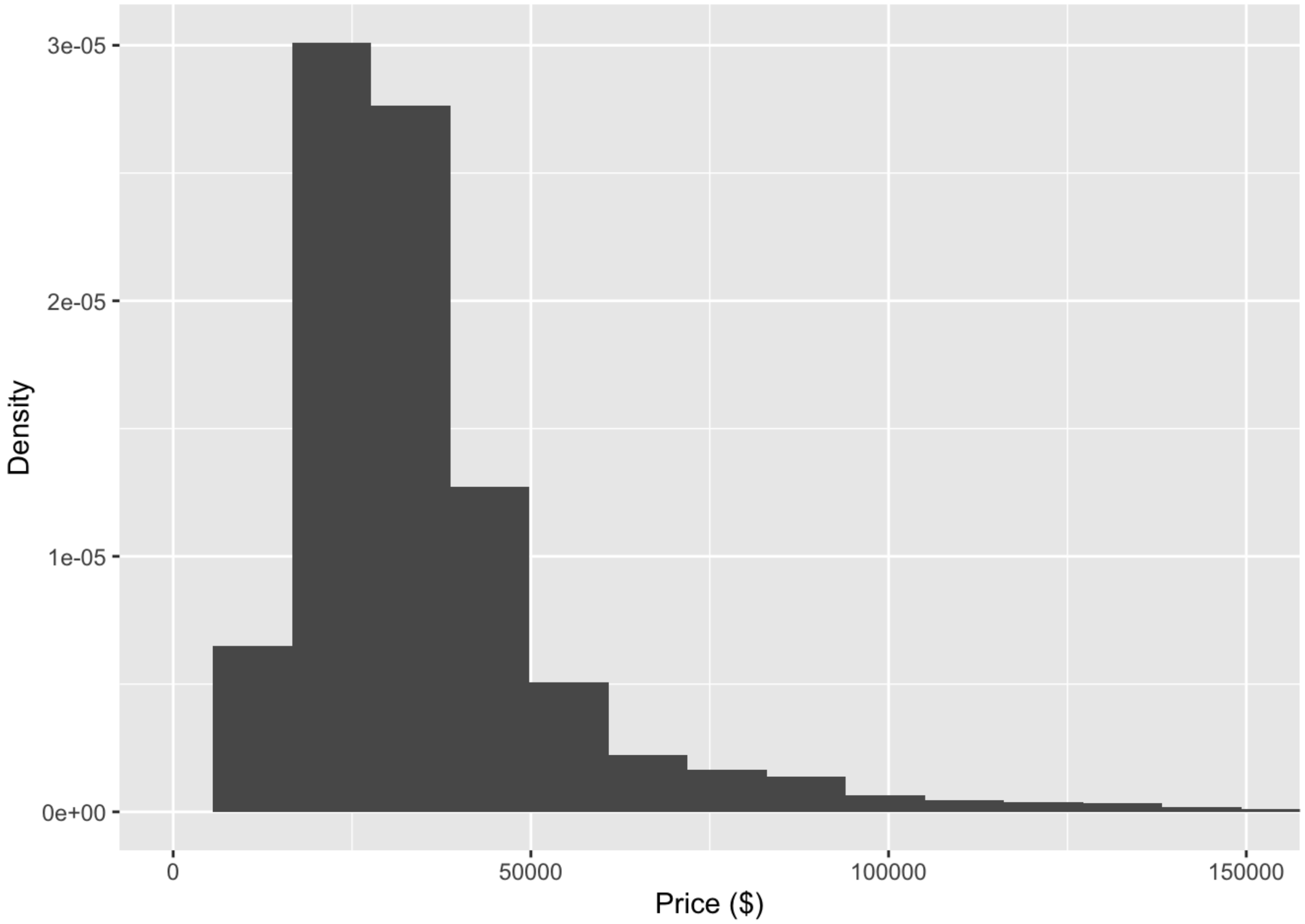
UsedCar_Pricing

```
#load data
library(tidyverse)
```

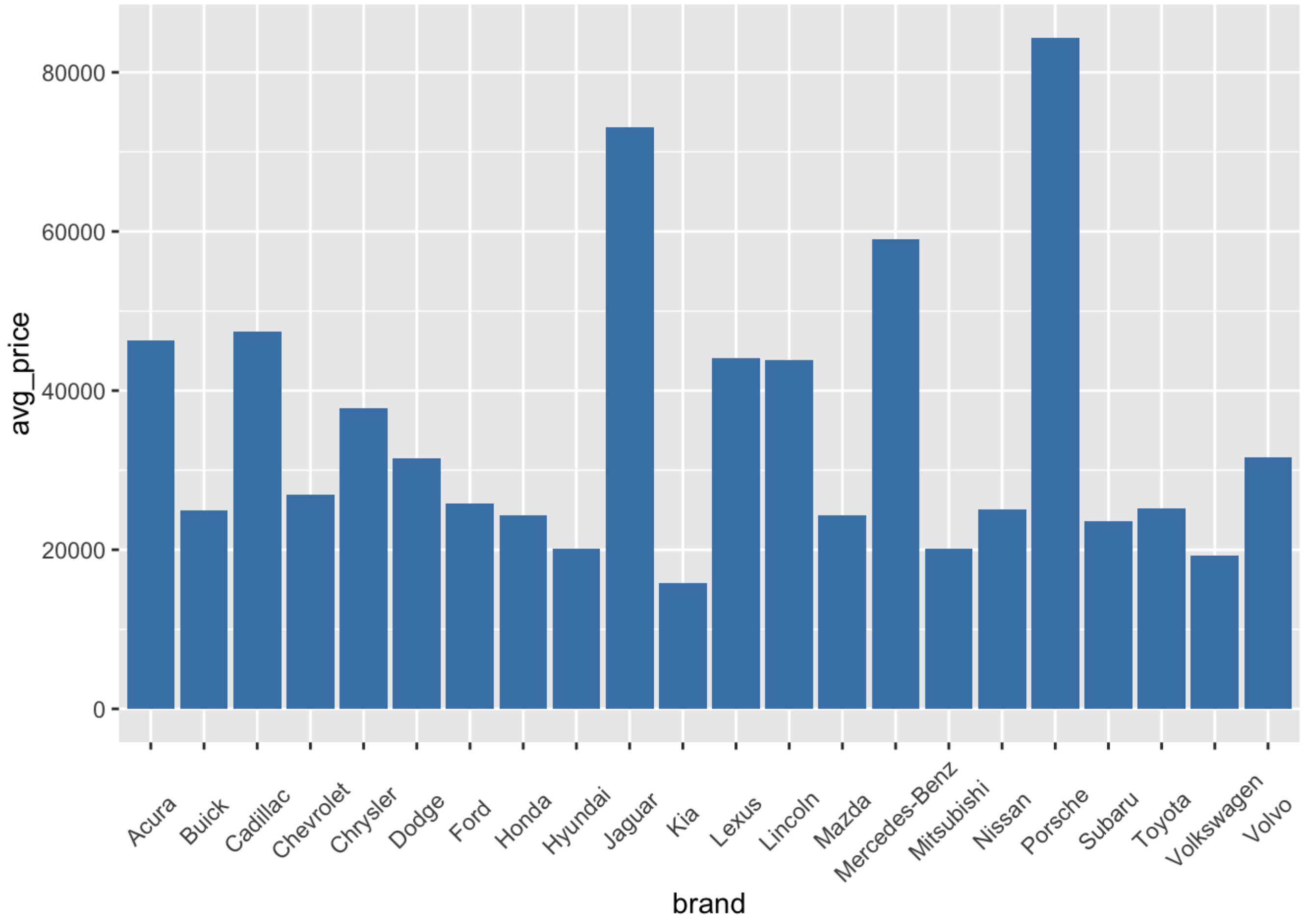
```
Introduction: The used car market has been active in recent years. The rapid development
of the used car market has quietly changed the way of buying behavior of customers. Peop
le who want to buy the used car prefer to buying the type of cars that have high price m
aintenance rate so that they can resell the cars without losing too much money. Does bra
nd actually affect the price maintenance rate? Are there other factors have an impact? S
o, what are the factors influence price maintenance rate, indeed? We will try to use lin
ear regression model to answer this question.
```

```
library(ggplot2)

# Our analysis includes 3 main parts:
# 1 — Data Characteristics and Visualization
# 2 — Modelling & Interpretation
# 3 — Recommendation & Conclusion

# 1.Data Characteristics and Exploration
# 1.1 Data Source 1:
# Used cars data, from Craigslist, found in https://www.kaggle.com/austinreese/craigslist-carstrucks-data.
# This dataset includes used cars price, manufacturer, model, year etc, , which might have impact on price maintenance rate.

# 1.2 Data Source 2:
# New cars data, from https://www.kaggle.com/prassanth/new-cars-price-2019?select=New_cars_cleaned.csv.
# This dataset includes original sale price, car brand, model, year, etc, which can be used to calculate price maintenance rate.



#read data
new <- read.csv("/Users/wuyin/Desktop/New_cars_price.csv",encoding = "UTF-8")
# View(new)
used <- read.csv("/Users/wuyin/Desktop/vehicles.csv",encoding = "UTF-8")
# View(used)


## clean data
new_sep <- new %>% separate(Model, c("year", "brand", "model", "other_descriptions"), " ")
```

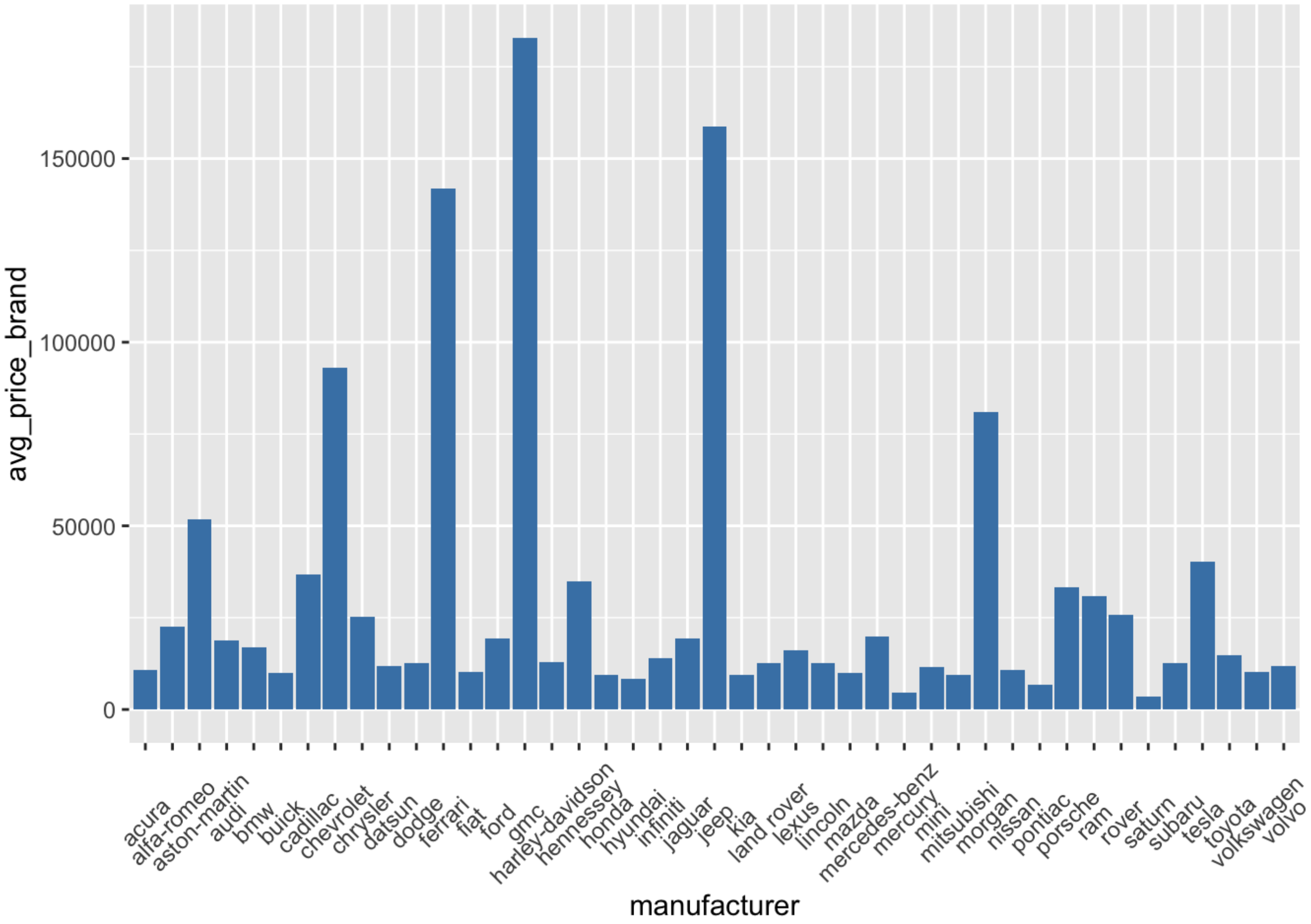
```
new_sep$MSRP <-  as.integer(gsub("[\\$,]", "", new_sep$MSRP))
# View(new_sep)

#visualize the new car data
#price distribution: g1 shows that most new car prices are below $100,000, and the plot is right skewed.
g1 <- ggplot(new_sep, aes(MSRP)) +
  geom_histogram(bins = 50, aes(y = ..density..)) +
  coord_cartesian(xlim=c(0,150000)) +
  labs(x = "Price ($)",
       y = "Density")
g1
```



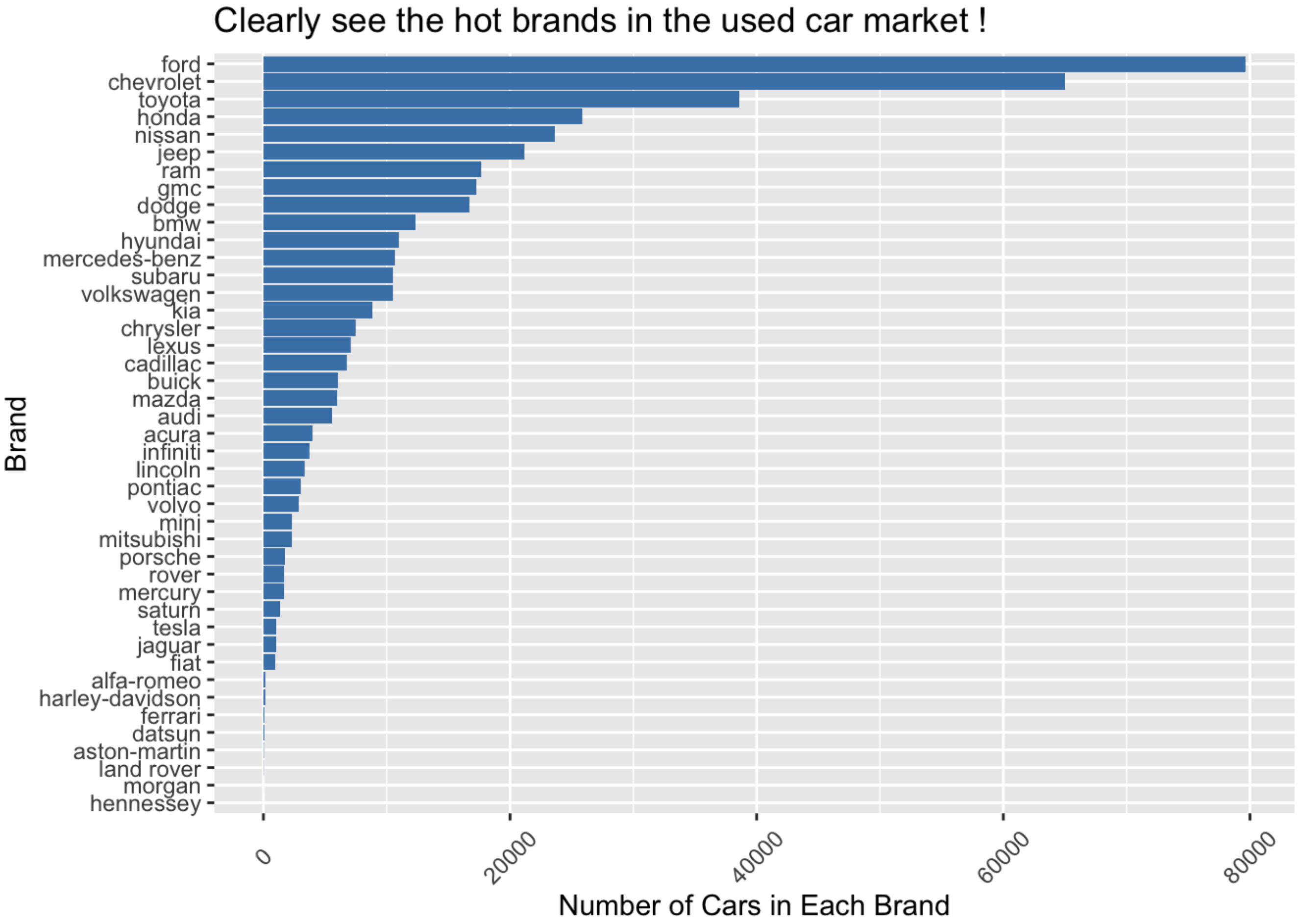
```
#compare the average price of different brand
#create a barplot for the average price of each brand
new1 <- new_sep %>% group_by(brand) %>% drop_na() %>% mutate(avg_price = mean(MSRP))
# View(new1)
new_avgPrice <- new1 %>% select(brand, avg_price)
new_avgPrice <- new_avgPrice[!duplicated(new_avgPrice$brand),]
# View(new_avgPrice)
g2 <- ggplot(new_avgPrice, aes(brand, avg_price)) +
  geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5))
g2
```



```
# g2 shows the rank of the average price of new cars for each brand;
# from the bar plot, we could see that the top 3 expensive brands of car are
# Porsche, Jaguar, Mercedes-Benz; the 3 most affordable brands are Kia, Volkswagen, Mitsubishi




#visualize the used car data
#draw a barplot for the average price of each manufacturer

#DATA LIMITATION 1: Used car price from this data are posted by sellers, but they might not be the final dealing price.
#Therefore, it might cause that our maintenance rate biased.

used1 <- used %>% group_by(manufacturer) %>% mutate(avg_price_brand = mean(price))
# View(used1)
used_avgPrice <- used1 %>% select(manufacturer, avg_price_brand)
used_avgPrice <- used_avgPrice[!duplicated(used_avgPrice$manufacturer),]
used_avgPrice <- used_avgPrice[!used_avgPrice$manufacturer == "", ]
# View(used_avgPrice)
g3 <- ggplot(used_avgPrice, aes(manufacturer, avg_price_brand)) +
  geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5))
g3
```



```
# g3 shows the rank of the average price of used cars for each brand;
# from the bar plot, we could see that the top 3 expensive brands of used car are
# gmc, jeep, ferrari; the 3 most affordable brands are saturn, mercury, pontiac


#draw a barplot for the number of used cars for each manufacture
num_cars <- used %>%
  group_by(manufacturer) %>%
  summarize(num_cars = n()) %>%
  select(manufacturer, num_cars)
num_cars <- num_cars[!duplicated(num_cars$manufacturer),]
num_cars <- num_cars[!num_cars$manufacturer == "", ]
# View(num_cars)
#use a horizontal barplot to see the hot brands in the used car market
g4 <- ggplot(num_cars, aes(reorder(manufacturer,num_cars), num_cars)) +
  geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5)) + coord_flip()
+
  labs(title = "Clearly see the hot brands in the used car market !",
       x = "Brand",
       y = "Number of Cars in Each Brand")
g4
```

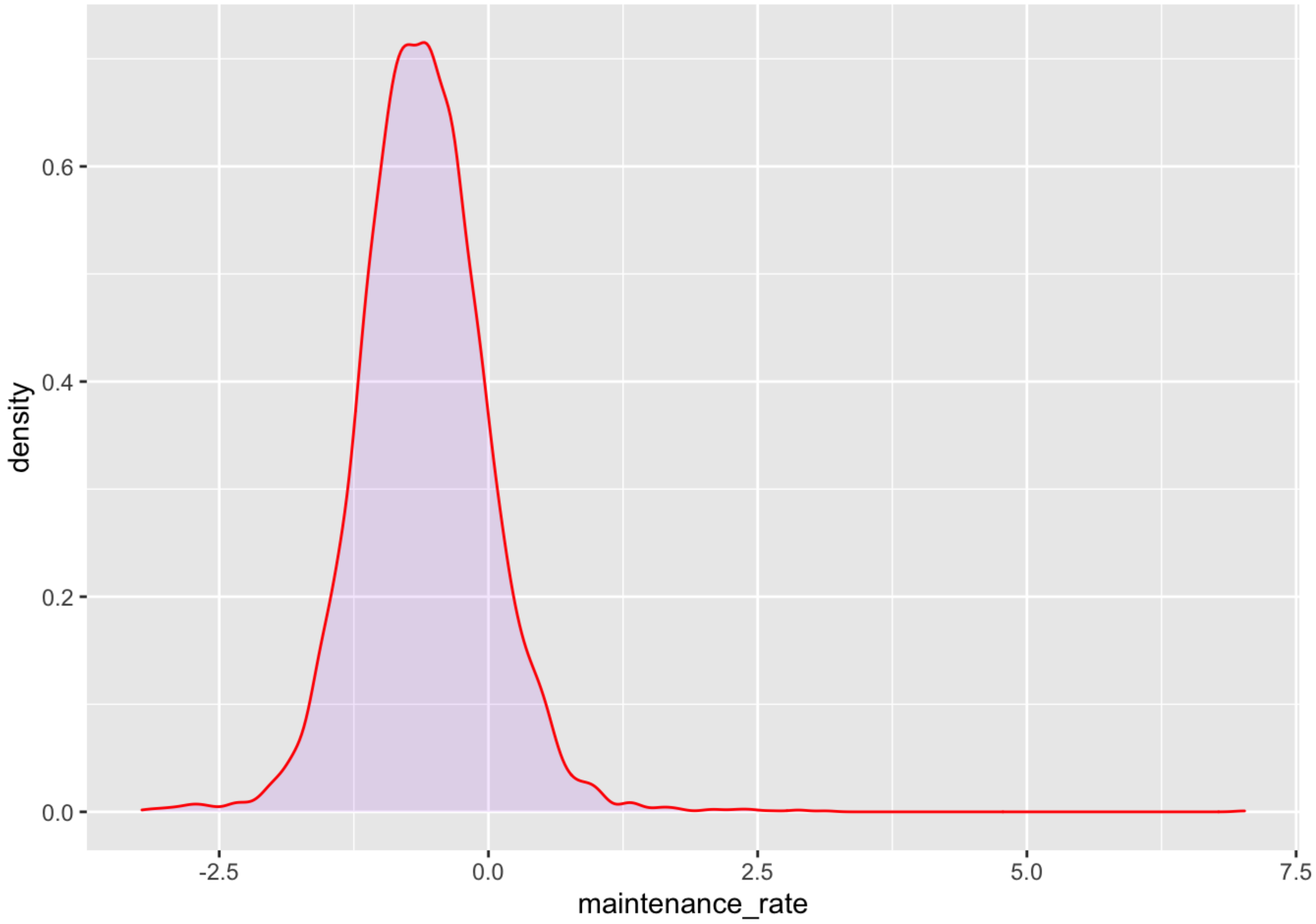

Clearly see the hot brands in the used car market !
Brand
ford
chevrolet
toyota
honda
nissan
jeep
ram
gmc
dodge
bmw
hyundai
mercedes-benz
subaru
volkswagen
kia
chrysler
lexus
cadillac
buick
mazda
audi
acura
infiniti
lincoln
pontiac
volvo
mini
mitsubishi
porsche
rover
mercury
saturn
tesla
jaguar
fiat
alfa-romeo
harley-davidson
ferrari
datsun
aston-martin
land rover
morgan
hennessey
0
20000
40000
60000
80000
Number of Cars in Each Brand

```
#By plotting a horizontal bar plot, we could see that Ford, Chevrolet and Toyota occupied the major used car market.


#data cleaning for joint data
used <- used %>% unite("CarYearModel", year:model, sep = " ") %>%
  select(price, CarYearModel, fuel, odometer, title_status, VIN, transmission, posting_date)
# View(used)

new <- new_sep %>%
  unite("CarYearModel", year:model, sep = " ")

new <- new %>%
  mutate(price = MSRP) %>%
  select(-MSRP)

new_brand <- function(x) {
  tolower(x)
}

new <- new %>%
  mutate(CarYearModel = map(.x = CarYearModel, .f = function(x) new_brand(x))) %>%
  mutate(CarYearModel = as.character(CarYearModel))
# View(new)


# 1.3 Data jointed
# By uniting manufacturer and model, we get “CarYearModel" from used car data. Then, we left join
# the used car data with the new car data by “CarYearModel” to get the jointed data.
# We calculated the price maintenance rate by used logarithm of  (used car price / new car price - used car price) .

#join by CarYearModel
all <- used %>% left_join(new, by = c("CarYearModel") )
# View(all)

#remove NA
all <- na.omit(all, cols = "price.y")
#remove duplicated
all <- all[!duplicated(all$VIN),]
# View(all)

#Calculate maintenance_rate
#Maintenance_rate Methodology:
#We calculated the price maintenance rate by used logarithm of (used car price / (new car price - used car price)).
#When the price maintenance rate is bigger than 0, the used car price is more than half of its original price.
#When the rate is less than 0, the used car price is less than half of its original price.
```

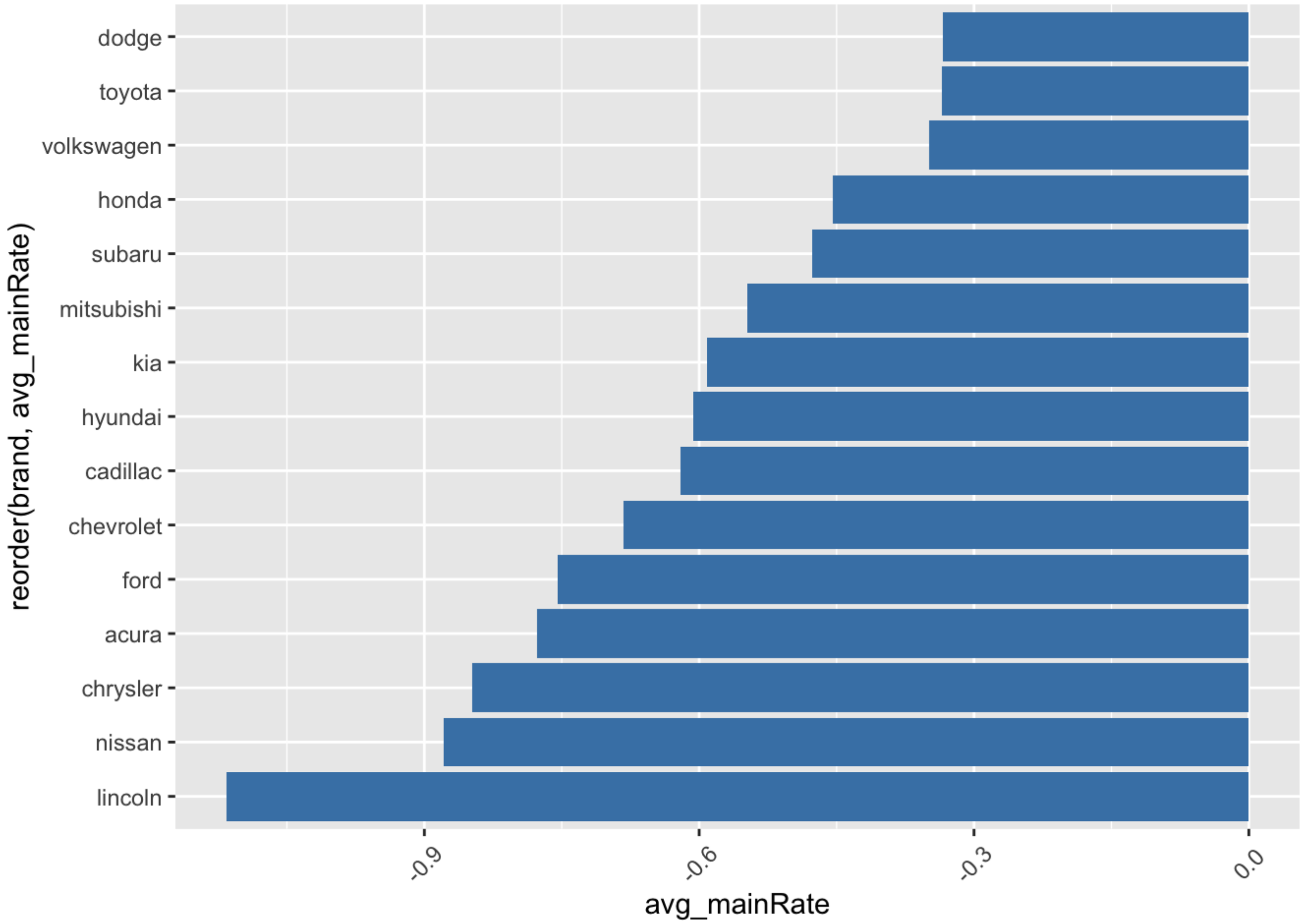
```
all1 <- all %>%
  separate(CarYearModel, c("model_year", "brand", "model"), " ") %>%
  rename(usedCarPrice = price.x, newCarPrice = price.y)  %>%
  filter(usedCarPrice > 1000) %>%  drop_na() %>%
  #DATA LIMITATION 2: There exists some unnormal data having price equal to 0 or too low
(<1000).
  #We deleted them as outliers.
  mutate(maintenance_rate = log(usedCarPrice/(newCarPrice-usedCarPrice)))
```

```
#Since the used cars data are posted in 2020, we used 2020-(model year) to get the numbe
r of years that used cars had been used.
#In order to control the influence of inflation and vintage cars on used car price, we f
ilter out the used cars older than 10 years.
all1 <- all1 %>%  separate(posting_date, c("posting_year", "month"), "-") %>%
  mutate(used_year = 2020 - as.integer(model_year)) %>%
  filter(used_year <= 10 ) %>%
  filter(odometer < 750000 & odometer > 0) %>%
  drop_na() %>%
  na.omit() %>%
  filter_all(all_vars(!is.infinite(.)))
```

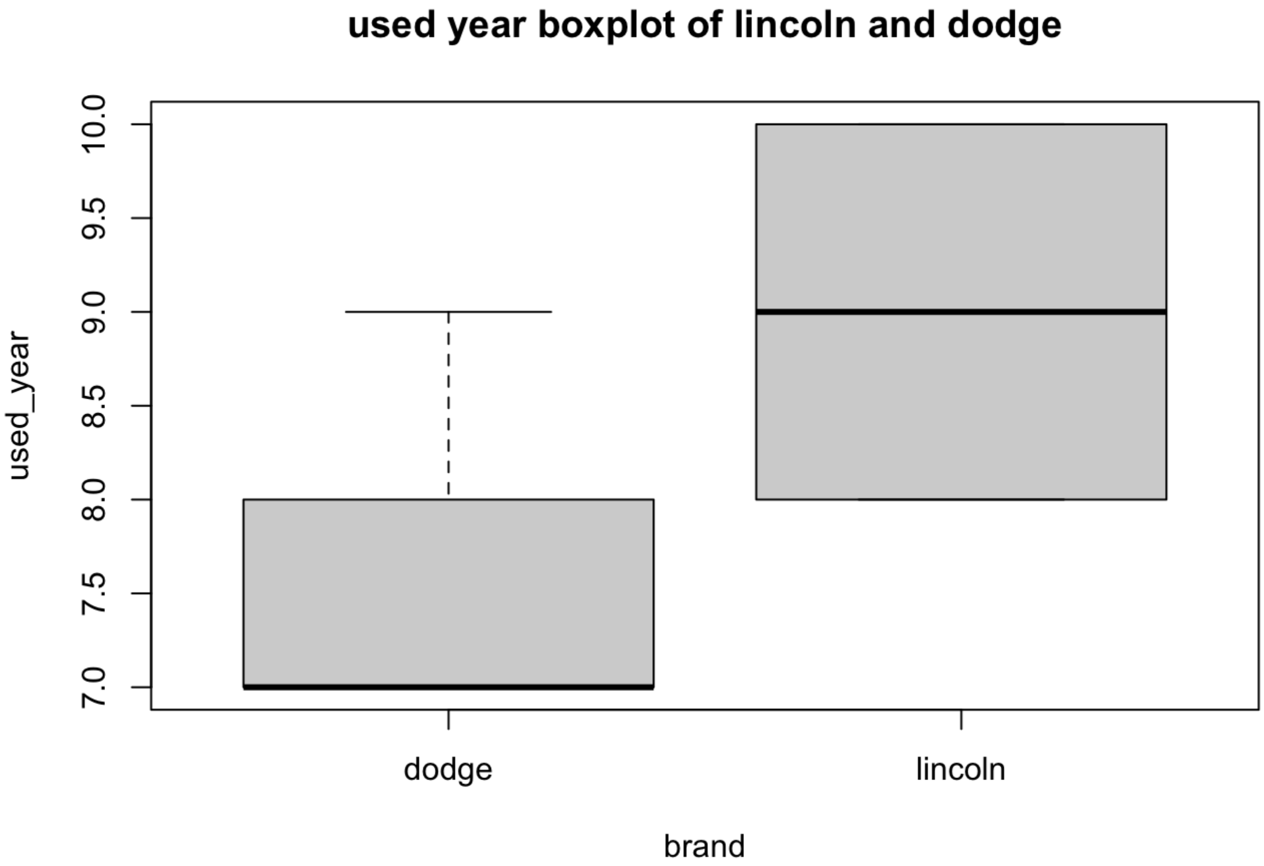
```
# View(all1)

## plot the density distribution of maintenance rate, normally distributed
g5 <- ggplot(all1, aes(x= maintenance_rate)) +
  geom_density(fill="purple",colour="red",alpha=0.1) +
  theme(plot.title = element_text(hjust = 0.5))
g5
```



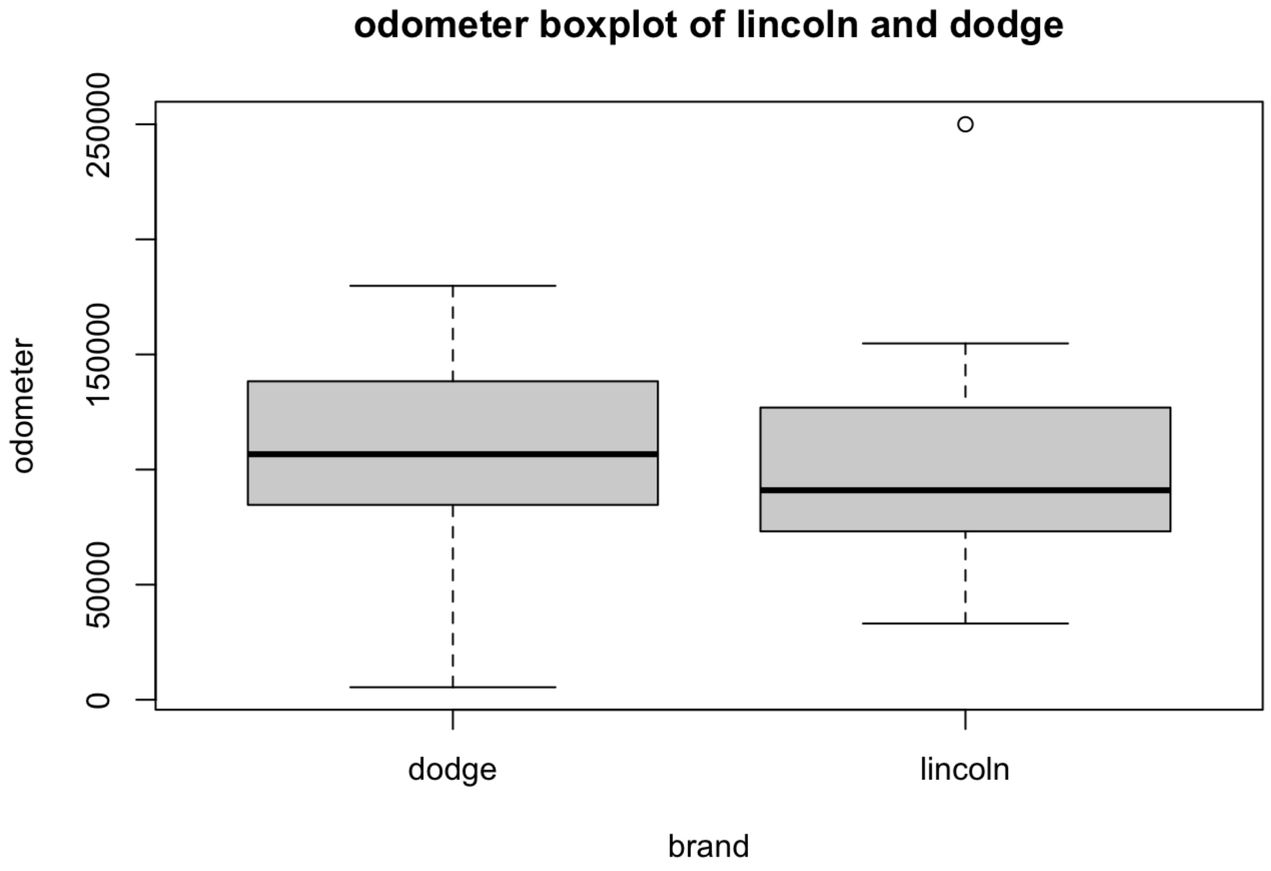
```
#g5 shows that the maintenance rate is approximately normally distributed,
# which indicates that linear regression could be applied here.
# Maintenance rate concentrates in the range of [-2.5, 2.5].



##draw a barplot for the average maintenance rate of each brand
all2 <- all1  %>%
  group_by(brand) %>% mutate(avg_mainRate = mean(maintenance_rate))
avg_mainRate <- all2 %>% select(brand, avg_mainRate)
avg_mainRate <- avg_mainRate[!duplicated(avg_mainRate$brand),]
View(avg_mainRate)
g6 <-  ggplot(avg_mainRate,aes(reorder(brand, avg_mainRate), avg_mainRate)) +
  geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5)) + coord_flip()
g6
```



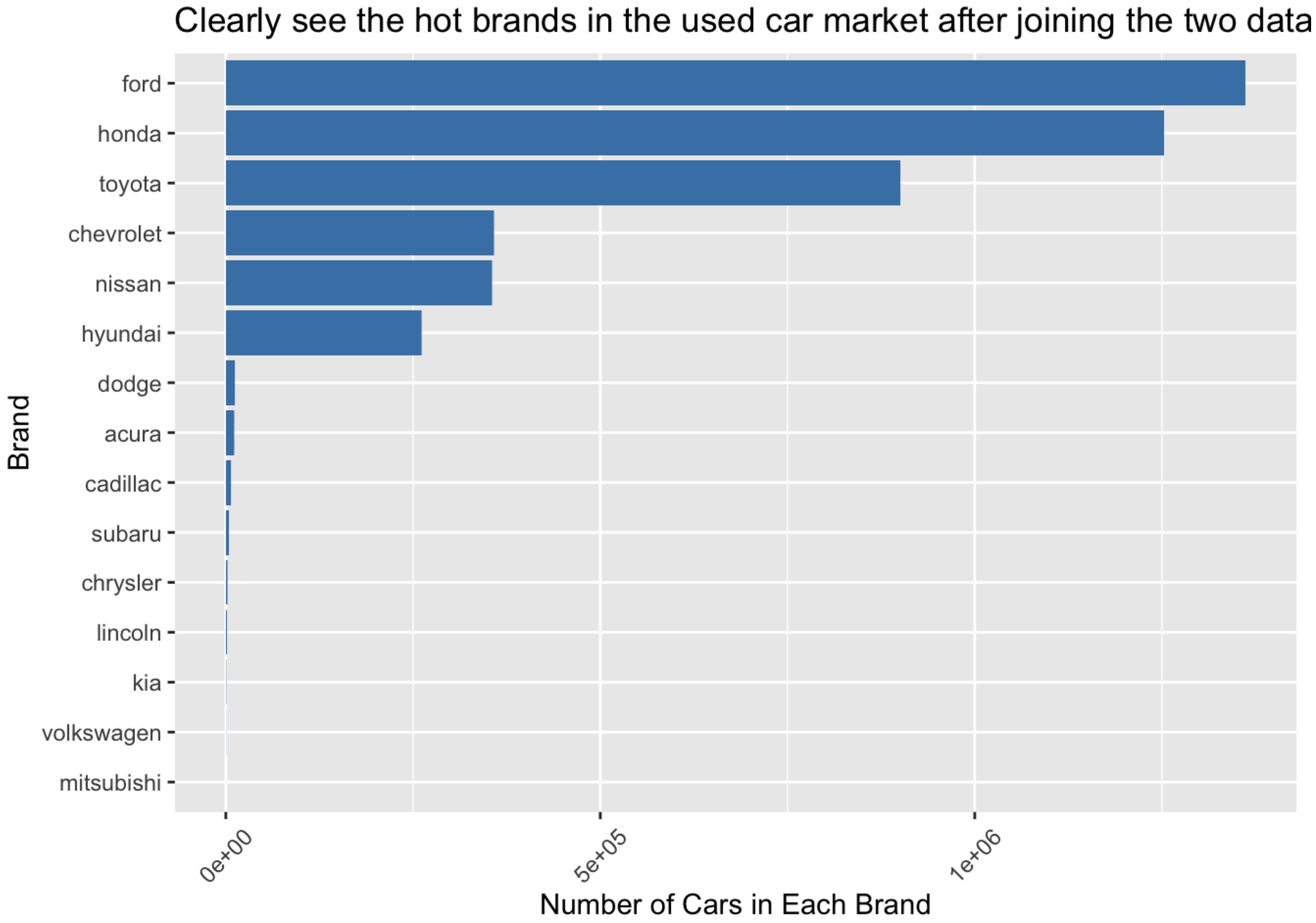
```
# By plotting a horizontal bar plot, we could see that dodge, toyota and volkswagen have
the highest average maintenance rate,
# while lincoln, nissan, chrysler have the lowest average maintenance rate.
#We guess that the major impact on the price are used year and odometer, besides brand.

#boxplot for the used year and odometer given two similar brand
two_brand <- all2 %>% filter(brand == "lincoln" | brand == "dodge")
View(two_brand)
boxplot(used_year ~ brand, data = two_brand, xlab = "brand",
              ylab = "used_year", main = "used year boxplot of lincoln and dodge")
```



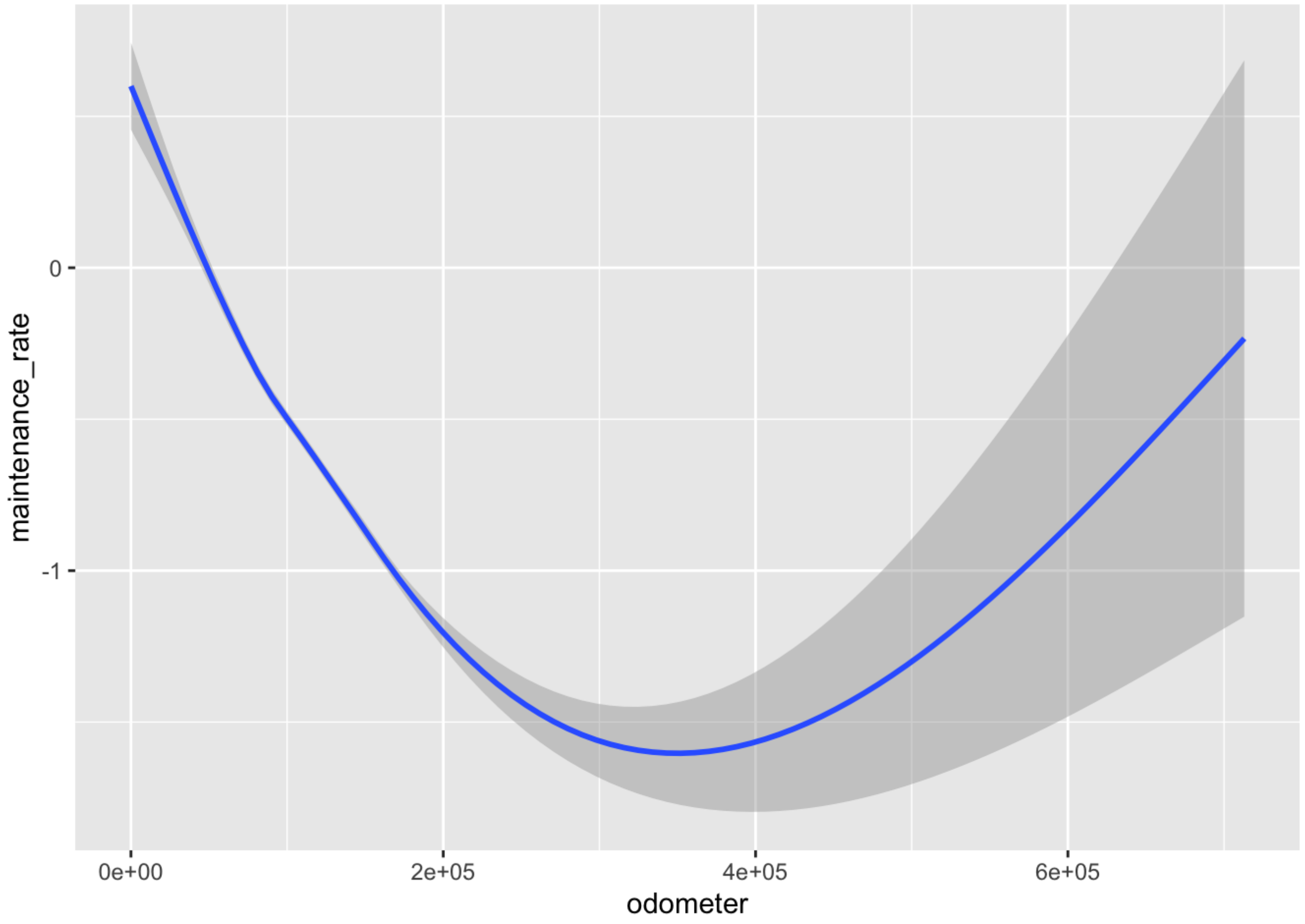
```
boxplot(odometer ~ brand, data = two_brand, xlab = "brand",
              ylab = "odometer", main = "odometer boxplot of lincoln and dodge")
```



```
#From the first boxplot, we could see that lincoln's used year > dodge's,
# which indicates that the used year would be the reason that the maintenance rate of lincoln is less than dodge.
#From the second boxplot, dodge's odometer > lincoln's,
# which indicates that the odometer doesn't influence the maintainence rate too much on these 2 brands.



##draw a barplot for the number of used cars for each brand, check the top 3 hot brands in the market
all3 <- all2 %>%
  group_by(brand) %>%
  mutate(count=n())
# View(all3)
g7 <-  ggplot(all3, aes(reorder(brand,count), count)) +
  geom_bar(stat="identity",fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5)) + coord_flip() +
  labs(title = "Clearly see the hot brands in the used car market after joining the two datasets!",
       x = "Brand",
       y = "Number of Cars in Each Brand")
g7
```



```
## see the relationship between odometer and maintenance_rate
g8 <- ggplot(all3, aes(odometer, maintenance_rate)) +
  geom_smooth()
g8
```



```
#g8 shows that maintenance rate is not only influenced by odometer


# select variables: We considered that fuel type and title status may have impact on the maintenance rate, so we converted these variables to dummy variables.
cars <- all1 %>%
  select(usedCarPrice, newCarPrice, used_year, odometer, brand, title_status, fuel, maintenance_rate) %>%
  mutate(fuel = ifelse(fuel == "gas", 1, 0) ) %>%
  mutate(title_status = ifelse(title_status == "clean", 1, 0))
# View(cars)




#correlation
library(Hmisc)
```

```
library(corrplot)
```

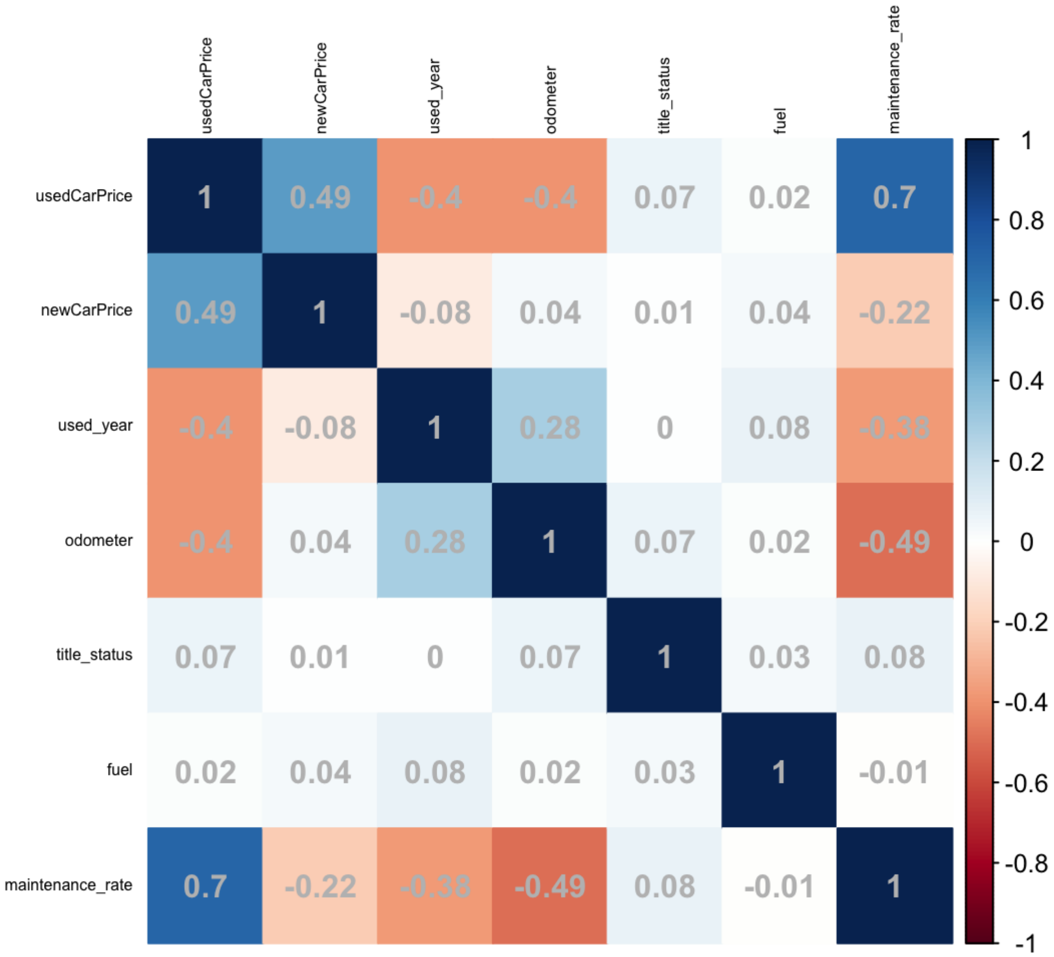
```
cars1 <- ungroup(cars) %>%
  select(-brand)
# View(cars1)
corr_cars <- cor(cars1)
corrplot(corr_cars, method = "color",
         addCoef.col = "gray",
         tl.cex = 0.5,
         tl.col = "black")
#no multicollinearity (linearity rate between maintenance rate and used car price is 0.
7,
# and linearity rate between maintenance rate and new car price is 0.49,
# this is due to the reason that maintenance rate is calculatd by new car price and used
car price)


#modelling
library(olsrr)
```

```
model <- lm(maintenance_rate ~. , data = cars)
summary(model)
```

```
## 
## Call:
## lm(formula = maintenance_rate ~ ., data = cars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.8779 -0.0517  0.0023  0.0649  4.7074 
## 
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)     -2.105e-01  3.300e-02   -6.377 1.95e-10 ***
## usedCarPrice     1.565e-04  9.099e-07  172.041  < 2e-16 ***
## newCarPrice     -6.069e-05  4.425e-07 -137.134  < 2e-16 ***
## used_year       -1.268e-02  2.029e-03   -6.249 4.46e-10 ***
## odometer        -9.327e-07  6.846e-08  -13.625  < 2e-16 ***
## brandcadillac   -1.922e-01  2.490e-02   -7.719 1.39e-14 ***
## brandchevrolet  -2.293e-01  1.883e-02  -12.177  < 2e-16 ***
## brandchrysler   -3.072e-02  3.042e-02   -1.010 0.312716    
## branddodge      -1.536e-01  2.318e-02   -6.625 3.81e-11 ***
## brandford       -1.046e-01  1.748e-02   -5.986 2.29e-09 ***
## brandhonda      -1.058e-01  1.792e-02   -5.906 3.71e-09 ***
## brandhyundai    -1.456e-01  1.915e-02   -7.603 3.39e-14 ***
## brandkia        -2.144e-01  3.868e-02   -5.543 3.12e-08 ***
## brandlincoln     9.889e-02  3.413e-02    2.898 0.003776 ** 
## brandmitsubishi -1.439e-01  6.119e-02   -2.351 0.018751 *  
## brandnissan     -1.365e-01  1.853e-02   -7.370 1.97e-13 ***
## brandsubaru     -1.083e-01  2.809e-02   -3.856 0.000117 ***
## brandtoyota     -9.668e-02  1.815e-02   -5.326 1.04e-07 ***
## brandvolkswagen -1.665e-01  4.623e-02   -3.603 0.000318 ***
## title_status     6.207e-02  1.318e-02    4.710 2.54e-06 ***
## fuel             2.207e-02  1.134e-02    1.946 0.051703 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1657 on 5387 degrees of freedom
## Multiple R-squared:  0.9231, Adjusted R-squared:  0.9228 
## F-statistic:  3232 on 20 and 5387 DF,  p-value: < 2.2e-16
```

```
# model$coefficients
# How model is fitted and its relationship to the problem:
# The adjusted R squared is 0.9228, which shows the high ratio of variation could be explained by the model.
# The F-test statistic is 3232 and the p-value of this model is very small, indicating that the regression model is significant
# as a whole.
# The coefficient estimates of the dependent variables such as original purchasing price, used year of cars, odometer, fuel type,
# and title status are shown above.
# The P values of the coefficients of each variable are all less than 0.05, which means that the influence
# of these variables on the price maintenance rate cannot be ignored.
# These variables in the model should be considered when choosing a used car with high price maintenance rate to buy.



#Uncertainty of parameter estimation and how this impacts your recommendation
all_un <- all %>%
  separate(CarYearModel, c("model_year", "brand", "model"), " ") %>%
  rename(usedCarPrice = price.x, newCarPrice = price.y)  %>%
  filter(usedCarPrice > 1000) %>%  drop_na() %>%
  mutate(maintenance_rate = log(usedCarPrice/(newCarPrice-usedCarPrice)))
```

```
all_un <- all_un %>%
  mutate(used_year = 2020 - as.integer(model_year)) %>%
  filter(used_year <= 10 ) %>%
  filter(odometer < 300000 & odometer > 0) %>%
  drop_na() %>%
  na.omit() %>%
  filter_all(all_vars(!is.infinite(.)))
cars_un <- all_un %>%
  select(usedCarPrice, newCarPrice, used_year, odometer, brand, title_status, fuel, maintenance_rate) %>%
  mutate(fuel = ifelse(fuel == "gas", 1, 0) ) %>%
  mutate(title_status = ifelse(title_status == "clean", 1, 0))
#correlation
library(Hmisc)
library(corrplot)
cars1_un <- ungroup(cars_un) %>%
  select(-brand)
# View(cars1)
corr_cars_un <- cor(cars1_un)
corrplot(corr_cars, method = "color",
         addCoef.col = "gray",
         tl.cex = 0.5,
         tl.col = "black")
```



```
#modelling
library(olsrr)
model_un <- lm(maintenance_rate ~. , data = cars_un)
summary(model_un)
```

```
##
## Call:
## lm(formula = maintenance_rate ~ ., data = cars_un)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8689 -0.0527  0.0023  0.0658  4.7092
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -2.019e-01  3.298e-02   -6.123 9.83e-10 ***
## usedCarPrice     1.560e-04  9.174e-07  170.054  < 2e-16 ***
## newCarPrice     -6.056e-05  4.439e-07 -136.421  < 2e-16 ***
## used_year       -1.270e-02  2.026e-03   -6.266 3.99e-10 ***
## odometer        -1.005e-06  7.202e-08  -13.954  < 2e-16 ***
## brandcadillac   -1.936e-01  2.485e-02   -7.789 8.07e-15 ***
## brandchevrolet  -2.303e-01  1.879e-02  -12.254  < 2e-16 ***
## brandchrysler   -3.186e-02  3.036e-02   -1.049 0.294098
## branddodge      -1.531e-01  2.313e-02   -6.620 3.94e-11 ***
## brandford       -1.053e-01  1.745e-02   -6.037 1.68e-09 ***
## brandhonda      -1.049e-01  1.789e-02   -5.865 4.75e-09 ***
## brandhyundai    -1.482e-01  1.912e-02   -7.750 1.10e-14 ***
## brandkia        -2.166e-01  3.860e-02   -5.610 2.13e-08 ***
## brandlincoln     9.598e-02  3.407e-02    2.817 0.004858 **
## brandmitsubishi -1.449e-01  6.106e-02   -2.373 0.017679 *
## brandnissan     -1.373e-01  1.849e-02   -7.425 1.31e-13 ***
## brandsubaru     -1.077e-01  2.803e-02   -3.843 0.000123 ***
## brandtoyota     -9.651e-02  1.812e-02   -5.326 1.04e-07 ***
## brandvolkswagen -1.686e-01  4.613e-02   -3.655 0.000260 ***
## title_status     6.397e-02  1.316e-02    4.861 1.20e-06 ***
## fuel             2.232e-02  1.132e-02    1.972 0.048632 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1654 on 5381 degrees of freedom
## Multiple R-squared:  0.9233, Adjusted R-squared:  0.923
## F-statistic:  3239 on 20 and 5381 DF,  p-value: < 2.2e-16
```

```
#We changed filter condition for odometer from 750000 to 300000 to check the uncertainty of parameter estimation,
#and found that coefficient of odometer increased very slightly.
# Therefore, this uncertainty does not necessarily influence our recommendation.

# Therefore, these variables in our model must be controlled when analyzing the influence of automobile brand
# on the price maintenance rate, which is also the reason why we establish the regression model.
```

```
Conclusion:
Regardless of the impact caused by brand, the most important factors affecting the price
of a used car are its title status, fuel type and used year, that is, the greater the us
ed year, the lower the maintenance rate; gas type car has higher maintenance rate; clean
-title car has higher maintenance rate.

Recommendation:
When choosing a used car, we should focus firstly on its brand and then title status, fu
el type and used year. Car brand recommended are Chrysler and Toyota. Also, we recommend
to choose gas-type, less used, clean-title cars.
```