Greg Fagan, Luke Rubin, Max Enabnit

Project Report

Features that Drive the Prices of Sports Cars

1. Introduction

Sports cars have long been admired for their speed, design, and exclusivity. But what factors determine their often high price tags? Beyond their appearance and brand reputation, specific measurable features play a key role in shaping their value. How do factors like mileage and year of manufacture influence prices? Which features are the most important? Which manufacturers dominate the luxury market, and which offer more budget-friendly options?

This project explores data from Kaggle's Sports Car Price Dataset and additional information gathered from CarMax. By analyzing variables such as engine size, horsepower, acceleration, mileage, make, and year, this project uncovers patterns and relationships that impact the pricing of sports cars. The goal is to provide insights into the luxury car market, highlight pricing factors, and compare manufacturers across different price ranges.

1. KaggleDataSet
2. CarMax

2. Data

In this project we used two data sources: Kaggle's Sports Car Price Dataset[1] that includes types of cars and their information and sales price, as well as Carmax.com[2], which includes current car listings of various cars and is frequently updated.

2.1 Sports Car Price Dataset

The first data source we worked with was Kaggle's Sports Car Price Dataset. This dataset contains detailed information about the prices of various sports cars from different manufacturers, including key attributes like engine size, horsepower, acceleration time, and the car's year of manufacture. It is designed for analyzing sports car prices and identifying market trends, making it ideal for our analysis.

For consistency, we filtered out electric cars and focused exclusively on traditional combustion engine vehicles. After filtering this, we were left with 947 sports cars and the data associated with them. When cleaning the data, we removed the column "Torque" and commas, changed the column titles to easier to understand names in snake case, and modified the data types of certain columns to make the data easier to work with.

2.2 CarMax

The website CarMax offers detailed information on cars for sale. In addition to including basic information about the car's make, model, and year, the website also includes the mileage of each car.

We primarily scraped this dataset to gain additional car data to add to the data frame to provide better insights about car statistics. CarMax includes many of the same variables as our Kaggle dataset, providing additional data to enhance the accuracy of our overall model. This data was found as advertisements for each individual car, with filtered results to only "Sports Cars". We created a web crawling code to scrape the first 1500 instances of data into a pandas data frame to continue our research with it. Rather than traditional by-page listings, the website includes a "See More Matches" button to expand the number of cars displayed. We had to consider this during scraping and were able to scrape the data with our code utilizing this button.

When cleaning this dataset, we normalized each column to create a seamless join with the Kaggle data. In the car price column, we standardized the data by removing symbols such as "*" and "$." For the car mileage column, we converted shorthand values like "8k" into full numbers (e.g., "8k" became "8000") and removed any commas for consistency. Lastly, we dropped the car trim column, as it did not add meaningful value to our project.

2.3 Combining the Data

Because we were looking to expand our number of cars totals, we combined the data sets by concatenating the values. However, the Kaggle dataset did not have the mileage total that was included in the CarMax dataset. To remove NaN values, we imputed the average mileage of the matching year from the remaining cars in the dataset. Additionally, engine size, horsepower, and acceleration time were not included in the CarMax dataset. To avoid including NaN values in this section, we imputed the average of each variable for each model. We then imputed the average of each variable for each matching make to provide values for car models that had not previously been included in the Kaggle portion of the dataset. Finally, we removed the remaining NaN values. A description of each of the variables we used can be found in Table 1 below.

Table 1 Data Dictionary for Combined Data sets

Column	Type	Source	Description
car_make	Text	Both	Brand or company that produced car
car_model	Text	Both	Model of sports car
year	Text	Both	Year of production of car
engine_size	Numeric	Kaggle	Size of engine in liters (L)
horsepower	Numeric	Kaggle	Horsepower (power output) of car's engine
acceleration_time	Numeric	Kaggle	Amount of time it takes (in seconds) for the car to accelerate from 0-60MPH
mileage	Numeric	CarMax	Mileage on car
price	Numeric	Both	Listing price of car

3. Analysis

3.1 Most Influential Features on Price

We set out to determine which car features most influence sales price. To do this, we calculated correlation coefficients for each feature and visualized them on a bar graph. Figure 1 shows that horsepower and acceleration time have the greatest impact, with acceleration's negative correlation reflecting lower acceleration times increase value. Year and engine size also have a moderate effect, as expected. Surprisingly, mileage shows only a weak negative correlation, suggesting that higher mileage likely affects the value of luxury and sports cars less than regular vehicles.

Figure 1

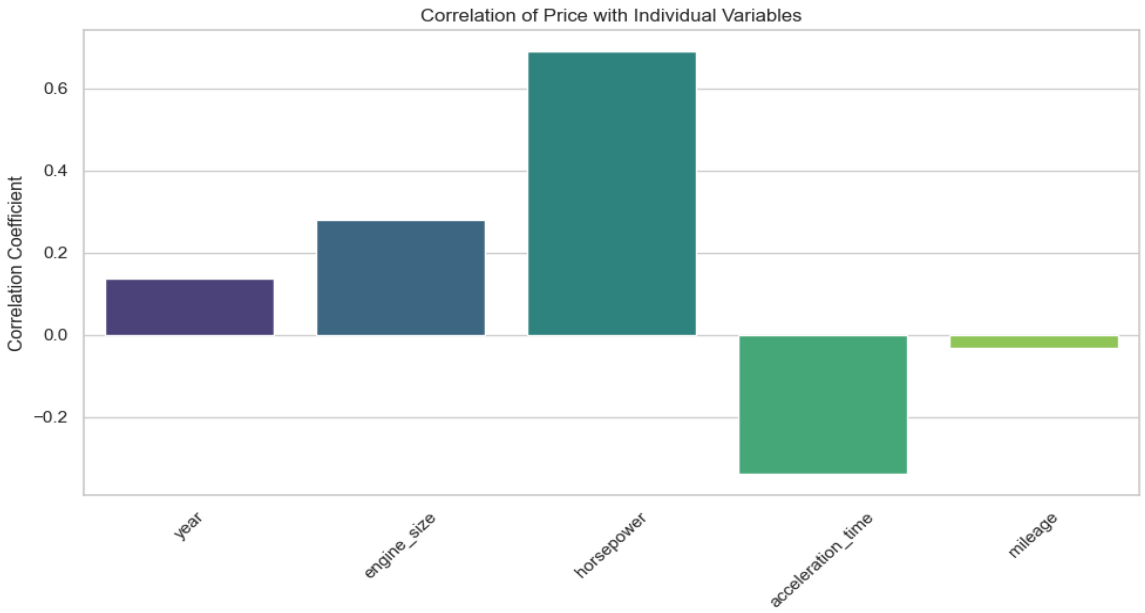
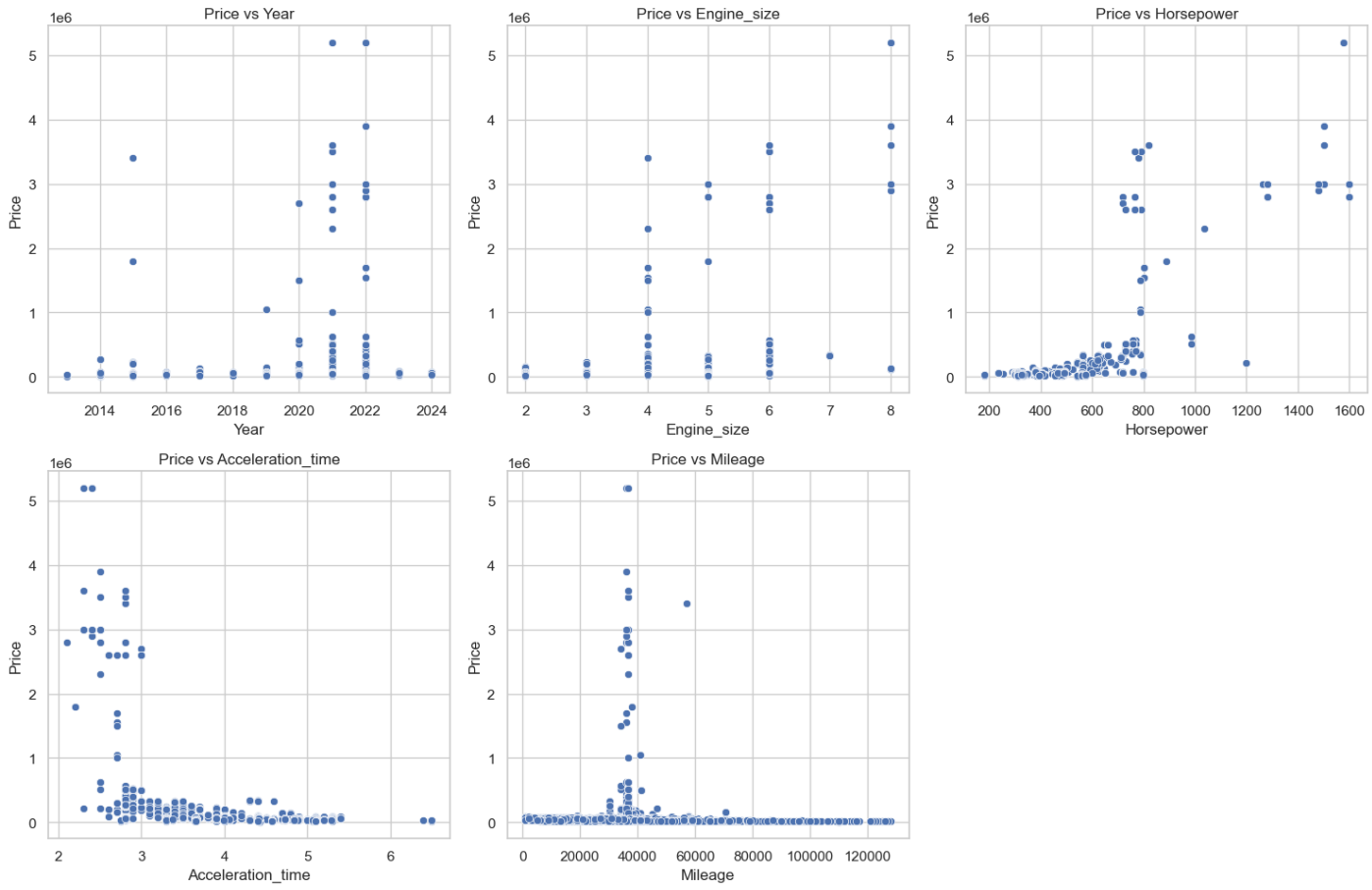


Figure 2 provides a more detailed analysis of the factors influencing the correlation coefficients. Each scatter plot compares one feature to price, revealing clear trends. In the Price vs. Horsepower plot, we observe a significant increase in price when horsepower exceeds 800, with only a few outliers dipping in price beyond this benchmark. Similarly, in Price vs. Acceleration Time, prices sharply rise when acceleration drops below three seconds. The Price vs. Year and Price vs. Engine Size plots align with expectations, showing a consistent upward trend in price as engine size and year increase. However, Price vs. Mileage yields again unusual results, with mileage appearing to have little effect on car value, except for a few outliers, likely representing lower-end sports cars.

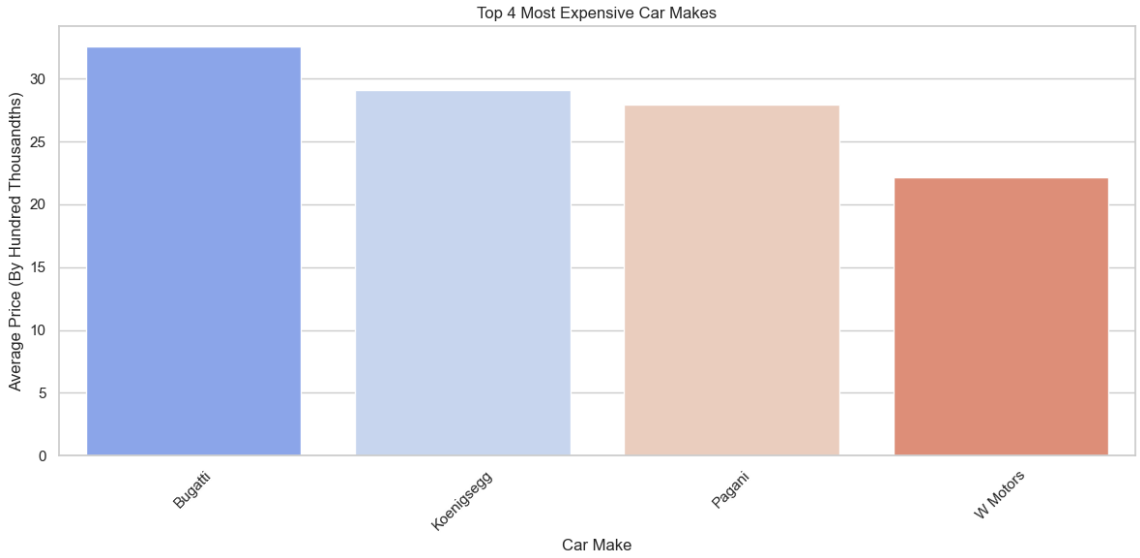
Figure 2



3.2 Least & Most Expensive Car makes

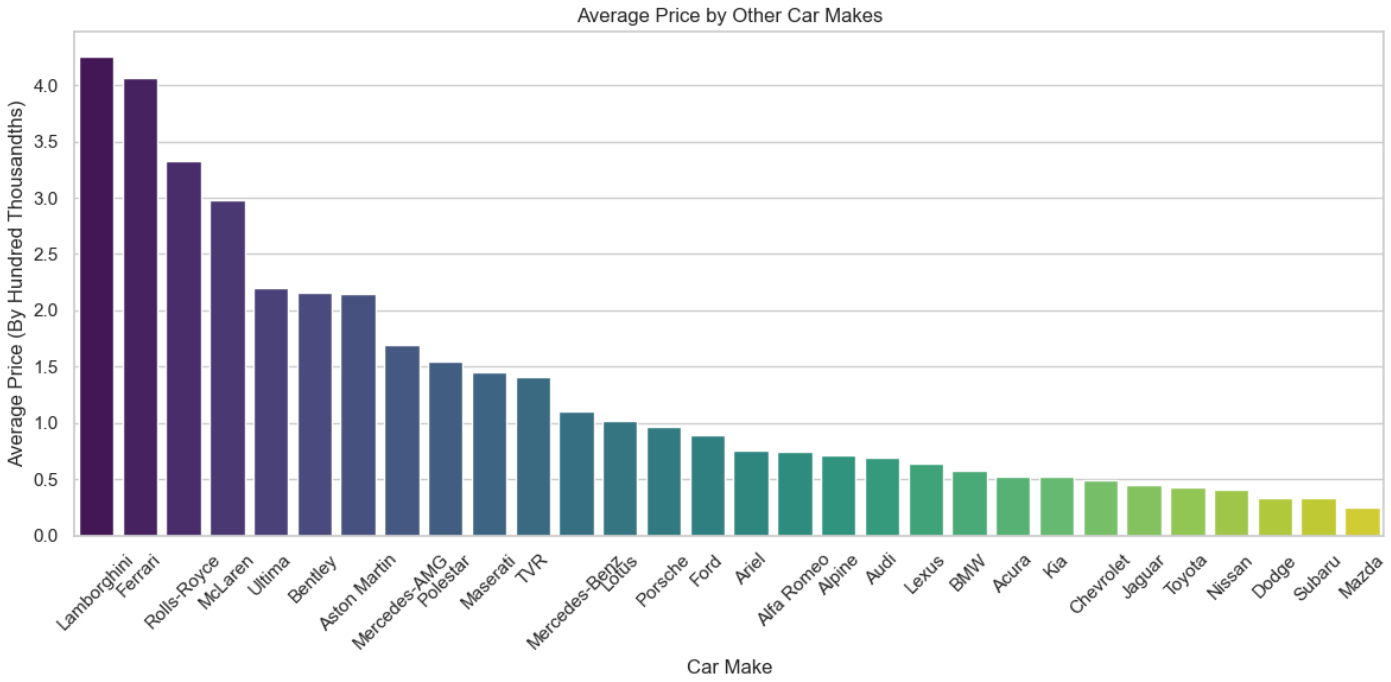
The aim here was to identify the least and most expensive car makes in the dataset. Initially, a single bar chart was created displaying all the makes, but the top four were significantly more expensive, which caused the chart to be overly skewed and difficult to interpret. To address this, we split it into two separate bar charts for clearer comparisons. Figure 3 highlights the top four makes—Bugatti, Koenigsegg, Pagani, and W Motors—whose cars prices average in the millions of dollars, solidifying their position as the most expensive brands.

Figure 3



In contrast, Figure 4 focuses on the remaining car makes, allowing for comparisons among practical consumer vehicles and cheaper sports cars. This visualization shows that Lamborghini, Ferrari, Rolls-Royce, and McLaren lead with the highest average prices among these makes. Meanwhile, Nissan, Dodge, Subaru, and Mazda rank at the bottom, representing the least expensive makes on average.

Figure 4



3.3 Price Distribution of Most Common Makes

In this section, we analyzed the price distributions of different car makes to identify key patterns among popular brands. Each figure highlights distinct characteristics that reflect varying market strategies and consumer targets.

Figure 5: BMW Price Distribution

The price distribution for BMW is right-skewed, with most cars priced between $20,000 and $70,000. A smaller proportion of vehicles exceed $100,000, representing the brand's high-end and specialty models. This range reflects BMW's focus on catering to both premium and ultra-luxury buyers.

Figure 6: Ford Price Distribution

Ford's price distribution is highly right-skewed, with most vehicles priced below $30,000. A few outliers above $100,000 likely represent limited-edition or performance models, while the core range highlights Ford's emphasis on affordability and mass-market appeal.

Figure 7: Chevrolet Price Distribution

Chevrolet's prices are concentrated between $20,000 and $40,000, with a sharp decline beyond this range. This suggests a focus on accessible models, aligning with the brand's reputation for delivering practical, cost-effective vehicles.

Figure 8: Mazda Price Distribution

Mazda's price distribution shows a more balanced shape, with most vehicles priced between $20,000 and $35,000. Prices taper off gradually for higher-end models, indicating Mazda's focus on affordability.

Figure 9: Chevrolet's Bimodal Price Pattern

Chevrolet's distribution has a bimodal pattern, with two peaks around $20,000 and $50,000. This reflects two primary market segments: entry-level models and higher-tier trucks or performance vehicles, showing perhaps a dual-market strategy.

The price distributions reveal different strategies across brands. BMW stands out with a wide price range targeting both premium and ultra-luxury markets. Ford and Chevrolet emphasize affordability while including higher-end options for niche buyers. Mazda maintains a narrower, balanced range to appeal to buyers seeking value and refinement. Chevrolet's unique bimodal distribution highlights its appeal across distinct buyer groups.

Figure 5

Figure 6

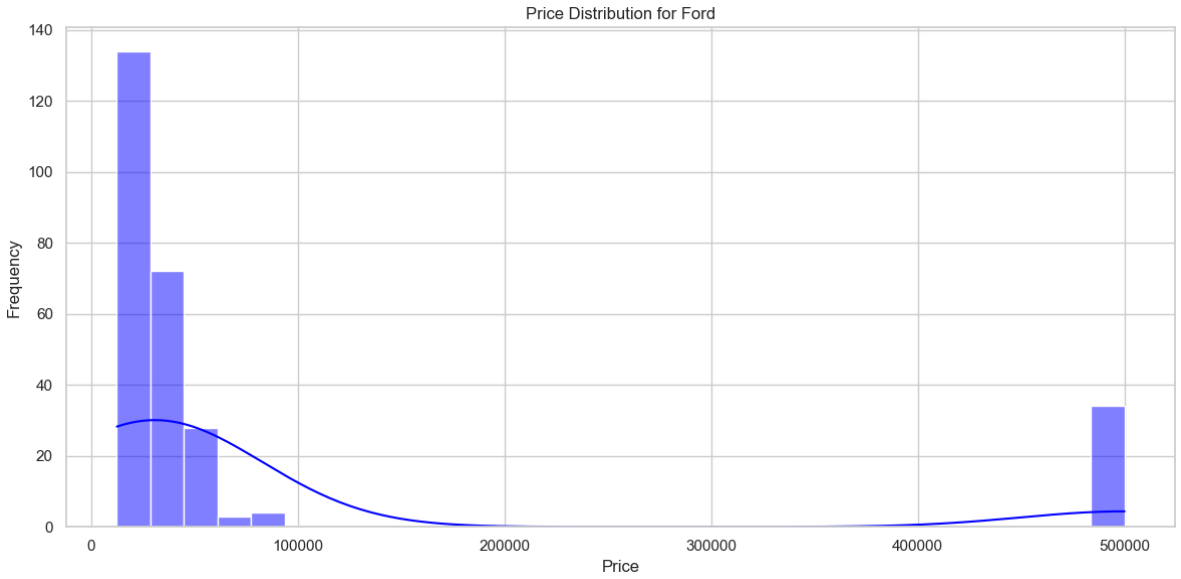


Figure 7

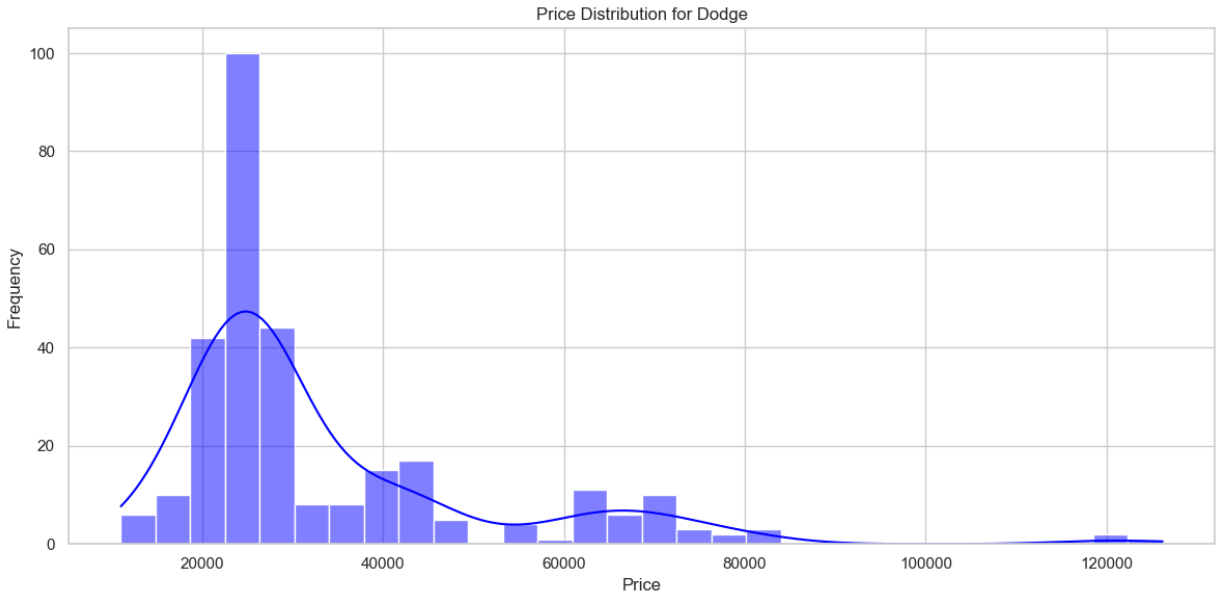
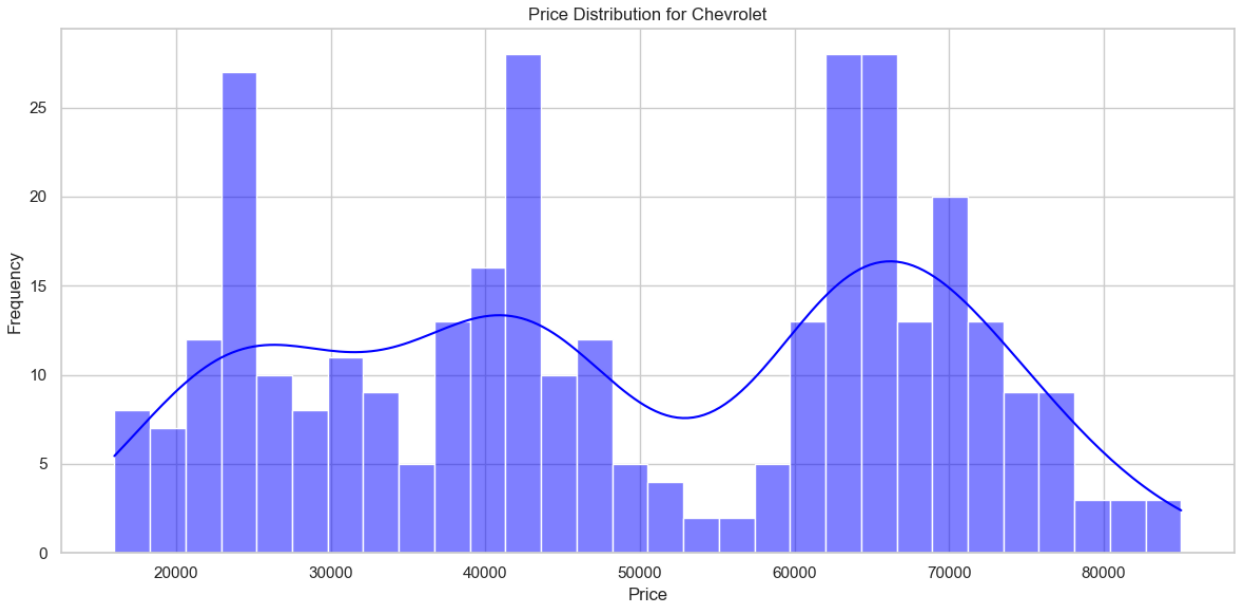


Figure 8

Figure 9

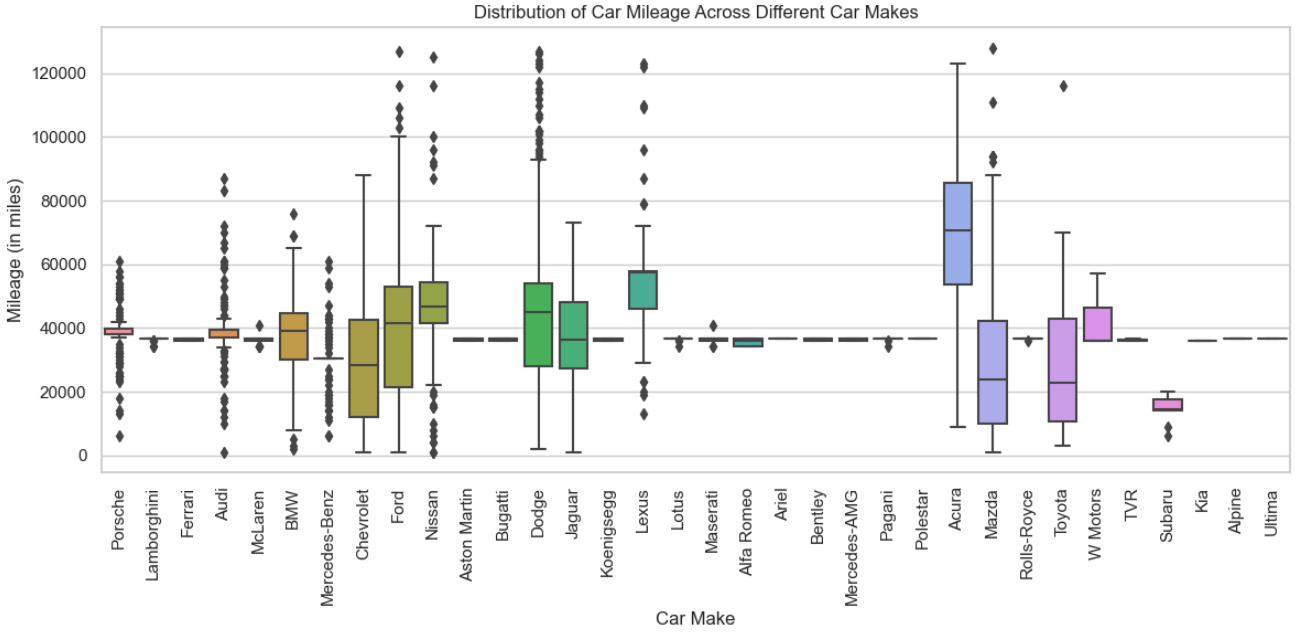


3.4 Distribution of Car Mileage Across Makes

Here, our objective was to examine the variation in car mileage across different makes. Mass-market brands like Ford, Chevrolet, and Dodge show higher median mileages (60,000–120,000) and wider distributions, reflecting their common use as daily drivers. In contrast, luxury and exotic brands like Koenigsegg, Pagani, Bentley, and Ferrari have lower median mileages and narrower distributions, emphasizing their limited usage and exclusivity.

Mid-range luxury brands like BMW and Mercedes-Benz have moderate median mileages and broader ranges, while high-end brands like Rolls-Royce cluster at lower mileages. Reliable brands like Toyota and Mazda show higher median mileages, reflecting everyday use. Niche manufacturers like Ultima and Ariel have consistently low mileages, while exotic brands like Ferrari and McLaren display outliers with unusually high mileages. Overall, the chart highlights that mass-market brands have more variability and higher usage, while luxury and exotic brands are used less frequently and in more specialized contexts.

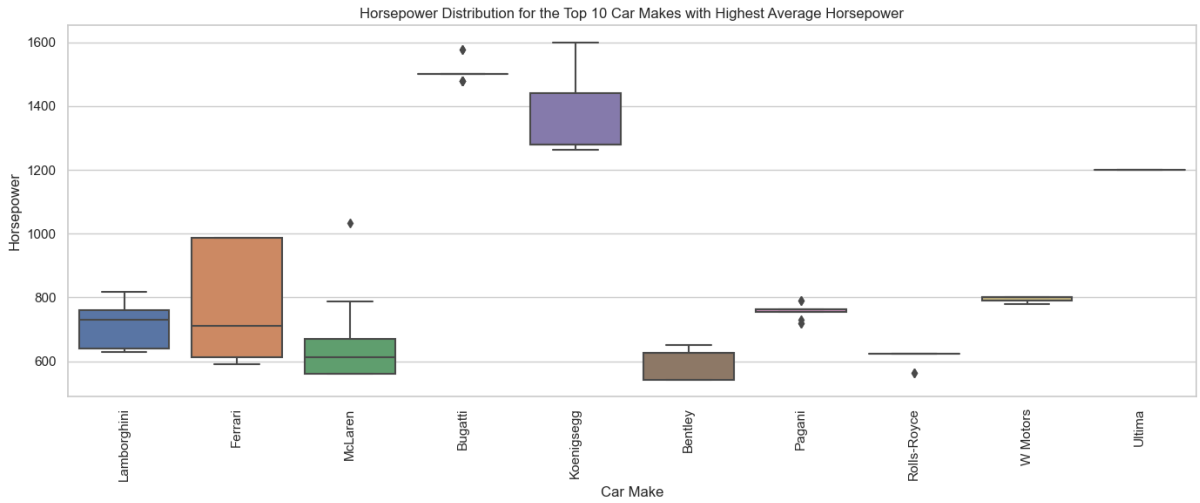
Figure 10



3.5 Distribution of Horsepower Across Most Expensive Makes

In this section, our goal was to analyze the distribution of horsepower among the top 10 car makes with the highest average horsepower. The box plot reveals that Bugatti and Koenigsegg lead with horsepower ranges between 1250–1600 HP, with medians well above 1400 HP, while Ultima stands out with a single value at 1200 HP, indicating no variability. Ferrari and McLaren show a wider spread of horsepower, ranging from around 550 HP to over 1000 HP, with occasional high outliers. In contrast, Lamborghini, Pagani, and W Motors exhibit tighter distributions, with medians between 700–800 HP. Bentley and Rolls-Royce cluster in a lower range, between 550–640 HP, with Rolls-Royce showing minor outliers. This graph highlights lots of variability among car makes, with Bugatti and Koenigsegg producing the most powerful cars, while brands like Bentley and Rolls-Royce prioritize more consistent, lower horsepower values.

Figure 11



3.6 Relationship Between Horsepower, Engine Size, and Acceleration Time

For this section, we decided to shift away from the theme of price and explore the innovation aspect of sports cars instead. We examined the relationships between horsepower, engine size, and acceleration time to provide insight into how these factors influence vehicle performance. Figure 12 demonstrates that as horsepower increases, acceleration time decreases, indicating that vehicles with higher horsepower achieve faster acceleration. Figure 13 examines engine size versus acceleration time, showing a similar trend: as engine size increases, acceleration time decreases, suggesting that larger engines typically allow for quicker acceleration. Figure 14 illustrates the positive correlation between engine size and horsepower, with larger engines generally producing higher horsepower.

From these observations, achieving peak acceleration requires balancing power output (horsepower) and weight (engine size). High horsepower results in faster acceleration, but larger engines that produce this horsepower often add weight, which can reduce overall efficiency. Therefore, the ideal engine for peak acceleration would be lightweight (a smaller engine, such as under 3 liters) while still capable of producing "big-engine horsepower" (400–600+ HP). This combination maximizes the power-to-weight ratio, delivering quick acceleration without the burden of additional weight.

Figure 12

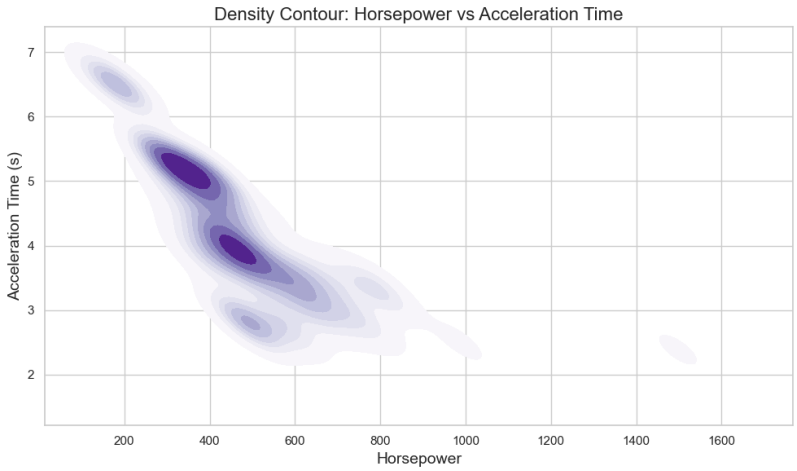


Figure 13

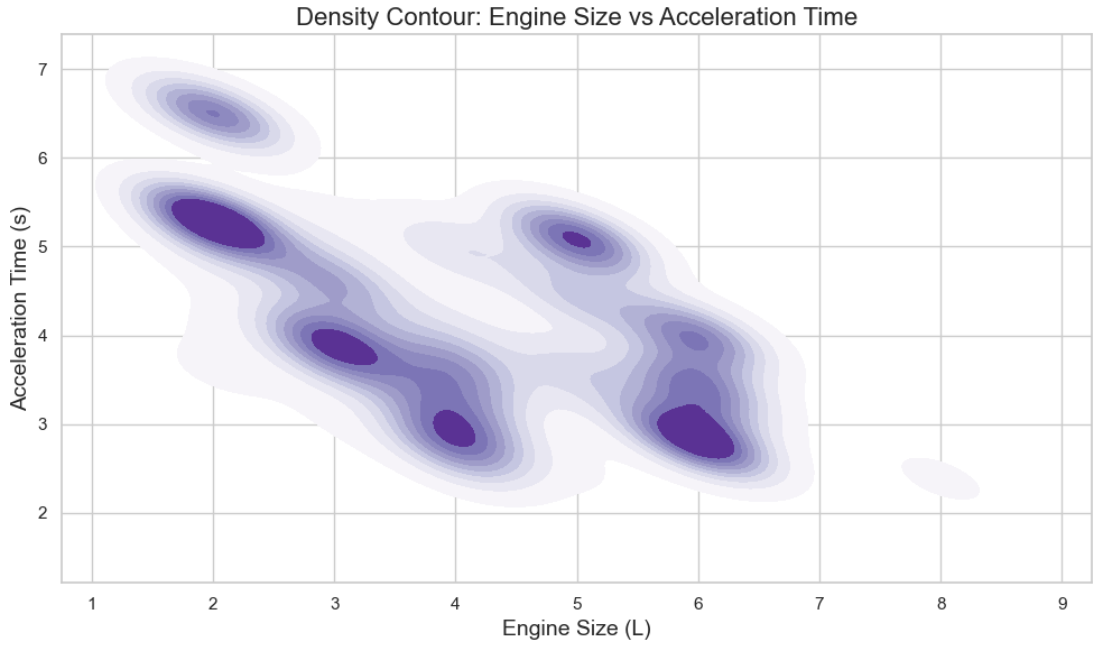
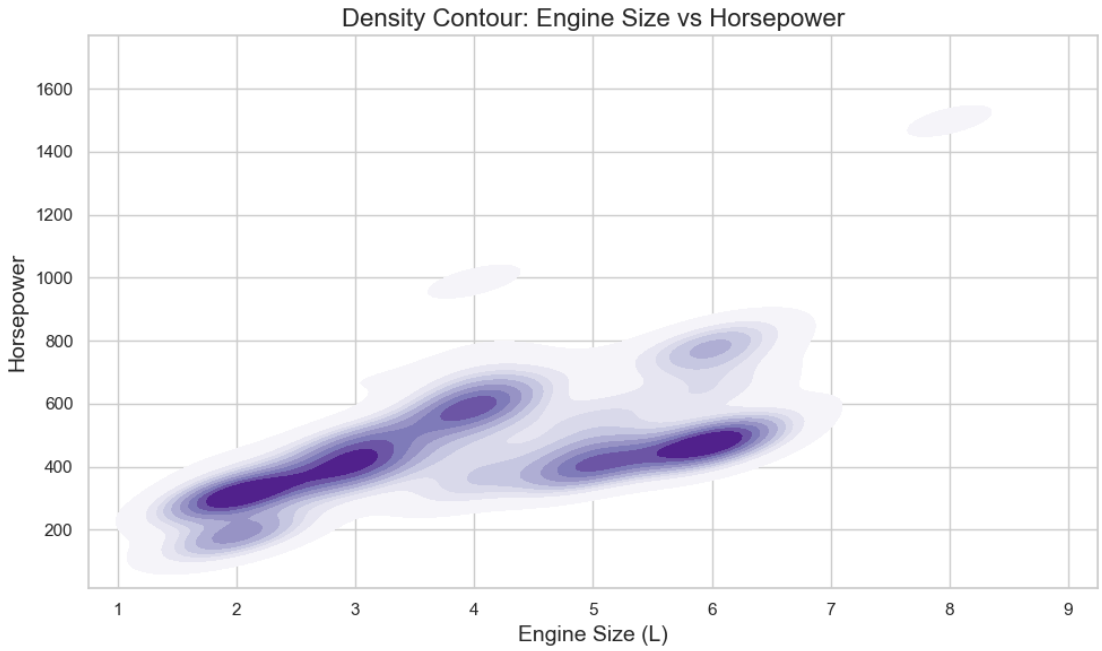


Figure 14



4. Conclusion

In this project, we analyzed pricing of sports cars and how much their basic features impact their sales price. From the questions presented in our project proposal, as well as a few extra questions we had when working more data, we found the following results.

1. *What factors have the biggest impact on sports car prices? Which factors have the least impact?*

   Horsepower has the strongest positive correlation with price, and acceleration time has the strongest negative correlation with price. Mileage has the weakest correlation with price and does not appear to have a strong impact on average in final asking price of a sports car.

2. *Which car "make" is the most expensive and least expensive on average?*

   The most expensive sports car "make" according to the data is Bugatti, with an average asking price of about $3.25 million. The least expensive sports car sold is a Mazda, with an average asking price of about $25,000.

3. *What is the price distribution of the most common makes?*
   The price distributions of the most common car makes reveal that luxury brands like BMW show broader, right-skewed price ranges with higher outliers, while mainstream brands such as Ford and Mazda show more concentrated, affordable price clusters. Chevrolet displays what appears to be two groupings within its distribution, suggesting it may have two separate pricing groups it markets its cars in.

4. *What is the distribution of car mileage across car makes?*
   Based on the charts, we concluded that mass-market brands have more variability and higher usage, while luxury and exotic brands are used less frequently and in more specialized contexts, leading to lower average car mileage listings.

5. *What is the distribution of horsepower across the most expensive car makes?*
   There is large variation in distribution of horsepower among the most expensive car makes, with brands like Bugatti and Koenigsegg dominating the high-performance range above 1400 HP, while brands such as Bentley and Rolls-Royce focus on lower horsepower levels around 550–640 HP. Each brand's market positioning is illustrated well with this variable, as the highest-end luxury brands feature the highest horsepower.

6. *What is the relationship between horsepower, engine size, and acceleration time?*
   The relationship between horsepower, engine size, and acceleration time reveals that higher horsepower and larger engine sizes generally result in faster acceleration, however larger engines can introduce weight that can offset performance. Achieving the highest acceleration requires balancing horsepower and engine size to achieve a high power-to-weight ratio, likely calling for smaller, lighter engines to reduce weight.

This project has some limitations, including the imputation of missing values using averages and the relatively small number of variables considered. Future work could expand the dataset to include more sports cars with complete information and incorporate additional data sources for greater variation. One potential avenue for further analysis is examining how factors such as sales location (e.g., area sold, online vs. in-person sales) may influence pricing.