CIS 5270 Project - 2 Analysis Based On Used Car Database Using RStudio

# Analysis Based On Used Car Database

## Using Rstudio

**Abhishek Shah CIN : 305084941**

**Jalashri Raval CIN : 305874197**

****

**A) Dataset URL's:**

# Dataset: <https://www.kaggle.com/orgesleka/used-cars-database>

This URL's contain the entire data regarding the used cars over 370000 used cars scraped with Scrapy from Ebay-Kleinanzeigen. As automobile sector is one of the fastest growing sectors at present, especially in United States. There is also a parallel sector growing along with the new cars market that is "Used Car Market". We are curious to study and analyse this huge dataset to understand the buyers' point of view, preferences and their ability to spend for the used cars and so as the automobile companies whose cars are more visible in the used cars market. Also, we are intended to study the price point level which varies on brand to brand. The dataset has total 20 columns and contains the information like, vehicle type, model, price, gearbox type, brand, kilometer travelled etc. We can even see the date and time the data was crawled. The study will give us the in detail idea about which are the factors that consumers pay major attention while buying a used car and vice versa. It is important to study this market as there are more and more companies penetrating into automobile sectors who are introducing different brands cars frequently. Now to study the impact on the market of Used cars due to launch of new cars, the target customers and the game of the price point are various crucial deciding factors.

The reasons behind why the new market has developed for the used cars, which was not much into attention before few years, what has let the people tilt to buy used cars (though the buying capacity and per head income has increased only).. these are the factors that can help to understand the overall changes in buyer's mentality and economic status of the country. With the help of Rstudio we have found some really helpful and informational analysis with graphical interpretation to find insights of the data.

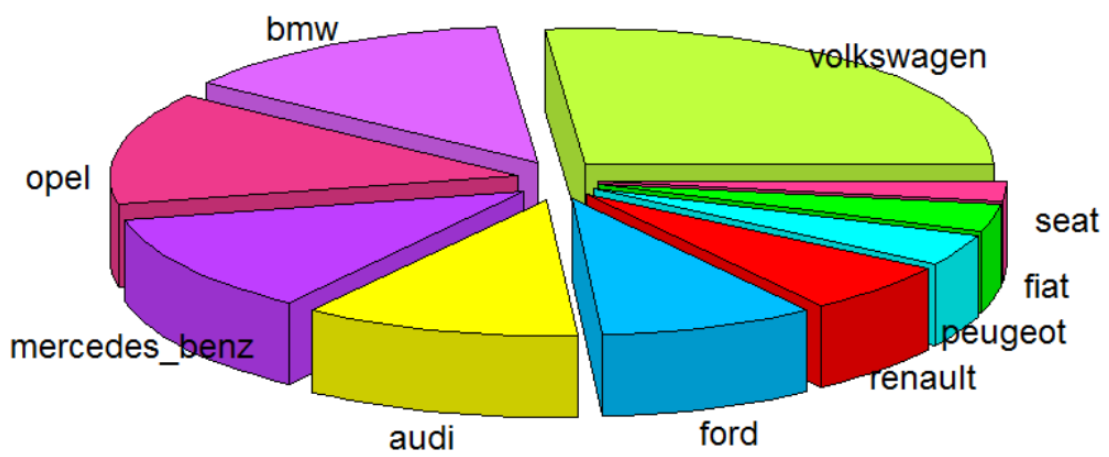
**B) Data Cleaning:**

|  |  |
| --- | --- |
| **Problem** | **Un-cleaned and Cleaned Data** |
| **Deleting**  **/Removing irrelevant columns** | C:\Users\Jalashri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\eliminating irrelevant columns.png  The dataset contained 20 columns. We removed unwanted columns and cleaned the data which contains 15 columns. |
| **Splitting Columns** | C:\Users\Jalashri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\before splitting.png  The "dateCrawled" column had a united date, but we required separated date by month, date, year and time. Hence we split the column into four.  C:\Users\Jalashri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\after splitting date.png |
| **Replacing Null-Values with Mean Value** | C:\Users\Jalashri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\before replacing null values.pngThere were NULL values in our dataset and we replaced them with the mean value of the price-variable.  C:\Users\Jalashri\AppData\Local\Microsoft\Windows\INetCache\Content.Word\after replacing null values with mean.png |

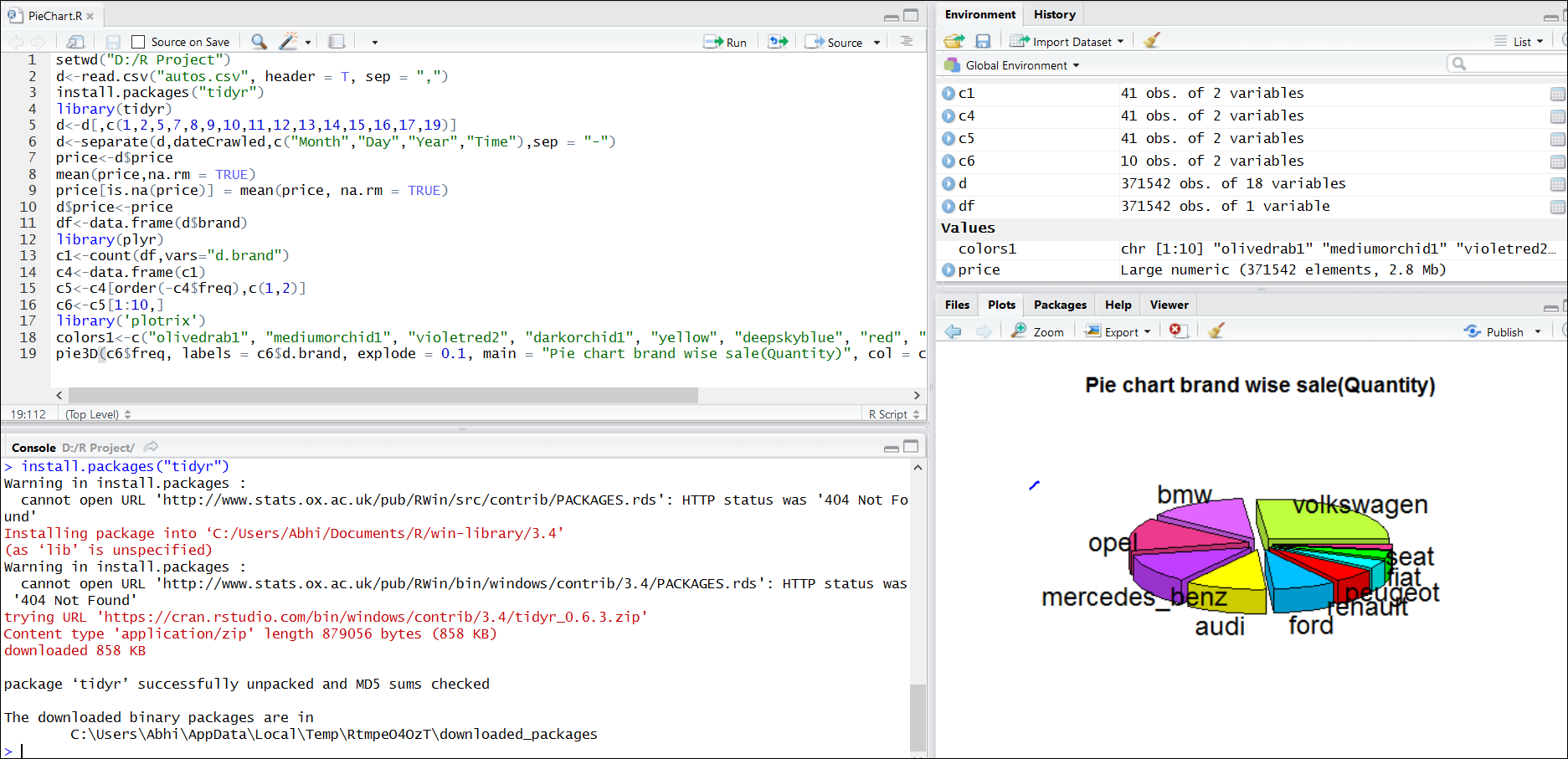
**C) Data Visualizations:**

**# Brand Wise Proportion Of Cars In Demand In "Used-Car Market"**

**C:\Users\Jalashri\Downloads\Pie Chart 3D 2.PNG**

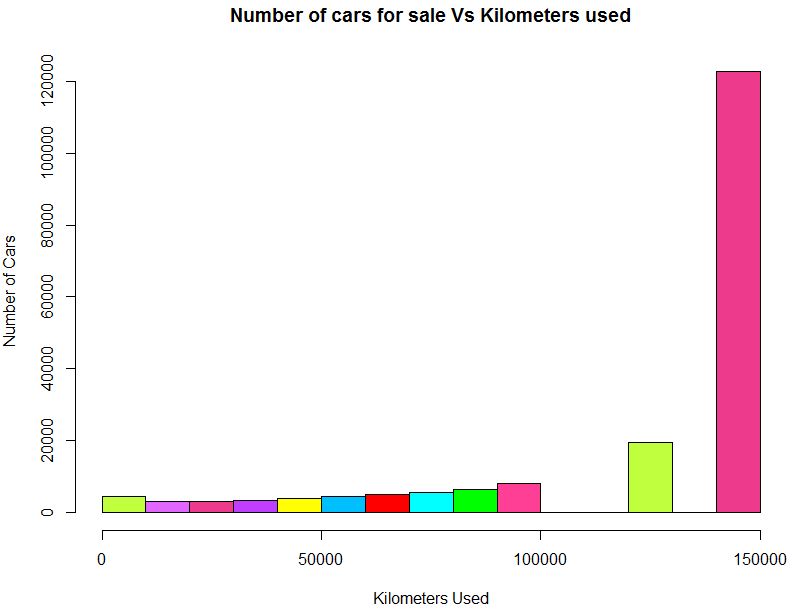
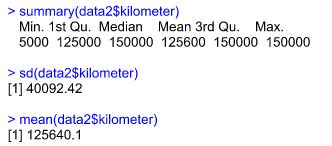
****

(Highlights from RScript –plotrix, plyr, pie chart)

****

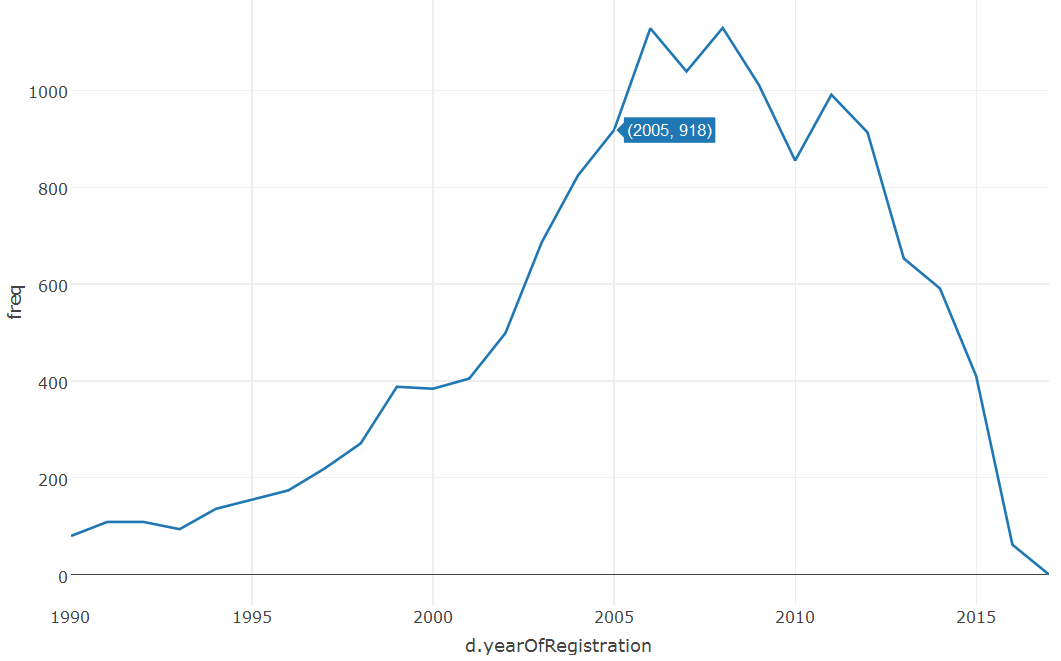
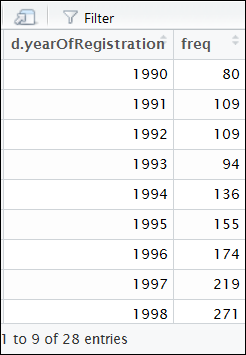
|  |
| --- |
| With the use of **R-Script** we have analyse the facts about brand wise sale proportion in above pie chart. It shows the total count of cars of individual brands for sale. We can see the brand for which the maximum cars are for sale in Used-Cars Market. Maximum is VOLKSWAGEN brand's cars in demand for sale followed by brands like, BMW, OPEL, MERCEDES, AUDI and etc. There are total 40687cars are for sale. Where, for BMW, there are 20545 cars are available. By looking at this statistics, we analyze that the performance and the expected price point criteria best matches for Volkswagen brand and people try to pursue their desire to buy the high end cars like BMW, Mercedes and Audi brands in Used-Cars Market. |

**# No. of Cars For Sale with its Average Odometer Reading**

In the above Histogram chart, We have tried to analyse the total number of cars available for sale in co-ordination with the kilometres travelled. We can analyse from this histogram that, maximum cars available for sale are used above 150000 kilometres. Almost more than 120000 cars are used for 150000Km, which are now available for sale. That way, less than 20000 cars are the ones that are used between 120000Km to 140000Km. And, less than 5000 cars are used for less than the 50000Km and now available for the sale. we can see in the summary, standard deviation for these data (40092.42) and the Mean (125640.1) along with other useful information for understanding the Histogram in detail.

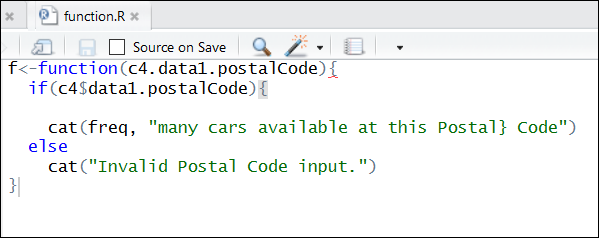
**#** **Which year's model of SUVs are highest in demand?**

This time series chart shows different models of SUV cars and its demand in Used-Car Market. We can clearly analyse the trend of SUV cars models & its sale. SUV cars registered between year 2005 to year 2010 are in highest demand. We can see in the table the year wise SUV sales, which is shown in the trend line above.

SUV models registered under the year 2005 are 918 cars which get sold. This time series chart help the manufacturing companies to predict the launch of their new car models. Also, we can analyse the customers' priority for car's registration model matching with their price points. This same analysis can be done to understand various brands car sale, model wise to perform future strategies more precisely. Manufacturing companies can make sure that which features are the customers exactly looking for in specific segment of cars in Used-car market.

**Function: function to find the number of cars with respect to postal codes**

****

🡺**R Code for all the Analysis and Visualizations:**

**DATA - CLEANING :**

setwd("D:/R Project")

data<-read.csv("autos.csv", header = T, sep = ",")

View(data)

data1<-data[,c(1,2,5,7,8,9,10,11,12,13,14,15,16,17,19)]

View(data1)

install.packages(‘tidyr’)

library(‘tidyr’)

data2<-separate(data1,dateCrawled,c("Month","Day","Year","Time"),sep = "-")

View(data2)

price<-data2$price

mean(price,na.rm = TRUE)

[1] 17295.44

price[is.na(price)] = mean(price, na.rm = TRUE)

data2$price<-price

**PIE - CHART : Brand Wise Proportion Of Cars In Demand In "Used-Car Market"**

df<-data.frame(d$brand)

library(plyr)

c1<-count(df,vars="d.brand")

c4<-data.frame(c1)

c5<-c4[order(-c4$freq),c(1,2)]

c6<-c5[1:10,]

library('plotrix')

colors1<-c("olivedrab1", "mediumorchid1", "violetred2", "darkorchid1", "yellow", "deepskyblue", "red", "cyan", "green1", "violetred1")

pie3D(c6$freq, labels = c6$d.brand, explode = 0.1, main = "Pie chart brand wise sale(Quantity)", col = colors1)

setwd(“C:/Users/Abhi/Desktop”)

source(“PieChart.R”)

**HISTOGRAM : No. of Cars For Sale with its Average Odometer Reading**

summary(data2$kilometer)

Min. 1st Qu. Median Mean 3rd Qu. Max.

5000 125000 150000 125600 150000 150000

sd(data2$kilometer)

[1] 40092.42

mean(data2$kilometer)

[1] 125640.1

colors1<-c("olivedrab1", "mediumorchid1", "violetred2", "darkorchid1", "yellow", "deepskyblue", "red", "cyan", "green1", "violetred1", "darkslategray1", "gold1")

hist(data2$kilometer, main = "Number of cars for sale Vs Kilometers used", xlab = 'Kilometers Used', ylab = 'Number of Cars', col = colors1)

**LINE-CHART : Which year's model of SUVs are highest in demand?**

year<-data.frame(d$yearOfRegistration)

View(year)

install.packages("plyr")

library(plyr)

c1<-count(year,vars="d.yearOfRegistration")

p1 <- plot\_ly(c1, x = ~d.yearOfRegistration, y = ~freq)%%

+ add\_lines(name = ~"car")

p1

**References:**

# 1. <http://www.cnbc.com/2014/08/22/used-car-market-showing-no-signs-of-slowdown.html>

# 2. <http://www.autoremarketing.com/trends/used-car-sales-expected-grow-2016>