default. The programming language R and the provided bankloan.csv data set are crucial to this process. The following columns will be used to determine if a loan is default: (x1) company age, (5) number of managers, (6) average manager age, (7) managers’ total stock, (11) the remaining months before borrowers to withdraw the obligations, (13) duration so far, and (14) duration between the last payment. The following boxplots represents the distribution of univariate outliers within each column.

A screenshot of a cell phone

Description automatically generated

In order to calculate which company obligators are default, I used the cluster and DMwR packages. I used the daisy function along with the outliers.ranking function to computer the top 15 accounts that are default. The image below contains the top 15 accounts that are default.

A screenshot of a cell phone

Description automatically generated