**CS544 Final Project\_report**

**Picking the Data Set**

Look into the following sites as an example and identify a data set that interests you.

1. https://www.kaggle.com/datasets
2. http://www.kdnuggets.com/datasets/index.html
3. Any other source of your choice

The data set named CAR DETAILS FROM CAR DEKHO.csv

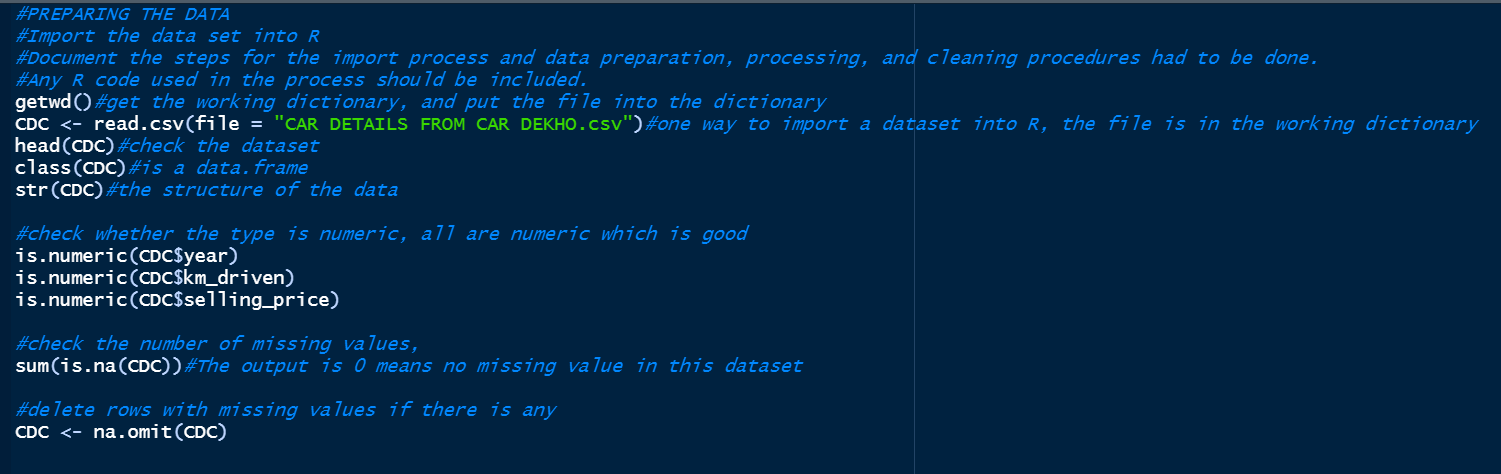
Number of tuples: 4340

Number of attributes: 8

Reference: source:https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho?select=CAR+DETAILS+FROM+CAR+DEKHO.csv

**Preparing the data (15 points)**

* Import the data set into R.
* Document the steps for the import process and data preparation, processing, and cleaning procedures had to be done. Any R code used in the process should be included.



I imported the data using read.csv (). So firstly, I got the working path by getwd(), then I put the file into the folder under the path. If the file is in the right dictionary, the read.csv () can read and import the file. The dataset is named CDC.

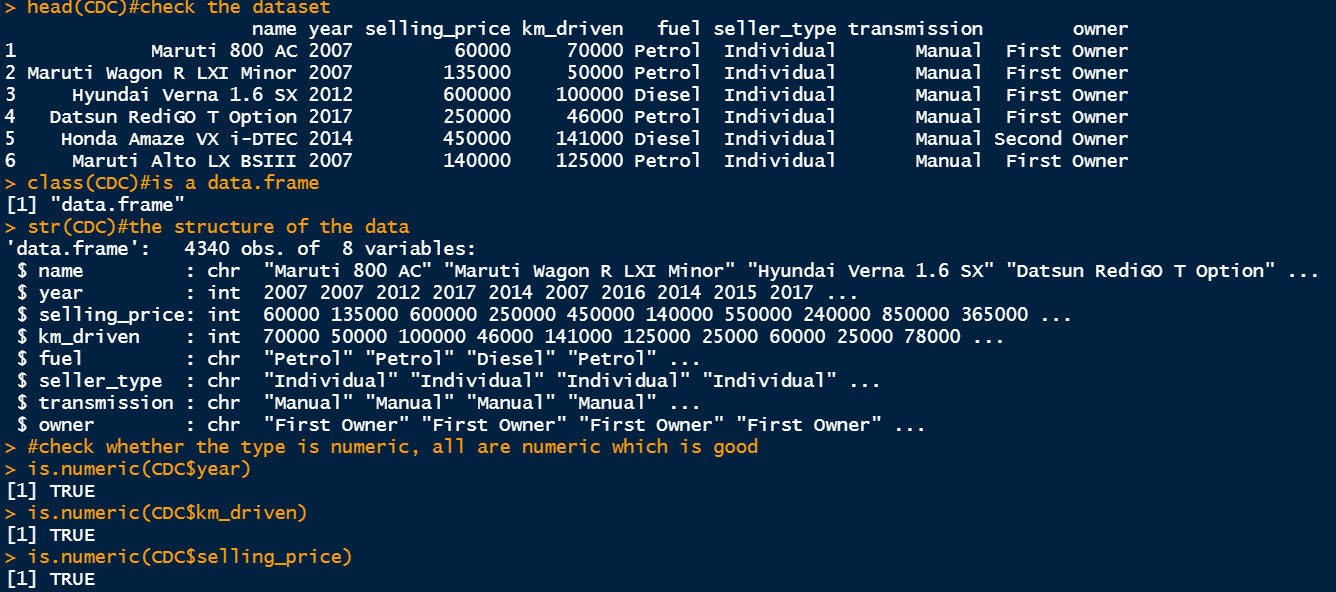
Then I check the class of the variables, for CDC$year, CDC$km\_driven and CDC\_selling\_price, I need them to be numeric. I used is.numeric to check the type, all returned TRUE which is good.

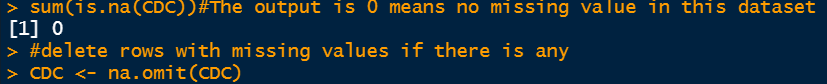
I checked to find out whether is there any NA value in this dataset by sum(is.na (CDC)). And remove any rows with NA values by na.omit(CDC).

Working dictionary:



Information about this dataset, and the variable type check:





There are no NAs in this dataset, I perform the na.omit() anyway.

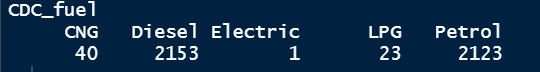
**Analyzing the data (50 points)**

* Do the analysis as in Module 3 for at least one categorical variable and at least one numerical variable. Show appropriate plots and properly label the plots. (10 points)



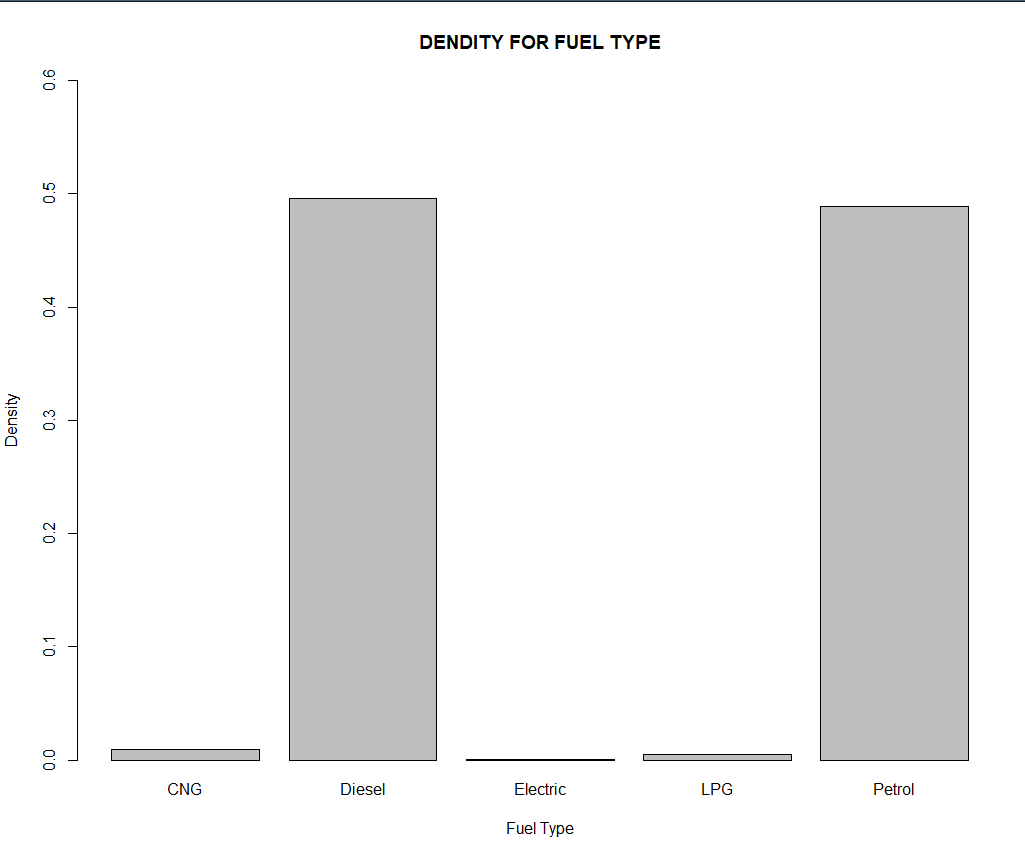
The categorical variable I chose is CDC\_fuel, it is a variable with 5 possible value.





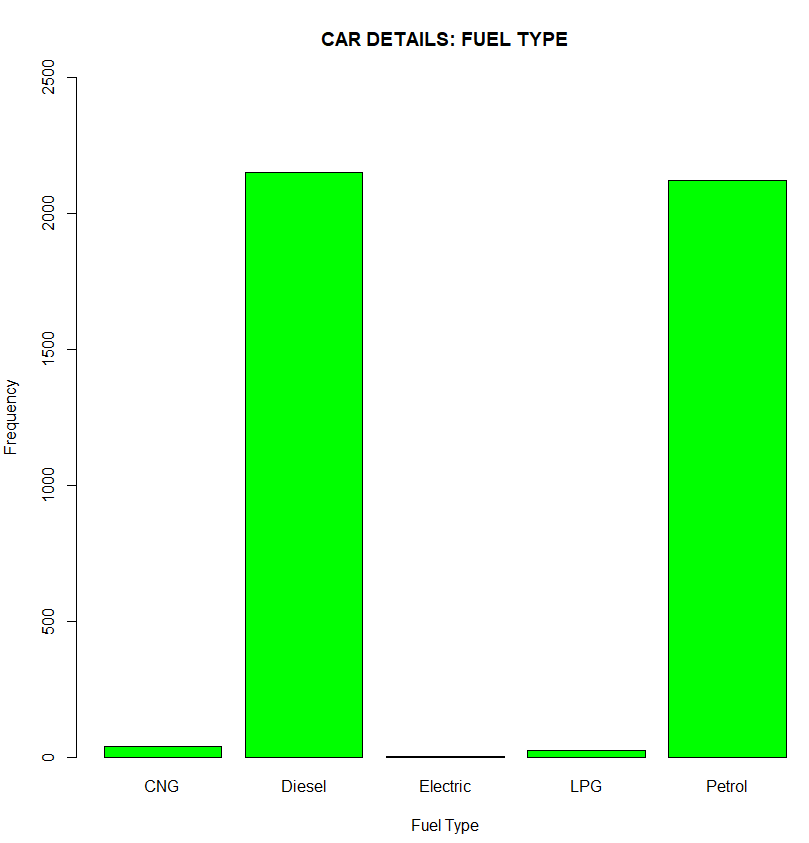
First, I made the barplot for density of this variable:





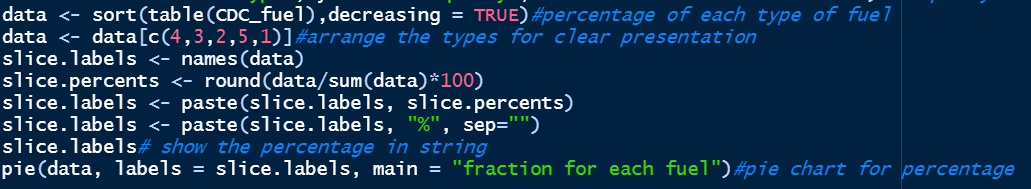
Then the barplot for amount for this variable:

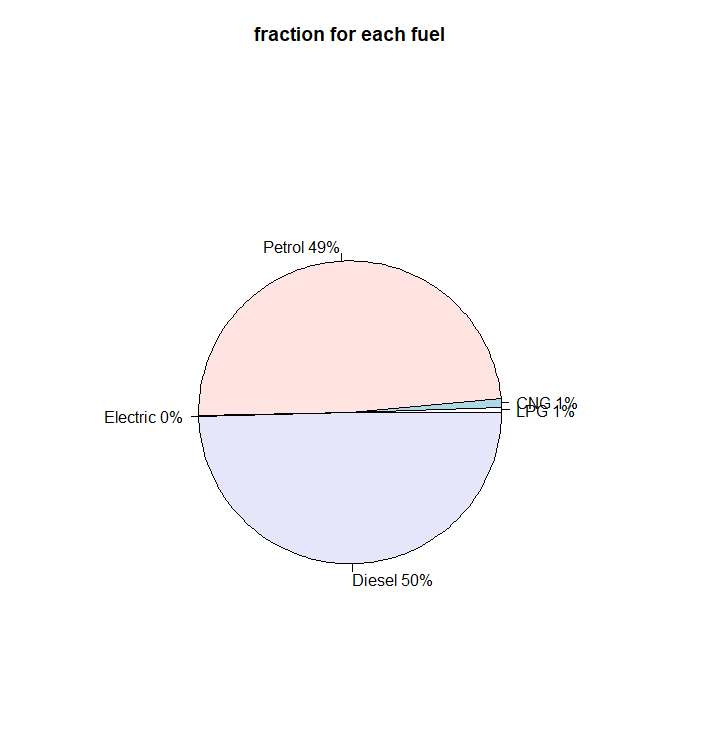




These two plots look identical. For the density plot, the plot represents percentage for each type of fuel for all data. For example, Diesel and Petrol are both close to 50% of all data, while others are all far less than 10%. The second plot represents the actual number for each type of fuel. Both diesel and petrol are between 2000 and 2500, and others are proximately less than 50. The goal for these two plots is to represents the relative amount between types of fuel.

They are another more direct way to represent the percentage of each type of fuel:



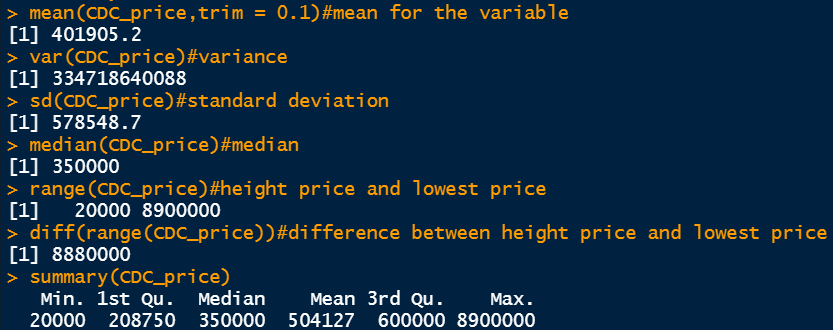


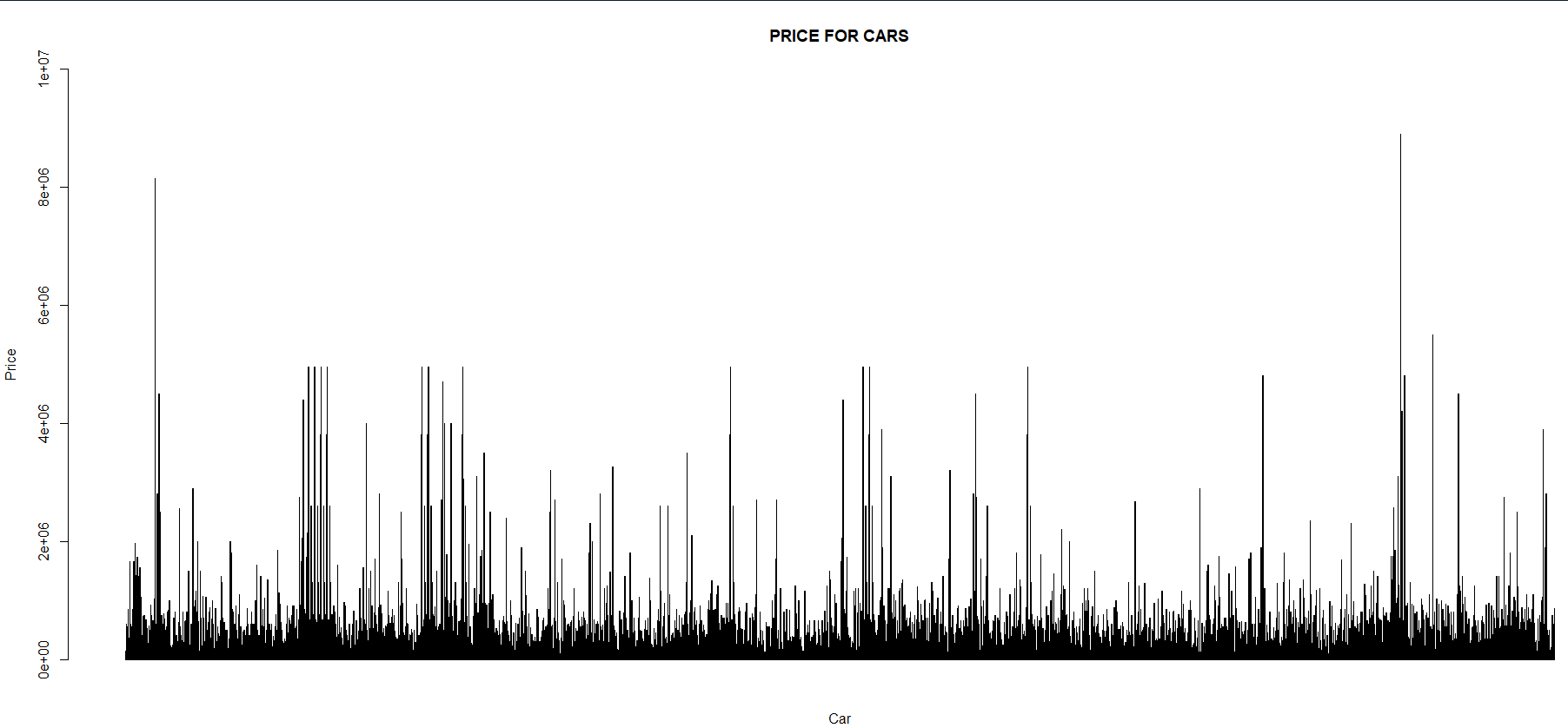
This is the pie plot for the CDC\_fuel variable. I sort the variable and arrange the variable before plotting to gain a clear plot. I put electric on the side away from CNG and LPG. From The plot, percentage for each type of fuel is shown accompanied with names of fuel.

The numeric variable I choose is CDC\_price which is the prices for all cars.

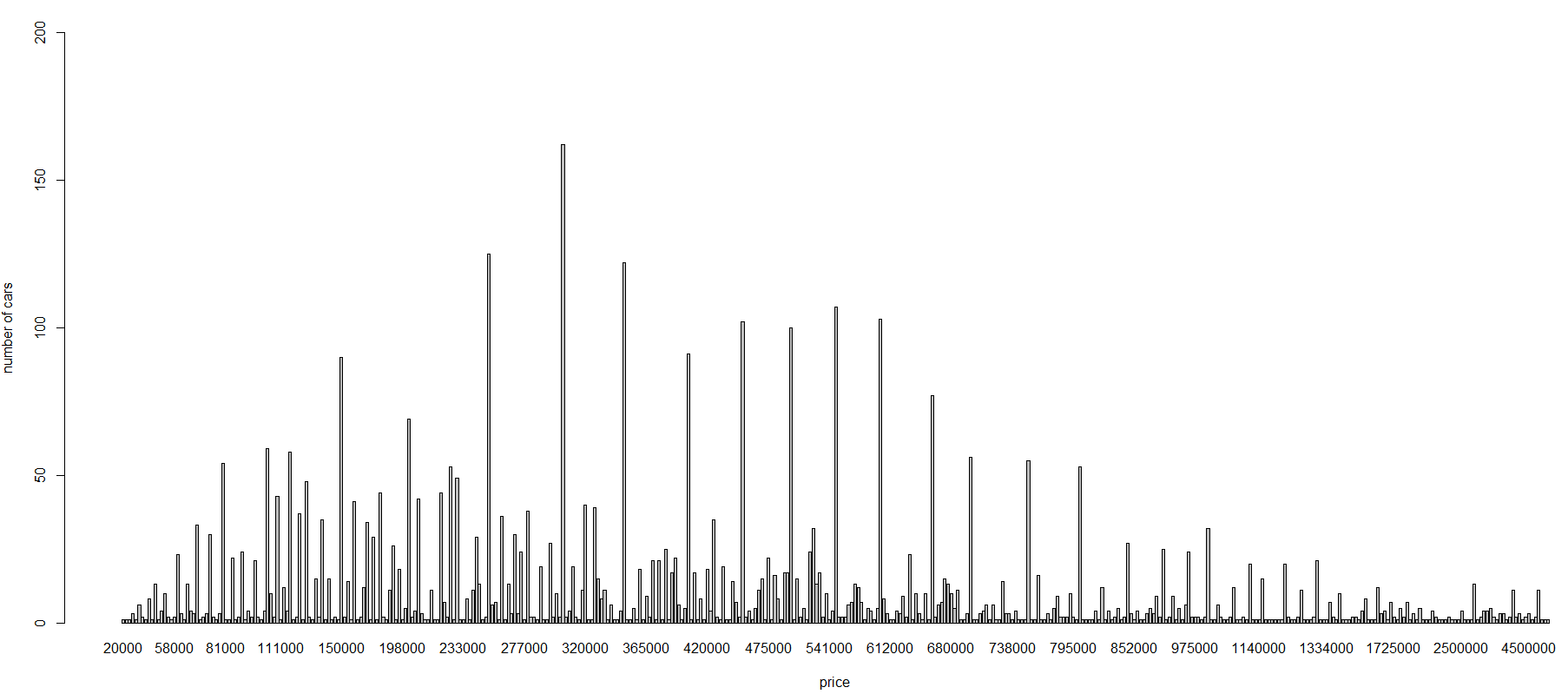


The basic information for this variable:

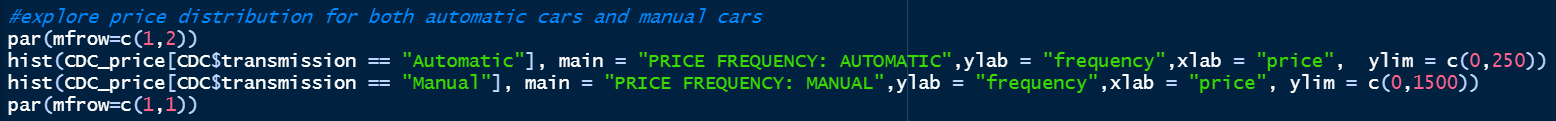


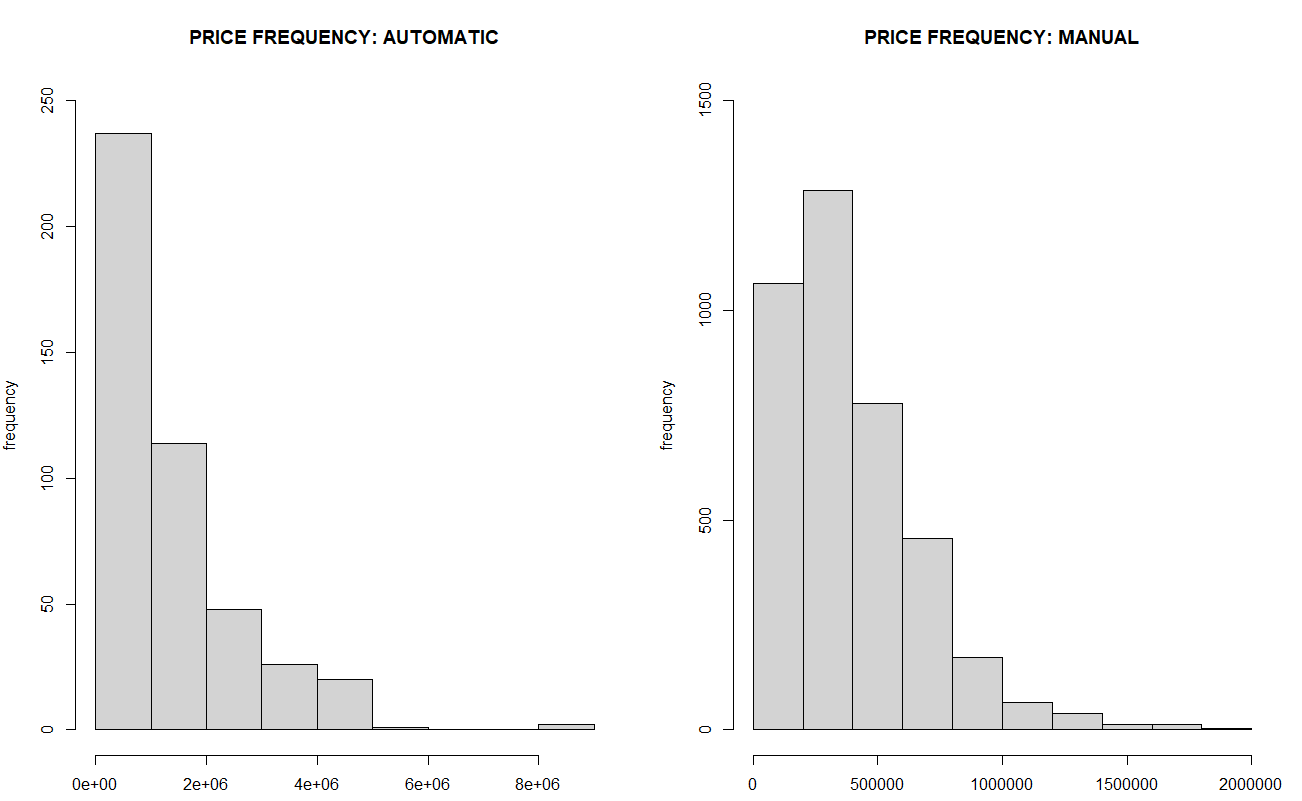


The barplot shows the prices for all cars. This plot offers brief information for prices. For example, most of cars have price less than 1,000,000 $, a few of them can be between 4,000,000$, very few cars can be more than 6,000,000$, the most expensive one is around 8,000,000$.

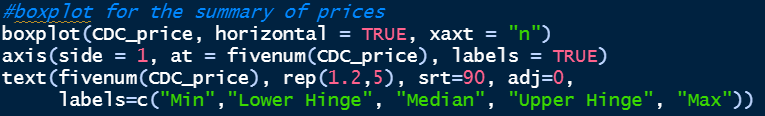


This graph represents density of prices. Using this graph, we can find out what is the common prices for the cars. The most common price is around 300,000$. Most of cars are between 81,000$ and 1,334,000$.





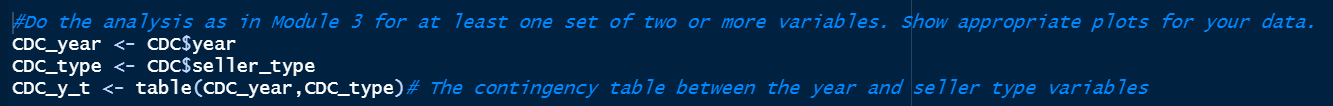
These two pots show the prices distribution for cars with automatic transmission and manual transmission. We can find out questions such as what kind of cars will have automatic transmission and what kind of cars can have manual transmission, and what are the prices if someone wants to get a manual car. In general, manual cars are less expensive than automatic cars. Almost all prices for automatic cars are less than 6,000,000$, a few of them can be more than 8,000,000$. There are many automatic cars over 2,000,000$. However, manual cars are cheaper. All of manual cars are less than 2,000,000$.

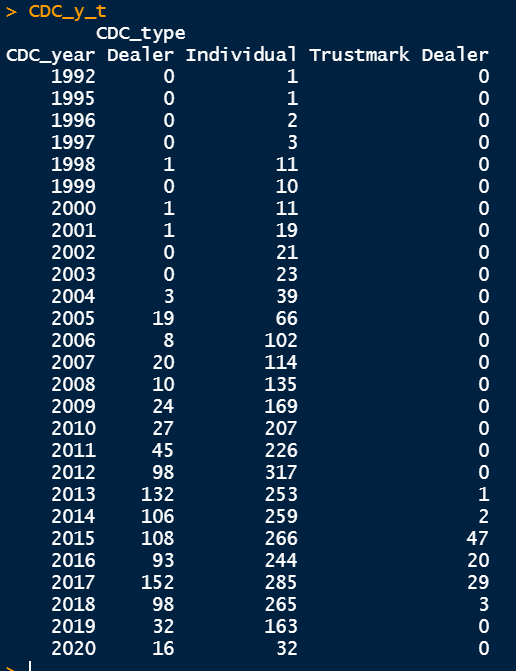




The bosplot shows the statistical information of the variable. There are lots of outliers represents extremely expensive cars. Most of prices are between 208,750$ and 600,000$, the cheapest car is 20,000$.

Do the analysis as in Module 3 for at least one set of two or more variables. Show appropriate plots for your data. (10 points)



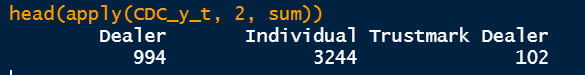


The variables show number of cars for each year and seller type. For example, dealer has no cars of 1992. Individual has 226 year 2011 cars.

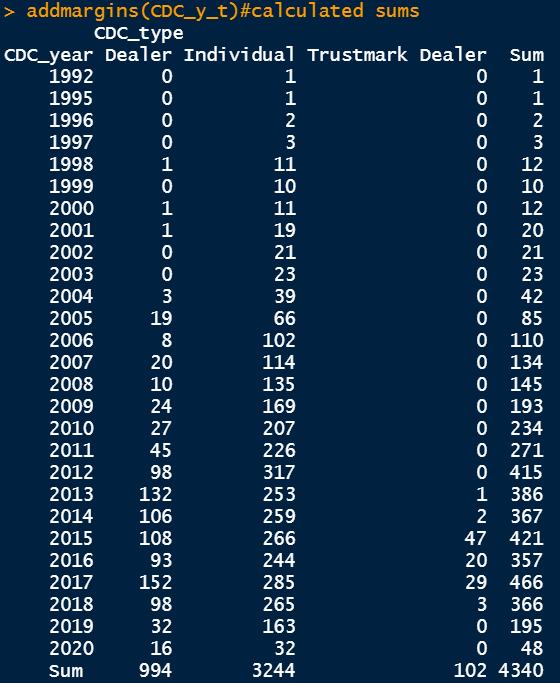
Number of cars for each year:

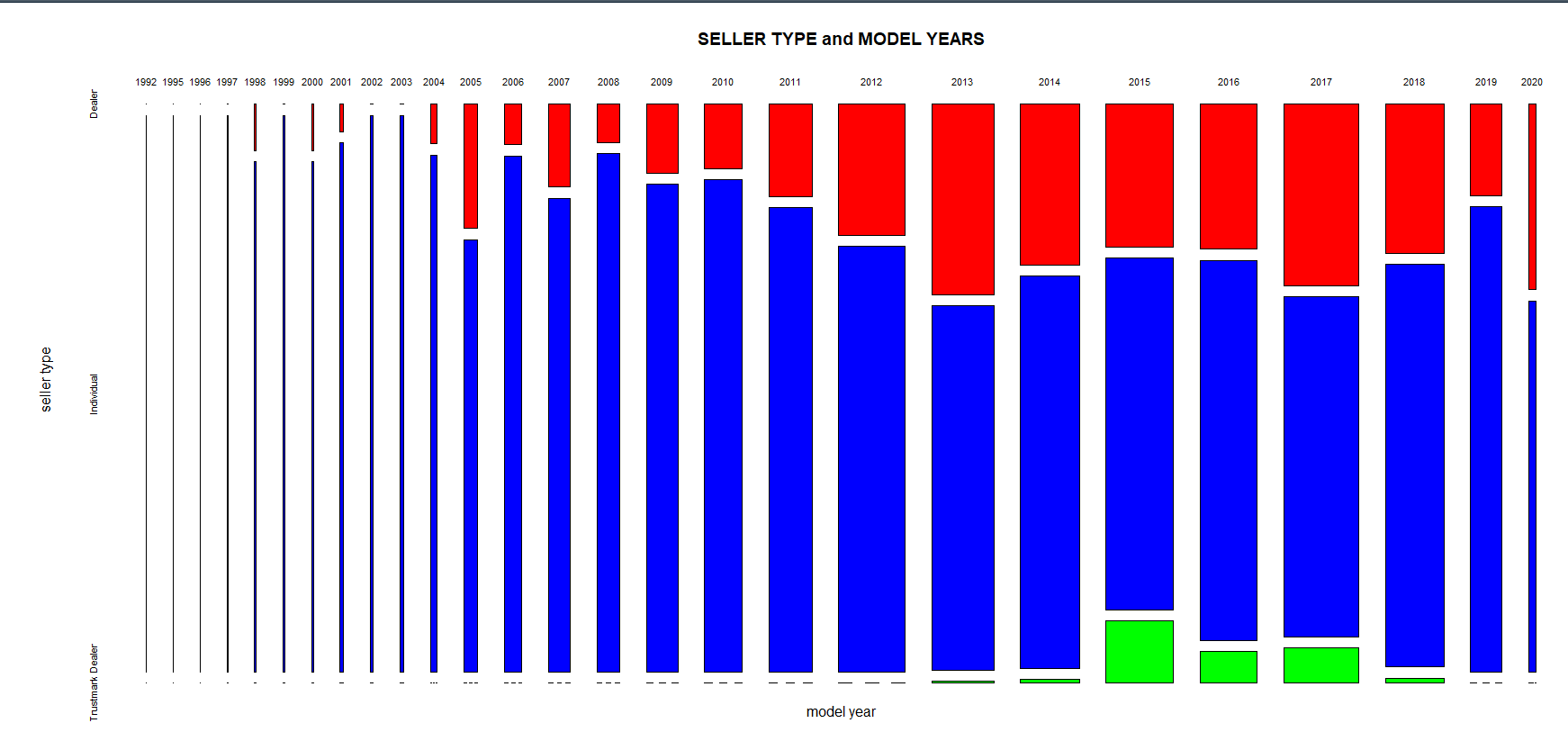


Number of cars for each type of seller:

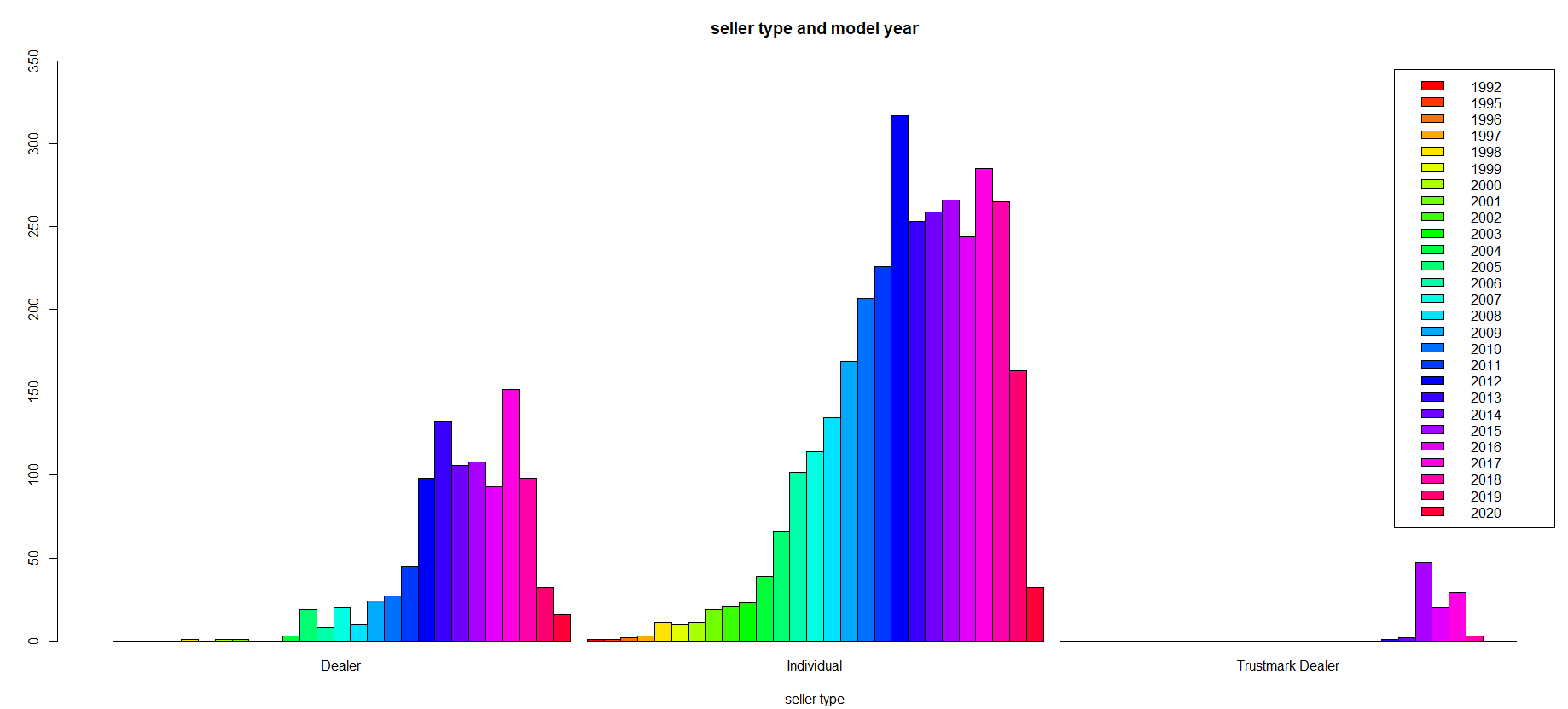


Sum for each year and each type of seller:



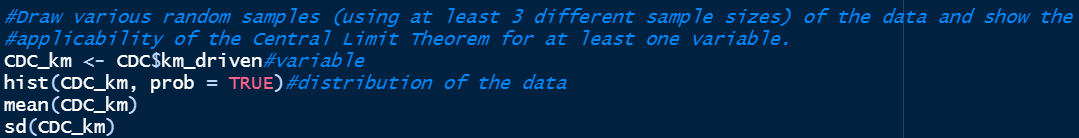


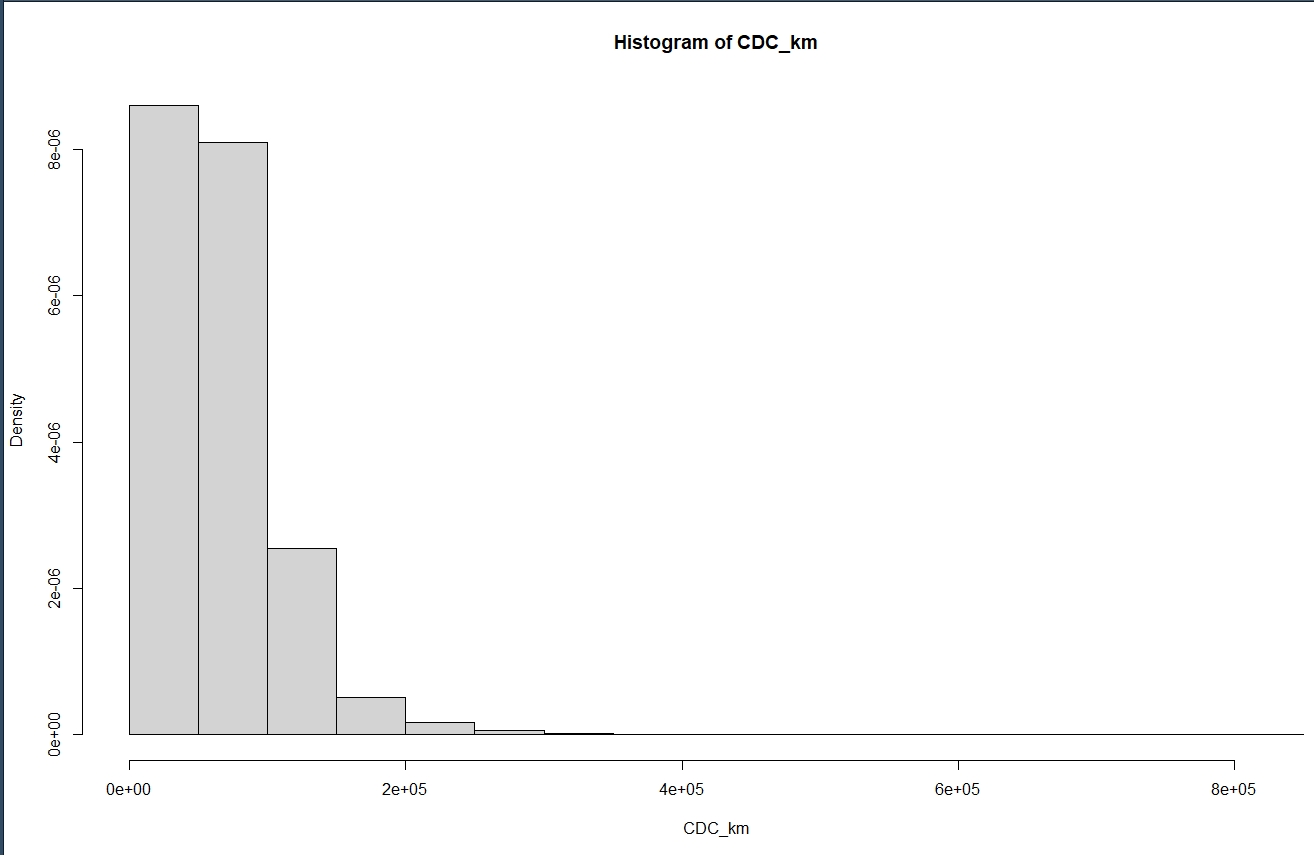
The mosaic plot shows the information of numbers of cars for each year and each type of seller. For example, individual has most of cars. Dealer has wide range of years for cars compare to Trustmark dealer. Trustmark dealer only has relative new cars.

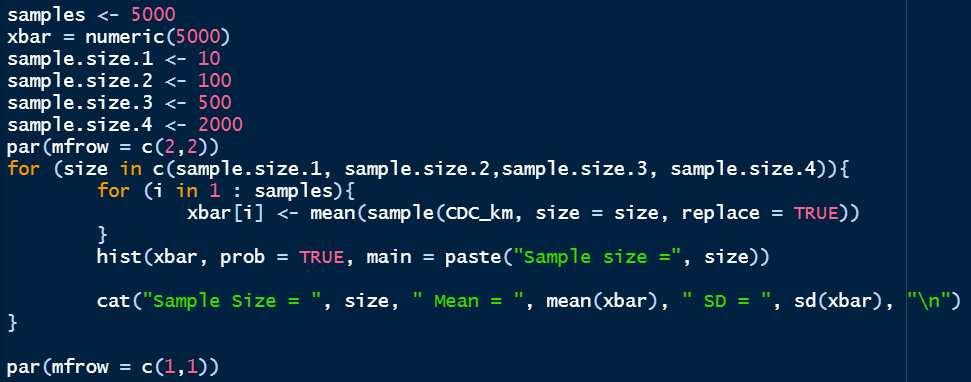


The barplot shows very detailed information. In general, individual have more cars than dealer and Trustmark dealer for every year of model. Similar to dealer, individuals’ cars are mostly between 2000 model and 2018 model, while Trustmark dealer only have cars latter than 2018.

* Draw various random samples (using at least 3 different sample sizes) of the data and show the applicability of the Central Limit Theorem for at least one variable. (15 points)

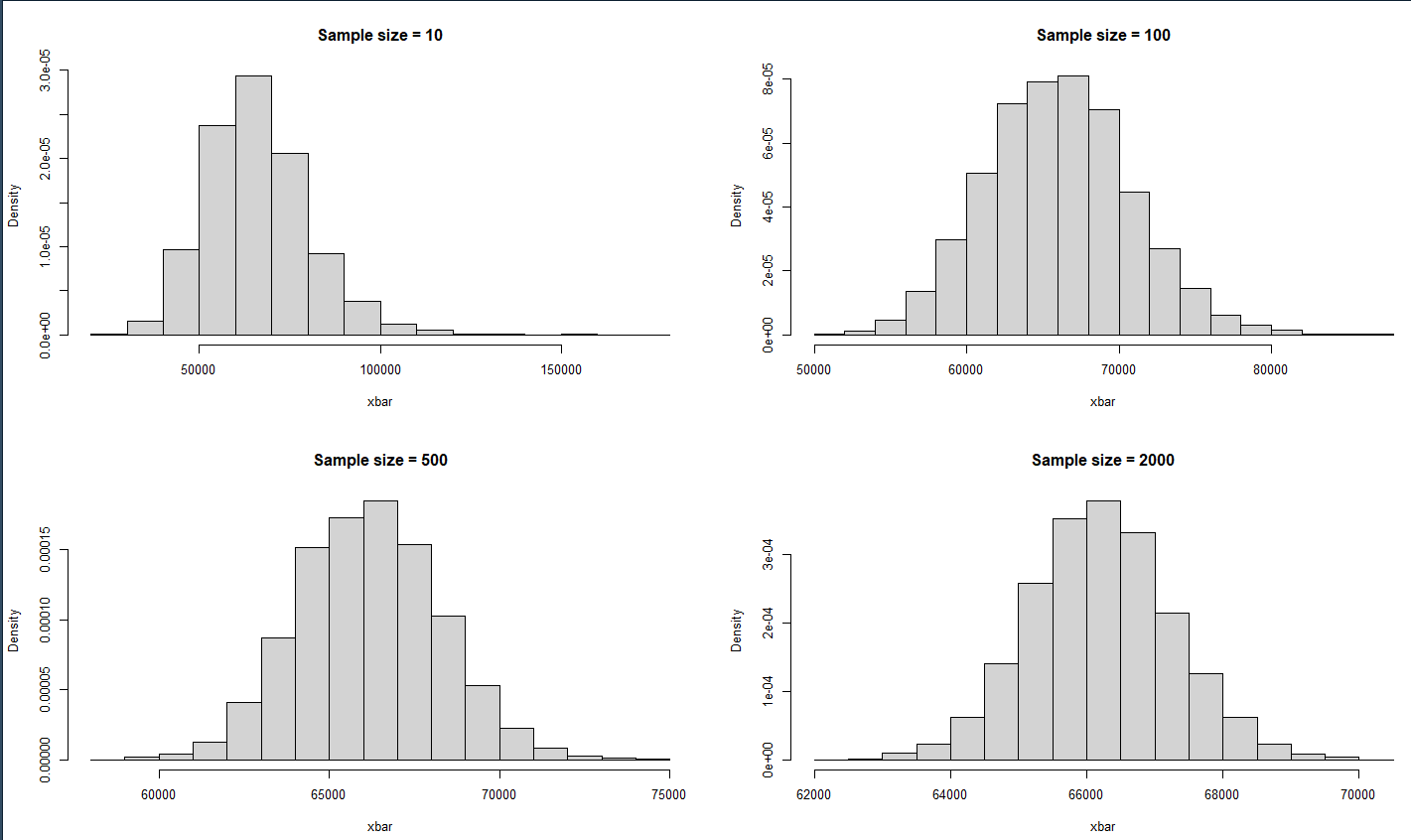


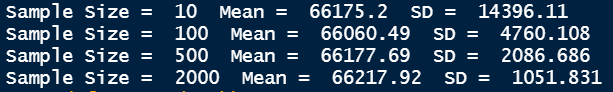




Draw 500 samples for each sample size. The sample size are 10, 100, 500 and 2000.

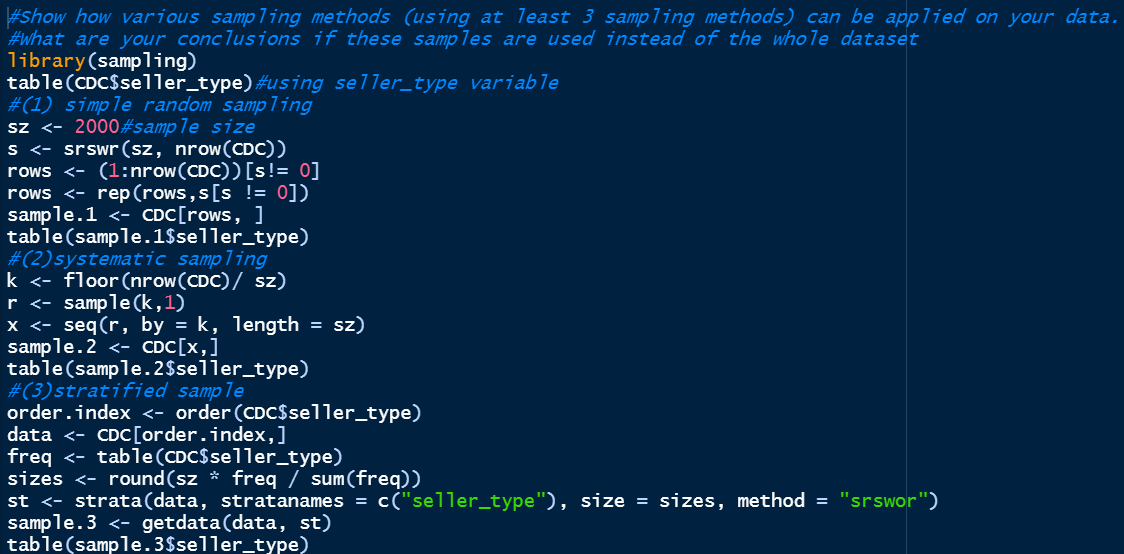
Plot:

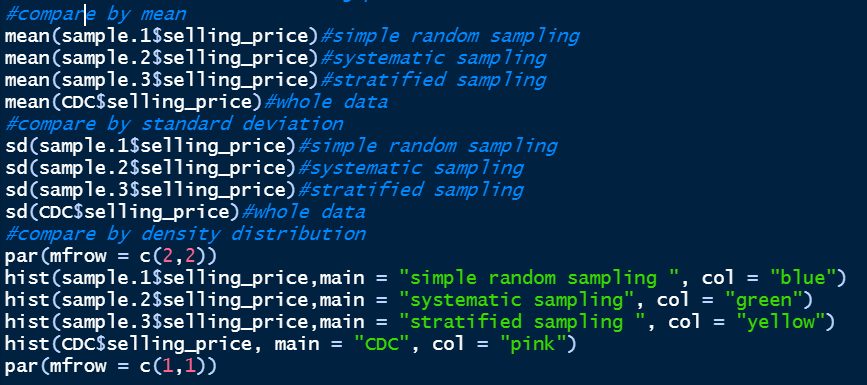




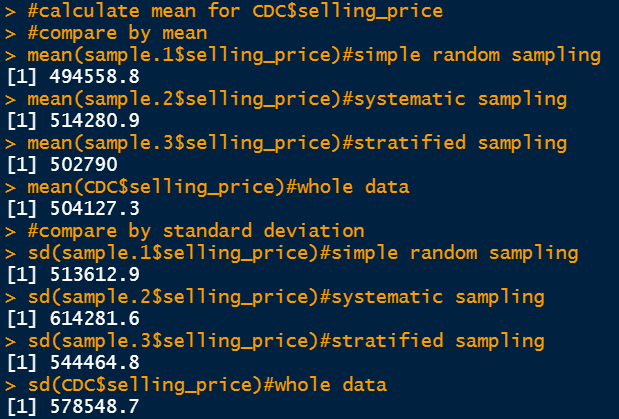
when the sample size increase, the graph is closer to normal distribution. With the sample size increase, the standard deviation decrease. The mean stays the same.

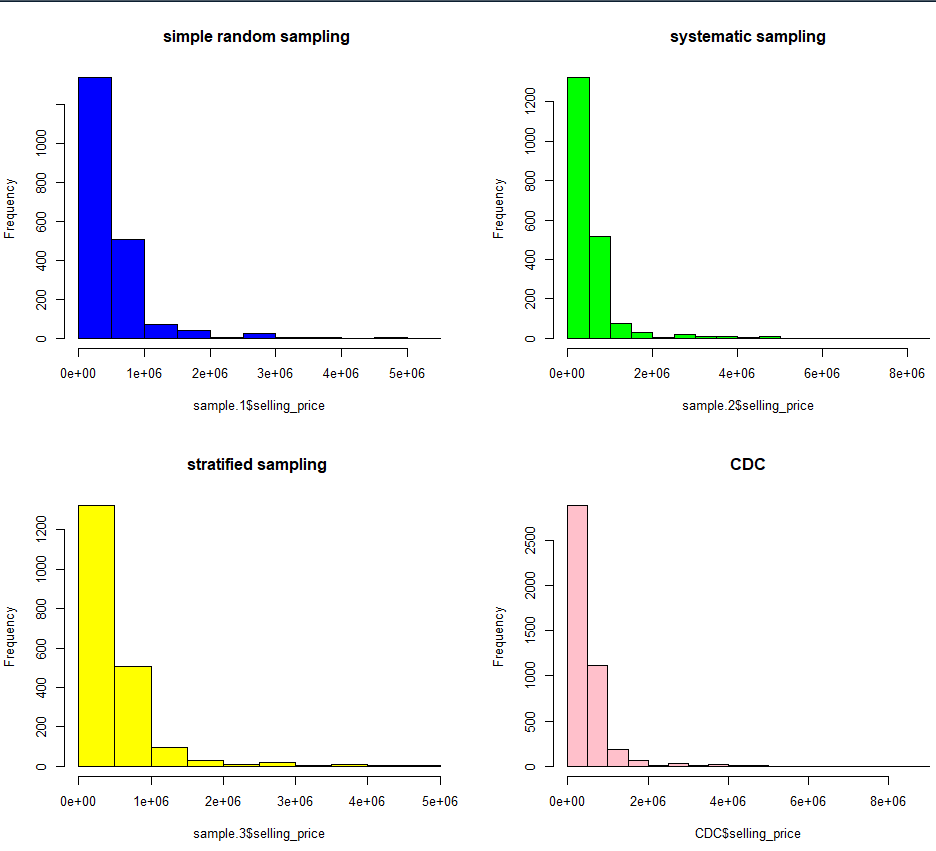
* Show how various sampling methods (using at least 3 sampling methods) can be applied on your data. What are your conclusions if these samples are used instead of the whole dataset? (15 points).





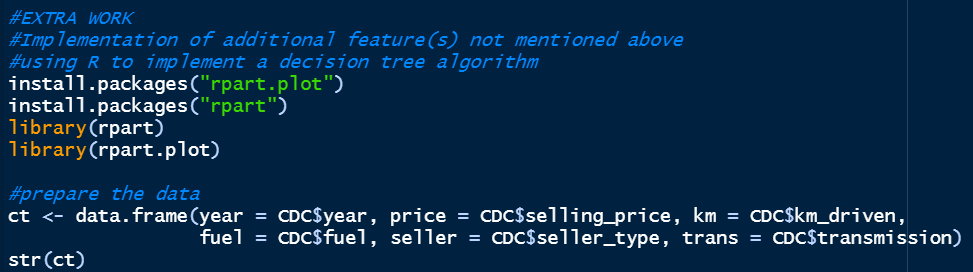
Result:





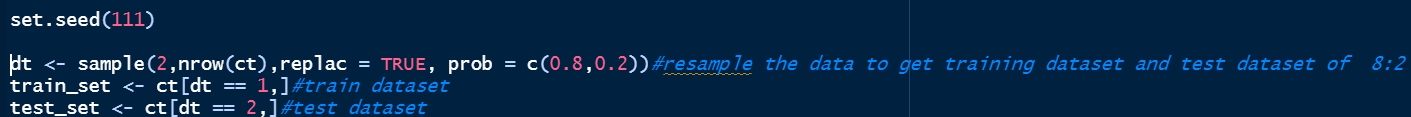
When sampling methods are applied on the data. Changes for mean and standard deviation for the numeric variable: selling\_price are observed. No matter what sampling method, the data will not stay the same. For the sampling methods used, stratified sampling has the closest mean and standard deviation compare to the original data. I would conclude that for selling\_price variable, the stratified sampling is the best way to be applied. From the graph, the distribution for the variable for all sampling methods and the original dataset are identical.

* Implementation of additional feature(s) not mentioned above (20 points)

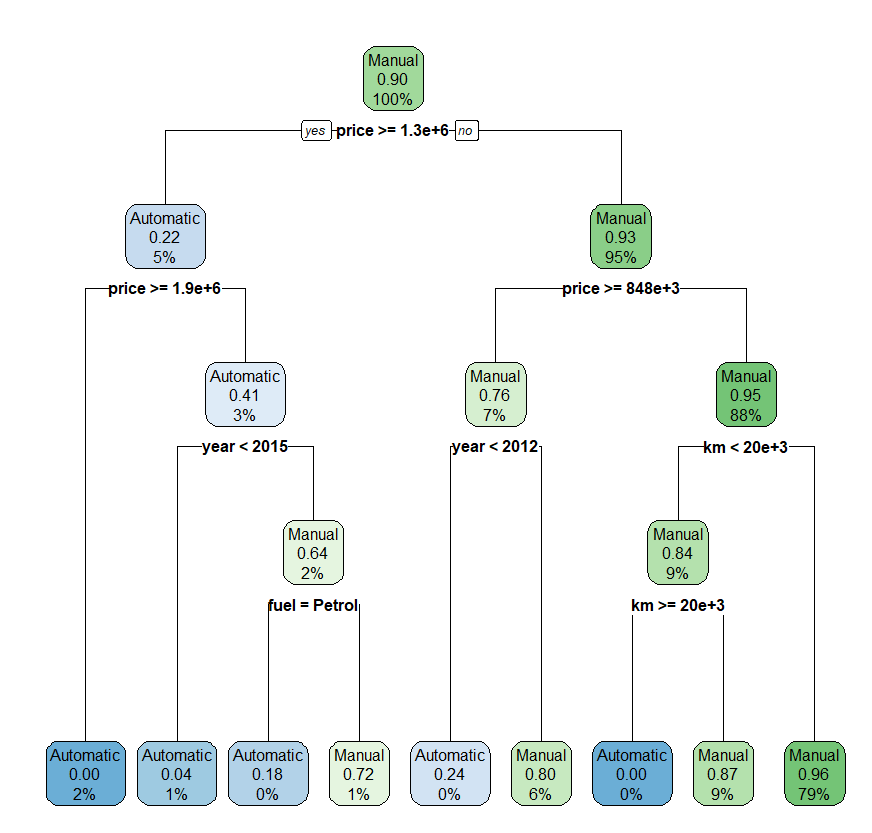


For the extra part, I did a decision tree.

Resample the date for training dataset and test dataset. The training dataset is used to train the model while the test dataset is for prediction. The ratio is 8 : 2. The class attribute is transmission.

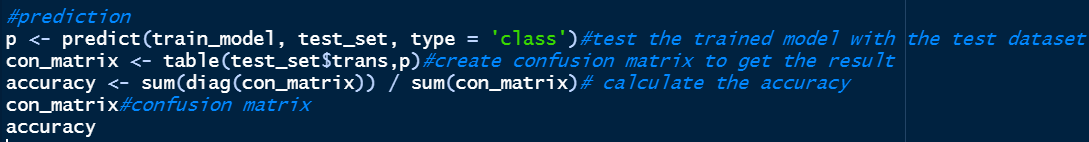


Build the mode and plot the tree:

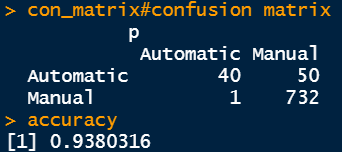


For the tree plot, if the price is less than 1,300,000, price less than 848,000, km >= 20000km, then the car is manual car.

Predict the test set and output confusion matrix and accuracy:



Result:



The confusion matrix shows that there are 40 automatic cars are correctly recognized as automatic, but 50 automatic cars are incorrectly recognized as manual. Only 1 manual car is incorrectly recognized as automatic. The accuracy is around 93.8%. The accuracy looks good, but since this is an unbalanced dataset, the true accuracy is not as good as the output The model successfully predicts the manual which is the majority class of the class attribution, but the automatic is poorly predicted. If the test class is manual, the model will have good accuracy, if the test class is automatic, the prediction will only have around 50% accuracy.