Exploratory Data Analysis (EDA) of Used Cars

Objective of the project

To analyze and visualize the data using EDA to gather insights about the data set and understand the diversity in the data and the range of every field. Various visulalization charts like bar chart, box plot, distribution graph, etc. used to explore how each feature varies and its relation with other features

Data Information

https://data.world/data-society/used-cars-data

Installing packages and Reading data(Place data file in R working directory)

```
library(data.table)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
auto<-fread("autos1.csv", stringsAsFactors = T)
```

Data Cleaning

```
auto$nrOfPictures <- NULL #Delete not required columns
auto$seller <- NULL
auto$offerType <- NULL
auto <- auto[price<150000&price>60] # Price between 60 and 150000 dollars
auto <- auto[yearOfRegistration>=1863&yearOfRegistration<2017]
auto <- auto[powerPS>0&powerPS<1100]
nom <- strsplit(as.character(auto$name),split = "_")
auto$model <- as.factor(sapply(nom,"[[",1))
```

Data Summary

```
summary(auto)
```

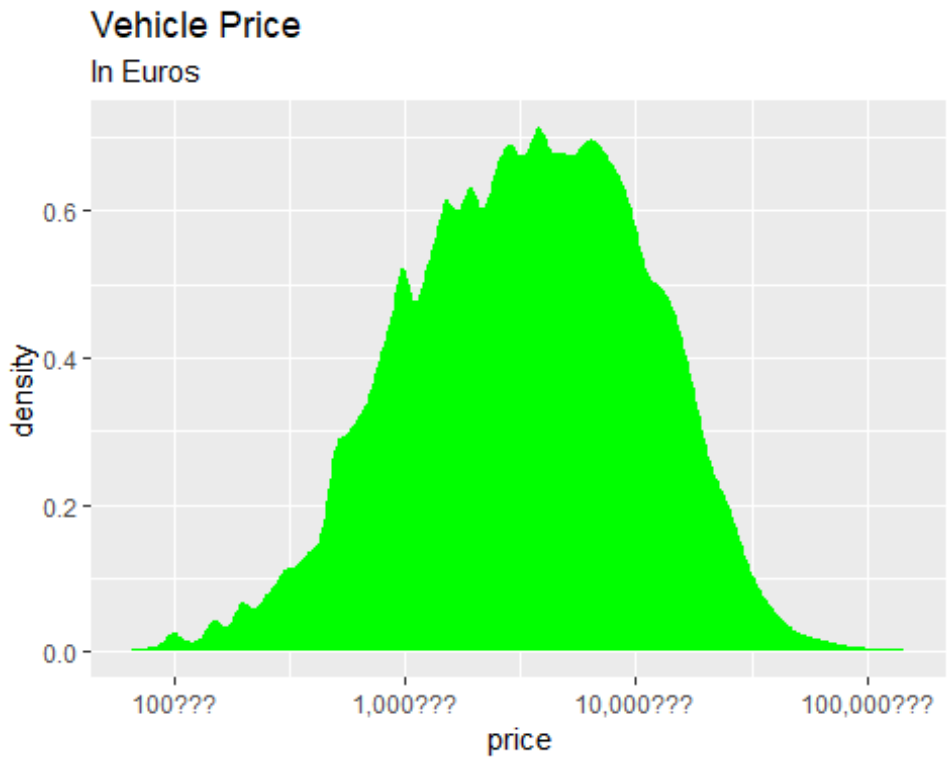
```
##            dateCrawled                          name            price          
##  05-03-2016 14:25:    61   BMW_318i                :   621   Min.   :    65  
##  05-03-2016 14:26:    53   Volkswagen_Golf_1.4:   581   1st Qu.:  1450  
##  05-03-2016 15:48:    53   BMW_316i                :   518   Median :  3500  
##  05-03-2016 17:49:    51   Ford_Fiesta             :   515   Mean   :  6260  
##  07-03-2016 16:50:    51   BMW_320i                :   488   3rd Qu.:  7999  
##  03-04-2016 16:49:    50   Opel_Corsa              :   463   Max.   :149999  
##  (Other)         :311656   (Other)                 :308789                   
##      abtest           vehicleType    yearOfRegistration
##         :     0   limousine :87245   Min.   :1910      
##  benzin :     0   kleinwagen:70054   1st Qu.:1999      
##  control:150364   kombi     :61491   Median :2003      
##  test   :161611   bus       :27558   Mean   :2003      
##                   cabrio    :21323   3rd Qu.:2008      
##                   coupe     :17206   Max.   :2016      
##                   (Other)   :27098                     
##               gearbox          powerPS              model       
##                   :  5416   Min.   :   1.0   Volkswagen: 35843  
##  25-03-2016 00:00:     0   1st Qu.:  80.0   BMW       : 29451  
##  automatik       : 69265   Median : 116.0   Opel      : 27268  
##  manuell         :237294   Mean   : 126.6   Mercedes  : 26118  
##                            3rd Qu.: 150.0   Audi      : 25975  
##                            Max.   :1090.0   Ford      : 18172  
##                                             (Other)   :149148  
##    kilometer      monthOfRegistration    fuelType      
##  150000 :198341   Min.   : 0.00       benzin :194162  
##  125000 : 33314   1st Qu.: 3.00       diesel : 96550  
##  100000 : 13811   Median : 6.00              : 15611  
##  90000  : 11228   Mean   : 6.01       lpg    :  4744  
##  80000  : 10002   3rd Qu.: 9.00       cng    :   486  
##  70000  :  8946   Max.   :12.00       hybrid :   245  
##  (Other): 36333                       (Other):   177  
##           brand       notRepairedDamage           dateCreated     
##  volkswagen   :66442       : 43281       03-04-2016 00:00: 12369  
##  bmw          :35406   ja  : 28871       04-04-2016 00:00: 11857  
##  opel         :32599   nein:239823       20-03-2016 00:00: 11497  
##  mercedes_benz:30379                     12-03-2016 00:00: 11407  
##  audi         :28920                     21-03-2016 00:00: 11089  
##  ford         :21016                     28-03-2016 00:00: 11033  
##  (Other)      :97213                     (Other)         :242723  
##    postalCode               lastSeen       
##  Min.   : 1067   07-04-2016 06:45:   623  
##  1st Qu.:31167   07-04-2016 06:16:   610  
##  Median :50765   06-04-2016 04:45:   607  
##  Mean   :51568   07-04-2016 07:16:   594  
##  3rd Qu.:72458   07-04-2016 08:45:   585
```

```
##  Max.   :99998   07-04-2016 05:45:   584
##                  (Other)         :308372
```

Smoothed version histograms

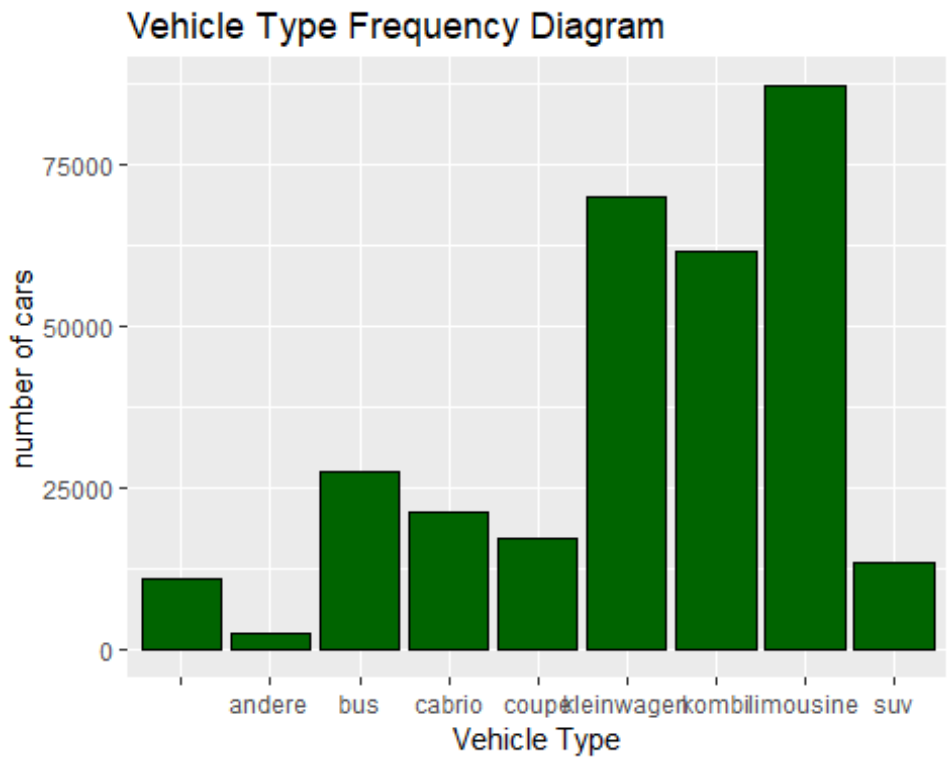
Vehicle Price

```
ggplot(auto,aes(price))+
  stat_density(fill="green")+scale_x_log10(labels =
scales::dollar_format(suffix = "???", prefix = ""))+
  labs(title="Vehicle Price",subtitle="In Euros")
```

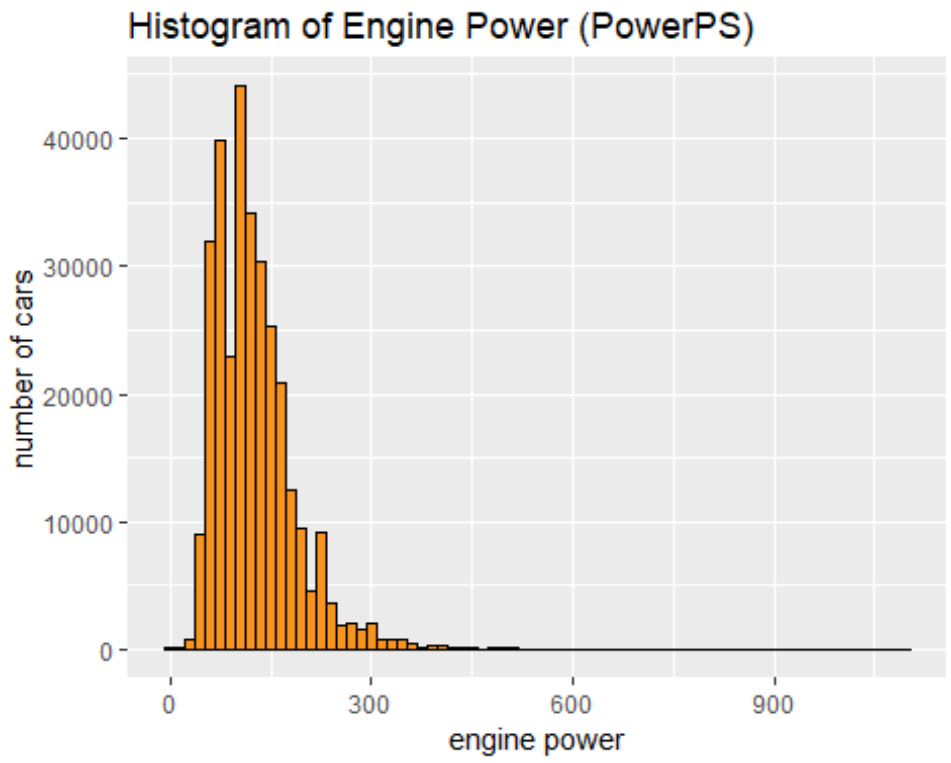


#Vehicle Type Frequency Diagram

```
ggplot(auto, aes(x=vehicleType)) +
  geom_bar(fill='darkgreen', color='black') +
  scale_fill_brewer(type= 'div') +
  labs(x= 'Vehicle Type', y= 'number of cars') +
  ggtitle('Vehicle Type Frequency Diagram')
```

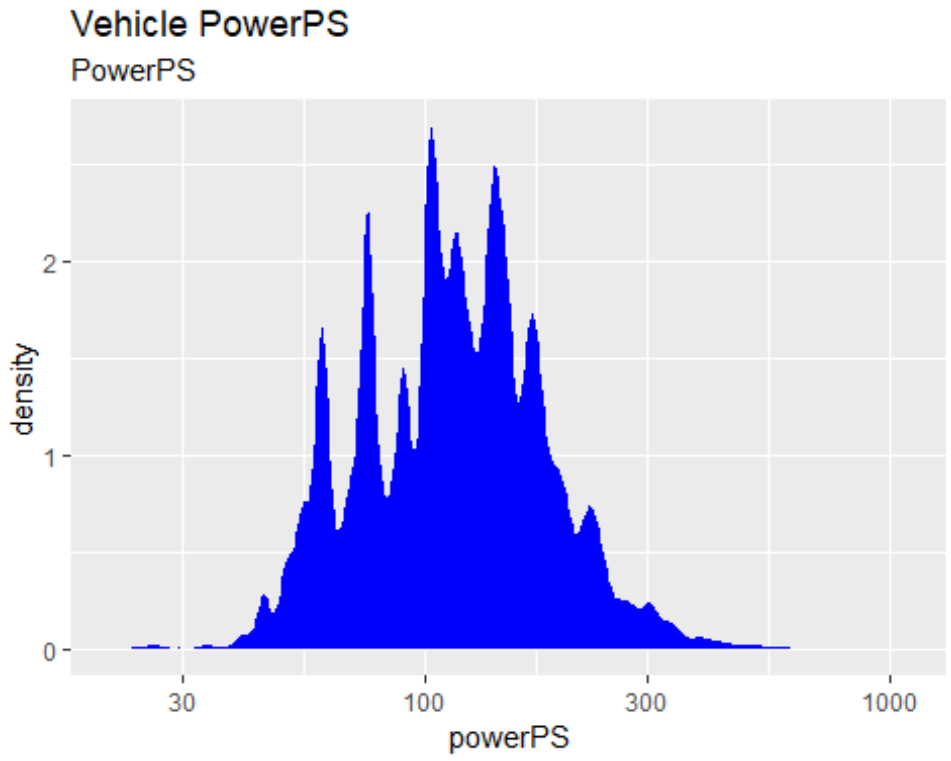



```r
ggplot(auto, aes(auto$powerPS)) +
  geom_histogram(fill= I('#F79420'), color='black', binwidth=15) +
  labs(x= 'engine power', y= 'number of cars') +
  ggtitle('Histogram of Engine Power (PowerPS)')
```



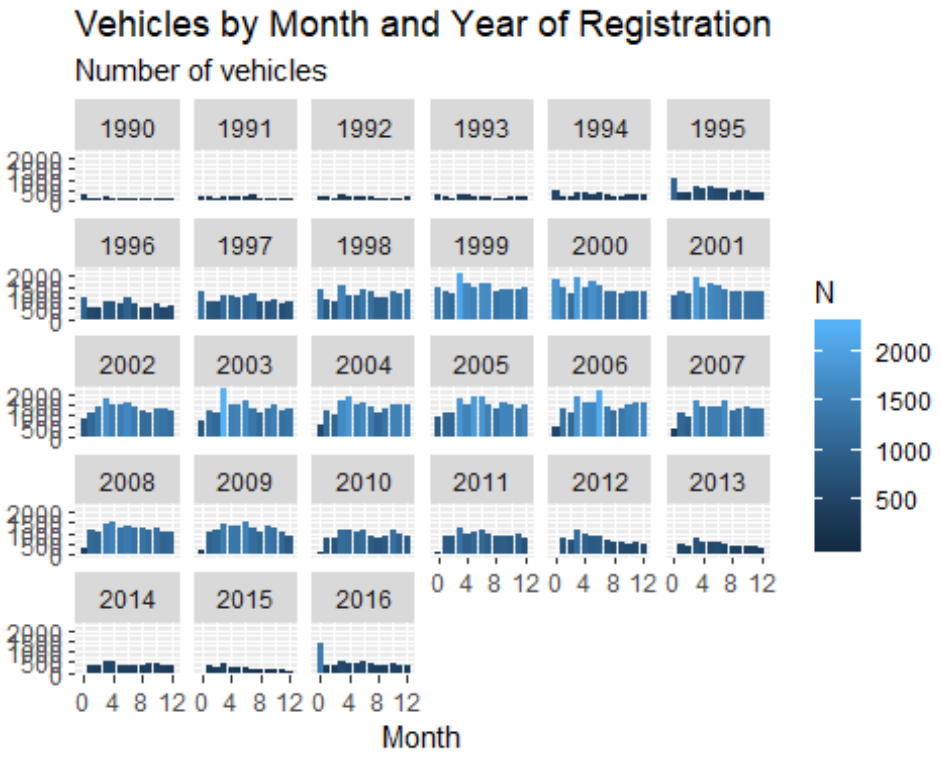
PowerPS

```
ggplot(auto[powerPS>20],aes(powerPS))+
  stat_density(fill="blue")+scale_x_log10()+
  labs(title="Vehicle PowerPS",subtitle="PowerPS")
```



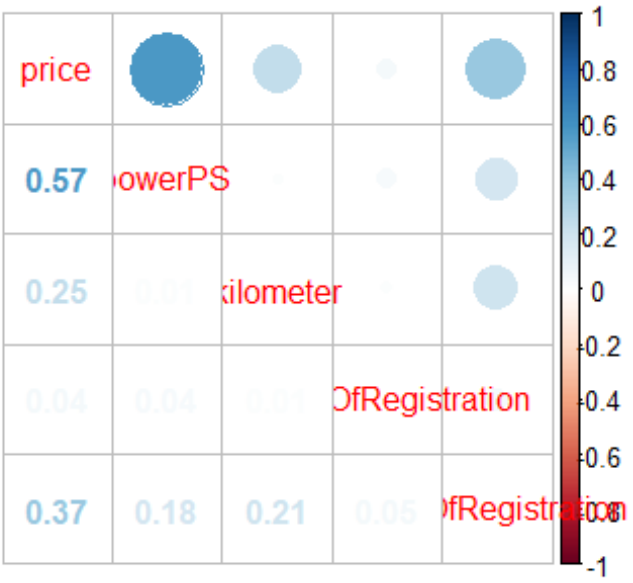
#Vehicles by month and year of registration

```
ggplot(auto[yearOfRegistration>1989,.N,by=.(monthOfRegistration,yearOfRegistration)],
       aes(x = monthOfRegistration,y = N,fill=N))+
  geom_bar(stat = "identity")+labs(title="Vehicles by Month and Year of Registration",
                                   subtitle="Number of vehicles")+
  xlab("Month")+ylab(NULL)+facet_wrap(~yearOfRegistration)
```



#Correlation plot

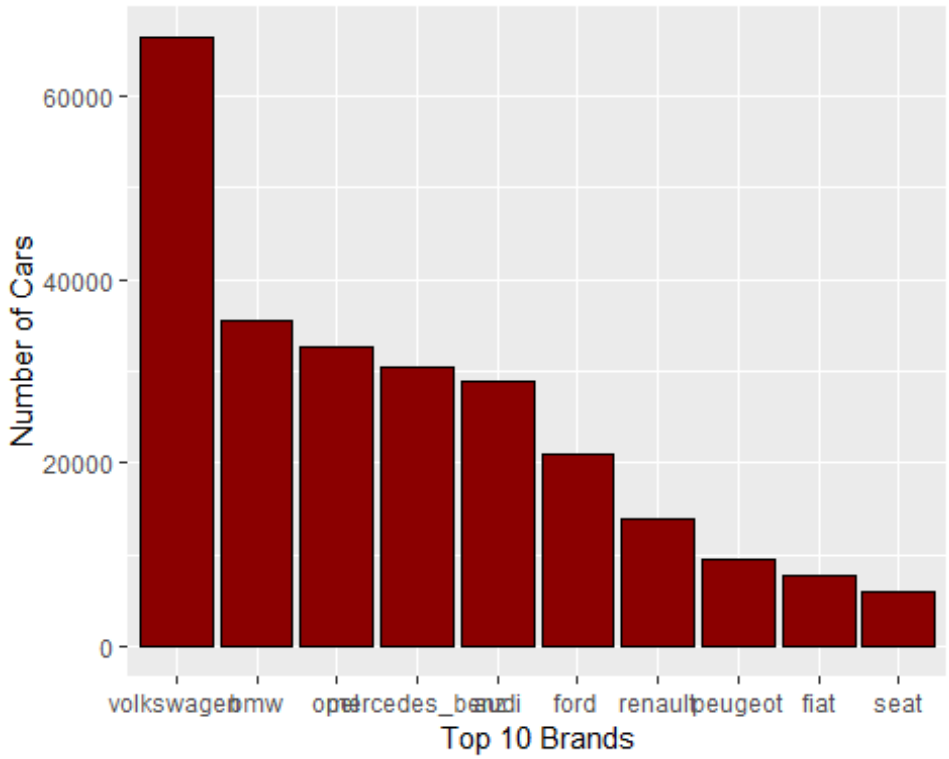
```
## corrplot 0.84 loaded
```



#Top 10 Brands

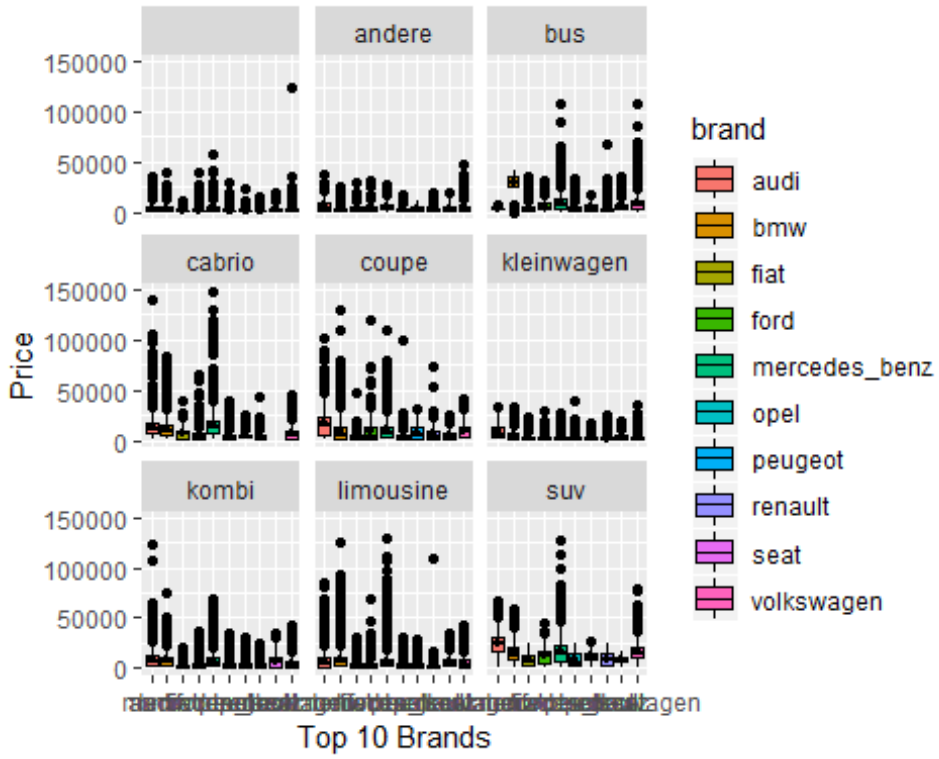
Frequency Diagram

```
auto_subbrand <- subset (auto, brand %in% c("seat", "fiat","peugeot"
,"renault", "ford", "audi", "mercedes_benz" , "opel", "bmw" , "volkswagen"))
ggplot(auto_subbrand, aes(x = reorder(brand, -table(brand)[brand]))) +
  geom_bar(color='black', fill= 'darkred') +
  labs(x= 'Top 10 Brands', y= 'Number of Cars', ggtitle= 'Top 10 Brands
Frequency Diagram')
```

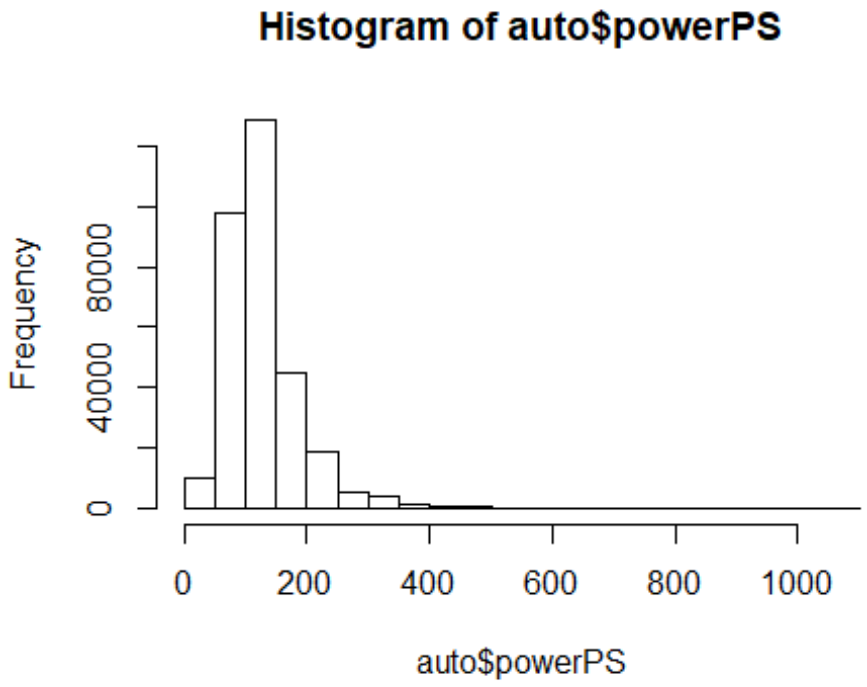


Price vs. Top 10 Brands by Vehicle Type

```
ggplot(auto_subbrand, aes(x=brand, y= price)) +
  geom_boxplot(aes(fill= brand), color= 'black') +
  stat_summary(fun.y = mean, geom="point", size=1) +
  facet_wrap(~vehicleType) +
  labs(x= 'Top 10 Brands', y= 'Price', ggtitle= 'Price vs. Top 10 Brands by
Vehicle Type')
```



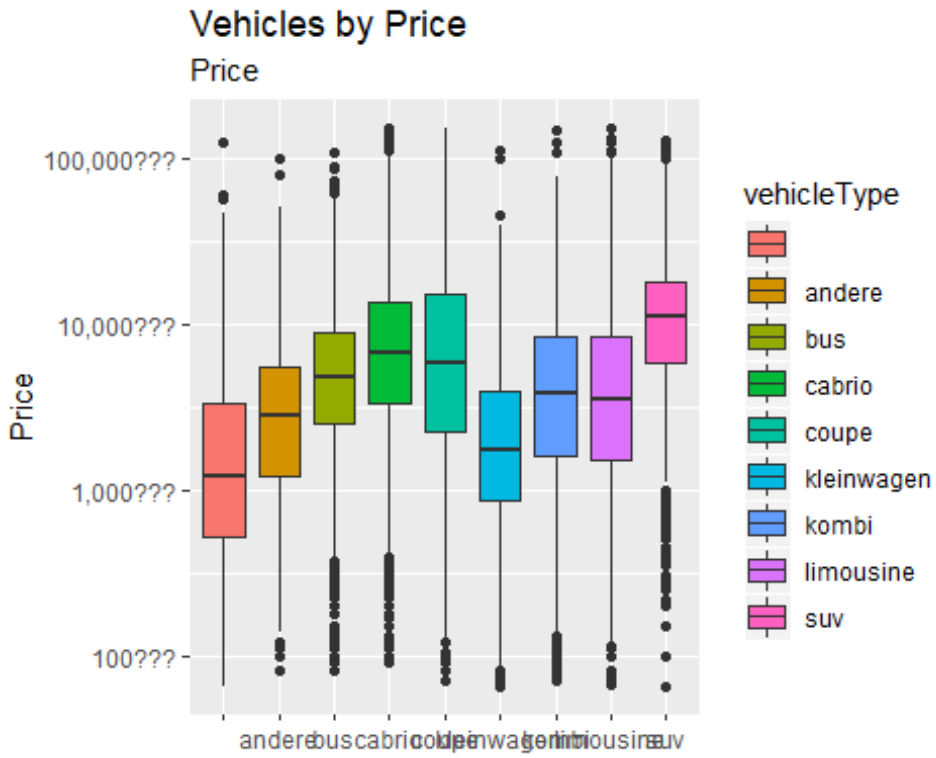

```r
hist(auto$powerPS)
```



Plot Price by...{.tabset}

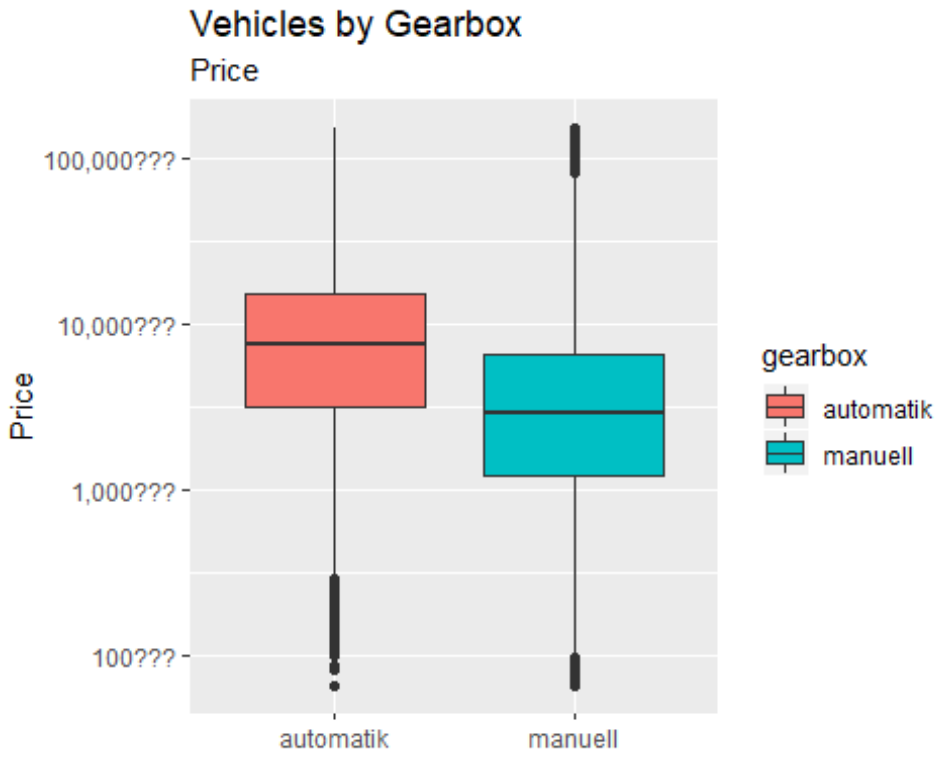
Vehicle Type

```
ggplot(auto,aes(y = price,x=vehicleType,fill=vehicleType))+
  geom_boxplot()+labs(title="Vehicles by
Price",subtitle="Price")+xlab(NULL)+ylab("Price")+scale_y_log10(labels =
scales::dollar_format(suffix = "???", prefix = ""))
```



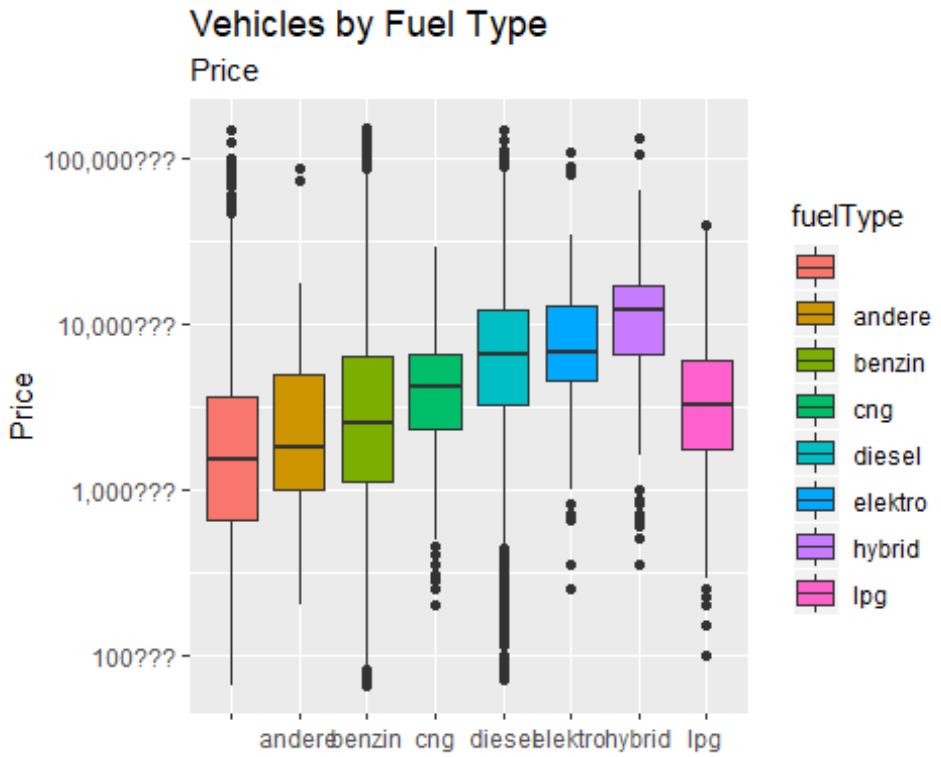
Gearbox

```
ggplot(auto[gearbox%in%c("automatik","manuell")],aes(y =
price,x=gearbox,fill=gearbox))+
  geom_boxplot()+labs(title="Vehicles by
Gearbox",subtitle="Price")+xlab(NULL)+ylab("Price")+scale_y_log10(labels =
scales::dollar_format(suffix = "???", prefix = ""))
```



Fuel Type

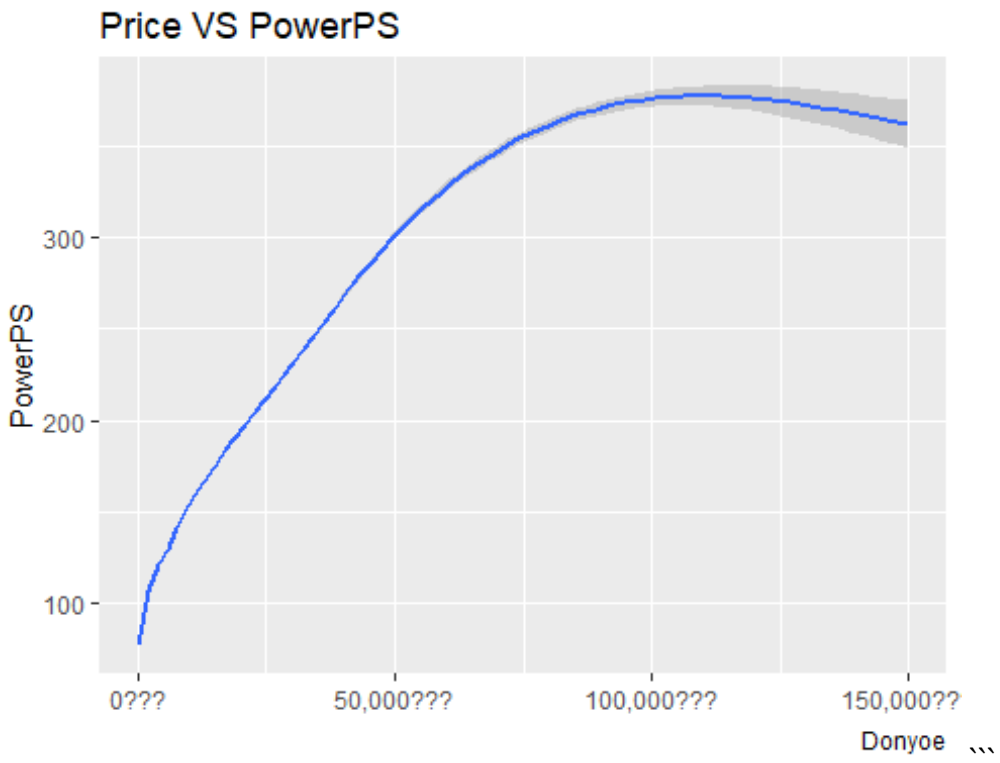
```
ggplot(auto,aes(y = price,x=fuelType,fill=fuelType))+
  geom_boxplot()+labs(title="Vehicles by Fuel
Type",subtitle="Price")+xlab(NULL)+ylab("Price")+scale_y_log10(labels =
scales::dollar_format(suffix = "???", prefix = ""))
```



#Price VS powerPS

```
ggplot(auto,aes(x = price,y = powerPS))+geom_smooth()+
  labs(title="Price VS
PowerPS",caption="Donyoe")+xlab(NULL)+ylab("PowerPS")+scale_x_continuous(labe
ls = scales::dollar_format(suffix = "???", prefix = ""))

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


Price VS PowerPS
PowerPS
300
200
100
0???
50,000???
100,000???
150,000??
Donyoe
```
```