- R for Data Analytics
- Presented by: Anna Liza Dela Cruz
- Presented to: Hamid Rajaee

R analytics (or **R** programming language) is a free, open-source software used for all kinds of data science, statistics, and visualization projects. It can be **used for data analysis** and **statistical** modeling. **R** is an environment for **statistical analysis**. **R** has various **statistical** and graphical capabilities. **R** is very important in **data science** because of its versatility in the field of statistics. **R** is usually **used** in the field of **data science** when the task requires special **analysis** of **data** for stand alone or distributed computing. **R** is also perfect for exploration.

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing the data to get a better understanding of the data and glean insight from it.

It is a statistical approach or technique for analyzing data sets in order to summarize their important and main characteristics.

The EDA approach can be used to gather knowledge about the following aspects of data:

Main characteristics or features of the data.

The variables and their relationships.

Finding out the important variables that can be used in our proble

I have my data set, Car details from Car Dekho, https://www.kaggle.com/datasets, it is about the setting or predictions of selling price of each certain cars, based on the type of sellers and owner itself. Which variables are independent to each other.
Before I started my analysis, I checked missing values if it is present in my data, so far there is no missing values, but there is outlier in selling price, when I did summarization, so I used IQR. Checked for duplicated values, there are some, but I did not do anything about it, when I checked the entire data set only in one column, but the other but the entire rows had different values as well, so I decided to keep it for the presentation purpose. I checked the head and tail to get the first 6 and the last 6 observations.

```
125          First Owner
 [ reached 'max' / getOption("max.print") -- omitted 4215 rows ]
> dim(data)
[1] 4340    8
> 
```

I have 8 variables. Selling price is my target variable.

```
 ...
> names(data)
[1] "name"          "year"          "selling_price" "km_driven"
[5] "fuel"          "seller_type"   "transmission"  "owner"
> 
```

- **Univariate analysis** is the simplest form of data **analysis** where the data being **analyzed** contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of **univariate analysis** is to describe the data and find patterns that exist within it.

```
> data_org1<-data
> summary(data$selling_price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20000  208750  350000  504127  600000 8900000 
> 
```

Univariate Analysis

The target column is price which is a numerical column, I summarize it by getting the five-number summary, as you notice there is a large interval with 3rd quarter and the maximum range. So, I will use the IQR.

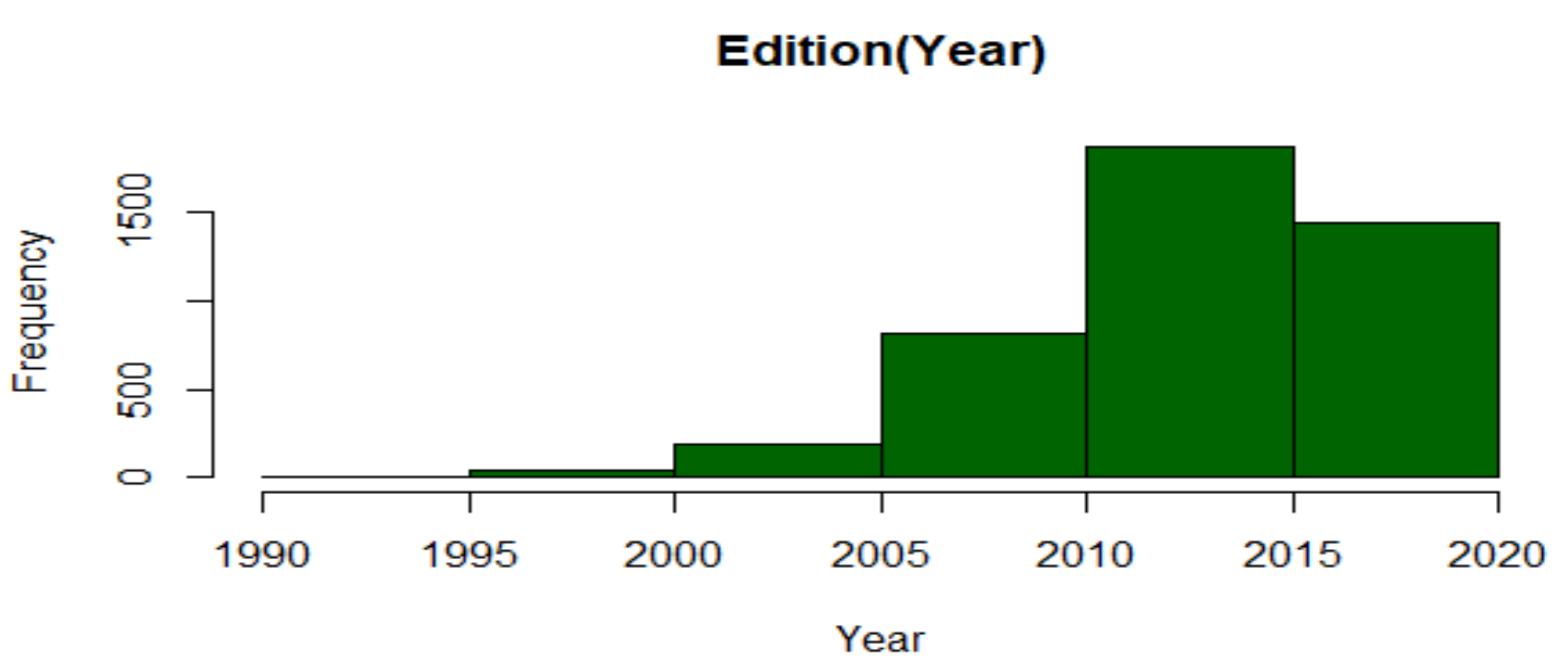
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20000  208750  350000  504127  600000 8900000 
> data_org2<-data
> IQR_dataRev <-600000-208750
> IQR_dataRev
[1] 391250
> Up_dataRev <-600000+1.5*IQR_dataRev
> Up_dataRev
[1] 1186875
```

```
> data$selling_price<-ifelse(data$selling_price>1186875, data$selling_price/5,
 data$selling_price)
> summary(data["selling_price"])
 selling_price    
 Min.   :  20000  
 1st Qu.: 208750  
 Median : 340000  
 Mean   : 395128  
 3rd Qu.: 550000  
 Max.   :1780000  
> 
```


Selling Price
Frequency
0
500
1000
1500
0
500000
1000000
1500000
Price

Univariate Analysis on Year

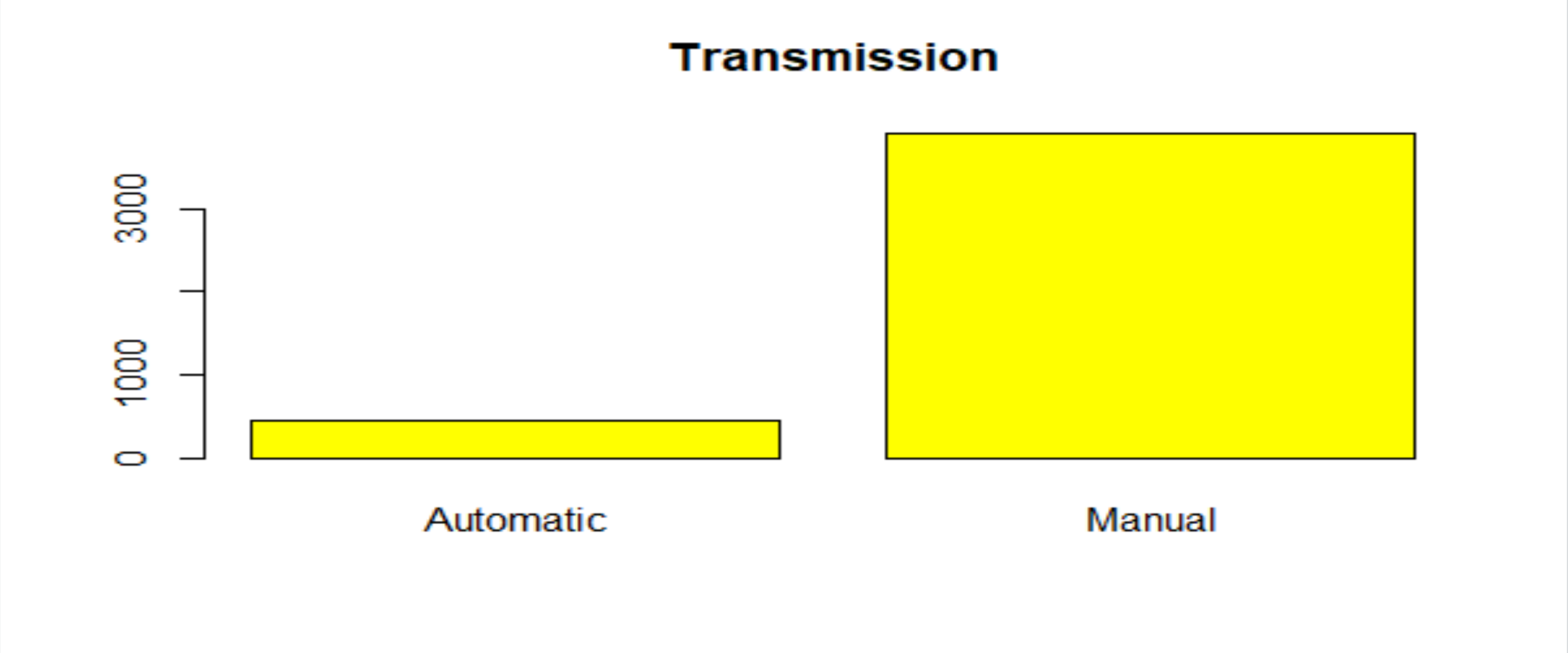
```
> summary(data$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1992    2011    2014    2013    2016    2020 
> hist(data$year, breaks = 5, main = "Edition(Year)",col="darkgreen", xlab="Ye
ar",ylab="Frequency")
> 
```



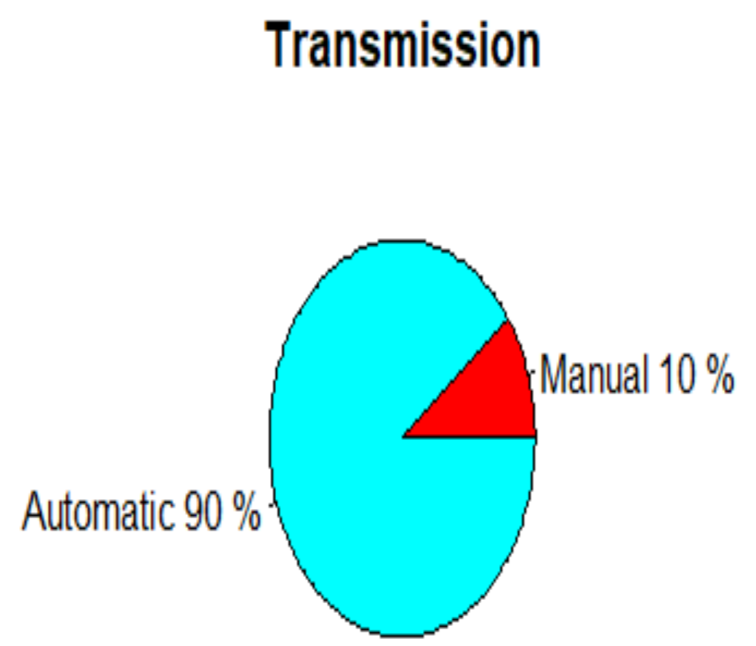
Transmission

Frequency of each levels on Transmission

```
> count<-table(data$transmission)
> count

Automatic    Manual
      448      3892
> barplot(c(count[1], count[2]), main="Transmission",
+         col = 'yellow',horiz = FALSE)
> 
```



```
+                    ind Owner"), las=2, main = "Owner", col = "orange")
> count<-table(data$transmission)
> count

Automatic    Manual
      448      3892
> freq1 <- c(count[1], count[2])
> lbls <- c("Manual", "Automatic")
> pct <- round(freq1/sum(freq1)*100)
> lbls <- paste(lbls, pct) # add percents to labels
> lbls <- paste(lbls,"%",sep=" ") # ad % to labels
> pie(freq1,labels = lbls, col=rainbow(length(lbls)),
+     main="Transmission")
>
```

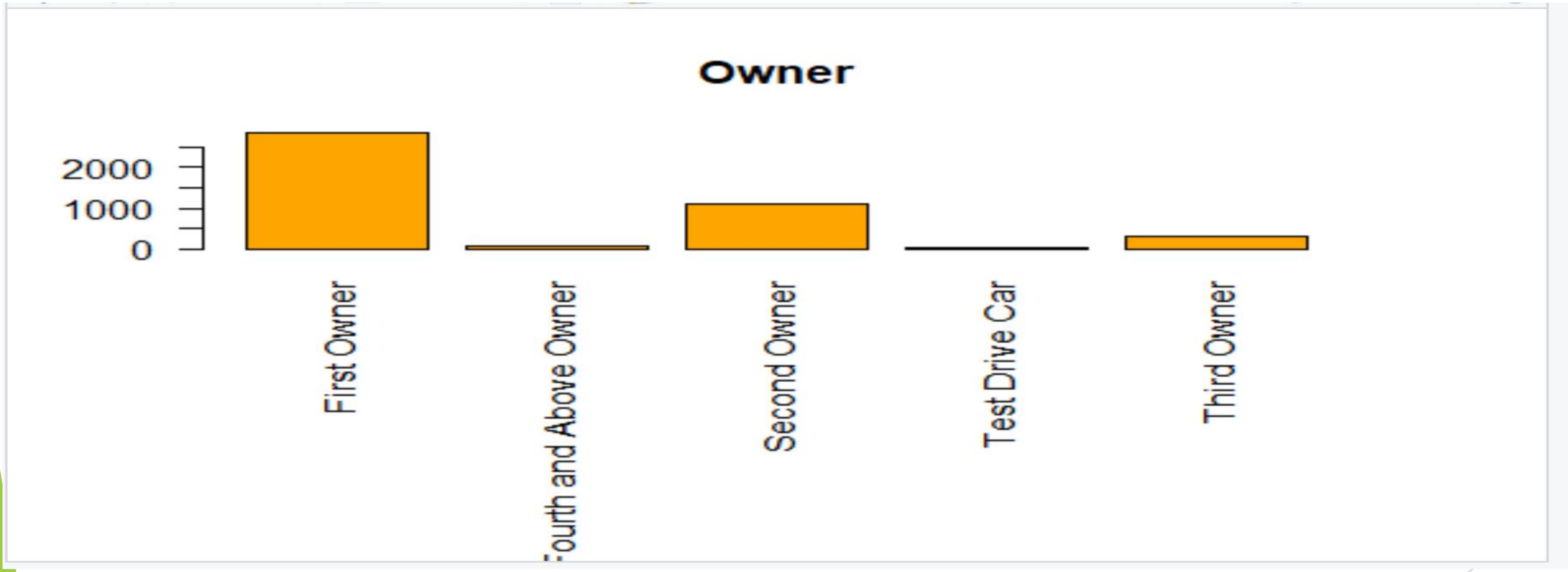


Owner

Frequency of each levels on Owner

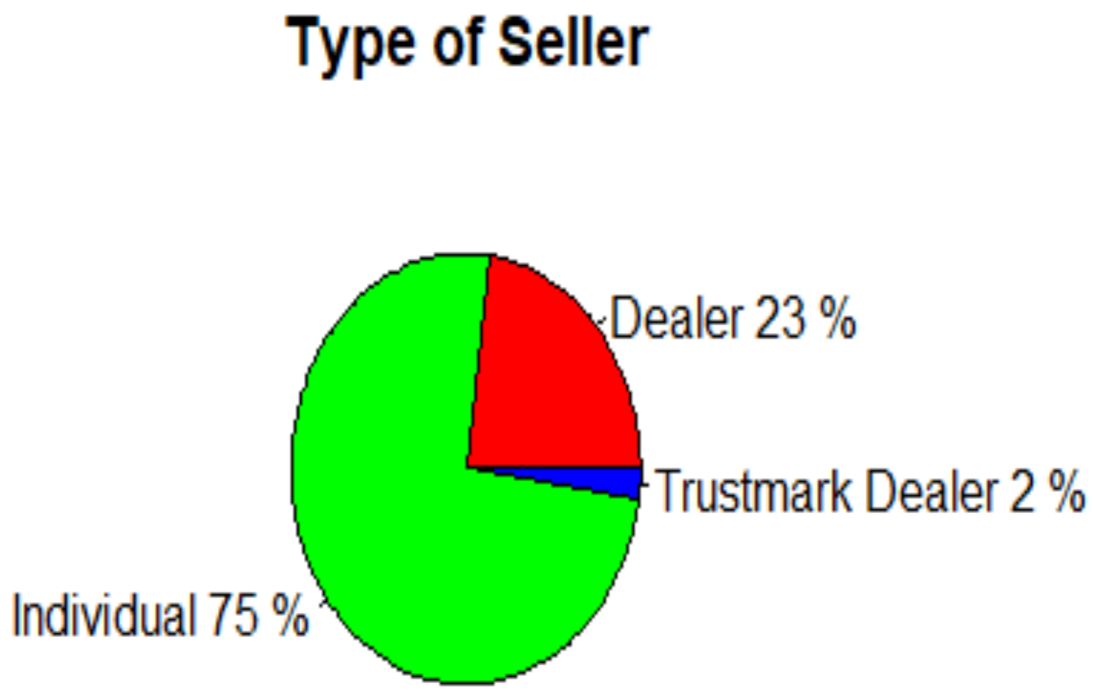
```
> count<-table(data$owner)
> count

          First Owner Fourth & Above Owner        Second Owner
                 2832                   81                1106
       Test Drive Car          Third Owner
                   17                  304
>
```

```
> barplot(c(count[1], count[2], count[3], count[4], count[5]),
+         names.arg=c("First Owner","Fourth and Above Owner",
+                     "Second Owner","Test Drive Car",
+                     "Third Owner"), las=2, main = "Owner", col = "orange" )
>
```

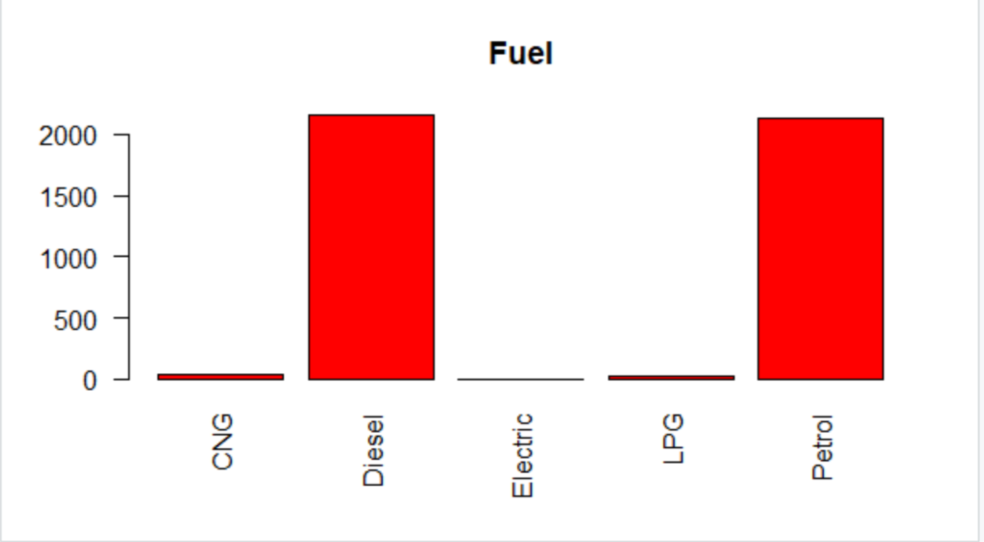


Seller Type
Frequency of each levels

```
       Dealer      Individual Trustmark Dealer
          994            3244            102
> freq1 <- c(count[1], count[2], count[3] )
> lbls <- c("Dealer", "Individual", "Trustmark Dealer")
> pct <- round(freq1/sum(freq1)*100)
> lbls <- paste(lbls, pct)
> lbls <- paste(lbls,"%",sep=" ")
> pie(freq1,labels = lbls, col=rainbow(length(lbls)),
+     main="Type of Seller")
>
```



```
> count<-table(data$fuel)
> count

     CNG    Diesel Electric       LPG    Petrol
      40      2153        1        23      2123
> barplot(c(count[1], count[2], count[3], count[4], count[5]),
+         names.arg=c("CNG","Diesel",
+                     "Electric","LPG",
+                     "Petrol"), las=2, main = "Fuel", col = "red" )
> 
```



- Fuel
- Frequency of each levels on Fuel

Bivariate Analysis

Bivariate analysis is when you are studying two variables. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values.

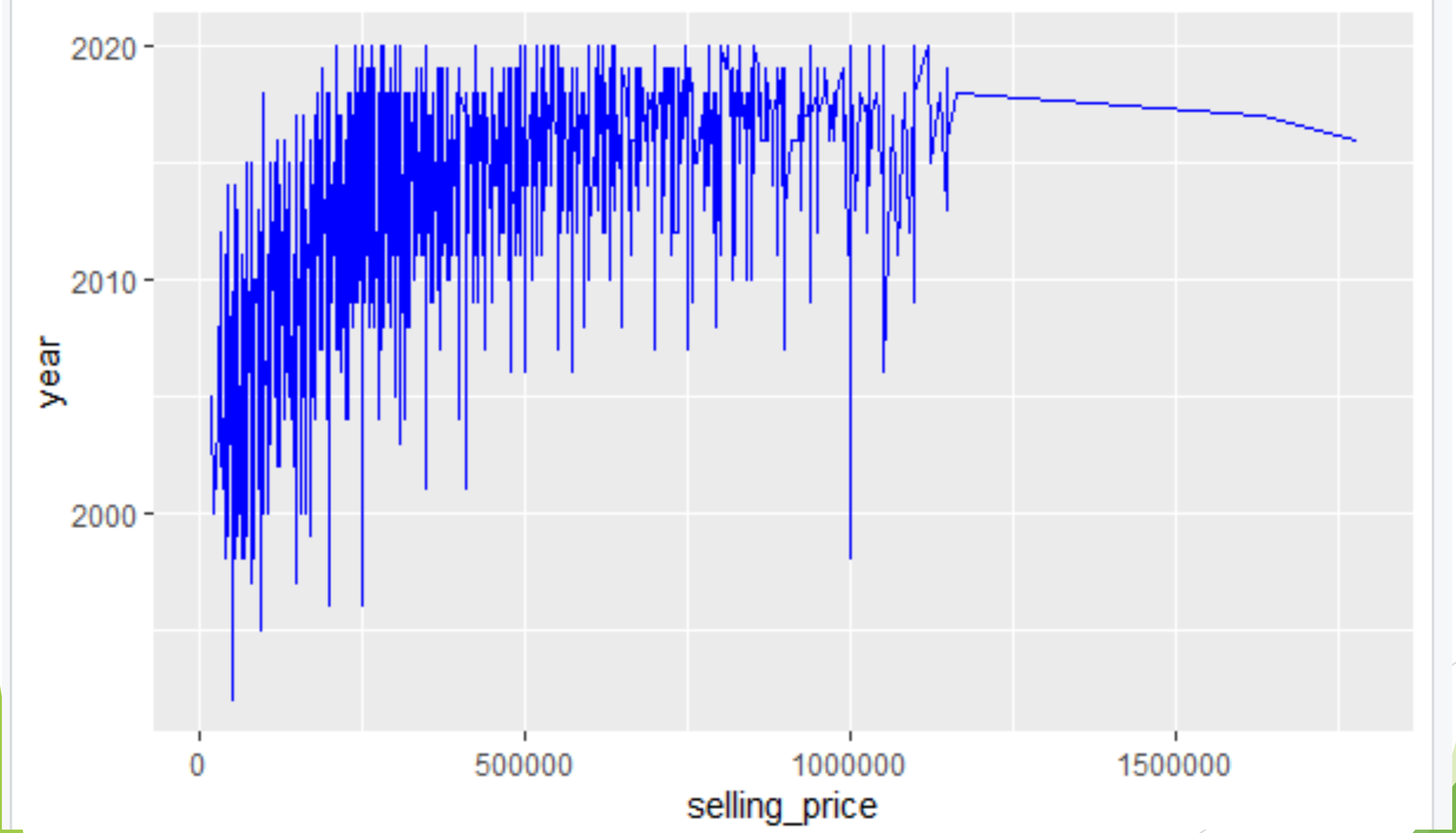
Continuous vs. Continuous

```
> #Continuous vs. Continuous
> sapply(data[,-c(1, 4:8)], quantile, na.rm=T)
     year selling_price
0%   1992       20000.0
25%  2011      208749.8
50%  2014      340000.0
75%  2016      550000.0
100% 2020     1780000.0
> cor(data$year,data$selling_price)
[1] 0.6050335
> 
```

```
> ggplot(data, aes(x=selling_price, y=year)) + geom_line(col= "blue")
> 
```

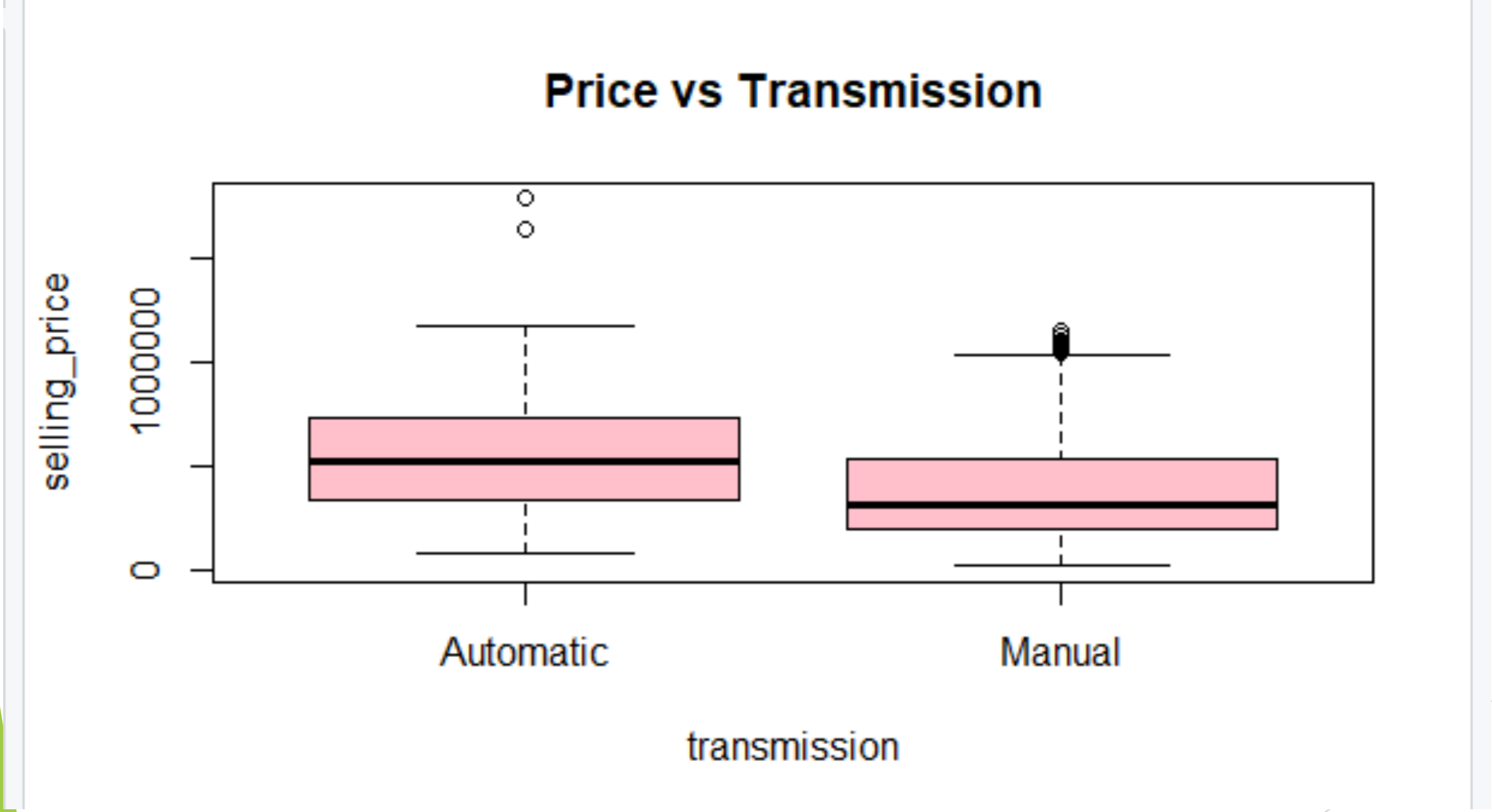
Type here to search

Price vs. Year



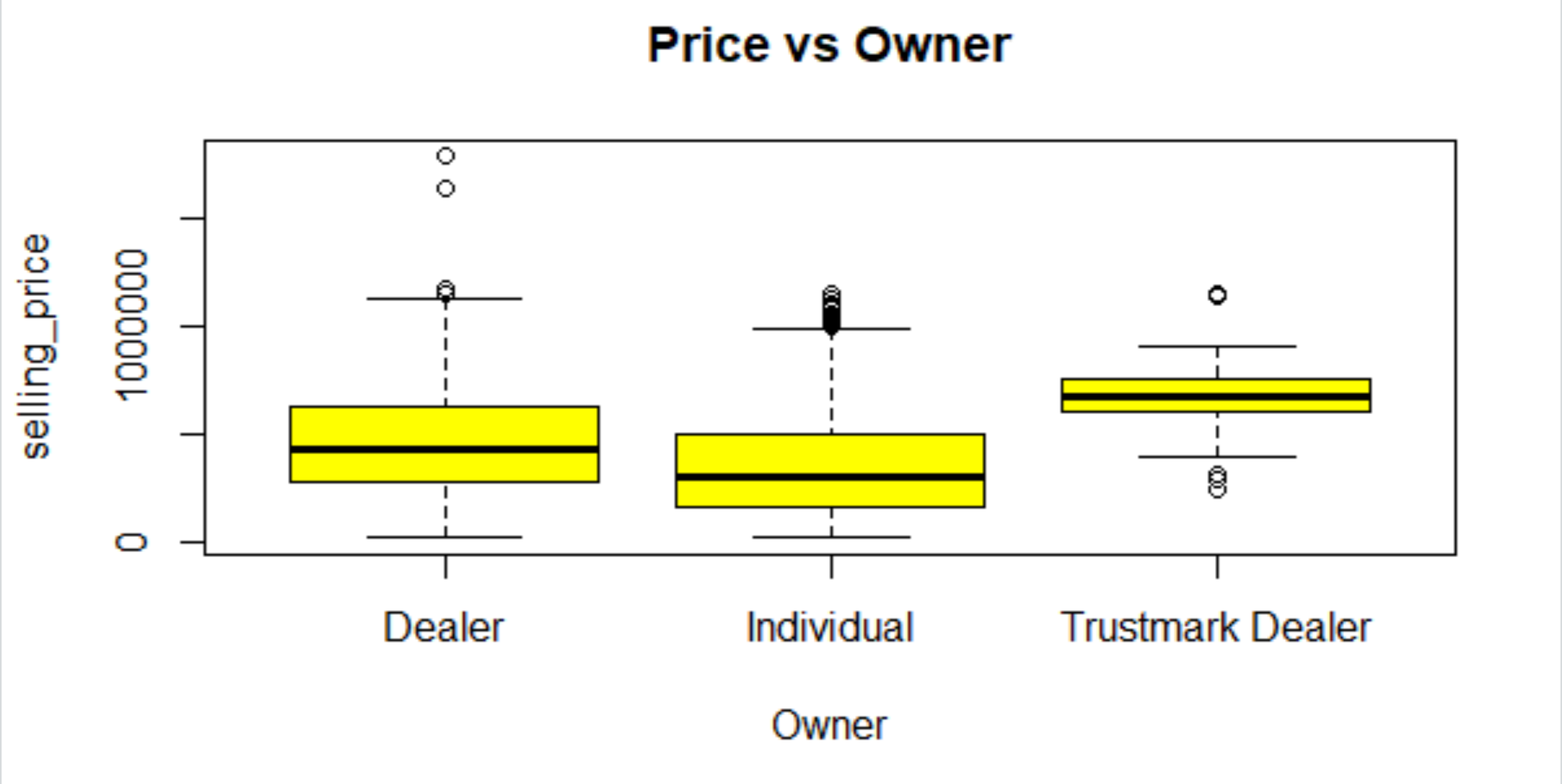
Continuous vs. Categorical

```
+      pch=21, cex=1.5)
> summaryBy(selling_price ~ transmission, data, FUN= quantile,na.rm=T)
  transmission selling_price.0% selling_price.25% selling_price.50%
1    Automatic            79000            340000            520000
2       Manual            20000            194500            320000
  selling_price.75% selling_price.100%
1            735000            1780000
2            530000            1151000
> 
```


Price vs Transmission
selling_price
0
1000000
Automatic
Manual
transmission

Test of Independency on Selling_Price vs. Transmission

```
2              530000              1151000
> t.test(selling_price~transmission, data=data)

        Welch Two Sample t-test

data:  selling_price by transmission
t = 13.289, df = 532.86, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 149213.3 200981.8
sample estimates:
mean in group Automatic    mean in group Manual
               552150.4                377052.9

> 
```



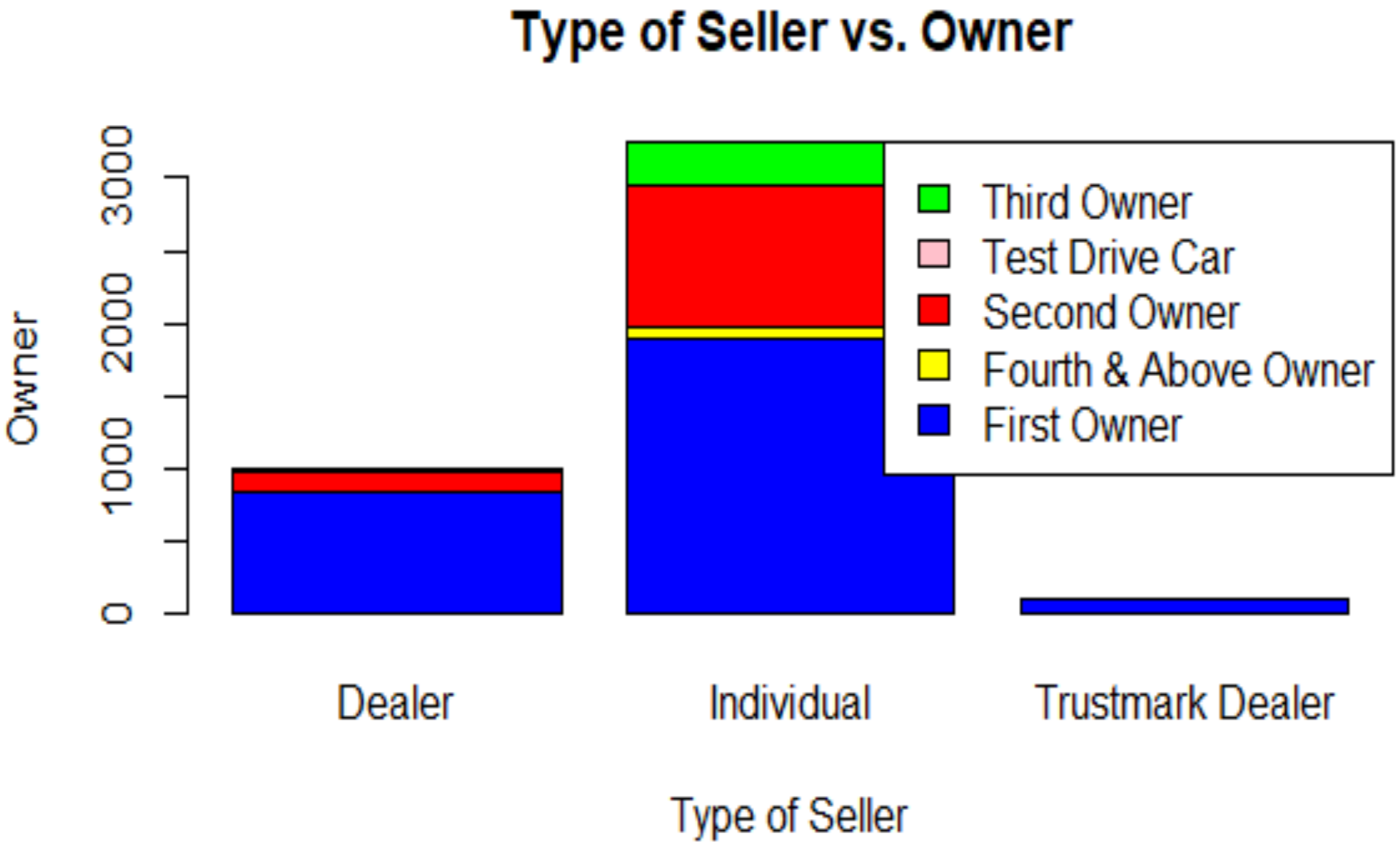
Price vs Owner
selling_price
0
10000000
Dealer
Individual
Trustmark Dealer
Owner

ANOVA test for Selling Price vs. Owner

```
> one.way <- aov(selling_price ~ owner, data = data)
> summary(one.way)
              Df    Sum Sq   Mean Sq F value Pr(>F)    
owner          4 3.070e+13 7.674e+12   141.1 <2e-16 ***
Residuals   4335 2.357e+14 5.438e+10                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Categorical vs. Categorical

```
                  First Owner Fourth & Above Owner Second Owner
  Dealer                  844                    2          122
  Individual             1890                   79          980
  Trustmark Dealer         98                    0            4

                  Test Drive Car Third Owner
  Dealer                      17           9
  Individual                   0         295
  Trustmark Dealer             0           0
> chisq.test(tbl)

	Pearson's Chi-squared test

data:  tbl
X-squared = 372.78, df = 8, p-value < 2.2e-16
```



Type of Seller vs. Owner
Owner
0
1000
2000
3000
Third Owner
Test Drive Car
Second Owner
Fourth & Above Owner
First Owner
Dealer
Individual
Trustmark Dealer
Type of Seller

Based on the analysis that I did, I figure out that the selling price as my target variable on this analysis has a strong relationship with the owner as it clearly shows when I did my Analysis of Variance on these variables which are selling price and owner. And after getting through with chi-squared test of type of seller against owner, it clearly shows that they also have a strong relationship with each other.

Some variables are independent to each other and does not have any relationship with the target variable which is the selling price, while some have strong relationship to it. I therefore conclude that through EDA, we can analyze and visualize the relationship of each variable and we are able to know which variable is independent to each other.

I therefore recommend that when setting price on each specific cars, price vs. owner vs. seller type has 5% significant to each other according to t-test and ANOVA test that I did. If you want to have a thorough analysis to which variable is independent to each other, running t-test and ANOVA is a very important type to each other. Car dealers set their price according to who sold each cars.