MERCEDES-BENZ PRICE PREDICTION

Personal Project

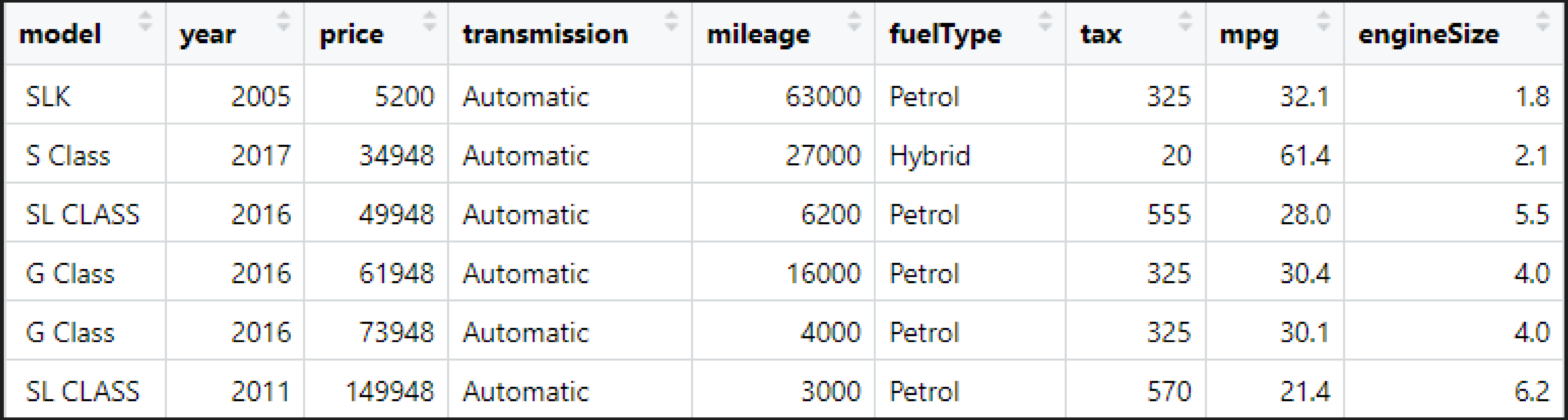
BUSINESS OBJECTIVES

The Mercedes-Benz company aims to empower junior salespeople with an advanced pricing tool to enhance their decision-making capabilities.

The primary objective is to develop a **predictive model** that accurately estimates the prices of used Mercedes-Benz cars.

DATASET

Data from kaggle the price range of listed Mercedes Used Car. The model year ranges between 1970-2020.



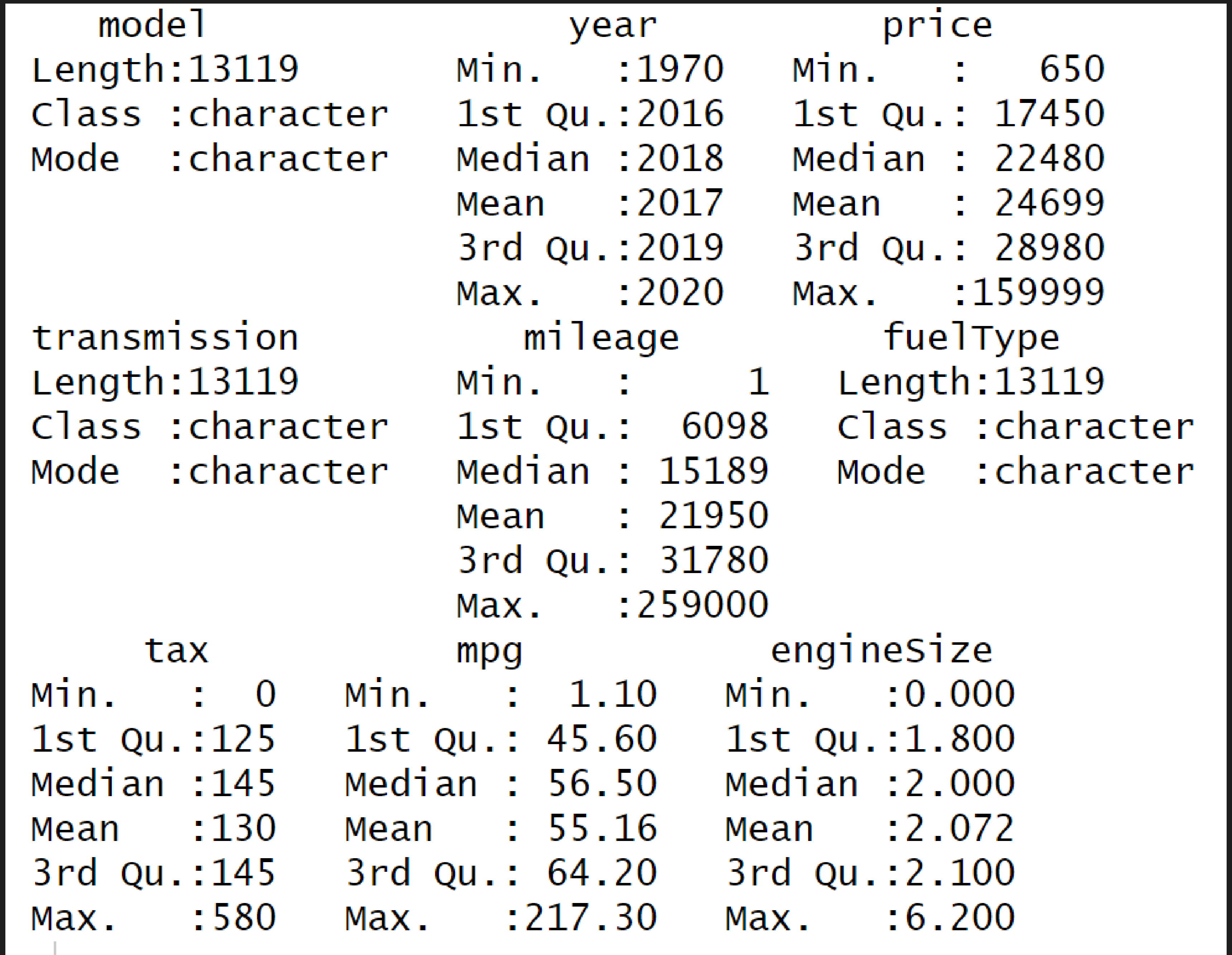
model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
SLK	2005	5200	Automatic	63000	Petrol	325	32.1	1.8
S Class	2017	34948	Automatic	27000	Hybrid	20	61.4	2.1
SL CLASS	2016	49948	Automatic	6200	Petrol	555	28.0	5.5
G Class	2016	61948	Automatic	16000	Petrol	325	30.4	4.0
G Class	2016	73948	Automatic	4000	Petrol	325	30.1	4.0
SL CLASS	2011	149948	Automatic	3000	Petrol	570	21.4	6.2

PROCESS

1. Explore data
2. Data preparation
3. Model training
4. Model evaluation
5. Conclusion and recommendations

DATA EXPLORATION & PREPARATION

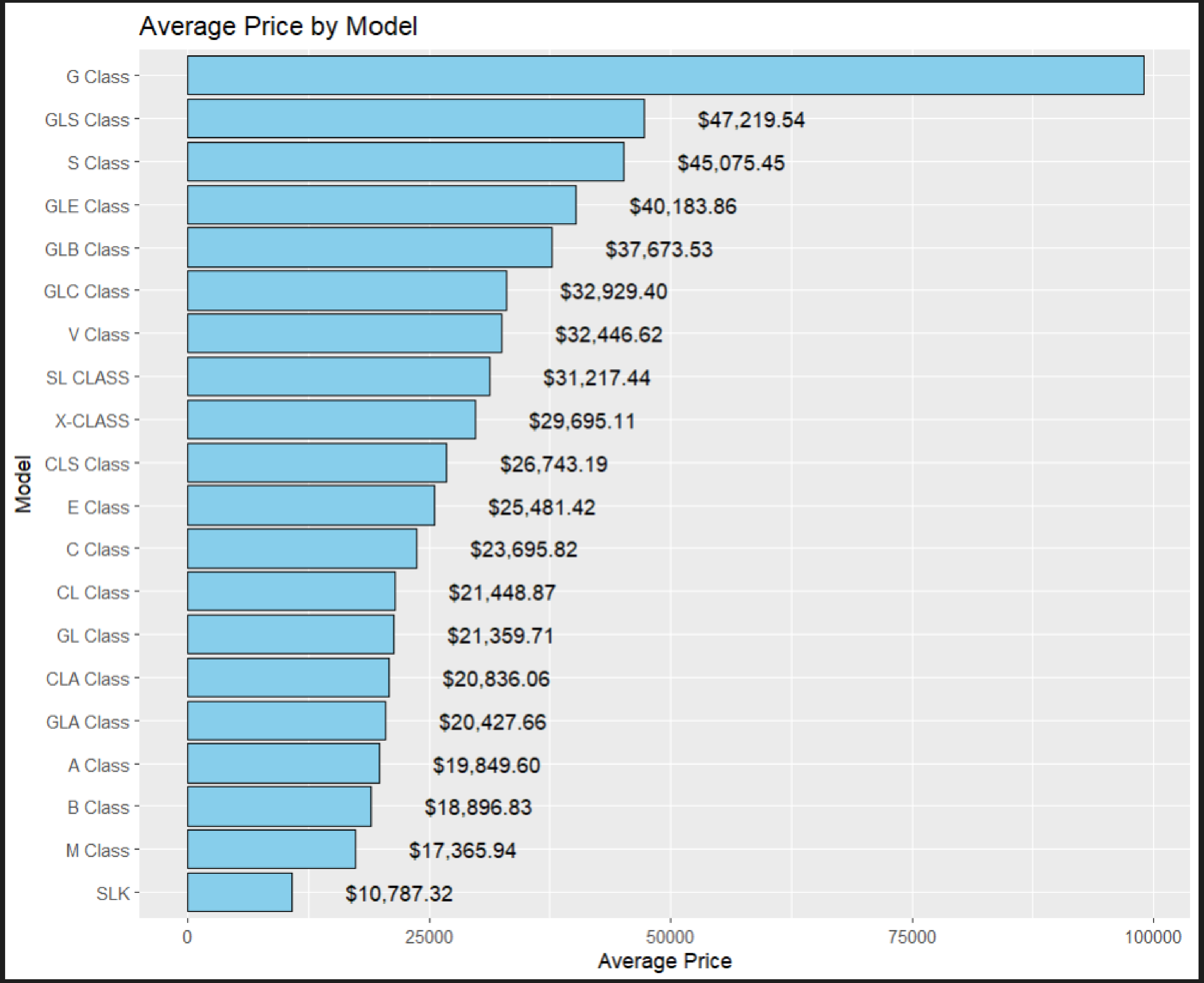
DATA EXPLORATION



```
   model                 year          price       
 Length:13119       Min.   :1970   Min.   :   650  
 Class :character   1st Qu.:2016   1st Qu.: 17450  
 Mode  :character   Median :2018   Median : 22480  
                    Mean   :2017   Mean   : 24699  
                    3rd Qu.:2019   3rd Qu.: 28980  
                    Max.   :2020   Max.   :159999  
 transmission          mileage         fuelType        
 Length:13119       Min.   :     1   Length:13119      
 Class :character   1st Qu.:  6098   Class :character  
 Mode  :character   Median : 15189   Mode  :character  
                    Mean   : 21950                     
                    3rd Qu.: 31780                     
                    Max.   :259000                     
      tax            mpg           engineSize   
 Min.   :  0   Min.   :  1.10   Min.   :0.000  
 1st Qu.:125   1st Qu.: 45.60   1st Qu.:1.800  
 Median :145   Median : 56.50   Median :2.000  
 Mean   :130   Mean   : 55.16   Mean   :2.072  
 3rd Qu.:145   3rd Qu.: 64.20   3rd Qu.:2.100  
 Max.   :580   Max.   :217.30   Max.   :6.200  
```

- The dataset is clean
  - 13119 rows
  - 9 columns
- 27 Mercedes-Benz models

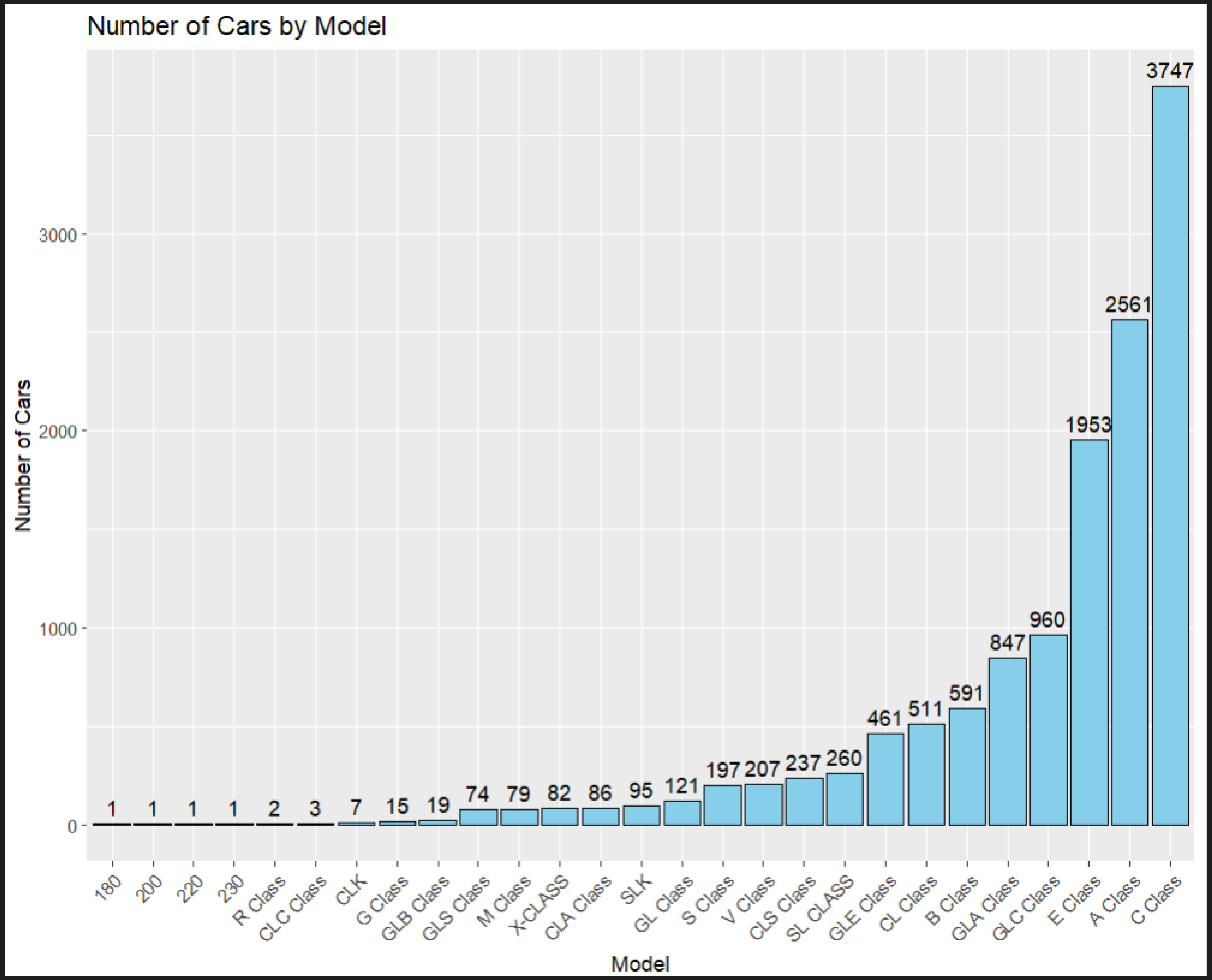
AVERAGE PRICE BY MODEL



- Model with highest average price is G Class, $98934
- Followed by GLS Class $47220 and S Class $245075

- Model with lowest average price is CLK, $3078
- Followed by 230 $4500 and CLC Class $5517

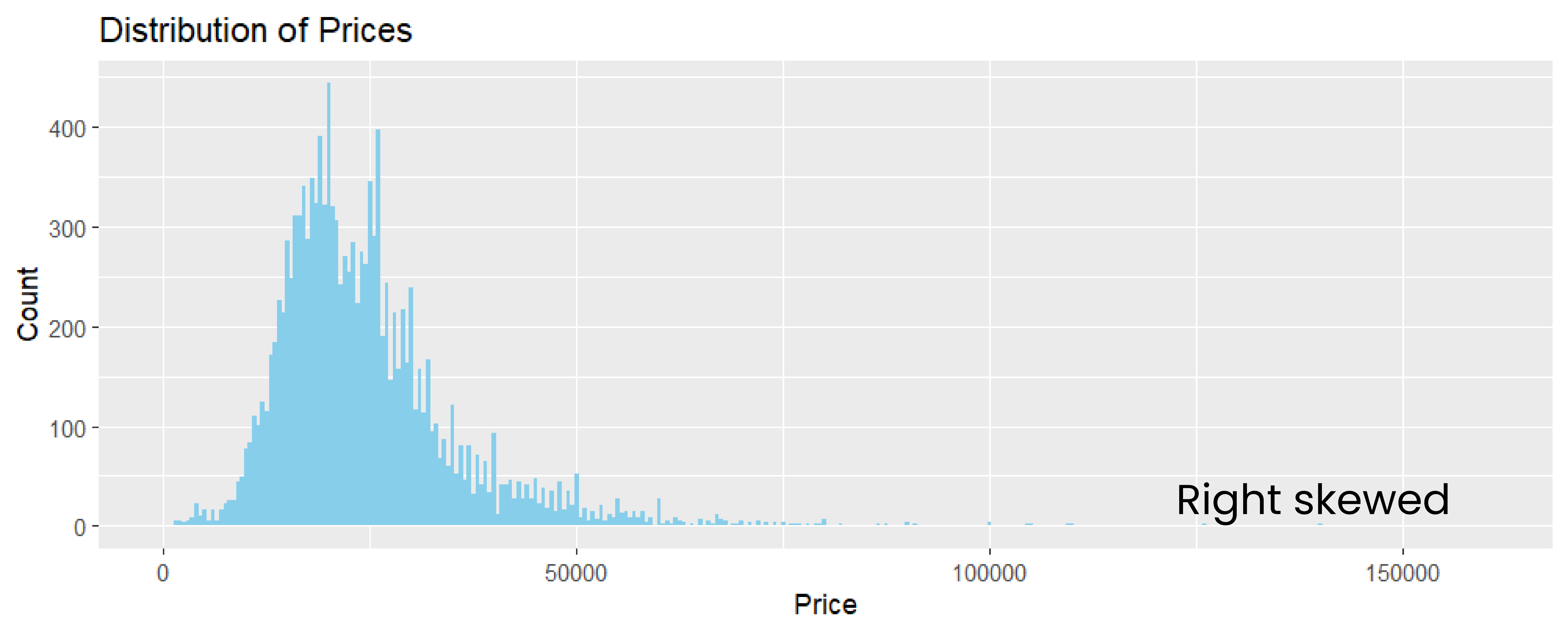
NUMBER OF CARS BY MODEL



- C Class, A Class and E Class are popular models with 63% of the samples in this dataset

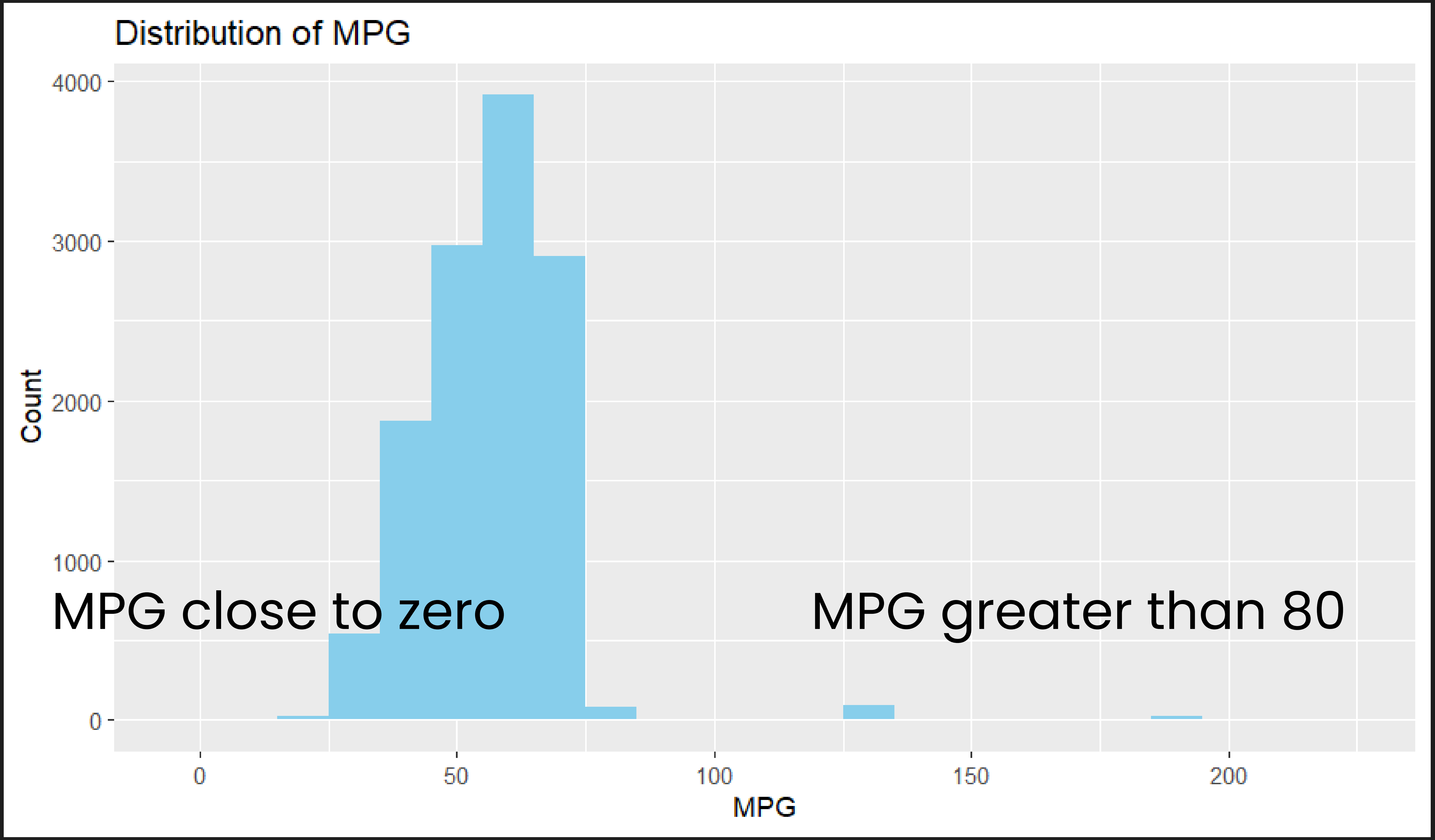
Some models have sample size n<50, which is quite small
I decide to filter these models out for the model training

DISTRIBUTION OF PRICES



Right skewed

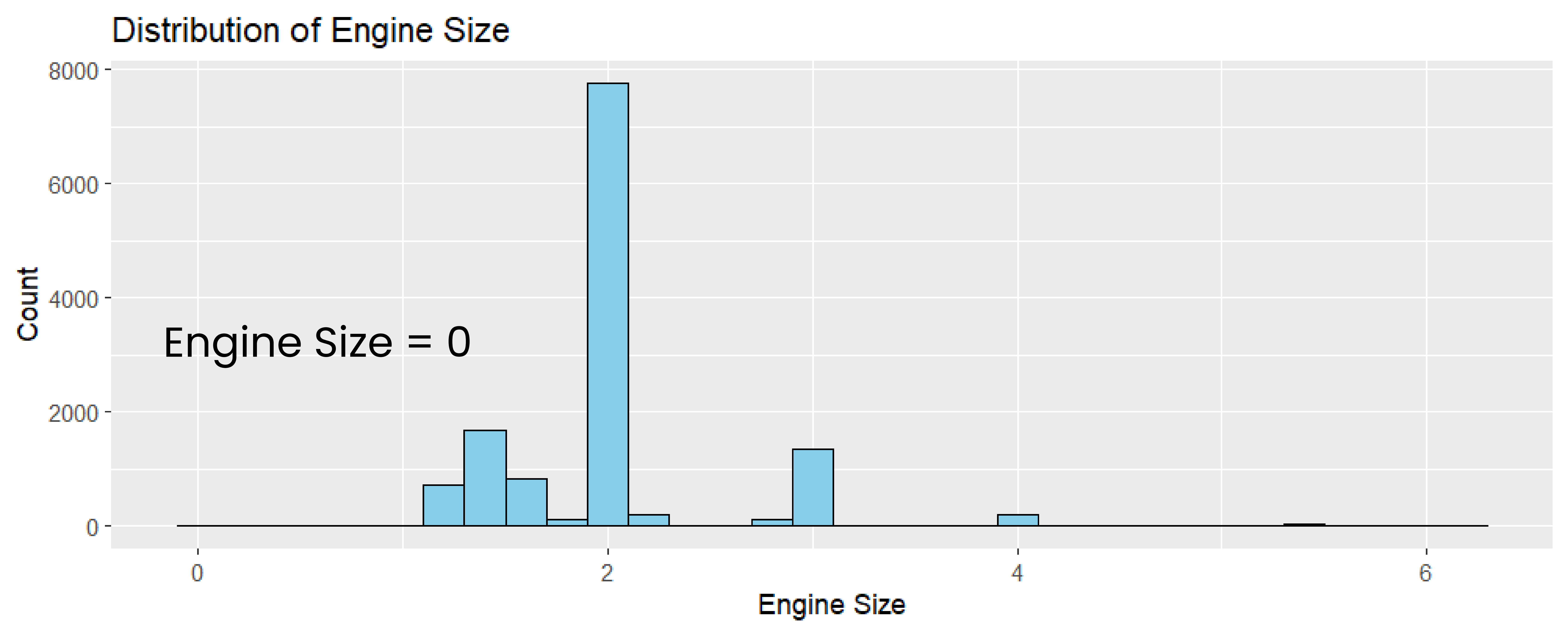
DISTRIBUTION OF MPG



MPG close to zero

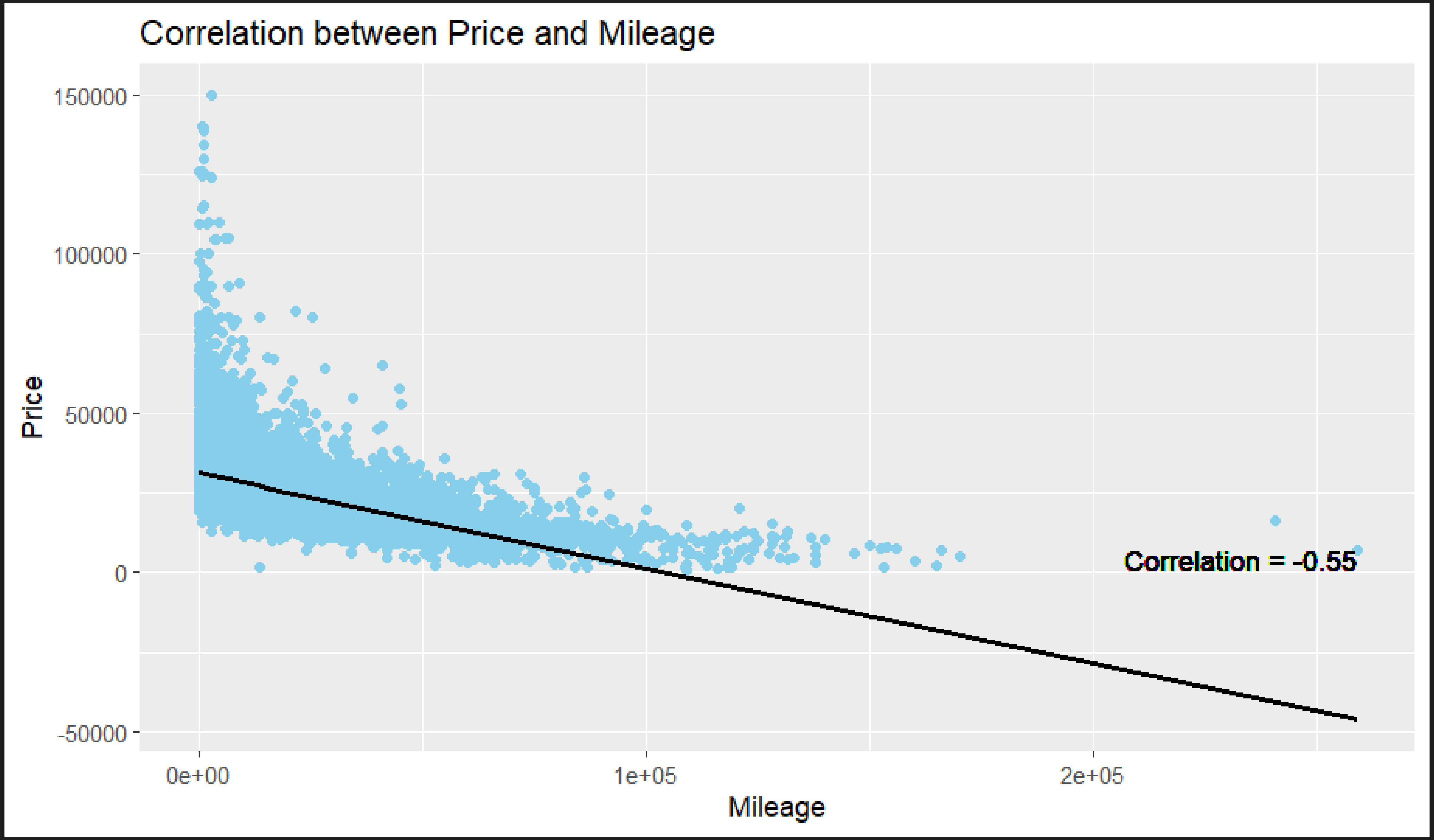
MPG greater than 80

DISTRIBUTION OF ENGINE SIZE



Engine Size = 0

CORRELATION PRICE AND MILEAGE

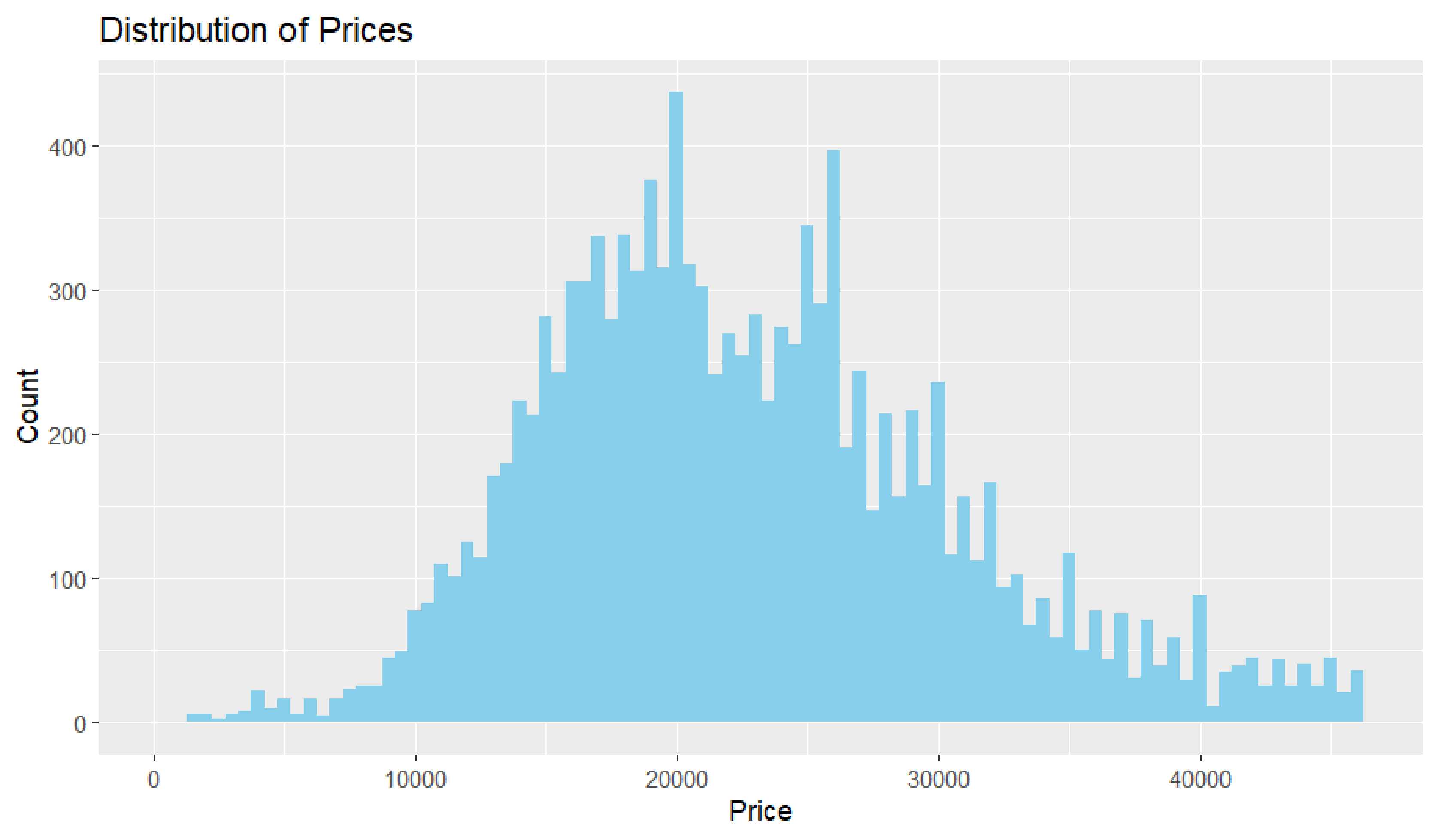


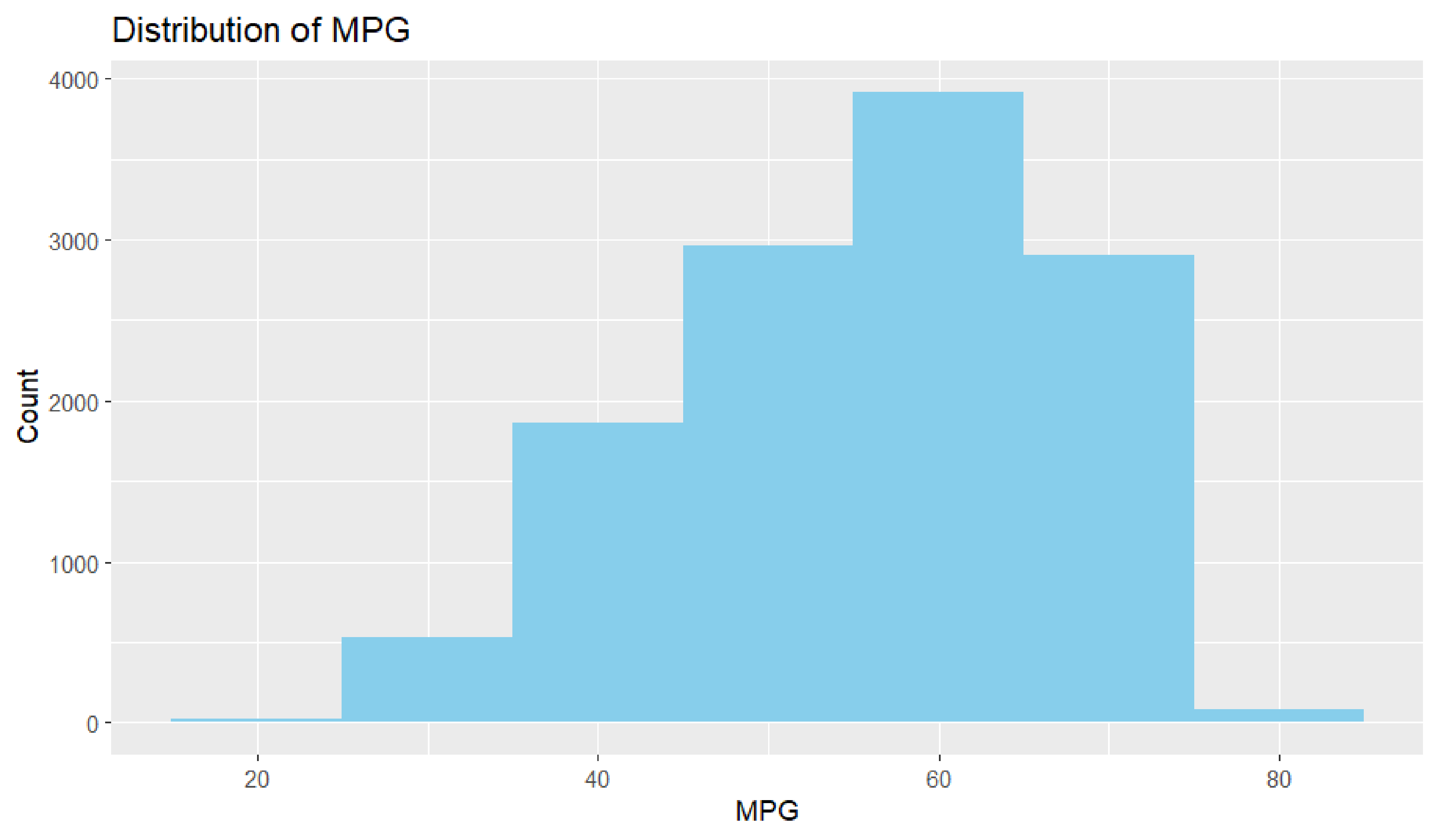
- Negative Correlation

DATA PREPARATION

1. Convert variables to the right types
2. Filter out models with a sample size less than 50
3. Filter out cars with engine size = 0
4. Identify and handle outliers

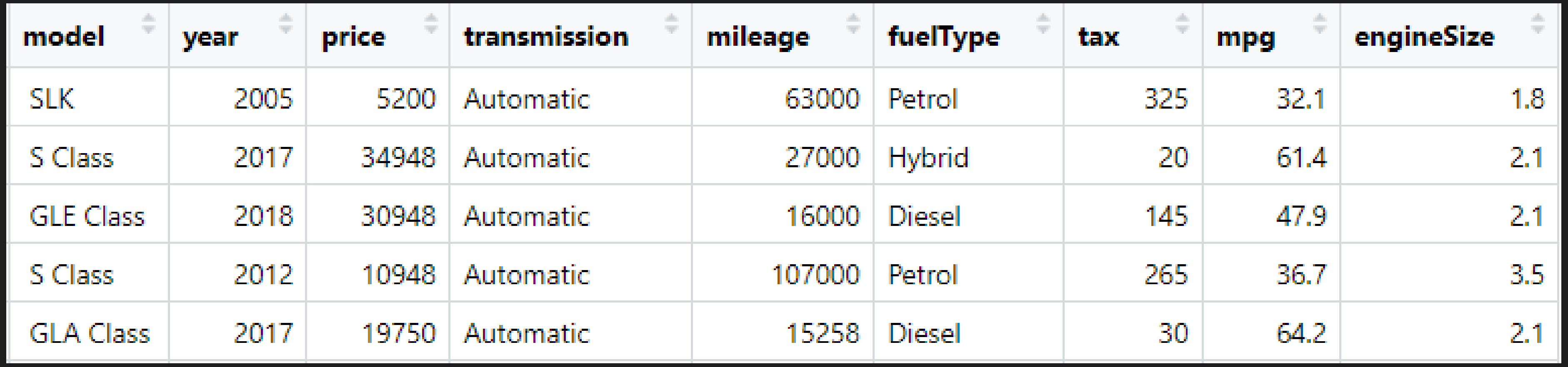
DISTRIBUTION OF PRICES & MPG



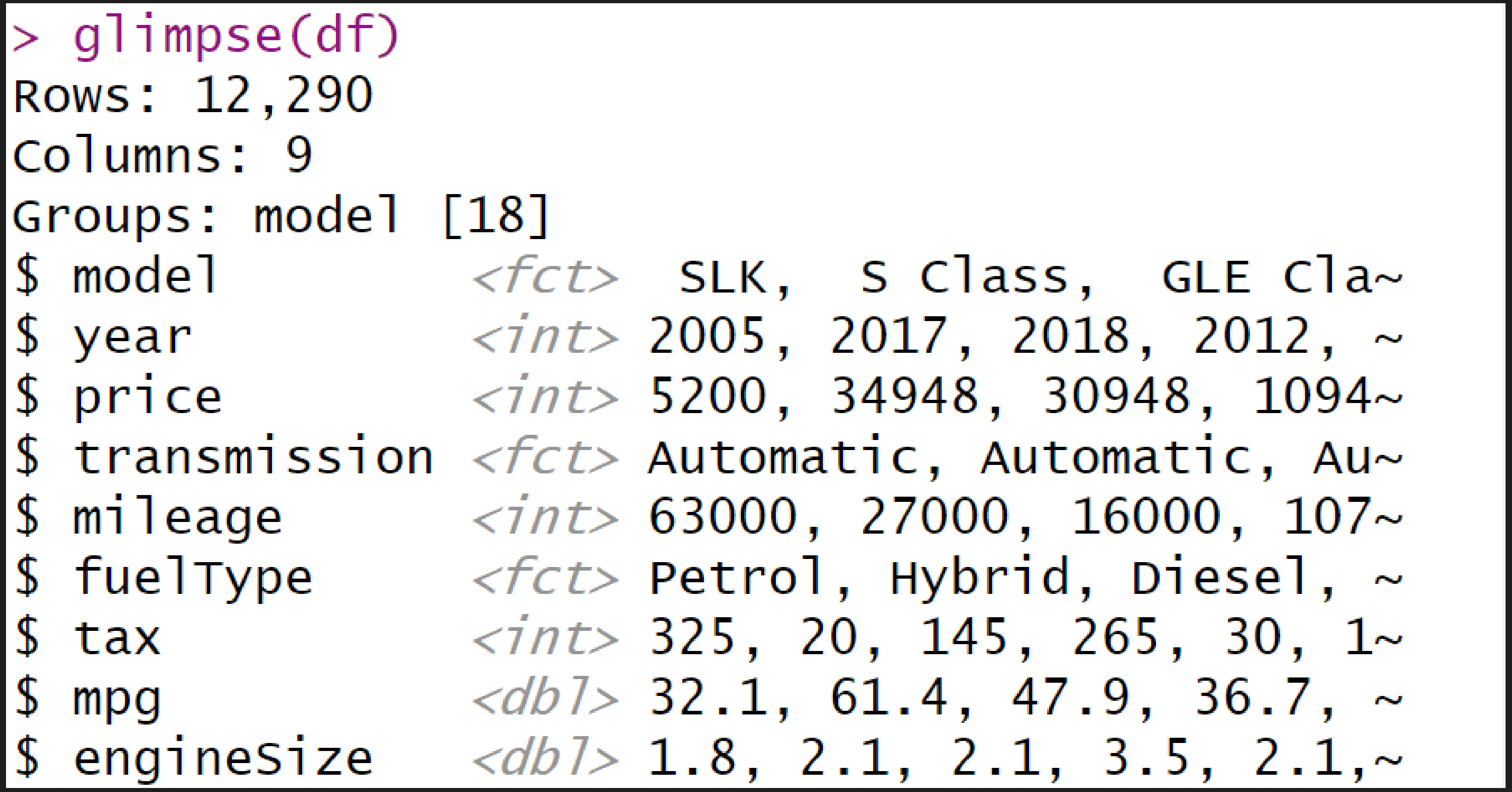


Distribution become quite normal
after outliers was removed

FINAL DATASET



model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
SLK	2005	5200	Automatic	63000	Petrol	325	32.1	1.8
S Class	2017	34948	Automatic	27000	Hybrid	20	61.4	2.1
GLE Class	2018	30948	Automatic	16000	Diesel	145	47.9	2.1
S Class	2012	10948	Automatic	107000	Petrol	265	36.7	3.5
GLA Class	2017	19750	Automatic	15258	Diesel	30	64.2	2.1



```
> glimpse(df)
Rows: 12,290
Columns: 9
Groups: model [18]
$ model        <fct>  SLK,  S Class,  GLE Cla~
$ year         <int> 2005, 2017, 2018, 2012, ~
$ price        <int> 5200, 34948, 30948, 1094~
$ transmission <fct> Automatic, Automatic, Au~
$ mileage      <int> 63000, 27000, 16000, 107~
$ fuelType     <fct> Petrol, Hybrid, Diesel, ~
$ tax          <int> 325, 20, 145, 265, 30, 1~
$ mpg          <dbl> 32.1, 61.4, 47.9, 36.7, ~
$ engineSize   <dbl> 1.8, 2.1, 2.1, 3.5, 2.1,~
```

- Final clean dataset are ready for model training
  - 12,290 rows
  - 9 columns
- Data types are in correct format

MODEL TRAINING & EVALUATION

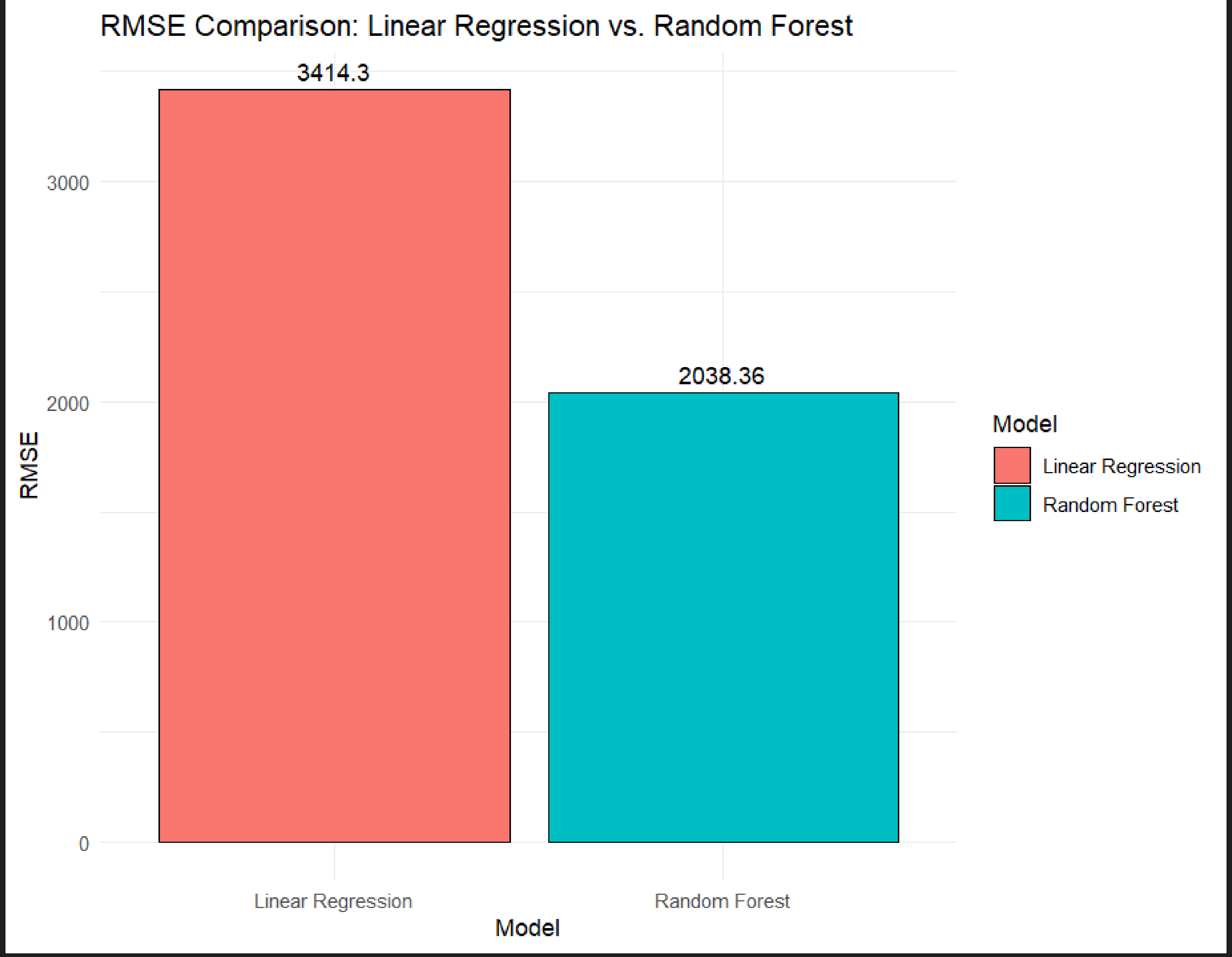
MODEL TRAINING

1. Train test split (80: 20)
2. Model training

   i. Linear regression as baseline model

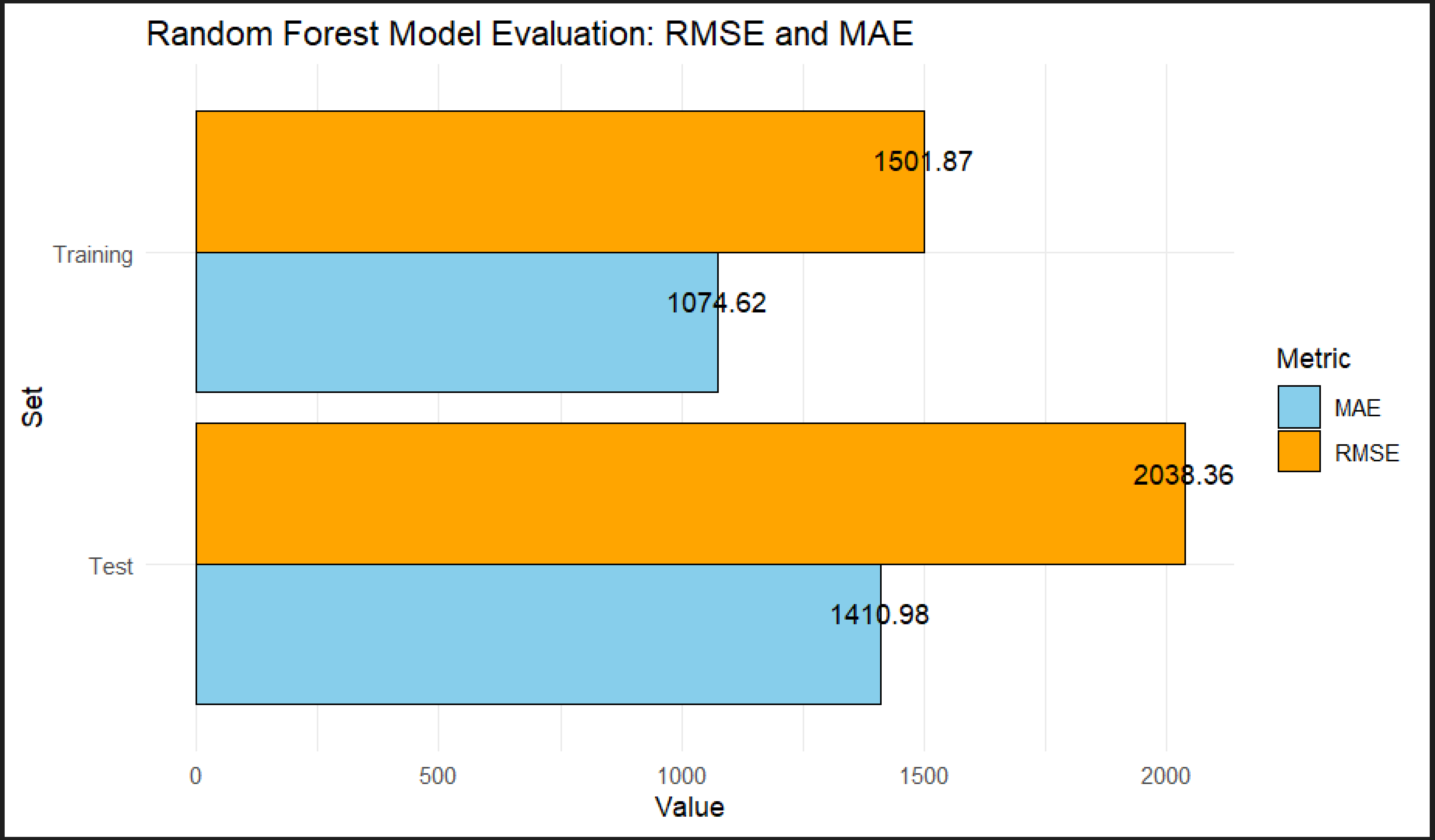
   ii. Random forest
3. Scoring
4. Model evaluation

MODEL EVALUATION



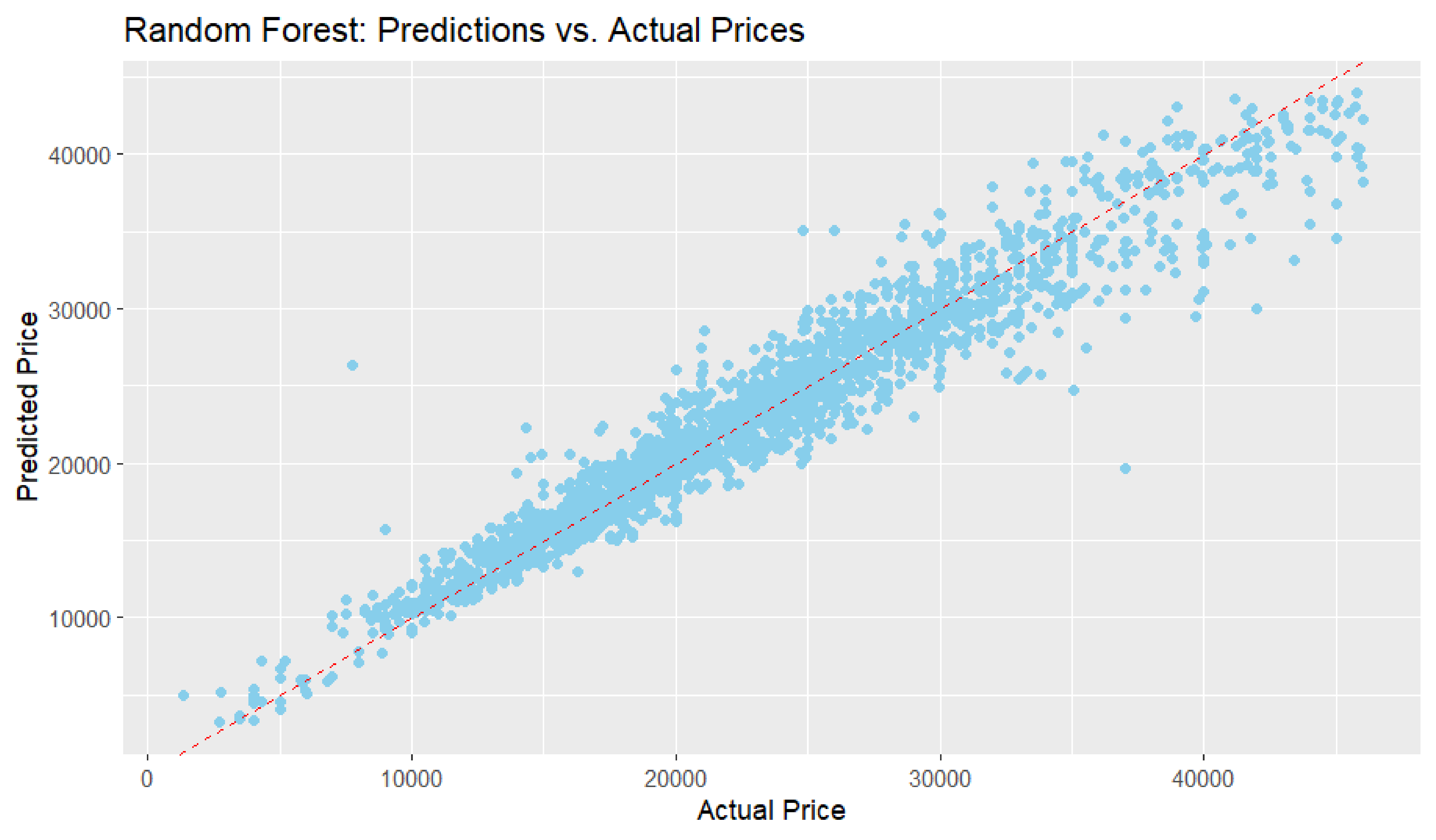
Random Forest outperforms Linear Regression with **lower RMSE.**

MODEL EVALUATION

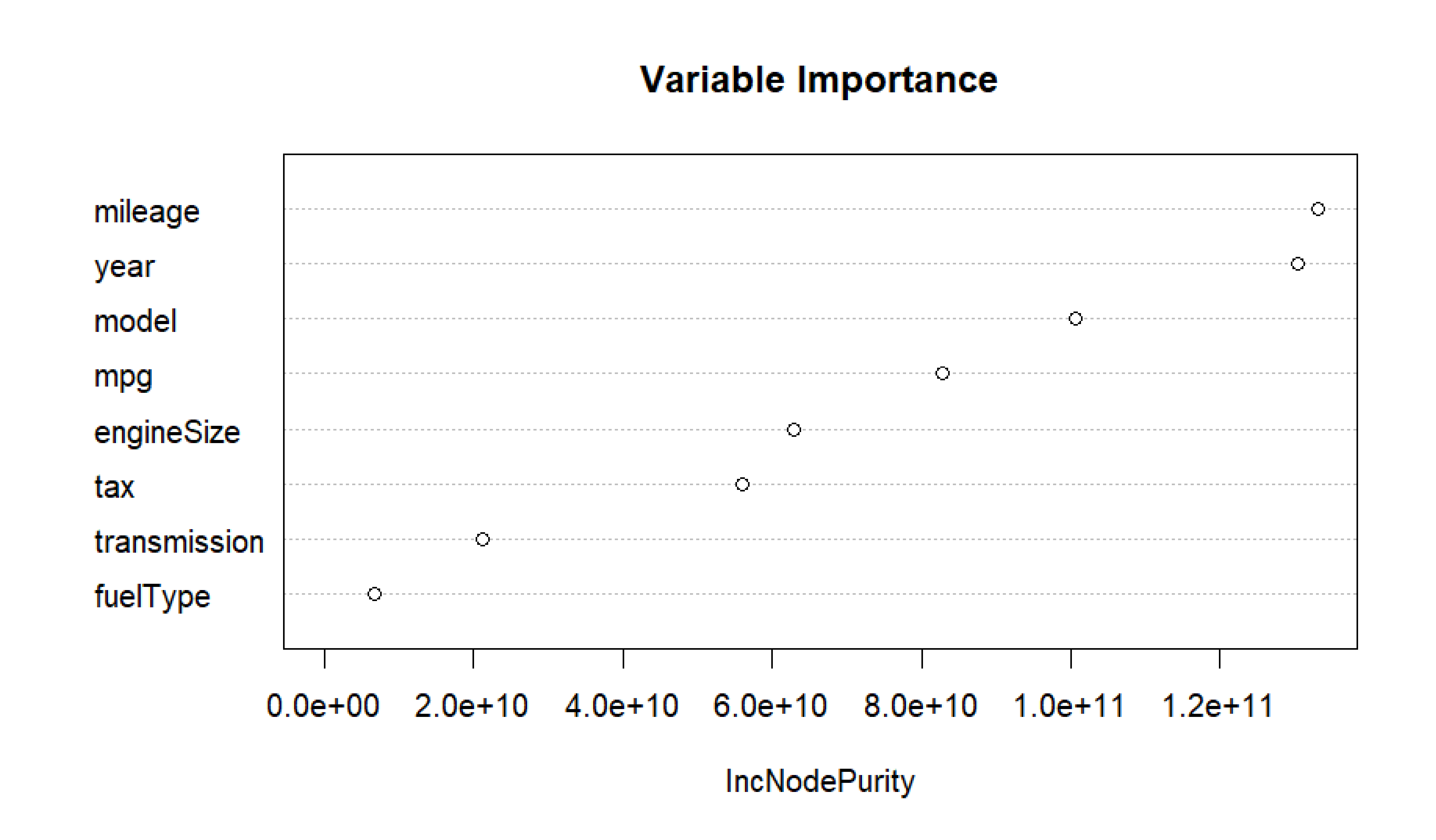


The model performed well with lower RMSE and MAE.

GOOD RESULTS

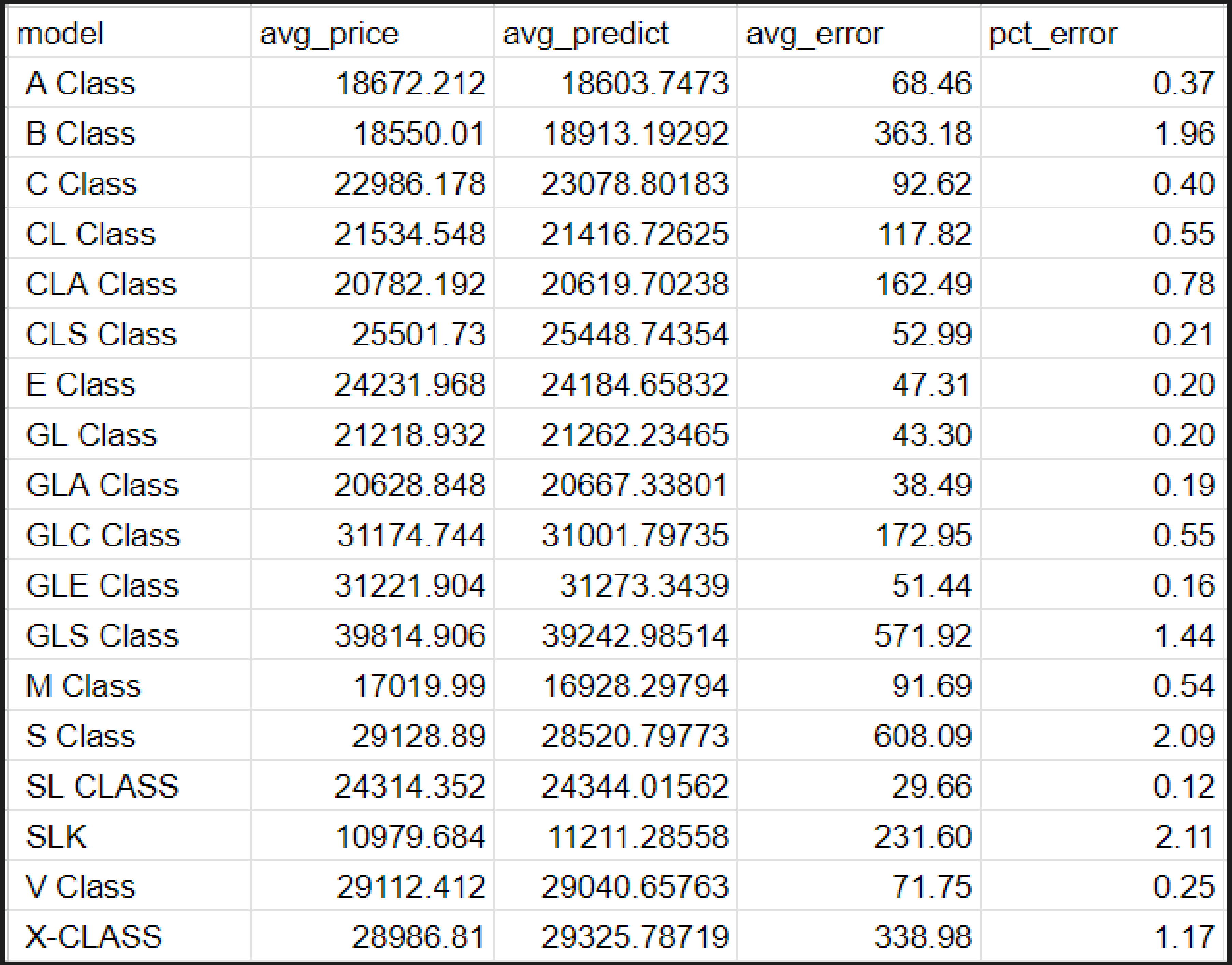


VARIABLE IMPORTANCE



mileage is the highest important feature.

ERROR BY MODELS



model	avg_price	avg_predict	avg_error	pct_error
A Class	18672.212	18603.7473	68.46	0.37
B Class	18550.01	18913.19292	363.18	1.96
C Class	22986.178	23078.80183	92.62	0.40
CL Class	21534.548	21416.72625	117.82	0.55
CLA Class	20782.192	20619.70238	162.49	0.78
CLS Class	25501.73	25448.74354	52.99	0.21
E Class	24231.968	24184.65832	47.31	0.20
GL Class	21218.932	21262.23465	43.30	0.20
GLA Class	20628.848	20667.33801	38.49	0.19
GLC Class	31174.744	31001.79735	172.95	0.55
GLE Class	31221.904	31273.3439	51.44	0.16
GLS Class	39814.906	39242.98514	571.92	1.44
M Class	17019.99	16928.29794	91.69	0.54
S Class	29128.89	28520.79773	608.09	2.09
SL CLASS	24314.352	24344.01562	29.66	0.12
SLK	10979.684	11211.28558	231.60	2.11
V Class	29112.412	29040.65763	71.75	0.25
X-CLASS	28986.81	29325.78719	338.98	1.17

Average % error for most models are under 5%

absolute error on average less than $700

RECOMMENDATIONS

1. Collect more data
2. Try different algorithms
3. Hyperparameter tuning

S CL 2362