# **WHAT CAR FEATURES ARE THE MAJOR DETERMINANT OF CAR PRICE VALUES**

# **INTRODUCTION**

# It goes without saying that a products appearance/aesthetic and functionality would play a major role in the successful sale and price of the product, hence in this project, a critical analysis have been carried out on a dataset that gives information on several car features alongside their price value (“Automobile Raw Data Prediction EDA & Modeling dataset) to determine the exact car feature that affects car prices and how much these features impact the price values.

# **METHODOLOGY**

In this project a critical analysis was carried out to determine the car features that affects car prices and how much these features impact the car price values. The first step taken in conducting this analysis as with all other forms of data programming task was the Data Acquistion Step/Data Discovery I.e., determining and thereafter obtaining the needed data necessary to successfully carry out the analysis that would answer our formulated question (Prażuch W., 2021). This dataset was obtained from Kaggle (<https://www.kaggle.com/razamh/automobile-raw-data-prediction-eda-modling?select=Update+Raw+Data+Prediction.csv>), the dataset contains car features and their respective price values, thus perfect for our analysis.

After the dataset was obtained, the second stage of data programming task; data cleaning and pre-processing was started (Prazuch W., 2021). This stage entails cleaning the dataset and putting it in good shape for the analysis, it is the step where errors and inconsistencies are removed from datasets to improve the quality of such data and thus forms a major part of data analysis ETL (Extract Transform Load) process (Rahm E., Hai Do H., 2000). This step which is also known as data wrangling basically entails handling missing values, removing outliers, data encoding e.t.c, however for the next step which is Exploratory Data Analysis, only the missing values were handled at this stage while the other forms of data cleaning was done before the modelling stage.

Exploratory, statistical and computational data analysis was thereafter carried out on the datasets to survey and determine useful insights that can help answer our question. This stage often involves use of visual aids (charts) which was generated using matplotlib (Hunter J. D., (2007) and seaborn packages of the python programming language (Van Rossum, G. & Drake, F.L., 2009).

Finally, we developed a model to predict car prices based on car features after performing further data cleaning tasks such as dropping some features (car\_comapany, car\_id), extracting an extra feature (company name) from the car company feature and converting all the categorical variables (fuel\_type, aspiration, door\_number, car\_body, drive\_wheel, engine\_location, wheel\_base, engine\_type, cylinder\_number, fuel\_system) in the dataset to numerical form using label encoding. The machine learning models used for the predictive analysis were built with the scikit learn package (Pedregosa *et al,* 2011) of the python programming language. The algorithms used were Multiple Linear regression and Random Forest Regression with all hyperparameters set to default values.

# **RESULTS**

In this section, we present the result of our findings after performing our Analysis.

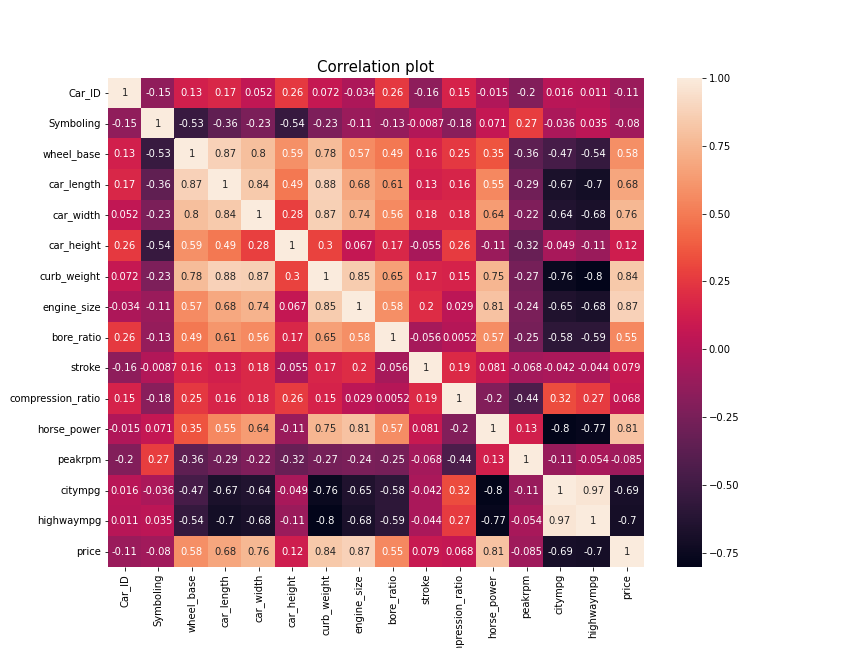


Fig 1: The correlation plot above shows the relationship that exists between the features, zooming into the price column which is of most importance to us, we discover that features such as curb\_weight, engine size and horse\_power strongly correlates with the price feature while features such as wheel base, car\_length, car width, bore ration, city/highway mpg mildly correlates with the price feature and other features bears little to no correlation at all with the price feature.

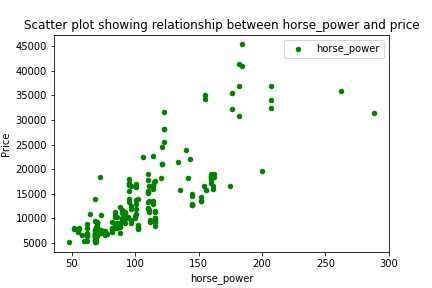
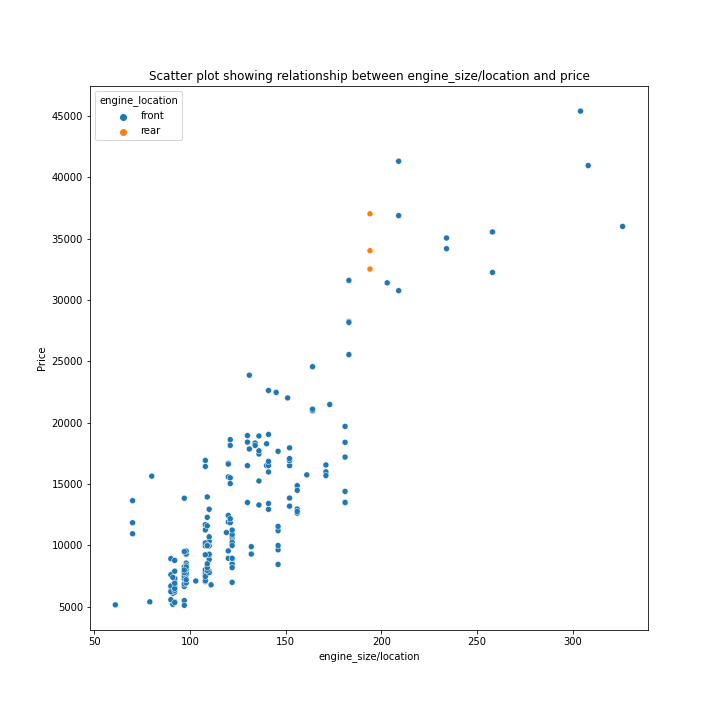


Fig 2 Fig 3

Fig 2 and 3 above are scatter plots that further shows the level of correlation that exists between the engine size (fig 2), horsepower (fig 3) and the car price values. The chart shows a positive (upward) trend, meaning that the car price value increases with engine size and horsepower. Fig 2 also shows that few vehicles have their engine at the rear position and they are all expensive.

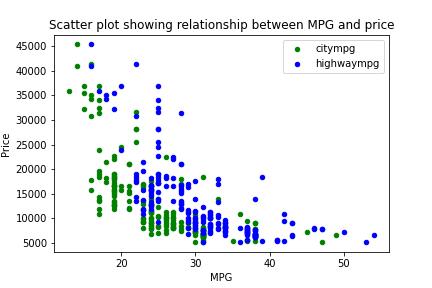
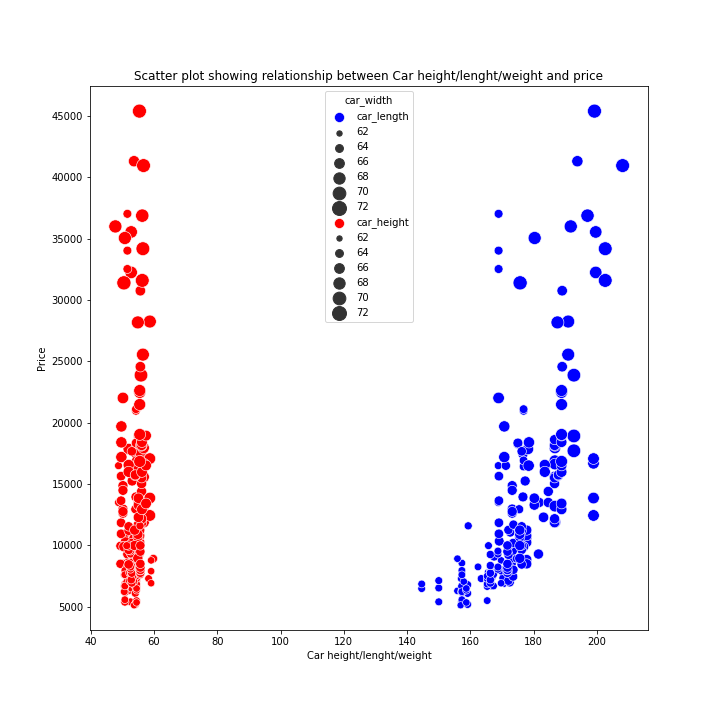


Fig 4 Fig 5

Fig 4 and 5 are also scatter plots that gives further information on the correlation between the features; high/city mpg, car height/ length/ width and the price features. Fig 4 shows that car length and width are slightly positively correlated with the price values I.e., higher length/width values will increase price values while car height has little to no correlation with the price values. Fig 5 shows that the mpg’s carries a negative correlation with the price features meaning that higher mpg’s implies lower price value and lower mpg’s implies higher price value.

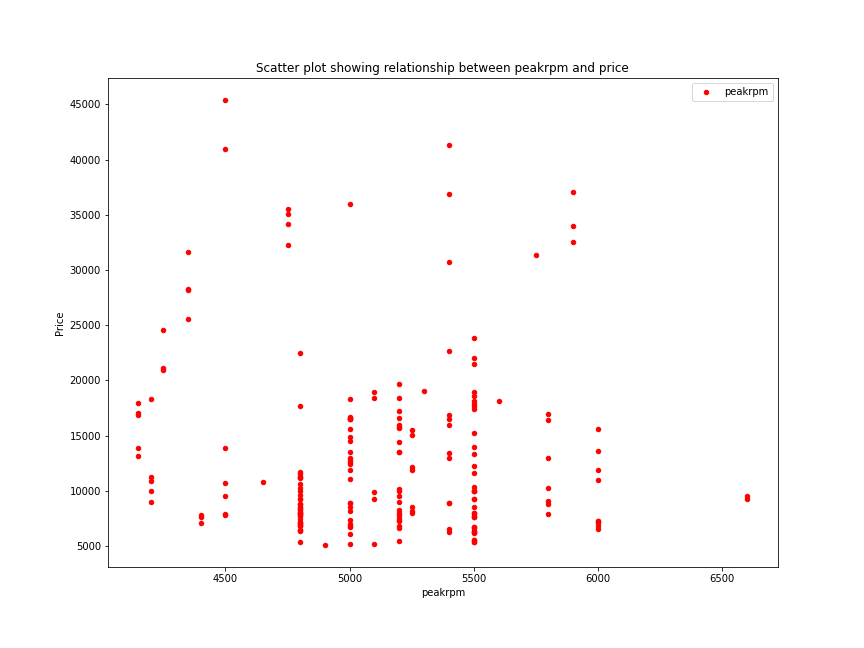
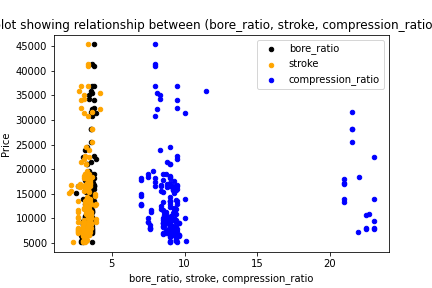
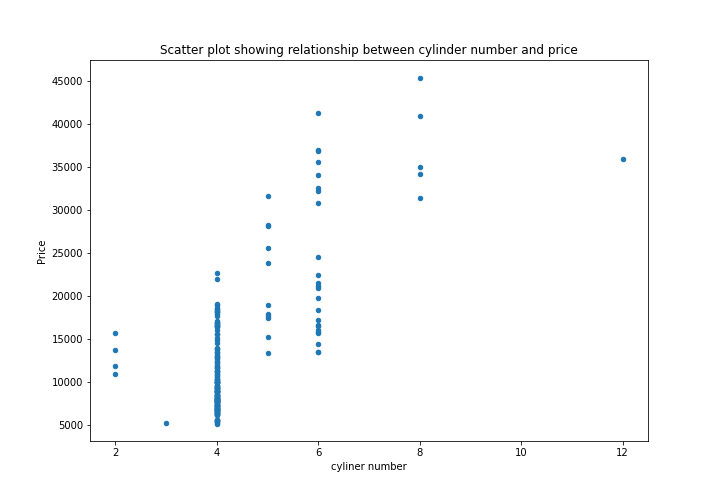
  

Fig 6 Fig 7 Fig 8

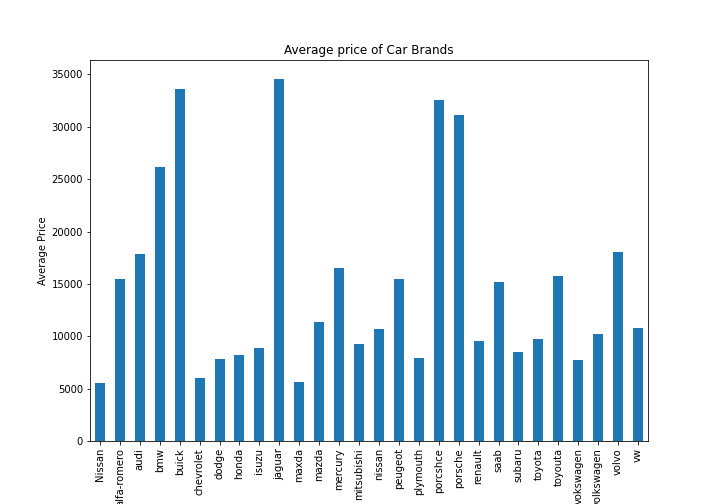
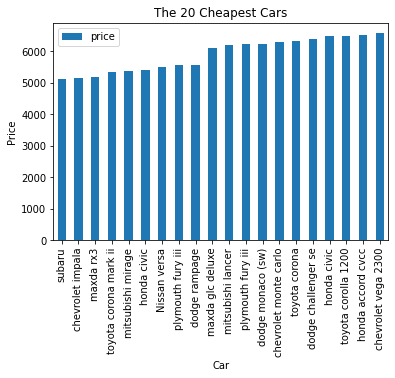
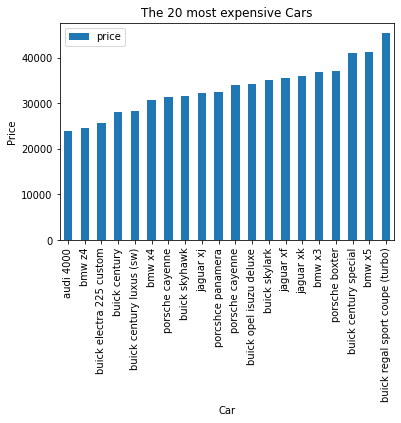
Fig 6, 7 and 8 above are all scatter plots revealing the state of randomness in correlation between [peakrpm (fig 6) Bore\_ratio, stroke, compression\_ratio (fig 7), cylinder number (fig 8)] and price feature, implying that this car features bears no impact on car price values. Fig 8 however reveals that although cylinder number carries little correlation with price value, higher cylinder number have better chances to positively impact car price values than lower cylinder number. 

Fig 9 is a bar chart showing the average price of all the cars based on the brands, The chart reveals that jaguar sells the most expensive cars on the average followed by buick, porshce, renault, bmw, audi in that order while Maxda, Nisaan, plymouth, chevrolet sells the cheapest car on average. The bar charts in Fig 10 and 11 below further reveals this fact as the top 20 expensive cars are either jaguar,audi, buick, porshce, bmw and many of the top 20 cheapest cars are chevrolet.

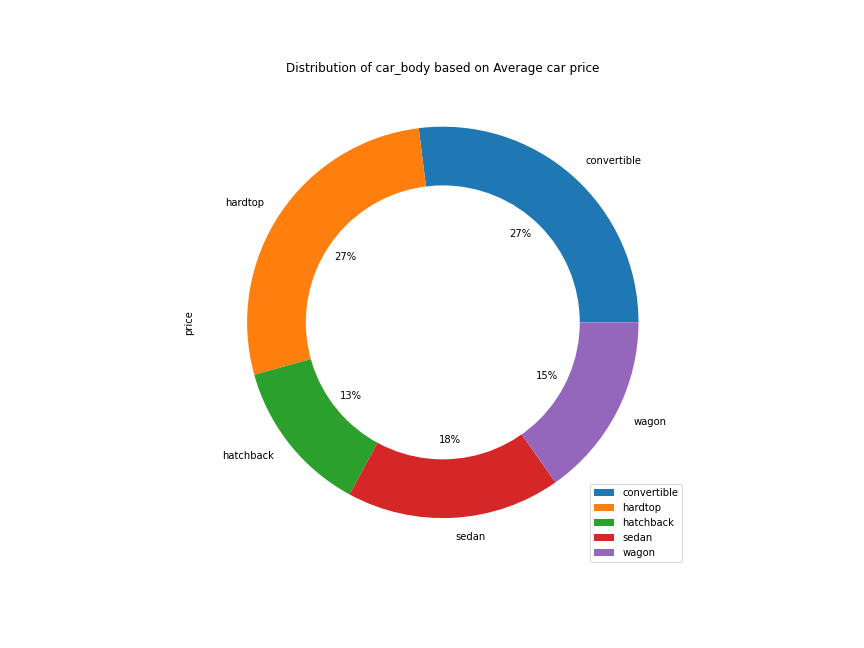
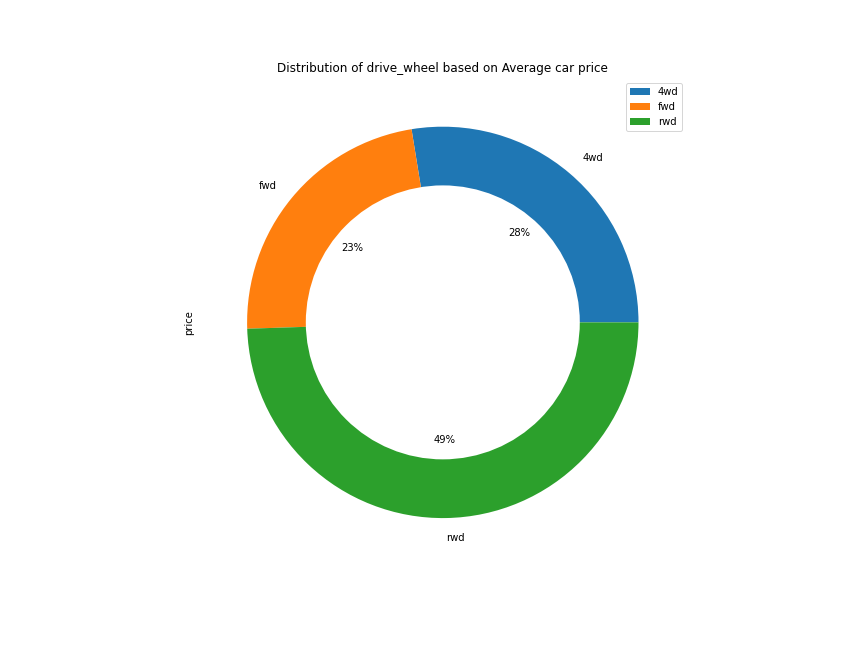
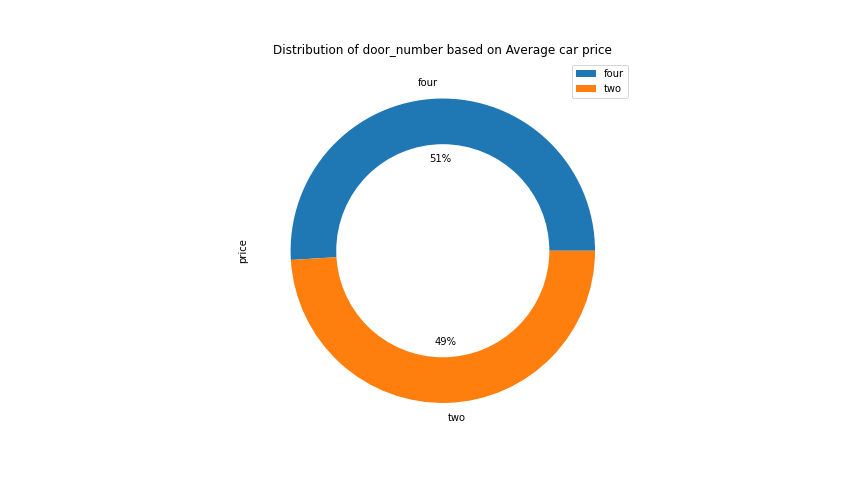
 

Fig 12 Fig 13 Fig 14

Fig 12, 13 and 14 are donut plots that shows the proportionate contribution of certain vehicle features (drive wheel, car body, door number) to the total car price. Fig 14 shows that vehicles with 4 doors contribute more to price values than that of two, fig 12 shows that vehicles with rewind drive wheel contributed more to total price value and fig 13 shows that vehicles with hardtop and convertible car bodies both contribute the highest to total price value. The grouped bar chart in fig 15 below shows

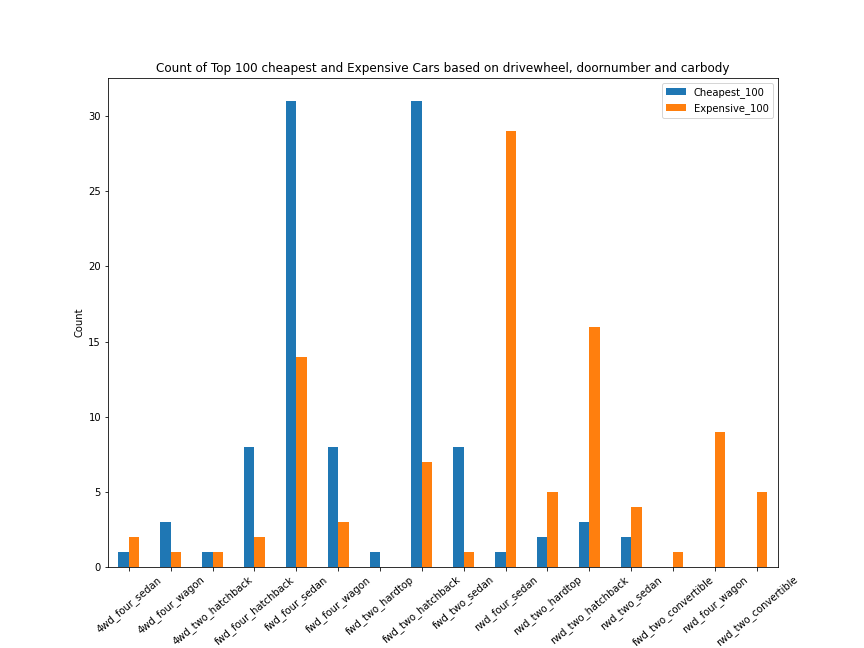
further insight into this

Fig 15: shows that many of the top 100 expensive cars are vehicles with either (rewind drive wheel, four doors and sedan car body), (rewind drive wheel, two doors and hatchback car body) or (forward drive wheel, four doors and sedan car body) while the cheapest are mostly (forward drive wheel, two doors and hatchback car body) or (forward drive wheel, four doors and sedan car body).

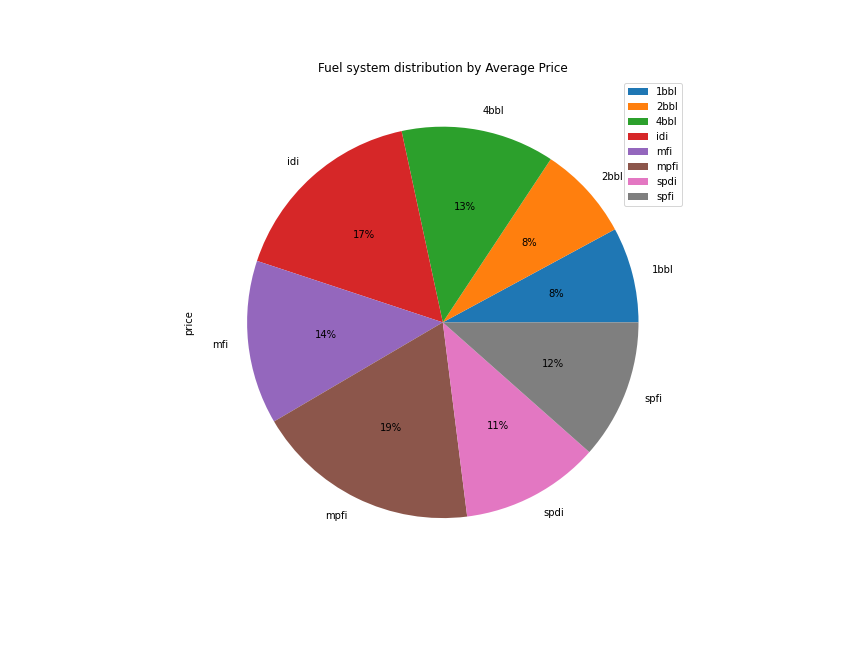
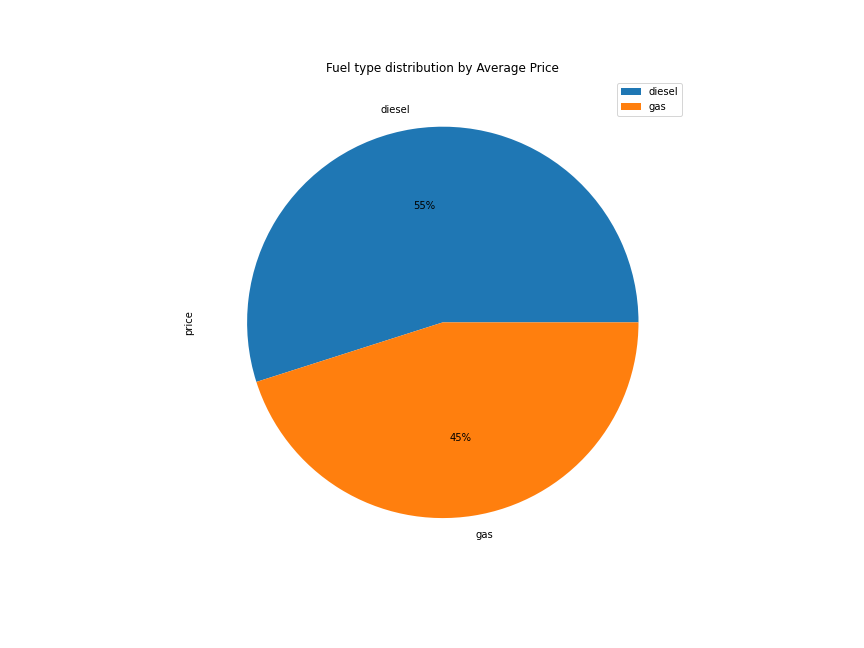


Fig 16 Fig 17

Fig 16 and 17 above are pie charts showing the proportionate contribution of fuel system and fuel type to the total car price. Fig 16 shows that vehicles with diesel fuel type contribute more to the total price, fig 18 below gives more insight.



Fig 18: shows that majority of the expensive 100 cars uses gas fuel type and mpfi fuel system while majority of the cheapest 100 cars use gas fuel type and 2bbl fuel system.

# **DISCUSSION**

Several features are known to determine price values of commodities, one of such features is the aesthetics/appearance which have been shown to play a major part in determining products price values. Audi reveals that as high as 60% of a consumer’s decision to purchase a vehicle is based on styling rather than technical performance (Ranscombe, C., Hicks, B., Mullineux, G., & Singh, B., 2012). Car brand have likewise been shown to affect consumers judgement during purchase (Page, C., & Herr, P. M. (2002). We thus set out on this project to determine what car features affects their price values and how much these features impact the price values.

The graph and table below show the summary of our predictive model

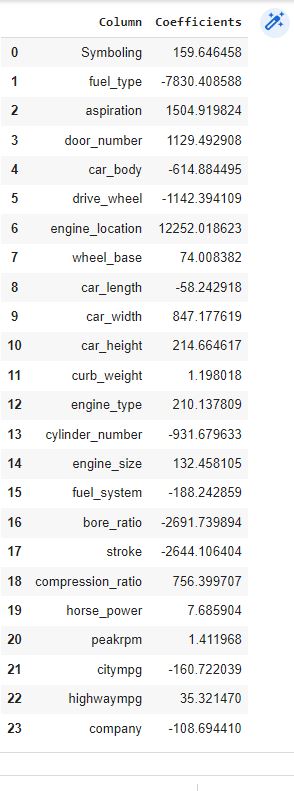
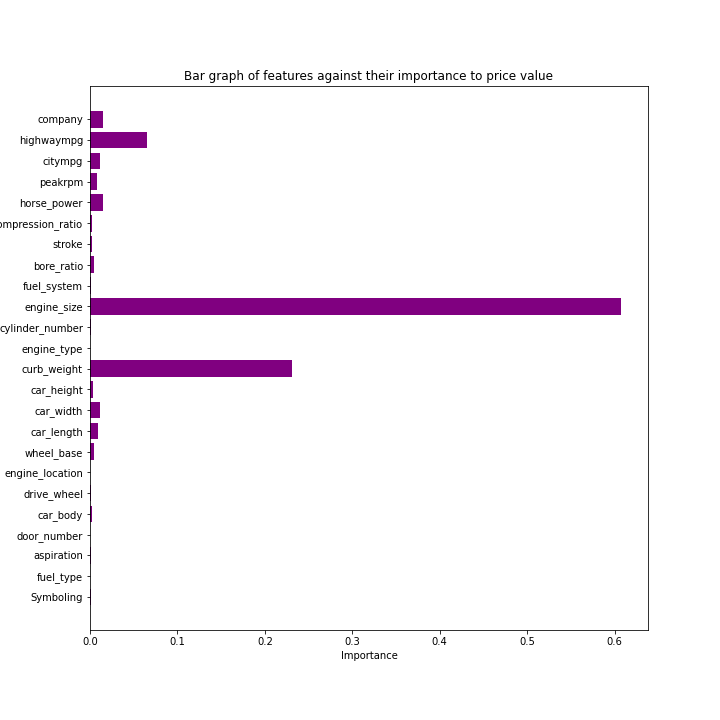


Fig 19: Feature Importance from Random Forest Table1: Coefficient Values from linear Regression

The Feature importance graph from our Random Forest algorithm which have an R2 score of 95% on the training set shows that engine size plays a very significant role in determining price values, followed by curb weight and mpg values while features like symboling, fuel\_type, aspiration, door number etc. have no effect. In the same vein, Table one shows the coefficient values of the independent variables from the linear regression model. It reveals engine location hugely impact price values positively, same as aspiration, door number, car width and compression ratio although not as high as that of engine location. The table also indicates that bore ratio, stroke car body, fuel type, cylinder number largely impacts price values although in the negative direction.

# **CONCLUSION**

In conclusion we have identified that all the following car features; 'car\_brand', 'fuel\_type', 'door\_number', 'car\_body', 'drive\_wheel', 'engine\_location', 'wheel\_base', 'car\_length', 'car\_width', 'car\_height', 'curb\_weight', 'engine\_type', 'cylinder\_number', 'engine\_size', 'fuel\_system', 'bore\_ratio', 'stroke', 'compression\_ratio', 'horse\_power', 'citympg', 'highwaympg' all contribute to the price values of automobiles. They are however not equally significant, the highly significant once are: 'engine\_size', 'engine\_location', 'horse\_power', 'citympg', 'highwaympg', ‘Car\_brand’, 'car\_width' and 'cylinder\_number'.

Lastly, a final reflection on this project brings me to the conclusion that doing these tasks have helped sharpen my data analytical skills, most especially, how to formulate an analytical research process around a given analytical or business problem.

**REFERENCES**

1. Prażuch W., (2021), 6 Phases of the Data Science Project Life Cycle
2. Rahm E., Hai Do H., (2000), Data Cleaning: Problems and Current Approaches
3. Hunter J. D., (2007) "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering.
4. Van Rossum, G. & Drake, F.L., 2009. *Python 3 Reference Manual*, Scotts Valley, CA: CreateSpace.
5. Pedregosa *et al,* 2011, [Scikit-learn: Machine Learning in Python.](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html)
6. Ranscombe, C., Hicks, B., Mullineux, G., & Singh, B. (2012). *Visually decomposing vehicle images: Exploring the influence of different aesthetic features on consumer perception of brand. Design Studies, 33(4), 319–341.*
7. Page, C., & Herr, P. M. (2002). An investigation of the processes by which product design and brand strength interact to determine initial affect and quality judgments