SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Capstone Project 1: Car Dataset Analysis

SIC – AI701
By: Hasnaa Hossam, Ahmed Ayman, Youssef Tamer
Supervised by: Eng: Aya Nada

Agenda

- Problem Definition
- Targeted audience
- Data Exploration
- Data Preprocessing
- Data Visualizations & Observations
- Features Encoding & Scaling
- Feature Engineering
- Conclusion

Problem Definition

Problem

Car prices are influenced by many factors—engine power, fuel efficiency, brand, and more. Without data-driven analysis, it’s difficult for dealers, and consumers to understand what drives pricing.

Target audience

Target audience

Dealers

Build smarter pricing strategies and competitive positioning.

Consumers

Gain transparency and confidence in price differences.

Understanding The Data

Data Columns

```
df.head()
```

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity	MSRP
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916	40650
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	29450
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916	34500

```
Shape: (11914, 16)
```

Data Preprocessing

Handling Duplicates & Missing Values

1. Handling Duplicates

- **There are 715 duplicate rows that need to be handled**

```
[591] df.duplicated().sum()
np.int64(715)
[592] df.drop_duplicates(inplace=True)
[593] df.duplicated().sum()
np.int64(0)
```

2. Handling Missing Values

- **(Engine Fuel Type) filled with Mode**
- **(Engine HP)filled with Median**
- **(Engine Cylinders):**
  - **Electric Cars: filled with Zero**
  - **Non-electric filled with Mode**
- **(Number of Doors) filled with Mode**
- **Market Category is Dropped**

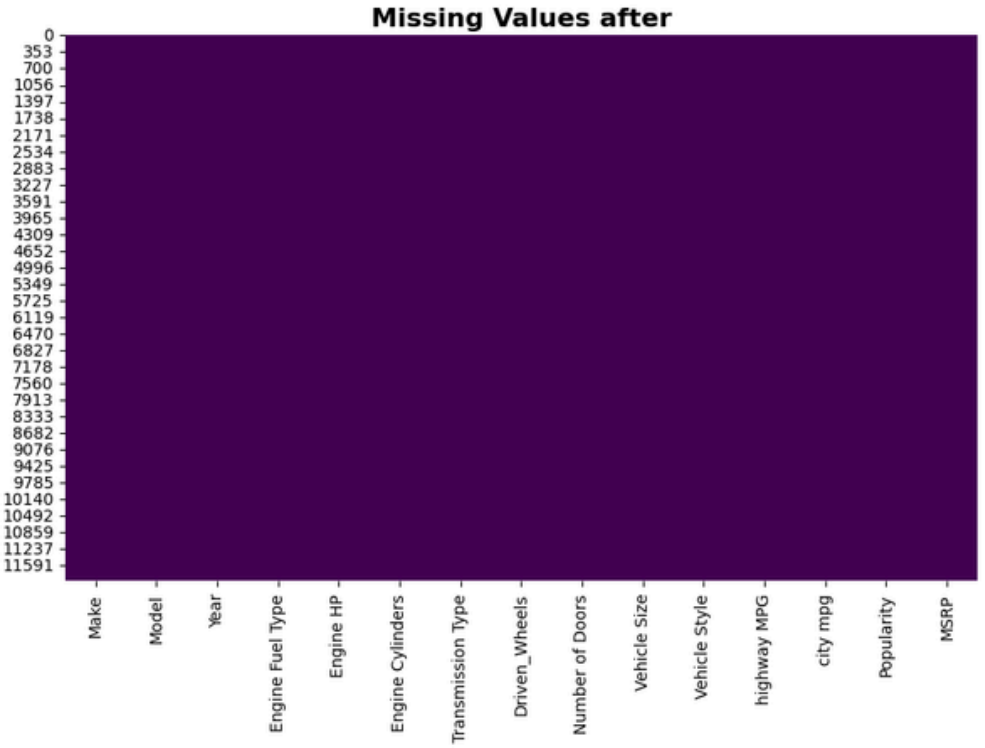
```
df.isnull().sum()
```

	0
Make	0
Model	0
Year	0
Engine Fuel Type	3
Engine HP	69
Engine Cylinders	30
Transmission Type	0
Driven_Wheels	0
Number of Doors	6
Market Category	3376
Vehicle Size	0
Vehicle Style	0
highway MPG	0
city mpg	0
Popularity	0
MSRP	0

```
df.isnull().sum()
```

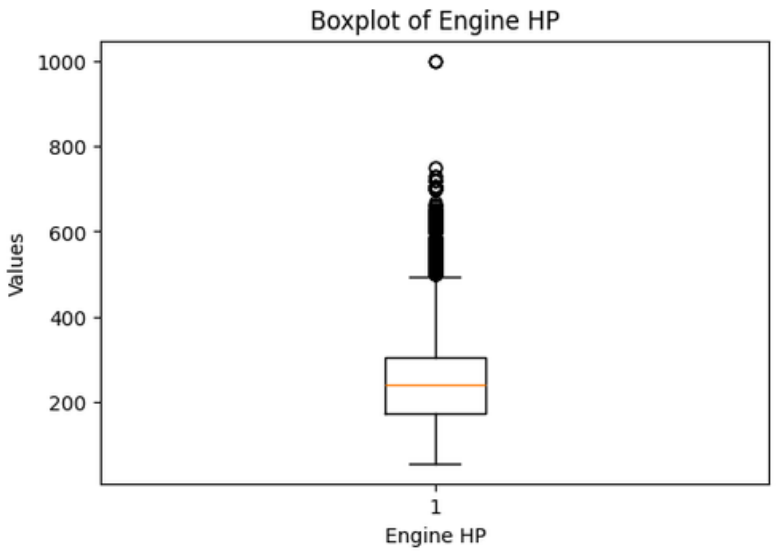
	0
Make	0
Model	0
Year	0
Engine Fuel Type	0
Engine HP	0
Engine Cylinders	0
Transmission Type	0
Driven_Wheels	0
Number of Doors	0
Vehicle Size	0
Vehicle Style	0
highway MPG	0
city mpg	0
Popularity	0
MSRP	0

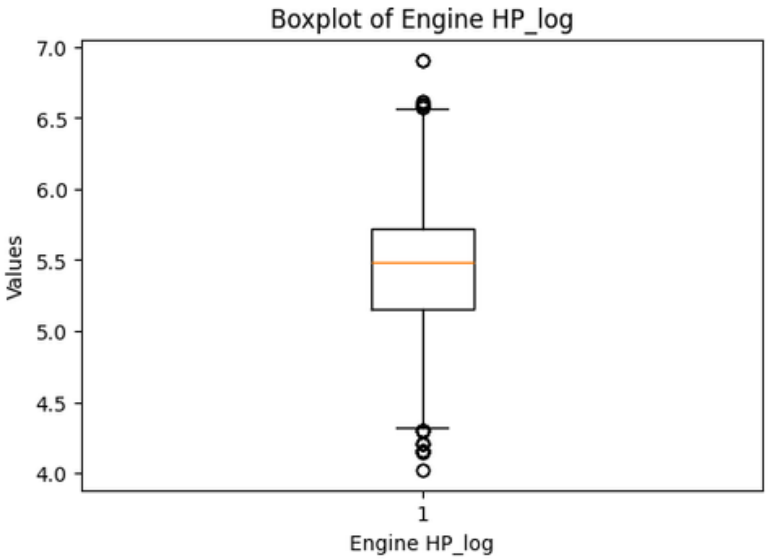
Handling Missing Values



Outliers Handling

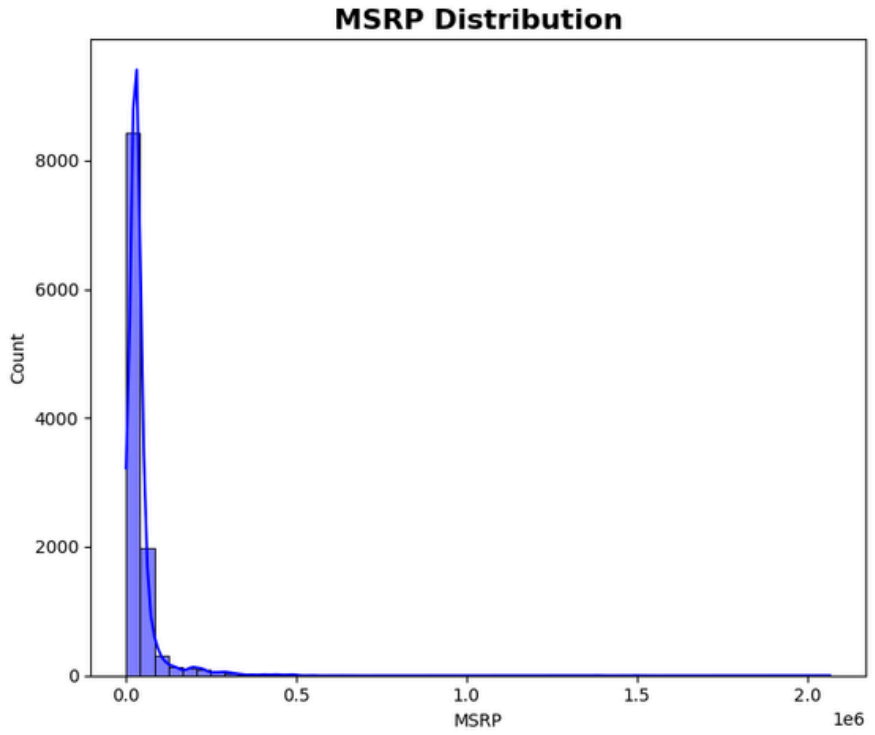
	Year	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg	Popularity	MSRP
count	11199.000000	11199.000000	11199.000000	11199.000000	11199.000000	11199.000000	11199.000000	1.119900e+04
mean	2010.714528	253.300205	5.657916	3.454416	26.610590	19.731851	1558.483347	4.192593e+04
std	7.228211	109.816822	1.803939	0.872804	8.977641	9.177555	1445.668872	6.153505e+04
min	1990.000000	55.000000	0.000000	2.000000	12.000000	7.000000	2.000000	2.000000e+03
25%	2007.000000	172.000000	4.000000	2.000000	22.000000	16.000000	549.000000	2.159950e+04
50%	2015.000000	239.000000	6.000000	4.000000	25.000000	18.000000	1385.000000	3.067500e+04
75%	2016.000000	303.000000	6.000000	4.000000	30.000000	22.000000	2009.000000	4.303250e+04
max	2017.000000	1001.000000	16.000000	4.000000	354.000000	137.000000	5657.000000	2.065902e+06

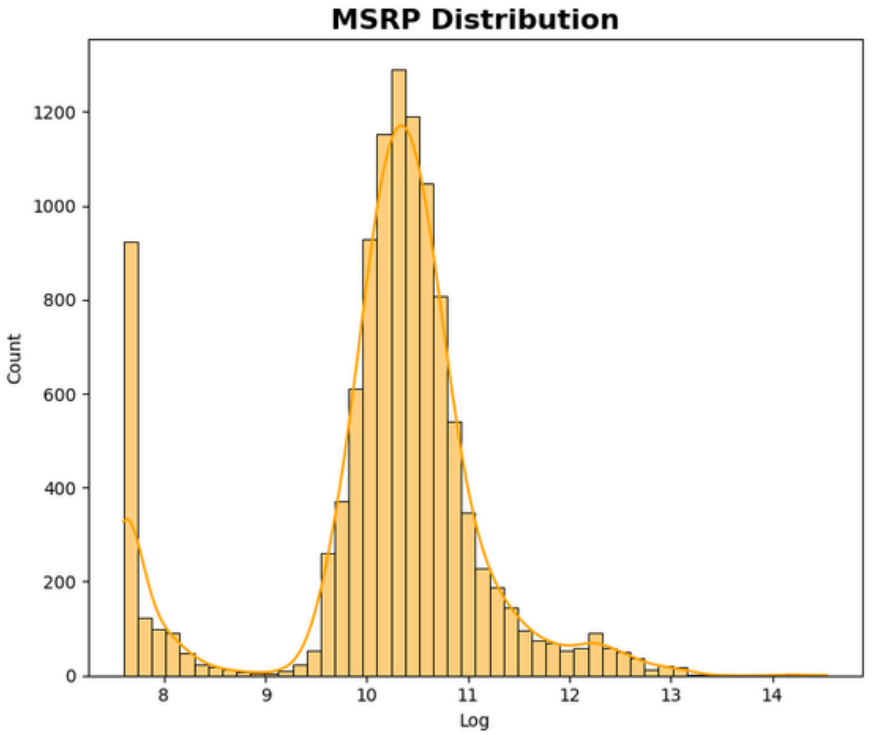




Handling Outliers: Log Transform

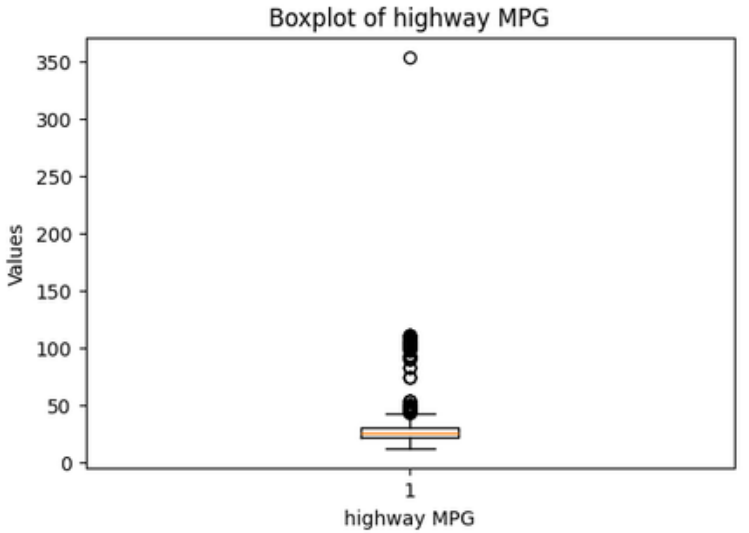
- Manufacturer’s Suggested Retail Price (MSRP)



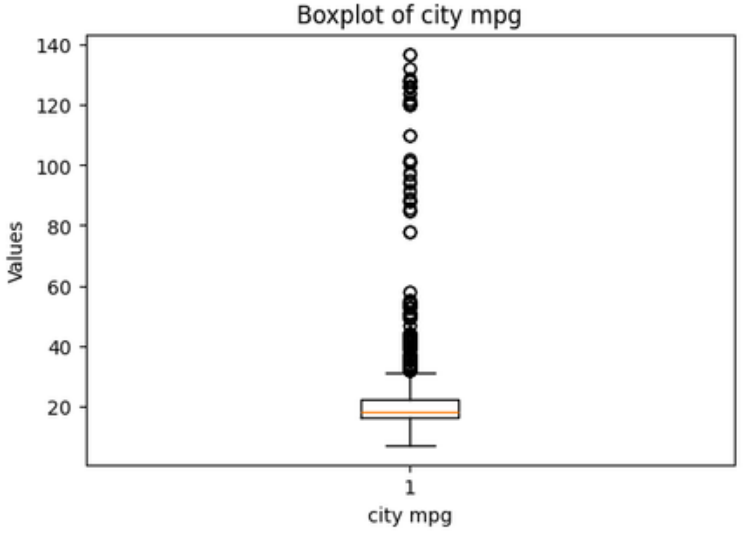


Handling Outliers: Highway & City MPG

- **Before:**
  - Highway MPG non-electric: 12 - 354
  - City MPG non-electric: 7 - 58
  - Highway MPG electric: 74 - 111
  - City MPG electric: 78 - 137



- **After:**
  - Highway MPG non-electric: 12 - 53
  - City MPG non-electric: 7 - 58
  - Highway MPG electric: 74 - 111
  - City MPG electric: 78 - 137

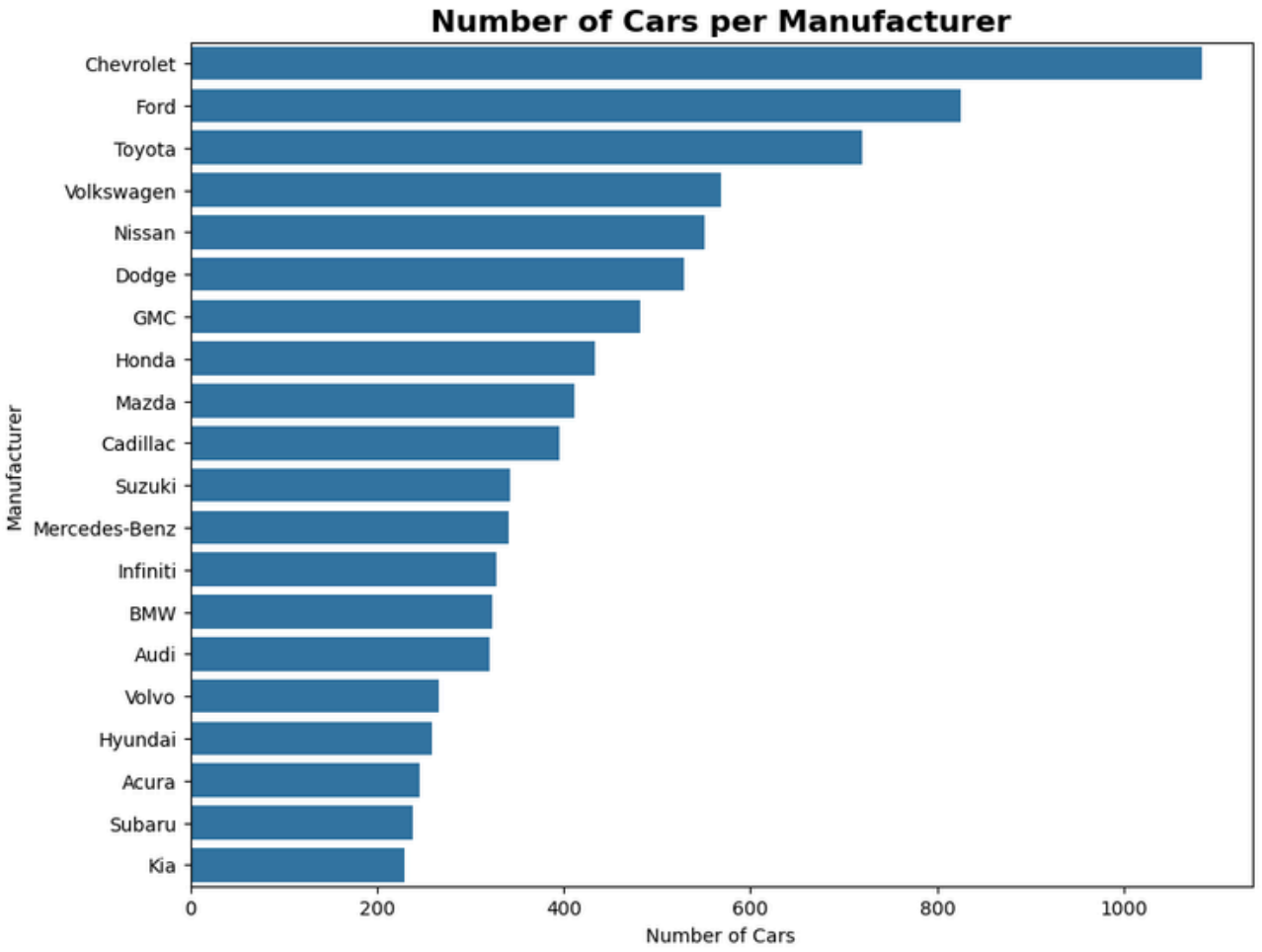


Visualizations

Car Production Count Distribution By Manufacturer

Observations:

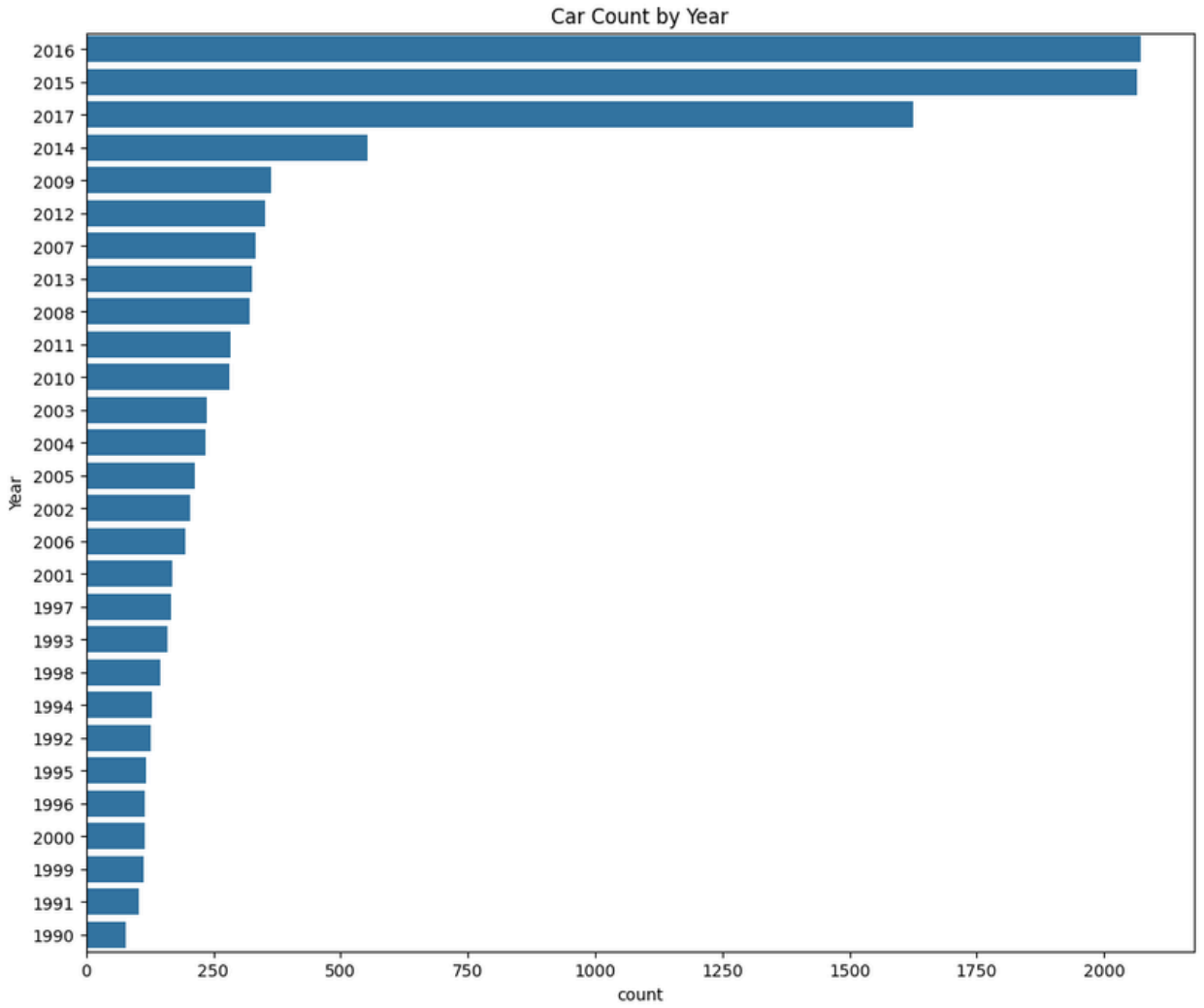
- The dataset is dominated by American brands like Chevrolet and Ford.
- International brands such as Kia and Hyundai are underrepresented.



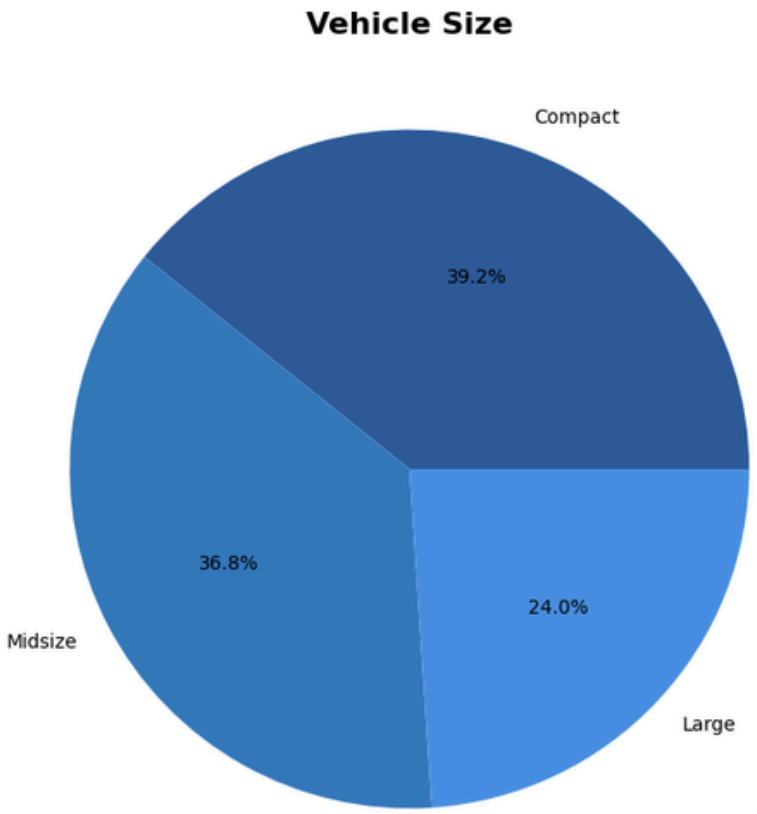
Car Production Count Distribution By Year

Observations:

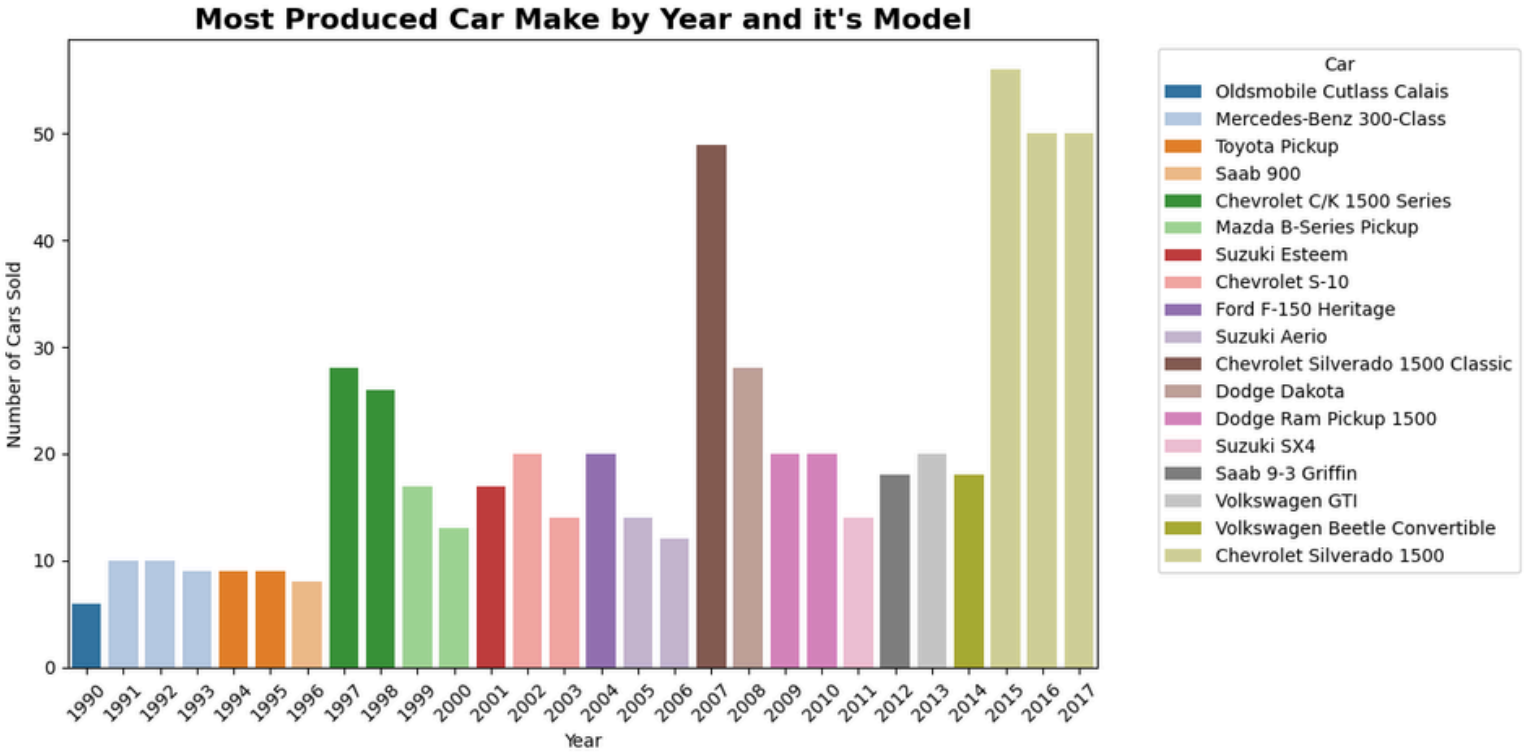
- Car counts rise sharply in 2014–2015, partly due to extra trims added.
- Sharp drop from 2016 → 2017, likely due to incomplete data for the final year.



Vehicle Size Distribution

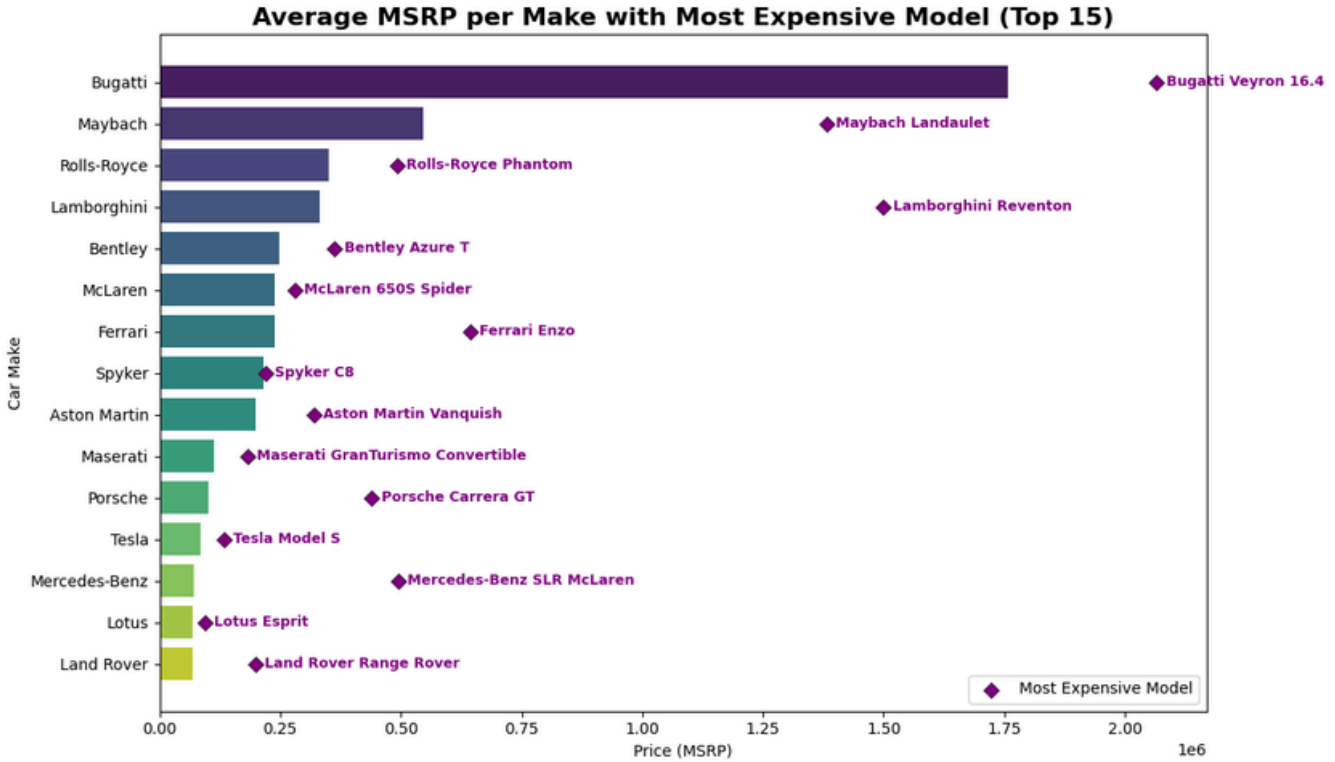


Most sold car Model by year



Chevrolet leads, especially with the **Silverado 1500** in **2015** (highest sales).
Pickup trucks (**Chevrolet**, **Dodge**, **Ford**) dominate across years.
1990s models show lower counts, while recent years record stronger volumes.

Highest average MSRP Brand and it's most expensive model

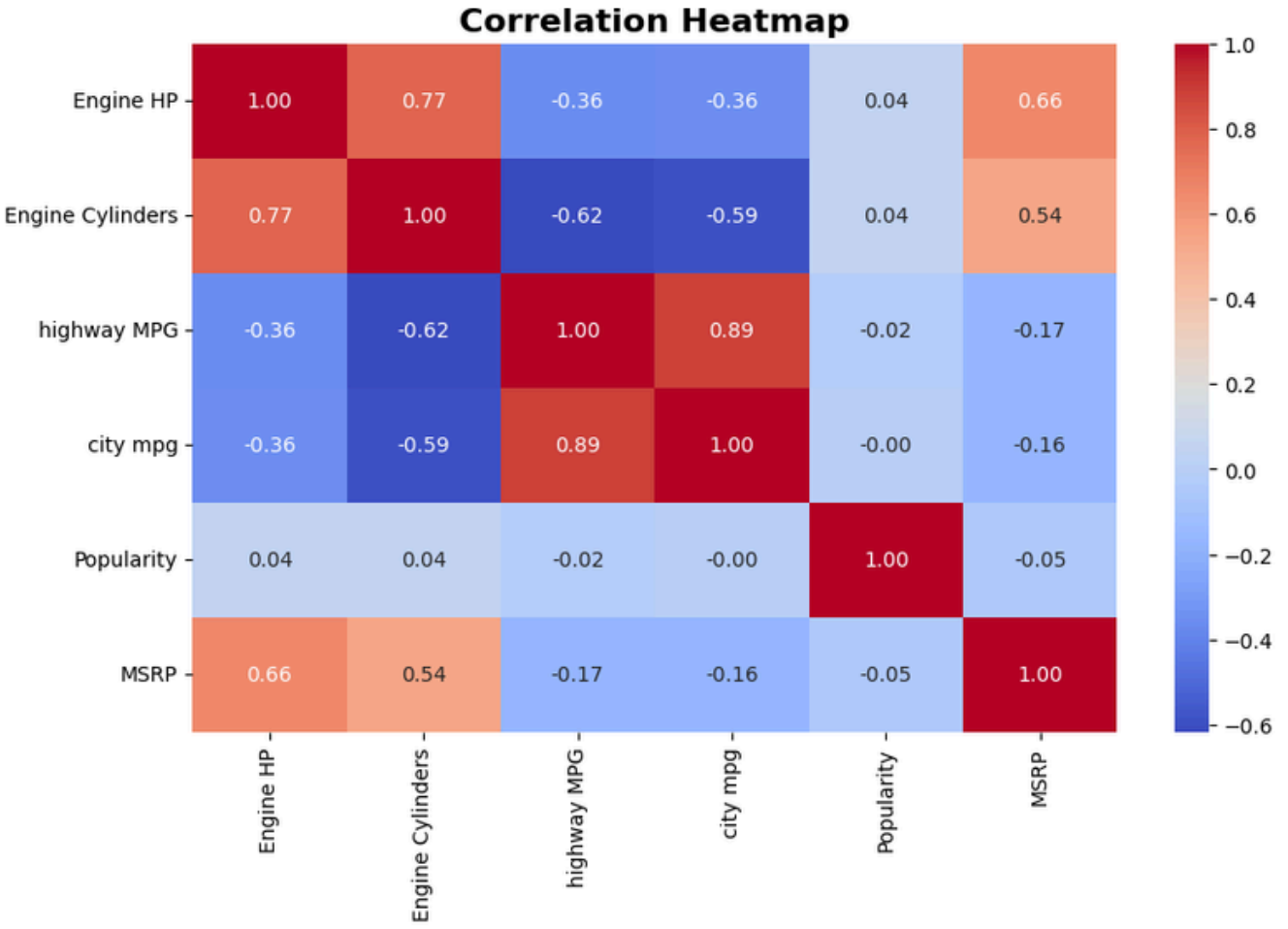


- **Bugatti has both the highest average MSRP and the most expensive model overall (Bugatti Veyron 16.4 at over $2M).**
- **For most luxury brands, the most expensive model is far above the brand's average price, especially for niche luxury brands like Maybach, Rolls-Royce and Lamborghini**
- **There is a clear gap between the average car price and the flagship models showing how a single ultra-luxury flagship can influence brand perception.**

Correlation Heatmap

Observations:

- Price increases with higher **horsepower** and more **cylinders**.
- Fuel efficiency has little effect on car price.
- Popularity does not influence car price linearly.



Encoding & Scaling

Encoding And Scaling

Encoding:

- **Label Encoding**
  - Vehicle Size

- **One Hot Encoding**
  - Engine Fuel Type

Standard Scaling:

Year, Engine HP, Engine Cylinders, highway MPG, city MPG

Feature Engineering & Additional Useful Visualizations

Feature Engineering

- Extracting features such as: Car Age, Horsepower per Cylinder, Price per Horsepower,

	Age	hp_per_cylinder	msrp_per_hp
0	6	55.833333	137.716418
1	6	50.000000	135.500000
2	6	50.000000	121.166667
3	6	38.333333	128.043478
4	6	38.333333	150.000000

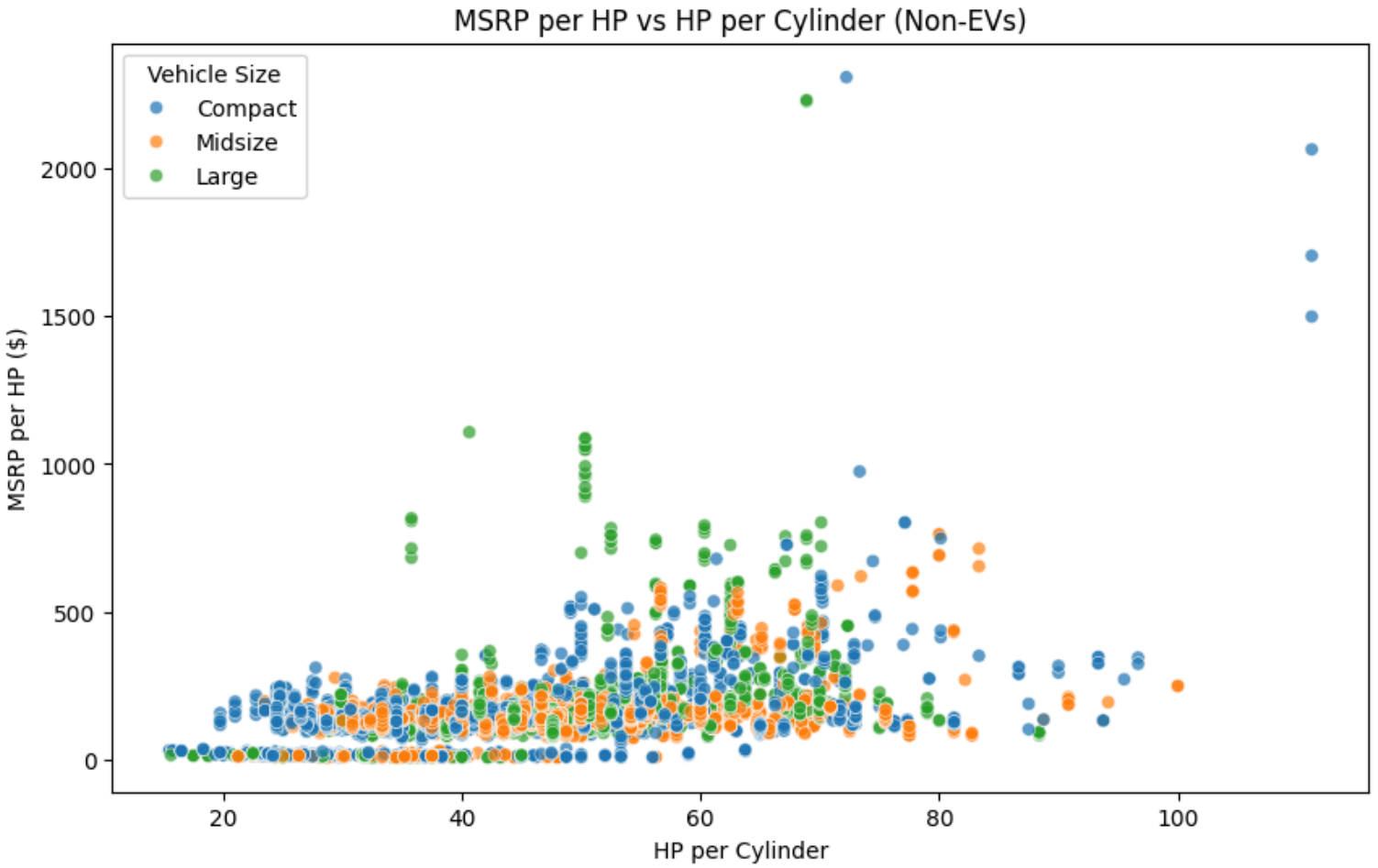
Best Cars Value for Money (EVs filtered out)

Observations

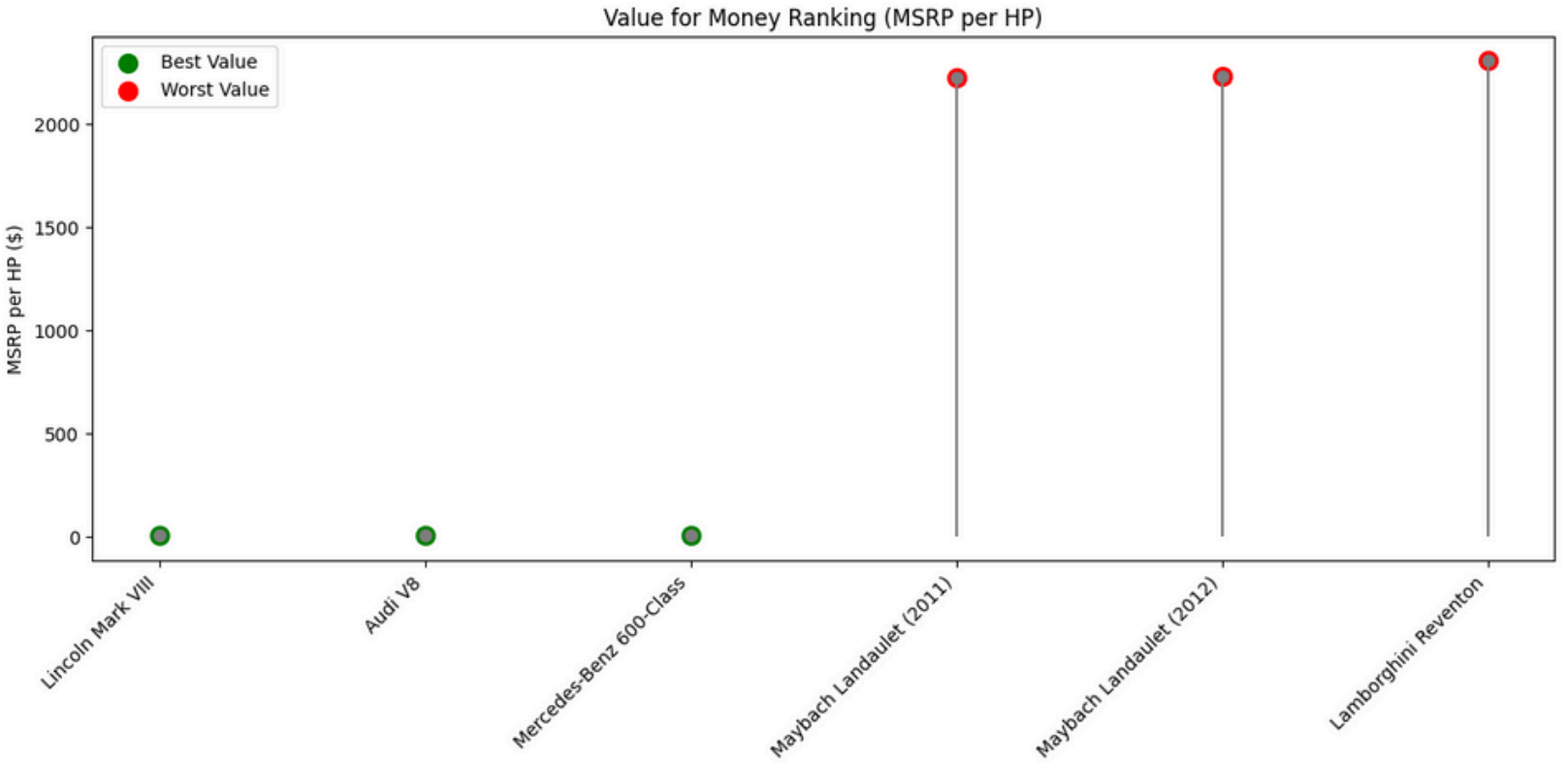
- Clear spread across vehicle sizes
- Compact & midsize = lower cost per HP
- Large vehicles = higher cost, less efficient
- Outliers = overpriced models

Summary

- Best value: compact & midsize
- Large cars = power at higher cost
- Outliers break the trend



Best Cars Value for Money (No Age or Price filtering)



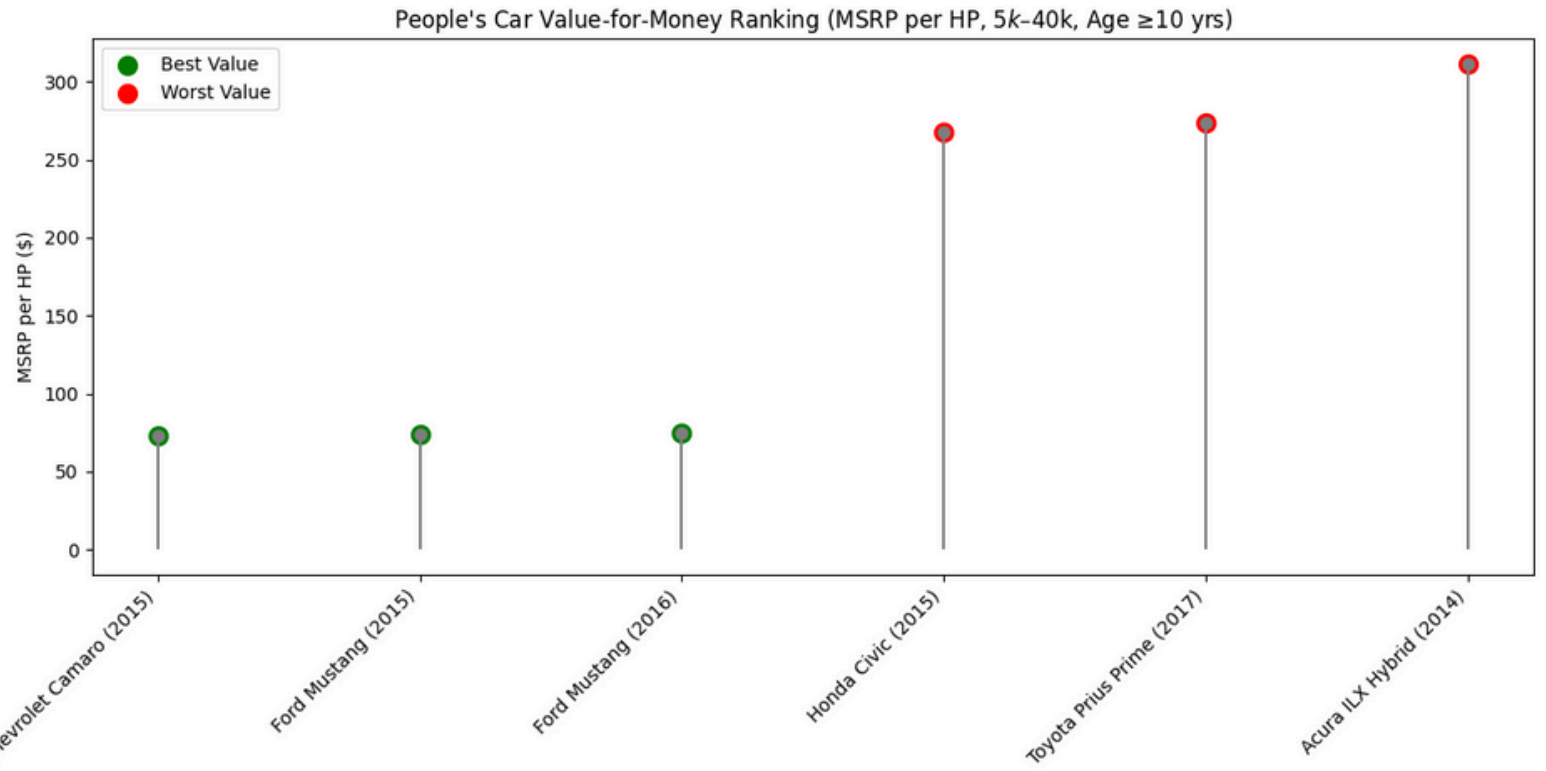
Observations

- Best value: ’90s sedans (Lincoln Mark VIII, Audi V8, Mercedes 600) $7/HP.
- Worst value: Lamborghini Reventon, Maybach Landaulet >$2200/HP.

Summary

- ’90s luxury sedans = high performance per dollar.
- Modern ultra-luxury = poor value, price driven by brand.

Best Cars Value for Money (Age & Price Filtered)



Observations

- Best value: Chevrolet Camaro (2015) $73/HP.
- Worst value: Acura ILX Hybrid (2014) >$310/HP.

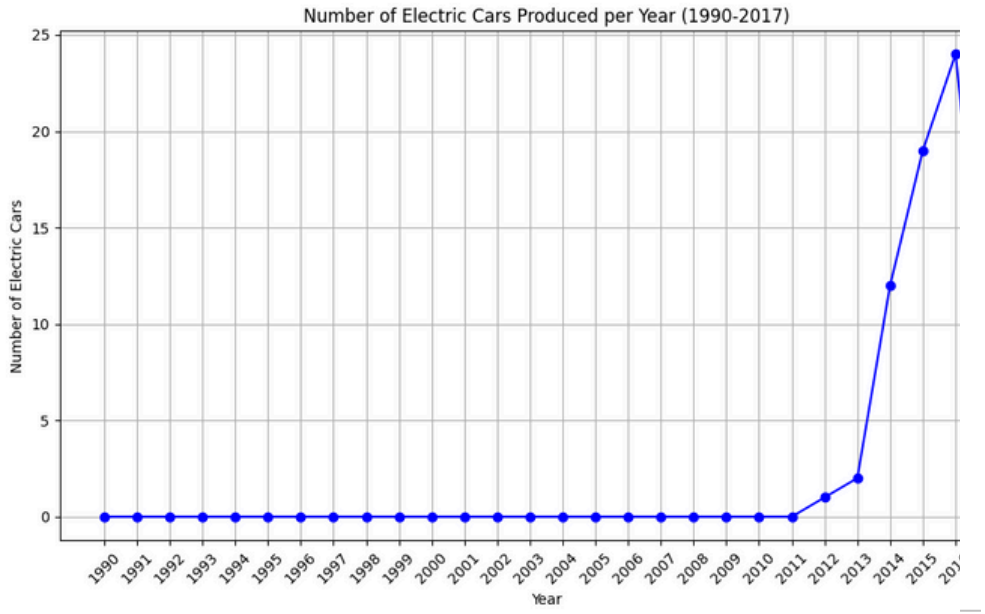
Summary

- Camaro = strong performance per dollar.
- ILX Hybrid = eco focus, weaker cost efficiency.

Conclusion & Final Takeaways

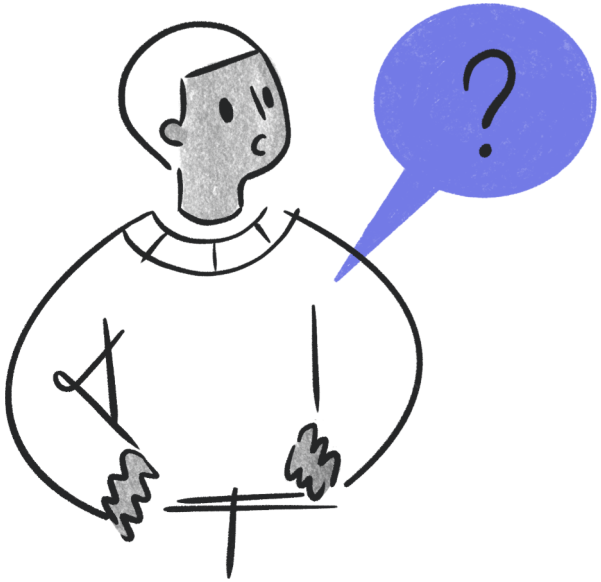
Conclusion

- Study highlights the tension between affordability, performance, and prestige in the car market.
- Electric vehicles emerge as outliers in efficiency, signaling market transition.



Final Takeaways

- Mainstream brands dominate in volume → accessible, practical choices.
- Luxury brands dominate in price → performance & exclusivity, but poor value-for-money.
- Value-for-money lens reveals surprising winners (1990s sedans, Camaro 2015) and losers (Maybach, Acura ILX Hybrid).
- The dataset mirrors real-world automotive trade-offs:
  - Cost vs. performance
  - Tradition vs. innovation
  - Prestige vs. practicality



Any Questions ?

SAMSUNG

Thank you

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations