STA141A Project

Youjia Wang, Yuehe Wen, Jiaye Chen

5/31/2021

1. Introduction

Buying or selling used cars has become more popular in recent years, considering the high costs of getting a brand new one. The goal of this project is to explore how different parameters, such as transmission, mileage, fuel type, etc., influence the selling price of used cars. In such a way, we are not only to predict how much a person should sell his/her old car or determine whether a listing car price is reasonable or not compared to the general market, but also to explore what are the key factors that affect the selling price of used cars.

We will be using ggplot2 and multiple linear regression to understand the relationship and contribution of car statistics to the price of used cars. We want to use this model to predict the price of a certain used car. We will also use variable selection to identify the most influential variable in the car statistics that contribute to its price. Additionally, we plan to analyze the correlation between each of the car statistics as well. Last but not least, we also plan to test on whether or not we can fit a model to predict the price of a used car.

2. Background

The data set that we will refer to was found from Kaggle, named 100,000 UK Used Car Data set. It was collected by Aditya, and incorporated by 100,000 scrapped used car listings in the British market.

Considering the above data set was incorporated by a number of different small data sets, and the huge amount of data in sum, we decide to choose 3 typical data sets from the above general data set. Additionally, considering the high possibility of finding a lot of missing data, we will perform a data cleaning to remove some unnecessary information, including replicates.

The variables that we will be using in this project include the information of cars' model, registration year, price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

All 9 variables are represented below: "model" – car's model "year" – registration year "price" – price in £ "transmission" – type of gearbox "mileage" – distance used "fuelType" – engine fuel "tax" – road tax "mpg" – miles per gallon "engineSize" – size in litres

3. Key Questions to be Addressed:

1. Is there any NA in each column? Which we should delete and which we should refill with useful information like mean or mode.
2. Is there any relationship between the variables? If so, what is it? Registration year vs price in £ type of gearbox vs price in £ distance used vs price in £ engine fuel vs price in £ road tax vs price in £ miles per gallon vs price in £ size in litres vs price in £
3. What's the most influential parameter variable affecting the response variable (price in £)?
4. What's the least influential parameter variable affecting the response variable (price in £)?
5. Can we predict the price of a used car given the model we build and how effective is it?

4. Data Cleaning

First of all, we do a data cleaning to ensure our future steps can be done smoothly. We also do it to address the question of whether there are any NA or not, and we decide whether we should delete or refill with useful information like mean or mode.

```
## # A tibble: 185 x 9
##    model  year price transmission mileage fuelType   tax   mpg engineSize
##    <chr> <dbl> <dbl> <chr>          <dbl> <chr>    <dbl> <dbl>      <dbl>
##  1 Q3     2019 34485 Automatic         10 Diesel     145  47.1        2
##  2 Q3     2019 34485 Automatic         10 Diesel     145  47.1        2
##  3 Q5     2019 31998 Semi-Auto        100 Petrol     145  33.2        2
##  4 Q3     2015 13995 Manual         35446 Diesel     145  54.3        2
##  5 Q2     2019 22495 Manual          1000 Diesel     145  49.6        1.6
##  6 Q2     2019 22495 Manual          1000 Diesel     145  49.6        1.6
##  7 Q3     2015 13995 Manual         35446 Diesel     145  54.3        2
##  8 Q5     2019 31998 Semi-Auto        100 Petrol     145  33.2        2
##  9 Q2     2019 22495 Manual          1000 Diesel     145  49.6        1.6
## 10 Q3     2019 32500 Automatic       4432 Petrol     145  31.4        2
## # ... with 175 more rows
```

```
## # A tibble: 6 x 10
##   model  year price transmission mileage fuelType   tax   mpg engineSize brand
##   <chr> <dbl> <dbl>        <dbl>   <dbl>    <dbl> <dbl> <dbl>      <dbl> <chr>
## 1 A1     2017 12500            1   15735        1   150  55.4        1.4 Audi
## 2 A6     2016 16500            2   36203        2    20  64.2        2   Audi
## 3 A1     2016 11000            1   29946        1    30  55.4        1.4 Audi
## 4 A4     2017 16800            2   25952        2   145  67.3        2   Audi
## 5 A3     2019 17300            1    1998        1   145  49.6        1   Audi
## 6 A1     2016 13900            2   32260        1    30  58.9        1.4 Audi
```

We find replicated rows. Since it's very rare of having the same model, same year,same mog and same size in the market, and we don't know the collection process. Therefore, in order to make our data more precise, we decide to delete the replicated rows.

After removing the unnecessary information, we generate our new data set by using rbind() to generate all 3 cleaned versions of data sets together.

5. Visualizations and Analysis

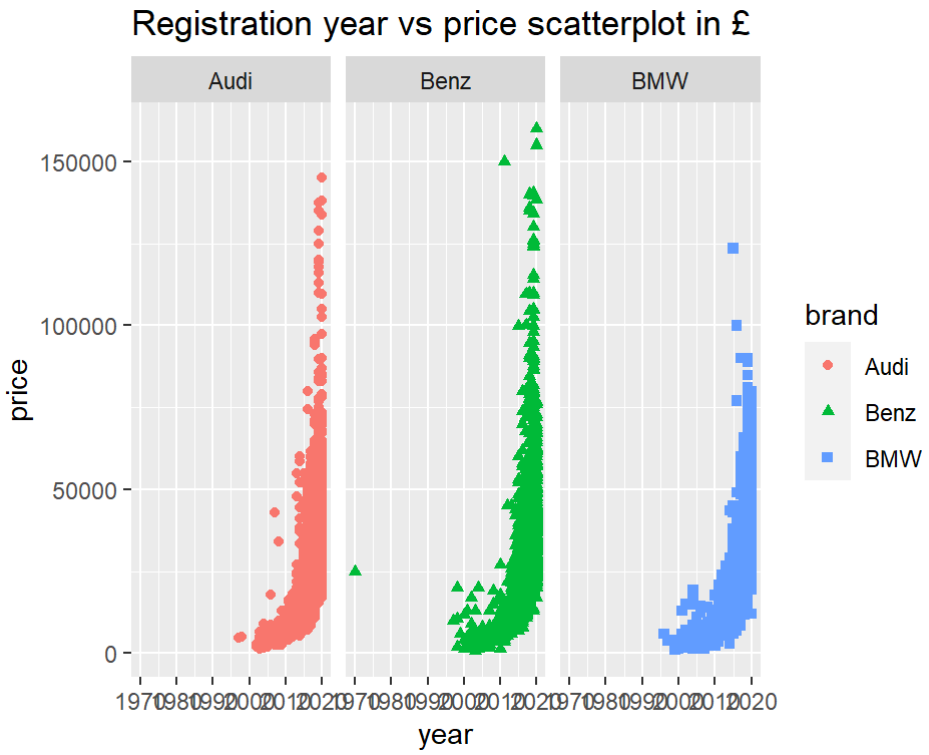
i. Relationship between the Variables

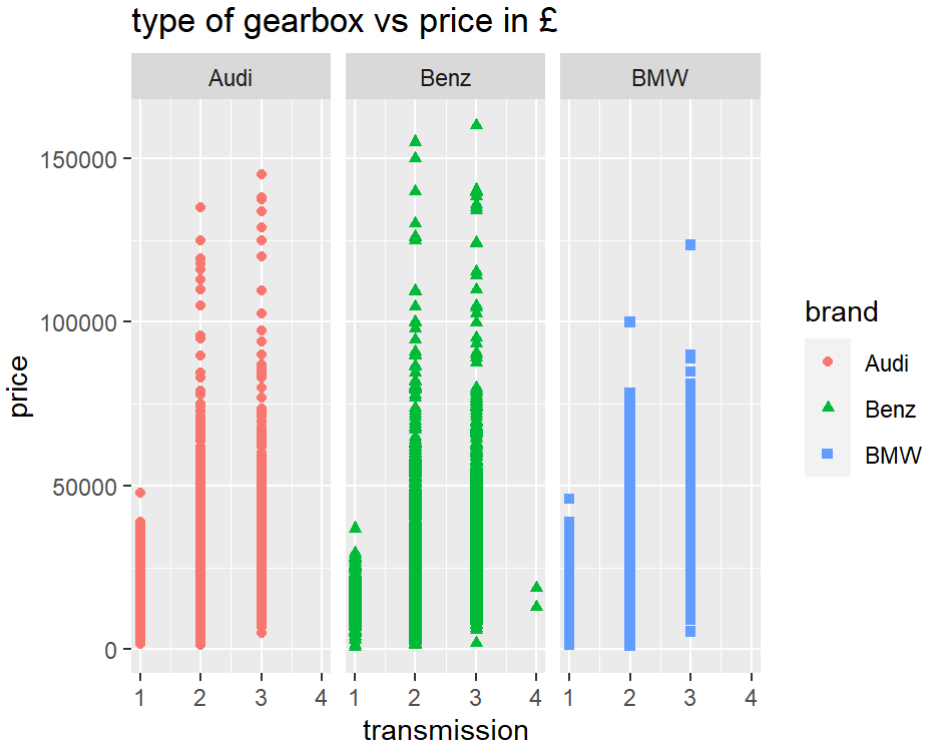
This project mainly focuses on exploring the relationship between the variables, and we pay close attention especially to find how/which explanatory variables affect the response variable.

We approach such a goal along with answering our second question of whether there's a relationship between the variables in several different ways.

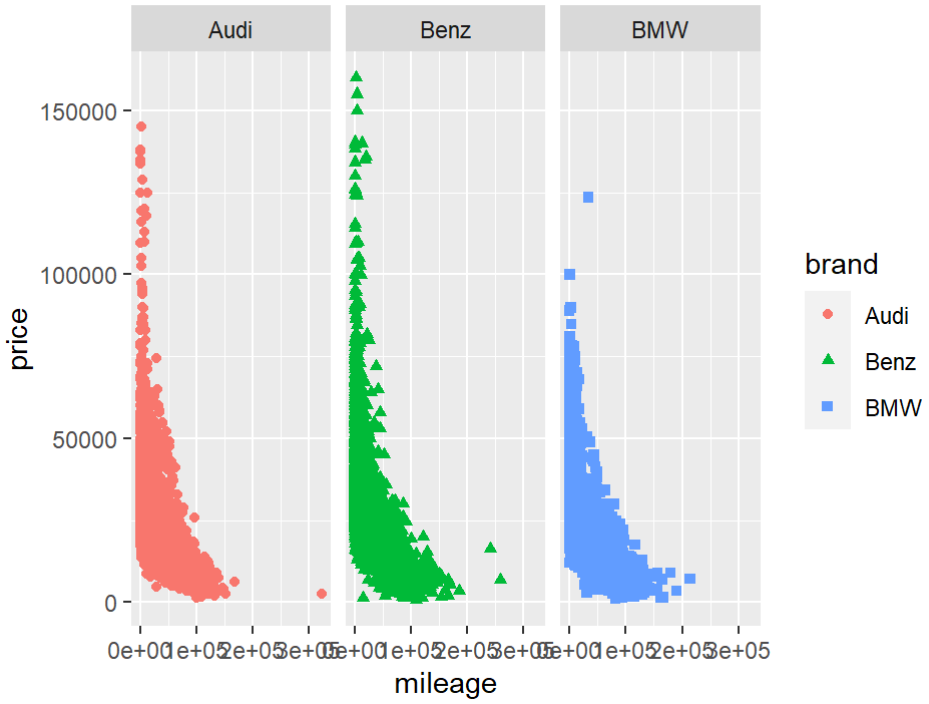
We first use *ggplot()* to generate dot/box plots between the explanatory variables with our response variable, price. We then approach this using the *biplot()* after using *principal component analysis (PCA)* and considering the correlation to have a clearer understanding on the relationship between the variables.

1) Dot/box plots using *ggplot()*:

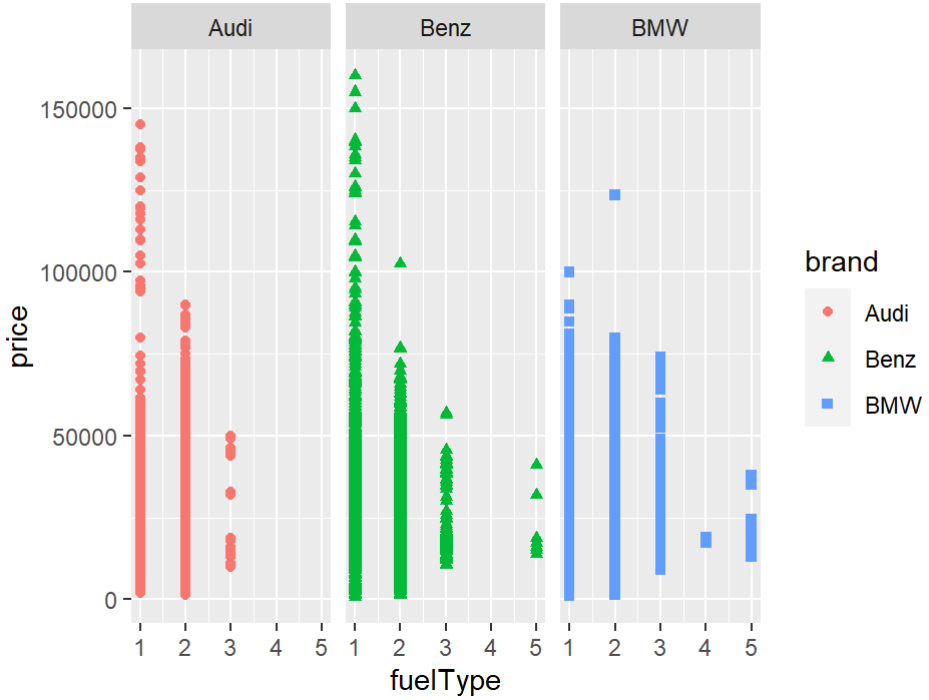

Registration year vs price scatterplot in £
Audi
Benz
BMW
150000
100000
50000
0
price
year
brand
Audi
Benz
BMW

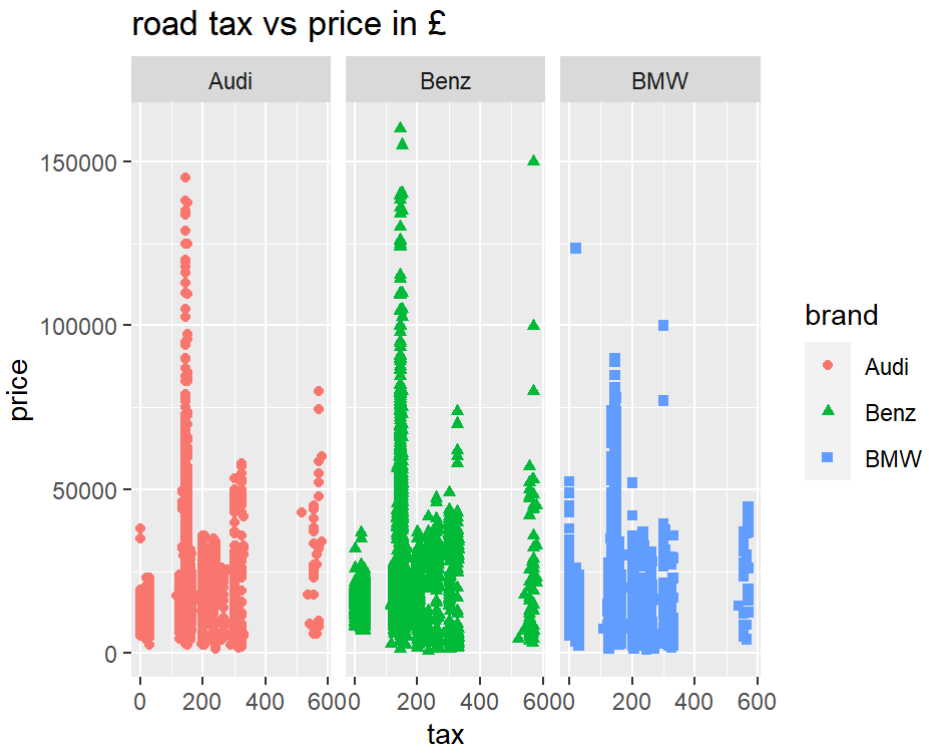

type of gearbox vs price in £
Audi
Benz
BMW
150000
100000
50000
0
price
1
2
3
4
transmission
brand
Audi
Benz
BMW

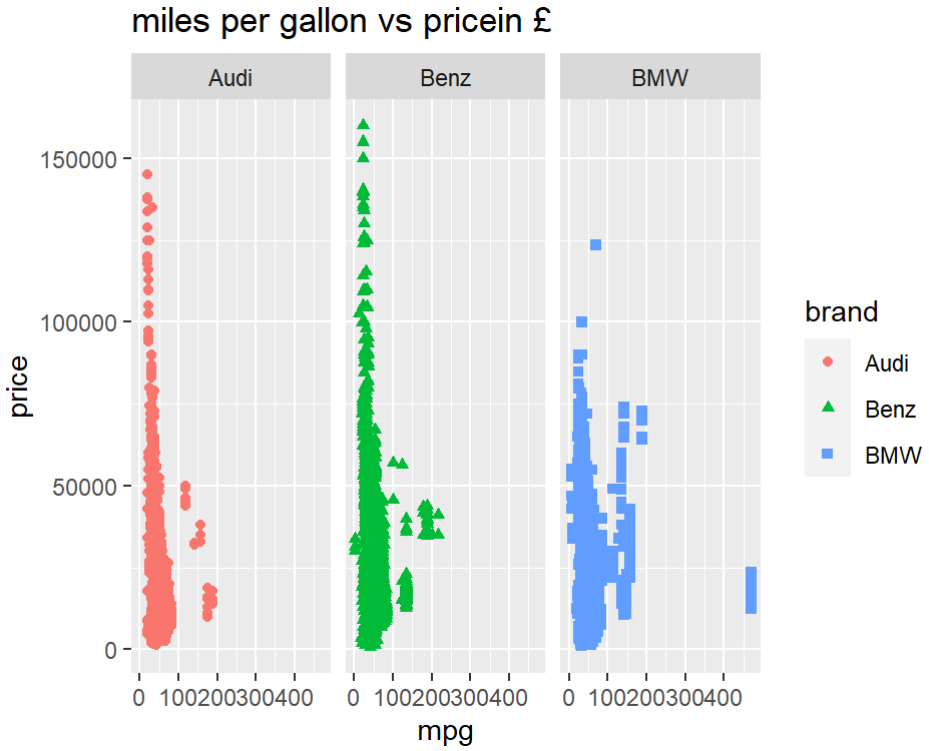
distance used vs price in £

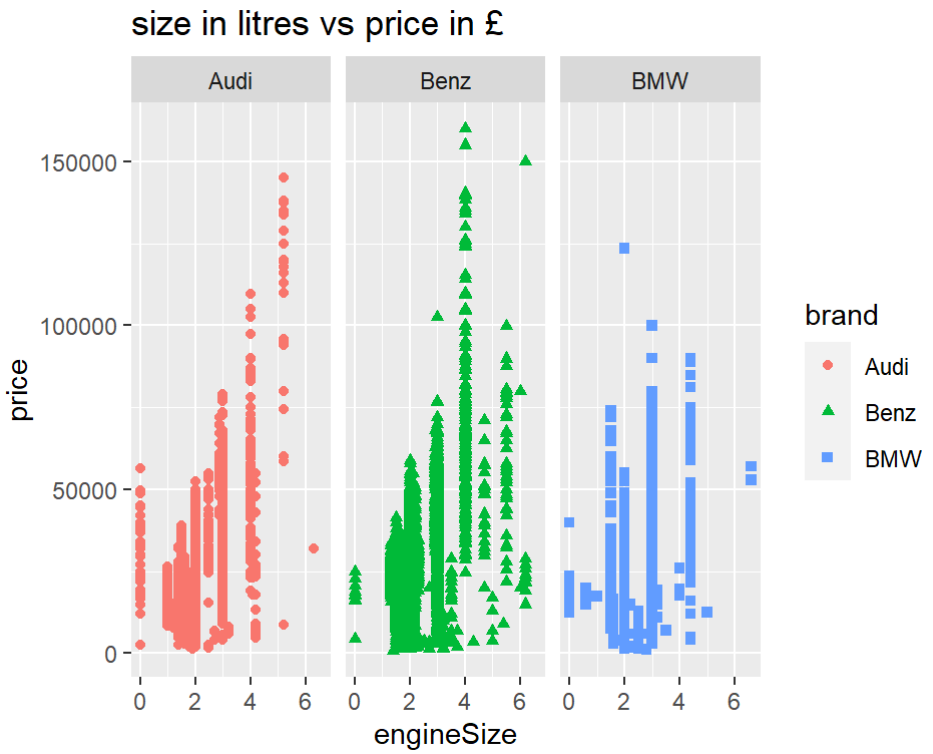


engine fuel vs price in £




road tax vs price in £
Audi
Benz
BMW
price
150000
100000
50000
0
0
200
400
6000
tax
brand
Audi
Benz
BMW


miles per gallon vs pricein £
Audi
Benz
BMW
price
150000
100000
50000
0
0
100
200
300
400
mpg
brand
Audi
Benz
BMW



Examing the relationship of each variable against the response variable price, we can see that for each brand, the trend for each variable is similar. The most correlated variable seems to be year, mileage, transmission, fueltype, mpg and engineSize. But since these are just rough graphs, we still need quantitative analysis to determine which variables are to be considered.

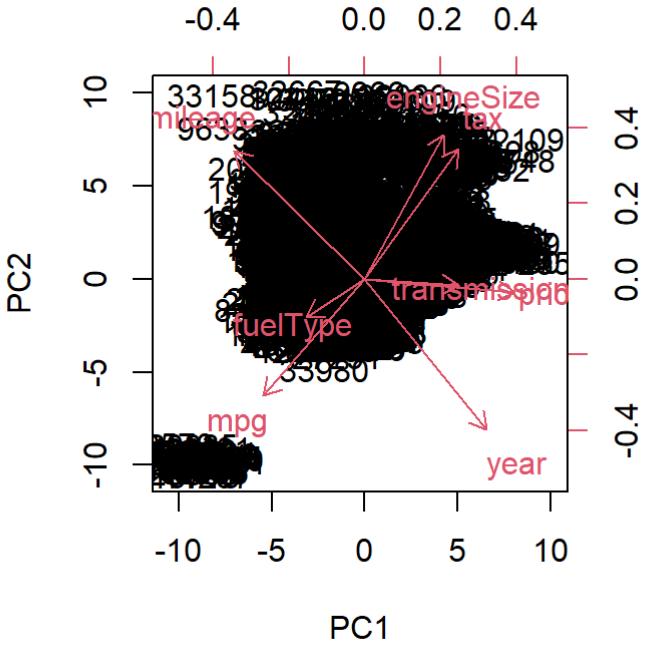
2) Variable selection based on correlation matrix

```
##                year price transmission mileage fuelType   tax   mpg engineSize
## year           1.00  0.58         0.28   -0.76    -0.11  0.04 -0.12       -0.07
## price          0.58  1.00         0.37   -0.56    -0.07  0.29 -0.32        0.52
## transmission   0.28  0.37         1.00   -0.27     0.04  0.22 -0.12        0.26
## mileage       -0.76 -0.56        -0.27    1.00     0.22 -0.16  0.18        0.04
## fuelType      -0.11 -0.07         0.04    0.22     1.00 -0.24  0.49        0.14
## tax            0.04  0.29         0.22   -0.16    -0.24  1.00 -0.41        0.39
## mpg           -0.12 -0.32        -0.12    0.18     0.49 -0.41  1.00       -0.32
## engineSize    -0.07  0.52         0.26    0.04     0.14  0.39 -0.32        1.00
```

From the correlation matrix, we can see that the variables that have highest correlations with price are year (0.58), mileage (absolute value 0.56), engineSize (0.52) and transmission (0.37). However, year has a correlation of 0.76 with mileage, which is rather high, indicating that these two variables are highly correlated with each other. Therefore, in order to avoid multicollinearity, we will drop the variable mileage and only includes year, engineSize and transmission in our model.

3) biplot() after performing PCA

We also do further analysis on discovering the correlation between explanatory variables and the response variable to verify our selection and to see if we miss some. The method we choose is the Principle Component Analysis. We will use *biplot()* after performing PCA to have a more direct and visualized result of the relationship between variables.



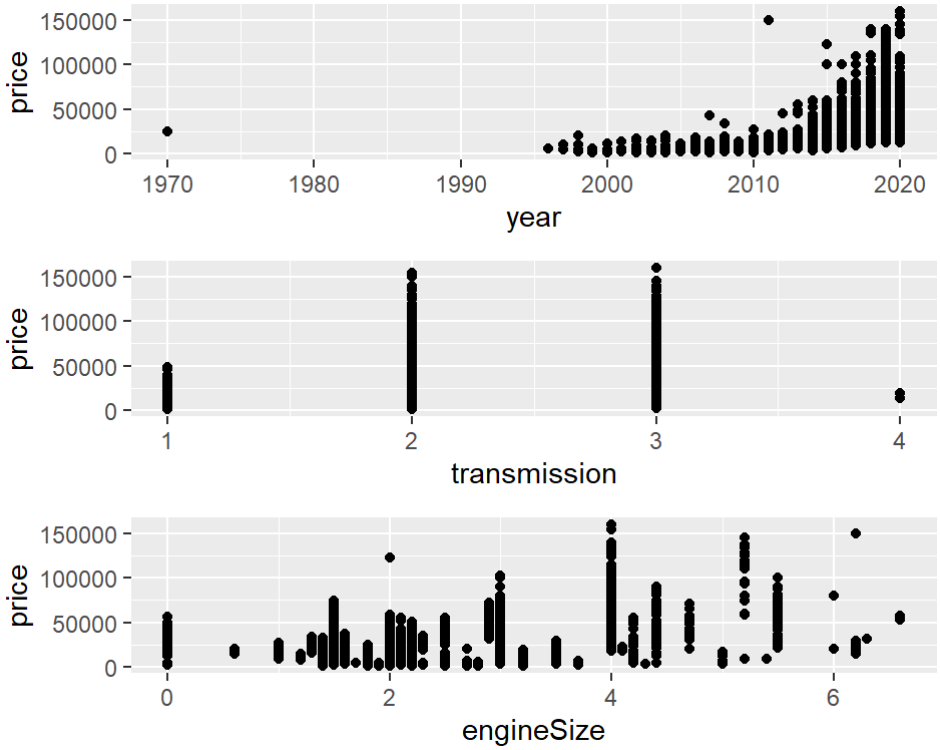
Here, we first do PCA on the dataset, from the graph, we can see that transmission, mileage, year and engineSize are close to price, showing that these four variables are highly correlated with price, which is the response variable. Variable year and mileage has a angle of almost 180 degree, indicating that they are negatively highly correlated

This result is consistent with the result in 2) that these four variables have the four highest correlation coefficients with price but year and mileage has multicollinearity so we will drop mileage.

ii. Linear Regression

Furthermore, to further address our third and fourth question, we use the multiple linear regression to help us discover whether some variables in the car statistics contribute significantly to the response variable, price. This will also help us simulate the regression model to predict the price.

Linear relationship between year vs price, transmission vs price, and engineSize vs price using *qplot()*: Continuing on the results we get above, we then use *qplot()* to continue exploring the relationship between the above three explanatory variables with the response variable, price, by demonstrating the linear relationship in between.



The above graph shows the linear relationship between our chosen variables and price. Therefore, we will first try the linear model : price ~ year + enginesize + transmission

```
##
## Call:
## lm(formula = price ~ year + engineSize + transmission, data = ds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37044  -3763   -699   2624 169743
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.232e+06  3.562e+04  -174.9   <2e-16 ***
## year          3.089e+03  1.767e+01   174.8   <2e-16 ***
## engineSize    1.087e+04  6.766e+01   160.6   <2e-16 ***
## transmission  9.610e+02  5.138e+01    18.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6972 on 34200 degrees of freedom
## Multiple R-squared:  0.646,  Adjusted R-squared:  0.646
## F-statistic: 2.081e+04 on 3 and 34200 DF,  p-value: < 2.2e-16
```

The model returns with results with F-statistic 2.081e+04 with d.f. 3, 34200, corresponding p-value of 0 and adjusted R2 0.646. Since the response variable price has large values, we consider taking the logrithm transformation on it.

Model with log transformation:

```
##
## Call:
## lm(formula = log(price) ~ year + engineSize + transmission, data = ds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5535 -0.1327 -0.0061  0.1238  8.0416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.965e+02  1.097e+00 -270.31   <2e-16 ***
## year          1.515e-01  5.440e-04  278.48   <2e-16 ***
## engineSize    3.464e-01  2.083e-03  166.26   <2e-16 ***
## transmission  7.663e-02  1.582e-03   48.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2147 on 34200 degrees of freedom
## Multiple R-squared:  0.7864, Adjusted R-squared:  0.7864
## F-statistic: 4.198e+04 on 3 and 34200 DF,  p-value: < 2.2e-16
```

The logarithmic transformation returns results with F-statistic 4.198e+04 with d.f. 3, 34200, corresponding p-value of 0 and increased adjusted R2 0.7864.

iii. ANOVA Analysis and Model Selection:

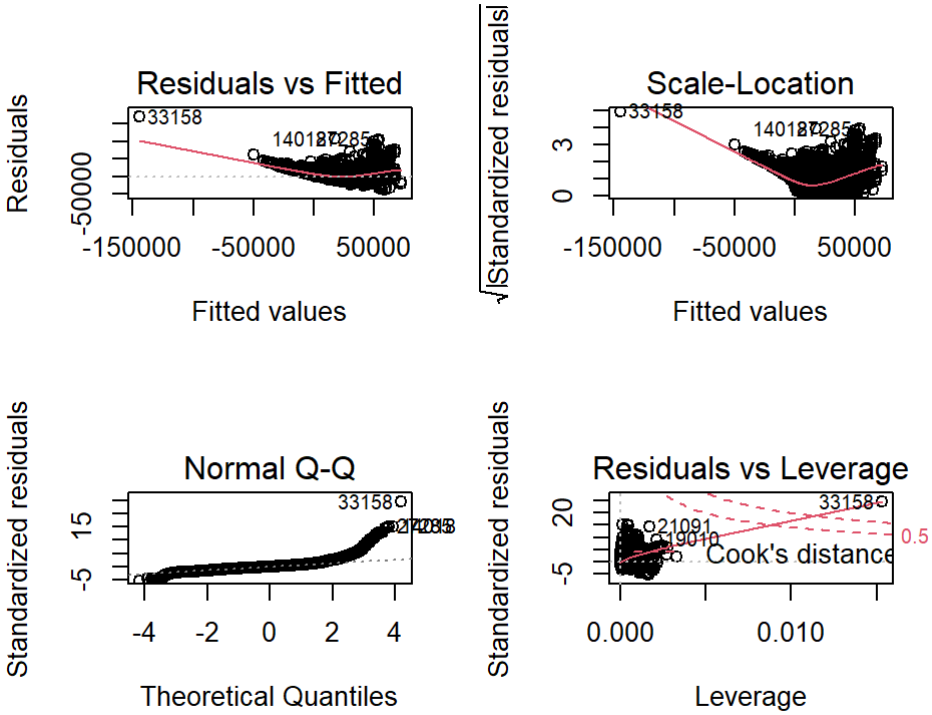
To decide which model is the ideal one, we do the ANOVA analysis:

```
## Analysis of Variance Table
##
## Response: price
##                 Df     Sum Sq    Mean Sq  F value    Pr(>F)
## year             1 1.5542e+12 1.5542e+12 31970.18 < 2.2e-16 ***
## engineSize       1 1.4631e+12 1.4631e+12 30096.28 < 2.2e-16 ***
## transmission     1 1.7009e+10 1.7009e+10   349.89 < 2.2e-16 ***
## Residuals    34200 1.6626e+12 4.8613e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

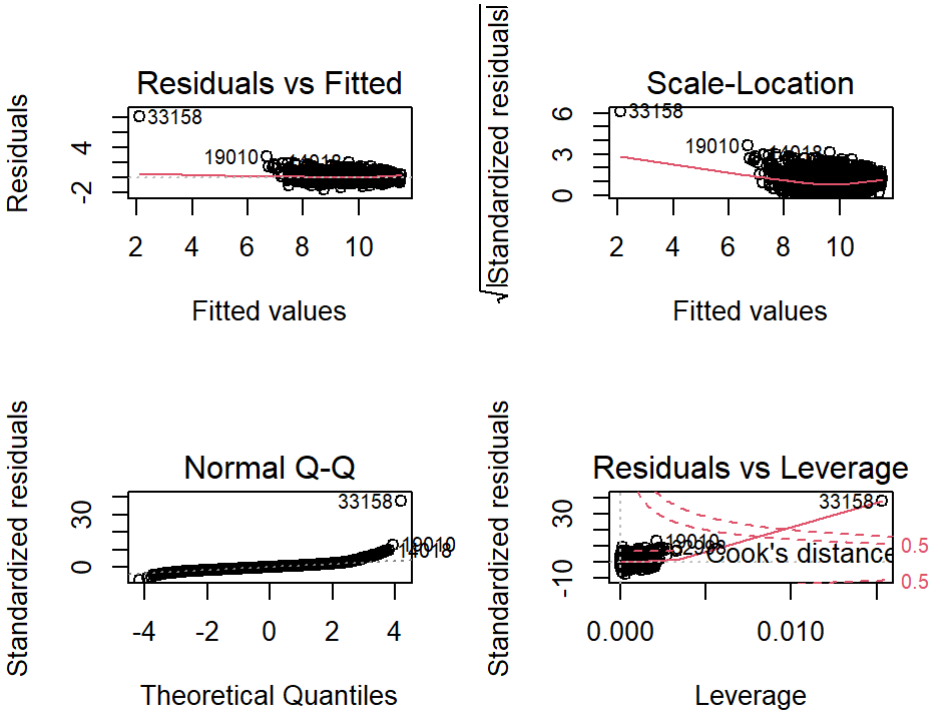
```
## Analysis of Variance Table
##
## Response: log(price)
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## year             1 4060.7  4060.7 88085.5 < 2.2e-16 ***
## engineSize       1 1636.9  1636.9 35507.7 < 2.2e-16 ***
## transmission     1  108.1   108.1  2345.9 < 2.2e-16 ***
## Residuals    34200 1576.6     0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the p-value, both models look significant, so we examine the qqplot and the residual vs fitted plot. Four plots of the model, qqplot, residual vs fitted, etc.:

The plot of the untransformed model:



The plot of the transformed model:



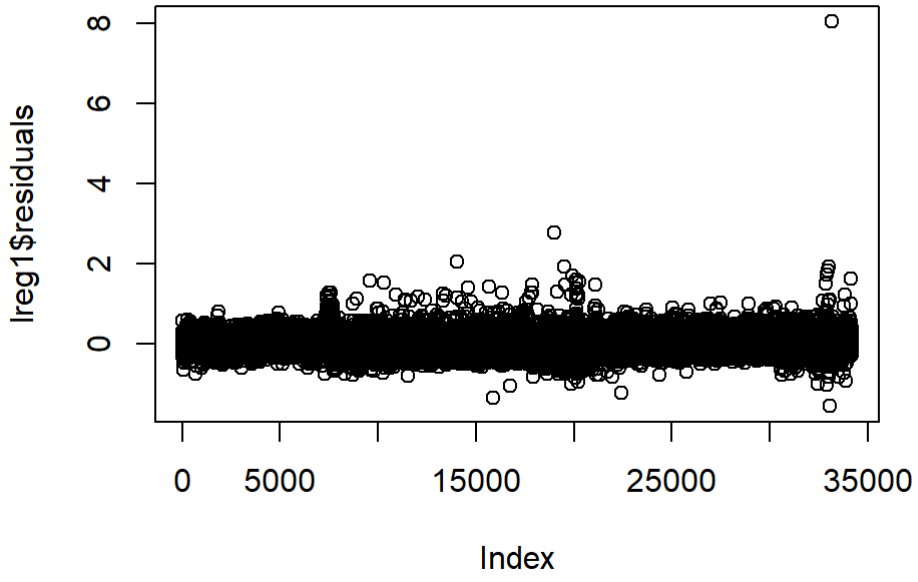
From the qqplot and residual vs fitted plot, we can clearly figure out that the ideal model is the model with logarithm transformation on the response variable price. Compared with the plots generated by the original model, the Residuals v Fitted plot of the transformed model flattens out significantly and the residuals better fit the reference line on the normal qqplot. Therefore, we will select the model with logarithm transformation, which is log(price) ~ year + transmission + engine size.

```
## Start:  AIC=-105240.3
## log(price) ~ year + engineSize + transmission
##
##                Df Sum of Sq    RSS     AIC
## <none>                      1576.6 -105240
## - transmission  1     108.1 1684.7 -102973
## - engineSize    1    1274.3 2850.9  -84980
## - year          1    3575.0 5151.6  -64743
```

```
##
## Call:
## lm(formula = log(price) ~ year + engineSize + transmission, data = ds)
##
## Coefficients:
##  (Intercept)          year    engineSize  transmission
##   -296.53343       0.15151       0.34641       0.07663
```

Besides, we also use stepwise algorithm to choose a model by AIC. And the result shows that the current model is the most ideal one.

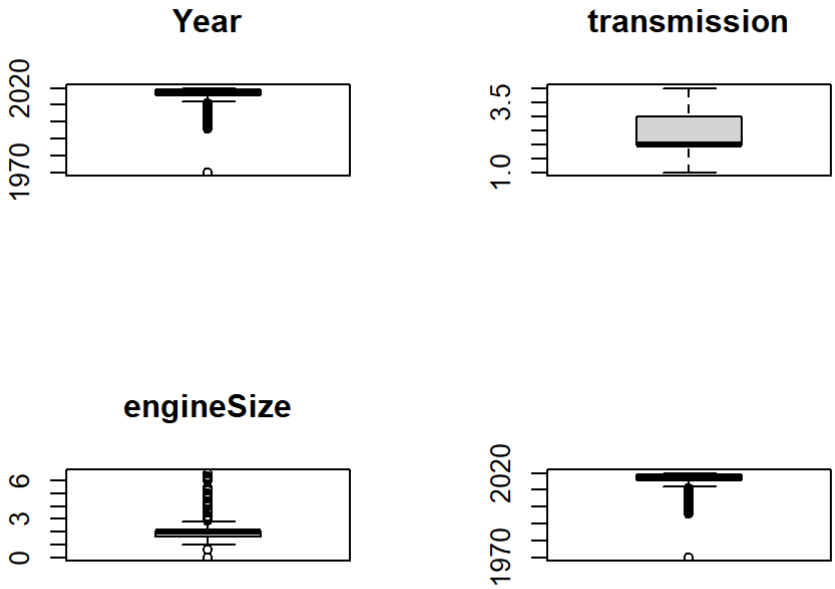
We then examine the residual plot since there are still roughly heteroscedastic residuals returned from the transformed model Residual plot for the transformed model:



From the residual plot, we can clearly see that there are a few outliers, so what we want do next is to remove the outliers and fit the model again.

iv. Outlier Removal

To better perform our data and get more accurate results. We test whether there are any outliers or not, and using *subset()* to remove if we find any outliers.



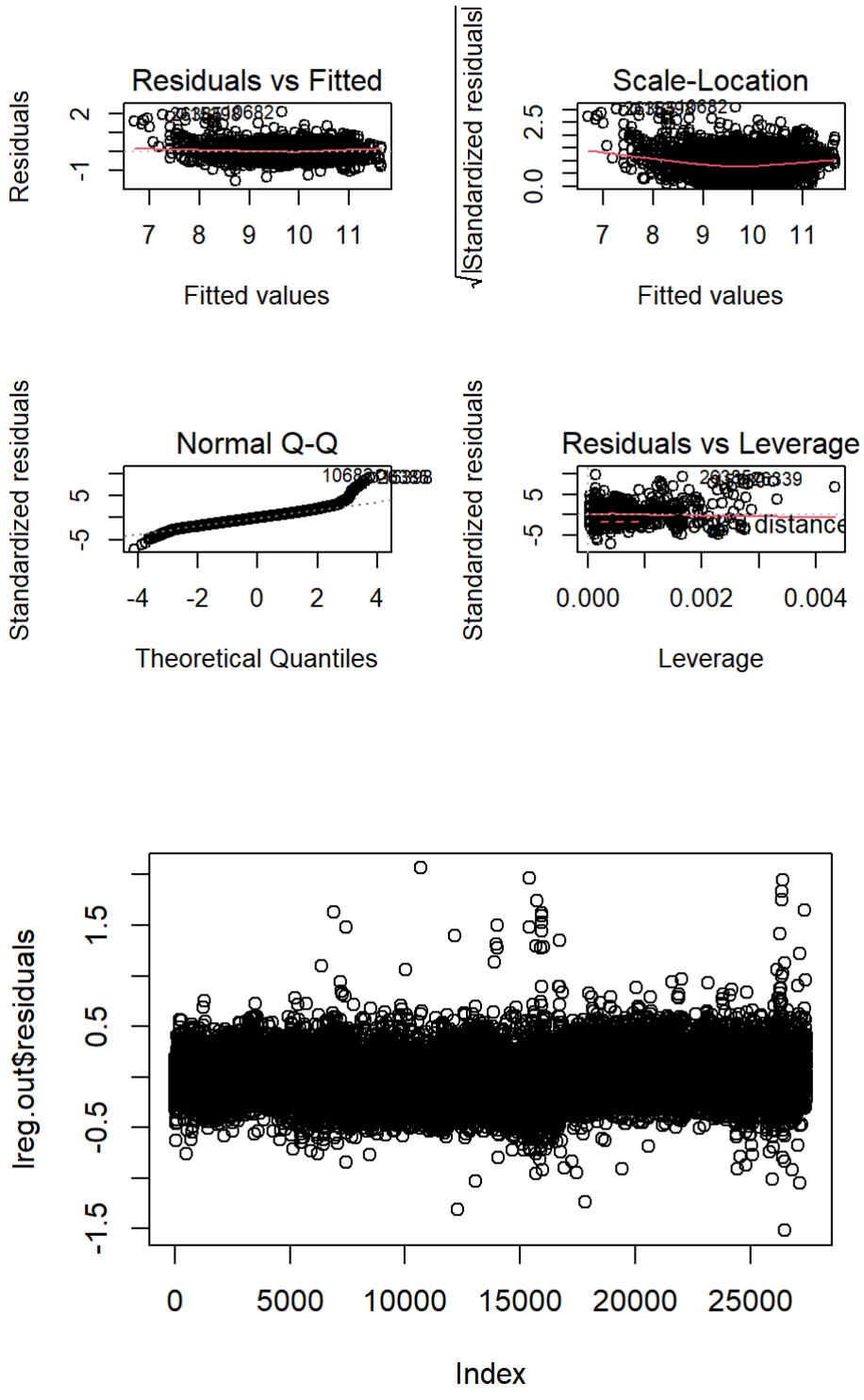

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1970    2016    2017    2017    2019    2020
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.60    2.00    2.06    2.10    6.60
```

The box plot above shows that there are some outliers in the variable year and engine size. We consider data points as outliers when they are beyond the 1st or the 3rd quantile. We then remove those who are considered to be outliers.

After removing the outliers, we fit the regression model again:

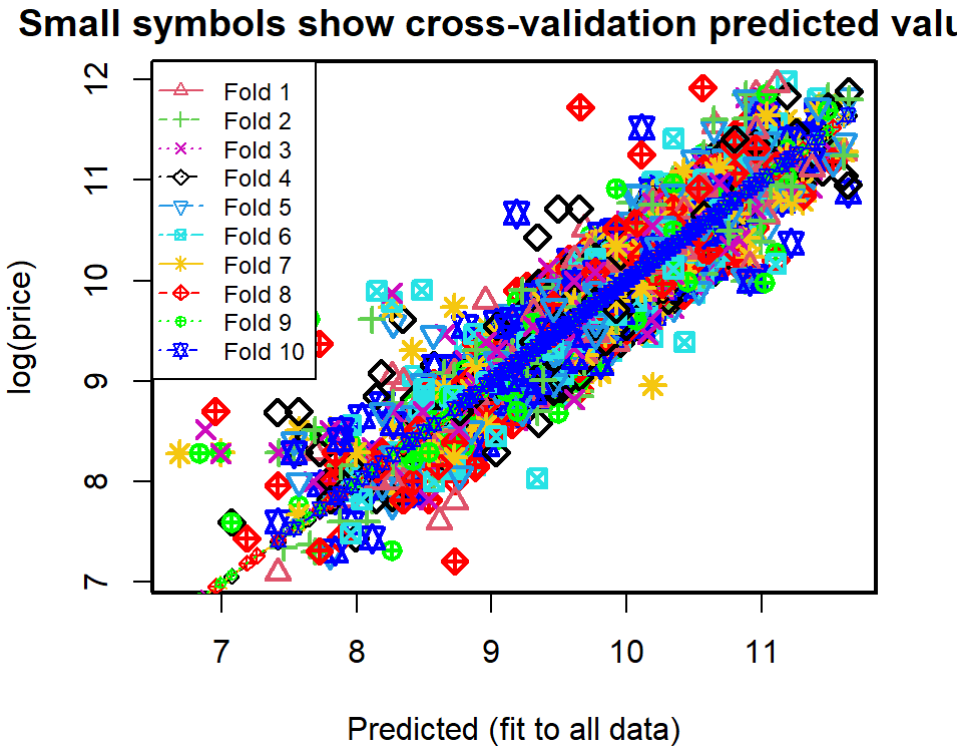
```
##
## Call:
## lm(formula = log(price) ~ year + engineSize + transmission, data = ds.rem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51845 -0.13608 -0.00142  0.12786  2.06713
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.016e+02  1.168e+00 -258.26   <2e-16 ***
## year          1.540e-01  5.794e-04  265.73   <2e-16 ***
## engineSize    3.817e-01  2.501e-03  152.58   <2e-16 ***
## transmission  8.016e-02  1.835e-03   43.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.212 on 27416 degrees of freedom
## Multiple R-squared:  0.8074, Adjusted R-squared:  0.8074
## F-statistic: 3.831e+04 on 3 and 27416 DF,  p-value: < 2.2e-16
```



The p-value of the new model is 0 and the adjusted R-squared is 0.8074, which is higher than the model without removing the outliers. The residual plot and normal Q-Q plot has slightly become more fitted and residuals are closer to 0, which indicating a better fit of the dataset. Therefore, by removing outliers, our model fits better.

v. Cross Validation

After model selection and outliers removal, we now want to assess our model and see what its performance on independent dataset is like. We choose the method Cross Validation to test whether the predicted result is close to the real result. Specifically, we use the k-fold cross validation with the choice that k = 10 since the dataset is large.



In the k-fold cross validation, for each fold, our model is fit to all observations that are not in the fold (that is, the 'training set') and prediction errors are calculated for the observations in the fold (that is, the 'test set'). From the graph, we can see that generally speaking, the majority of the predictions in each fold are close to the real observations, as the trend of the graph is a straight line. It seems like our model fits well based on the result of k-fold cross validation. To verify in a more quantitative way, we will calculate the Root Mean Square Error (RMSE), which is the prediction error.

```
## [1] 0.212
```

The RMSE is 0.212, which is rather small, it shows that our model fits the known data as well as unknown data.

6. Results and Interpretation

Working with the data of used car, we have conducted analysis on what affects the price of used cars. At first sight of all the explanatory variables, based on common knowledge, people will think that all of them may play important roles in considering price of used cars and the first-step ggplot also shows some trend between each variable and price. Therefore, one of our important tasks is to identify which variables make the most contributions to the response variable price. To reach our goal, we use Principle Component Analysis to see which variables are more correlated with the variable price. And by examining the angles between price and other variables, we are able to make decisions. Combined with the correlation matrix, we are able to address our questions of interest. We can conclude that year affects the price of used cars most as it has the highest correlation coefficients and smallest angles with price in PCA. While fuel type is the least effective to price, as it has smallest correlation coefficients and large angle with price.

After understanding the relationship between each variable with the response variable price, we move on to discover whether we can fit a model to predict the price of a used car. From the boxplot we plotted, we find out that most cars are in the year between 2016 and 2020, with a mean of 2017, but there are a few cars that are very old. The oldest is in 1970. The price of old cars are sometimes not accurate and they may have different criterion compared to relative new cars.They will act as outliers in the model so we decide to remove them. Similarly, for engine size, cars with very large or small engine size are rare and will have abnormal prices, so we also remove them. Having cleaned the data and chosen the model, we move on to validate the effectiveness of our model. We choose the method of k-fold cross validation to test if the predicted results are

close to real observations. The result of the test, shown by the graph and a very low prediction error, indicating that predictions based on our model are almost the same as the real price, suggesting a good fit. Thus, we can address our last question of interest that by using year, transmission and engine size, it is feasible to predict the price of a used car.

7. Conclusion

Although buying or selling used cars is now becoming a fashion, as it's not only economical but also eco-friendly in terms of reducing resources' consumption, what to consider when one has the intention to buy/sell a car is still a challenging question. In other words, many people are deterred from entering the used car market because they have no idea on what to look at when appraising the used cars. Therefore, given such a fact, we got inspired to generate this project. As we have mentioned previously, we are studying how different factors affect the price of used cars, and what is the necessary information in order to predict the price of a used car. In such a way, we are aiming to give the readers a sense on what they should pay close attention to when valuing the price of cars in the used car market. This can also potentially be viewed as a way to encourage people to enter the used cars market by providing information that people might be interested in under the support of statistics.

From the results, we can conclude that all of our explanatory variables affect the response variable, price to a certain degree. However, among all of the explanatory variables that we designed, year affects most to the price of used car and the fuel type is the least effective to price. In addition, we state that it's sufficient to predict the price of a used car by having the information of the year, transmission and engine size.

In general, our explanatory variables based on the dataset we have contain most of the critical factors that are potentially related to the response variable, price. However, some more factors, such as cars' color and cars' damages are also worth adding into the dataset, and looking into in future research. Specifically, some new cars' price varies based on color, so whether used cars' color will affect the price or not would be an interesting question to investigate. Additionally, the condition of the cars would definitely affect the price, but considering the damages are vary, including scratches, dents and so on, it would also be interesting to test on the relationship between the different types of damages and price. Both of these require additional data collection, and the data collection for car's damages would potentially be complex. Since it might potentially be a challenge to correctly categorize the type of damages and define damages based on the same standard. For further research, we can improve our studies by expanding our dataset to have more explanatory variables to have a more comprehensive understanding on explanatory variables' effects on price.

Member Contributions:

Youjia Wang: Responsible for data set cleaning and data visualization.

Yuehe Wen: Responsible for statistical analysis including linear regression and variable selection.

Jiaye Chen: Responsible for result interpretation and report integration.

References:

https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes
(https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes)

Appendix

```
knitr::opts_chunk$set(
    error = FALSE,
    message = FALSE,
    warning = FALSE,
    echo = FALSE, # hide all R codes!!
    fig.width=5, fig.height=4,#set figure size
    fig.align='center',#center plot
    options(knitr.kable.NA = ''), #do not print NA in knitr table
    tidy = FALSE #add line breaks in R codes
)
library(tidyverse)
library(tidyr)
library(dplyr)
library(ggplot2)
audi <- read_csv("audi.csv")  #checking wether there are NAs in the csv file
#0 means no NA in audi's used car markets data collection

audi[duplicated(audi) | duplicated(audi, fromLast=TRUE),]

#replicate rows
audirepeat <- audi[duplicated(audi) | duplicated(audi, fromLast=TRUE),]
audi <- anti_join(audi,audirepeat)
# the new audi dataset aleady remove the repeat rows.
audi$brand <- rep(c("Audi"), times = 10483)
# creat a brand columns in order to combine with orther brands csv.
#final useful dataset

#do the same cleaning process for other data
bmw <- read_csv("bmw.csv")

bmw.dup <- bmw[duplicated(bmw) | duplicated(bmw, fromLast=TRUE),]
bmwrepeat <- bmw[duplicated(bmw) | duplicated(bmw, fromLast=TRUE),]
bmw <- anti_join(bmw,bmwrepeat)
bmw$brand <- rep(c("BMW"), times = dim(bmw)[1])



merc <- read_csv("merc.csv")

merc.dup <- merc[duplicated(merc) | duplicated(merc, fromLast=TRUE),]
# Although the number of rows in the merc.csv is larger than the previous two
# csv, it shows that there is more repeat rows in this file as well.
mercrepeat <- merc[duplicated(merc) | duplicated(merc, fromLast=TRUE),]
merc <- anti_join(merc,bmwrepeat)
merc$brand <- rep(c("Benz"), times = dim(merc)[1])



ds <- rbind(audi,bmw, merc)

trans.unq <- unique(ds$transmission)

ds$transmission[ds$transmission == "Manual"] = 1
ds$transmission[ds$transmission == "Automatic"] = 2
ds$transmission[ds$transmission == "Semi-Auto"] = 3
ds$transmission[ds$transmission == "Other"] = 4
ds$transmission <- as.numeric(ds$transmission)
```

```
ds <- na.omit(ds)

fuel.unq <- unique(ds$fuelType)

ds$fuelType[ds$fuelType == "Petrol"] = 1
ds$fuelType[ds$fuelType == "Diesel"] = 2
ds$fuelType[ds$fuelType == "Hybrid"] = 3
ds$fuelType[ds$fuelType == "Electric"] = 4
ds$fuelType[ds$fuelType == "Other"] = 5
ds$fuelType <- as.numeric(ds$fuelType)

numericds <-cbind(ds[,2],ds[,3],ds[,4],ds[,5],ds[,6],ds[,7],ds[,8],ds[9])
cor.ds <- round(cor(numericds),digits = 2)
head(ds)
library(ggplot2)
ggplot(data = ds) +
  geom_point(mapping = aes(x = year, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "Registration year vs price scatterplot in £")
ggplot(data = ds) +
  geom_point(mapping = aes(x = transmission, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "type of gearbox vs price in £")
ggplot(data = ds) +
  geom_point(mapping = aes(x = mileage, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "distance used vs price in £")
ggplot(data = ds) +
  geom_point(mapping = aes(x = fuelType, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "engine fuel vs price  in £")
ggplot(data = ds) +
  geom_point(mapping = aes(x = tax, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "road tax vs price in £")
ggplot(data = ds) +
  geom_point(mapping = aes(x = mpg, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "miles per gallon vs pricein £")

ggplot(data = ds) +
  geom_point(mapping = aes(x = engineSize, y = price, shape=brand, color=brand))+
  facet_wrap(~brand)+
  labs(title = "size in litres vs price in £")
cor.ds
pr.out <- prcomp(numericds,scale.=TRUE)
biplot(pr.out,scale=0)
year_sp <- qplot(x = year, y = price, data = ds)
trans_sp <- qplot(x = transmission, y = price, data = ds)
engine_sp <- qplot(x = engineSize, y = price, data = ds)

gridExtra::grid.arrange(year_sp, trans_sp,engine_sp,ncol=1) #This graph shows the linear rela
tionship between each variable and price.
lreg <- lm(price ~ year + engineSize + transmission, data = ds)
summary(lreg)
lreg1 <- lm(log(price) ~ year + engineSize + transmission, data = ds) ## model with log trans
```

```
formation
summary(lreg1)
anova(lreg)
anova(lreg1)
layout(matrix(c(1,2,3,4),2,2))
plot(lreg)## four plots of the model, qqplot, residual vs fitted, etc.
par(mfrow=c(1,1))


## From the qq plot, we can see that the transformed model is better.
layout(matrix(c(1,2,3,4),2,2))
plot(lreg1) ## qqplot and other three plots of the transformed model
par(mfrow=c(1,1))
step(lreg1)
## residual plot for the transformed model
plot(lreg1$residuals)
### Outlier removal
par(mfrow=c(2,2))
boxplot(ds$year,main="Year")
boxplot(ds$transmission,main="transmission")
boxplot(ds$engineSize,main="engineSize") ## boxplot showing that there are some outliers
boxplot(ds$year)
par(mfrow=c(1,1))

summary(ds$year) ## remove outliers that are beyond 1st quantile
ds.rem <- subset(ds, ds$year >= 2016)

summary(ds$engineSize) ## remove outliers that are beyond 1st quantile
ds.rem <- subset(ds, ds$engineSize >= 1.600)
lreg.out <- lm(log(price) ~ year + engineSize + transmission, data = ds.rem)
summary(lreg.out)
layout(matrix(c(1,2,3,4),2,2))
plot(lreg.out)

par(mfrow=c(1,1))
plot(lreg.out$residuals)

install.packages("DAAG", repos = "https://cran.rstudio.com")
install.packages("lattice", repos = "https://cran.rstudio.com")
library(lattice)
library(DAAG)
ds.rem <- as.data.frame(ds.rem)
cross.v <- cv.lm(ds.rem, lreg.out, 10)

RMSE<-sqrt(mean((cross.v$cvpred - cross.v$`log(price)`)^2))
RMSE
```