UNIVERSITY OF PADUA

Department of Mathematics

Master's Degree in Data Science

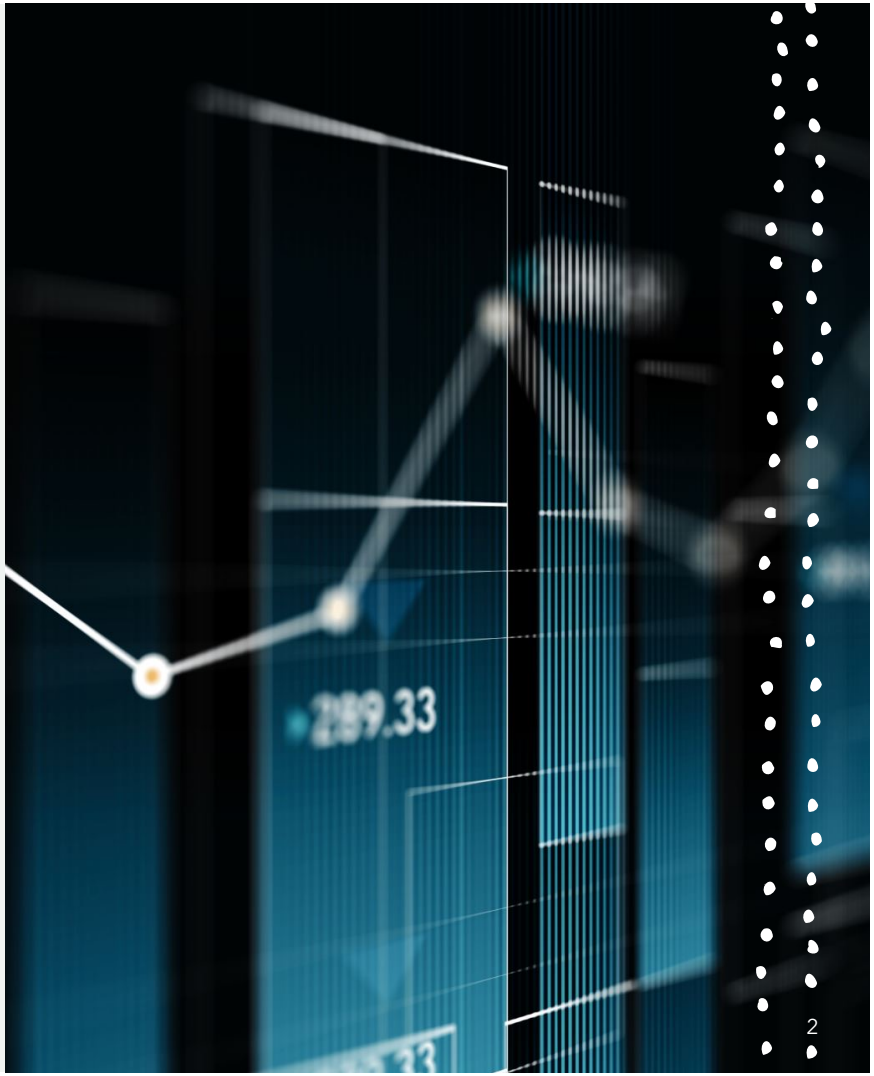
Statistical Learning Project (Mod B)

Car Price Prediction

Roya Ghamari - Alireza Saberi – Nahid Jahanianarange

July 2023

1. Introduction
2. Preparation of the Dataset
3. Exploratory Data Analysis (EDA)
4. Correlation Plot of Features
5. Modeling the Data
6. Conclusion



1. Introduction:

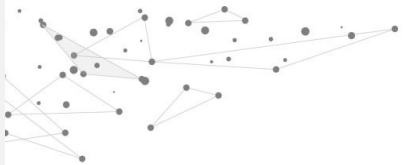
1.1 goal of project

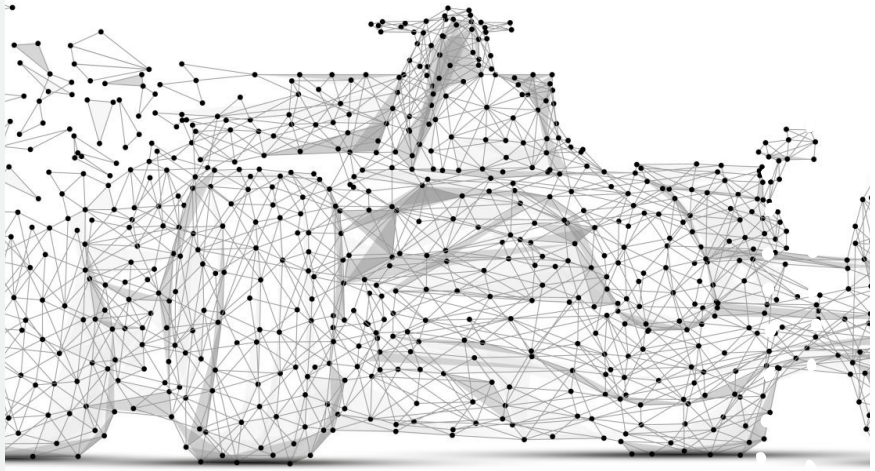
1.2 explanation about dataset

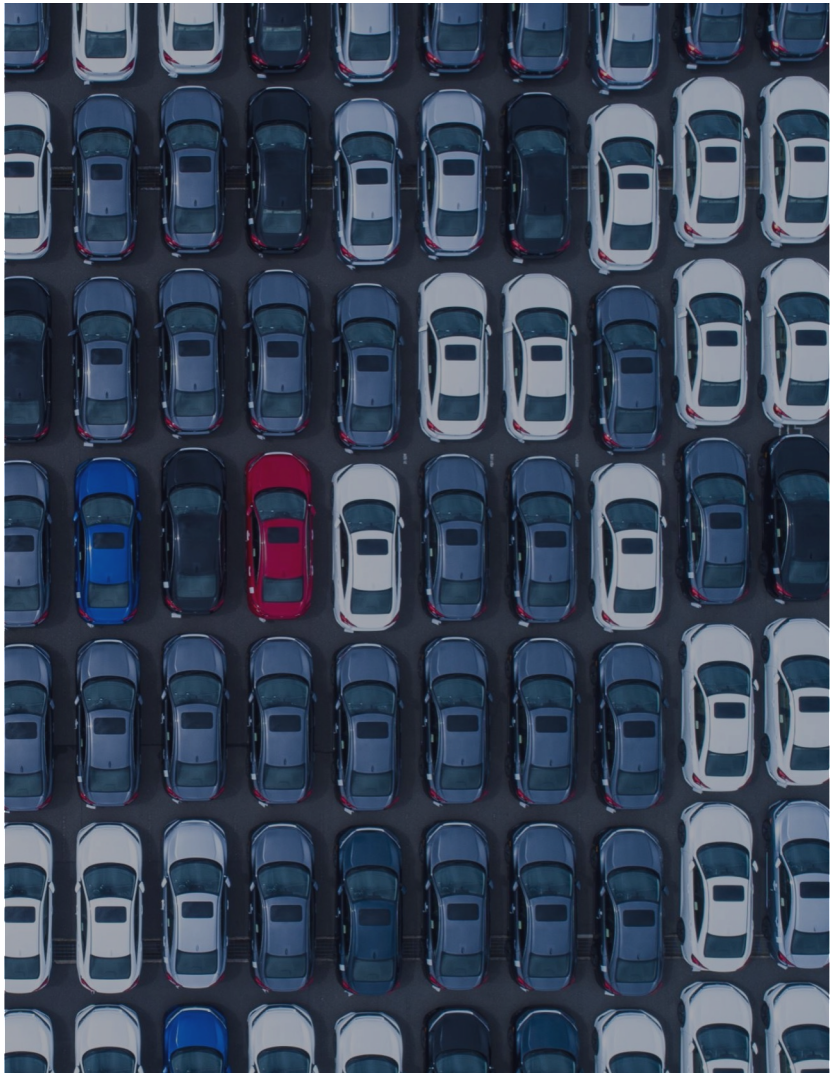
1.3 Path of the project

1.1 Goal of the project

- To predict how much cars cost using statistical techniques.
- Creating a smart system that can guess the price of a car
- This will help people who want to buy or sell cars
- to enhance transparency and efficiency in the automotive market







1.2 explanation about dataset

- Dataset of Car Dekho is about used cars and various features of them
- To get more information you can visit https://www.cardekho.com
- This dataset consists of sale prices of 8128 cares sold between 1983 and 2020.

1.3 Path of the project

1. Our main goal is to figure out what factors affect the price of used cars the most
2. Discovering correlations between data by using simple linear regression model
3. Using linear , multiple, polynomial regressions and lasso to find a proper model to predict the price.

2. Preparation of the Dataset

2.1 Variables description

2.2 Preprocessing

2.3 Cleaning the Data

2.1 variable description

- **Name**
- **Year**
- **Selling_price**
- **Km_driven**
- **Fuel**
- **Seller_type**
- **Owner**
- **Transmission**
- **Mileage**
- **Engine**
- **Max_power**
- **Torque**
- **Seats**

An overview of dataset

name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5
Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5
Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5
Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5
Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5
Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm	5
Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3 km/ kg	1061 CC	57.5 bhp	7.8@ 4,500(kgm@ rpm)	5
Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1 kmpl	796 CC	37 bhp	59Nm@ 2500rpm	4
Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59 kmpl	1364 CC	67.1 bhp	170Nm@ 1800-2400rpm	5
Ford Figo Diesel Celebration Edition	2013	200000	169000	Diesel	Individual	Manual	First Owner	20.0 kmpl	1399 CC	68.1 bhp	160Nm@ 2000rpm	5
Renault Duster 110PS Diesel RxL	2014	500000	68000	Diesel	Individual	Manual	Second Owner	19.01 kmpl	1461 CC	108.45 bhp	248Nm@ 2250rpm	5
Maruti Zen LX	2005	92000	100000	Petrol	Individual	Manual	Second Owner	17.3 kmpl	993 CC	60 bhp	78Nm@ 4500rpm	5
Maruti Swift Dzire VDi	2009	280000	140000	Diesel	Individual	Manual	Second Owner	19.3 kmpl	1248 CC	73.9 bhp	190Nm@ 2000rpm	5
Maruti Swift 1.3 VXi	2007	200000	80000	Petrol	Individual	Manual	Second Owner					
Maruti Wagon R LXI Minor	2009	180000	90000	Petrol	Individual	Manual	Second Owner	18.9 kmpl	1061 CC	67 bhp	84Nm@ 3500rpm	5
Mahindra KUV 100 mFALCON G80 K8 5str	2016	400000	40000	Petrol	Individual	Manual	First Owner	18.15 kmpl	1198 CC	82 bhp	115Nm@ 3500-3600rpm	5
Maruti Ertiga SHVS VDI	2016	778000	70000	Diesel	Individual	Manual	Second Owner	24.52 kmpl	1248 CC	88.5 bhp	200Nm@ 1750rpm	7
Hyundai i20 1.4 CRDi Asta	2012	500000	53000	Diesel	Individual	Manual	Second Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5
Maruti Alto LX	2002	150000	80000	Petrol	Individual	Manual	Second Owner	19.7 kmpl	796 CC	46.3 bhp	62Nm@ 3000rpm	5
Hyundai i20 2015-2017 Asta 1.4 CRDi	2016	680000	100000	Diesel	Individual	Manual	First Owner	22.54 kmpl	1396 CC	88.73 bhp	219.7Nm@ 1500-2750rpm	5
Mahindra Verito 1.5 D4 BSIII	2011	174000	100000	Diesel	Individual	Manual	Second Owner	21.0 kmpl	1461 CC	64.1 bhp	160Nm@ 2000rpm	5
Honda WR-V i-DTEC VX	2017	950000	50000	Diesel	Individual	Manual	First Owner	25.5 kmpl	1498 CC	98.6 bhp	200Nm@ 1750rpm	5
Maruti Swift Dzire ZDI	2015	525000	40000	Diesel	Individual	Manual	First Owner	26.59 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5
Maruti SX4 ZDI	2012	600000	72000	Diesel	Individual	Manual	First Owner	21.5 kmpl	1248 CC	88.8 bhp	200Nm@ 1750rpm	5

2.2 Preprocessing

- 1. Dimension of Dataset:

  8128 , 13

- 2. The amount of null values:

  221

- 3. Inspecting the response variable, in this case, ”Selling_price”:

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  29999   254999   450000   638272   675000 10000000
```

- 4. log Transformation

2.3 Cleaning the Data

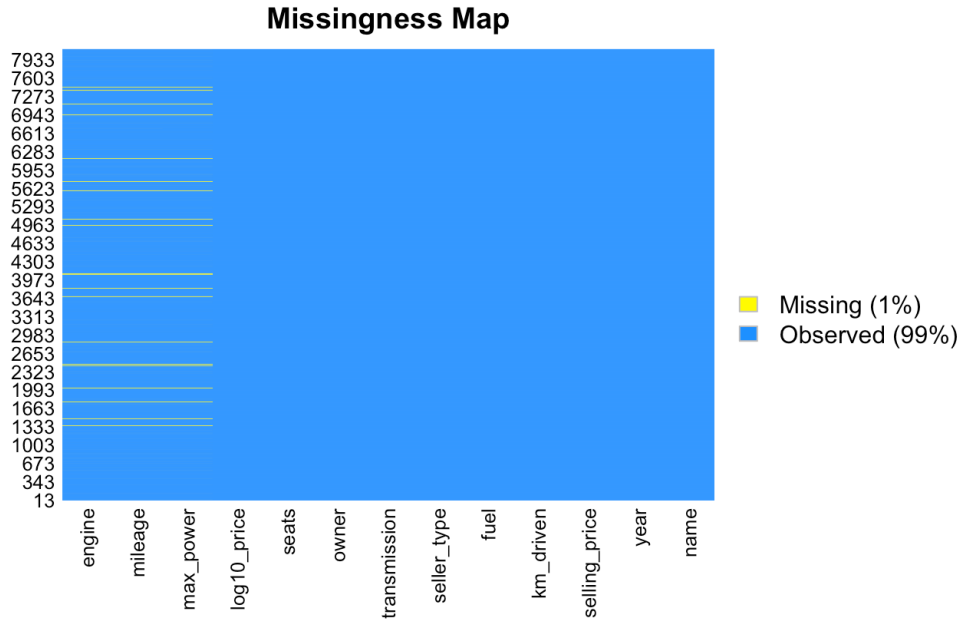
Removing string values

mileage	engine	max_power
23.4 kmpl	1248 CC	74 bhp
21.14 kmpl	1498 CC	103.52 bhp
17.7 kmpl	1497 CC	78 bhp
23.0 kmpl	1396 CC	90 bhp
16.1 kmpl	1298 CC	88.2 bhp
20.14 kmpl	1197 CC	81.86 bhp

Encoding data

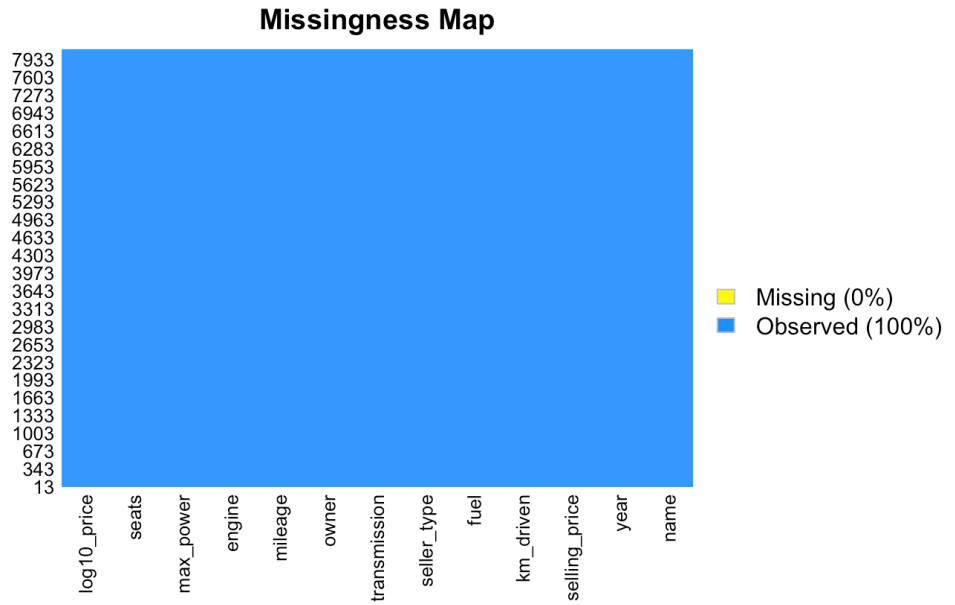
fuel	seller_type	transmission	owner
Diesel	Individual	Manual	First Owner
Diesel	Individual	Manual	Second Owner
Petrol	Individual	Manual	Third Owner
Diesel	Individual	Manual	First Owner
Petrol	Individual	Manual	First Owner
Petrol	Individual	Manual	First Owner
LPG	Individual	Manual	First Owner
Petrol	Individual	Manual	Second Owner
Diesel	Individual	Manual	First Owner

Replacing null values which are not “NA” type like white space and then Checking for missing values:



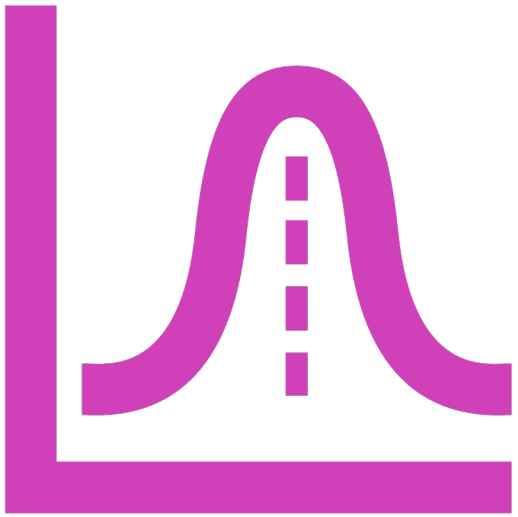
name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats	log10_price
0	0	0	0	0	0	0	0	221	221	215	0	0

Replacing the missing values with the mean value



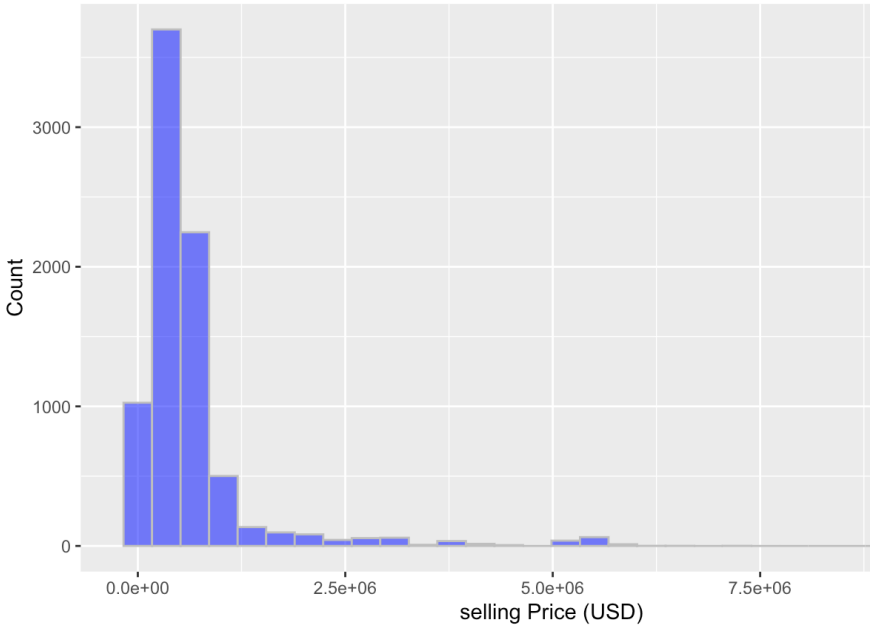
name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats	log10_price
0	0	0	0	0	0	0	0	0	0	0	0	0

3. Exploratory Data Analysis (EDA)



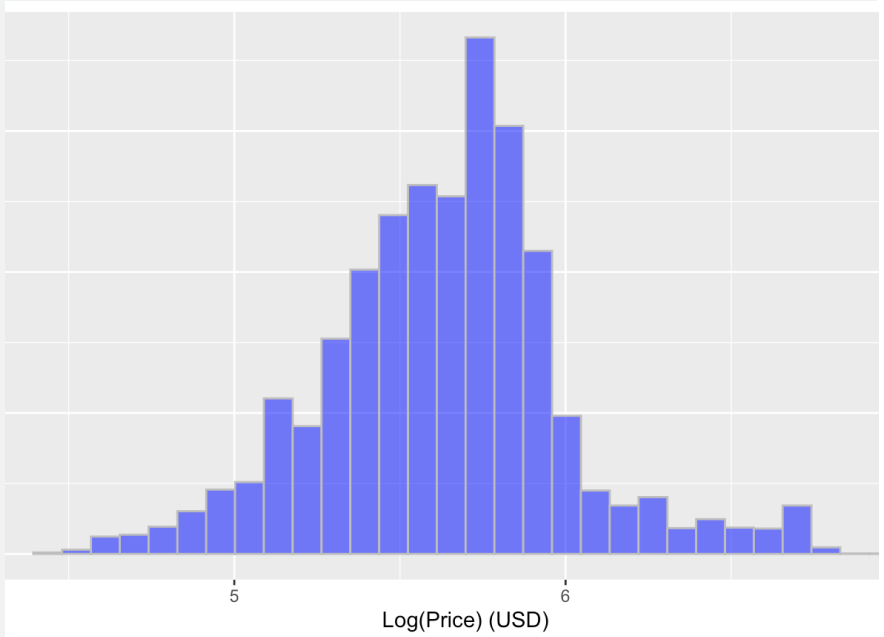
Plotting the distribution of ”Selling Price”

- Selling price of majority of cars is less than two and half million dollars
- The distribution of the target variable price is right-skewed



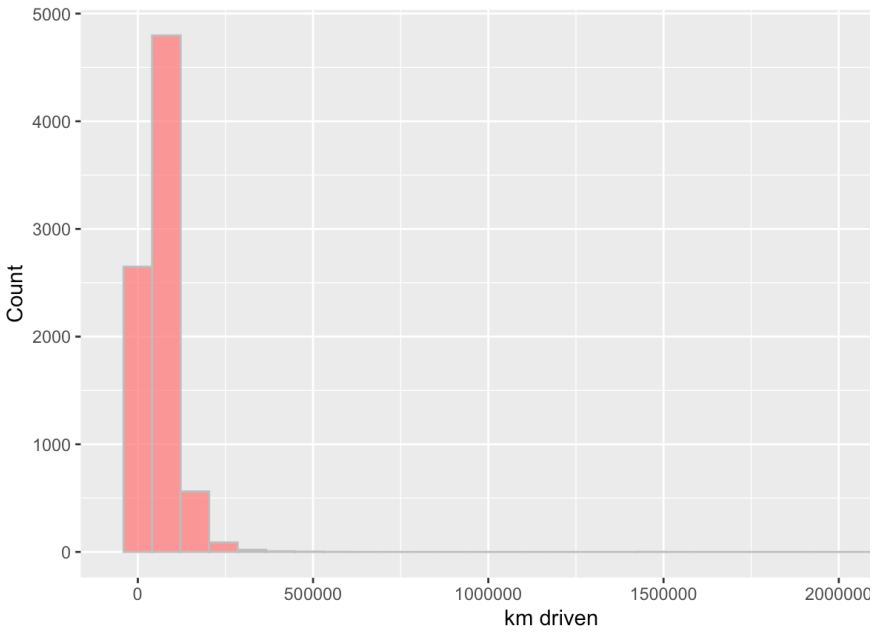
Applying a logarithm transformation on “Selling Price” Distribution

- The distribution of the logarithmic transformation of price becomes bell-shaped



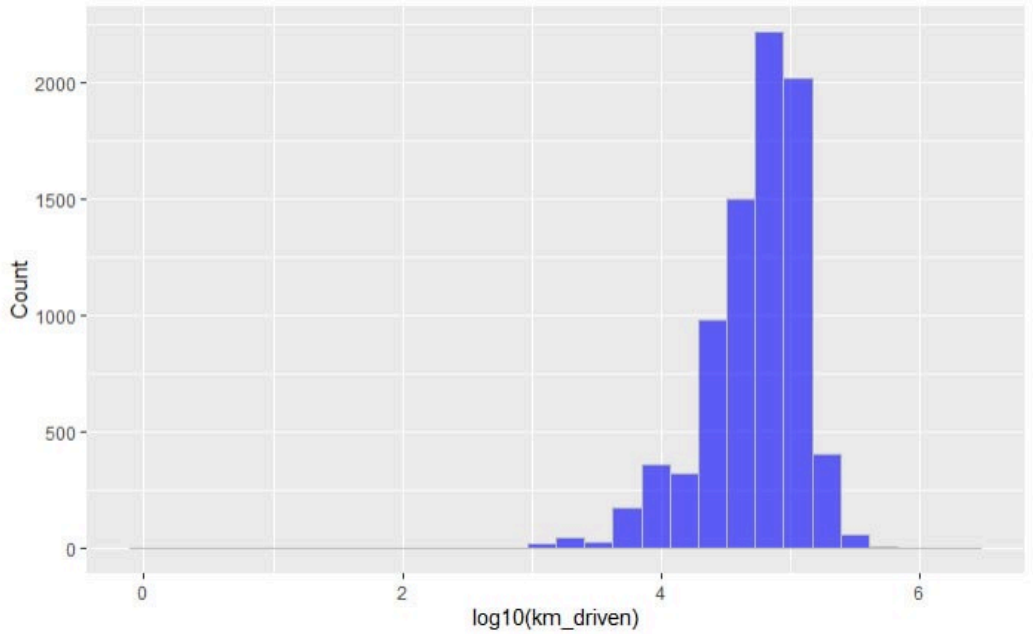
Plotting the distribution of ” km_driven”

- Most of the cars have been driven less than 500000 kms
- The graph is right-skewed



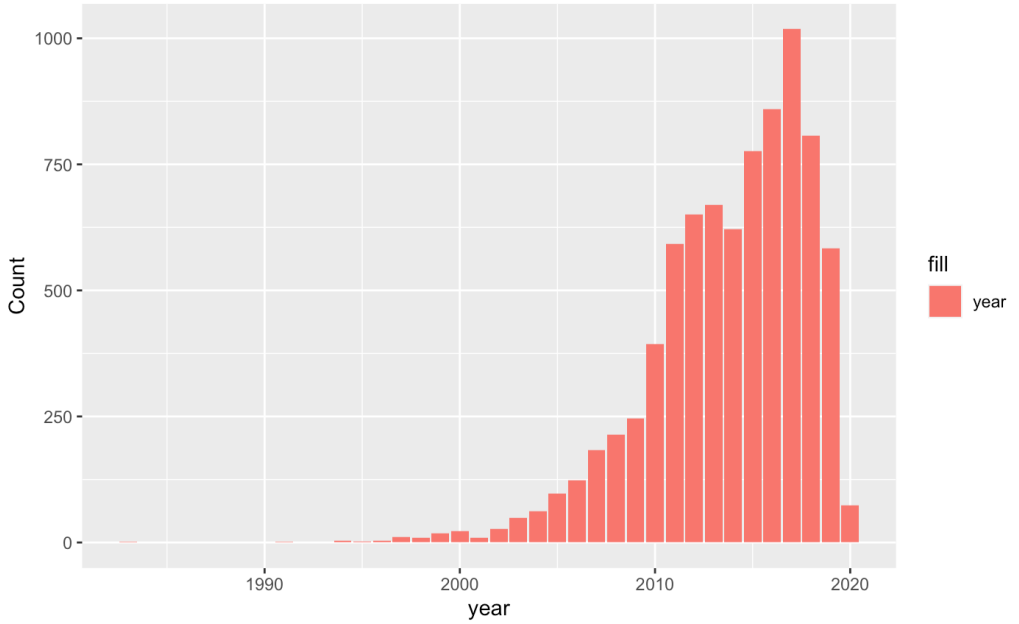
Applying a logarithm transformation on “km_driven” Distribution

- The distribution of the logarithmic transformation of Kms becomes bell-shaped



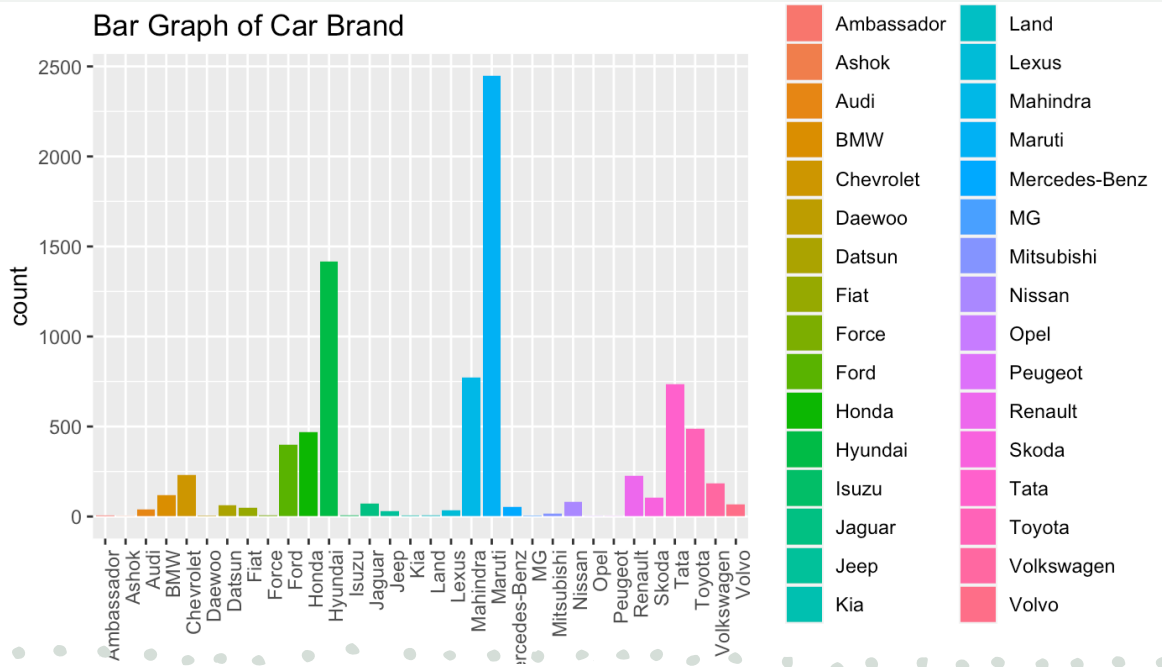
Plotting the distribution of ”Year”

- Most of the sold cars in our dataset is after 2010
- Our graph is left-skewed



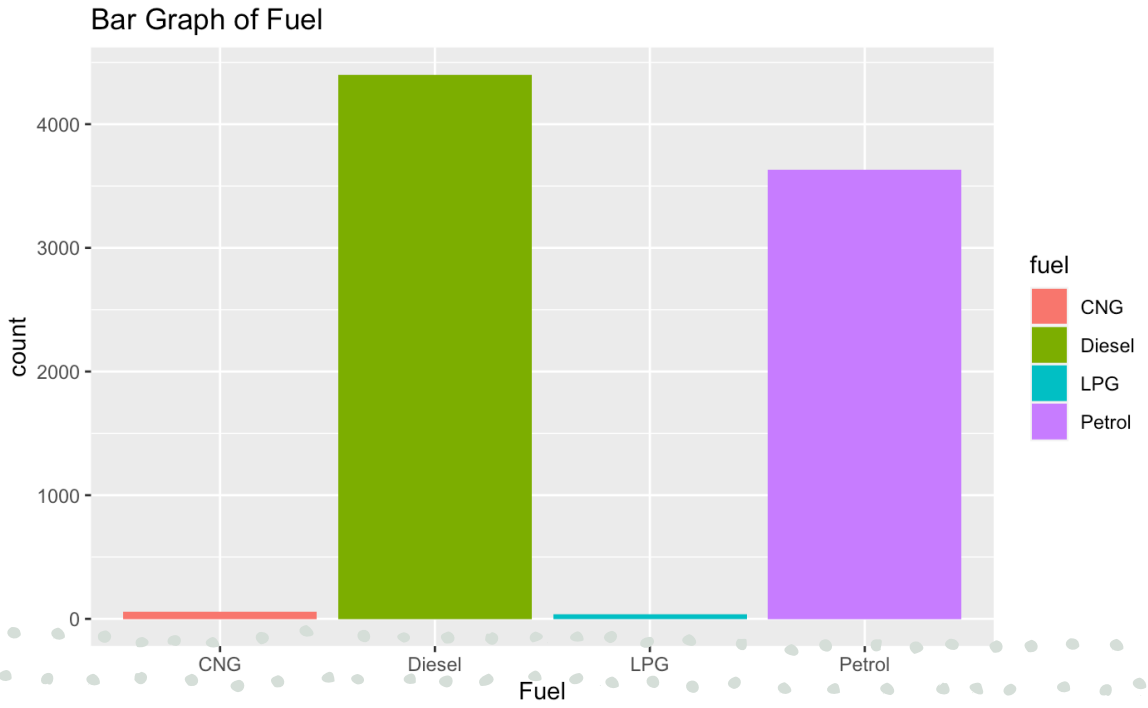
Bar graph of Car Brand

- Highest number of cars fall into “Marutti” followed by “Hyundai” , “Mahindro” and “Tata” brans in order.



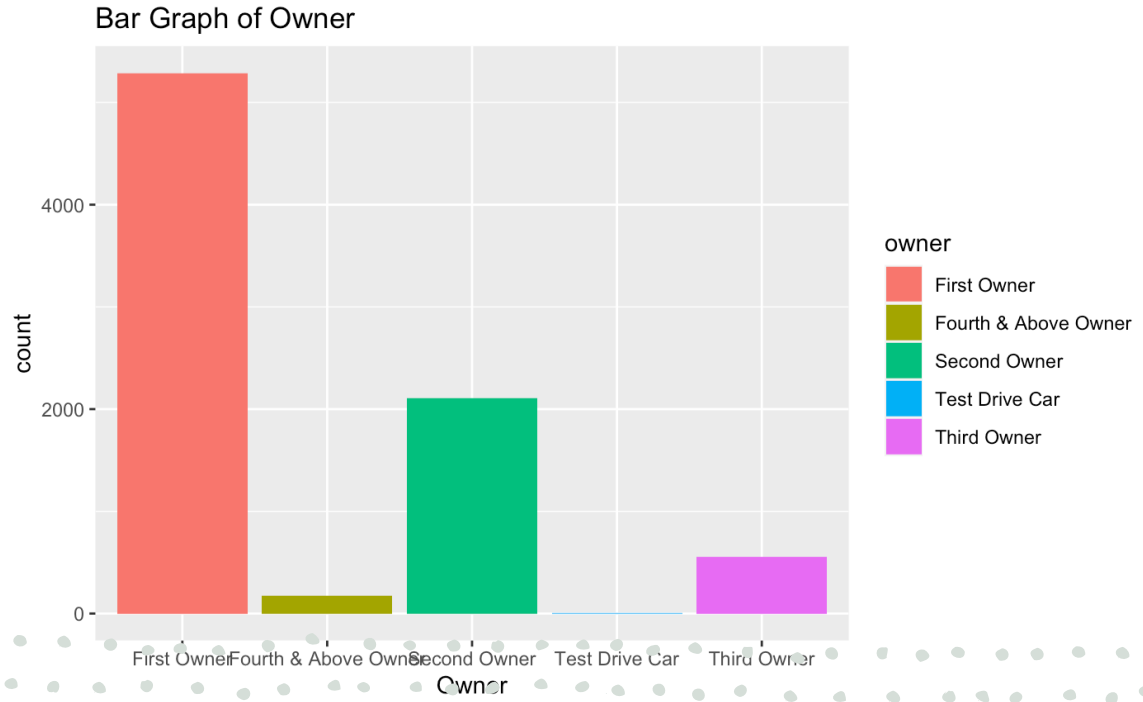
Bar graph of Fuel

- Diesel and Petrol have the highest ownership for the fuel types
- Most of the cars fall into Diesel category followed by Petrol
- Very few cars fall into CNG and LPG category



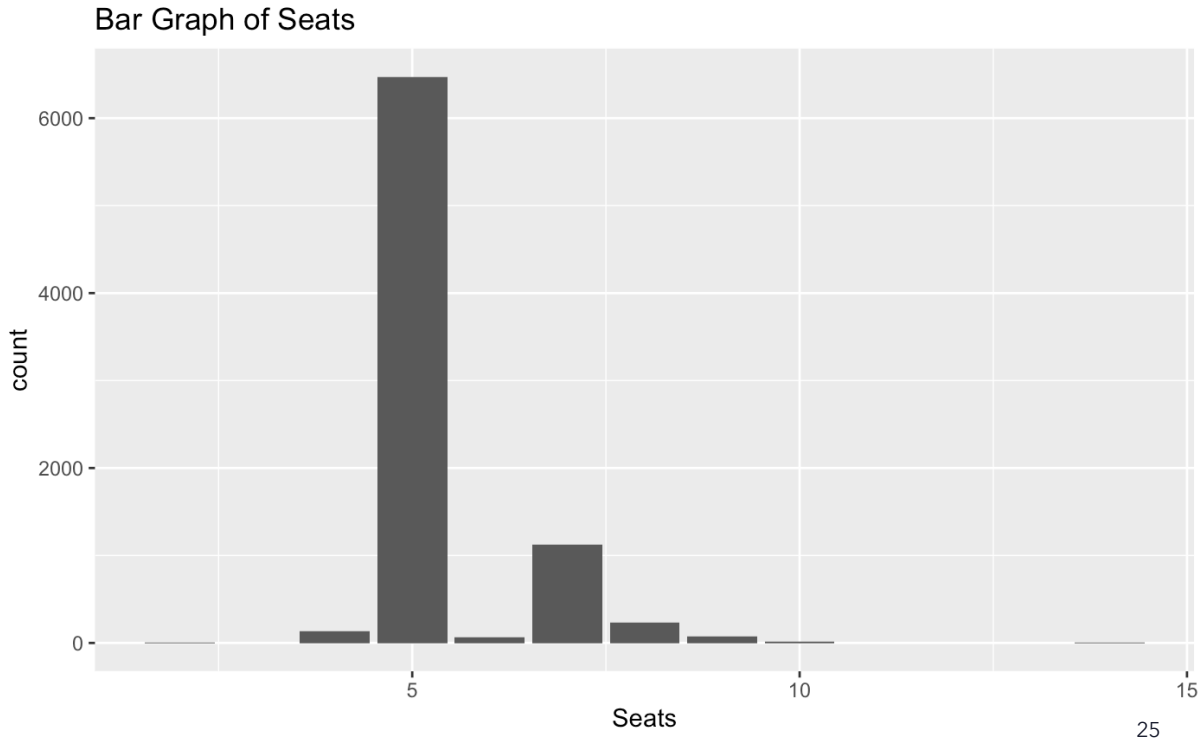
Bar graph of Owner

- Most of the cars are owned by first owners

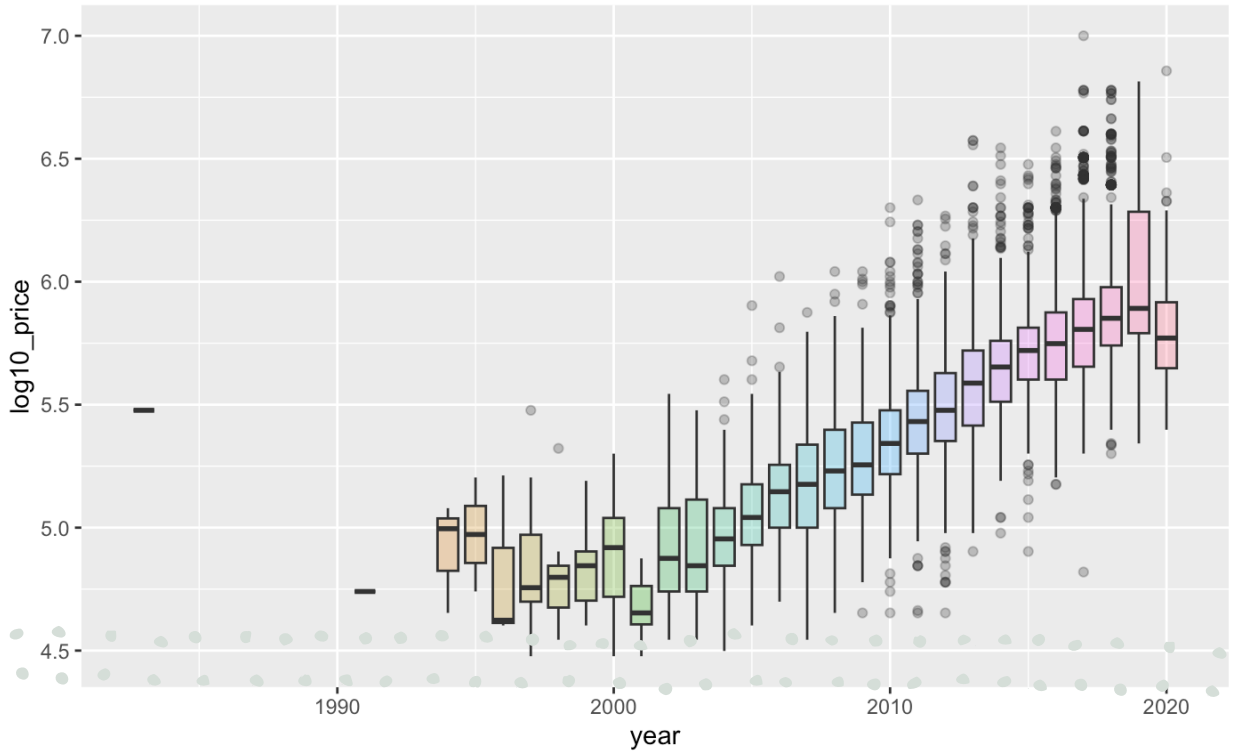


Bar graph of Seats

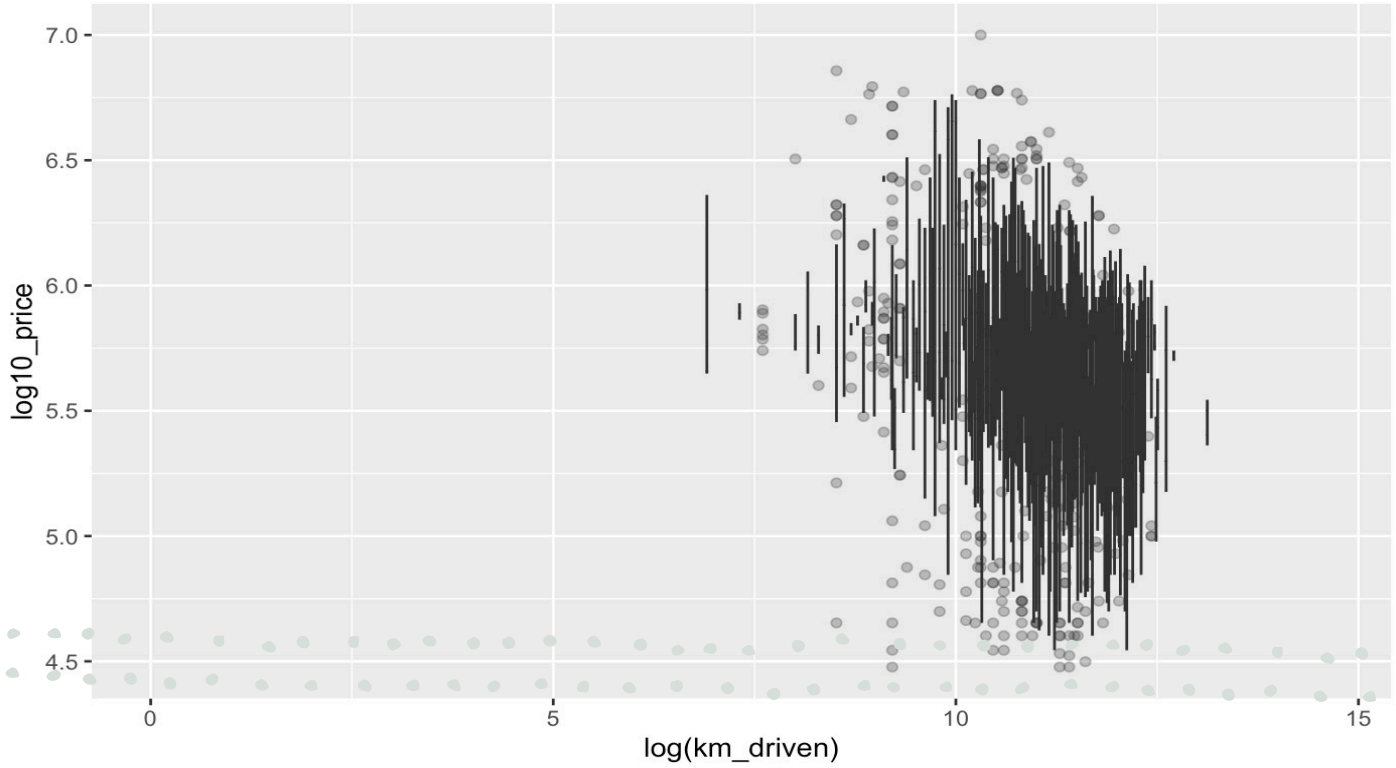
- Majority of the seats in the cars are 5
- compact cars are the most dominant one in the car market.



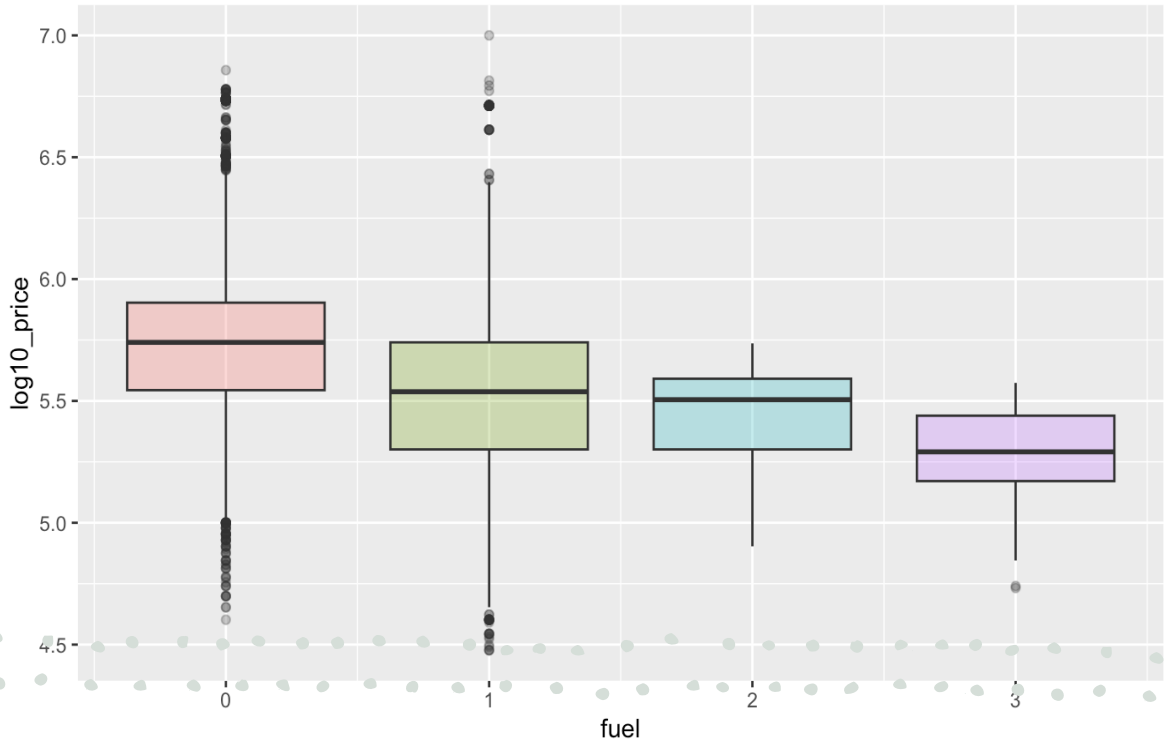
Box plot of year



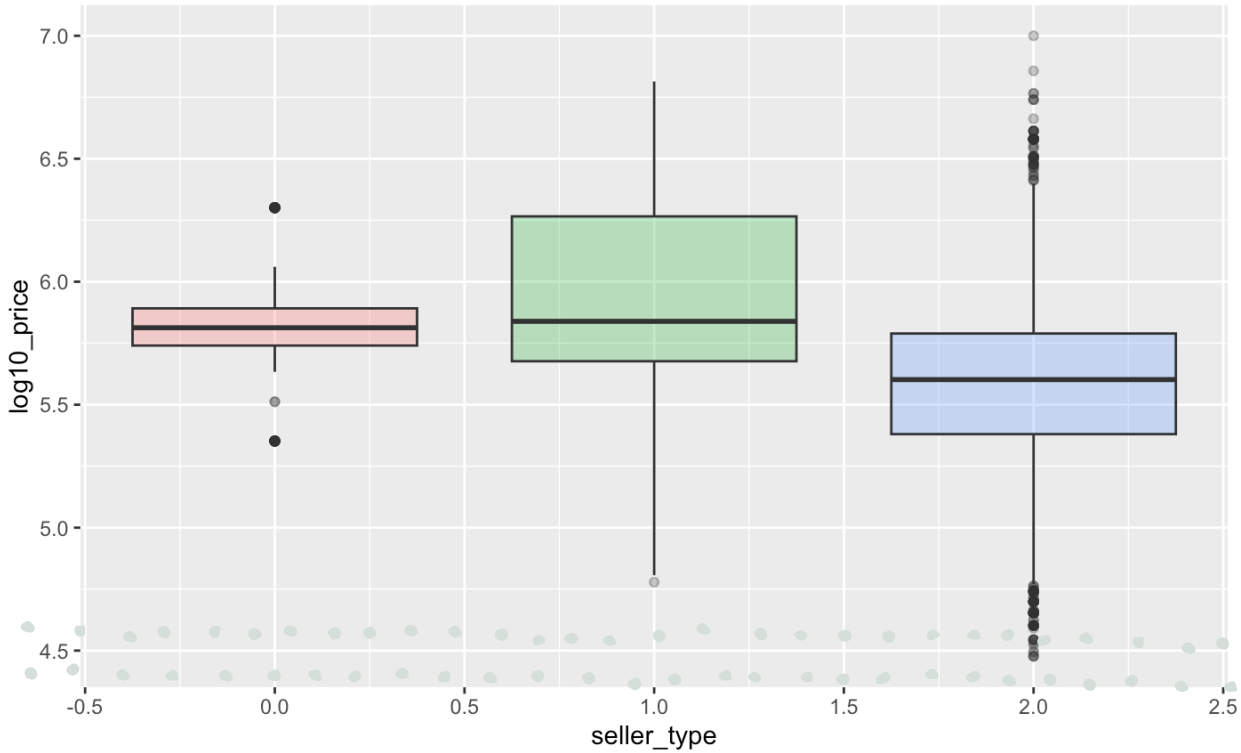
Box plot of Km Driven



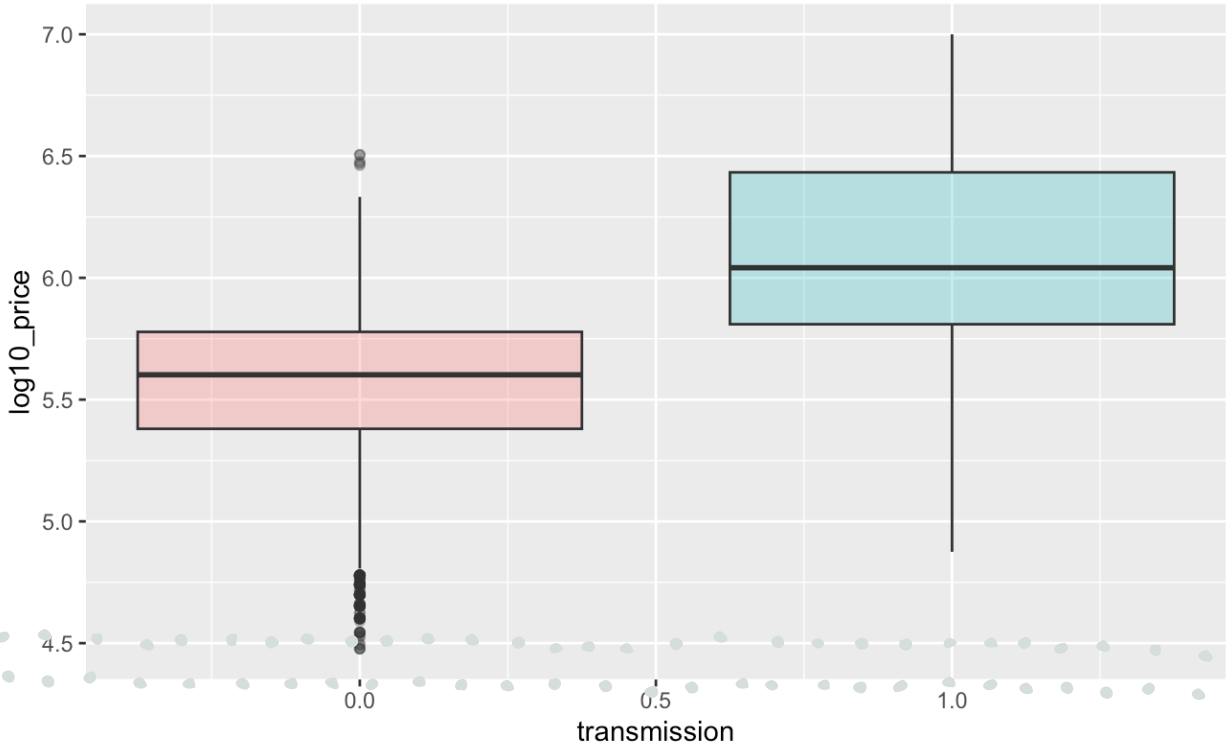
Box plot of fuel



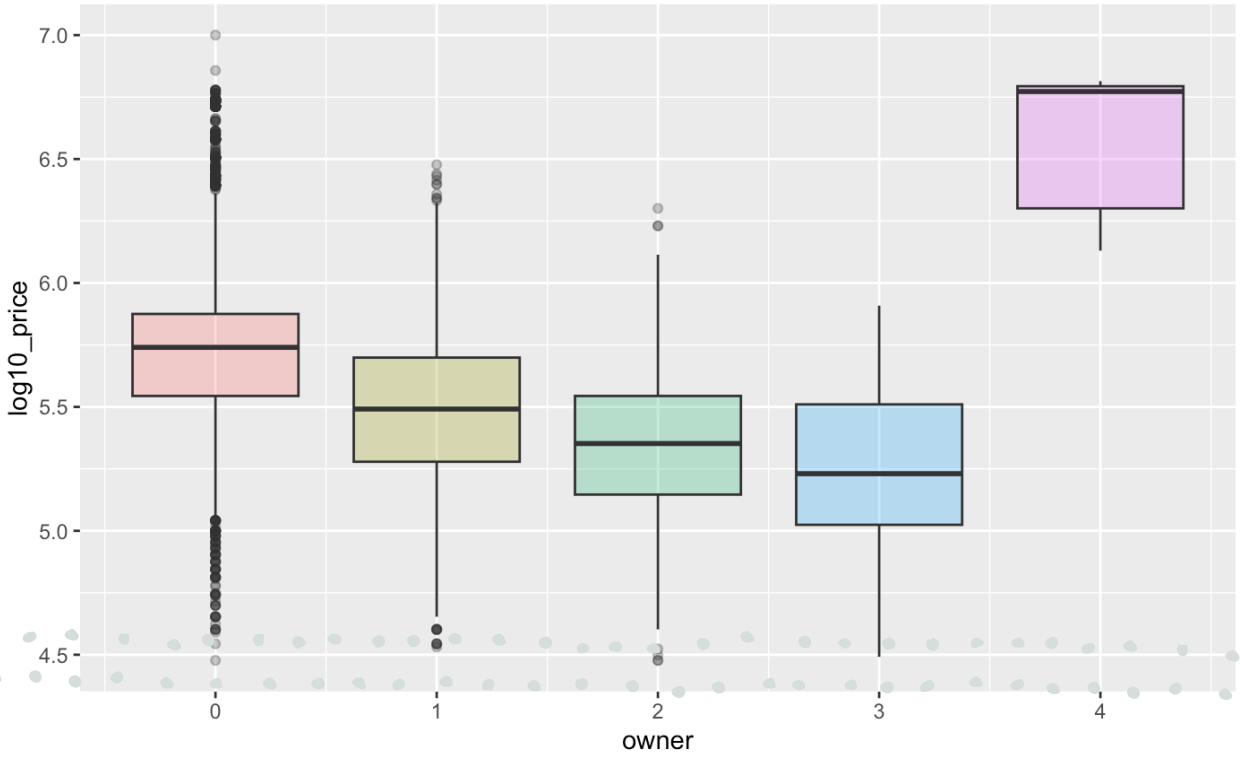
Box plot of seller type



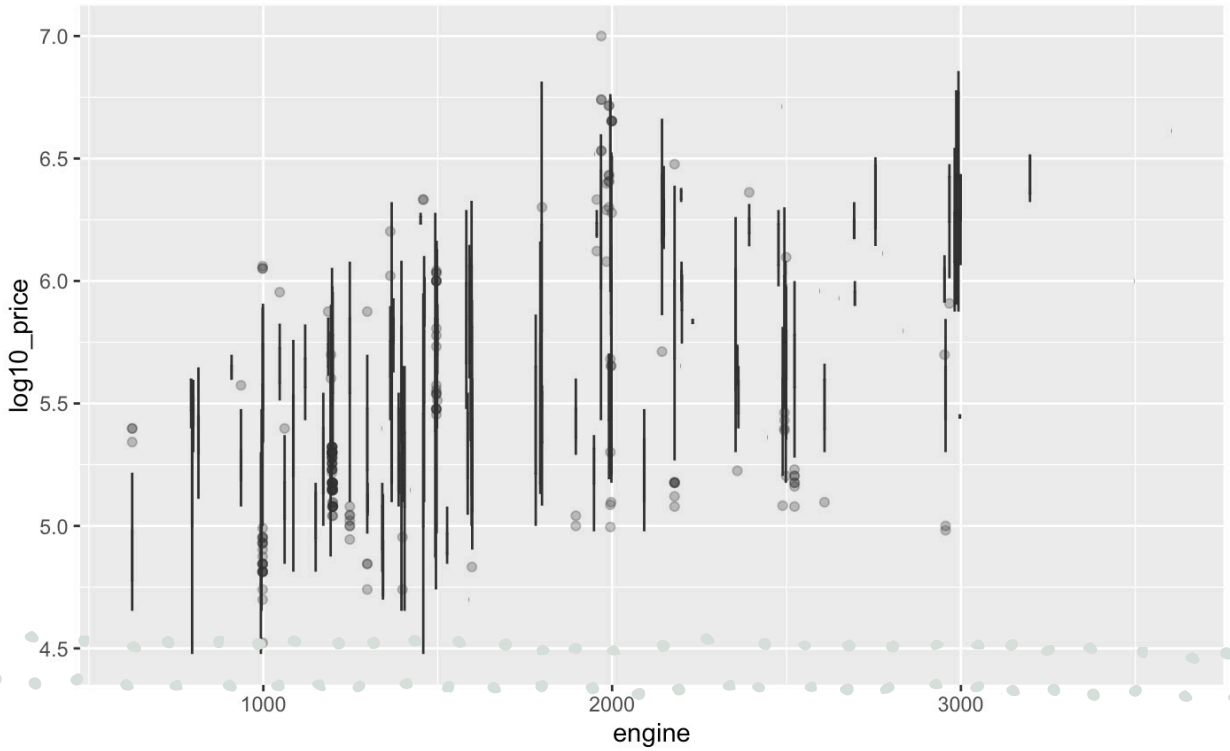
Box plot of Transmission



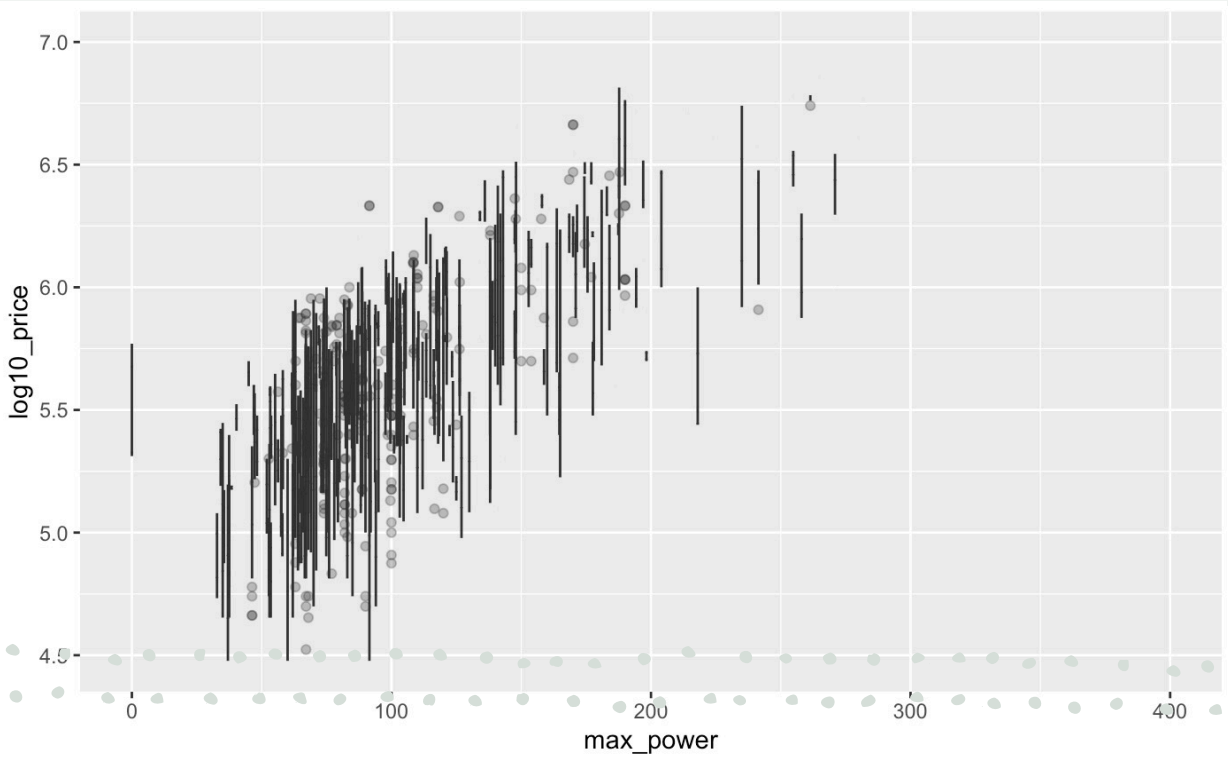
Box Plot of the Owner



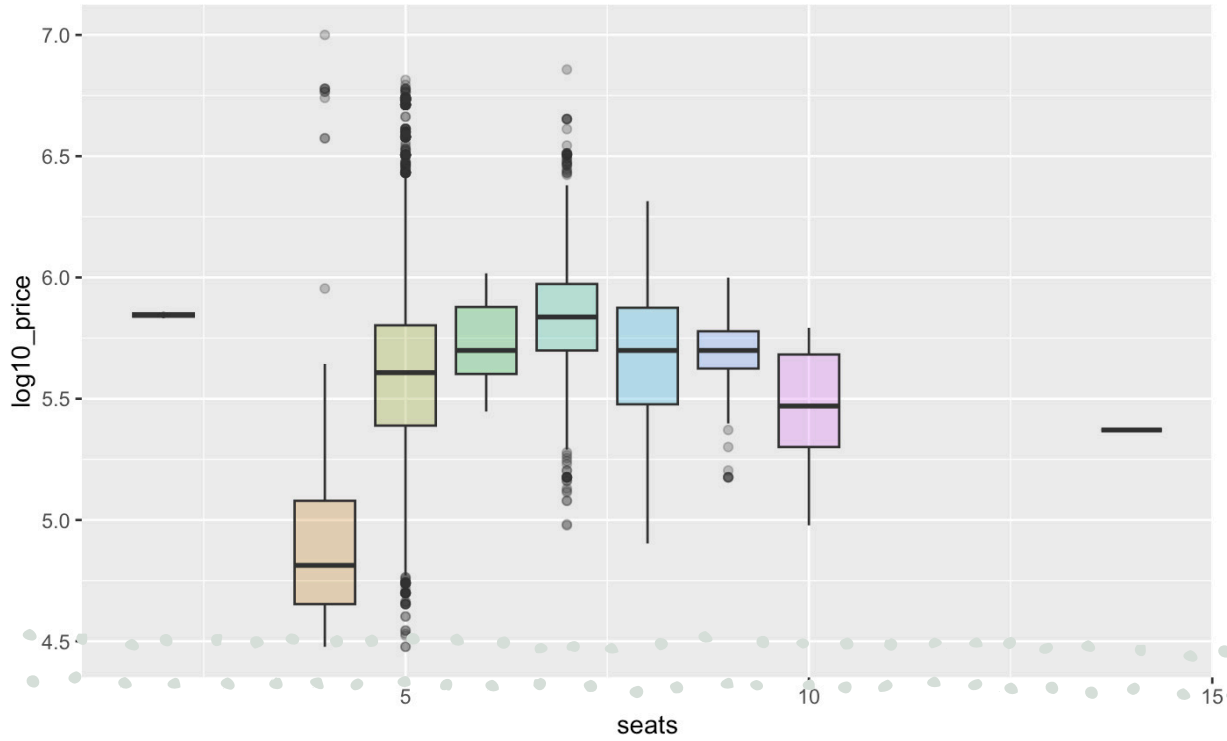
Box plot of engine

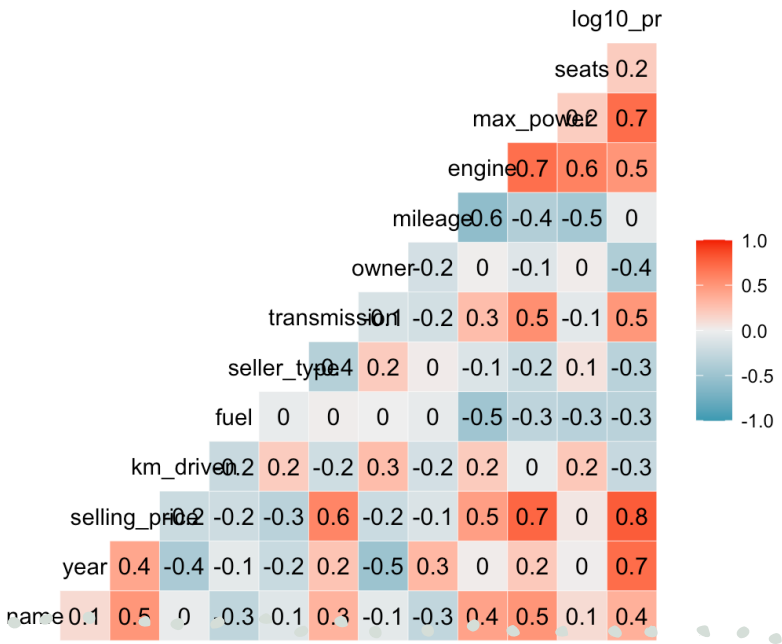


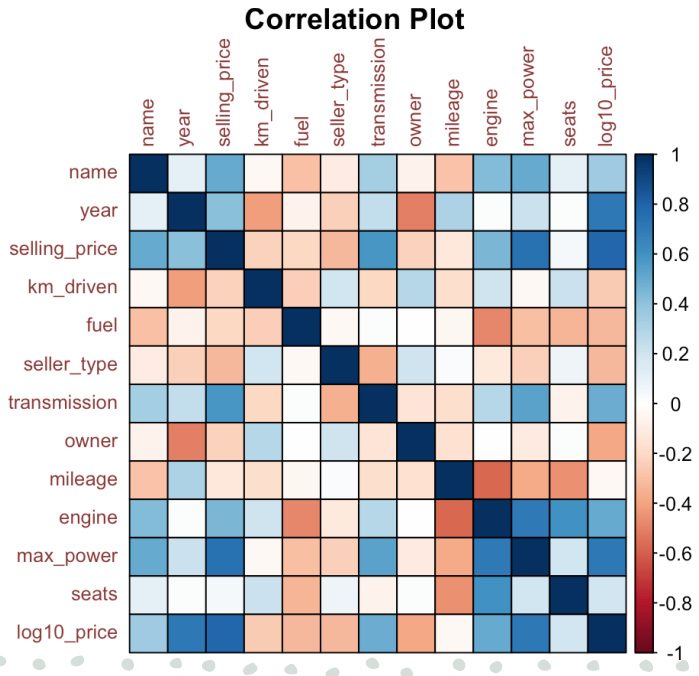
Box plot of max power



Box plot of seats





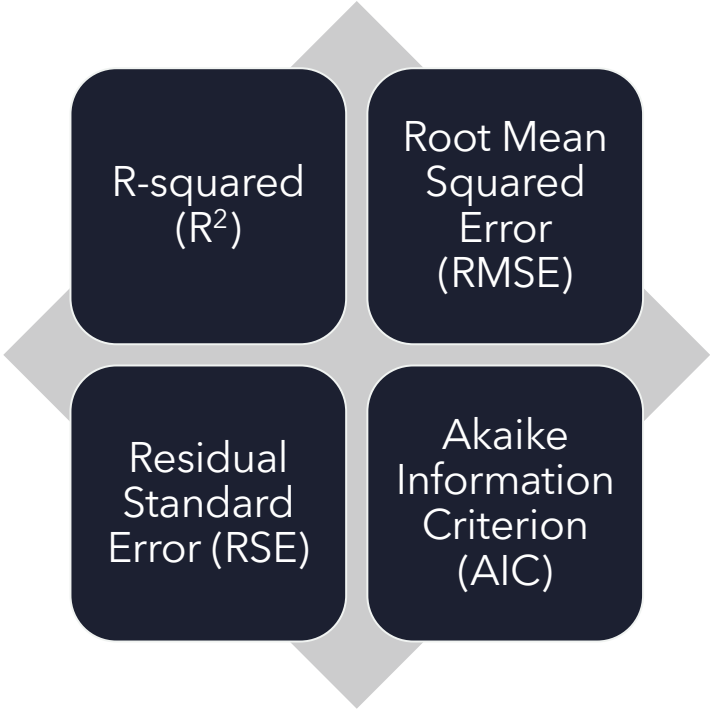


4. Correlation Plot of features

5. Modeling the data

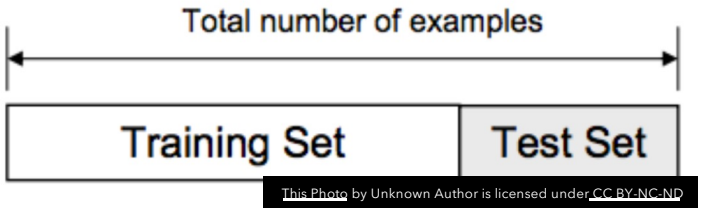
- 5.1. Accuracy metrics
- 5.2. Splitting of train data and test data
- 5.3. Linear Regression 1
- 5.4. Linear Regression 2
- 5.5. Adequacy checking of Linear Regression
- 5.6. Log Linear Regression
- 5.7.Adequacy checking of Log Linear Regression
- 5.8. Polynomial regression
- 5.9. Polynomial Regression 2
- 5.10. Models Evaluation
- 5.11. Lasso model
- 5.12. The Evaluation of Lasso Model

5.1. Accuracy metrics



5.2. Splitting of train data and test data

- Partition the data set into Train (70%) and Test (30%) set



5.3. Linear Regression 1

- we build the simplest model using all the available features
- we have an F-statistic of 1182 and a p-value almost equal to 0

```
Residuals:
     Min       1Q   Median       3Q      Max
-2464480  -211652    -5224   167360  3952046

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.854e+07  4.256e+06 -13.753  < 2e-16 ***
name         2.444e+04  1.426e+03  17.134  < 2e-16 ***
year         2.862e+04  2.127e+03  13.455  < 2e-16 ***
km_driven   -1.585e+00  1.608e-01  -9.855  < 2e-16 ***
seller_type -9.833e+04  1.407e+04  -6.990 3.07e-12 ***
mileage      2.360e+04  2.422e+03   9.744  < 2e-16 ***
transmission 4.348e+05  2.307e+04  18.844  < 2e-16 ***
max_power    1.243e+04  3.001e+02  41.427  < 2e-16 ***
engine       9.283e+01  2.682e+01   3.461 0.000542 ***
fuel         1.777e+04  1.532e+04   1.160 0.246009
owner       -3.004e+03  9.705e+03  -0.310 0.756893
seats       -1.293e+04  9.401e+03  -1.376 0.169011
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 449100 on 5679 degrees of freedom
Multiple R-squared:  0.696,     Adjusted R-squared:  0.6954
F-statistic:  1182 on 11 and 5679 DF,  p-value: < 2.2e-16
```

5.4. Linear Regression 2

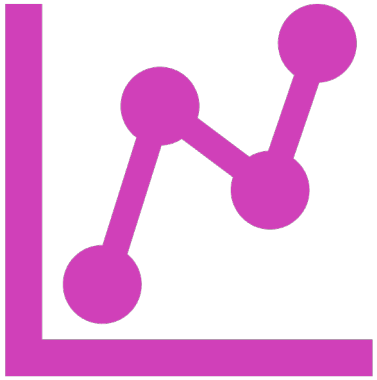
- The linear model without the features fuel, owner, and seats
- All of them are significant on selling price because of their p-values
- Transmission has the biggest coefficient of all features.
- A manual type car would have less price than an automatic one by $442,300

```
Residuals:
     Min       1Q   Median       3Q      Max
-2451313  -212110    -4713   164808  3963632

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.752e+07  3.834e+06 -15.002  < 2e-16 ***
name         2.430e+04  1.380e+03  17.606  < 2e-16 ***
year         2.811e+04  1.913e+03  14.696  < 2e-16 ***
km_driven   -1.648e+00  1.562e-01 -10.554  < 2e-16 ***
seller_type -1.015e+05  1.394e+04  -7.281 3.76e-13 ***
mileage      2.314e+04  2.057e+03  11.250  < 2e-16 ***
transmission 4.423e+05  2.274e+04  19.450  < 2e-16 ***
max_power    1.256e+04  2.848e+02  44.114  < 2e-16 ***
engine       6.006e+01  2.067e+01   2.905  0.00369 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 449200 on 5682 degrees of freedom
Multiple R-squared:  0.6958,    Adjusted R-squared:  0.6954
F-statistic:  1625 on 8 and 5682 DF,  p-value: < 2.2e-16
```

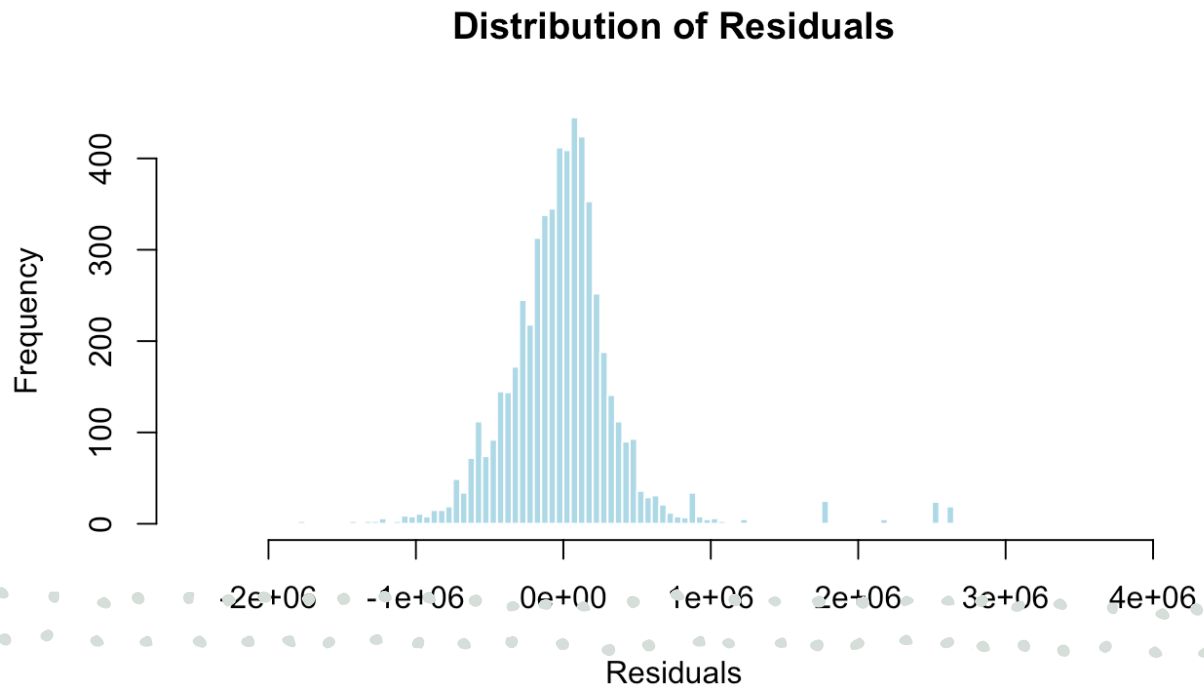
5.5. Adequacy checking of Linear Regression

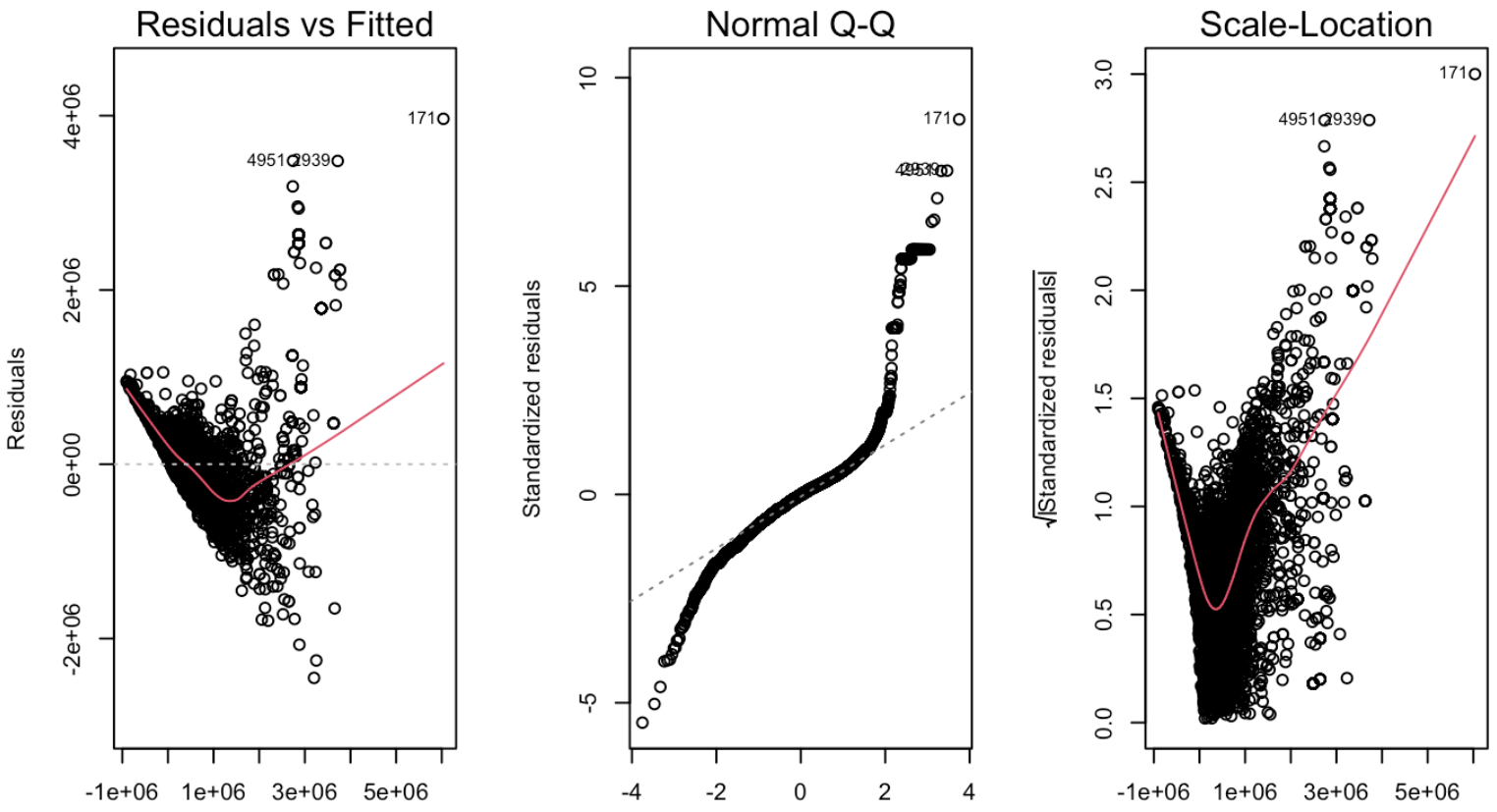
we check some assumptions of linear regression such as:

- Histogram of residuals
- Linearity of data using Residuals x Fitted plot
- Normality of residuals using Normal QQ plot
- Homogeneity of residuals variance, residuals with constant variance using scale location plot



Histogram of residuals



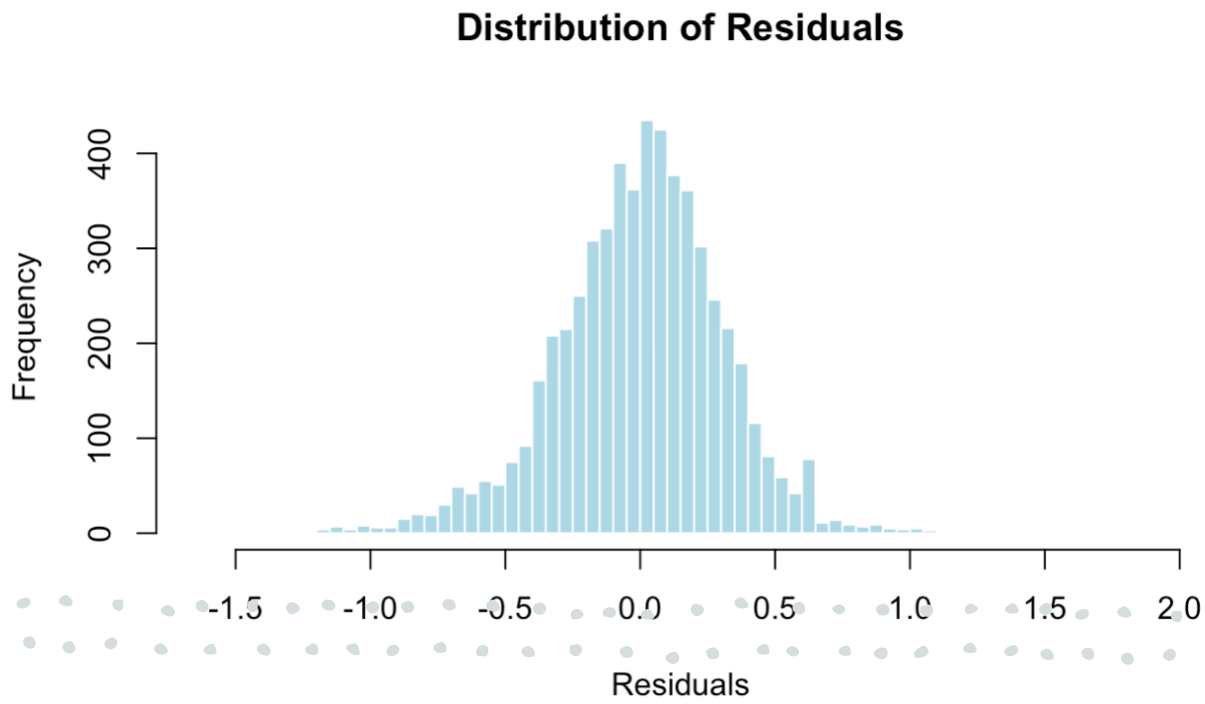

Residuals vs Fitted
Residuals
Fitted values
171
4951
2939
-2e+06
0e+00
2e+06
4e+06
-1e+06
1e+06
3e+06
5e+06
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-5
0
5
10
-4
-2
0
2
4
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0

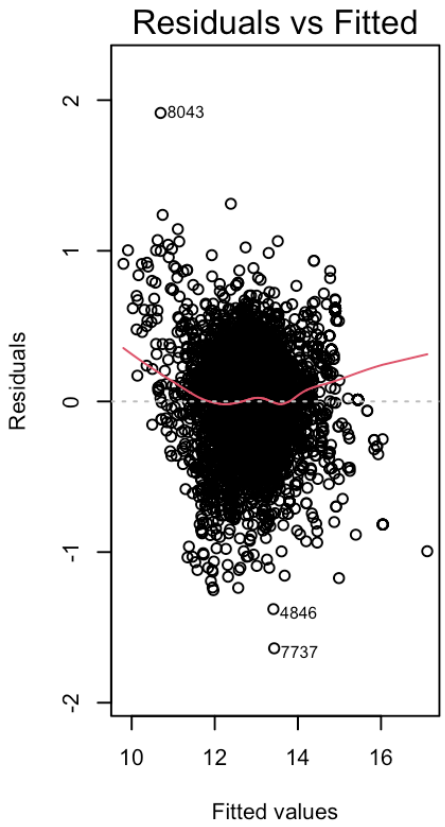
5.6. Log Linear Regression

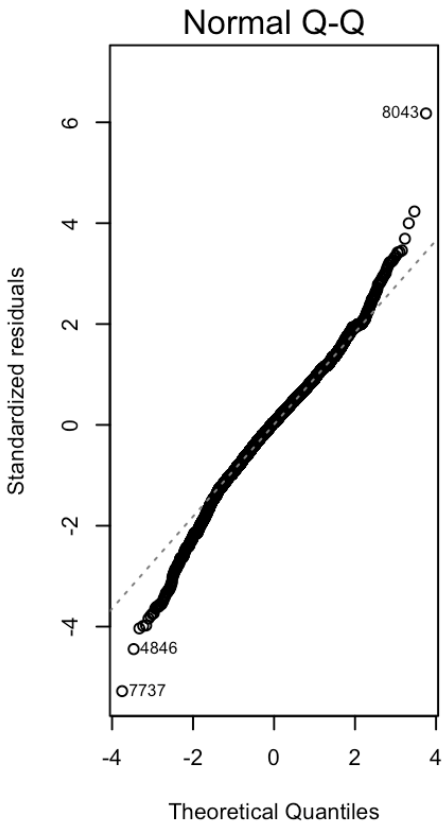
- We have an adjusted R-squared near 1 which indicates that the independent variables in the model are good predictors of the dependent variable
- we have a good model fit or overfitting

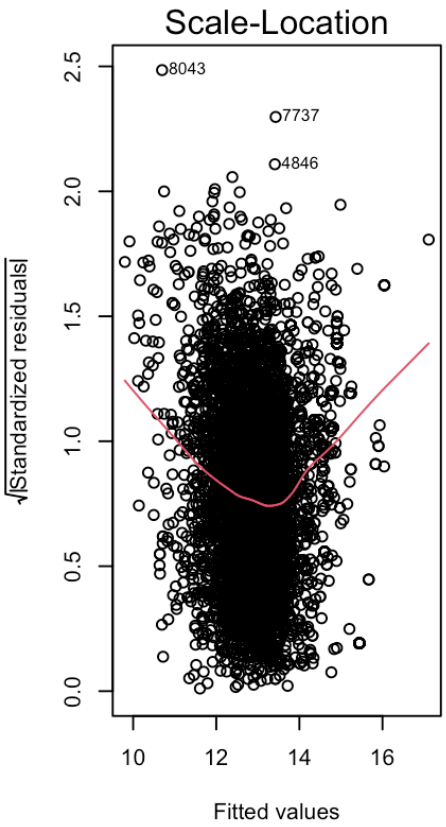
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.63890 -0.18819  0.01647  0.19385  1.91515

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.214e+02  2.652e+00 -83.484  < 2e-16 ***
name         -5.487e-03  9.549e-04  -5.747 9.58e-09 ***
year          1.155e-01  1.323e-03  87.324  < 2e-16 ***
km_driven    -4.763e-07  1.080e-07  -4.409 1.06e-05 ***
seller_type  -6.980e-02  9.641e-03  -7.240 5.09e-13 ***
mileage       1.964e-02  1.422e-03  13.810  < 2e-16 ***
transmission  1.859e-01  1.573e-02  11.817  < 2e-16 ***
max_power     9.861e-03  1.970e-04  50.066  < 2e-16 ***
engine        4.196e-04  1.430e-05  29.344  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3107 on 5682 degrees of freedom
Multiple R-squared:  0.862,     Adjusted R-squared:  0.8618
F-statistic:  4435 on 8 and 5682 DF,  p-value: < 2.2e-16
```

5.7.Adequacy checking of Log Linear Regression
Histogram of residuals




Residuals vs Fitted
Residuals
Fitted values
8043
4846
7737


Normal Q-Q
Standardized residuals
Theoretical Quantiles
8043
4846
7737


Scale-Location
√|Standardized residuals|
Fitted values
8043
7737
4846

5.8. Polynomial regression

- We tried to generate a new feature matrix consisting of all polynomial combinations of the features with degree 2
- We don’t want selling price to be included in the process of generating the polynomial combinations
- we take out selling price from train and test
- we know that feature seats has no correlation with selling price. Therefore, we can drop it
- Our new datasets train_poly and test_poly now have 133 columns.
- We iteratively remove one feature at a time and re-fit the model, calculating the AIC at each step. The best model is the one with lowest AIC

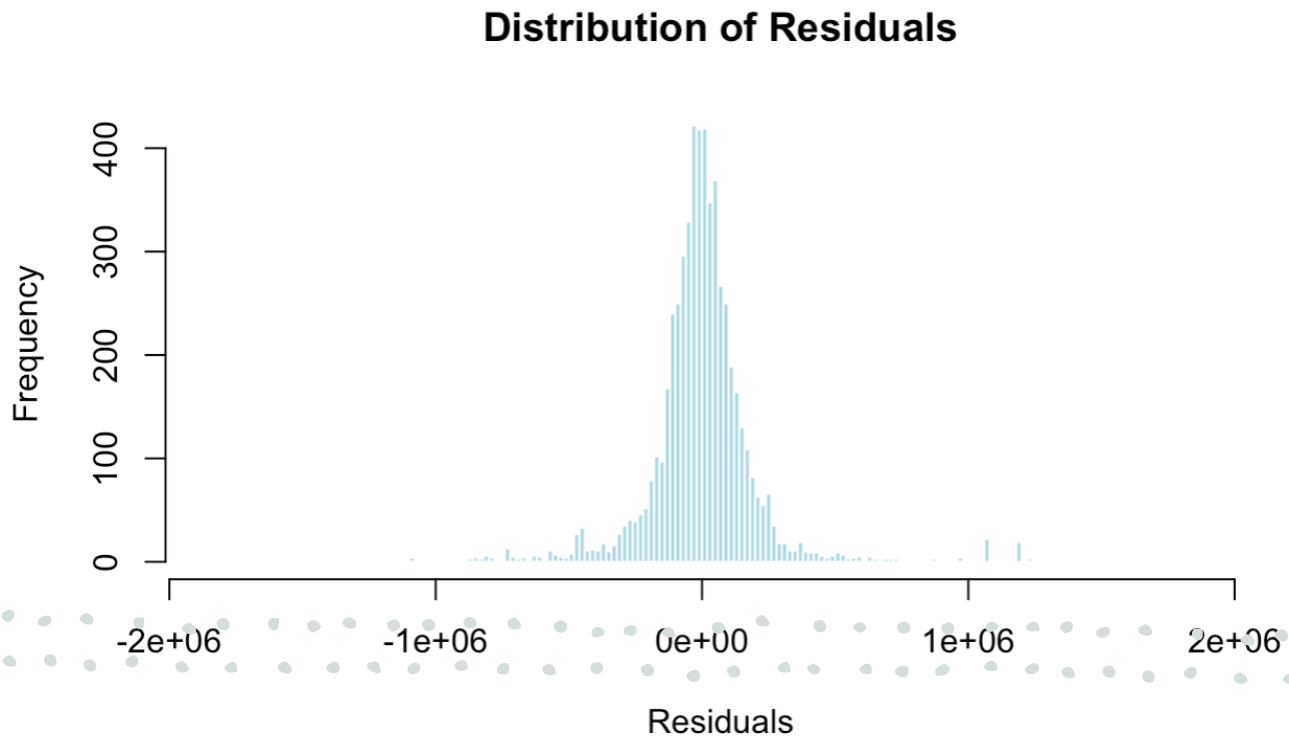
5.8. Polynomial regression

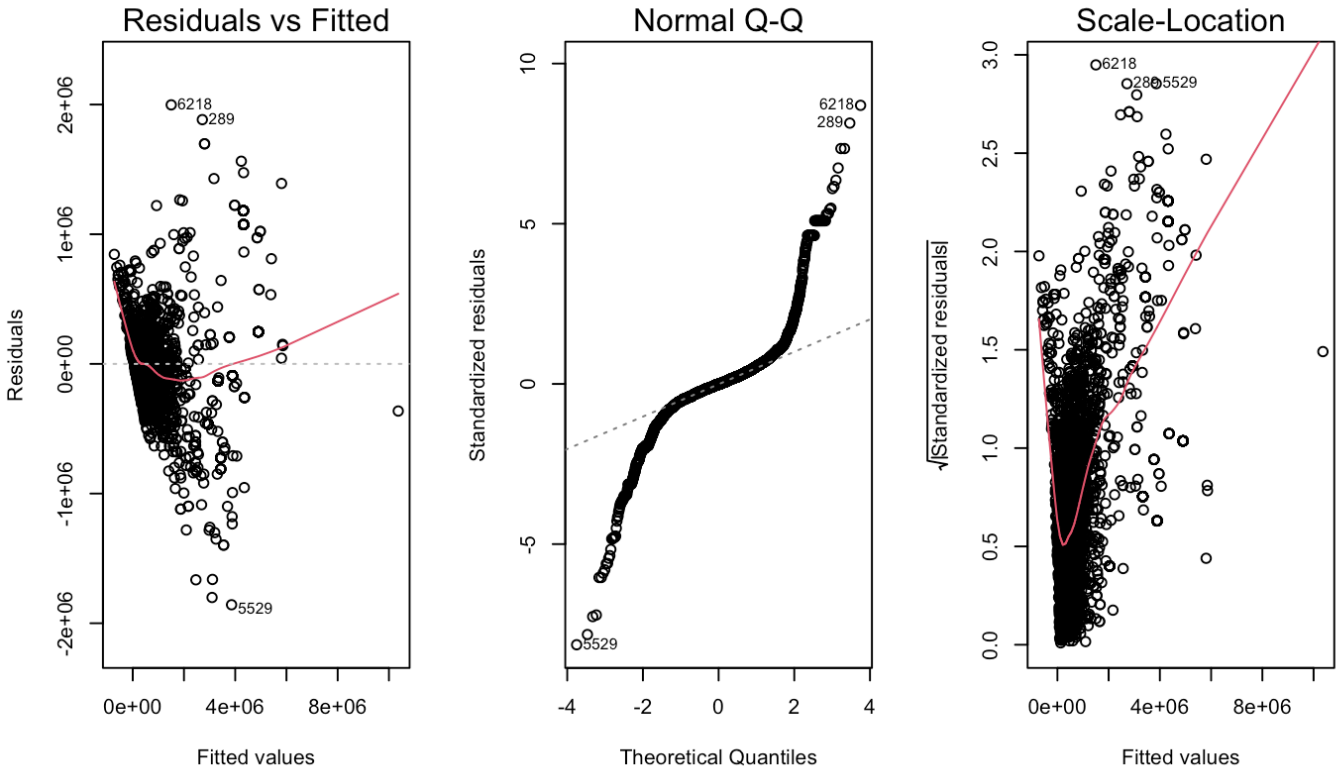
- We save the best model as l_p, and after that we predict and calculate the metrics

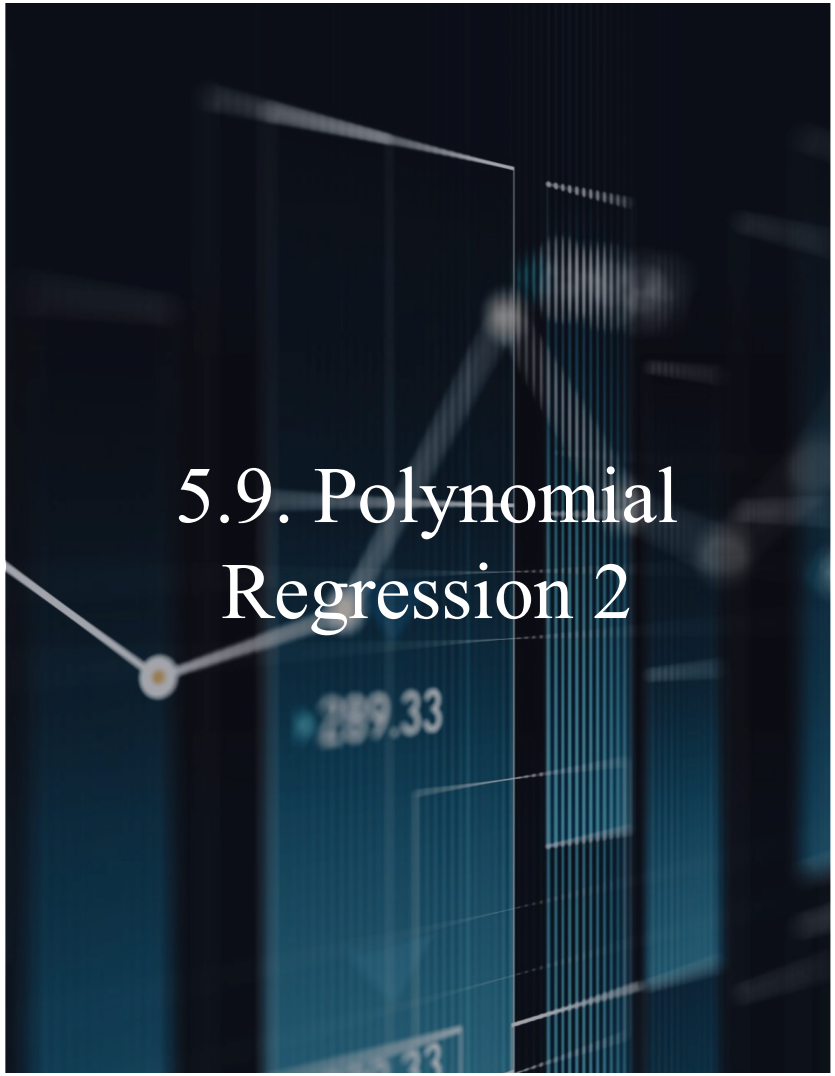
```
Residuals:
     Min       1Q   Median       3Q      Max
-1857447   -81240    -2194    77208  1997530

Residual standard error: 232800 on 5645 degrees of freedom
Multiple R-squared:  0.9188,    Adjusted R-squared:  0.9182
F-statistic:  1420 on 45 and 5645 DF,  p-value: < 2.2e-16
```

- There are 3 features (mileage, km_driven * km_driven and year * mileage) that are not significant and 42 are.
- Intercept has the biggest coefficient of all features which may be a sign of overfitting in this model

Adequacy checking of Polynomial Regression
Histogram of residuals




Residuals vs Fitted
Residuals
Fitted values
6218
289
5529
-2e+06
-1e+06
0e+00
1e+06
2e+06
0e+00
4e+06
8e+06
Normal Q-Q
Standardized residuals
Theoretical Quantiles
6218
289
5529
-5
0
5
10
-4
-2
0
2
4
Scale-Location
√|Standardized residuals|
Fitted values
6218
289
5529
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0e+00
4e+06
8e+06



5.9. Polynomial Regression 2

- This time we use the logarithm of selling price as target variable because:

1. The relationship between our target variable and independent variables seems to be non-linear
2. This can be especially useful if the target variable exhibits heteroscedasticity, meaning that its variance is not constant across its range.
3. Help to mitigate the impact of outliers

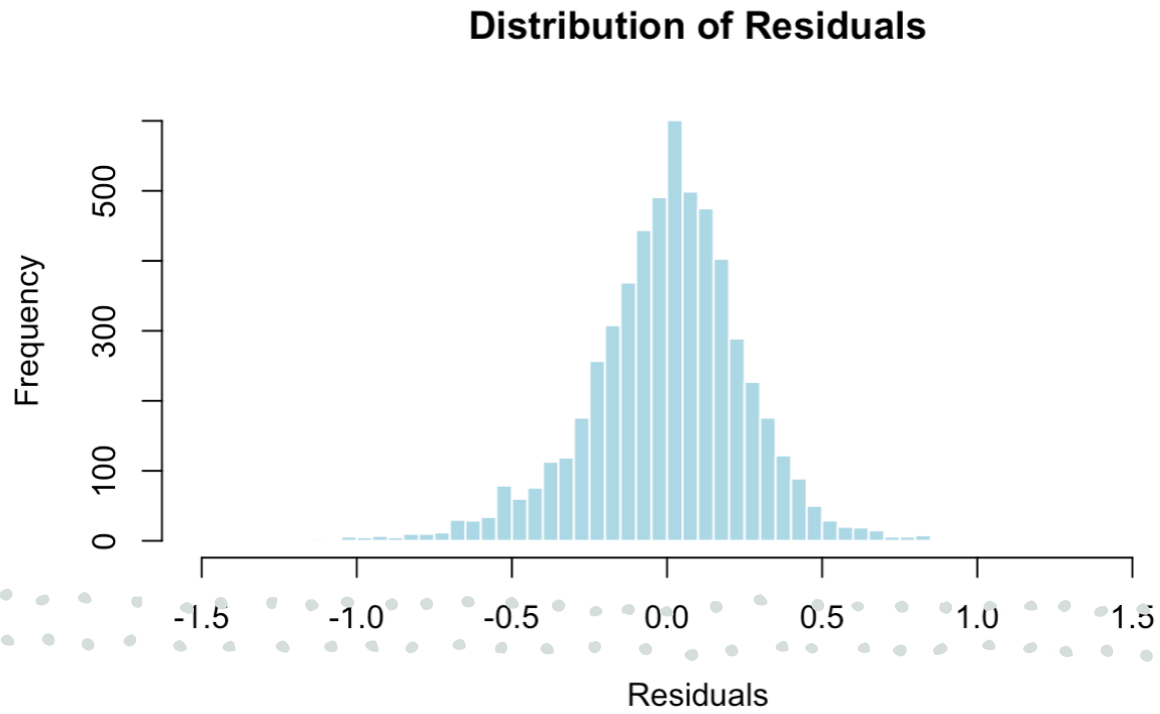
5.9. Polynomial Regression 2

- Our datasets train_poly and test_poly have 133 columns.
- We calculate the AIC at each step.
- Save the best model as l_p_log, then predict. After that, calculate the metrics.

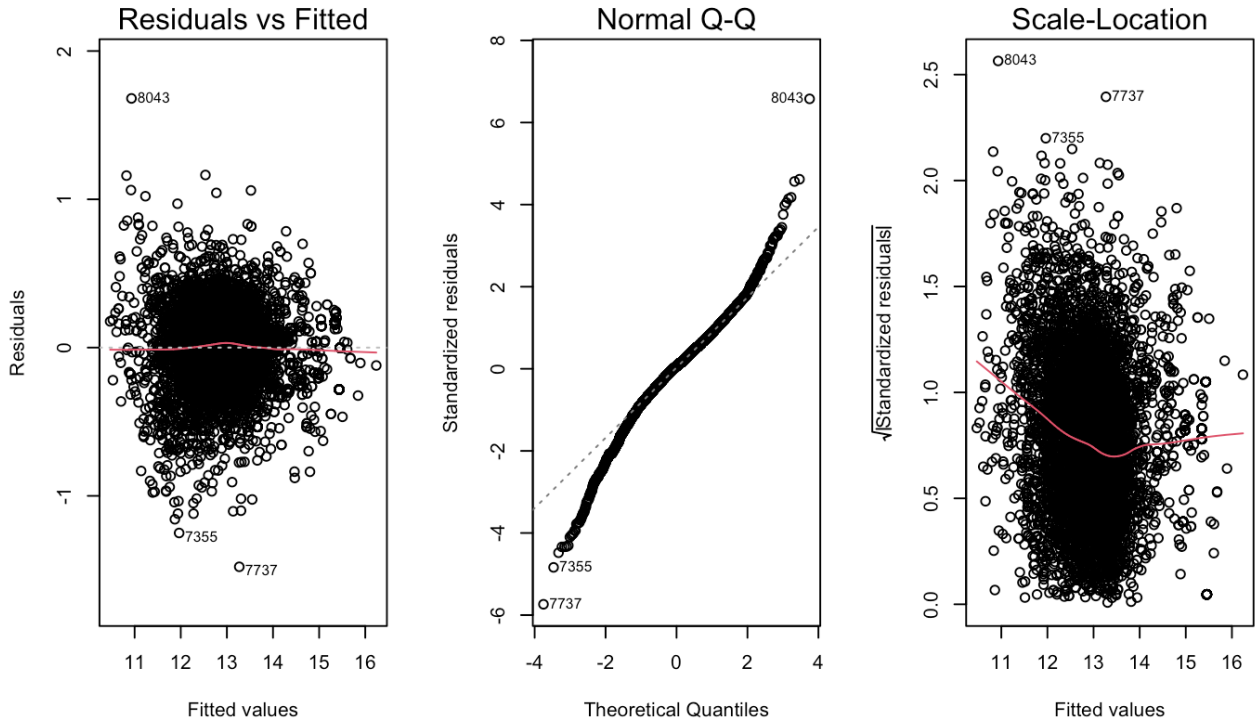
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.47833 -0.14068  0.01663  0.15690  1.68031
Residual standard error: 0.2589 on 5644 degrees of freedom
Multiple R-squared:  0.9048,    Adjusted R-squared:  0.904
F-statistic:  1166 on 46 and 5644 DF,  p-value: < 2.2e-16
```

- There are 4 features (km_driven * max_power, seller_type * engine, seller_type * max_power and mileage * max_power) that are not significant and 46 that are
- Intercept has the biggest coefficient of all features
- considering all other features to be zero, selling price will be equal to 4039

Histogram of residuals



Adequacy checking of Polynomial Regression 2



5.10. Models Evaluation

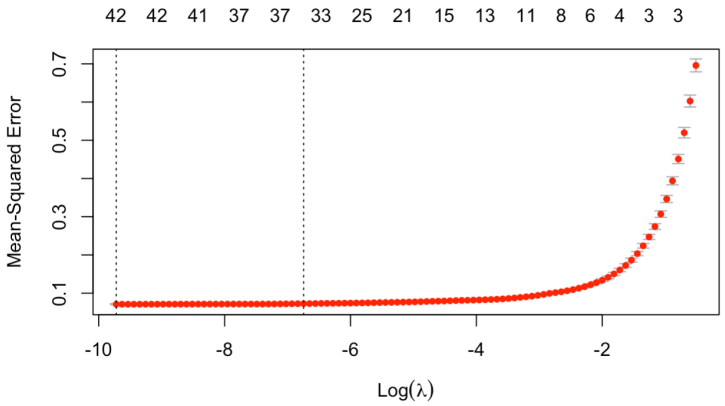
	Adjusted R sq	RSE	RMSE	AIC
Linear Regression 1	0.6954480	4.491287e+05	4.590137e+05	164301.8136
Linear Regression 2	0.6954036	4.491614e+05	4.592924e+05	164299.6481
Log Linear Regression	0.8617610	3.106861e-01	3.249474e-01	2856.1085
Polynomial Regression	0.9181999	2.327647e+05	2.877652e+05	156854.4604
Polynomial Regression 2	0.9039901	2.589193e-01	3.997758e-01	819.3611

5.11. Lasso Model

- Lasso regularization: Shrinks coefficients to zero, performs feature selection, and reduces model complexity.
- Goal: Assess the restrictiveness of the Lasso model in feature selection and coefficient control.
- Procedure: Extract coefficient estimates based on the optimal lambda using the provided code.
- Objective: Determine the number of variables considered zero by the Lasso model.

5.12. The Evaluation of Lasso Model

- Its response variable and its features are the same as our second Polynomial Regression model
- We used the cross validation to determine the optimal lambda.
- RMSE for Lasso model: Slightly higher than the second polynomial model, indicating similar predictive accuracy
- Our lasso model has considered the coefficient of only 4 variables out of our total 47 variable equal to zero. These variables are owner, name:year, year:seller_type and year:owner



6. Conclusion

1. Correlation with Price: Number of seats has low correlation with price, while Max Power has the highest correlation.

2. Inter-feature Correlation: Features do not have strong correlation with each other.

3. Logical Trends: Feature impacts on price follow logical trends.

4. Model Comparison: Our log linear model without the features fuel, owner, and seats was a good model in case of Adjusted R2 and AIC and our polynomial regression model which considered the log of price had even a bit better results in case of these metrics too.

5. Lasso Regularization: Lasso model results in few zero coefficients, indicating stability.