Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics & Computer Science Programmes

Car price analisys

Econometrics course final report

Authors:
Dmytro Strus
Taras Martsin

April 2024

1 Introduction

The automotive industry is dynamic and highly competitive, with various factors influencing car prices. Understanding these factors is crucial for both consumers and manufacturers, which will be our main audience. Our objective is to help consumers make educated decisions about what to buy and to help manufacturers optimize their pricing strategies by offering insightful information on the factors that influence automotive prices. Through research and development in this field, we hope to improve market efficiency and enable improved decision-making for all parties involved.

2 The aim and the tasks

The main goal of our research is to analyze car prices, identify the key factors driving price variations and estimate some relations between car features. To achieve this goal, our task will be to choose the proper model with the most significant attributes,test different hypotheses on the impact of different variables on car price. We will consider a linear regression model with dependent variable price and independent variables - car features for investigating impact of different variables on car price.

Our tasks will be following:

- Preliminary data exploration for our aim
- Creating and testing different regression models
- Model corrections and describing
- Providing results and making conclusions

Questions:

1. What features affect the car price? Is popularity significant in estimating car price or not and has a huge impact on price?
2. There is no significant negative correlation between vehicle size and fuel efficiency.
3. The choice of transmission type is not significantly associated with the vehicle size.

3 Literature

[3] Sciencedirect Article: This article discusses the relationship between energy consumption and economic growth

[4] Sagepub Article: This article seems to delve into a specific research study related to cars

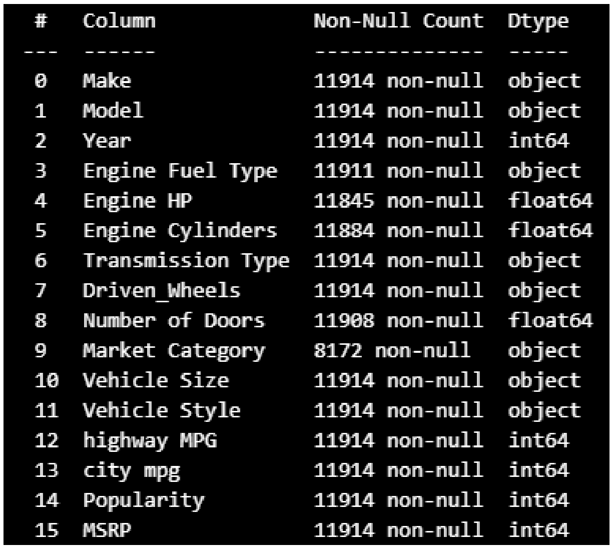
[5] Slideshare Presentation: The presentation covers a car price prediction model or related analysis. ResearchGate Publication: This publication focuses on regression analysis of count data

[6] [7] The websites provides information on car performance, reliability, and customer satisfaction, which could be useful for understanding consumer preferences and their influence on car pricing including automotive research and data analytics services.

[8] [9] helps to estimate inflation rate from 1990 to 2024

4 Data Analysis

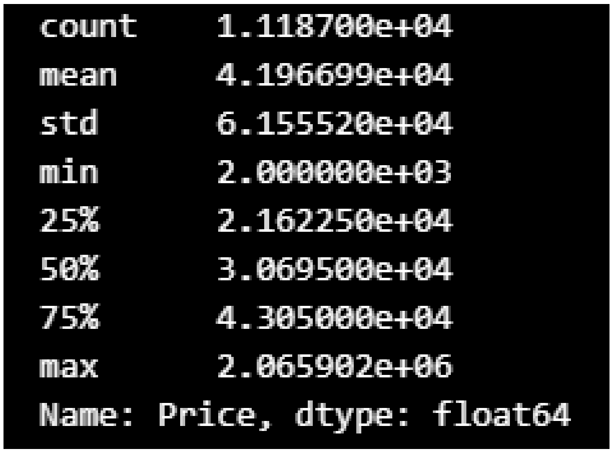
We took a dataset “Car features and MRSP” from Kaggle. It covers the production of cars from 1990 to 2017 with a lot of car features and its MSRP (Price). Notes: All columns explanation is in the end of the report We have renamed several columns: Make -¿ Firm, Engine HP -¿ Horsepower, Engine Fuel Type -¿ Fuel Type, MSRP -¿ Price. In section data analysis words ‘common’ and ‘popular’ means that are used to refer to the majority (bulk) of the cars that have been produced. In order not to confuse ‘popularity’ from the dataset column we will indicate that we made some results based on this column.



```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Make               11914 non-null  object
 1   Model              11914 non-null  object
 2   Year               11914 non-null  int64
 3   Engine Fuel Type   11911 non-null  object
 4   Engine HP          11845 non-null  float64
 5   Engine Cylinders   11884 non-null  float64
 6   Transmission Type  11914 non-null  object
 7   Driven_Wheels      11914 non-null  object
 8   Number of Doors    11908 non-null  float64
 9   Market Category    8172 non-null   object
 10  Vehicle Size       11914 non-null  object
 11  Vehicle Style      11914 non-null  object
 12  highway MPG        11914 non-null  int64
 13  city mpg           11914 non-null  int64
 14  Popularity         11914 non-null  int64
 15  MSRP               11914 non-null  int64
```

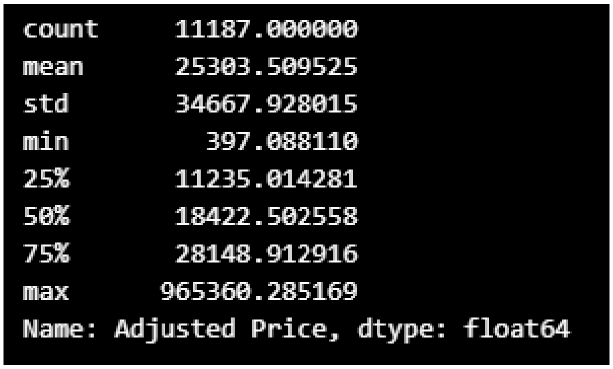
4.1 Cleaning and fixing data

After inspecting the data, we decided to clear our data. We dropped all duplicated rows from our dataset. Next step was dealing with NAN values. For Market Category we have got 3000+ NAN values which is more than 20% of our dataset records so we decided to drop this column. For Engine Cylinders we filled with 0 number of cylinders (because they are electric cars) and the rest were browsing for exact car models and filling corresponding values. For Engine Horsepower we filled horsepower NAN values which matched the firm and engine fuel type with average horsepower for exact firm and engine fuel type. For Engine fuel type we browsed for the exact model and filled corresponding values. For the number of doors we filled the mode number of doors in the dataset. The rows with UNKNOWN Transmission Type were dropped. Another important part was fixing time dependency of price. Since our dataset contains a long period of time (from 1990 to 2017) it was decided to take into account inflation. It was on average 4.87% according to website provided in literature and we recalculated adjusted price in order to remove time dependency on our price Before:



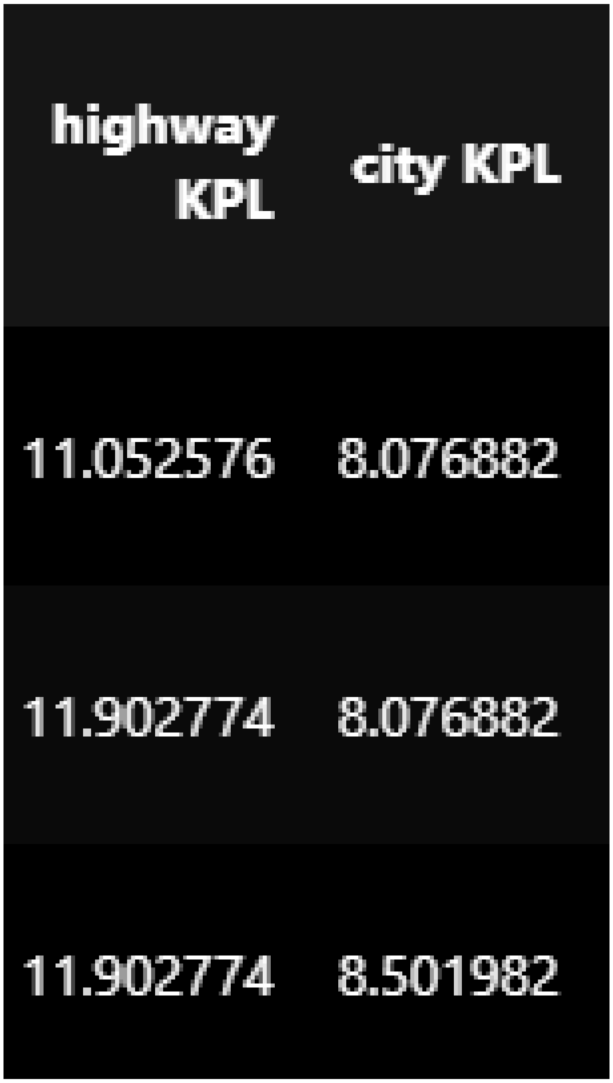
```
count    1.118700e+04
mean     4.196699e+04
std      6.155520e+04
min      2.000000e+03
25%      2.162250e+04
50%      3.069500e+04
75%      4.305000e+04
max      2.065902e+06
Name: Price, dtype: float64
```

After:



```
count     11187.000000
mean      25303.509525
std       34667.928015
min         397.088110
25%       11235.014281
50%       18422.502558
75%       28148.912916
max      965360.285169
Name: Adjusted Price, dtype: float64
```

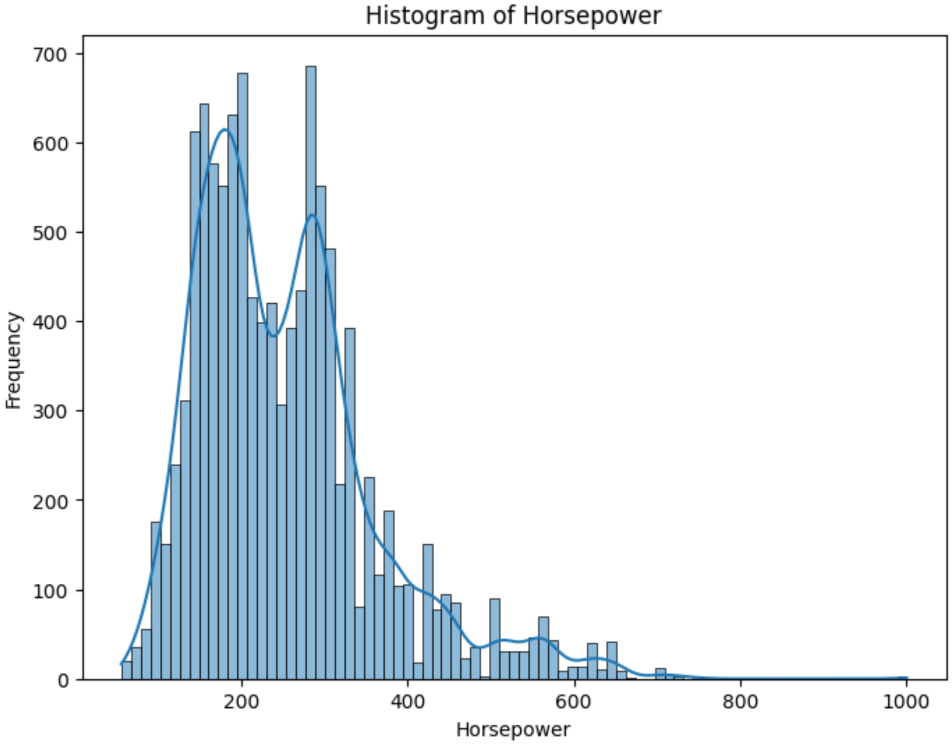
We transformed miles per gallon into kilometers per litre to make it more common.

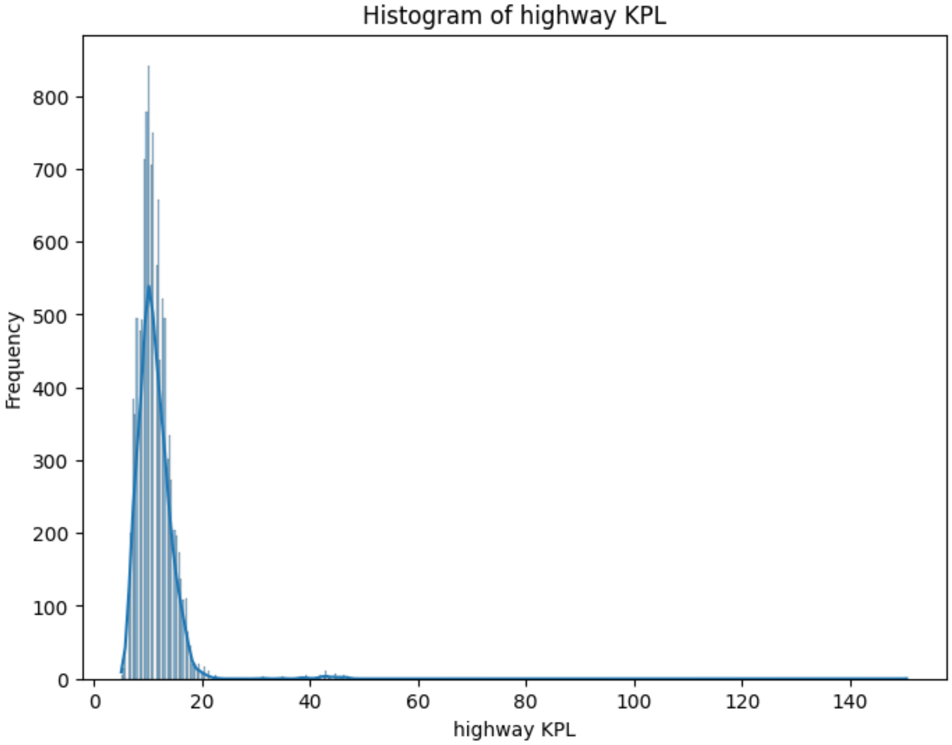


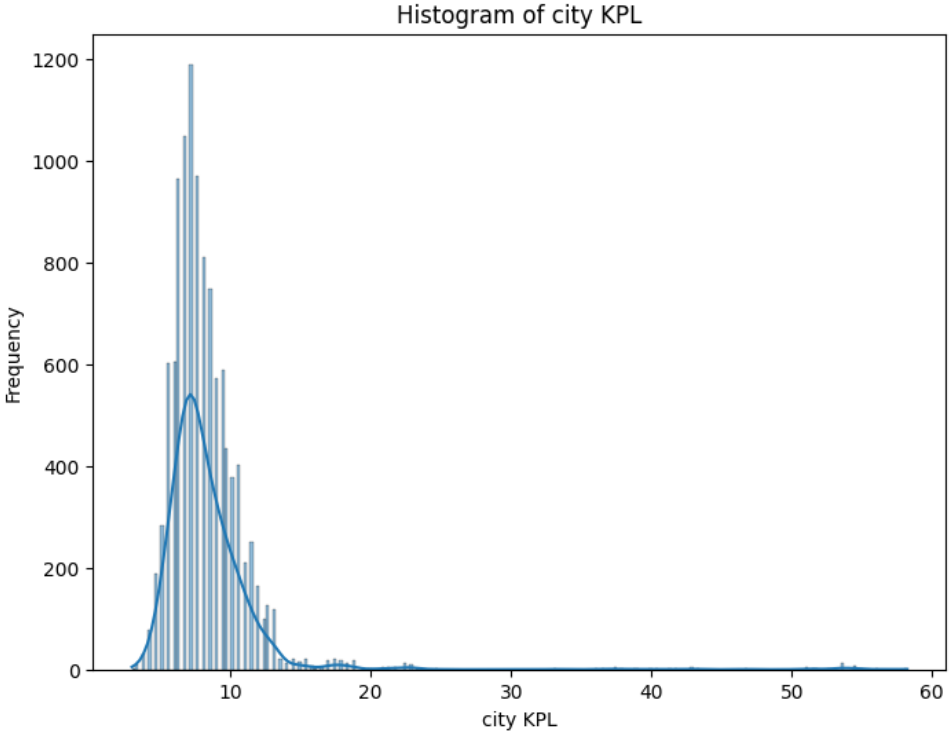
highway KPL	city KPL
11.052576	8.076882
11.902774	8.076882
11.902774	8.501982

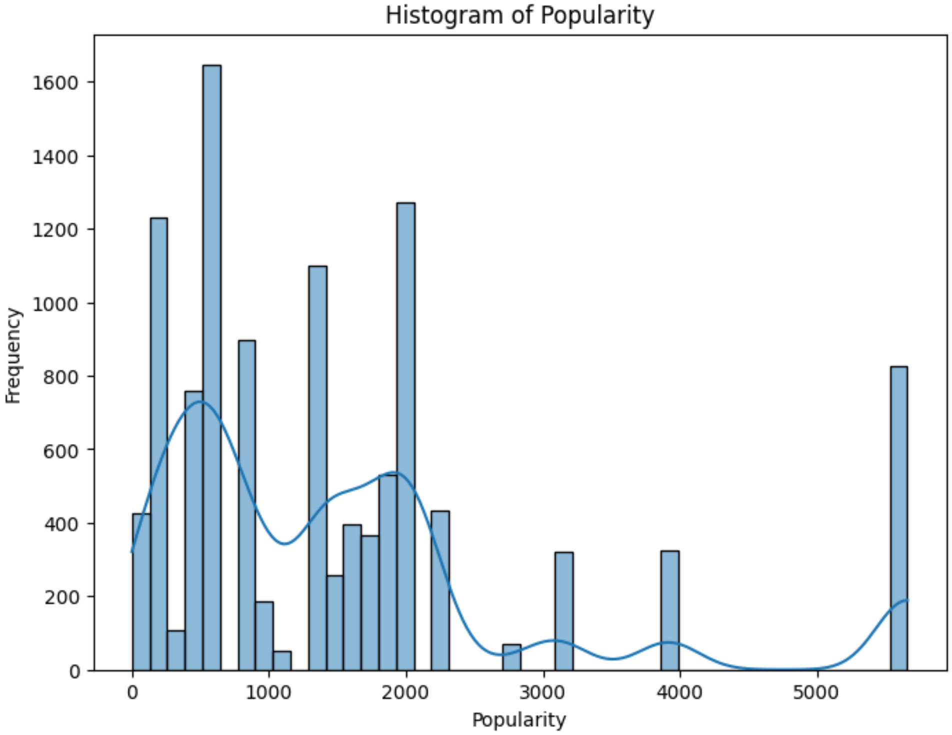
4.2 Insights

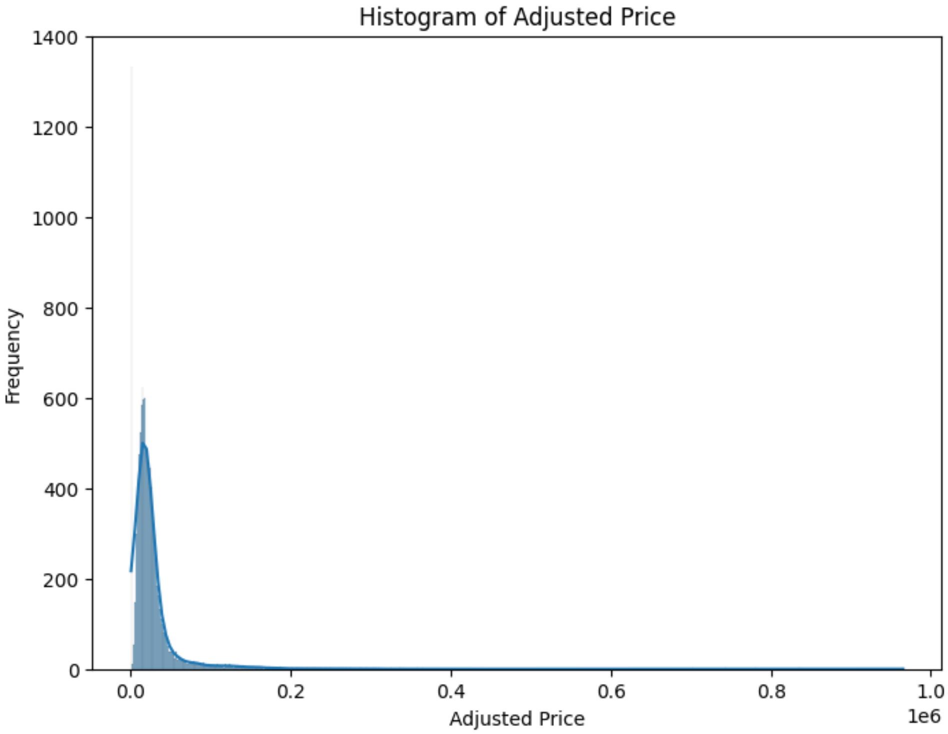
All next steps we will work with already adjusted price We decided to provide a further analysis of the 'Price' column separately, describing its statistics and calculating its 85th percentile value. Additionally, we identified the row(s) with the maximum 'Price' value to gain insights into the highest-priced car(s) in the dataset. And we get that 85% of the prices are below 35 500$ and the other 15% are higher. There are some cars with 1 mln$ which are making strong outliers. Data Visualization: We created histograms to visualize the distributions of numerical columns ('Engine HP', 'highway MPG', 'city mpg', 'Popularity', 'Price'). These visualizations help understand the data distribution and identify any potential outliers.


Histogram of Horsepower
Frequency
0
100
200
300
400
500
600
700
200
400
600
800
1000
Horsepower

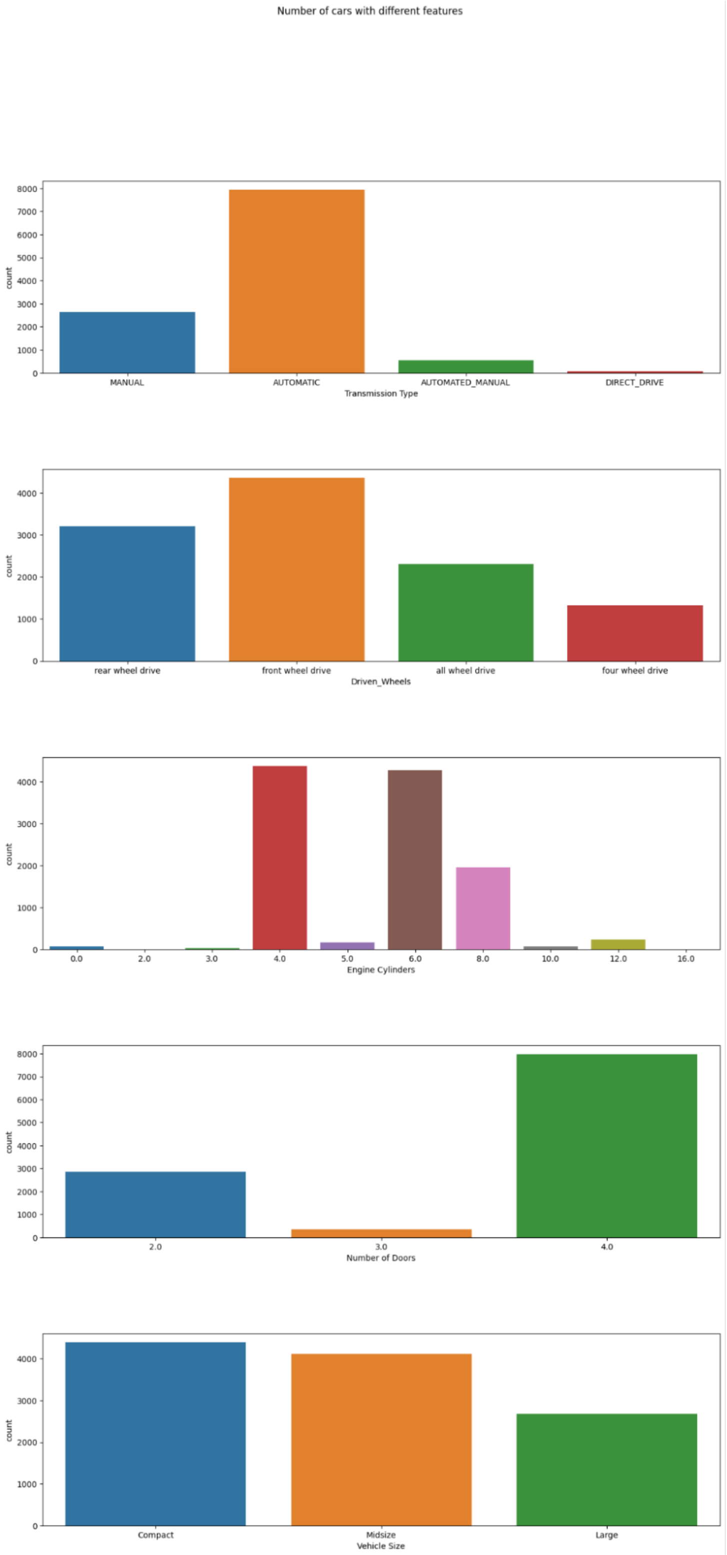

Histogram of highway KPL
Frequency
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120
140
highway KPL


Histogram of city KPL
Frequency
0
200
400
600
800
1000
1200
city KPL
10
20
30
40
50
60

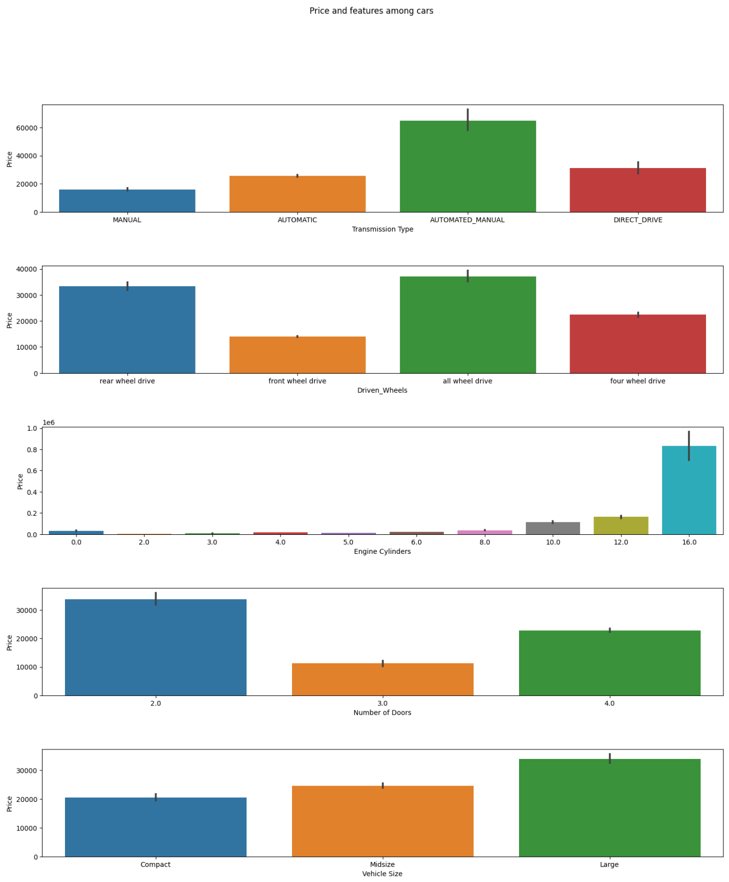

Histogram of Popularity
Frequency
0
200
400
600
800
1000
1200
1400
1600
Popularity
0
1000
2000
3000
4000
5000



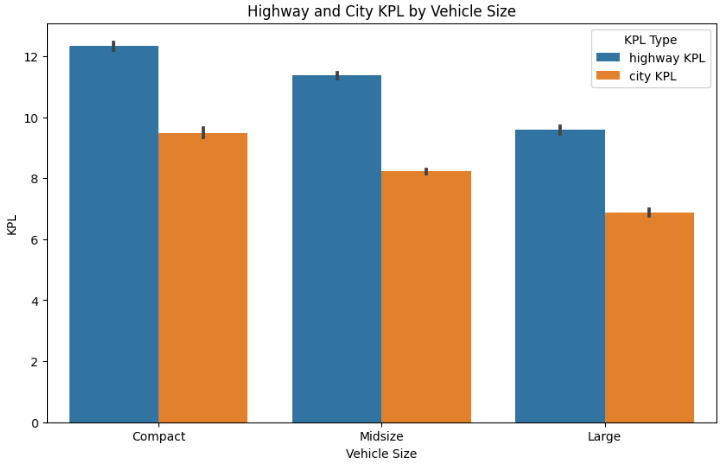
The most common horsepower of cars are varied from 100 to 400 horsepower. City mpg is varied from 10 to 40 when the highway mpg is in range from 10 to 50. The popularity of cars has a kind of random variation so we can not at first glance say what the common value for it. And for price we can see the most concentrated number of price from $2000 to 30000$ and the luxury and sport cars making strong outliers distort the plot.



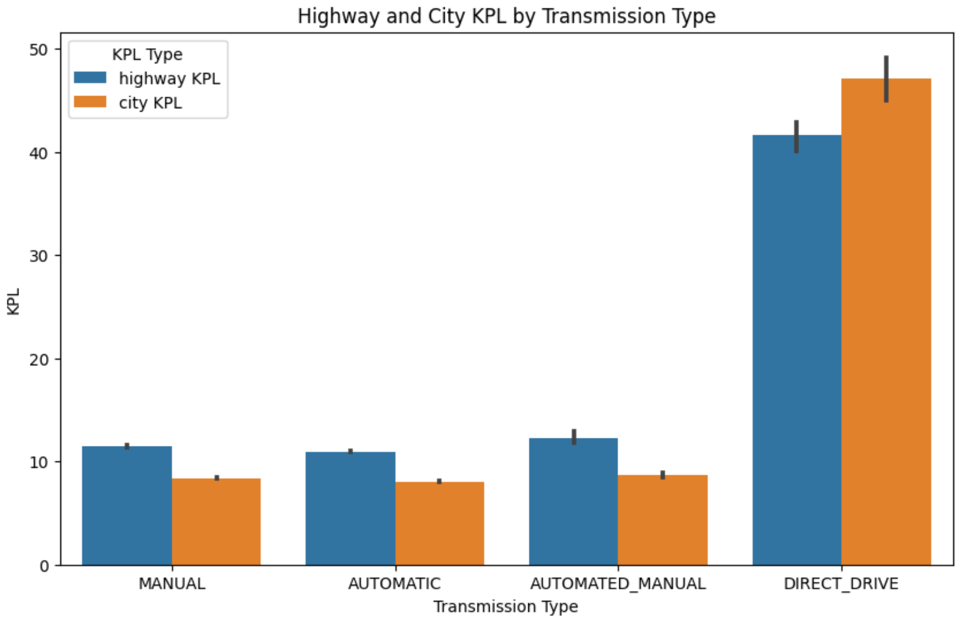
The most produced cars of firms are with four doors with different sizes approximately equally distributed. The automatic transmission type has a dominant position in cars. The main driven wheels for cars are front wheels but there's quite a few cars with all rear wheels and front rear wheels. The most common number of engine cylinders for cars are 4, 6 and 8 which have an influence on horsepower of engine.



The bar plots with average price of different features provides us preliminary insights that mixed transmission type is the most expensive, larger cars are more expensive that other, but cars with only 2 doors have more price rather than with 3 or 3 doors. The cars with 16 engine cylinders obviously are the most expensive because more cylinders used in more powerful engines - hence they are sport cars or premium (luxury).



At first glance, there is approximately no difference between compact and midsize in KPLs but large size has less consumption of petrol

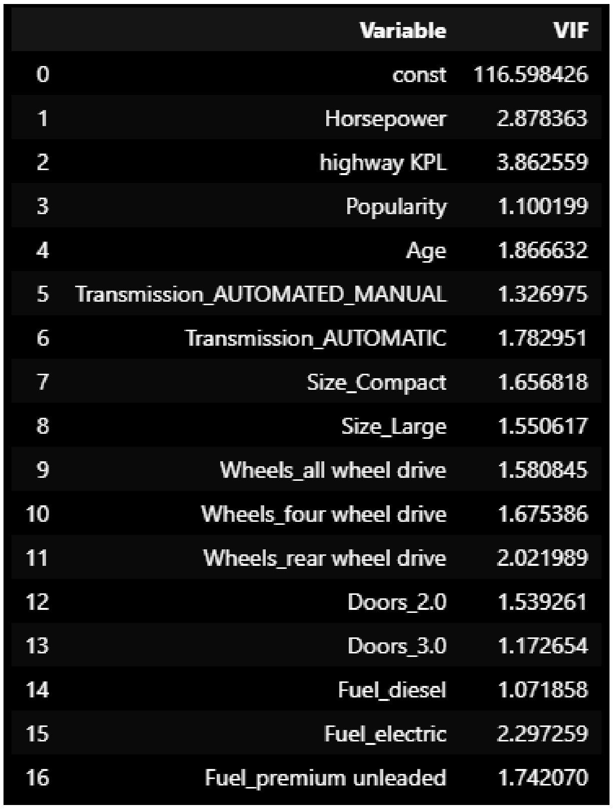


Direct drive cars are the most demanding for energy (KPL here is transformed already because direct drive transmission type have electric cars)

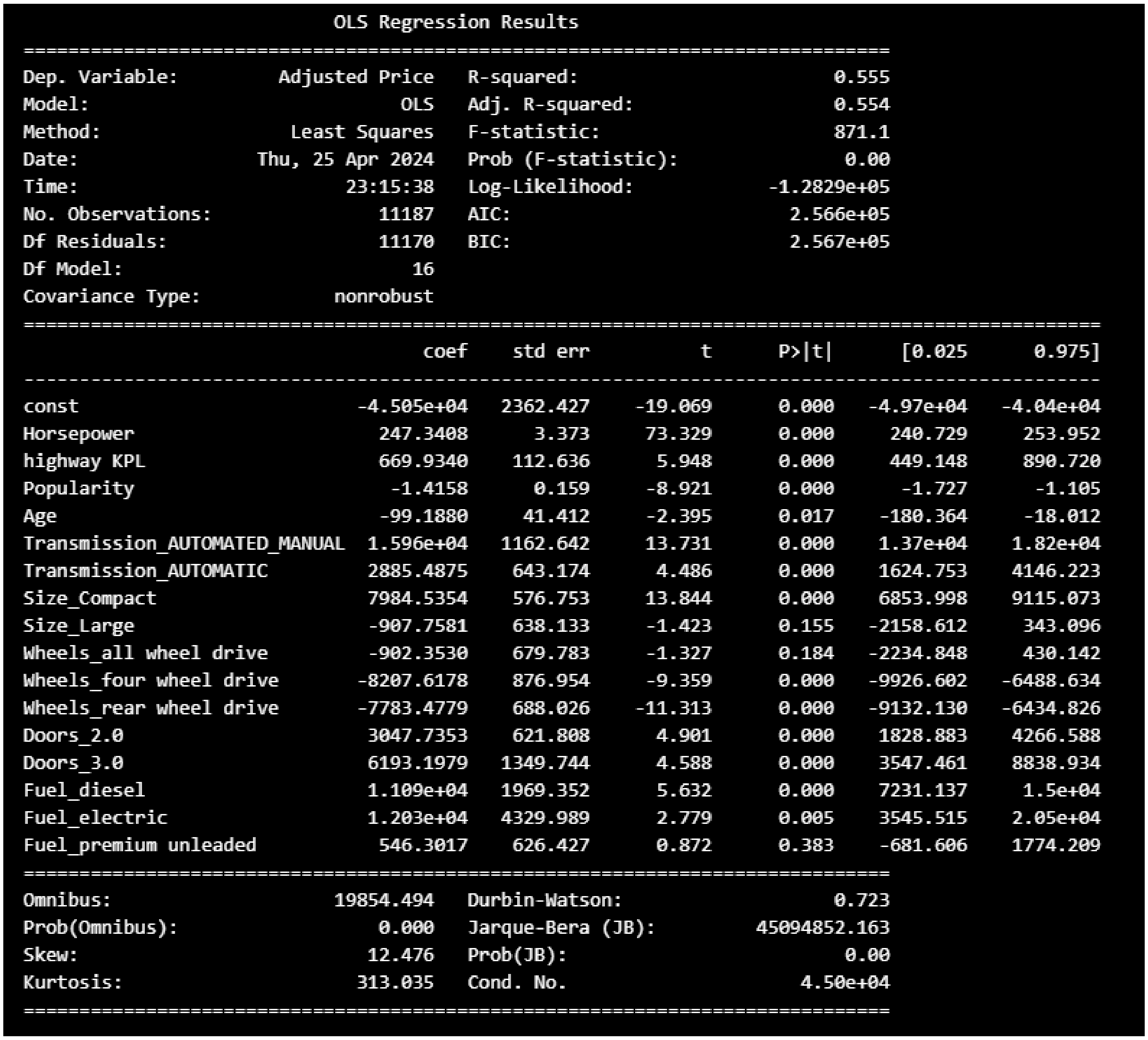
5 Methods

5.1 What features affect the car price? Is popularity significant in estimating car price or not and has a huge impact on price?

We have chosen the OLS regression model to answer the question what affects car price and is popularity really important for car price. Firstly, we have transformed several independent variables into dummy variables There are 4 groups of fuel type: premium unleaded, regular unleaded, electric, diesel. 3 groups of size: compact, midsize, large size 4 groups of driven wheels: all wheels, front wheels, rear wheels and four wheels 3 groups of doors: 2, 3 or 4 doors in car 4 groups of transmission type: manual, automated, direct drive and automated-manual All of them were included into the model and also considering numerical variables: age (a special variable that was created in a way that from the current 2024 year was subtracted from the year of production of the car), highway KPL, city KPL, horsepower, engine cylinders, popularity. Our base group was a car with 4 doors, midsize, on regular unleaded fuel, with manual transmission type. Corresponding variables were dropped from the model before. Next we checked multicollinearity with Variance Inflation Factor (VIF) criterion and dropped some variables that had larger than 5 VIF score(it is the recommended threshold). We have dropped direct drive (it had very strong correlation with electric cars because all direct drive cars are electric cars), engine cylinders (it had strong correlation with horsepower because more horsepowers - more powerful engine - more cylinders needed), city KPL (it had strong correlation with highway KPL because it almost the same but represents efficiency on highways not in city).



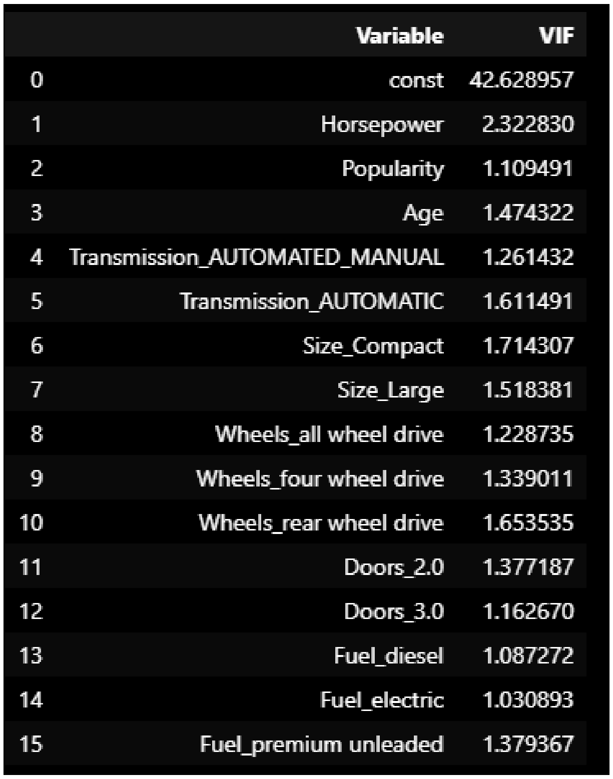
	Variable	VIF
0	const	116.598426
1	Horsepower	2.878363
2	highway KPL	3.862559
3	Popularity	1.100199
4	Age	1.866632
5	Transmission_AUTOMATED_MANUAL	1.326975
6	Transmission_AUTOMATIC	1.782951
7	Size_Compact	1.656818
8	Size_Large	1.550617
9	Wheels_all wheel drive	1.580845
10	Wheels_four wheel drive	1.675386
11	Wheels_rear wheel drive	2.021989
12	Doors_2.0	1.539261
13	Doors_3.0	1.172654
14	Fuel_diesel	1.071858
15	Fuel_electric	2.297259
16	Fuel_premium unleaded	1.742070



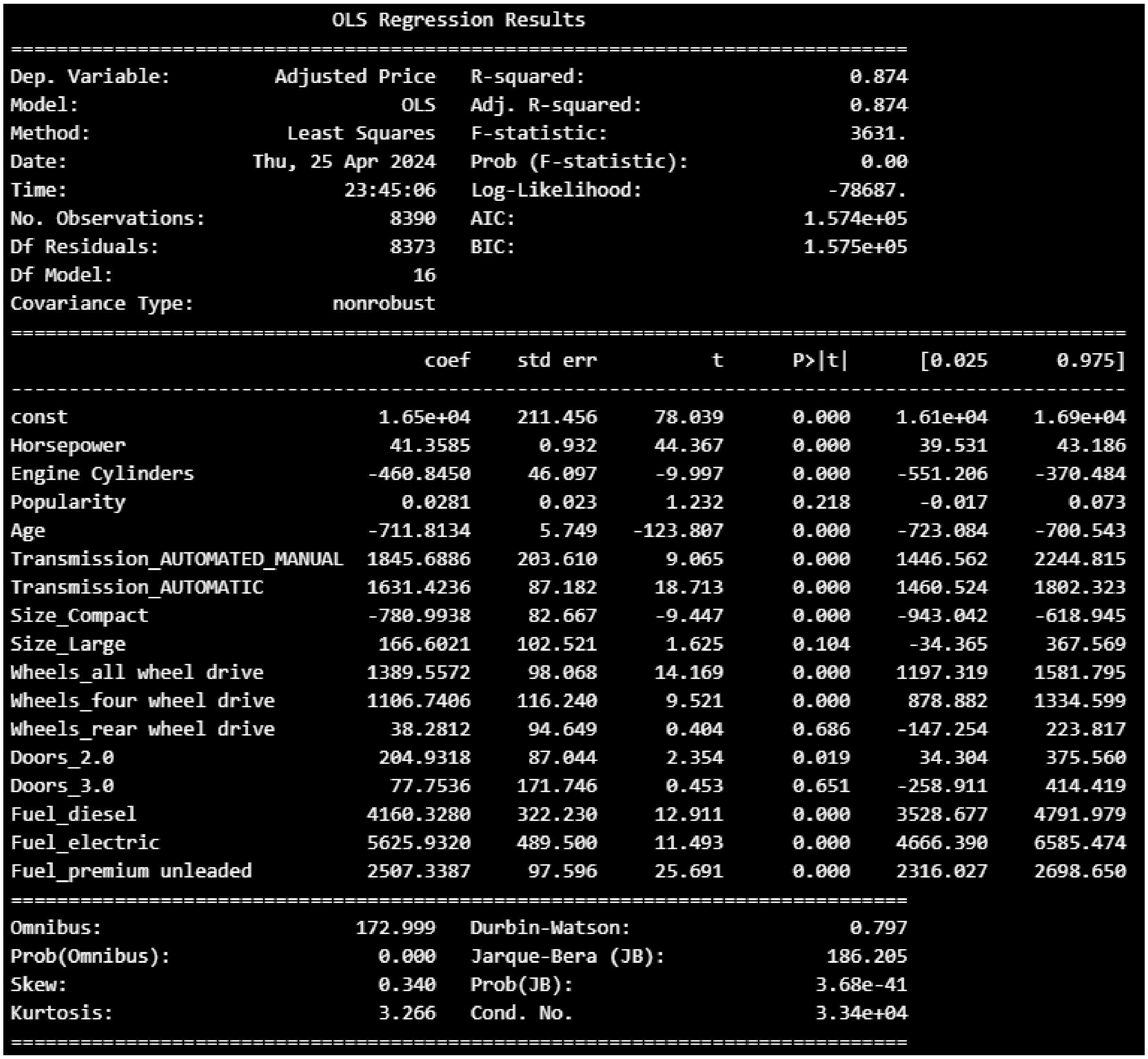
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Adjusted Price   R-squared:                       0.555
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     871.1
Date:                Thu, 25 Apr 2024   Prob (F-statistic):               0.00
Time:                        23:15:38   Log-Likelihood:            -1.2829e+05
No. Observations:               11187   AIC:                         2.566e+05
Df Residuals:                   11170   BIC:                         2.567e+05
Df Model:                          16
Covariance Type:            nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                         -4.505e+04   2362.427    -19.069      0.000   -4.97e+04   -4.04e+04
Horsepower                      247.3408      3.373     73.329      0.000     240.729     253.952
highway KPL                     669.9340    112.636      5.948      0.000     449.148     890.720
Popularity                       -1.4158      0.159     -8.921      0.000      -1.727      -1.105
Age                             -99.1880     41.412     -2.395      0.017    -180.364     -18.012
Transmission_AUTOMATED_MANUAL  1.596e+04   1162.642     13.731      0.000    1.37e+04    1.82e+04
Transmission_AUTOMATIC         2885.4875    643.174      4.486      0.000    1624.753    4146.223
Size_Compact                   7984.5354    576.753     13.844      0.000    6853.998    9115.073
Size_Large                     -907.7581    638.133     -1.423      0.155   -2158.612     343.096
Wheels_all wheel drive         -902.3530    679.783     -1.327      0.184   -2234.848     430.142
Wheels_four wheel drive       -8207.6178    876.954     -9.359      0.000   -9926.602   -6488.634
Wheels_rear wheel drive       -7783.4779    688.026    -11.313      0.000   -9132.130   -6434.826
Doors_2.0                      3047.7353    621.808      4.901      0.000    1828.883    4266.588
Doors_3.0                      6193.1979   1349.744      4.588      0.000    3547.461    8838.934
Fuel_diesel                    1.109e+04   1969.352      5.632      0.000    7231.137     1.5e+04
Fuel_electric                  1.203e+04   4329.989      2.779      0.005    3545.515    2.05e+04
Fuel_premium unleaded           546.3017    626.427      0.872      0.383    -681.606    1774.209
==============================================================================
Omnibus:                    19854.494   Durbin-Watson:                   0.723
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         45094852.163
Skew:                          12.476   Prob(JB):                         0.00
Kurtosis:                     313.035   Cond. No.                     4.50e+04
==============================================================================
```

Using t-test and setting our null hypothesis that the variable is statistically insignificant, we have dropped consequently from the model several variables: premium unleaded, all wheel drive, large size. All of them had larger p-value than 0.05.

We were not satisfied with the results because we took a look at the popularity coefficient. It has value -1.44 which means one unit change in popularity decreases the car's price by 1.44$ dollars. It is nonsense, but this variable is statistically significant. So we decided to make another model but removing all cars that have more that 75% percentile of the car's price from the dataset. We repeated all steps with manual and VIF multicollinearity checking and have got another model.



	Variable	VIF
0	const	42.628957
1	Horsepower	2.322830
2	Popularity	1.109491
3	Age	1.474322
4	Transmission_AUTOMATED_MANUAL	1.261432
5	Transmission_AUTOMATIC	1.611491
6	Size_Compact	1.714307
7	Size_Large	1.518381
8	Wheels_all wheel drive	1.228735
9	Wheels_four wheel drive	1.339011
10	Wheels_rear wheel drive	1.653535
11	Doors_2.0	1.377187
12	Doors_3.0	1.162670
13	Fuel_diesel	1.087272
14	Fuel_electric	1.030893
15	Fuel_premium unleaded	1.379367



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Adjusted Price   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     3631.
Date:                Thu, 25 Apr 2024   Prob (F-statistic):               0.00
Time:                        23:45:06   Log-Likelihood:                -78687.
No. Observations:                8390   AIC:                         1.574e+05
Df Residuals:                    8373   BIC:                         1.575e+05
Df Model:                          16
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                           1.65e+04    211.456     78.039      0.000    1.61e+04    1.69e+04
Horsepower                       41.3585      0.932     44.367      0.000      39.531      43.186
Engine Cylinders               -460.8450     46.097     -9.997      0.000    -551.206    -370.484
Popularity                        0.0281      0.023      1.232      0.218      -0.017       0.073
Age                            -711.8134      5.749   -123.807      0.000    -723.084    -700.543
Transmission_AUTOMATED_MANUAL  1845.6886    203.610      9.065      0.000    1446.562    2244.815
Transmission_AUTOMATIC         1631.4236     87.182     18.713      0.000    1460.524    1802.323
Size_Compact                   -780.9938     82.667     -9.447      0.000    -943.042    -618.945
Size_Large                      166.6021    102.521      1.625      0.104     -34.365     367.569
Wheels_all wheel drive         1389.5572     98.068     14.169      0.000    1197.319    1581.795
Wheels_four wheel drive        1106.7406    116.240      9.521      0.000     878.882    1334.599
Wheels_rear wheel drive          38.2812     94.649      0.404      0.686    -147.254     223.817
Doors_2.0                       204.9318     87.044      2.354      0.019      34.304     375.560
Doors_3.0                        77.7536    171.746      0.453      0.651    -258.911     414.419
Fuel_diesel                    4160.3280    322.230     12.911      0.000    3528.677    4791.979
Fuel_electric                  5625.9320    489.500     11.493      0.000    4666.390    6585.474
Fuel_premium unleaded          2507.3387     97.596     25.691      0.000    2316.027    2698.650
==============================================================================
Omnibus:                      172.999   Durbin-Watson:                   0.797
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              186.205
Skew:                           0.340   Prob(JB):                     3.68e-41
Kurtosis:                       3.266   Cond. No.                     3.34e+04
==============================================================================
```

And using t-test and step by step dropping insignificant variables - popularity, large size, rear wheels drive, three doors - and have got another which had no more popularity in variables because it was insignificant

5.2 There is no significant negative correlation between vehicle size and fuel efficiency.

For testing the hypothesis if there is correlation between the Vehicle Size and Fuel Efficiency, the OLS model was used. Firstly, we mapped the Vehicle Size: “Compact” - 1; “Midsize” - 2; “Large” - 3 After that, we used “highway KPL” and “city KPL” as metrics for calculating the Fuel Efficiency. After performing the OLS with metrics as independent variables, the coefficient and the intercept can be interpreted as the change in ”highway KPL” for a one-unit increase in ”Vehicle Size (Mapped)” (from Compact to Midsize, or from Midsize to Large) and the intercept represents the expected value of ”highway KPL” when all independent variables are zero. In our case, it’s the estimated highway KPL for a ”Vehicle Size (Mapped)” of zero (which doesn’t correspond to a real vehicle size but serves as a mathematical anchor). For making sure of it, one can get an OLS summary, where p-value will indicate the sagnificance.

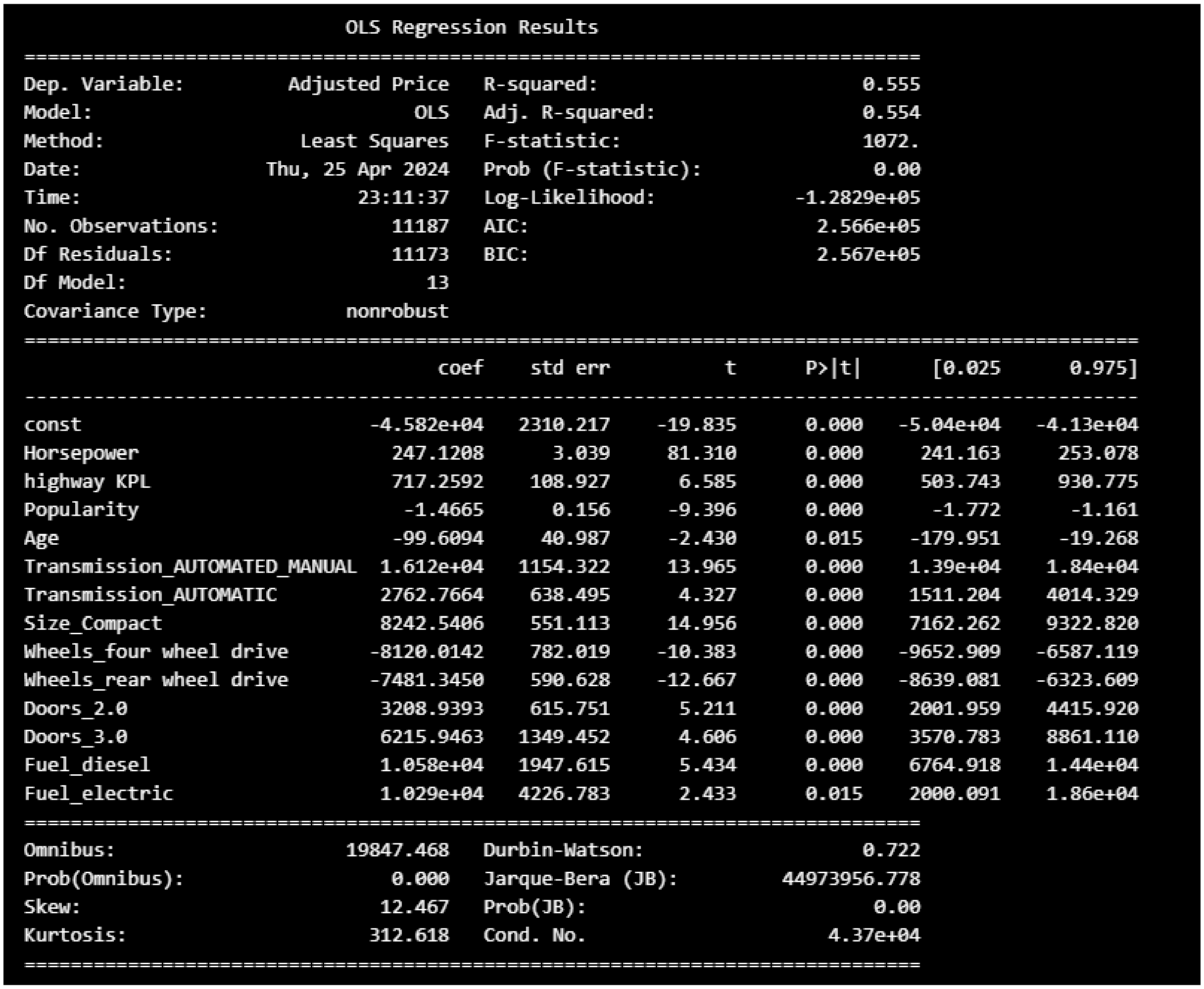
5.3 The choice of transmission type is not significantly associated with the vehicle size.

Same methodology will work for testing the correlation between the Vehicle Size and Transmission Type. Same mapping goes for Vehicle size After that, we need to transform TransmissionType into dummy variables and perform the OLS test where Vehicle size is TransmissionType is independent variable. We can also treat Constant as the expected value of the dependent variable when all independent variables are zero. In our case, when none of the dummy variables representing transmission types are present (all transmission types are zero), the constant represents the expected vehicle size. And the p-value will show us the probability of observing the estimated coefficient.

6 Results

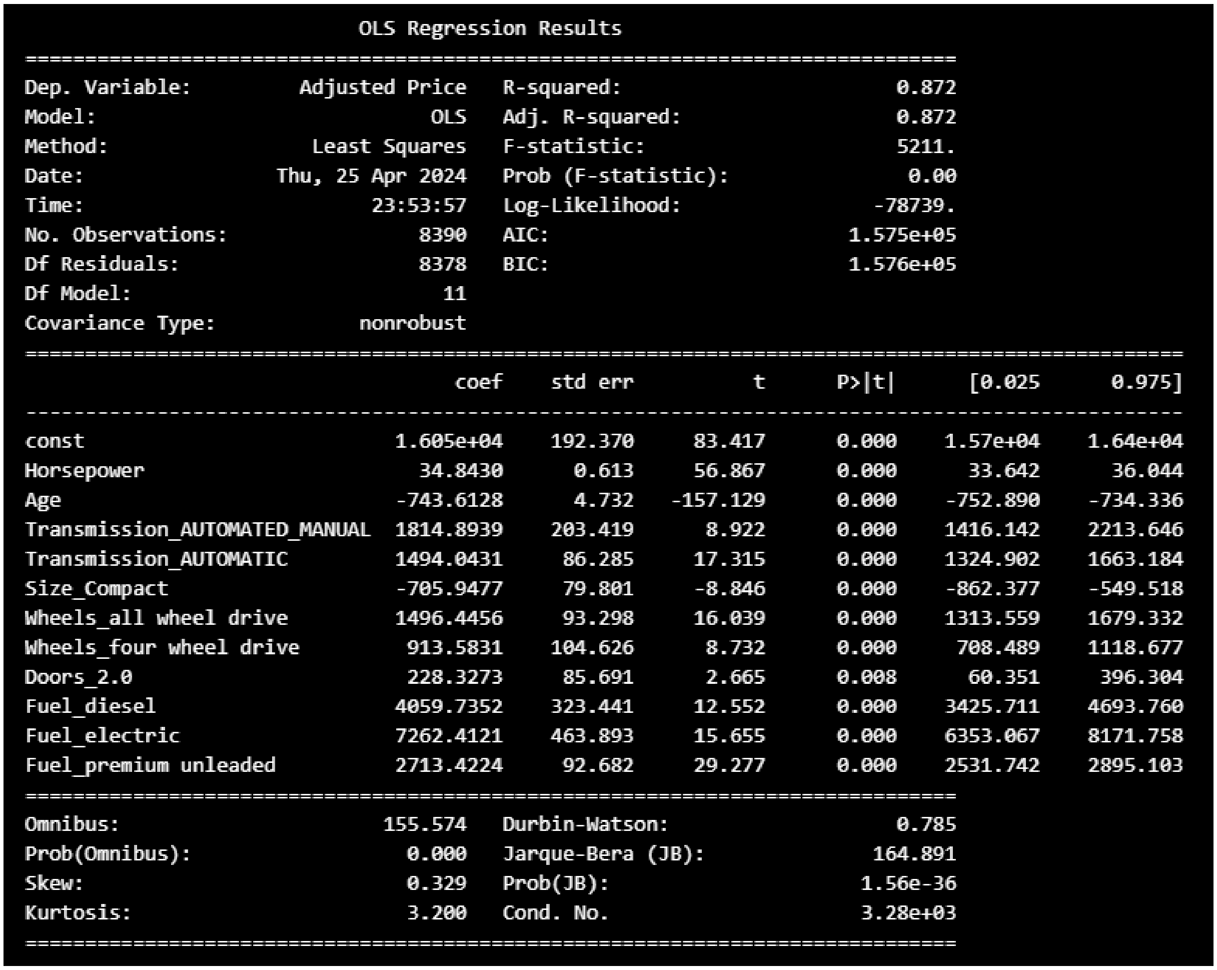
6.1 Car price

Preliminary analysis revealed interesting insights into the dataset. Further analysis focused on the 'Price' column, revealing that 85% of prices fell below $35,000, with outliers reaching as high as $1 million. Horsepower, city and highway mpg, and car popularity showed varying distributions, with concentrations and outliers affecting the plots. The leading car manufacturers primarily produce four-door vehicles across various sizes, with automatic transmission being the preferred choice among consumers. While front-wheel drive is predominant, there's a notable presence of both rear-wheel and all-wheel drive configurations. The most common engine cylinder configurations are 4, 6, and 8, which significantly impact horsepower. Analyzing average prices across different features reveals that cars with mixed transmission types tend to be the priciest. Larger vehicles generally command higher prices, although surprisingly, two-door models often carry a higher price tag compared to three or four-door variants. Notably, cars with 16 cylinders stand out as the most expensive due to their association with high-performance or luxury vehicles. Initial observations suggest minimal differences in fuel efficiency between compact and midsize cars, with larger vehicles generally consuming less petrol. However, cars with direct drive transmission types, particularly electric vehicles, exhibit higher energy demands, as reflected in their kilometers per liter (KPL) figures. The corresponding final models:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Adjusted Price   R-squared:                       0.555
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     1072.
Date:                Thu, 25 Apr 2024   Prob (F-statistic):               0.00
Time:                        23:11:37   Log-Likelihood:            -1.2829e+05
No. Observations:               11187   AIC:                         2.566e+05
Df Residuals:                   11173   BIC:                         2.567e+05
Df Model:                          13
Covariance Type:            nonrobust
================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       -4.582e+04   2310.217    -19.835      0.000   -5.04e+04   -4.13e+04
Horsepower                    247.1208      3.039     81.310      0.000     241.163     253.078
highway KPL                   717.2592    108.927      6.585      0.000     503.743     930.775
Popularity                     -1.4665      0.156     -9.396      0.000      -1.772      -1.161
Age                           -99.6094     40.987     -2.430      0.015    -179.951     -19.268
Transmission_AUTOMATED_MANUAL  1.612e+04   1154.322     13.965      0.000    1.39e+04    1.84e+04
Transmission_AUTOMATIC        2762.7664    638.495      4.327      0.000    1511.204    4014.329
Size_Compact                  8242.5406    551.113     14.956      0.000    7162.262    9322.820
Wheels_four wheel drive      -8120.0142    782.019    -10.383      0.000   -9652.909   -6587.119
Wheels_rear wheel drive      -7481.3450    590.628    -12.667      0.000   -8639.081   -6323.609
Doors_2.0                     3208.9393    615.751      5.211      0.000    2001.959    4415.920
Doors_3.0                     6215.9463   1349.452      4.606      0.000    3570.783    8861.110
Fuel_diesel                   1.058e+04   1947.615      5.434      0.000    6764.918    1.44e+04
Fuel_electric                 1.029e+04   4226.783      2.433      0.015    2000.091    1.86e+04
==============================================================================
Omnibus:                    19847.468   Durbin-Watson:                   0.722
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         44973956.778
Skew:                          12.467   Prob(JB):                         0.00
Kurtosis:                     312.618   Cond. No.                     4.37e+04
==============================================================================
```

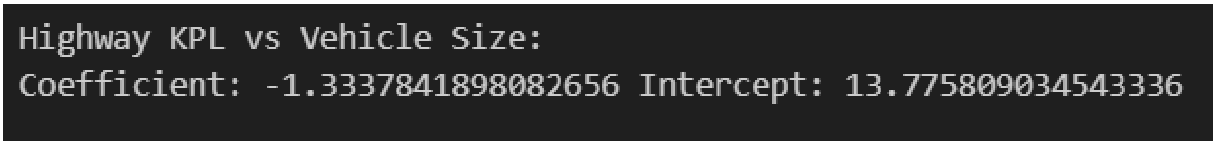
Horsepower: For each additional unit of horsepower, the Adjusted Price is estimated to increase by $247.12. highway KPL: A one-unit increase in highway kilometers per liter is associated with an estimated increase in Adjusted Price of $717.26. Popularity: For each one-unit increase in Popularity, the Adjusted Price is estimated to decrease by $1.47. Age: A one-unit increase in Age is associated with an estimated decrease in Adjusted Price of $99.61. Transmission_AUTOMATED_MANUAL: Compared to manual transmission, vehicles with an AUTOMATED MANUAL (mixed) transmission have an estimated increase in Adjusted Price of $16,120. Transmission_AUTOMATIC: Compared to manual transmission, vehicles with an AUTOMATIC transmission have an estimated increase in Adjusted Price of $2762.77. Size_Compact: Compact-sized vehicles have an estimated increase in Adjusted Price of $8242.54 compared to midsize cars. Wheels_four wheel drive: Vehicles with four-wheel drive have an estimated decrease in Adjusted Price of $8120.01 compared to vehicles with front wheels drive. Wheels_rear wheel drive: Vehicles with rear-wheel drive have an estimated decrease in Adjusted Price of $7481.35 compared to vehicles with front wheel drive. Doors_2.0: Vehicles with 2 doors have an estimated increase in Adjusted Price of $3208.94 compared to vehicles with 4 doors. Doors_3.0: Vehicles with 3 doors have an estimated increase in Adjusted Price of $6215.95 compared to vehicles with 4 doors. Fuel_diesel: Vehicles running on diesel fuel have an estimated increase in Adjusted Price of $10,580 compared to vehicles running on regular unleaded fuel. Fuel_electric: Electric vehicles have an estimated increase in Adjusted Price of $10,290 compared to vehicles running on regular unleaded fuel. Since we dropped insignificant variables there is no difference in price between premium unleaded and regular unleaded, all wheel drive and front wheel drive cars, large size and midsize cars.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Adjusted Price   R-squared:                       0.872
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                     5211.
Date:                Thu, 25 Apr 2024   Prob (F-statistic):               0.00
Time:                        23:53:57   Log-Likelihood:                -78739.
No. Observations:                8390   AIC:                         1.575e+05
Df Residuals:                    8378   BIC:                         1.576e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          1.605e+04    192.370     83.417      0.000    1.57e+04    1.64e+04
Horsepower                       34.8430      0.613     56.867      0.000      33.642      36.044
Age                            -743.6128      4.732   -157.129      0.000    -752.890    -734.336
Transmission_AUTOMATED_MANUAL  1814.8939    203.419      8.922      0.000    1416.142    2213.646
Transmission_AUTOMATIC         1494.0431     86.285     17.315      0.000    1324.902    1663.184
Size_Compact                   -705.9477     79.801     -8.846      0.000    -862.377    -549.518
Wheels_all wheel drive         1496.4456     93.298     16.039      0.000    1313.559    1679.332
Wheels_four wheel drive         913.5831    104.626      8.732      0.000     708.489    1118.677
Doors_2.0                       228.3273     85.691      2.665      0.008      60.351     396.304
Fuel_diesel                    4059.7352    323.441     12.552      0.000    3425.711    4693.760
Fuel_electric                  7262.4121    463.893     15.655      0.000    6353.067    8171.758
Fuel_premium unleaded          2713.4224     92.682     29.277      0.000    2531.742    2895.103
==============================================================================
Omnibus:                      155.574   Durbin-Watson:                   0.785
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              164.891
Skew:                           0.329   Prob(JB):                     1.56e-36
Kurtosis:                       3.200   Cond. No.                     3.28e+03
==============================================================================
```

Horsepower: For each additional unit of horsepower, the Adjusted Price is estimated to increase by $34.84. Age: For each additional year of age, the price is estimated to decrease by $743.61. Transmission_AUTOMATED_MANUAL: Compared to manual transmission, vehicles with an AUTOMATED MANUAL (mixed) transmission have an estimated increase in Adjusted Price of $1814.89. Transmission_AUTOMATIC: Compared to manual transmission, vehicles with an AUTOMATIC transmission have an estimated increase in Adjusted Price of $1494.04. Size_Compact: Vehicles categorized as Compact have an estimated decrease in Adjusted Price of $705.95 compared to midsize cars. Wheels_all wheel drive: Vehicles with all-wheel drive have an estimated increase in Adjusted Price of $1496.45 compared to front wheels drive. Wheels_four wheel drive: Vehicles with four-wheel drive have an estimated increase in Adjusted Price of $913.58 compared to vehicles to front wheels drives. Doors_2.0: Vehicles with 2 doors have an estimated increase in Adjusted Price of $228.33 compared to vehicles with 4 doors. Fuel_diesel: Vehicles running on diesel fuel have an estimated increase in Adjusted Price of $4059.74 compared to vehicles running on regular unleaded fuel. Fuel_electric: Electric vehicles have an estimated increase in Adjusted Price of $7262.41 compared to vehicles running on regular unleaded fuel. Fuel_premium unleaded: Vehicles running on premium unleaded fuel have an estimated increase in Adjusted Price of $2713.42 compared to vehicles running on regular unleaded fuel. We dropped insignificant variables such as popularity means that price does not depend on price, large size has no difference in price with midsize cars, rear wheel drive does not differ in price with front wheel drive, cars with three doors has no difference in price with cars with four doors.

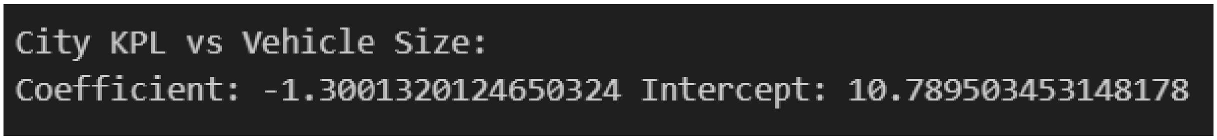
6.2 Vehicle-Fuel Efficiency



```
Highway KPL vs Vehicle Size:
Coefficient: -1.3337841898082656 Intercept: 13.775809034543336
```

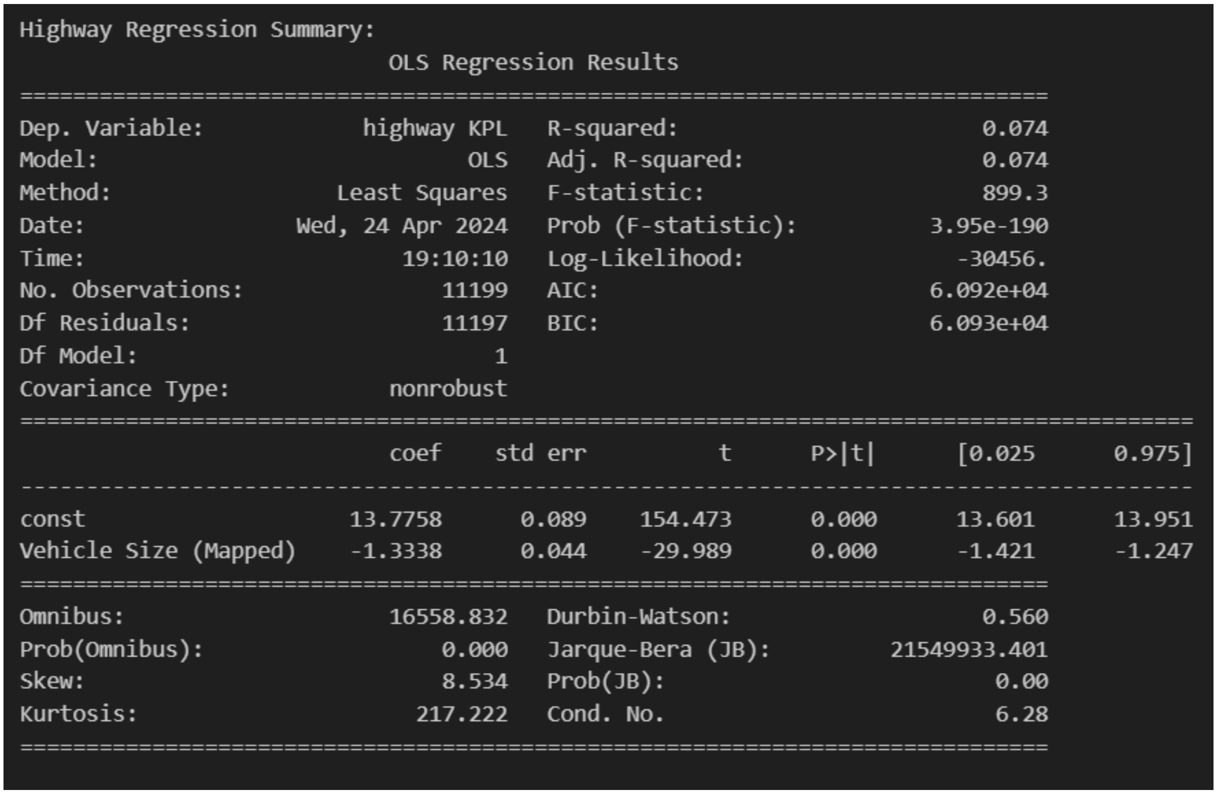
The coefficient is negative (-1.33378), indicating that as the vehicle size increases, the highway kilometers per liter (KPL) decreases. A decrease in KPL suggests reduced fuel efficiency. The intercept is 13.775, which is the expected highway KPL for the smallest vehicle size (Compact). For every one-unit increase in vehicle size (e.g., from ’Compact’ to ’Midsize’), highway KPL decreases by approximately 1.33 units. Thus, as vehicle size increases, highway MPG tends to decrease, suggesting that larger vehicles are less fuel-efficient on highways.

6.3 Vehicle-Fuel Efficiency



```
City KPL vs Vehicle Size:
Coefficient: -1.3001320124650324 Intercept: 10.789503453148178
```

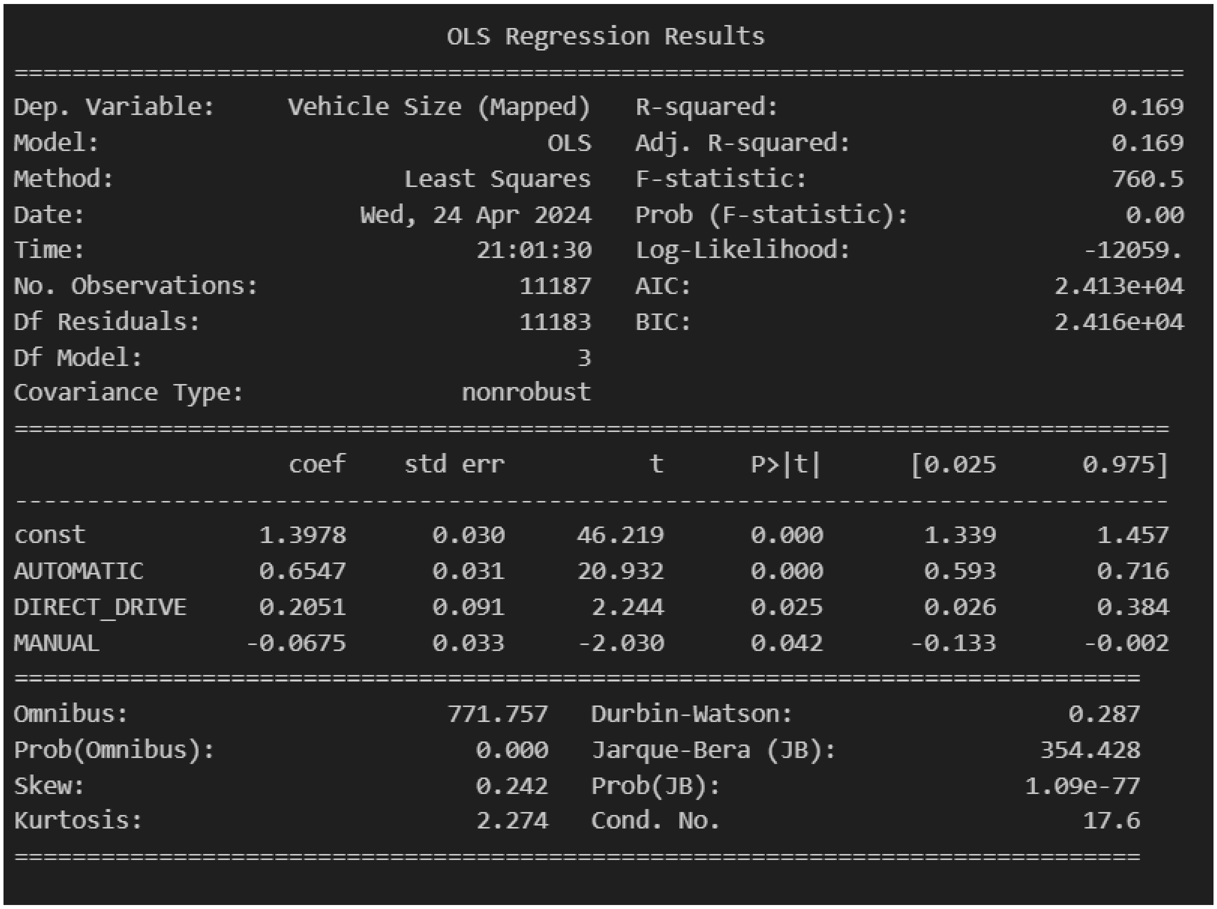
Similar to the highway results, the negative coefficient (-1.30013) suggests that as vehicle size increases, city KPL decreases, indicating reduced fuel efficiency. For every one-unit increase in vehicle size, city KPL decreases by about 1.30 units. The intercept is 10.789, indicating the expected city KPL for the smallest vehicle size.



```
Highway Regression Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:            highway KPL   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.074
Method:                 Least Squares   F-statistic:                     899.3
Date:                Wed, 24 Apr 2024   Prob (F-statistic):          3.95e-190
Time:                        19:10:10   Log-Likelihood:                -30456.
No. Observations:               11199   AIC:                         6.092e+04
Df Residuals:                   11197   BIC:                         6.093e+04
Df Model:                           1
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    13.7758      0.089    154.473      0.000      13.601      13.951
Vehicle Size (Mapped)    -1.3338      0.044    -29.989      0.000      -1.421      -1.247
==============================================================================
Omnibus:                    16558.832   Durbin-Watson:                   0.560
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         21549933.401
Skew:                           8.534   Prob(JB):                         0.00
Kurtosis:                     217.222   Cond. No.                         6.28
==============================================================================
```

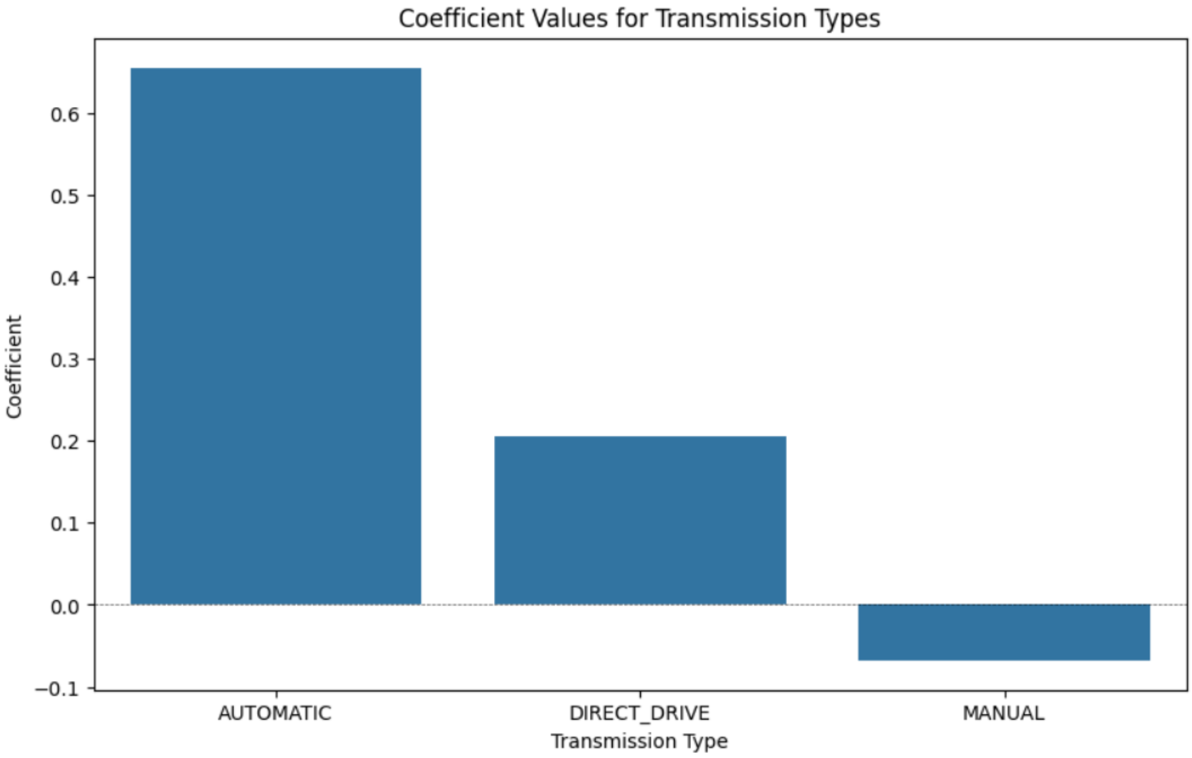
Low R-squared value indicates that the model explains only a small portion of the variance in highway KPL. This suggests that other factors, beyond vehicle size, likely play a significant role in determining fuel efficiency on the highway. However, according to our Coefficient and P-value, we can make sure in our linear models results with p-value being 0.000, indicating that the negative relationship between Fuel Efficiency and Vehicle Size is significant.

6.4 Vehicle-Transmission Type



```
                            OLS Regression Results
===================================================================================
Dep. Variable:     Vehicle Size (Mapped)   R-squared:                         0.169
Model:                               OLS   Adj. R-squared:                    0.169
Method:                    Least Squares   F-statistic:                       760.5
Date:                   Wed, 24 Apr 2024   Prob (F-statistic):                 0.00
Time:                           21:01:30   Log-Likelihood:                  -12059.
No. Observations:                  11187   AIC:                           2.413e+04
Df Residuals:                      11183   BIC:                           2.416e+04
Df Model:                              3
Covariance Type:               nonrobust
==================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            1.3978      0.030     46.219      0.000       1.339       1.457
AUTOMATIC        0.6547      0.031     20.932      0.000       0.593       0.716
DIRECT_DRIVE     0.2051      0.091      2.244      0.025       0.026       0.384
MANUAL          -0.0675      0.033     -2.030      0.042      -0.133      -0.002
==============================================================================
Omnibus:                      771.757   Durbin-Watson:                   0.287
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              354.428
Skew:                           0.242   Prob(JB):                     1.09e-77
Kurtosis:                       2.274   Cond. No.                         17.6
==============================================================================
```

R-squared value of 0.169 indicates that about 16.9% of the variability in ”Vehicle Size (Mapped)” can be explained by the independent variables (”Transmission Type”). This isn’t a very high proportion, but as we are only considering the association between the size of vehicle and its transmission type, one can make sure that: ”AUTOMATIC” is positively associated with larger vehicle sizes - automatic transmissions are more commonly found in larger vehicles. (The coefficient is 0.6547, p-value 0.000) ”DIRECT_DRIVE” has a positive but smaller association with vehicle size - vehicles with direct drive transmissions are more likely to be midsize to large in size. (The coefficient is 0.2051, p-value 0.025) ”MANUAL” has a slight negative association with vehicle size - vehicles with manual transmissions are more likely to be smaller in size. (The coefficient is -0.0675, p-value 0.042) This aligns with the notion that certain transmission types are more likely in larger vehicles, while others (like manual) are more common in smaller ones.



7 Conclusions

In conclusion, our interim analysis lays the foundation for deeper exploration into the factors influencing car prices. We identified hypotheses to test, formulated questions that are interesting for us and conducted preliminary data analysis to understand the relationships between variables and car prices. Several conclusions based on first insights and data analysis were made:

1. Understanding that 85% of car prices fall below $35,000 highlights the affordability range for most consumers. To buy a new good car you must have up to 35, 000$

2. With most cars having horsepower between 100 and 400 consumers can gauge performance expectations. Similarly knowing the typical city and highway KPLs ranges (10-20 and 10-40 respectively) helps in assessing fuel efficiency.

3. For the business side there are several notes on how to earn more money and attract major customers. To capture a broader market share and prioritize features and technologies that align with consumer preferences and market demands businesses can produce more automatic transmission type cars because there is a tendency that using such cars is increasing. Still need to produce a different number of machines of different sizes with different driven wheels to capture market share. To keep a balance between horsepower, price and customers preferences (the more horsepower your car has - the more cool you look like :)) it is common to put 6 or 8 cylinders for the engine in order not to overboard in any one feature.

Also there is a global tendency that production of cars are only growing, which is clearly visible from plots in recent years. That implies more people are buying cars, more congested roads in cities are and more exhaust gases from factories and machines are released into the atmosphere which already has certain negative consequences. So for business it may have sense to be the first to come up with a new type of engine or update the current one in an interesting way to capture a new market share and attract new customers.

Making conclusions from regression models we can say that taking into account all cars in dataset (including premium and luxury) there is a small change in price over age because more luxury cars - more years it will be fashionable and less deduction in price will have. Most of the coefficients are intuitive but some are interesting like the positive coefficient for compact size cars in full model (with all dataset) says that sports cars are included in the dataset and they are very expensive compared to midsize default cars.

Creating model without cars with price higher than 75% percentile gives interesting results that popularity is not significant because popularity is number how many times car was measured in Twitter. Hence, more luxury and expensive cars are more popular in different videos. And for people it does not make sense to buy a car on popularity measured in such way. If people want more comfortable car with automatic transmission they have pay extra only 1500$ than for manual transmission car. Overall, the second model with excluded percentile is more appropriate for common people when it comes to buying a car because it removes premium and sport cars. It has logically correct signs of coefficients and can be used in measuring car comparing to based one which was described (4 doors, regular unleaded, midsize, front wheel).

If you are buying an automatic and electric car be ready to have a larger vehicle rather than in manual transmission where it tends to be smaller sizes. With larger vehicle size fuel efficiency decreases and it is obviously but important to know in order to service larger machines is more costly. Next steps for our models can be their improvement with regularization and considering other models for predicting prices - regressor trees, SVM machines - which are out of our scope of our course.

8 Dataset

Columns description:

1. Firm: This column represents the make or manufacturer of the vehicle.

2. Model: This column represents the model name of the vehicle. For example, '1 Series M' or '1 Series' in the provided data.

3. Year: This column represents the manufacturing year of the vehicle.

4. Engine Fuel Type: This column indicates the type of fuel the vehicle's engine requires.

5. Engine HP (Horsepower): This column represents the engine horsepower, a measure of the power of the vehicle's engine.

6. Engine Cylinders: This column represents the number of cylinders in the vehicle's engine.

7. Transmission Type: This column indicates the type of transmission the vehicle has, such as 'MANUAL'.

8. Driven Wheels: This column indicates the type of wheels the vehicle is driven by, such as 'rear wheel drive'.

9. Number of Doors: This column represents the number of doors on the vehicle.

10. Vehicle Size: This column represents the size category of the vehicle, such as 'Compact' in the provided data.

11. Vehicle Style: This column represents the style or body type of the vehicle, such as 'Coupe' or 'Convertible'.

12. highway MPG: This column represents the vehicle's fuel efficiency in miles per gallon (MPG) on the highway.

13. city mpg: This column represents the vehicle's fuel efficiency in miles per gallon (MPG) in the city.

14. Popularity: number of times the car was mentioned in a Twitter stream

15. MSRP (PRICE): This column represents the manufacturer's suggested retail price (MSRP) of the vehicle. It indicates the price at which the manufacturer suggests that the vehicle be sold.

A Useful auxiliary facts

[0] GitHub:
https://github.com/Baredal/car_price_analysis

[1] CCar Features and MSRP Dataset:
https://www.kaggle.com/datasets/CooperUnion/cardataset/data

[3] ScienceDirect Article:
https://www.sciencedirect.com/science/article/abs/pii/S1361920923001359

[4] SagePub Article:
https://journals.sagepub.com/doi/10.1177/21582440221120647?icid=int.sj-full-text.similar-articles.4

[5] SlideShare Presentation:
https://www.slideshare.net/NAVINCHACKO1/car-price-predictionpptx

[6] AuctionExport Website:
https://www.auctionexport.com/uk

[7] J.D. Power Website:
https://www.jdpower.com/

[8] MacroTrends Website:
https://www.macrotrends.net/global-metrics/countries/USA/united-states/inflation-rate-cpi

[9] Statista Website:
https://www.statista.com/statistics/256598/global-inflation-rate-compared-to-previous-year/