Sports Car Data Analysis

What makes a “sports car” a “*sports car”*? In this project, we utilized concepts learned throughout the course. Our goal was to conduct a thorough comparative analysis of various components/features within sports cars in order to determine their respective contributions to the overall price of these types of vehicles. We utilized two data sets found from Kaggle.com. One data set consisted of performance and functionality features while the other consisted of design and configuration features. We demonstrated data cleaning and preprocessing, feature engineering, exploratory and statistical analysis, creating visualizations, and model building for predictions in order to display various methods of data exploration that supports what makes a sports car valuable.

We decided to choose datasets that would focus on the contributions to the price and decision-making process in order to perform a comprehensive data analysis of the features. We hope to gain a holistic understanding of the interplay between function and design aspects. You can say that these models would help those who might be looking into buying a sports car or even a dealership that is trying to sell a sports car.

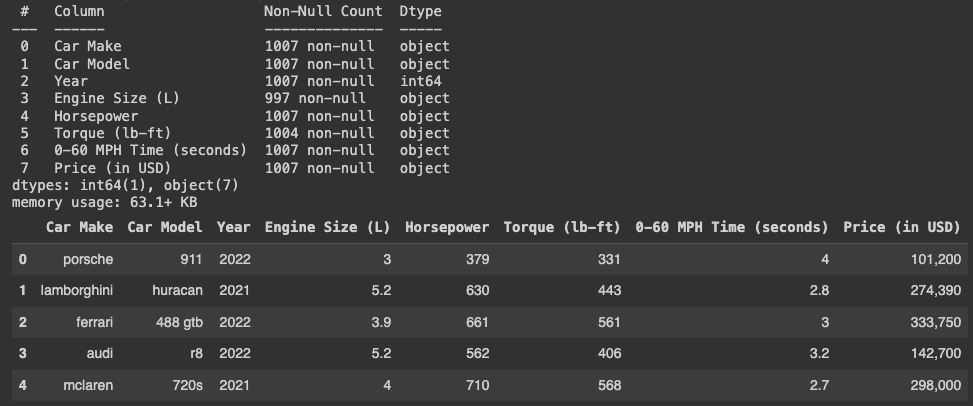


Figure 1: Dataset 1. sports-car-prices-dataset

To start, the sports-car-prices–dataset required data cleaning to make the features more usable and easier to utilize for future manipulation and feature engineering. All features were converted to lowercase letters since it makes the data easier to use when needing to find strings later on. For features Horsepower, Torque(lb-ft), 0-60 MPH Time (seconds), there were special characters that had to be dealt with in order to make all the data consistent. Special characters like ‘+’, ‘-’, ‘<’, ‘>’, ‘,’ were considered to just drop. Engine Size(L) found NaN values that instead of removing from the data set and making the data size samples smaller, minor data integration were through outside sources. It turned out that all the NaN valued Car Makes were electric cars meaning we gave it value of 0.0 L. Engine Size (L) that contained the string word ‘electric’ was given 0.0 L and any car value that contained 1.5 + electric or hybrid, was given a hard coded value of 1.5 acting as the base value for hybrid sports cars. This allowed for Engine Size (L) to be completely numeric which makes it easier to use later on.

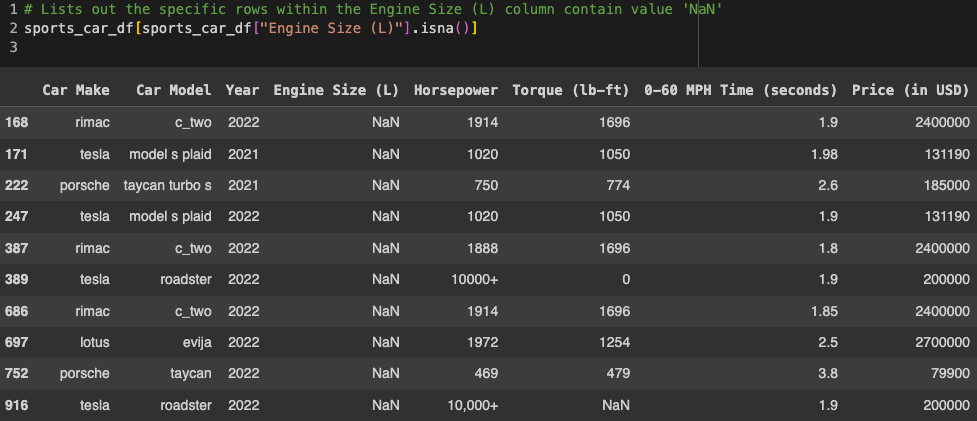
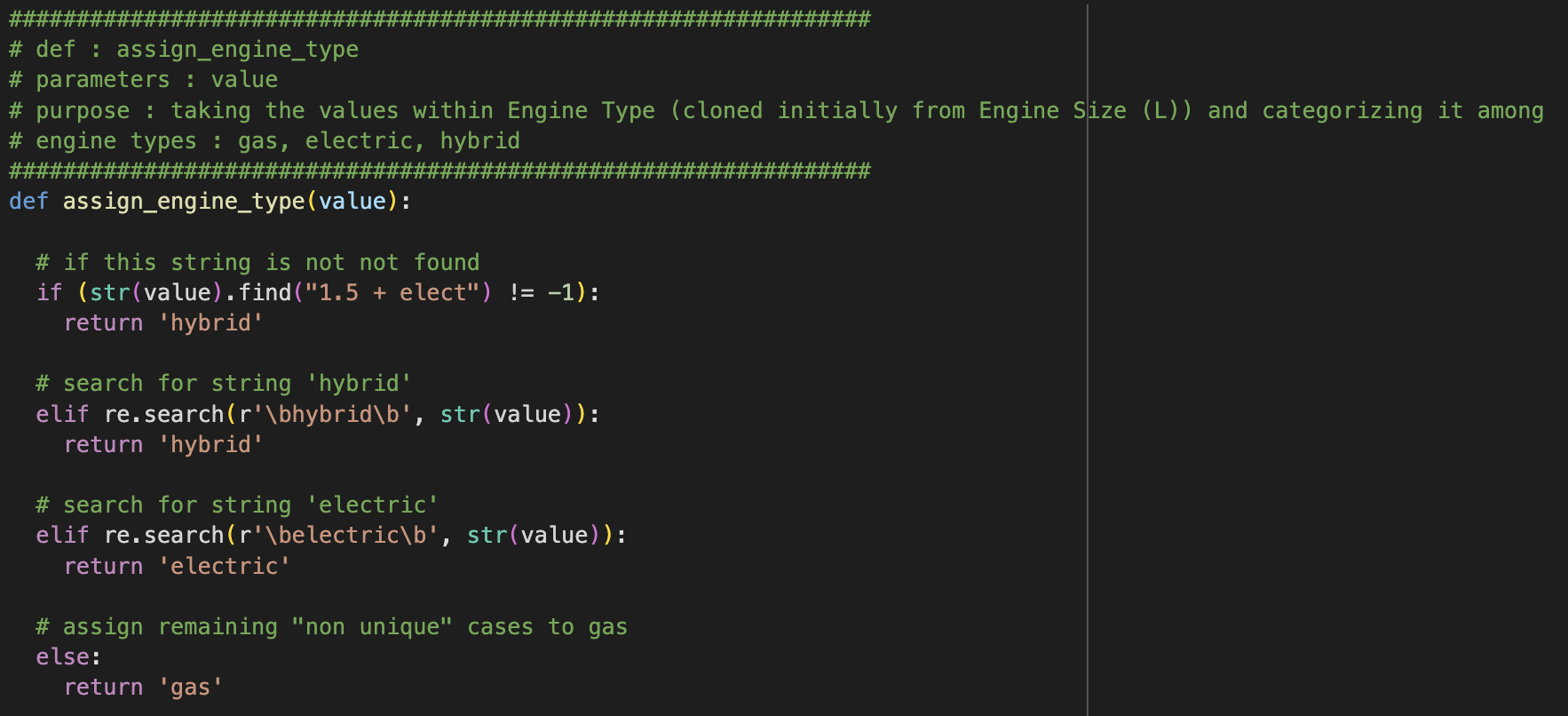


Figure 2: Showing Engine Size (L) unique cases that have NaN. All NaN car makes and models were determined to have value 0.0 since they were found to be all electric.

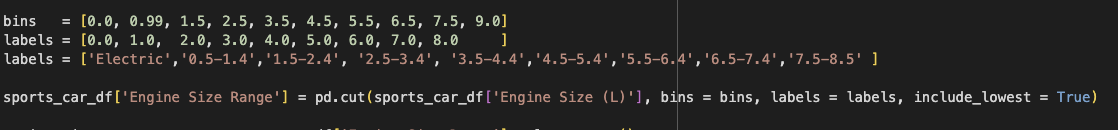
To expand our data set, we decided to utilize the engineering technique, feature engineering. The reason behind this decision was to see any additional trends to draw conclusions that may or may not support our project's goal.

Feature #1 that was engineered was ‘Engine Type’. This was based on the original ‘Engine Size (L)’ feature. This new feature had three values, gas, electric and ‘hybrid’. The reason behind creating this feature was to see if there was a relationship or trend of the type of car compared to the price of sports cars. Any Engine Size (L) that had a value of 0.0 was assigned to ‘electric’, anything with a special case that read a string ‘hybrid’ was assigned ‘hybrid’, and the rest of the non special cases were assigned ‘gas’. See code snippet for clarification that takes care of the special cases related to the electric and hybrid sports car cases.

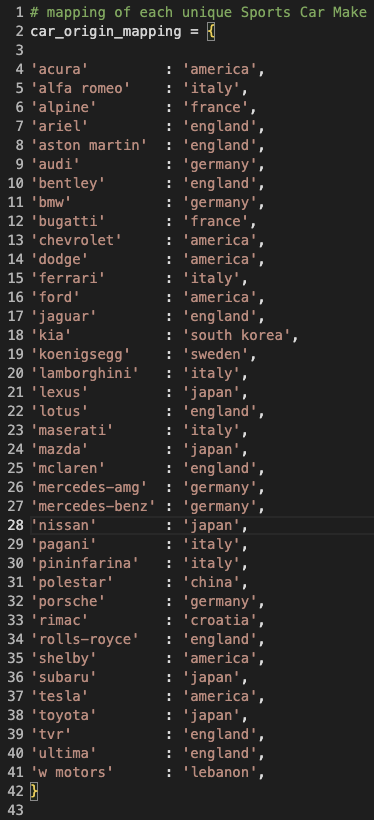


Code Snippet 1: this was the code to take care of any unique cases that were in the Engine Size (L) feature. The main purpose was to determine the hybrid and electric cars. The ‘else’ resulted in gas since it was a valid float value representing an accurate value for Engine Size (L).

Feature #2 that was engineered was ‘Engine Size Range’. When looking at the Distribution of the Engine Size (L) feature in Figure 4 and the plot comparing Engine SIze (L) to Price (in USD), there were too many unique Engine Sizes (L). We decided to print out the count of each unique Engine Size (L) to see the sample size of each size. We noticed that there were a lot of sizes that had a count of 10 or less. The range was count value of 1 to 219. There were 16 out of 34 unique Engine Size (L) count values that were 10 or less providing a very small sample size. that the data did not provide the best visualization and trends. By creating a range of engine sizes, it provided a better visualization to analyze trends that we will dive into later.

Code Snippet 2: this created the bins and labels for the newly created feature ‘Engine Size Range’.

Feature #3 that was engineered was ‘Origin’.This was done to see if there is an impact of the sports car’s price based on the origin country. Through data integration, outside sources confirmed the origin country of each unique Car Make. Each Car Make was mapped to its respected Origin. For example, Bugatti is from France, Porsche is from Germany, Pagani is from Italy, etc. See snippet for reference:



Code Snippet 3: Mapping of all the unique Car Makes within the dataset. This was done by data integration, utilizing outside sources to confirm the respective origin country.

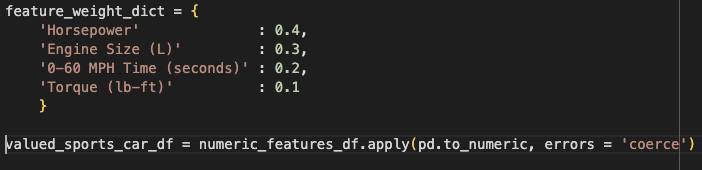
Feature #4 that was engineered was ‘$ to horsepower’. This was done to just see the relationship between the cost per horsepower. This allowed us to see each car's cost of each horsepower and compare it to the total price. This allowed us to see the relationship if there is a car with a high $ cost per horsepower, the total price should be higher. This was not too crucial to our conclusions but provided confidence in the horsepower contribution.



Code Snippet : Creating new feature $ per Horsepower through the existing features ‘Price (in USD)’ and ‘Horsepower’

Feature #5 that was engineered was ‘Score’. The purpose of this feature was to give each numeric feature a weight based on the importance of the features. The reason we did not include the categorical features within the dataset was due to the need to give them numeric values so that they can be multiplied by the weighted values. If we were to utilize one hot encoding and give each unique case a unique binary value, it would just be 1\*weight and not having a true impact or meaning to the score. If we were to utilize label encoder, this would give each unique case a value but the value given would not provide any background or meaning. Label Encoding would skew the score for each car so we did not carry with having any categorical features that have an impact on the score feature.

The value of the weights were based on a few opinions when asking a very small pool of family or friends of what features they thought were more important. The feature weight values mapping



Code snippet 4: Mapping of numeric features to determine weight along with the new feature ‘Score’ created and how the math is implemented to come up with the score of each car.

Now with all the original features cleaned and with newly engineered features, we would utilize familiar techniques learned through the course to create visualizations of the relationships between the features and Price and try to provide conclusions.

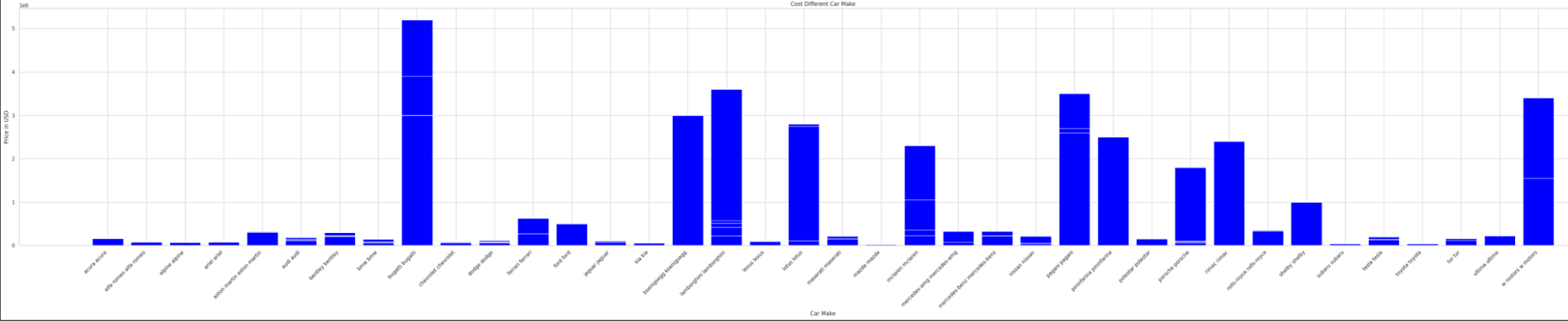


Figure 3: Car Make vs Price

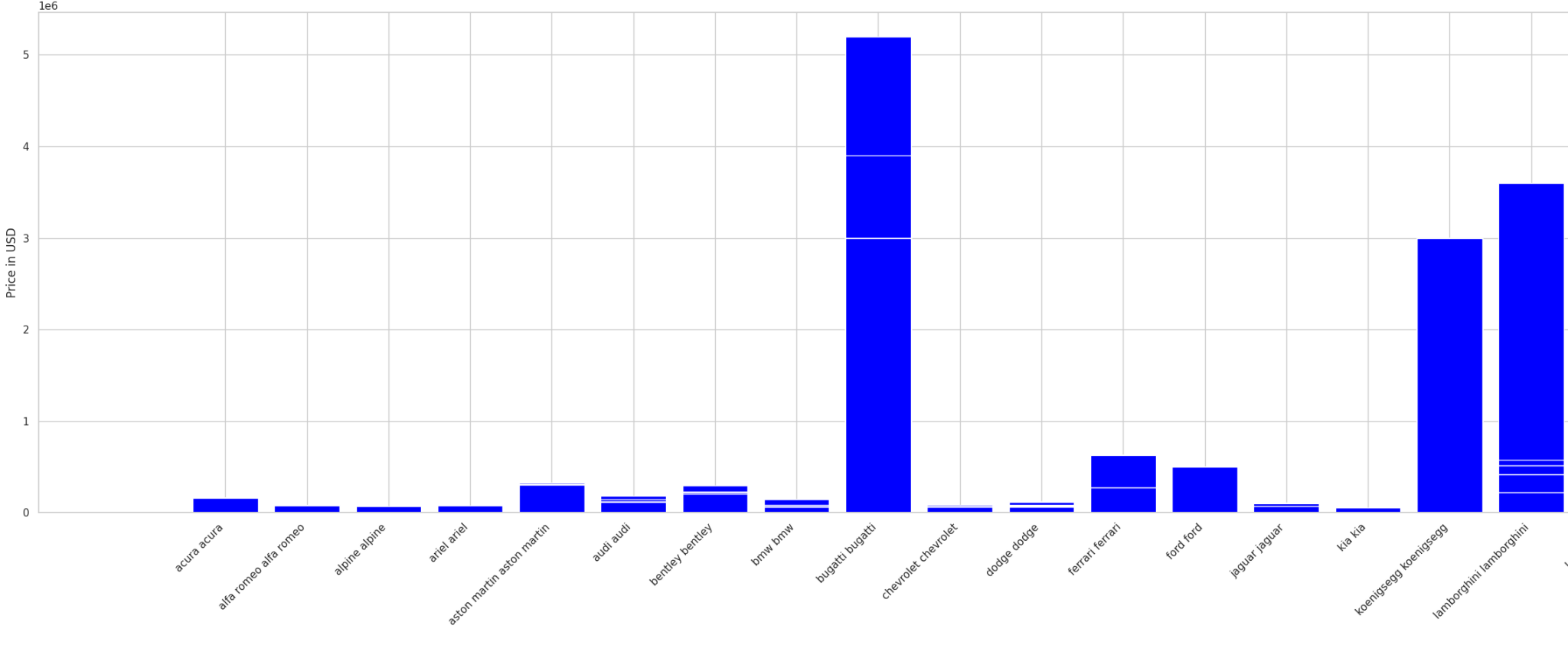


Figure 3a: Car Make vs Price (zoomed in section 1)

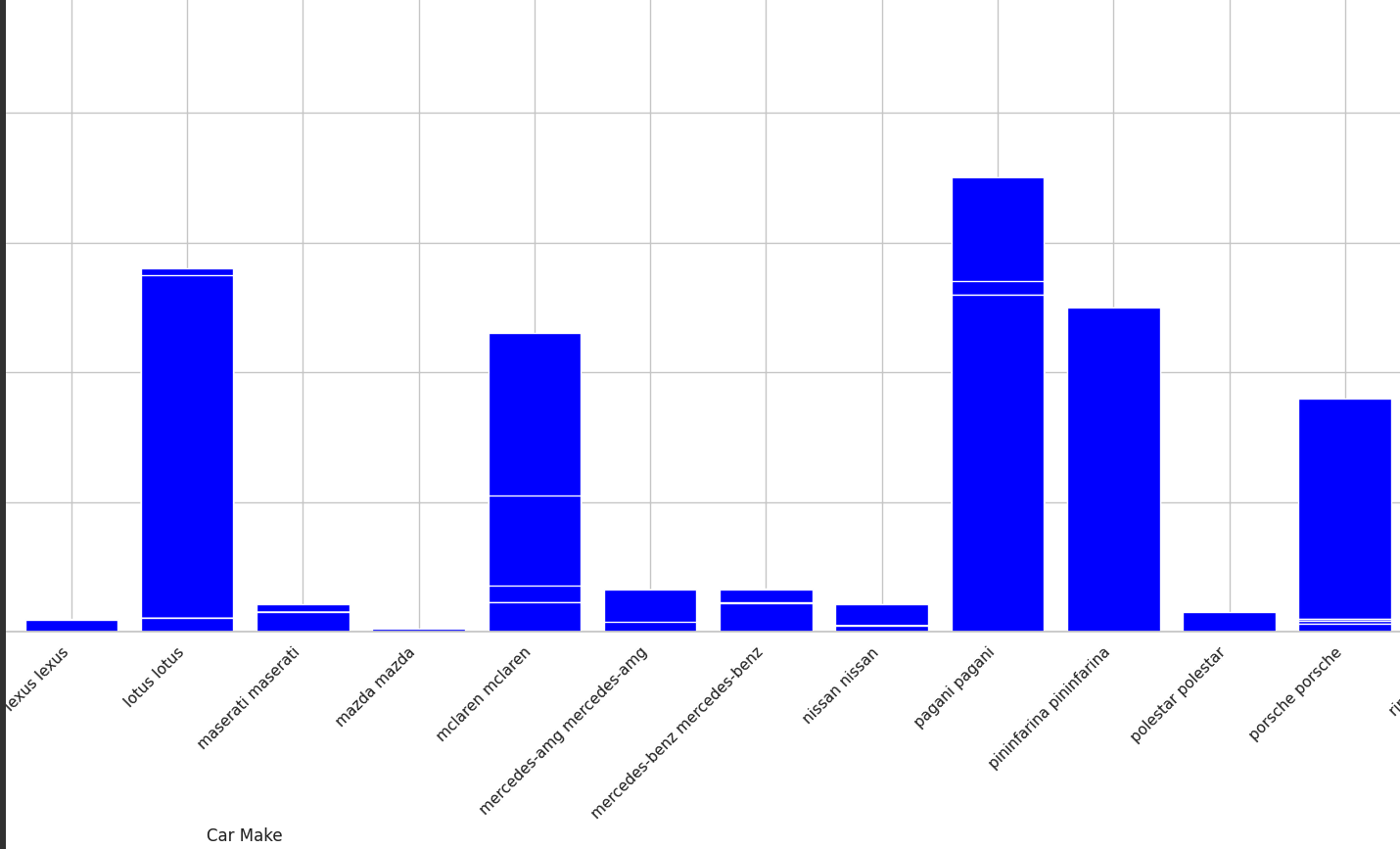


Figure 3b: Car Make vs Price (zoomed in section 2)

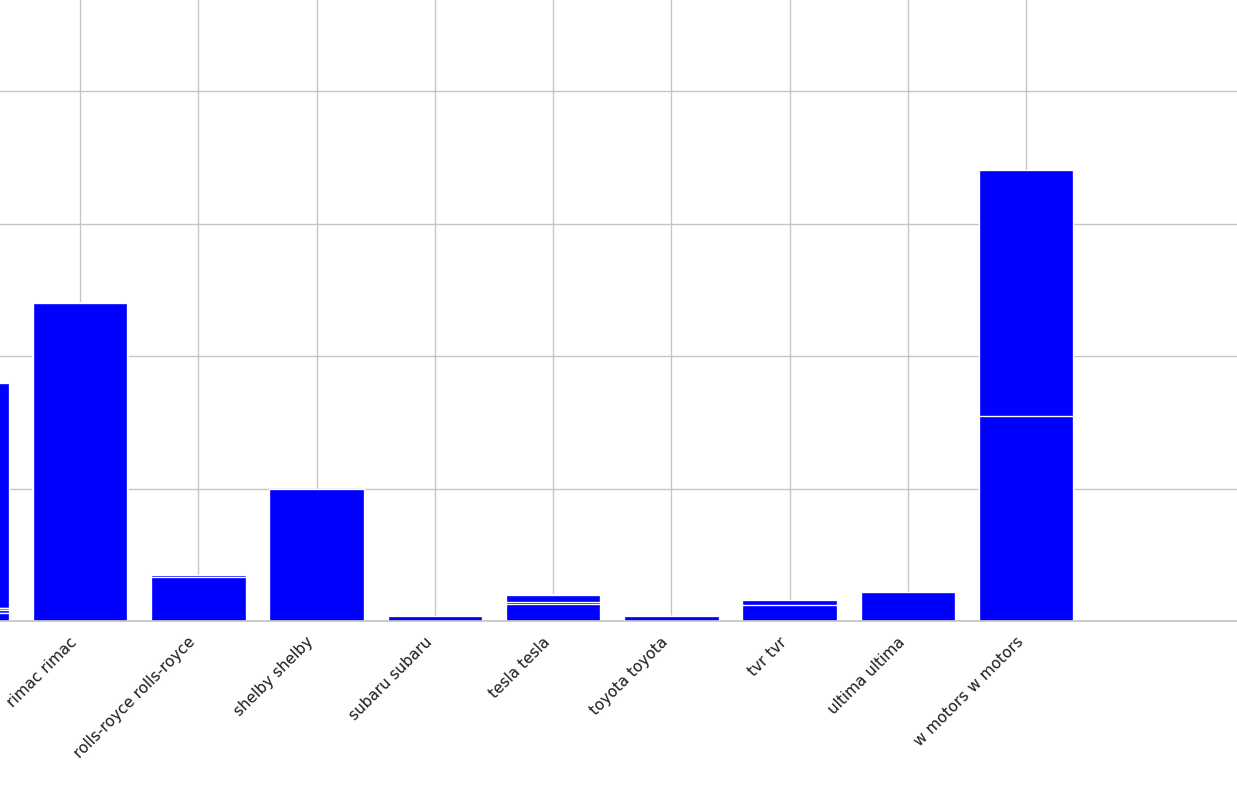


Figure 3c: Car Make vs Price (zoomed in section 3)

The reason for Figure 3 was to visually represent which Car Makes are the most expensive and also look at the cars that are the least expensive. We can see that out of the unique Car Makes, we can analyze that Bugatti is the most expensive followed by Lambogiuini, Pagani, Koenigsegg, and W Motors. For the least expensive cars, we can say Mazda, Chevrolet, and Kia. This plot is something we keep in mind as we continue with our work since we look for these cars specifically when seeing if their features impact their price.

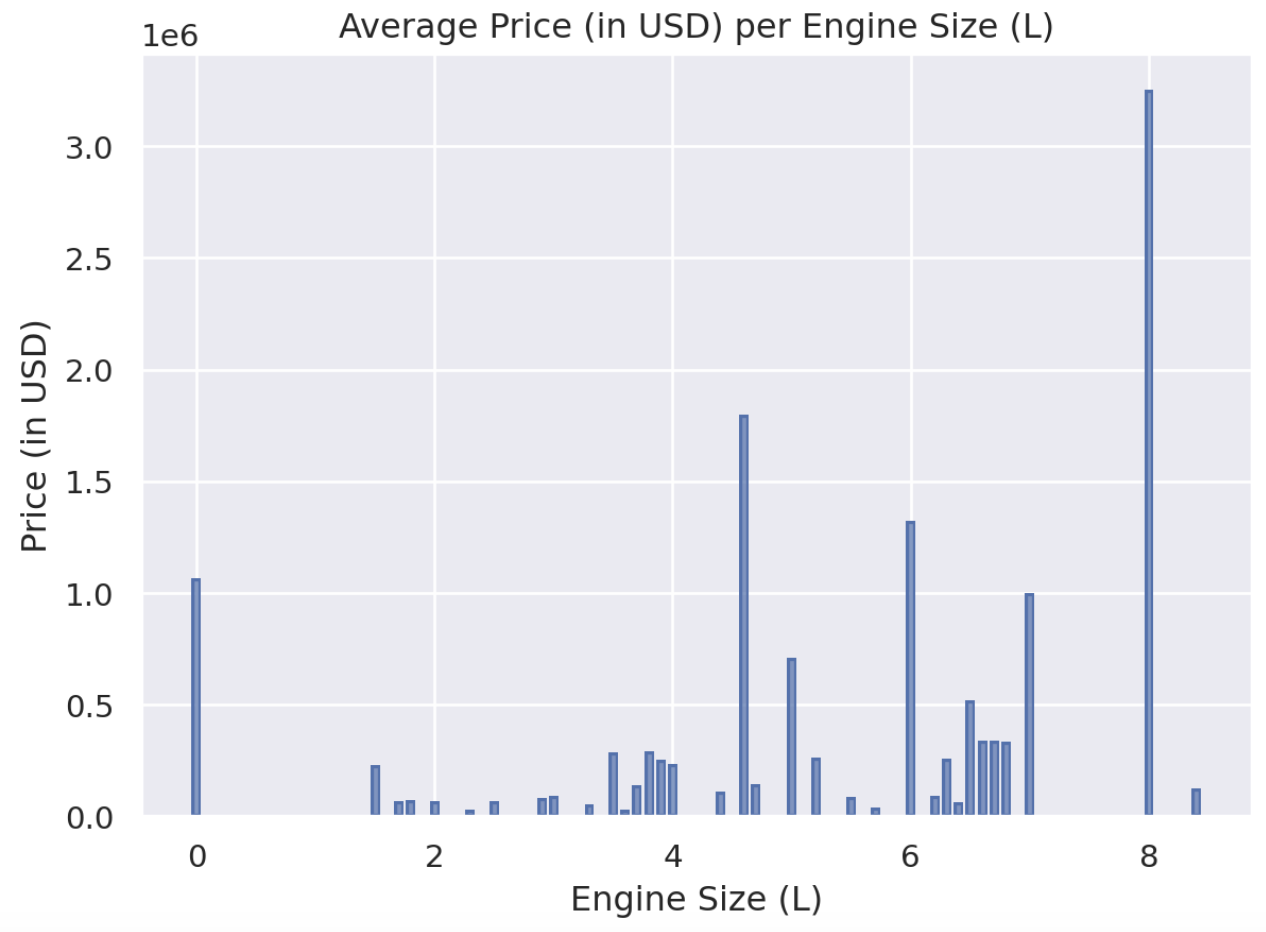
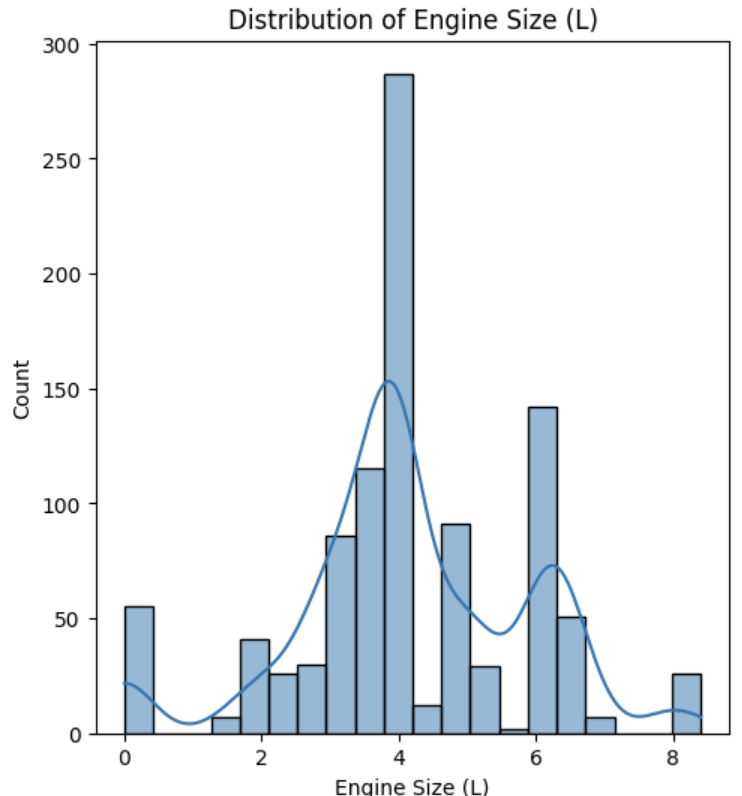


Figure 4: Distribution of Engine Size (L) Figure 5: Engine Size (L) vs Price

Figure 4 shows the distribution of Engine Size (L) and we can see a large amount of cars fall within 4 L. Reminder that 0.0 L engine size are the cars that are Electric and have motors, not engines. 4.0 L had the highest count of cars with that size which lines up with the distribution plot of Figure 4. Figure 5 shows the Engine Size (L) vs Price and it is sort of hard to see any possible trends. Due to the struggle to spot any trends, it did spark the creation of the new feature ‘Engine Size Range’ which was explained earlier in the report.

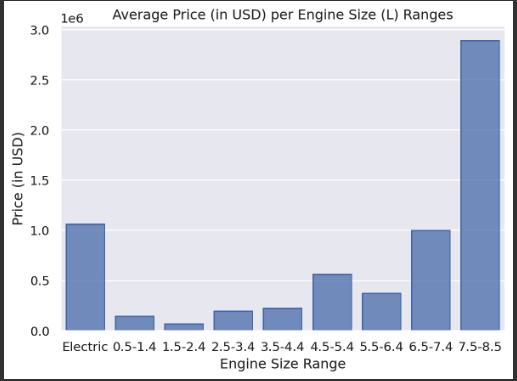
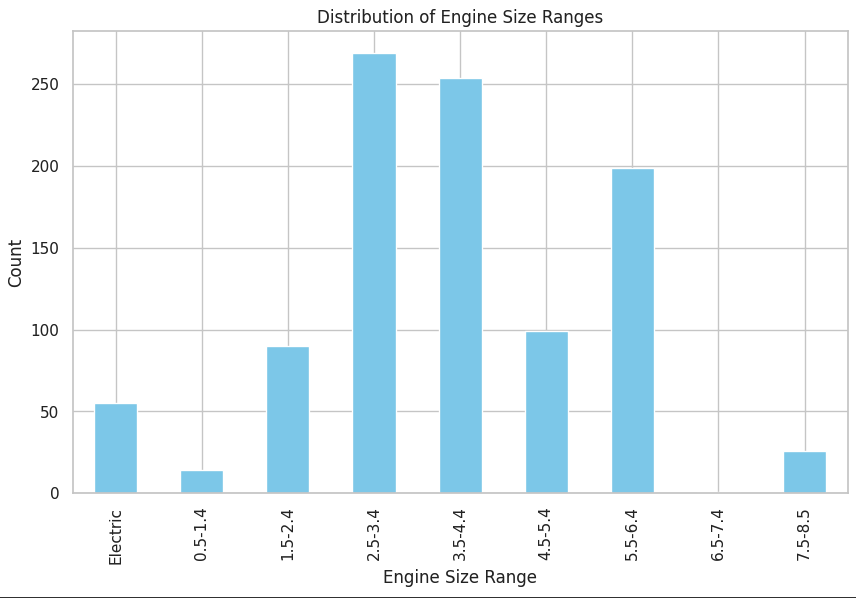


Figure 6 & 7 : Distribution of Engine Size (L) & Engine Size (L) vs Price

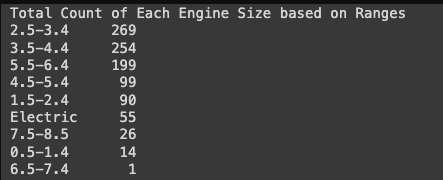


Figure 8: Total Count of Each Engine Size based on Ranges

Figure 6 shows the distribution of Engine Size Range and we can still see the distribution is heavy near the 4.0 L. Given the new sample sizes (Figure 8) being better than the unique value cases, we were able to plot (Figure 7) and notice some possible trends. The trends we see in Figure 7 are upward trends as the engine sizes increase, so do the prices. There might have been a few cars within the ranges that skew the data in 0.5 L - 1.4 L and 4.5 L - 5.4 L. We think it is fair to make an assumption based on the trend that the Price does increase when a sports car’s engine size is larger.

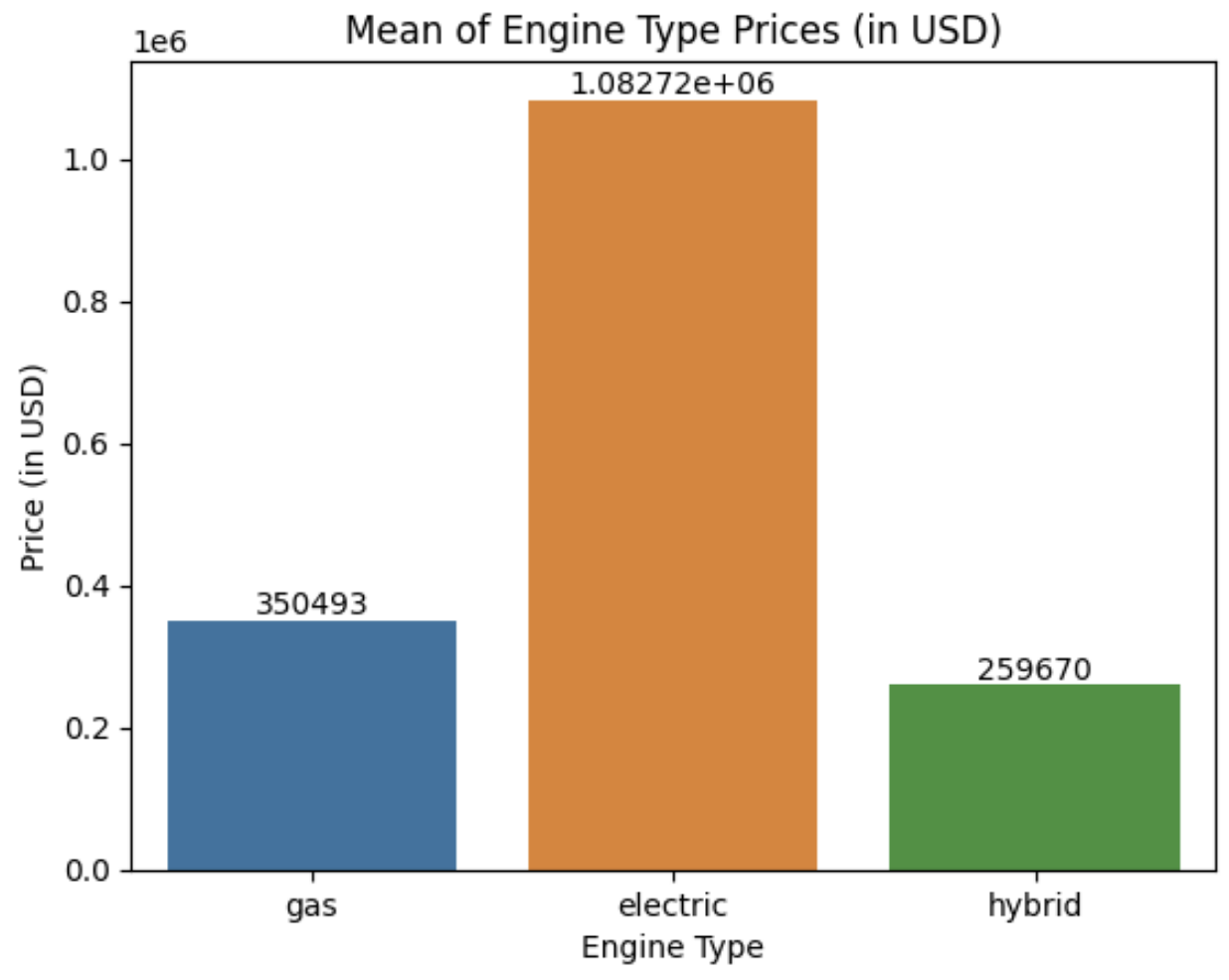
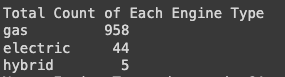


Figure 9: Average Prices of each Engine Type (gas, electric, hybrid)



With this newly created feature, we thought that this feature would bring important value to see the impact of the sports car’s prices. Due to the lack of cars within the electric or hybrid engine type, it would not be valid to make conclusions. With gas having 958 different cars that are gas, it provides a more accurate representation of the price based on its type. Electric and Hybrid have such a small count and it suggests looking at just a few cars to make a determination of its price. In order to make conclusions per this feature, it would require a much larger and even count sample size for each type. Though this would have been an important feature to see the impact on the prices, we will not utilize this in our conclusions.

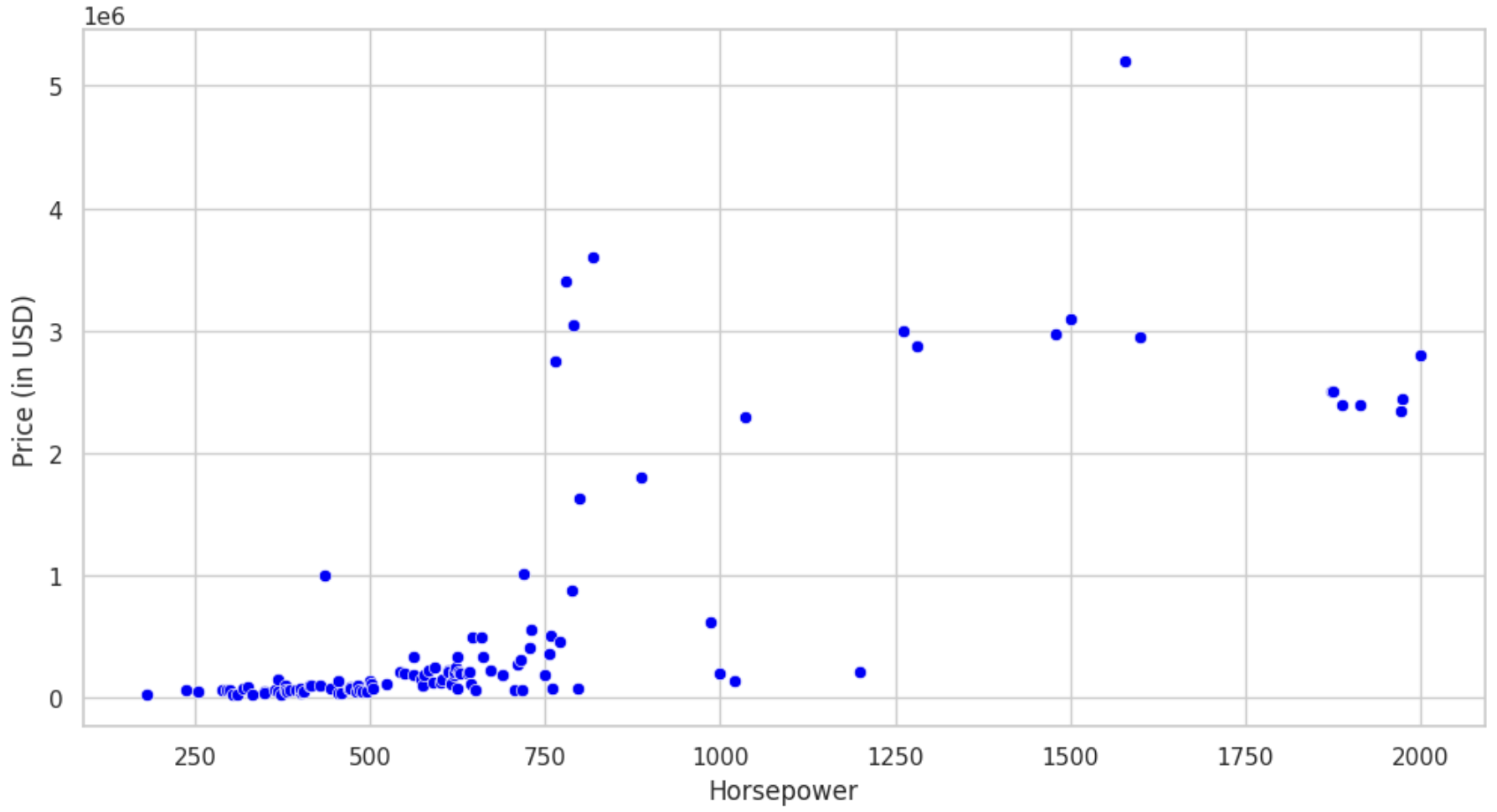


Figure 10: Horsepower vs Price (in USD)

Figure 10 shows us the impact Horsepower has on the price. We see that there is a distinct upward trend within the 250-750 horsepower range meaning, as the horsepower increases, so does the price. From 750 - 2000 horsepower, we lack the amount of data points to show a trend but we do notice that the prices do increase as the horsepower increases. If there were more data within the data set, we would expect a much more obvious trend throughout the range 0-2000 horsepower. Based on this figure and the analysis we made, we are confident to say that as the car’s horsepower value increases, so does the price in USD.

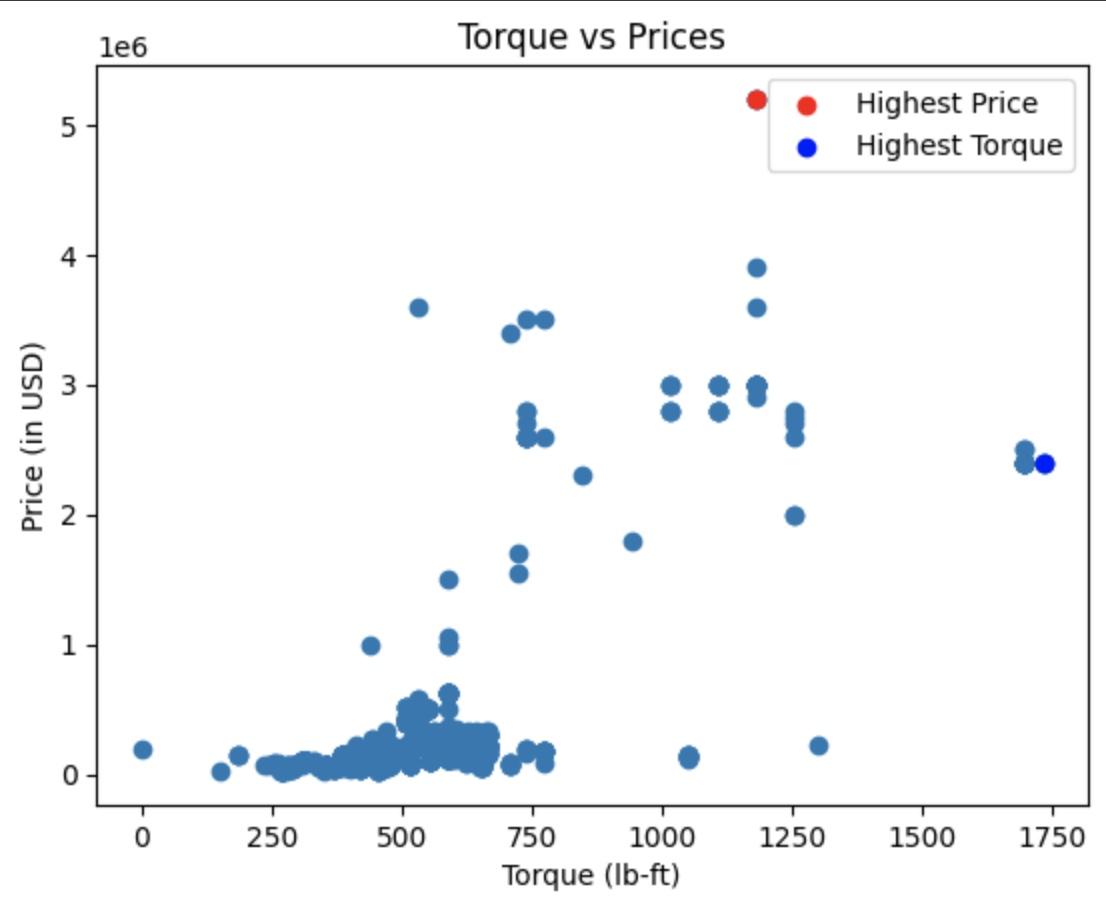
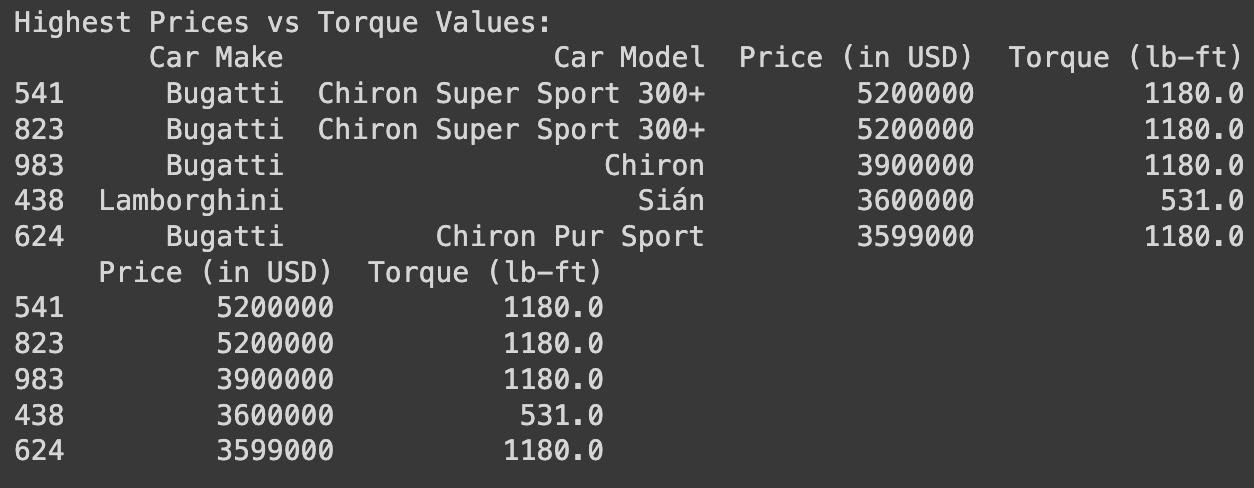
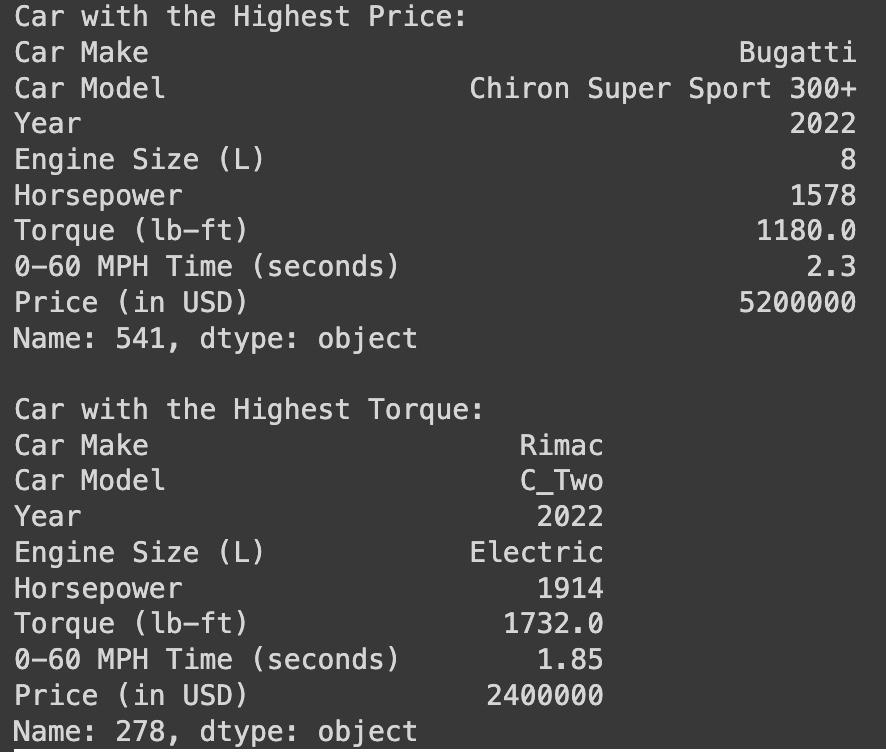
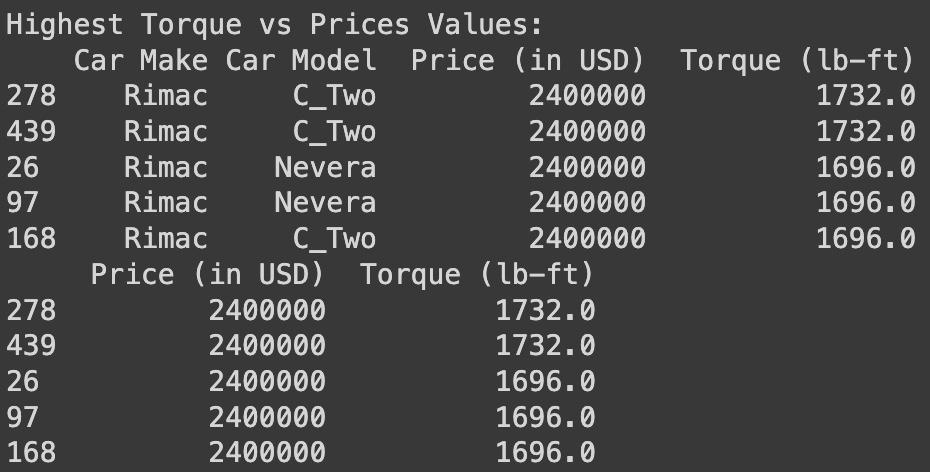


Figure 11: Torque (lb-ft) vs Price (in USD)



There is a concentration of data points within the 250-750 Torque range. Although, majority of the data in this range, we notice an upwards trend as the torque increases, the price increases. We can interpret the data and see the highest Torque for a car was an electric car. It had a horsepower of 1914 ft-lb/s and the torque was 1732 lb-ft, coming at a price of 2.4 million USD. We can interpret the data for the highest price for a sports car from the dataset which has a Horsepower of 1578 ft-lb/s and torque of 1180 lb-ft, coming at a price of 5.2 million USD. Torque is an important factor in understanding the performance of a sports car. It directly correlates with the amount of air flowing through the engine. In the context of engine dynamics we see that larger engines have the capacity to pump more air and this will result in production of higher torque.

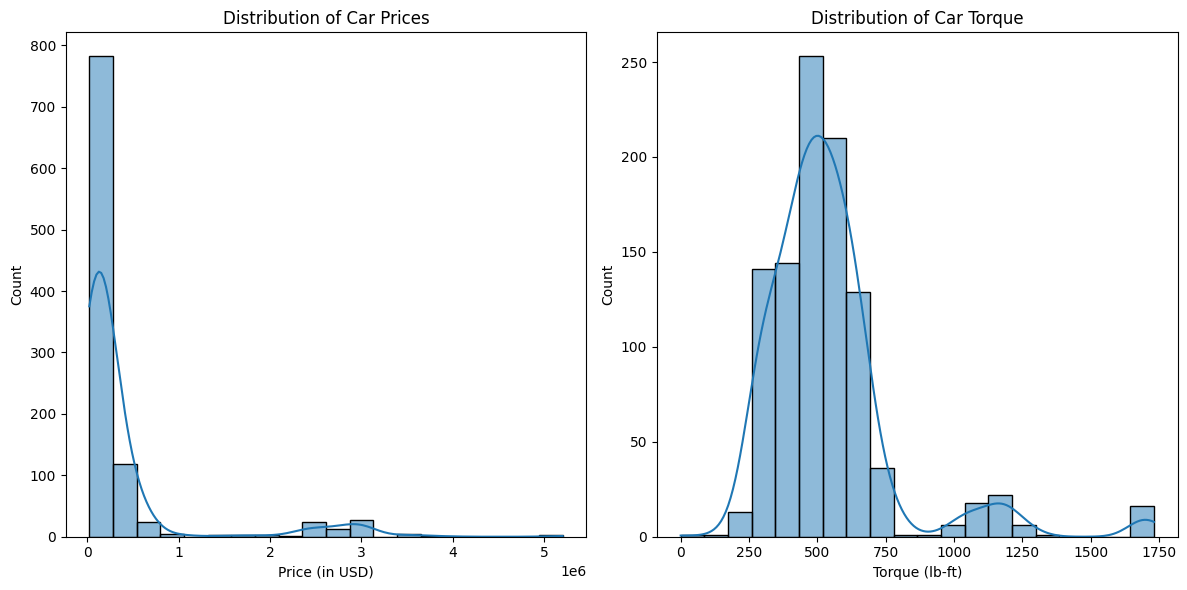


Figure 12: Distribution plot of sports car prices in relation to the distribution of torque

According to the distribution plots, the car torque follows a normal distribution. Although it follows a relative bell shape the graph is skewed slightly to the right denoting that the average is higher than the median torques. The car prices are completely skewed left indicating an average prices being less than the range of pricier cars.

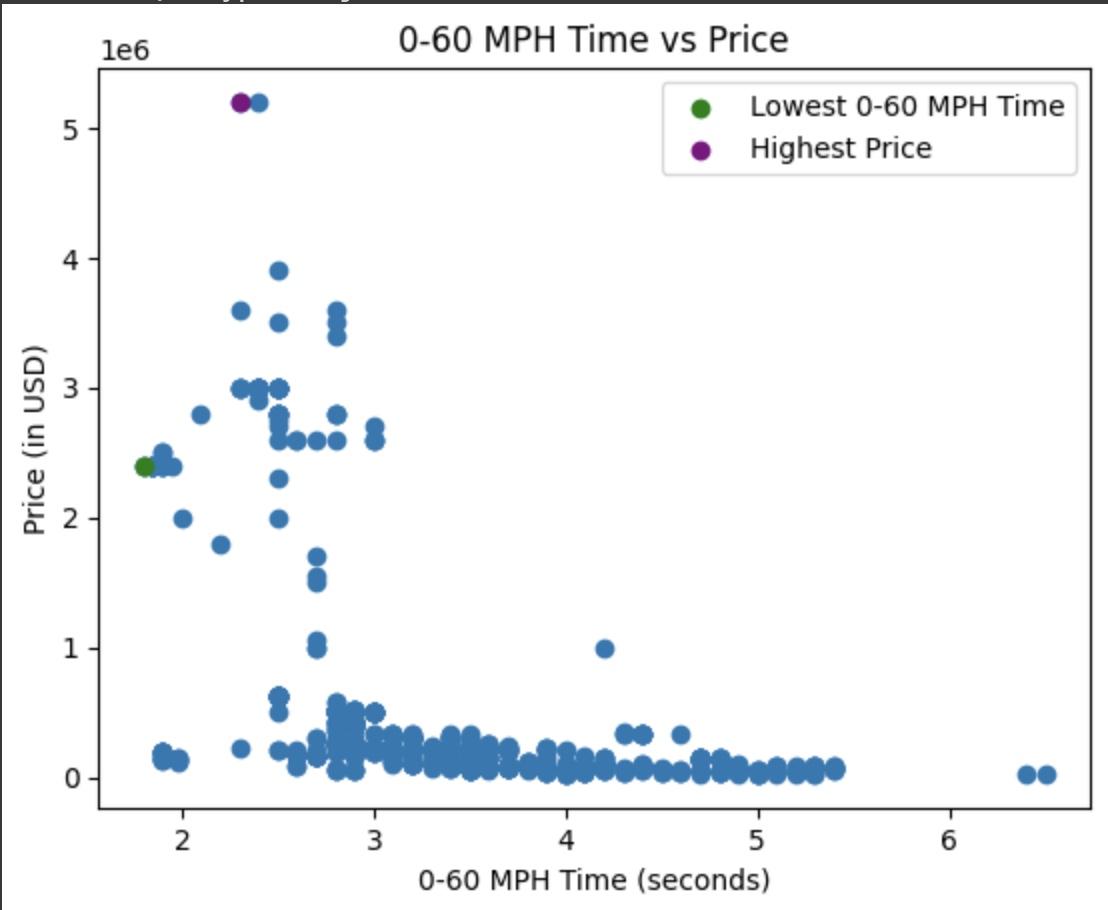
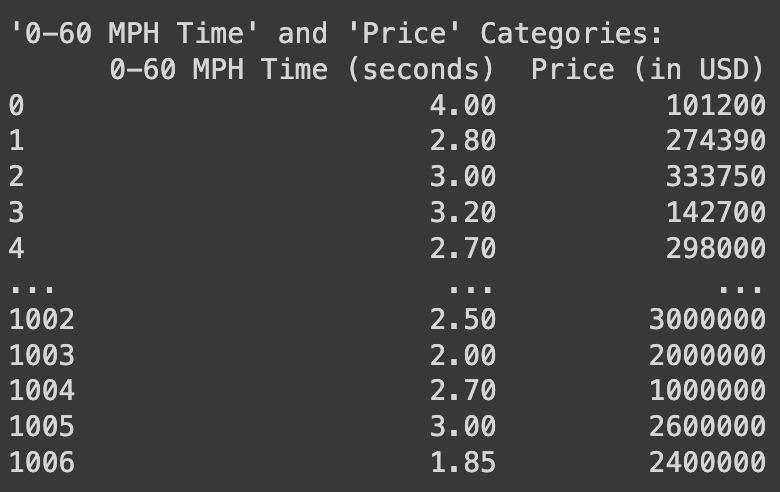
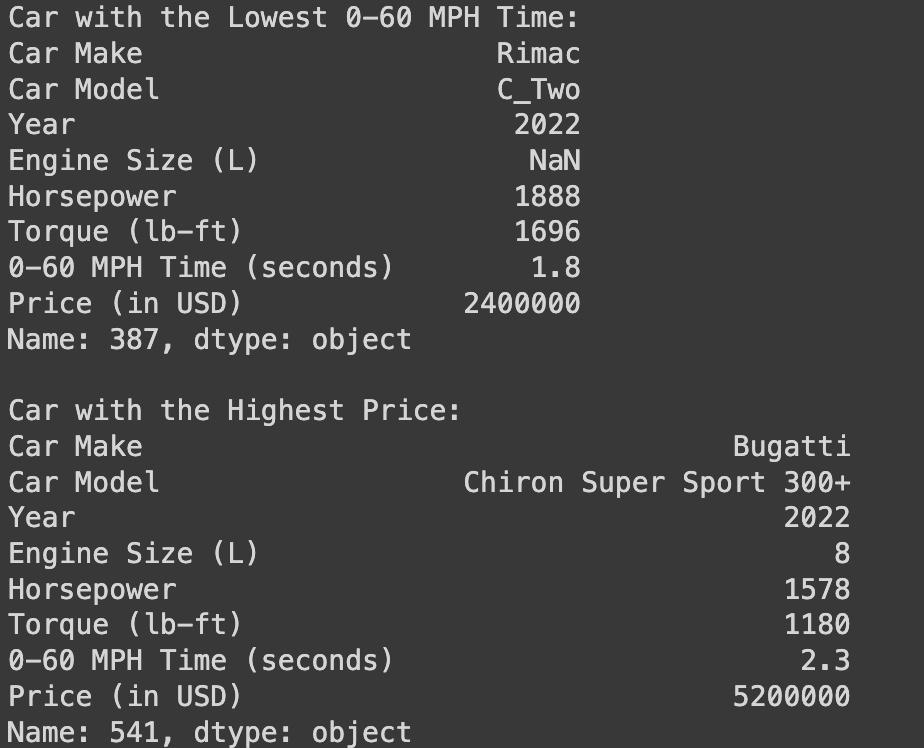
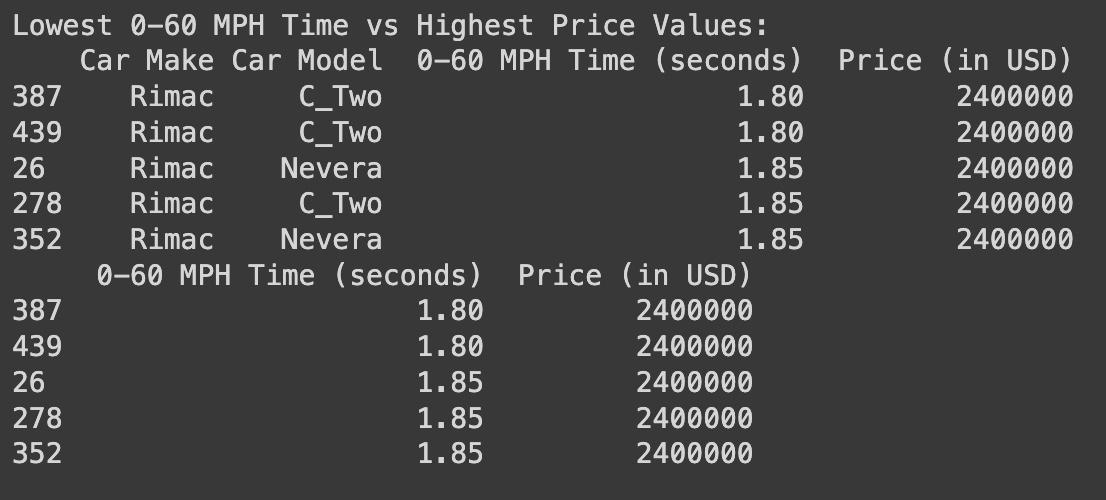
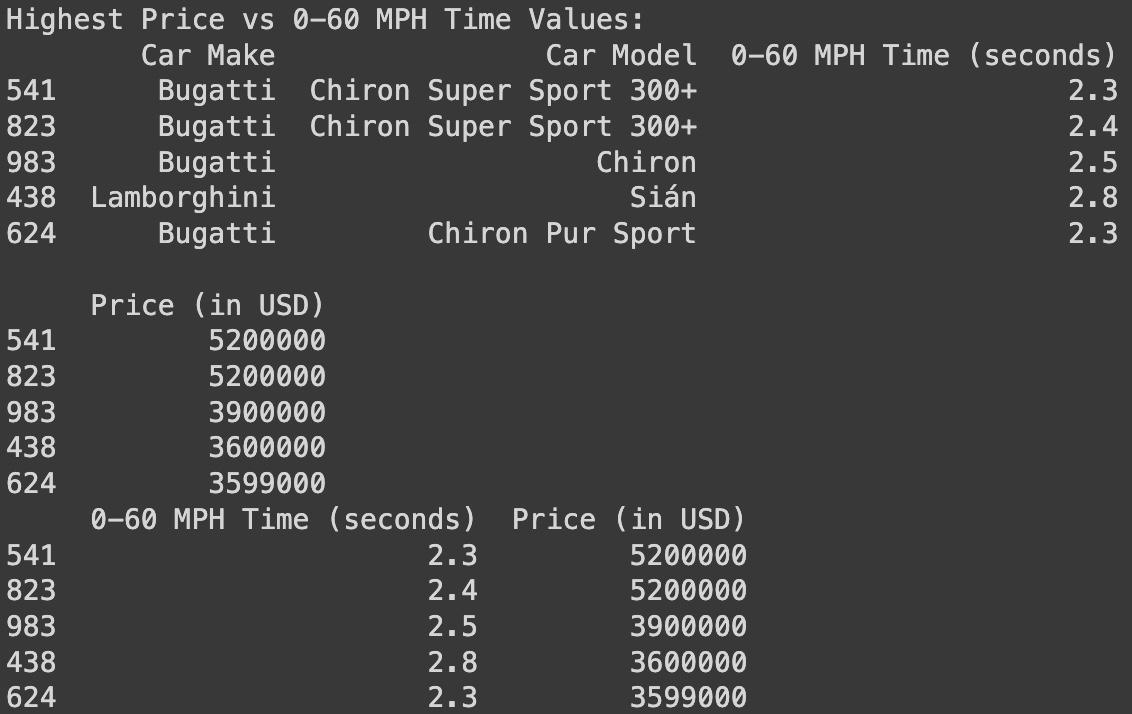


Figure 13: 0-60 MPH Time vs Price



This data shows the faster the acceleration (lower 0-60 time) the higher the price. Price ranges indicate variability in acceleration performance while the prices show a diverse range among the car models. The lowest 0-60 times are from the same manufacturer, which has the lowest time for 0-60 mph acceleration of 1.8 seconds. This was also recorded for the highest torque in the previous data analysis. It is also among the higher priced car range. The highest priced car demonstrates a competitive 0-60 MPH time of 2.3 seconds which balances the luxury and performance for a sports car. The diversity of the dataset suggests some cars will achieve exceptional acceleration at relatively lower prices while other cars combine high performance with luxury at a premium price point.

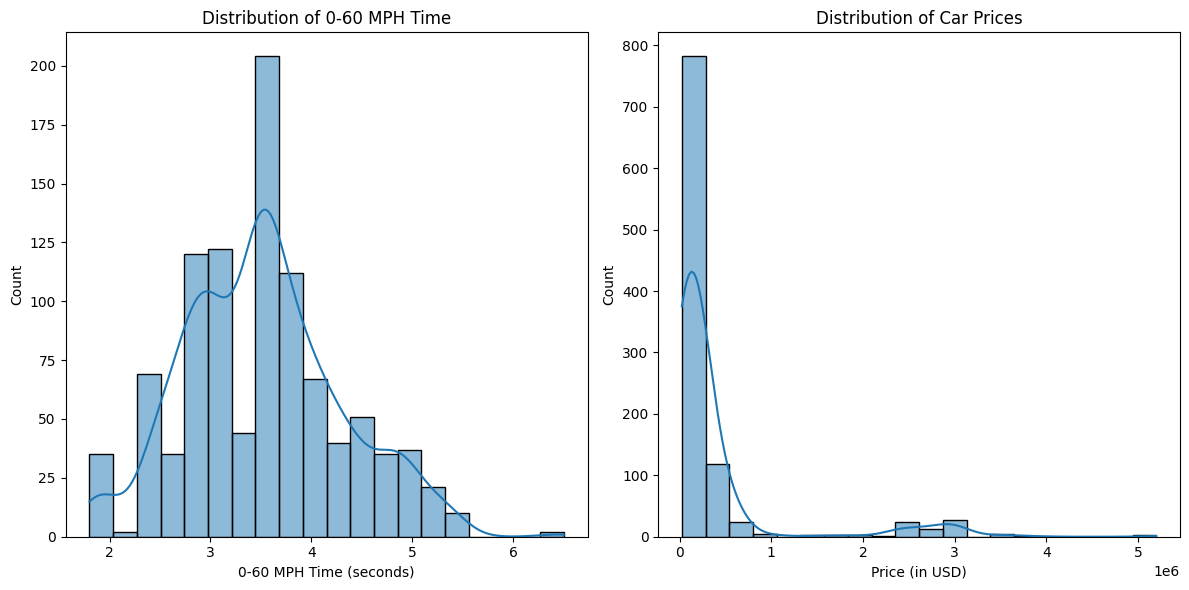


Figure 14 & 15 : Distribution 0-60 MPH Time & Distribution of Car Prices

The normal distribution of the 0-60 mph shows a better for a bell shaped curve and has a relative median, mean, and mode because the data is concentrated towards the center while being slightly skewed to the left similar to the torque distribution.

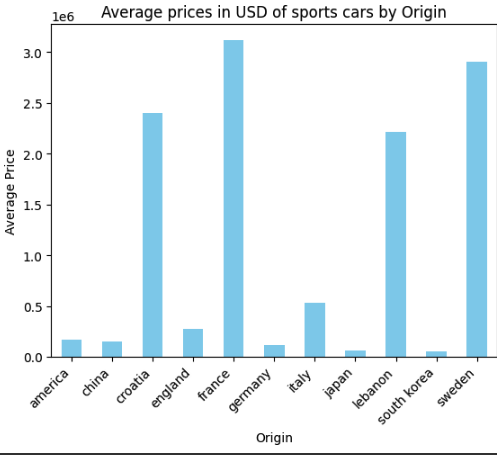


Figure 16 : Average Prices (in USD) based on Origin (Country)

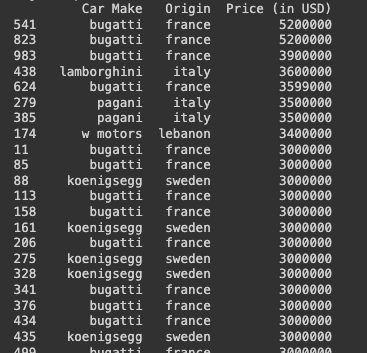


Figure 16 shows the newly created feature ‘Origin’ and the average prices of their sports cars. There is a vast range of numbers of cars from their origin country within this data set. Although this may skew some of the values, this plot and information is still important. We can see from the plot in Figure 13 that France has the most expensive sports car price on average, followed by Sweden, Croatia, Lebanon, and Italy. This makes sense to analyze because if we look at the top ~20 most expensive cars, we notice the origin of each are in the list of the top 5 countries we just listed. Through this plot and analysis of the images containing the most expensive and the average country price, we can conclude that the country of origin has an impact on the price of sports cars. Sports cars from specific countries tend to carry a higher price tag.

Finally from Dataset #1 , the last feature we created and wanted to explore was ‘Score’ where we gave the previously mentioned weighted values to the numeric features.

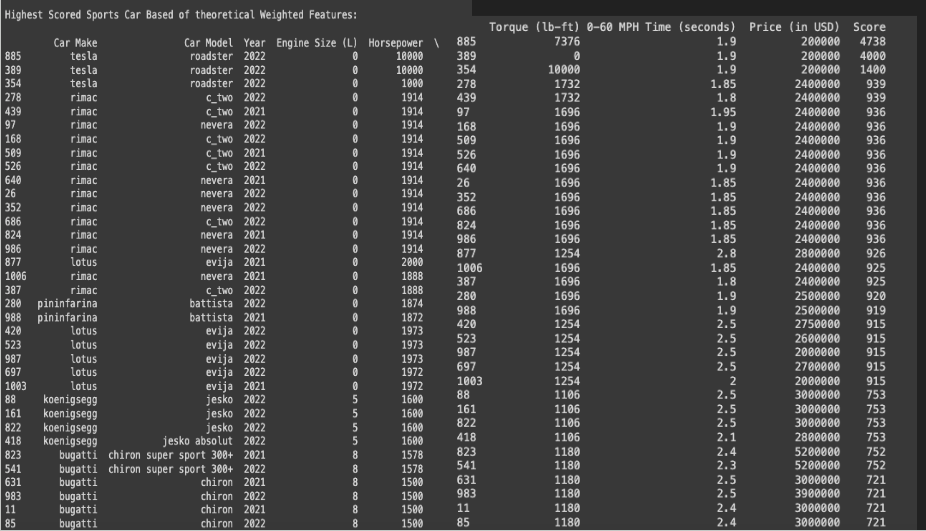
cor

Figure 17: List of the Top Scored Sports Car given the Score Calculation via feature weights

Weights:

1. Horsepower = 0.4

2. engine size (L) = 0.3

3. 0-60MPH Time (seconds) = 0.2

4. torque (lb-ft) = 0.1

Calculation:

Score = Xn,1 \* 0.4 + Xn,2 \* 0.3 + Xn,3 \* 0.2 + Xn,4 \* 0.1

where n represents the row and 1 - 4 represents the respective weight feature.

The weight values were theoretical and based on a small sample of opinions. We understand that the weight of importance of these features would be different per person asked. Based on the feedback we received, we determined the weight values to be as provided. Given the weights and the calculation, we can see in Figure 14 the highest scored sports cars. A common trend we noticed was that many of the highest scored sports cars are in the list of highest priced sports cars. A common theme is that the highest priced cars have a horsepower higher than 1200 Horsepower. Since horsepower was weighted with the highest importance, we notice that the highest scored cars were the ones with the highest horsepower within the data set. If we were to change the order of importance and change the weight differently, we would notice a different list. While the provided list may serve as a useful guideline / reference, it's essential to recognize that it represents a specific set of opinions and priorities.

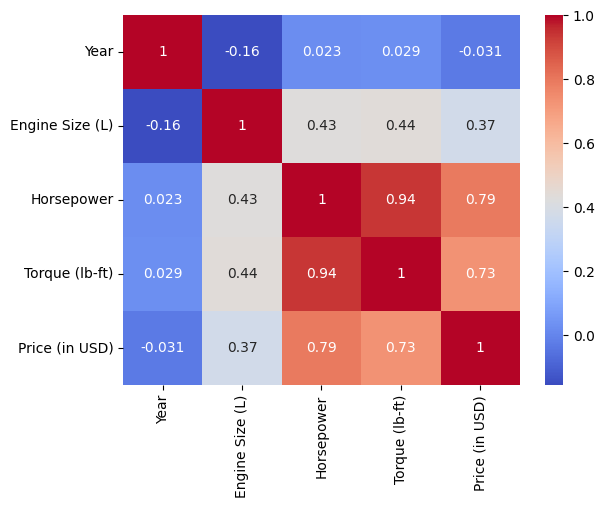
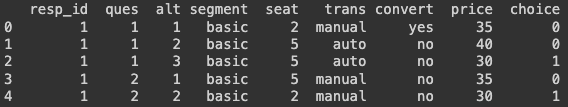


Figure 18: Heatmap for Feature Correlation

The heatmap is to understand how these features are correlated and identify any patterns or trends that provide insights that influence prices of sports cars. We only included engine size, horsepower, torque, price and year to give us a range of heat saturation. Year and Engine size having the darkest blue block suggest that the newer sports cars tend to have smaller engine sizes. Year feature has a lot of negative correlation with all the features which indicates that the year alone does not strongly predict changes in features. Horsepower and torque have the darkest red correlations which tells us that it has a strong positive correlation between the two features, higher horsepower also means higher torque. Price has the highest correlation with horsepower which indicates that higher priced sports cars also have higher horsepower followed by higher torque.



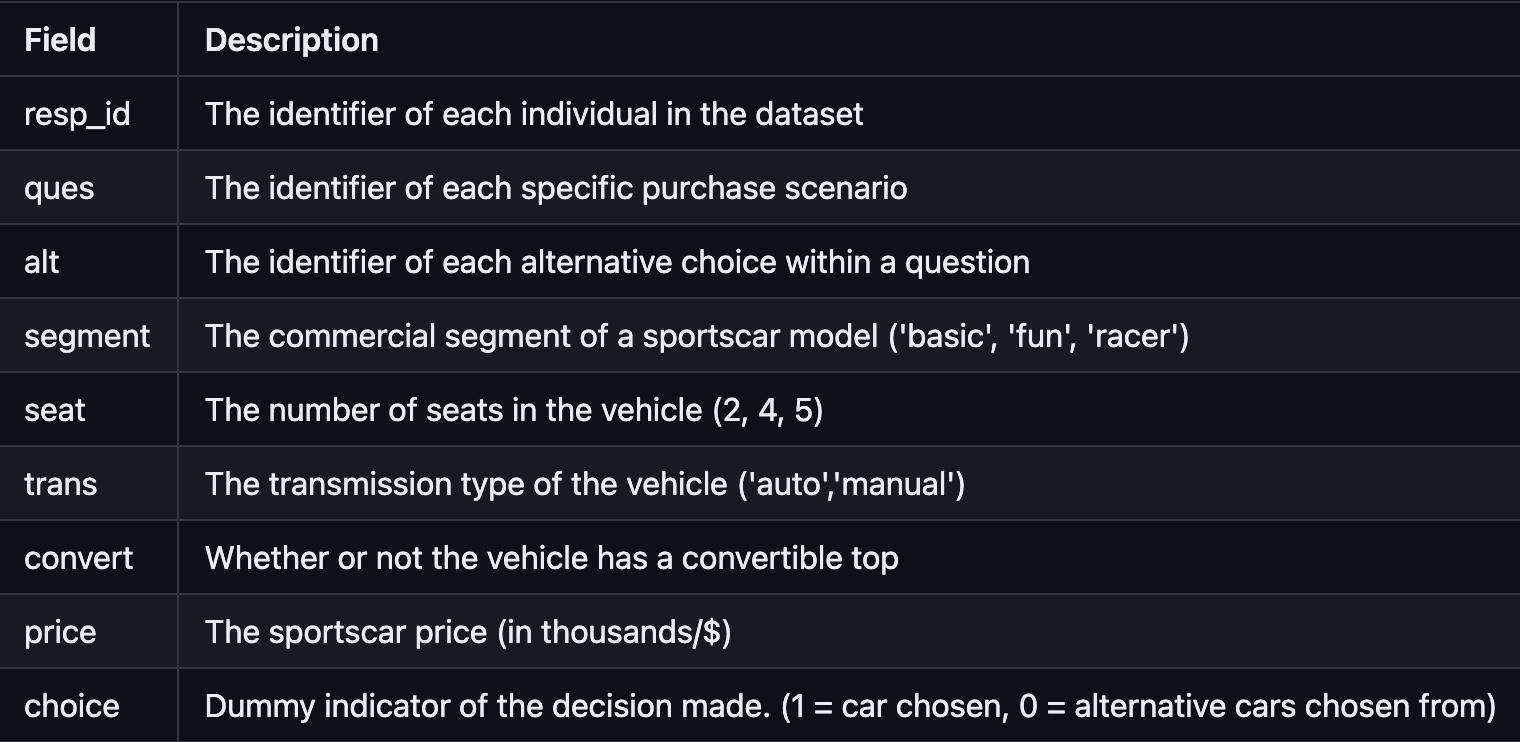


Figure 19: Dataset 2 descriptions

This dataset explores the decision making of 200 individuals when it came to sports cars design and configuration. This dataset was set up to be in a long, choice model format that includes identifiers and a dummy indicator of the choice column. Some of the categories of sports car design include commercial segment, seat number, transmission type, convertible top, and price.

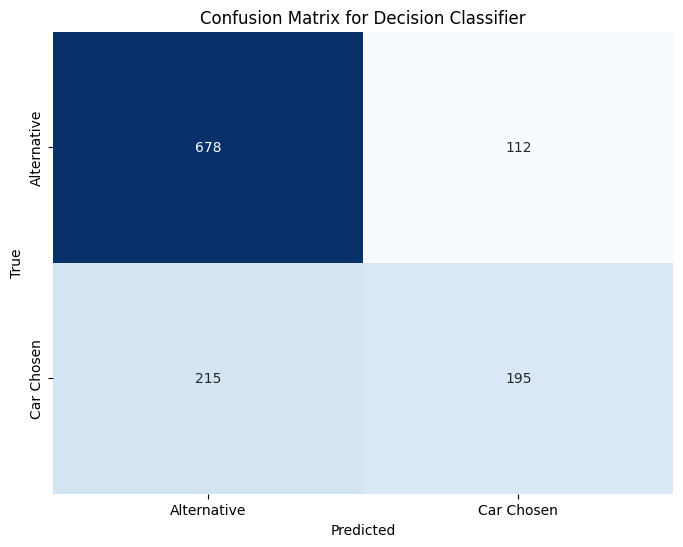


Figure 20 : Confusion Matrix

This model takes the data of the sports car choices dataset and classifies them either Car Chosen or Alternative based on features. This is based on the choice column that predicts whether a user might choose a car or an alternative, it helps to assess the model’s performance. 678 for true positives show alternative instances for predictions that were true for alternative. 112 instances show the model incorrectly predicted the car chosen while the true class was alternative for false positives. 215 represents the model incorrectly predicted alternative while the true class was the car chosen. 195 instances where the model indicates that the model correctly classified the car chosen when the true class was also the car chosen as true negatives. The higher numbers of true positives and true negatives show that it was effective at classification. The false positives and negatives show the misclassification which can be insightful into patterns that can be improved for the model.

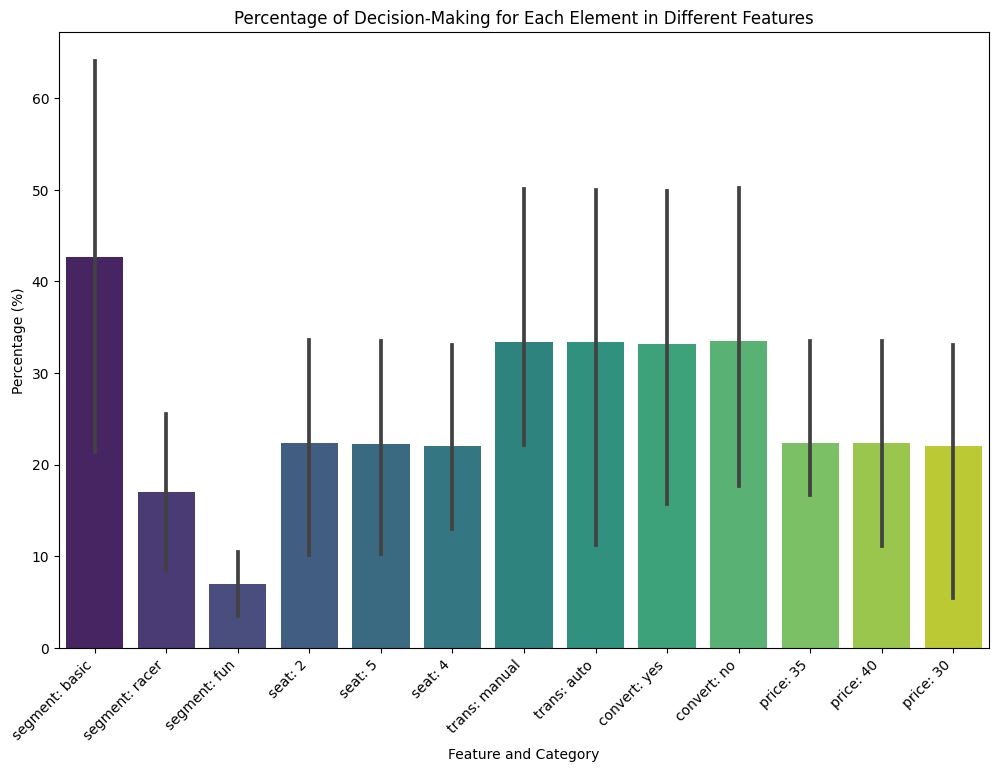
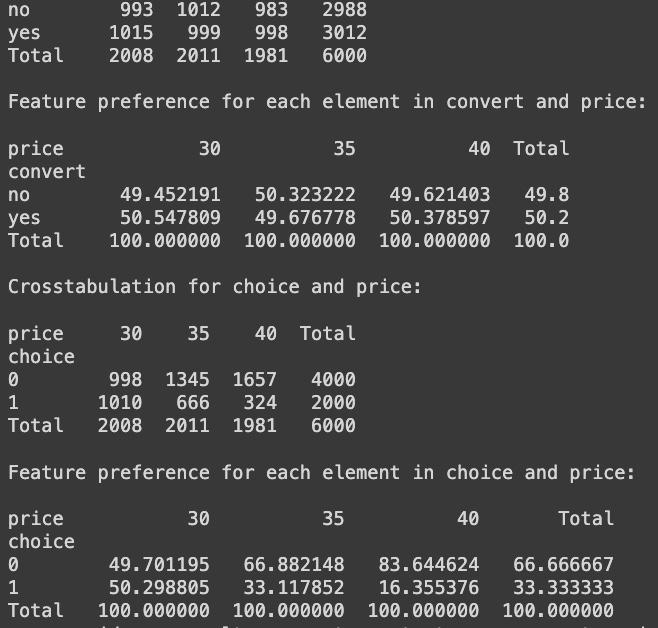
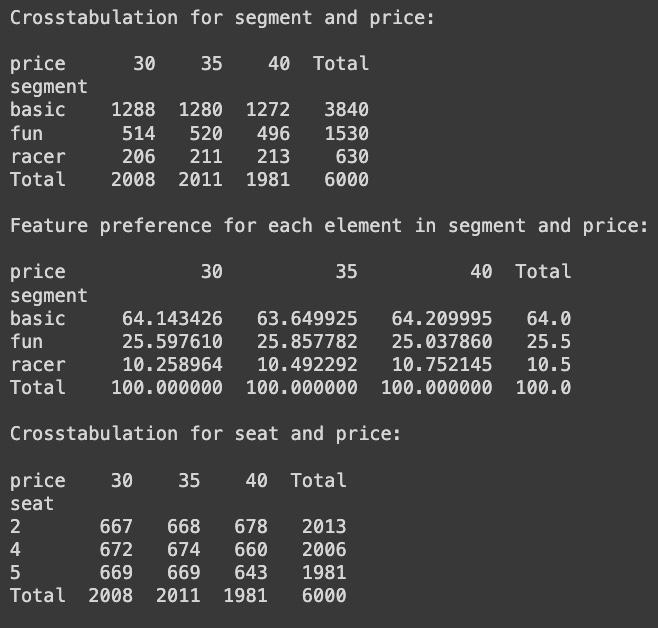
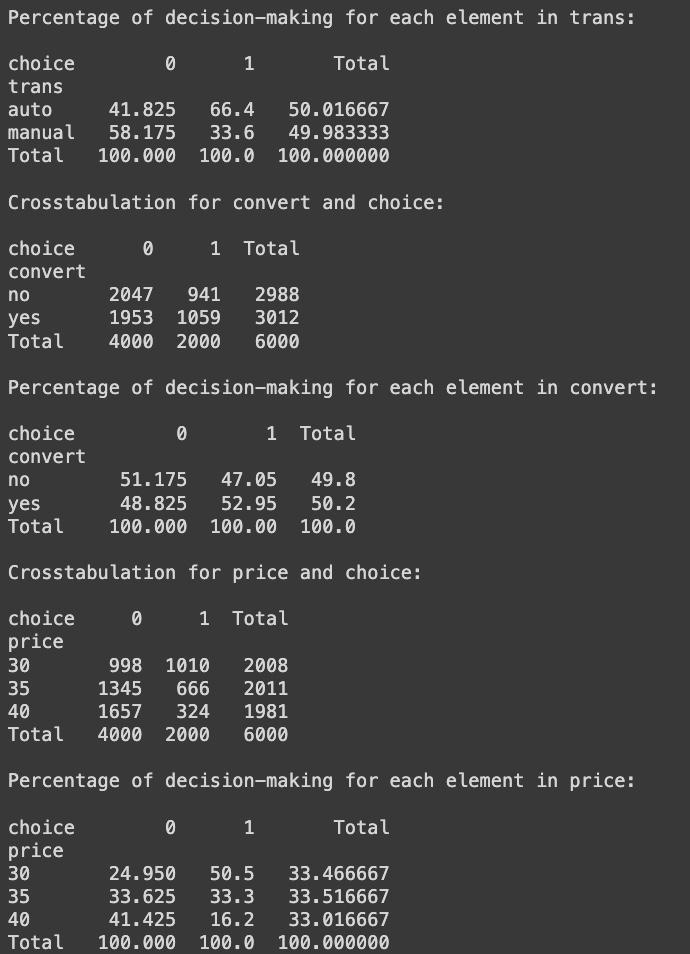
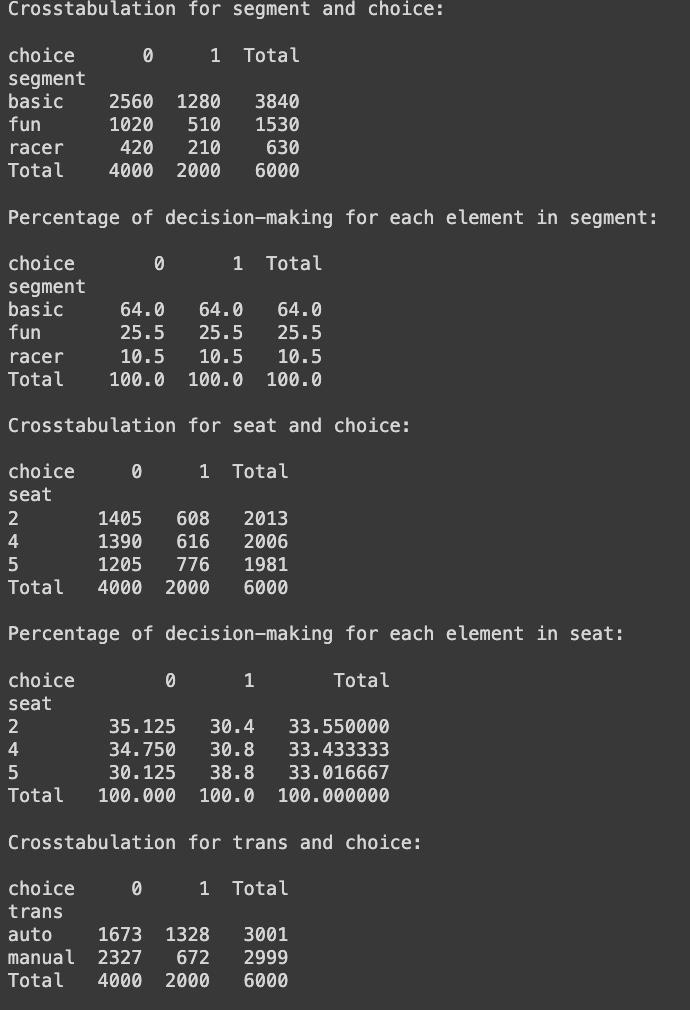


Figure 21: Percentage Decision-Making

Commercial segmentation of a basic car is a key factor influencing whether a car would be chosen or not, carrying the highest weight in terms of choice. However, transmission has the highest weight in influencing choosing a car if it was automatic, but lower in weight when you choose an alternative car.



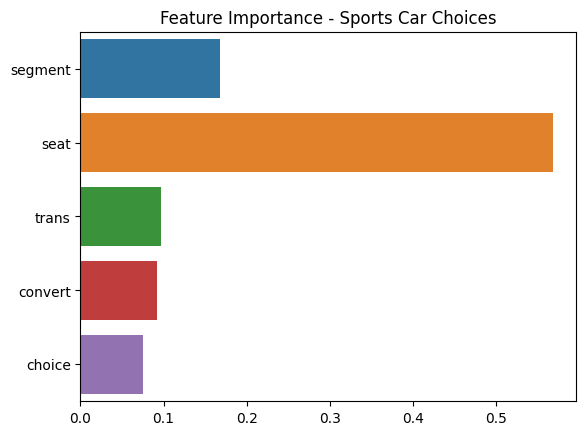
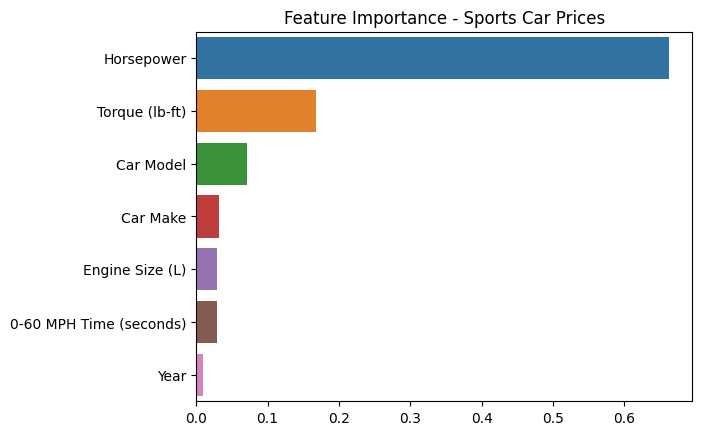


Figure 22: Feature Importance for both datasets

In order to visualize side by side feature importance of the two datasets, we can see that horsepower dominates in the performance and functionality aspects of sports cars. In Sports Car choices, the dataset seat emerges as the highest. The x-axis represents the importance score of each feature in influencing the price in this case. This also suggests that engine power represented through horsepower contributes to determining the price. As expected, the year has the least amount of significance when it comes to the importance score. Seat configuration stands the most surprising, as it influences consumers choices the most when choosing a sports car. The commercial segment of the car follows along with transmission, which were expected outcomes.

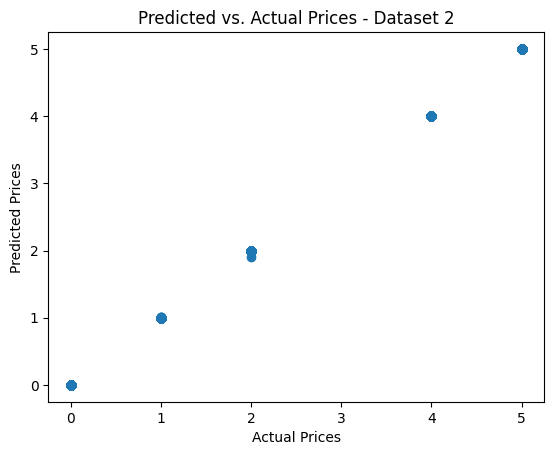
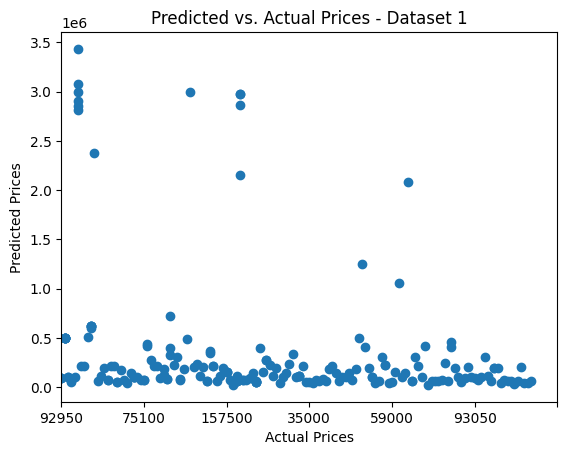


Figure 22 & 23 : Predicted vs Actual Prices for both Dataset 1 & 2

The lack of a clear pattern and scattered points suggest that the model for Dataset 1 may need further investigation and refinement. There are outliers where the model's predictions deviate significantly from the actual prices. The model for Dataset 1 might not be capturing the relationships between features and prices as effectively. The scattered points and wider range of predicted prices indicate that the model in Dataset 1 may be having difficulties making accurate predictions for a broader set of scenarios. The linear pattern in a diagonal line indicates a well-performing model for predicting car prices in Dataset 2. It means that the model's predictions are generally in line with the actual prices. The majority of the data points align along a diagonal line, indicating a close match between predicted and actual prices. There are a few outliers where the model's predictions deviate from the actual prices.

Roles:

In the collaborative effort to analyze the importance and impact of sports car features to the overall price, James and Zannate each played crucial roles, contributing their skills to distinct aspects of the project. James took care of initial data cleaning, feature engineering, and data integration. He also focused on data analysis for his respected features and provided insight on the results found within the data. Zannate focused on cleaning and preprocessing second dataset and exploratory analysis. She also delved into feature modeling to analyze the importance of decision and design aspects. Both actively contributed to the shaping of the presentation and write-up creating a thorough encapsulation of their findings and insights.

Conclusions:

With some skewed data preprocessing methods, our technique worked to perform the designated tasks that we intended in order to highlight key importances when it comes to the value of sports cars. We considered how an individual might choose a sports car and performance features that play a huge role in functionality of sports cars. We expected some outcomes through our hypothesis but cannot fully conclude if that satisfies the conclusion but we noticed the trends. Feature importance in a sports car is very hard to define, since several factors contribute to the overall cost, the importance of these features vary from one individual to another. When creating these models and forming predictions we found a majority result of the functional feature of Horsepower being a common theme. Commercial segmented ‘basic’ model or automatic transmission is desired or considered more important in design factors. However, a surprising discovery showed that the number of seats is a huge percentage of decision-making that is impacted when choosing a sports car. We looked at other features in regards to prices and choices as feature importance when assessing the value of a sports car to see if rarity, popularity, resale value, and/or brand plays a role. In summary, our comprehensive analysis gave us insight for sports car evaluations and highlighted the several defining features for a sports car that influence individuals’ choices and price dynamics. Although we believe that the value of a sports car is varied through differences of opinion, we have reached a consensus through our own research that horsepower is a common trend when it comes to shaping the idea of what makes a sports car a sports car.