Project Step 2

Kiran Komati

5/22/2021

How to import and clean my data

The data set for my project is Car Dekho Dataset from kaggle.com. 1. It is from 3 files and in csv format. We can import the csv data set into R easily by using read.csv function. 2. Data set 3 has additional columns. I need to create a new data set from this without these additional columns in order to be able to merge the 3 data sets without any issues. 3. Dataset 2 has one additional column,Current_price , I need to remove it before merging. 4. Dataset 1 has the price in multiples of 100k rupees. I need to convert to a price by multiplying it with 100000. 5. Rename data sets 1 and Data set 3 to match the names of Data set2. 6. I need to merge the 3 data sets into a single data set. 7. Perform Exploratory Data Analysis to understand more about the columns and if its normally distributed.

```
library('dplyr')
CD.df1Raw <- read.csv("CarDekho1.csv")
CD.df2Raw <- read.csv('CarDekho2.csv')
CD.df3Raw <- read.csv('CarDekho3.csv')
CD.df1<-select(CD.df1Raw, -Present_Price) %>%
        rename(name=Car_Name,
                     year=Year,
                     selling_price=Selling_Price,
                     km_driven=Kms_Driven,
                     fuel=Fuel_Type,
                     seller_type=Seller_Type,
                     transmission=Transmission,
                     owner=Owner) %>%
        mutate(owner=recode(owner,`0`="First Owner",
                            `1`="Second Owner",
                            `3`="Fourth Owner"))

CD.df1$selling_price<- CD.df1$selling_price*100000

CD.df2 <- CD.df2Raw
CD.df3 <- select(CD.df3Raw,-engine,-max_power,-torque,-seats,-mileage)
CarData <- rbind(CD.df1,CD.df2,CD.df3)
```

What does the final data set look like?

```
head(CarData)
```

```
##            name year selling_price km_driven   fuel seller_type transmission
## 1          ritz 2014        335000     27000 Petrol      Dealer       Manual
## 2           sx4 2013        475000     43000 Diesel      Dealer       Manual
## 3          ciaz 2017        725000      6900 Petrol      Dealer       Manual
## 4       wagon r 2011        285000      5200 Petrol      Dealer       Manual
## 5         swift 2014        460000     42450 Diesel      Dealer       Manual
## 6 vitara brezza 2018        925000      2071 Diesel      Dealer       Manual
##         owner
## 1 First Owner
## 2 First Owner
## 3 First Owner
## 4 First Owner
## 5 First Owner
## 6 First Owner
```

```
summary(CarData)
```

```
##      name                year      selling_price        km_driven      
##  Length:12769       Min.   :1983   Min.   :   10000   Min.   :      1  
##  Class :character   1st Qu.:2011   1st Qu.:  239000   1st Qu.:  34000  
##  Mode  :character   Median :2014   Median :  415000   Median :  60000  
##                     Mean   :2014   Mean   :  588620   Mean   :  67820  
##                     3rd Qu.:2017   3rd Qu.:  650000   3rd Qu.:  90000  
##                     Max.   :2020   Max.   :10000000   Max.   :2360457  
##      fuel           seller_type        transmission          owner          
##  Length:12769       Length:12769       Length:12769       Length:12769      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
## 
```

Questions for future steps.

1. Is there are relation between the variables.
2. Is there a positive correlation or negative correlation between the variables?
3. Which model do we need to use fo the price prediction?
4. is there any specific variable that we can start the model with?
5. Once the model is finalized, what variables need to be used?
6. How can we validate the performance of the model?

What information is not self-evident?

It is not self evident if there is any relation between the variables. we need to use the R functions to understand the same which we learned in the past few weeks.

What are different ways you could look at this data?

1. We can look at each variable to see how they are distributed.

2. We can check if they are left or right skewed.
3. We can check at the mean selling price to understand and compare our model against.

How do you plan to slice and dice the data?

I have already done this as part of the data cleaning. I have eliminated the fields that are not common to the three data sets and selected(diced) only the columns tt are needed. Filtering out(slicing) is not needed as there are no columns that have null or na values.

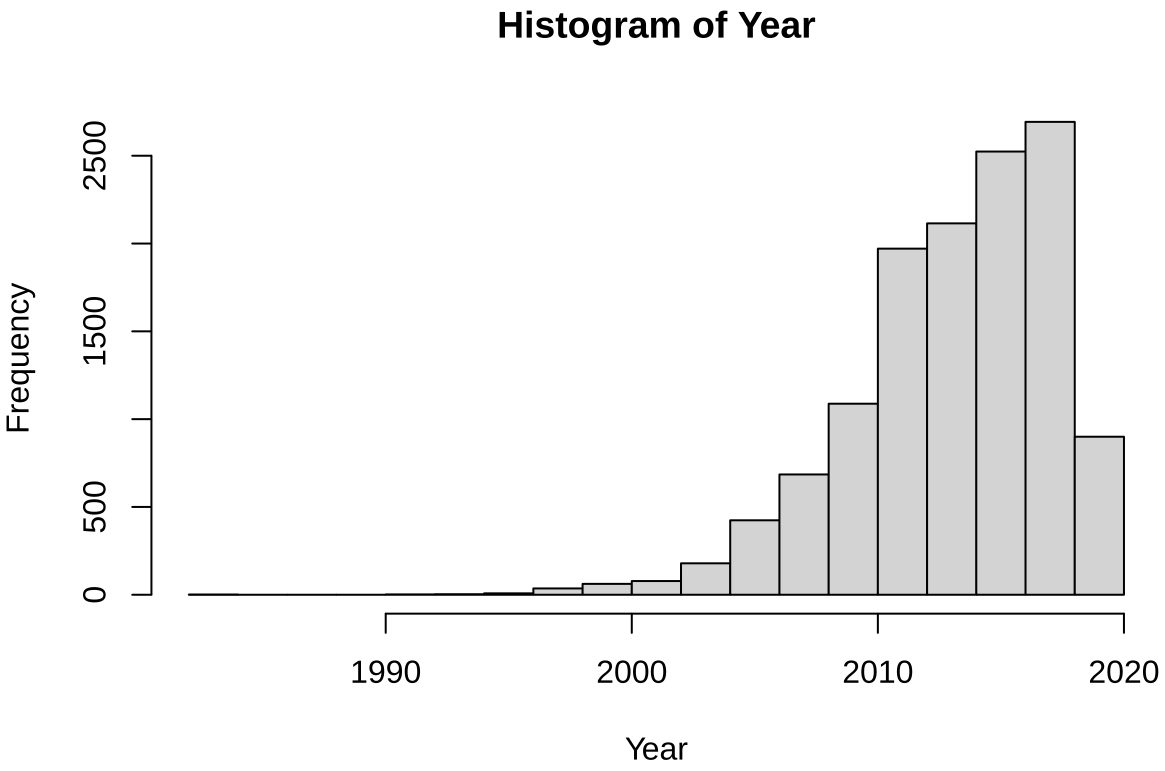
How could you summarize your data to answer key questions?

For the numerical values we can get the mean,median,variance,standard deviation,minimum,maximum,range values and for the categorical columns we can get the counts and percentages to understand the distribution of the data.We can use boxplots to understand if there are any anomalies in the data.

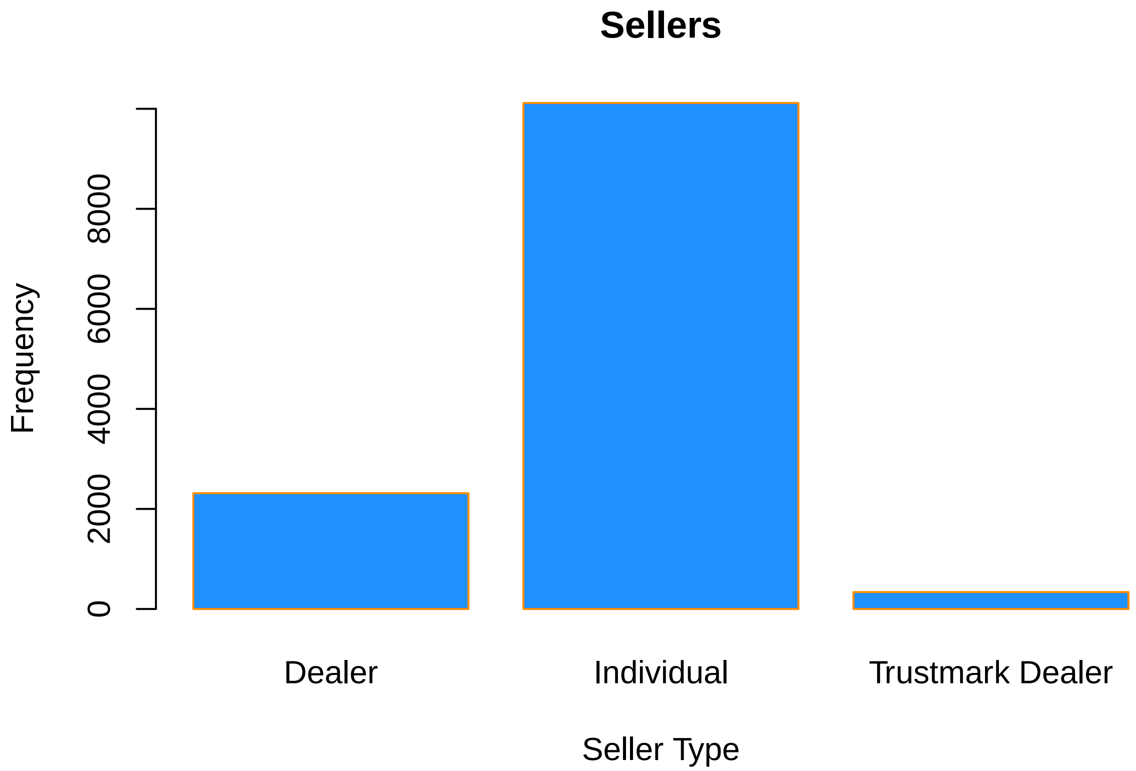
What types of plots and tables will help you to illustrate the findings to your questions?

Plots such as boxplots,bar charts,Histograms, scatter plots will help in understanding the data distribution. Sample ones are included below. I will ggplot for them in the next step.

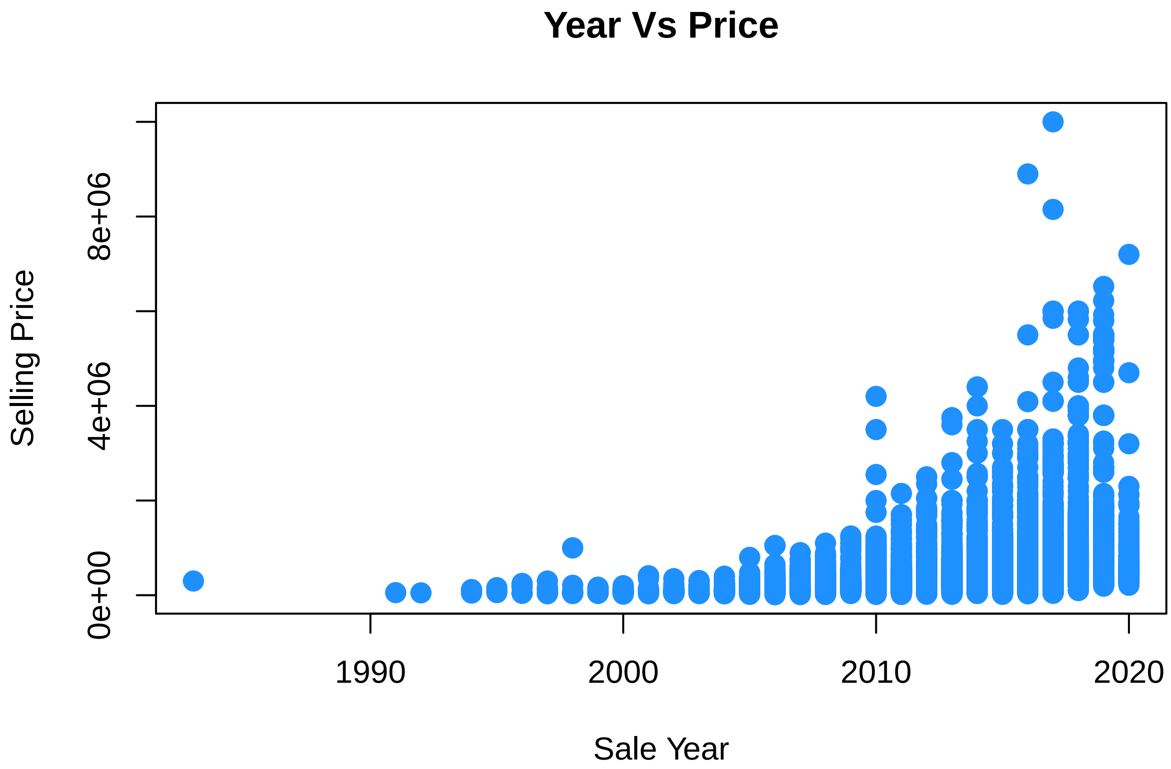
```r
hist(CarData$year,xlab = "Year",main = "Histogram of Year")
```



```
barplot(table(CarData$seller_type),
        xlab   = "Seller Type",
        ylab   = "Frequency",
        main   = "Sellers",
        col    = "dodgerblue",
        border = "darkorange")
```



```
plot(selling_price ~ year,CarData,
     xlab = "Sale Year",
     ylab = "Selling Price",
     main = "Year Vs Price",
     pch  = 20,
     cex  = 2,
     col  = "dodgerblue")
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I'm planning to use linear regression method to predict the prices. I will use simple linear regression method as my first model and then add the predictors one by one to determine the optimal model for the price prediction.

Questions for future steps

1. How to identify the covariance, correlation.
2. Which method best helps us in predicting the prices.
3. What variables to start the linear model with.
4. Can we use multiple regression?
5. If yes, what variables needs to be used for multiple regression.
6. How to compare the models against each other.