James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

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.

```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Car Make               1007 non-null   object
 1   Car Model              1007 non-null   object
 2   Year                   1007 non-null   int64
 3   Engine Size (L)        997 non-null    object
 4   Horsepower             1007 non-null   object
 5   Torque (lb-ft)         1004 non-null   object
 6   0-60 MPH Time (seconds)  1007 non-null   object
 7   Price (in USD)         1007 non-null   object
dtypes: int64(1), object(7)
memory usage: 63.1+ KB
```

	Car Make	Car Model	Year	Engine Size (L)	Horsepower	Torque (lb-ft)	0-60 MPH Time (seconds)	Price (in USD)
0	porsche	911	2022	3	379	331	4	101,200
1	lamborghini	huracan	2021	5.2	630	443	2.8	274,390
2	ferrari	488 gtb	2022	3.9	661	561	3	333,750
3	audi	r8	2022	5.2	562	406	3.2	142,700
4	mclaren	720s	2021	4	710	568	2.7	298,000

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.

```
1 # Lists out the specific rows within the Engine Size (L) column contain value 'NaN'
2 sports_car_df[sports_car_df["Engine Size (L)"].isna()]
3
```

	Car Make	Car Model	Year	Engine Size (L)	Horsepower	Torque (lb-ft)	0-60 MPH Time (seconds)	Price (in USD)
168	rimac	c_two	2022	NaN	1914	1696	1.9	2400000
171	tesla	model s plaid	2021	NaN	1020	1050	1.98	131190
222	porsche	taycan turbo s	2021	NaN	750	774	2.6	185000
247	tesla	model s plaid	2022	NaN	1020	1050	1.9	131190
387	rimac	c_two	2022	NaN	1888	1696	1.8	2400000
389	tesla	roadster	2022	NaN	10000+	0	1.9	200000
686	rimac	c_two	2022	NaN	1914	1696	1.85	2400000
697	lotus	evija	2022	NaN	1972	1254	2.5	2700000
752	porsche	taycan	2022	NaN	469	479	3.8	79900
916	tesla	roadster	2022	NaN	10,000+	NaN	1.9	200000

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

```
#############################################################
# def : assign_engine_type
# parameters : value
# purpose : taking the values within Engine Type (cloned initially from Engine Size (L)) and categorizing it among
# engine types : gas, electric, hybrid
#############################################################
def assign_engine_type(value):

  # if this string is not not found
  if (str(value).find("1.5 + elect") != -1):
    return 'hybrid'

  # search for string 'hybrid'
  elif re.search(r'\bhybrid\b', str(value)):
    return 'hybrid'

  # search for string 'electric'
  elif re.search(r'\belectric\b', str(value)):
    return 'electric'

  # assign remaining "non unique" cases to gas
  else:
    return 'gas'
```

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.

```
bins   = [0.0, 0.99, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 9.0]
labels = [0.0, 1.0,  2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0     ]
labels = ['Electric','0.5-1.4','1.5-2.4', '2.5-3.4', '3.5-4.4','4.5-5.4','5.5-6.4','6.5-7.4','7.5-8.5' ]

sports_car_df['Engine Size Range'] = pd.cut(sports_car_df['Engine Size (L)'], bins = bins, labels = labels, include_lowest = True)
```

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:

```
# mapping of each unique Sports Car Make
car_origin_mapping = {

'acura'         : 'america',
'alfa romeo'    : 'italy',
'alpine'        : 'france',
'ariel'         : 'england',
'aston martin'  : 'england',
'audi'          : 'germany',
'bentley'       : 'england',
'bmw'           : 'germany',
'bugatti'       : 'france',
'chevrolet'     : 'america',
'dodge'         : 'america',
'ferrari'       : 'italy',
'ford'          : 'america',
'jaguar'        : 'england',
'kia'           : 'south korea',
'koenigsegg'    : 'sweden',
'lamborghini'   : 'italy',
'lexus'         : 'japan',
'lotus'         : 'england',
'maserati'      : 'italy',
'mazda'         : 'japan',
'mclaren'       : 'england',
'mercedes-amg'  : 'germany',
'mercedes-benz' : 'germany',
'nissan'        : 'japan',
'pagani'        : 'italy',
'pininfarina'   : 'italy',
'polestar'      : 'china',
'porsche'       : 'germany',
'rimac'         : 'croatia',
'rolls-royce'   : 'england',
'shelby'        : 'america',
'subaru'        : 'japan',
'tesla'         : 'america',
'toyota'        : 'japan',
'tvr'           : 'england',
'ultima'        : 'england',
'w motors'      : 'lebanon',
}

```

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.

```
# Feature Engineering to create a new analysis of the cost per horsepower for the given sports cars
sports_car_df['$ per Horsepower'] = sports_car_df['Price (in USD)'].astype(float) / sports_car_df['Horsepower'].astype(float)
```

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

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

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

```
feature_weight_dict = {
    'Horsepower'              : 0.4,
    'Engine Size (L)'         : 0.3,
    '0-60 MPH Time (seconds)' : 0.2,
    'Torque (lb-ft)'          : 0.1
    }

valued_sports_car_df = numeric_features_df.apply(pd.to_numeric, errors = 'coerce')
```

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.

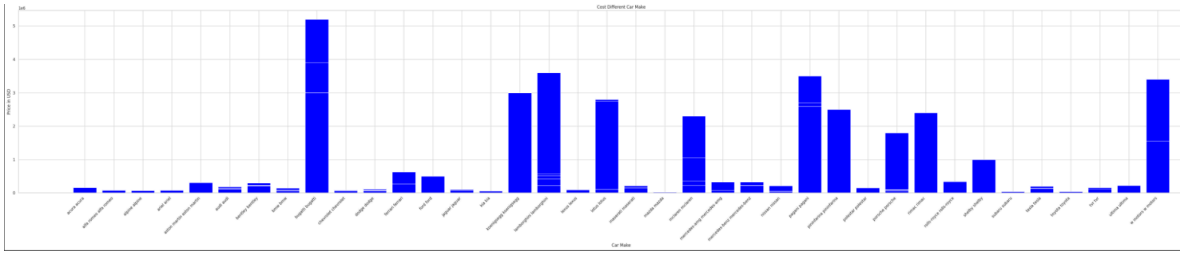


Figure 3: Car Make vs Price

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

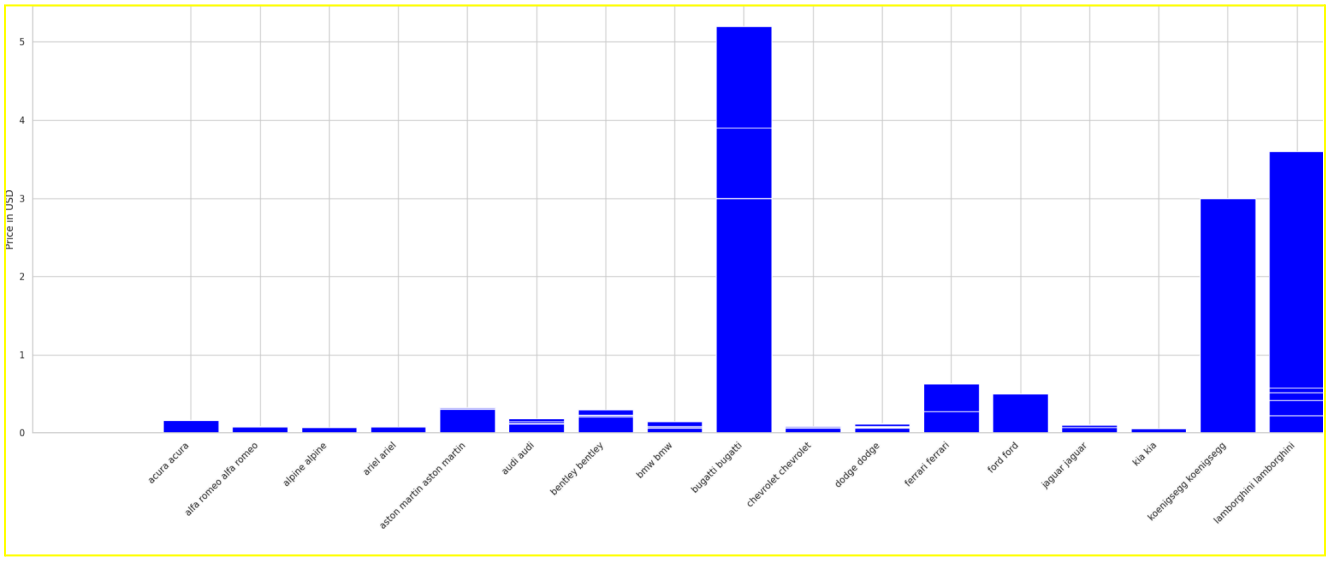


Figure 3a: Car Make vs Price (zoomed in section 1)

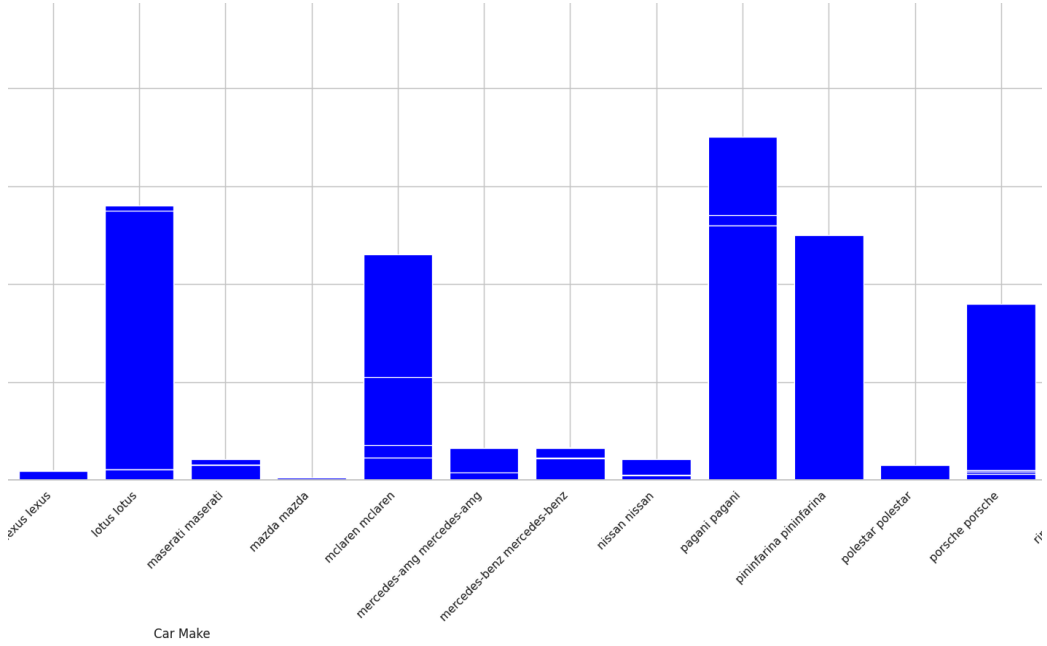


Figure 3b: Car Make vs Price (zoomed in section 2)

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

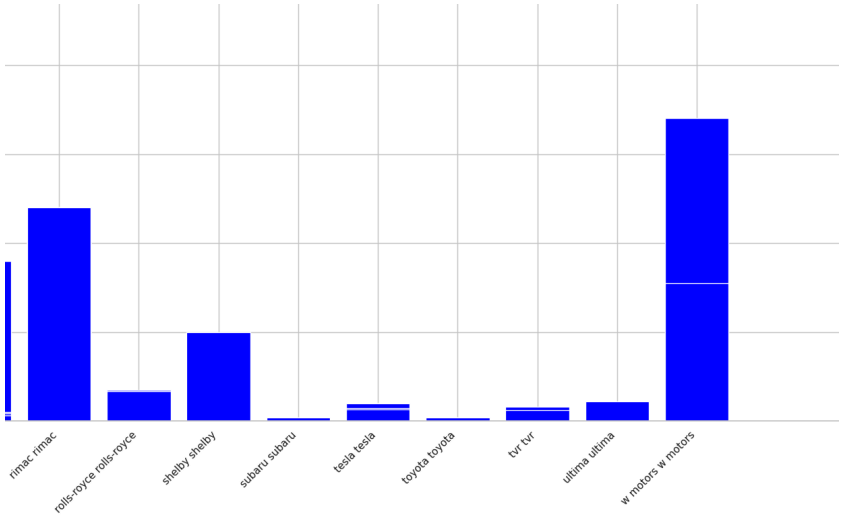


Figure 3c: Car Make vs Price (zoomed in section 3)

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.

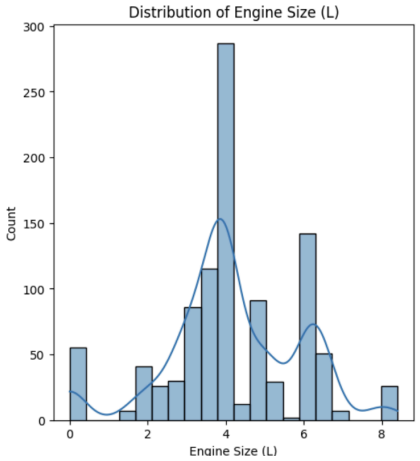


Figure 4: Distribution of Engine Size (L)

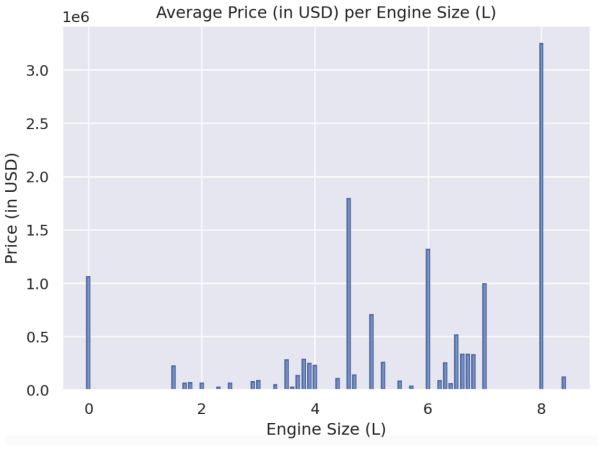


Figure 5: Engine Size (L) vs Price

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

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.

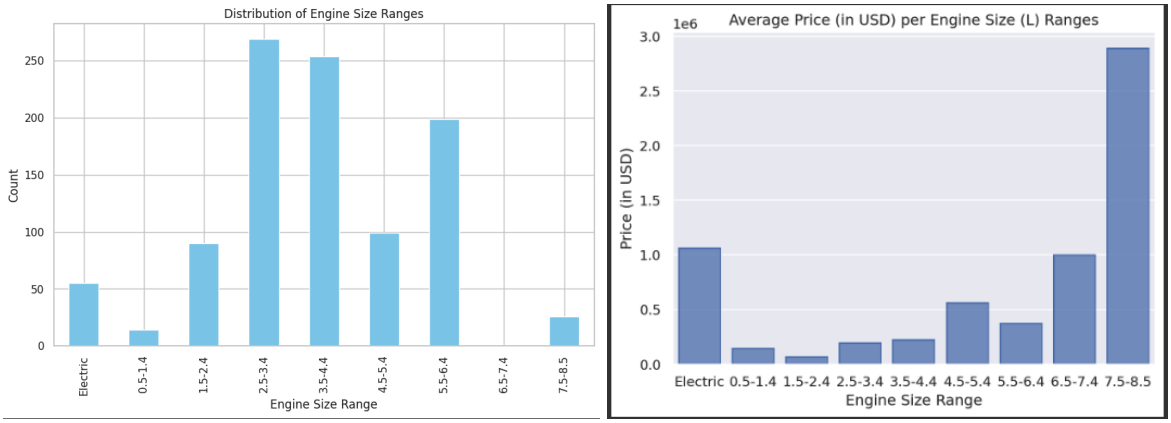


Figure 6 & 7 : Distribution of Engine Size (L) & Engine Size (L) vs Price

```
Total Count of Each Engine Size based on Ranges
2.5-3.4     269
3.5-4.4     254
5.5-6.4     199
4.5-5.4      99
1.5-2.4      90
Electric     55
7.5-8.5      26
0.5-1.4      14
6.5-7.4       1
```

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

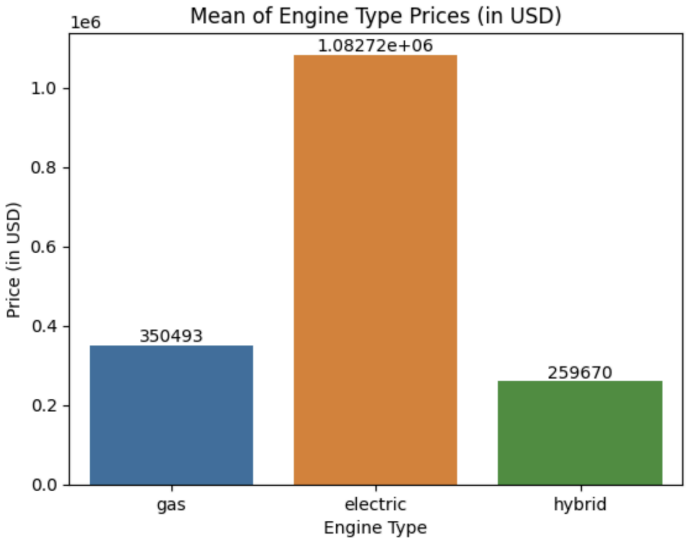


Figure 9: Average Prices of each Engine Type (gas, electric, hybrid)

```
Total Count of Each Engine Type
gas          958
electric      44
hybrid         5
```

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

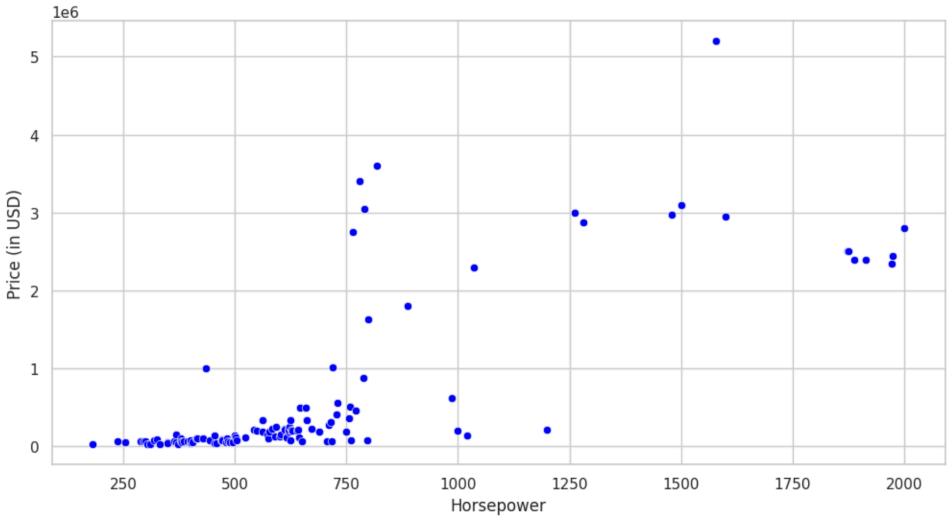


Figure 10: Horsepower vs Price (in USD)

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

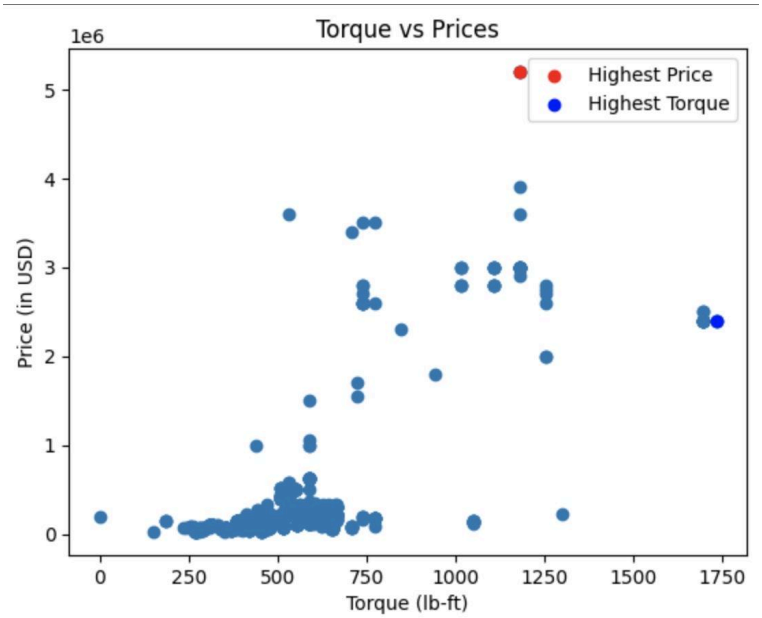


Figure 11: Torque (lb-ft) vs Price (in USD)

```
Highest Torque vs Prices Values:
    Car Make Car Model  Price (in USD)  Torque (lb-ft)
278    Rimac     C_Two         2400000          1732.0
439    Rimac     C_Two         2400000          1732.0
26     Rimac    Nevera         2400000          1696.0
97     Rimac    Nevera         2400000          1696.0
168    Rimac     C_Two         2400000          1696.0
     Price (in USD)  Torque (lb-ft)
278         2400000          1732.0
439         2400000          1732.0
26          2400000          1696.0
97          2400000          1696.0
168         2400000          1696.0
```

```
Car with the Highest Price:
Car Make                                Bugatti
Car Model                 Chiron Super Sport 300+
Year                                         2022
Engine Size (L)                                 8
Horsepower                                   1578
Torque (lb-ft)                             1180.0
0-60 MPH Time (seconds)                       2.3
Price (in USD)                            5200000
Name: 541, dtype: object

Car with the Highest Torque:
Car Make                                  Rimac
Car Model                                 C_Two
Year                                       2022
Engine Size (L)                         Electric
Horsepower                                 1914
Torque (lb-ft)                           1732.0
0-60 MPH Time (seconds)                    1.85
Price (in USD)                          2400000
Name: 278, dtype: object
```

```
Highest Prices vs Torque Values:
       Car Make                Car Model  Price (in USD)  Torque (lb-ft)
541     Bugatti  Chiron Super Sport 300+         5200000          1180.0
823     Bugatti  Chiron Super Sport 300+         5200000          1180.0
983     Bugatti                  Chiron         3900000          1180.0
438  Lamborghini                    Sián         3600000           531.0
624     Bugatti          Chiron Pur Sport         3599000          1180.0
     Price (in USD)  Torque (lb-ft)
541         5200000          1180.0
823         5200000          1180.0
983         3900000          1180.0
438         3600000           531.0
624         3599000          1180.0
```

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

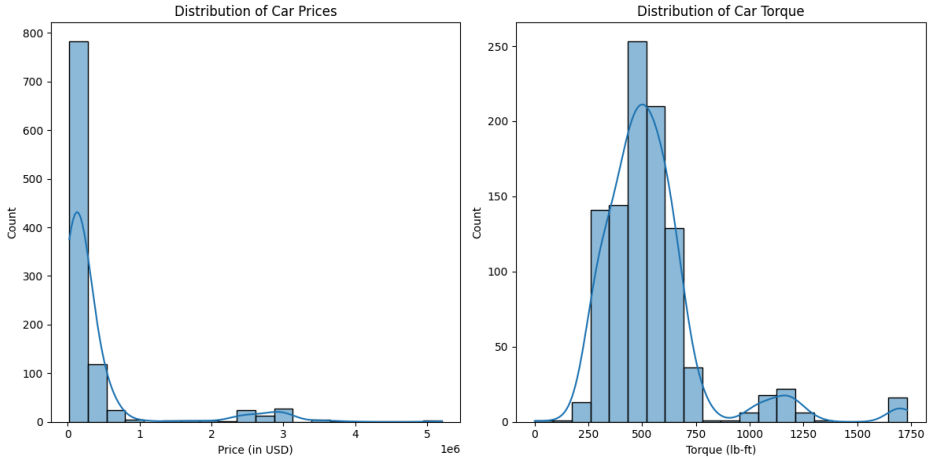


Figure 12: Distribution plot of sports car prices in relation to the distribution of torque

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.

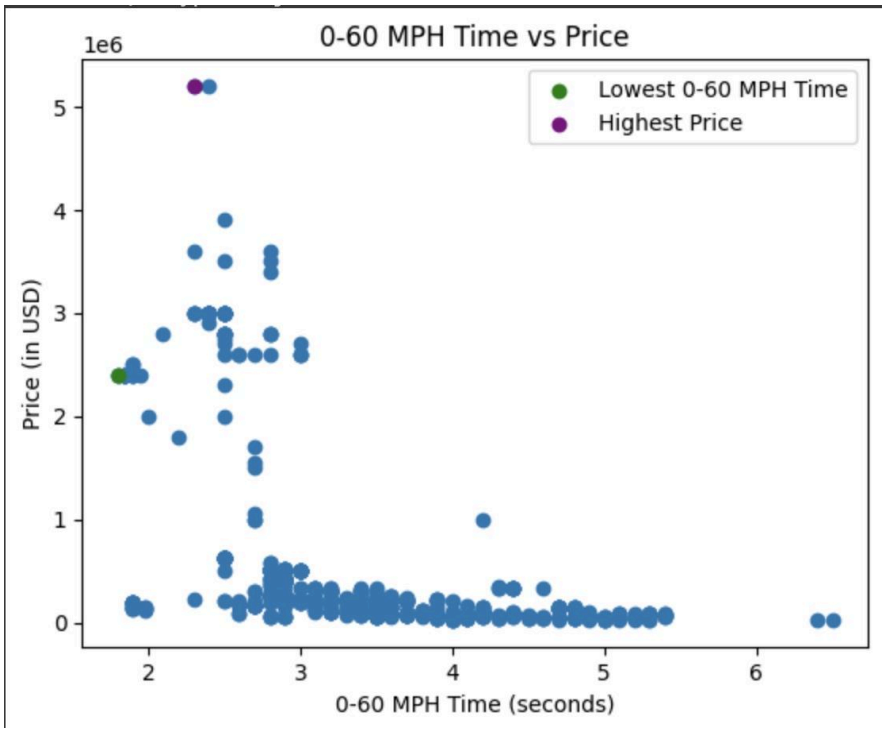


Figure 13: 0-60 MPH Time vs Price

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

```
Highest Price vs 0-60 MPH Time Values:
     Car Make            Car Model  0-60 MPH Time (seconds)
541   Bugatti  Chiron Super Sport 300+                    2.3
823   Bugatti  Chiron Super Sport 300+                    2.4
983   Bugatti              Chiron                        2.5
438  Lamborghini            Sián                        2.8
624   Bugatti         Chiron Pur Sport                    2.3

    Price (in USD)
541        5200000
823        5200000
983        3900000
438        3600000
624        3599000
     0-60 MPH Time (seconds)  Price (in USD)
541                      2.3        5200000
823                      2.4        5200000
983                      2.5        3900000
438                      2.8        3600000
624                      2.3        3599000
```

```
Lowest 0-60 MPH Time vs Highest Price Values:
    Car Make Car Model  0-60 MPH Time (seconds)  Price (in USD)
387    Rimac     C_Two                      1.80        2400000
439    Rimac     C_Two                      1.80        2400000
26     Rimac    Nevera                      1.85        2400000
278    Rimac     C_Two                      1.85        2400000
352    Rimac    Nevera                      1.85        2400000
     0-60 MPH Time (seconds)  Price (in USD)
387                     1.80        2400000
439                     1.80        2400000
26                      1.85        2400000
278                     1.85        2400000
352                     1.85        2400000
```

```
Car with the Lowest 0-60 MPH Time:
Car Make                    Rimac
Car Model                   C_Two
Year                         2022
Engine Size (L)               NaN
Horsepower                   1888
Torque (lb-ft)               1696
0-60 MPH Time (seconds)       1.8
Price (in USD)            2400000
Name: 387, dtype: object

Car with the Highest Price:
Car Make                                 Bugatti
Car Model                Chiron Super Sport 300+
Year                                        2022
Engine Size (L)                                8
Horsepower                                  1578
Torque (lb-ft)                              1180
0-60 MPH Time (seconds)                      2.3
Price (in USD)                           5200000
Name: 541, dtype: object
```

```
'0-60 MPH Time' and 'Price' Categories:
      0-60 MPH Time (seconds)  Price (in USD)
0                        4.00          101200
1                        2.80          274390
2                        3.00          333750
3                        3.20          142700
4                        2.70          298000
...                       ...             ...
1002                     2.50         3000000
1003                     2.00         2000000
1004                     2.70         1000000
1005                     3.00         2600000
1006                     1.85         2400000
```

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.

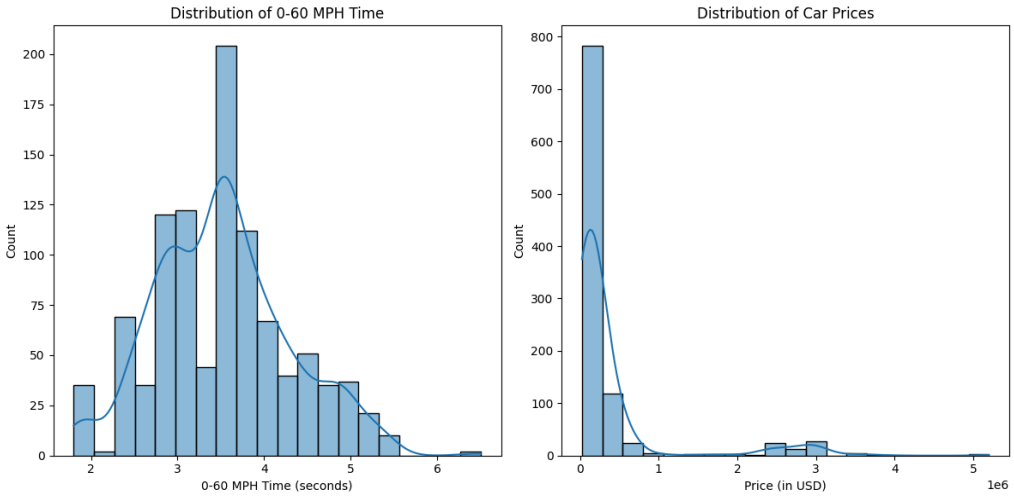


Figure 14 & 15 : Distribution 0-60 MPH Time & Distribution of Car Prices

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

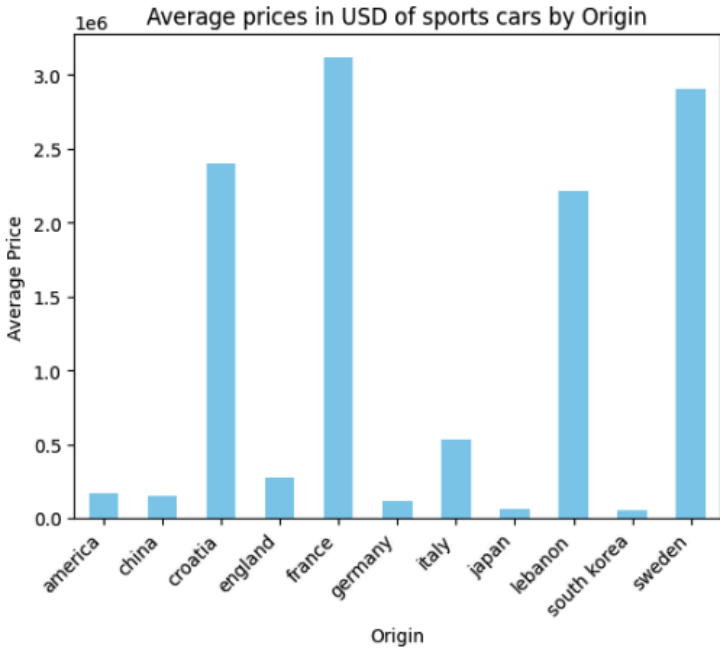


Figure 16 : Average Prices (in USD) based on Origin (Country)

```
Total Count of Origins
germany        287
england        229
america        185
italy          176
japan           72
france          24
sweden          15
croatia         14
lebanon          3
china            1
south korea      1
Name: Origin, dtype: int64


Average Prices of Sports Cars By Origin:
Origin
america        165096
china          155000
croatia       2400000
england        273701
france        3119438
germany        117987
italy          534350
japan           64701
lebanon       2216667
south korea     52200
sweden        2906667
```

	Car Make	Origin	Price (in USD)
541	bugatti	france	5200000
823	bugatti	france	5200000
983	bugatti	france	3900000
438	lamborghini	italy	3600000
624	bugatti	france	3599000
279	pagani	italy	3500000
385	pagani	italy	3500000
174	w motors	lebanon	3400000
11	bugatti	france	3000000
85	bugatti	france	3000000
88	koenigsegg	sweden	3000000
113	bugatti	france	3000000
158	bugatti	france	3000000
161	koenigsegg	sweden	3000000
206	bugatti	france	3000000
275	koenigsegg	sweden	3000000
328	koenigsegg	sweden	3000000
341	bugatti	france	3000000
376	bugatti	france	3000000
434	bugatti	france	3000000
435	koenigsegg	sweden	3000000

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

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.

Highest Scored Sports Car Based of theoretical Weighted Features:

	Car Make	Car Model	Year	Engine Size (L)	Horsepower	Torque (lb-ft)	0-60 MPH Time (seconds)	Price (in USD)	Score
885	tesla	roadster	2022	0	10000	7376	1.9	200000	4738
389	tesla	roadster	2022	0	10000	0	1.9	200000	4000
354	tesla	roadster	2022	0	1000	10000	1.9	200000	1400
278	rimac	c_two	2022	0	1914	1732	1.85	2400000	939
439	rimac	c_two	2021	0	1914	1732	1.8	2400000	939
97	rimac	nevera	2022	0	1914	1696	1.95	2400000	936
168	rimac	c_two	2022	0	1914	1696	1.9	2400000	936
509	rimac	c_two	2021	0	1914	1696	1.9	2400000	936
526	rimac	c_two	2022	0	1914	1696	1.9	2400000	936
640	rimac	nevera	2021	0	1914	1696	1.9	2400000	936
26	rimac	nevera	2022	0	1914	1696	1.85	2400000	936
352	rimac	nevera	2022	0	1914	1696	1.85	2400000	936
686	rimac	c_two	2022	0	1914	1696	1.85	2400000	936
824	rimac	nevera	2021	0	1914	1696	1.85	2400000	936
986	rimac	nevera	2022	0	1914	1696	1.85	2400000	936
877	lotus	evija	2021	0	2000	1254	2.8	2800000	926
1006	rimac	nevera	2021	0	1888	1696	1.85	2400000	925
387	rimac	c_two	2022	0	1888	1696	1.8	2400000	925
280	pininfarina	battista	2022	0	1874	1696	1.9	2500000	920
988	pininfarina	battista	2021	0	1872	1696	1.9	2500000	919
420	lotus	evija	2022	0	1973	1254	2.5	2750000	915
523	lotus	evija	2022	0	1973	1254	2.5	2600000	915
987	lotus	evija	2022	0	1973	1254	2.5	2000000	915
697	lotus	evija	2022	0	1972	1254	2.5	2700000	915
1003	lotus	evija	2021	0	1972	1254	2	2000000	915
88	koenigsegg	jesko	2022	5	1600	1106	2.5	3000000	753
161	koenigsegg	jesko	2022	5	1600	1106	2.5	3000000	753
822	koenigsegg	jesko	2022	5	1600	1106	2.5	3000000	753
418	koenigsegg	jesko absolut	2022	5	1600	1106	2.1	2800000	753
823	bugatti	chiron super sport 300+	2021	8	1578	1180	2.4	5200000	752
541	bugatti	chiron super sport 300+	2022	8	1578	1180	2.3	5200000	752
631	bugatti	chiron	2021	8	1500	1180	2.5	3000000	721
983	bugatti	chiron	2022	8	1500	1180	2.5	3900000	721
11	bugatti	chiron	2021	8	1500	1180	2.4	3000000	721
85	bugatti	chiron	2022	8	1500	1180	2.4	3000000	721

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 = $X_{n,1} * 0.4 + X_{n,2} * 0.3 + X_{n,3} * 0.2 + X_{n,4} * 0.1$

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.

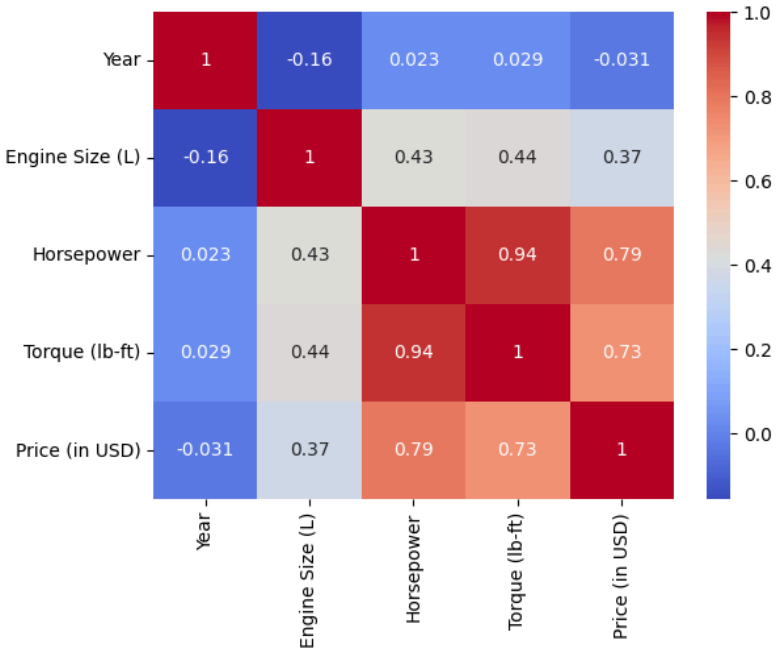


Figure 18: Heatmap for Feature Correlation

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

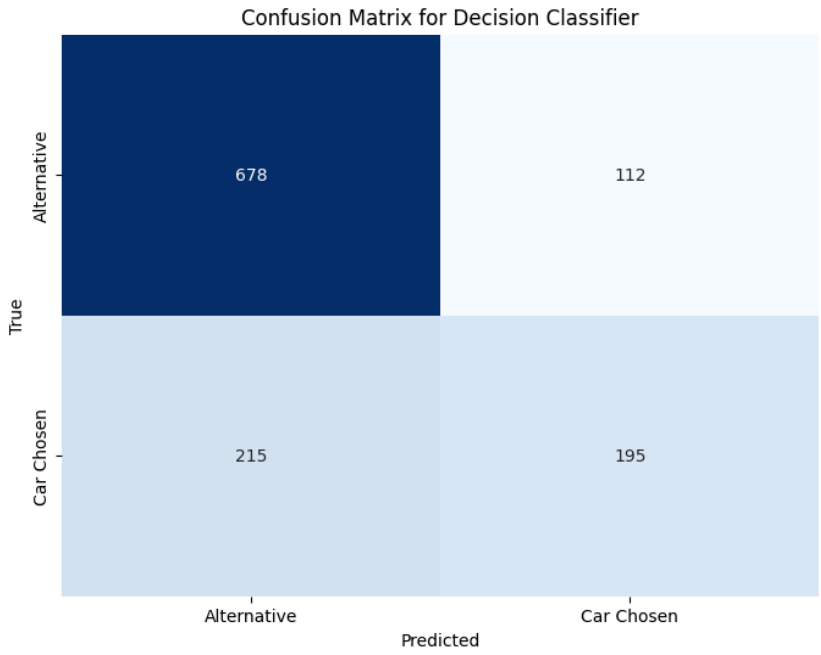
followed by higher torque.

	resp_id	ques	alt	segment	seat	trans	convert	price	choice
0	1	1	1	basic	2	manual	yes	35	0
1	1	1	2	basic	5	auto	no	40	0
2	1	1	3	basic	5	auto	no	30	1
3	1	2	1	basic	5	manual	no	35	0
4	1	2	2	basic	2	manual	no	30	1

Field	Description
resp_id	The identifier of each individual in the dataset
ques	The identifier of each specific purchase scenario
alt	The identifier of each alternative choice within a question
segment	The commercial segment of a sportscar model ('basic', 'fun', 'racer')
seat	The number of seats in the vehicle (2, 4, 5)
trans	The transmission type of the vehicle ('auto','manual')
convert	Whether or not the vehicle has a convertible top
price	The sportscar price (in thousands/$)
choice	Dummy indicator of the decision made. (1 = car chosen, 0 = alternative cars chosen from)

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.



James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

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.

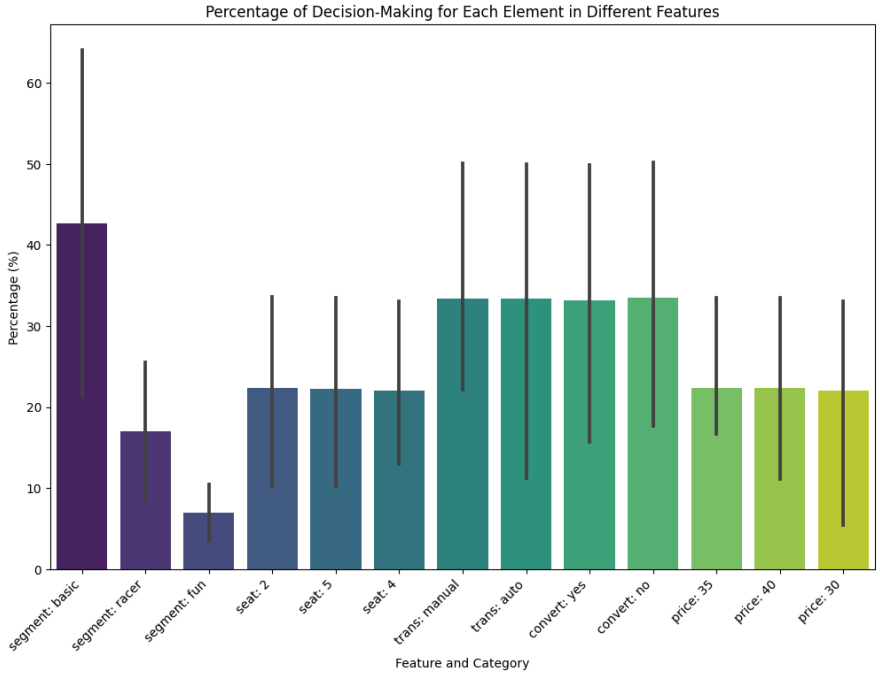


Figure 21: Percentage Decision-Making

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.

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

```
Crosstabulation for segment and choice:

choice      0     1  Total
segment
basic    2560  1280   3840
fun      1020   510   1530
racer     420   210    630
Total    4000  2000   6000

Percentage of decision-making for each element in segment:

choice       0      1  Total
segment
basic     64.0   64.0   64.0
fun       25.5   25.5   25.5
racer     10.5   10.5   10.5
Total    100.0  100.0  100.0

Crosstabulation for seat and choice:

choice     0     1  Total
seat
2       1405   608   2013
4       1390   616   2006
5       1205   776   1981
Total   4000  2000   6000

Percentage of decision-making for each element in seat:

choice        0      1       Total
seat
2        35.125   30.4   33.550000
4        34.750   30.8   33.433333
5        30.125   38.8   33.016667
Total   100.000  100.0  100.000000

Crosstabulation for trans and choice:

choice     0     1  Total
trans
auto    1673  1328   3001
manual  2327   672   2999
Total   4000  2000   6000
```

```
Percentage of decision-making for each element in trans:

choice        0      1       Total
trans
auto     41.825   66.4   50.016667
manual   58.175   33.6   49.983333
Total   100.000  100.0  100.000000

Crosstabulation for convert and choice:

choice      0     1  Total
convert
no       2047   941   2988
yes      1953  1059   3012
Total    4000  2000   6000

Percentage of decision-making for each element in convert:

choice         0       1  Total
convert
no       51.175   47.05   49.8
yes      48.825   52.95   50.2
Total   100.000  100.00  100.0

Crosstabulation for price and choice:

choice     0     1  Total
price
30       998  1010   2008
35      1345   666   2011
40      1657   324   1981
Total   4000  2000   6000

Percentage of decision-making for each element in price:

choice        0      1       Total
price
30       24.950   50.5   33.466667
35       33.625   33.3   33.516667
40       41.425   16.2   33.016667
Total   100.000  100.0  100.000000
```

```
Crosstabulation for segment and price:

price      30    35    40  Total
segment
basic    1288  1280  1272   3840
fun       514   520   496   1530
racer     206   211   213    630
Total    2008  2011  1981   6000

Feature preference for each element in segment and price:

price             30           35           40  Total
segment
basic     64.143426   63.649925   64.209995   64.0
fun       25.597610   25.857782   25.037860   25.5
racer     10.258964   10.492292   10.752145   10.5
Total    100.000000  100.000000  100.000000  100.0

Crosstabulation for seat and price:

price    30    35    40  Total
seat
2       667   668   678   2013
4       672   674   660   2006
5       669   669   643   1981
Total  2008  2011  1981   6000
```

```
no         993  1012   983   2988
yes       1015   999   998   3012
Total     2008  2011  1981   6000

Feature preference for each element in convert and price:

price             30           35           40  Total
convert
no        49.452191   50.323222   49.621403   49.8
yes       50.547809   49.676778   50.378597   50.2
Total    100.000000  100.000000  100.000000  100.0

Crosstabulation for choice and price:

price     30    35    40  Total
choice
0        998  1345  1657   4000
1       1010   666   324   2000
Total   2008  2011  1981   6000

Feature preference for each element in choice and price:

price            30           35           40        Total
choice
0         49.701195   66.882148   83.644624   66.666667
1         50.298805   33.117852   16.355376   33.333333
Total    100.000000  100.000000  100.000000  100.000000
```

James David
Zannate Malik
ENPM 808W-DE01
Dec 9, 2023

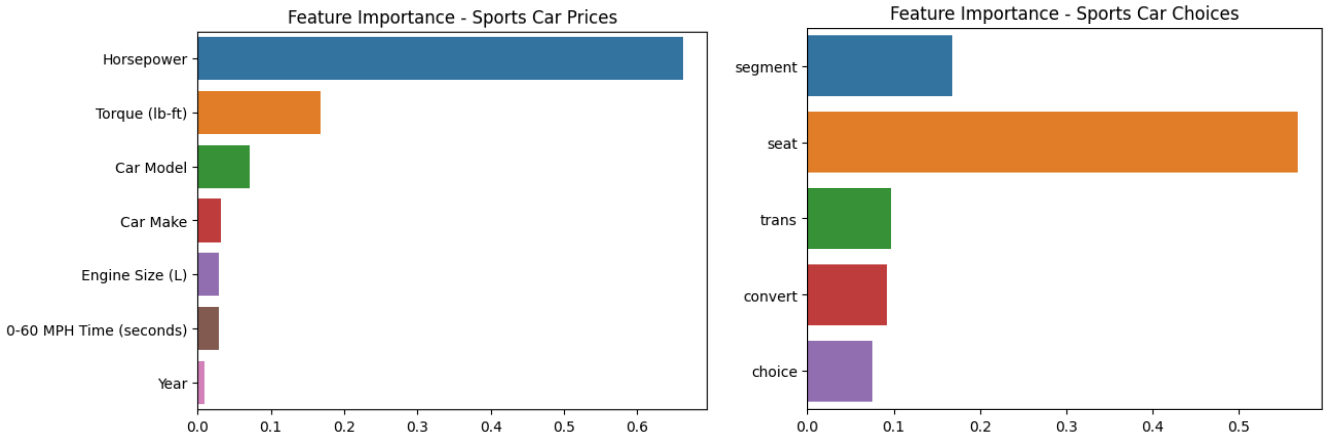


Figure 22: Feature Importance for both datasets

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.

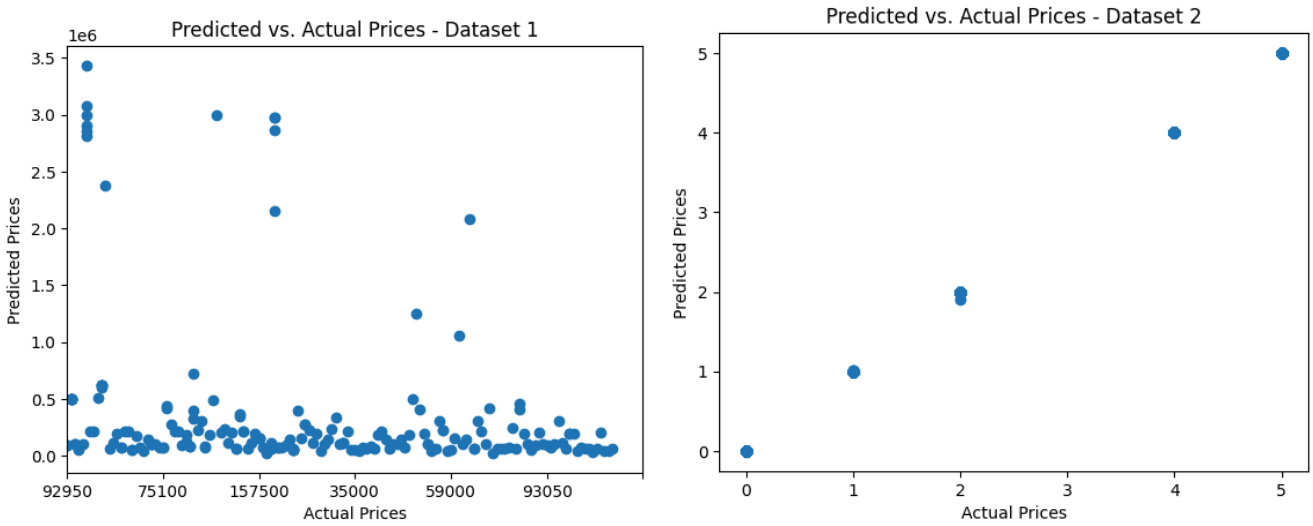


Figure 22 & 23 : Predicted vs Actual Prices for both Dataset 1 & 2

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.