College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 2 Midweek Project

Submitted on:

Jan 25, 2024

Submitted to:	**Submitted by:**
Prof: BEHZAD AHMADI	NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The Carprice dataset is a collection of data on various cars, with a range of attributes that describe their specifications and features. It consists of 205 entries, each representing a unique car, and is structured into 26 columns

Column Details:

Numerical Columns:

- car_ID: A unique identifier for each car.
- symboling: An insurance risk rating.
- wheelbase, carlength, carwidth, carheight: Dimensions of the car in various aspects.
- curbweight: The weight of the car without passengers or cargo.
- enginesize: Size of the car's engine.
- boreratio, stroke, compressionratio: Engine specifications.
- horsepower, peakrpm: Engine power and performance metrics.
- citympg, highwaympg: Fuel efficiency in the city and on the highway, respectively.
- price: The price of the car.

Categorical Columns:

- CarName: The name of the car.
- fueltype: The type of fuel the car uses (gas or diesel).
- aspiration: The type of aspiration (standard or turbo).
- doornumber: The number of doors in the car.
- carbody: The body style of the car (e.g., sedan, hatchback etc).
- drivewheel: The type of drivewheel (e.g., front-wheel drive).
- enginelocation: Location of the engine (front or rear).
- enginetype: The type of engine (e.g., ohc, ohcf etc).
- cylindernumber: The number of cylinders in the engine.
- fuelsystem: The fuel system of the car (e.g., mpfi, 2bbl etc).

This dataset is a great tool for conducting market analysis, and price prediction modeling in the automotive sector.

Analysts can leverage it to gain insights into how different factors like brand reputation, technical specifications, and market trends influence car prices. It can also be used to identify patterns and correlations within the automotive market, providing valuable information for strategic decision-making in business.

DATA CLEANING

- **Checking for the number of missing values in the dataset**
  The variables do not seem to have any null or NA values.

- **Eliminating columns**
  Column “Car_ID” was removed because this column does not contribute to the analysis or the insights.

- **Adding a column “CompanyName”**
  The 'CompanyName' was extracted from the 'CarName' column by using the first word, categorizing cars by their manufacturer. Then, the 'CarName' column was dropped to make the data simpler.

- **Replacing data values**
  Misspellings in the 'CompanyName' column were corrected to ensure consistency and accuracy in the dataset

- **Removing duplicate rows**
  Duplicate rows in the dataset can lead to inconsistencies which may affect the accuracy of the analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.
  There are no duplicate rows.

DATA EXPLORATION AND EDA

Observations from the EDA:

Figure 1:

Here's an overview of the histograms for different variables -

- **Price:** The price histogram is right-skewed, showing that most cars in the dataset are in the lower price bracket, with luxury or high-end cars being less frequent.
- **Engine Size:** There is a right skew in engine size, implying that smaller engines are more common than larger ones within this dataset.
- **Horsepower:** Similarly, the horsepower histogram is also right-skewed, showing that cars with lower horsepower are more prevalent, which could correlate with the larger number of smaller-engine cars.

Figure 2:

Engine size, curb weight, and horsepower are identified as three key factors that strongly influence car prices based on the correlation analysis. Understanding the importance of these attributes can help car manufacturers and dealers make data driven decisions about product development, marketing efforts and pricing strategies.

Figure 3:

The bubble chart visually confirms that there is a positive relationship between car price and both horsepower and engine size. As horsepower and engine size increase, car prices tend to rise.

The chart can also help in identifying different segments within the market. For instance, there may be a cluster of small bubbles with low horsepower and small engine sizes, indicating a segment of affordable, compact cars. On the other hand, larger bubbles with high horsepower and engine sizes represent premium and high-performance cars.

Figure 4:

When visualizing the average prices of cars for each company, it becomes evident that Jaguar has the highest average price, while Nissan has the lowest average price among the companies.

LINEAR REGRESSION MODEL

We started with encoding the categorical values present in the data. Encoding is a crucial data preprocessing step in machine learning and statistical analysis, particularly when working with categorical variables. Without encoding, models may not consider categorical data, leading to errors. This process also ensures consistency in representing categories and plays a vital role in improving model performance by allowing models to learn from categorical data.
There are various encoding methods and I have used below :

- **Manual Label Mapping:**
  In this step, numerical values were assigned to categorical variables with a limited number of unique categories. For example, we mapped 'doornumber' categories 'two' and 'four' to 2 and 4, respectively.

- **Label Encoding using sci-kit-learn:**
  Label encoding is particularly useful when dealing with categorical variables that have ordinal relationships, meaning there is a natural order or hierarchy among the categories.

Then, the dataset (Figure 5 – after encoding) is divided into two parts: one containing the independent variables and the other containing the target variable.
The dataset is further divided into training and testing sets – (80% train set and 20% test set).

The feature data was standardized to ensure that all features are on a similar scale and help to prevent features with large scales from dominating the model.

A linear regression model is constructed. This model aims to learn the associations between the car variables and their prices. It does so by finding the best-fit line that minimizes the differences between predicted and actual prices.

To evaluate the model's accuracy, metrics such as Mean Squared Error (MSE) and R-squared (R2) are used. MSE quantifies the average squared differences between predicted and actual prices, while R2 measures how well the model explains the variance in car prices.

CONCLUSION

The three most significant variables based on their coefficients in the linear regression model **(Figure 6)** are:

- Engine Size
- Curb Weight
- Horsepower

Of these three significant variables, Engine Size (engine size) had the greatest positive influence on car prices. This means that as the engine size increases, car prices tend to increase as well.

The linear regression model used to predict car prices based on various attributes performs reasonably well **(Figure 7)**. The MSE value of approximately 11,775,815.14 suggests that the model's predictions are relatively close to the actual car prices on average. Additionally, the R^2 value of 0.851 indicates that the model can explain approximately 85.08% of the variance in car prices, which signifies a good level of predictability.

REFERENCES

Avcontentteam. (2023, December 19). *How to perform label encoding in Python?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/07/label-encoding-in-python/

Pandas to replace value using map function. (n.d.). Stack Overflow. https://stackoverflow.com/questions/57873297/pandas-to-replace-value-using-map-function

APPENDIX

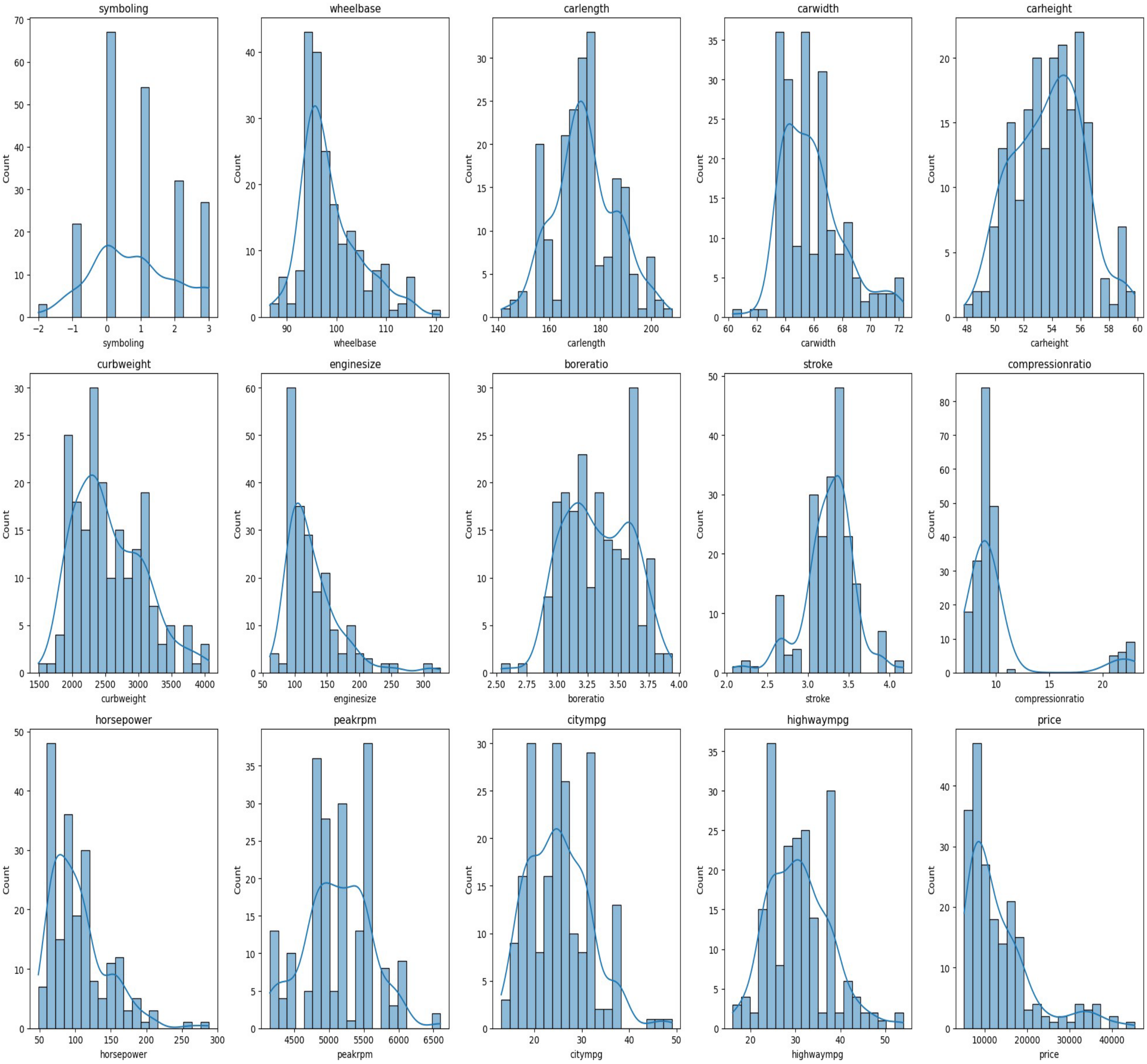
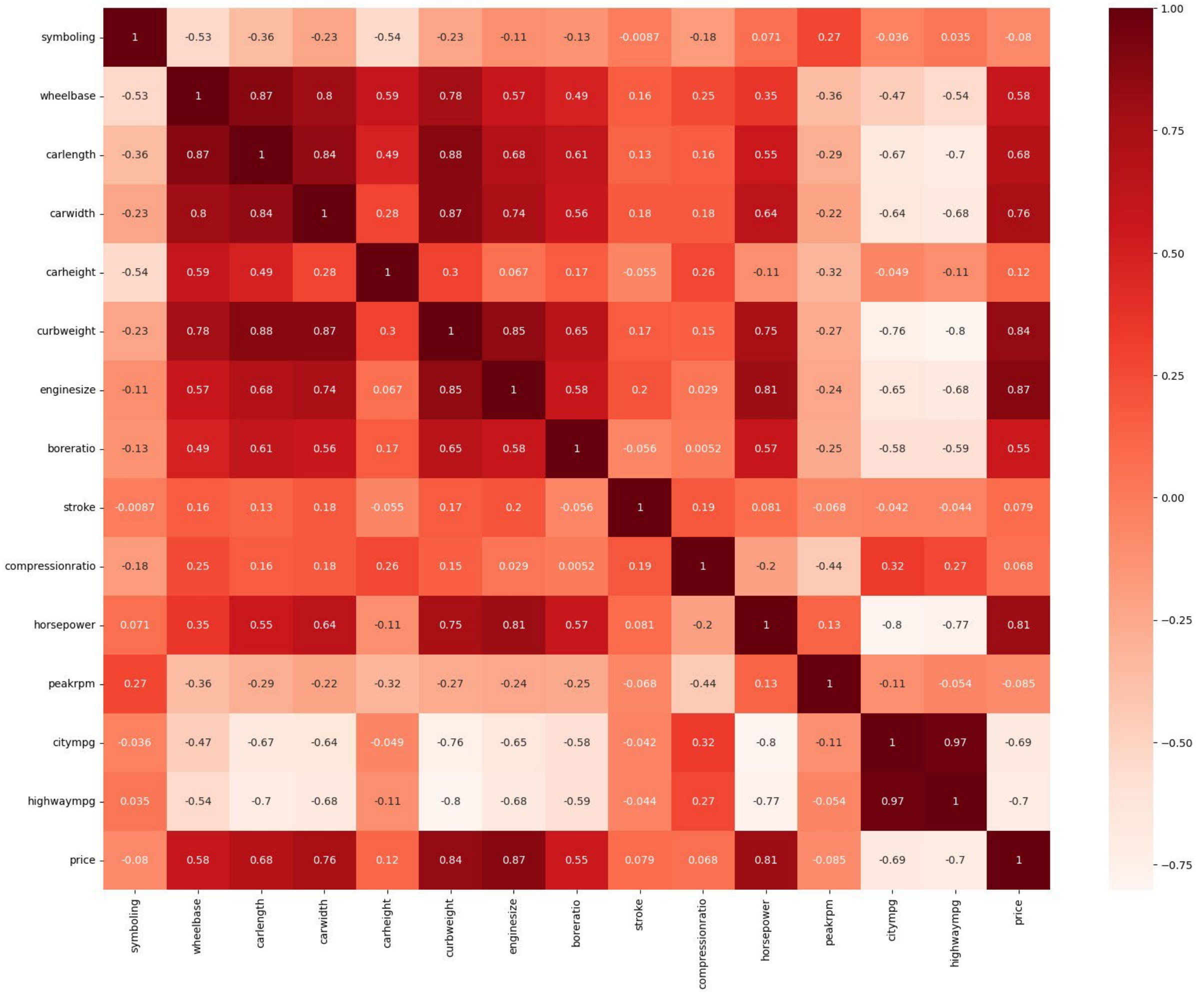


Figure 1 – Histograms for understanding distribution



	symboling	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
symboling	1	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.13	-0.0087	-0.18	0.071	0.27	-0.036	0.035	-0.08
wheelbase	-0.53	1	0.87	0.8	0.59	0.78	0.57	0.49	0.16	0.25	0.35	-0.36	-0.47	-0.54	0.58
carlength	-0.36	0.87	1	0.84	0.49	0.88	0.68	0.61	0.13	0.16	0.55	-0.29	-0.67	-0.7	0.68
carwidth	-0.23	0.8	0.84	1	0.28	0.87	0.74	0.56	0.18	0.18	0.64	-0.22	-0.64	-0.68	0.76
carheight	-0.54	0.59	0.49	0.28	1	0.3	0.067	0.17	-0.055	0.26	-0.11	-0.32	-0.049	-0.11	0.12
curbweight	-0.23	0.78	0.88	0.87	0.3	1	0.85	0.65	0.17	0.15	0.75	-0.27	-0.76	-0.8	0.84
enginesize	-0.11	0.57	0.68	0.74	0.067	0.85	1	0.58	0.2	0.029	0.81	-0.24	-0.65	-0.68	0.87
boreratio	-0.13	0.49	0.61	0.56	0.17	0.65	0.58	1	-0.056	0.0052	0.57	-0.25	-0.58	-0.59	0.55
stroke	-0.0087	0.16	0.13	0.18	-0.055	0.17	0.2	-0.056	1	0.19	0.081	-0.068	-0.042	-0.044	0.079
compressionratio	-0.18	0.25	0.16	0.18	0.26	0.15	0.029	0.0052	0.19	1	-0.2	-0.44	0.32	0.27	0.068
horsepower	0.071	0.35	0.55	0.64	-0.11	0.75	0.81	0.57	0.081	-0.2	1	0.13	-0.8	-0.77	0.81
peakrpm	0.27	-0.36	-0.29	-0.22	-0.32	-0.27	-0.24	-0.25	-0.068	-0.44	0.13	1	-0.11	-0.054	-0.085
citympg	-0.036	-0.47	-0.67	-0.64	-0.049	-0.76	-0.65	-0.58	-0.042	0.32	-0.8	-0.11	1	0.97	-0.69
highwaympg	0.035	-0.54	-0.7	-0.68	-0.11	-0.8	-0.68	-0.59	-0.044	0.27	-0.77	-0.054	0.97	1	-0.7
price	-0.08	0.58	0.68	0.76	0.12	0.84	0.87	0.55	0.079	0.068	0.81	-0.085	-0.69	-0.7	1

Figure 2 – Correlation matrix

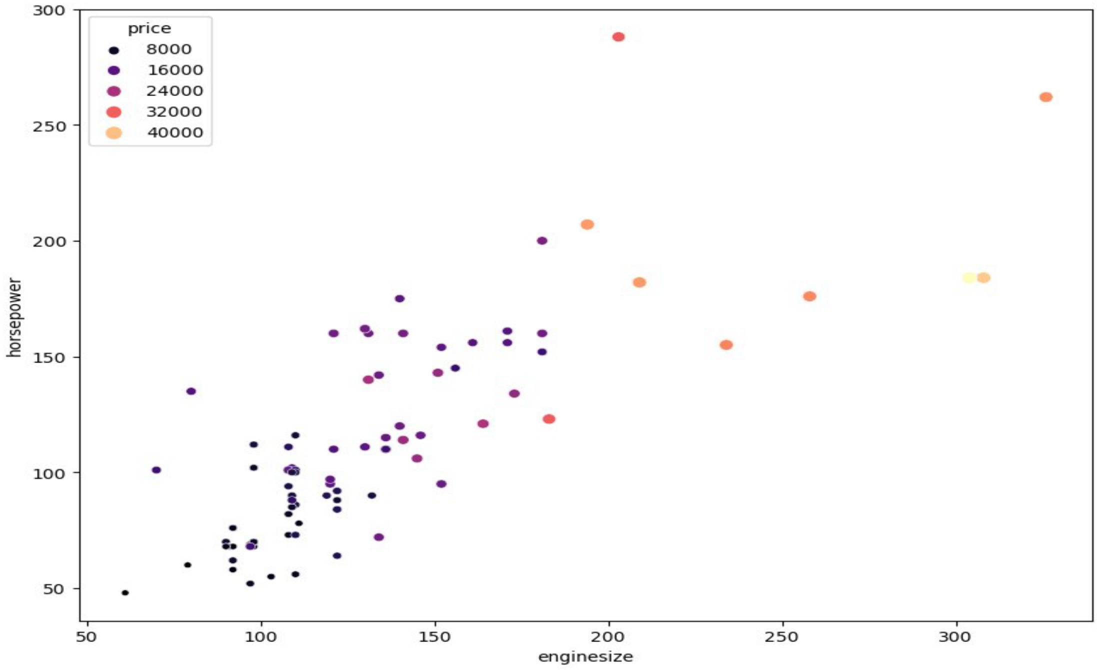


Figure 3 – HP vs Engine size (with prices)

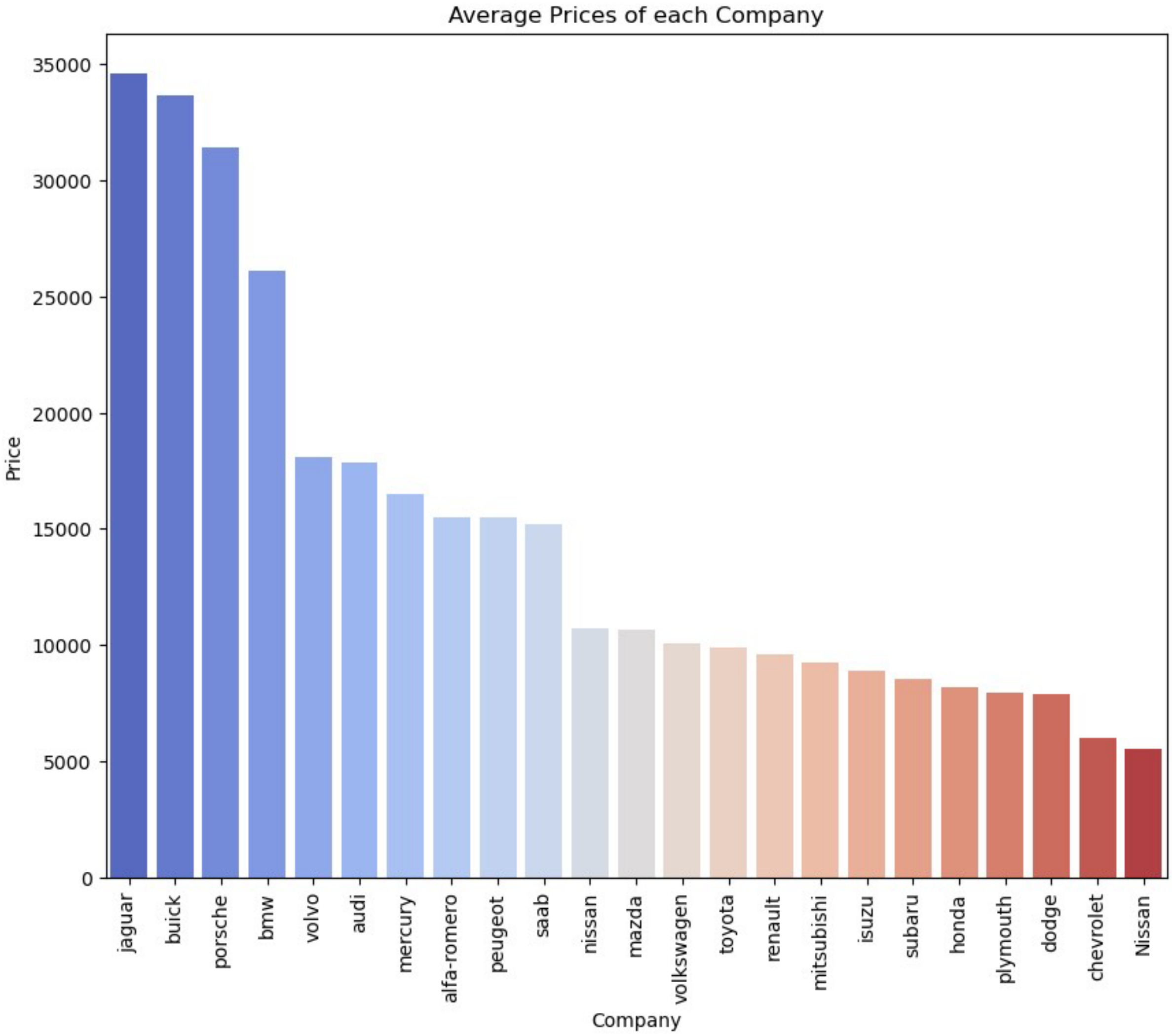
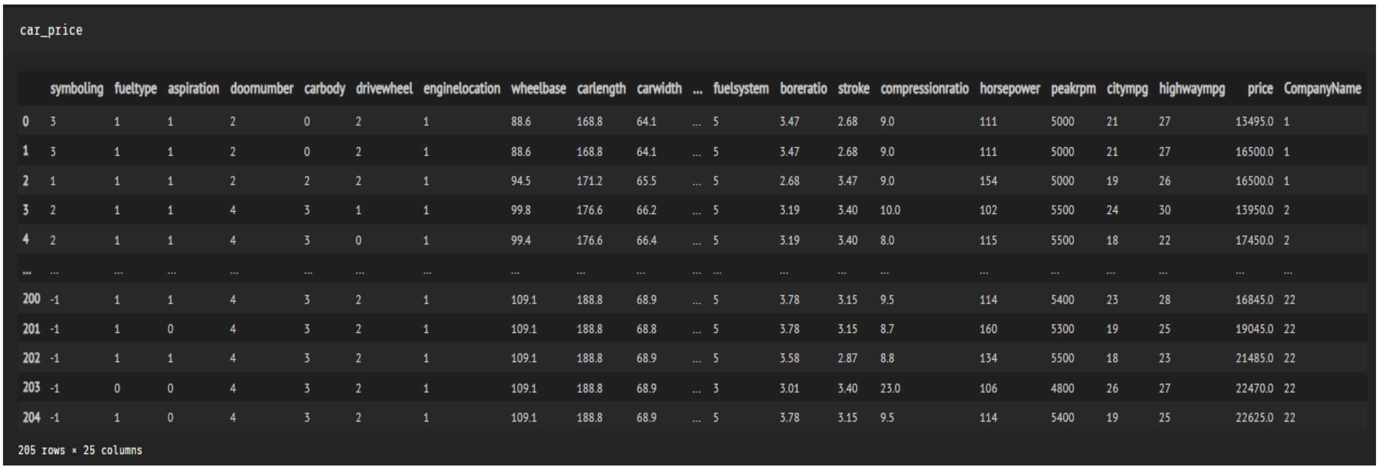


Figure 4 – Barplot (average car prices for different companies)

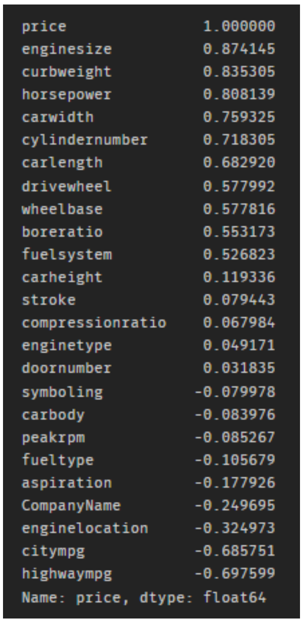


car_price

	symboling	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	...	fuelsystem	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price	CompanyName
0	3	1	1	2	0	2	1	88.6	168.8	64.1	...	5	3.47	2.68	9.0	111	5000	21	27	13495.0	1
1	3	1	1	2	0	2	1	88.6	168.8	64.1	...	5	3.47	2.68	9.0	111	5000	21	27	16500.0	1
2	1	1	1	2	2	2	1	94.5	171.2	65.5	...	5	2.68	3.47	9.0	154	5000	19	26	16500.0	1
3	2	1	1	4	3	1	1	99.8	176.6	66.2	...	5	3.19	3.40	10.0	102	5500	24	30	13950.0	2
4	2	1	1	4	3	0	1	99.4	176.6	66.4	...	5	3.19	3.40	8.0	115	5500	18	22	17450.0	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
200	-1	1	1	4	3	2	1	109.1	188.8	68.9	...	5	3.78	3.15	9.5	114	5400	23	28	16845.0	22
201	-1	1	0	4	3	2	1	109.1	188.8	68.8	...	5	3.78	3.15	8.7	160	5300	19	25	19045.0	22
202	-1	1	1	4	3	2	1	109.1	188.8	68.9	...	5	3.58	2.87	8.8	134	5500	18	23	21485.0	22
203	-1	0	0	4	3	2	1	109.1	188.8	68.9	...	3	3.01	3.40	23.0	106	4800	26	27	22470.0	22
204	-1	1	0	4	3	2	1	109.1	188.8	68.9	...	5	3.78	3.15	9.5	114	5400	19	25	22625.0	22

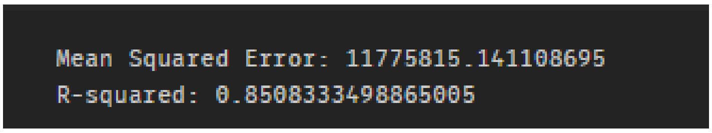
205 rows × 25 columns

Figure 5 – Data after encoding



```
price               1.000000
enginesize          0.874145
curbweight          0.835305
horsepower          0.808139
carwidth            0.759325
cylindernumber      0.718305
carlength           0.682920
drivewheel          0.577992
wheelbase           0.577816
boreratio           0.553173
fuelsystem          0.526823
carheight           0.119336
stroke              0.079443
compressionratio    0.067984
enginetype          0.049171
doornumber          0.031835
symboling          -0.079978
carbody            -0.083976
peakrpm            -0.085267
fueltype           -0.105679
aspiration         -0.177926
CompanyName        -0.249695
enginelocation     -0.324973
citympg            -0.685751
highwaympg         -0.697599
Name: price, dtype: float64
```

Figure 6 – Most correlated features with price



```
Mean Squared Error: 11775815.141108695
R-squared: 0.8508333498865005
```

Figure 7 – Linear regression model results