PROJECT REPORT

SIC - AI702

SAMSUNG INNOVATION CAMPUS

PROJECT 1 - SEPTEMBER 2025

Team Members

- Mariam Hany
- Shrouk Sayed
- Lojayn Khaled

Project Overview & Data

Project Overview

The focus of this project is to explore a dataset of cars. This will be achieved by cleaning the data, performing an exploratory analysis through various graphs and visualizations, and ultimately drawing key conclusions and insights from the findings.

Data Used

The data used is a Kaggle data set.
The Data consists of:

- 11,914 Rows
- 16 Columns.

Dataset Key Features

Column Name	Description	Example Value
Make	The brand of the vehicle	Toyota, BMW, Ford
Model	The specific product name of the brand	Camry, X5, Mustang
Year	The model year of the vehicle	2020, 2015, 1998
Engine Fuel Type	The type of fuel the engine is designed to use	Regular unleaded, electric, premium unleaded, and diesel. Premium unleaded, Diesel
Engine HP	The power output of the engine is measured in horsepower.	150, 320, 450
Engine Cylinders	The number of cylinders in the vehicle's engine.	4, 6, 8, 12
Transmission Type	The type of gearbox that transfers power from the engine to the wheels.	Manual, Automatic, Automated manual, Direct Drive and Unknown
Driven wheels	Which wheels receive power from the engine	front wheel drive, rear wheel drive, all-wheel

		drive and four-wheel drive
Number of doors	The number of doors in the vehicle	2, 3,4
Market Category	Marketing tags used to classify the vehicle	Crossover, Luxury, Hatchback and Diesel
Vehicle Size	Vehicle physical dimensions	Compact, Midsize, Large
Vehicle Style	The body type or design of the vehicle	Sedan, Coupe, SUV, Pickup, Wagon
Highway mpg	The vehicle's fuel efficiency on highway driving, in miles per gallon.	26, 28, 24
City mpg	The vehicle's fuel efficiency in city driving, in miles per gallon.	17, 18, 19, 26
Popularity		2,26 and 61
MSRP	Manufacturer's Suggested Retail Price.	25995, 36350 and 92600

Cleaning & preparing data

1) **Data Ingestion**

The raw data was loaded into a pandas Data Frame for processing.

2) **Column Name Standardization:**

The original column names contained inconsistencies. We standardized them by converting all names to a consistent format using spaces for readability

3) **Deleting Duplicate Rows**

- Identified and removed all duplicate rows from the dataset to ensure the integrity and accuracy of the analysis.
- 1332 rows were found to be duplicates, so we dropped them.

4) **Handling Missing Value:**

Missing values were addressed on a column-by-column basis using the following strategies:

- **Engine Fuel Type & Number of Doors:** A small number of missing values were present (3 and 6, respectively). These were imputed by looking up the table for the same Make and Model combination within the dataset.
- **Engine HP & Engine Cylinders:** As these represent discrete numerical values, missing entries were filled using the median value of their respective columns to minimize the influence of any remaining outliers.
- **Market Category:** This field presented a unique challenge. It contained a high number of nulls and its values are multi-label categorical strings (e.g., "Luxury, Performance, Hatchback")
    - **Step 1:** Initial null values were filled with the overall mode (most common single value).
    - **Step 2:** To make this data usable for analysis, we transformed the Market Category column. Each unique category within the comma-separated strings was extracted and converted into a new binary feature column (e.g., Luxury, Performance, Hatchback). A value of 1 indicates the vehicle belongs to that category, while 0 indicates it does not.

5) **Correcting Data Types**

Converted the Number of Doors and Engine Cylinders columns from float to integer (int) data types, as they represent discrete, countable values.

6) **Feature Engineering**

Created a new Fuel Efficiency feature by calculating the average of the City MPG and Highway MPG columns. This consolidated two related features into one more general metric and helped reduce the number of features.

7) **Dropping Unnecessary Columns**
   - Analyzed a correlation heatmap to identify the relationship between numerical features and the car price.
   - Removed all columns with a very low correlation to the price, as they were deemed unlikely to significantly impact on our analysis.
   - The following columns were removed:
     - Diesel
     - Flex Fuel
     - Hybrid
     - Number of Doors
     - Hatchback

Exploratory Data Analysis (EDA)

Data Source: Power BI Dashboards (Car Popularity & Car Price Analysis)

A. Descriptive Statistics & Key Findings

1. Overall Price Distribution: Extreme Range with High-Value Outliers

- **Average Price (Mean):** ~$41,930
- **Median Price:** $30,680
- **Price Range:** $2,068,000 (Min: $2,000 - Max: $2,070,000)

Interpretation & Insight:
The significant difference between the mean and median price is the most critical insight. It reveals that the **average is heavily skewed upwards by a small number of ultra-luxury and hypercars**. Therefore, the **median price of $30,680 is a more accurate representation of a "typical" car's price** in this market. The existence of vehicles priced over $2 million (e.g., Bugatti Veyron) creates an extreme long-tail distribution.

Dashboard Evidence: The "Average of MSRP by Model" chart explicitly shows models like the **Veyron 16.4** and **Phantom** at the extreme high end (>$1.5M), which pulls the overall average up.

2. Engine Power (HP) and Cylinders: The Primary Performance Correlates

- **Engine Cylinders:** The data includes configurations from 3 to 16 cylinders.
  - **Most Common:** 4, 6, and 8 cylinders are the most frequent, representing mainstream vehicles.
  - **Relationship with Price:** There is a very strong positive correlation. The "Average of MSRP by Engine Cylinders" chart shows a clear trend: **price increases exponentially with cylinder count**. Cars with 12+ cylinders have an average MSRP exceeding $1.5 million.

- **Engine Horsepower (HP):**
    - **Average HP:** 253.3 HP
    - **Maximum HP:** 1,001 HP (aligning with hypercars like Bugatti)
    - **Relationship with Price:** High correlation. The "Engine HP Analytics" section confirms that HP is a key metric for segmenting "Performance" and "Exotic" vehicles, which command premium prices.

Interpretation & Insight: Cylinder count and horsepower are not just specs; they are **direct proxies for vehicle performance and luxury tier**. The jump from 8 to 12 cylinders represents a fundamental shift from high-performance sports cars to the ultra-exclusive hypercar segment.

3. Market Segmentation: Luxury/Performance vs. Mainstream

The dashboards are built around a key segmentation strategy:

- **Standard Vehicles:** Make up the bulk of the volume. Median price ~$30k.
- **Performance Vehicles:** Higher HP and cylinders, higher price point (~$60K - $100K+ based on "Average of MSRP by PerformanceCategory").
- **Exotic Vehicles:** Represent the extreme end of the market. The "Average of MSRP by ExoticCategory" chart shows an **average price approaching $200,000**, with individual models far exceeding this.

Interpretation & Insight: Segmenting by "Performance" and "Exotic" categories is highly effective. It isolates the high-value outliers that distort overall averages and allows for separate analysis of mainstream and luxury markets.

B. Brand and Model Analysis: Volume vs. Value

4. Most Popular Car Brands (by Volume)

- **Top Brands:** Chevrolet, Toyota, Nissan, Volkswagen, Ford.
- **Insight:** These high-volume manufacturers dominate the market in terms of number of models and units sold, focusing on the affordable to mid-range segments.

Dashboard Evidence: Confirmed by the "Most Famous car brands" card in the Car Popularity dashboard.

5. Most Popular Car Models (by Volume)

- **Top Models:** F-150, Silverando 1500, Tundra, Frontier, Sierra 1500.
- **Insight:** This suggests a strong market presence for **pickup trucks and SUVs**, indicating high consumer demand for these vehicle types.

6. Highest Value Car Brands (by Average Price)

- **Top Luxury Brands:** Bugatti, Rolls-Royce, Lamborghini, McLaren, Ferrari, Aston Martin, Spyker.
- **Insight:** These brands have a focused strategy on the high-margin, low-volume luxury and performance segment. Their average MSRP is orders of magnitude higher than volume brands.

Dashboard Evidence: The "Average of MSRP by Make" chart clearly lists these brands with average prices in the hundreds of thousands to millions.

7. Highest Value Car Models (by Absolute Price)

- **Top Models:** Bugatti Veyron 16.4, Mercedes-Benz SLR McLaren, Rolls-Royce Phantom, Lamborghini Aventador.
- **Insight:** These models represent the pinnacle of automotive luxury and performance, serving as "halo products" for their brands and defining the upper extreme of the price range.

C. Additional Key Insights from Your Analysis

8. Impact of Vehicle Size and Type

- **Finding:** The "Average of MSRP by Vehicle Size" chart confirms a relationship.
- **Insight: Larger vehicles (e.g., Full-Size trucks, SUVs) command a higher average price** than compact or midsize cars. This is likely due to higher materials cost, more features, and consumer willingness to pay a premium for utility and space.

9. Transmission Type

- **Finding:** The "Averages of MSRP by Transmission Type" chart shows a variance.
- **Insight:** While automatic transmissions are most common, **specialized transmissions (e.g., AUTOMATED_MANUAL, DIRECT_DRIVE)** are often found in higher-performance, higher-priced vehicles, indicating another feature correlating with cost.

Data Visualization & Questions

A. Key Questions & Findings

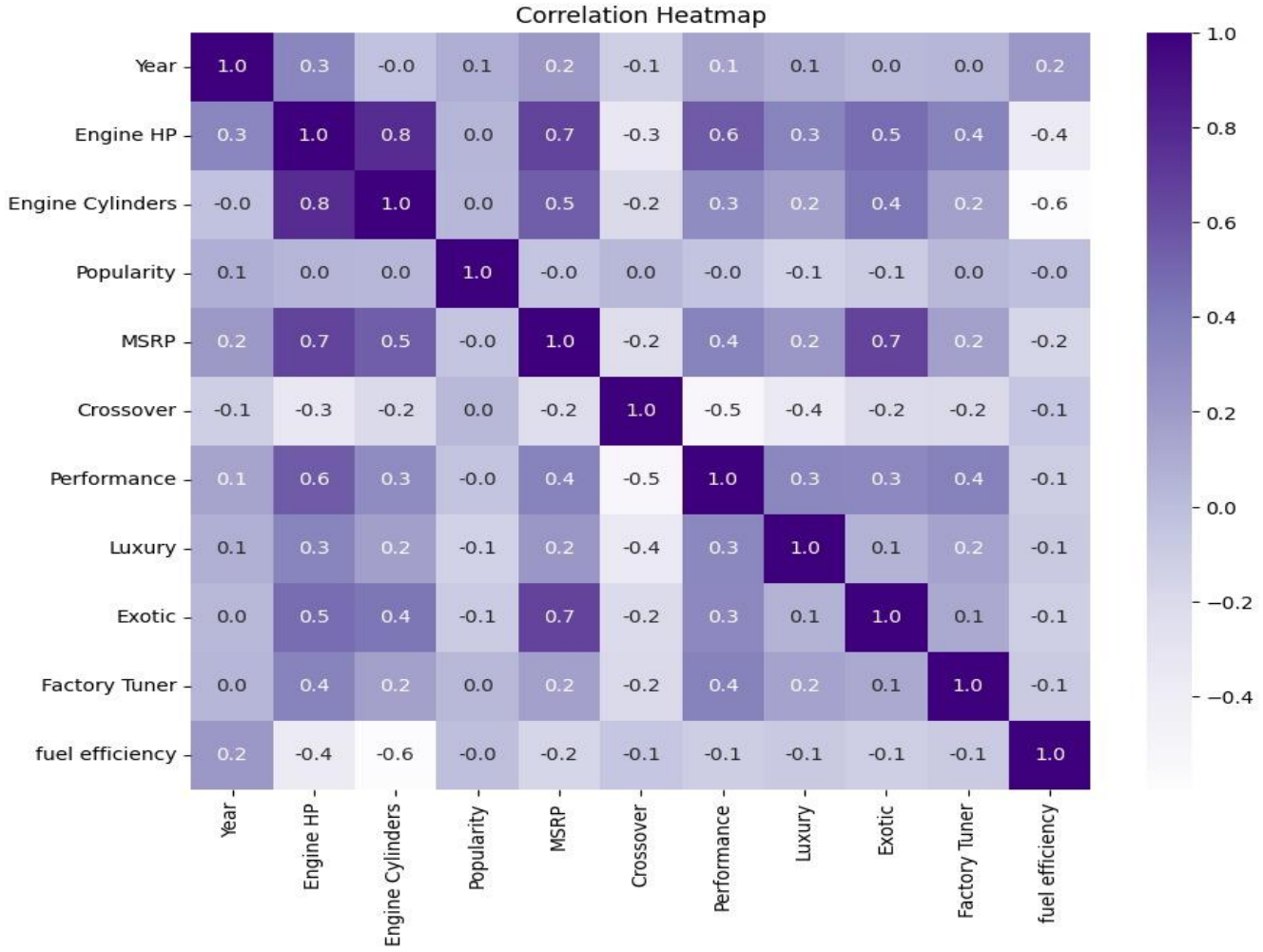
The analysis was driven by several key questions about what factors influence a vehicle's price.

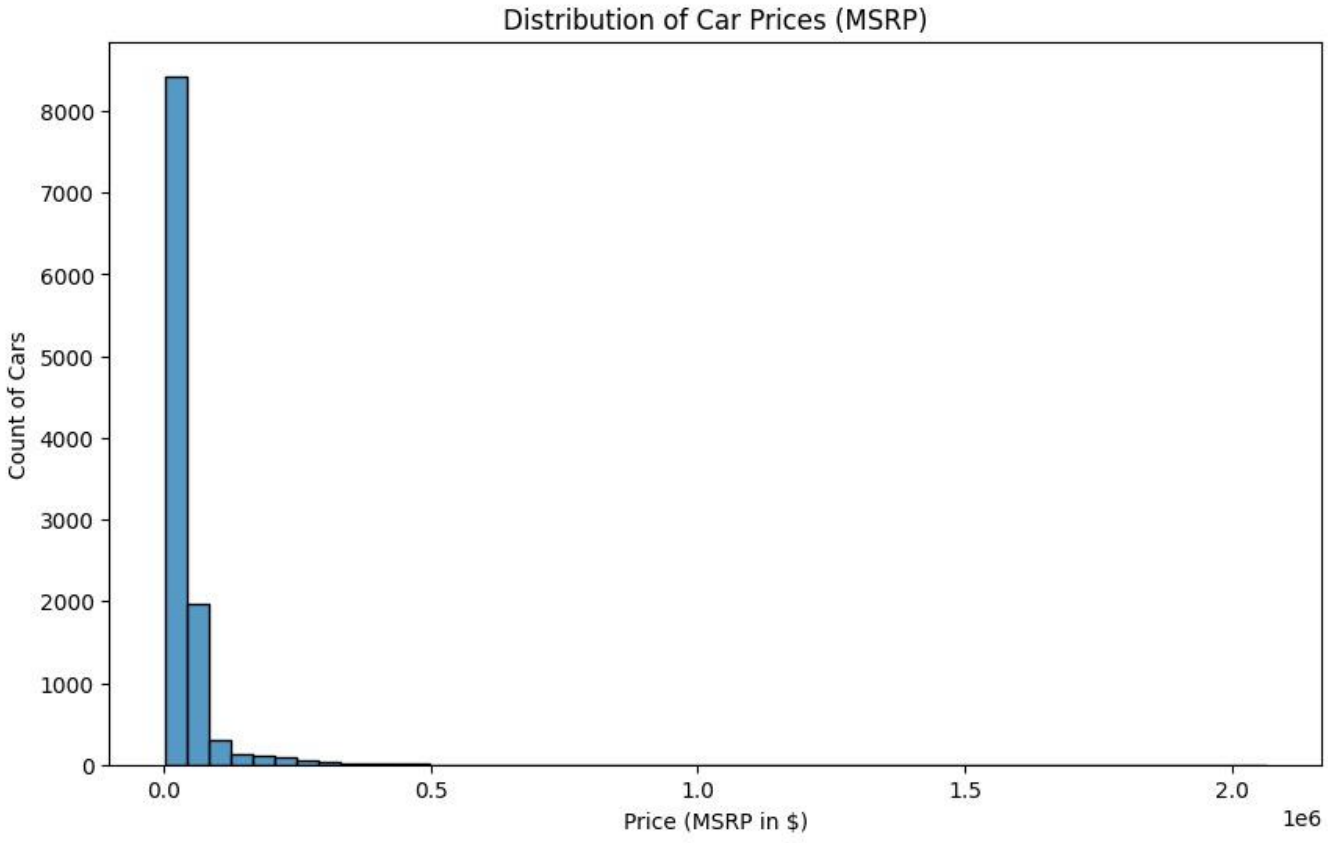
Question	What We Found
Which car make have the highest average price?	**Exotic and ultra-luxury brands** (Bugatti, Rolls-Royce, Lamborghini, McLaren) command the highest prices, driven by extreme performance, exclusivity, and hand-crafted engineering. This is the single biggest differentiator in price.
Is there a relation between vehicle size & average price?	**Yes.** Larger vehicles (e.g., full-size SUVs, trucks) have a higher average price than compact or midsize cars, justified by more features, space, and larger engines.
Is there a relation between transmission type & average price?	**Yes.** Specialized transmissions (**Automated Manual** and **Direct Drive**) are strong price indicators, commanding the highest averages as they are featured in high-performance models. Traditional automatics and manuals are associated with more affordable vehicles.
What are the strongest technical predictors of price?	**Engine Horsepower (HP)** and **Engine Cylinders** have the highest correlation with MSRP (0.7 and 0.5, respectively). They are the primary technical drivers of cost.

B. Data Visualization Strategy

We employed specific visualizations to effectively communicate these insights.

Visualization	Type	Why It Was Chosen
Average MSRP by Make	Bar Chart	To allow for a direct, easy comparison of average prices across brands, instantly identifying premium and volume manufacturers.
Average MSRP by Vehicle Size	Bar Chart	To clearly show the ordinal relationship between a categorical size (Compact, Midsize, Large) and its corresponding average price.
Average MSRP by Engine Cylinders	Bar Chart	To effectively visualize the strong positive correlation and exponential price growth as the number of cylinders increases.
Correlation Heatmap	Heatmap	To provide a single, comprehensive overview of the linear relationships between all numeric variables, quickly identifying key price drivers like Engine HP and Exotic status
MSRP Distribution	Histogram	To reveal the underlying **power-law distribution** of the data, showing that most cars are affordable while exponentially fewer ultra-expensive cars form a long tail. This explains why the median price is a better measure of centrality than the mean.


Correlation Heatmap
Year 1.0 0.3 -0.0 0.1 0.2 -0.1 0.1 0.1 0.0 0.0 0.2
Engine HP 0.3 1.0 0.8 0.0 0.7 -0.3 0.6 0.3 0.5 0.4 -0.4
Engine Cylinders -0.0 0.8 1.0 0.0 0.5 -0.2 0.3 0.2 0.4 0.2 -0.6
Popularity 0.1 0.0 0.0 1.0 -0.0 0.0 -0.0 -0.1 -0.1 0.0 -0.0
MSRP 0.2 0.7 0.5 -0.0 1.0 -0.2 0.4 0.2 0.7 0.2 -0.2
Crossover -0.1 -0.3 -0.2 0.0 -0.2 1.0 -0.5 -0.4 -0.2 -0.2 -0.1
Performance 0.1 0.6 0.3 -0.0 0.4 -0.5 1.0 0.3 0.3 0.4 -0.1
Luxury 0.1 0.3 0.2 -0.1 0.2 -0.4 0.3 1.0 0.1 0.2 -0.1
Exotic 0.0 0.5 0.4 -0.1 0.7 -0.2 0.3 0.1 1.0 0.1 -0.1
Factory Tuner 0.0 0.4 0.2 0.0 0.2 -0.2 0.4 0.2 0.1 1.0 -0.1
fuel efficiency 0.2 -0.4 -0.6 -0.0 -0.2 -0.1 -0.1 -0.1 -0.1 -0.1 1.0
Year
Engine HP
Engine Cylinders
Popularity
MSRP
Crossover
Performance
Luxury
Exotic
Factory Tuner
fuel efficiency
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Distribution of Car Prices (MSRP)
Count of Cars
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
Price (MSRP in $)
1e6

C. Main Conclusions of the Analysis

1. **The Market is Bimodal:** The data reveals two distinct markets: a **high-volume, affordable mainstream market** (median ~$30k) dominated by brands like Toyota and Chevrolet, and a **low-volume, high-margin luxury performance market** defined by exotic brands.
2. **Price is Driven by Performance and Exclusivity:** The strongest predictors of a high MSRP are not basic features but metrics of performance (**Engine HP, Cylinders**) and market positioning (**Exotic** or **Performance** category).
3. **The Data is Heavily Skewed:** Car prices follow a power-law distribution. Using the **mean price ($41,930) is misleading**; the **median price ($30,680)** is a more accurate representation of what a typical car costs. This skew is caused by extreme outliers in the luxury segment.
4. **Feature Hierarchy Matters:** When predicting price, 'Exotic' status and engine specs are paramount. Other features like vehicle size and transmission type are secondary indicators.

D. Business Recommendations

1. **For Marketing & Sales Teams:**
    - **Segment Your Audience:** Develop distinct marketing campaigns for performance/luxury buyers (focused on horsepower, exclusivity, engineering) vs. mainstream buyers (focused on reliability, value, utility).
    - **Value-Based Pricing:** For luxury models, pricing can be aggressive and based on positioning against competitors like Ferrari or McLaren. For mainstream models, pricing must be competitive within a tight range.
2. **For Product Strategy & Manufacturing:**
    - **Focus on HP and Performance Trims:** Since engine power is a key purchase driver, investing in performance variants of existing models can create higher-margin products.
    - **Understand the Luxury Premium:** The analysis quantifies the "exotic premium." This value can justify investments in bespoke materials, limited production runs, and advanced engineering for high-end marques.
3. **For Data Analysis Teams:**
    - **Build Segmented Models:** Do not build a single model to predict price for a Ferrari and a Ford. **Split the data** by vehicle category (e.g., standard, performance, exotic) before building predictive or analytical models to ensure accuracy.
    - **Use Robust Metrics:** Always report **median alongside mean** for financial metrics like price to avoid the skewing effect of ultra-luxury vehicles.

E. Limitations of the Analysis

1. **Dated Dataset:** The most significant limitation is that the analysis is based on **car model years from 1990 to 2017**. The market has evolved substantially since then, with the rise of electric vehicles (which have different price drivers like battery range), new luxury brands, and shifts in consumer preference towards SUVs. The conclusions may not fully reflect the current market dynamics.
2. **Causality vs. Correlation:** The analysis identifies correlations (e.g., more cylinders = higher price) but cannot prove causation. The high price of exotic cars is also driven by brand prestige, craftsmanship, and marketing, which are not fully captured in the data.
3. **Data Scope:** The dataset may not include all possible price-influencing factors, such as maintenance costs, brand perception scores, or specific technology packages, which could improve the model's accuracy.