# Exploratory Data Analysis (EDA) Report

*Generated on: 2026-01-11 16:42*

## Data Loading and Cleaning

- Removed 715 duplicate rows.
- Filled 3 missing values in 'Engine Fuel Type' using mode.
- Filled 69 missing values in 'Engine HP' using median.
- Filled 30 missing values in 'Engine Cylinders' using median.
- Filled 6 missing values in 'Number of Doors' using median.
- Filled 3376 missing values in 'Market Category' using mode.
- Dropped high-cardinality column 'Model' (too many unique categories).
- Dropped high-cardinality column 'Market Category' (too many unique categories).

## EDA Assumptions

- Assumes the dataset represents a single consistent population without major distribution shifts.
- Assumes missing values are Missing At Random (MAR) and can be imputed.
- Assumes rows are independent observations (no time dependency unless stated).
- Outliers are treated as valid extreme behavior unless explicitly removed.

## Data Summary and Descriptive Statistics

### *Column Types*

| column | dtype | unique_values |
|---|---|---|
| Make | object | 48 |
| Model | object | 915 |
| Year | int64 | 28 |
| Engine Fuel Type | object | 10 |
| Engine HP | float64 | 356 |
| Engine Cylinders | float64 | 9 |
| Transmission Type | object | 5 |
| Driven_Wheels | object | 4 |
| Number of Doors | float64 | 3 |
| Market Category | object | 71 |
| Vehicle Size | object | 3 |
| Vehicle Style | object | 16 |
| highway MPG | int64 | 59 |
| city mpg | int64 | 69 |
| Popularity | int64 | 48 |
| MSRP | int64 | 6049 |

### *Summary Statistics (Numeric)*

| feature | count | mean | std | min | 25% | 50% | 75% | max | missing_count | missing_% |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | 11199.0 | 2010.71 | 7.23 | 1990.0 | 2007.0 | 2015.0 | 2016.0 | 2017.0 | 0 | 0.0 |
| Engine HP | 11199.0 | 253.3 | 109.82 | 55.0 | 172.0 | 239.0 | 303.0 | 1001.0 | 0 | 0.0 |
| Engine Cylinders | 11199.0 | 5.67 | 1.79 | 0.0 | 4.0 | 6.0 | 6.0 | 16.0 | 0 | 0.0 |
| Number of Doors | 11199.0 | 3.45 | 0.87 | 2.0 | 2.0 | 4.0 | 4.0 | 4.0 | 0 | 0.0 |
| highway MPG | 11199.0 | 26.61 | 8.98 | 12.0 | 22.0 | 25.0 | 30.0 | 354.0 | 0 | 0.0 |
| city mpg | 11199.0 | 19.73 | 9.18 | 7.0 | 16.0 | 18.0 | 22.0 | 137.0 | 0 | 0.0 |
| Popularity | 11199.0 | 1558.48 | 1445.67 | 2.0 | 549.0 | 1385.0 | 2009.0 | 5657.0 | 0 | 0.0 |
| MSRP | 11199.0 | 41925.93 | 61535.05 | 2000.0 | 21599.5 | 30675.0 | 43032.5 | 2065902.0 | 0 | 0.0 |

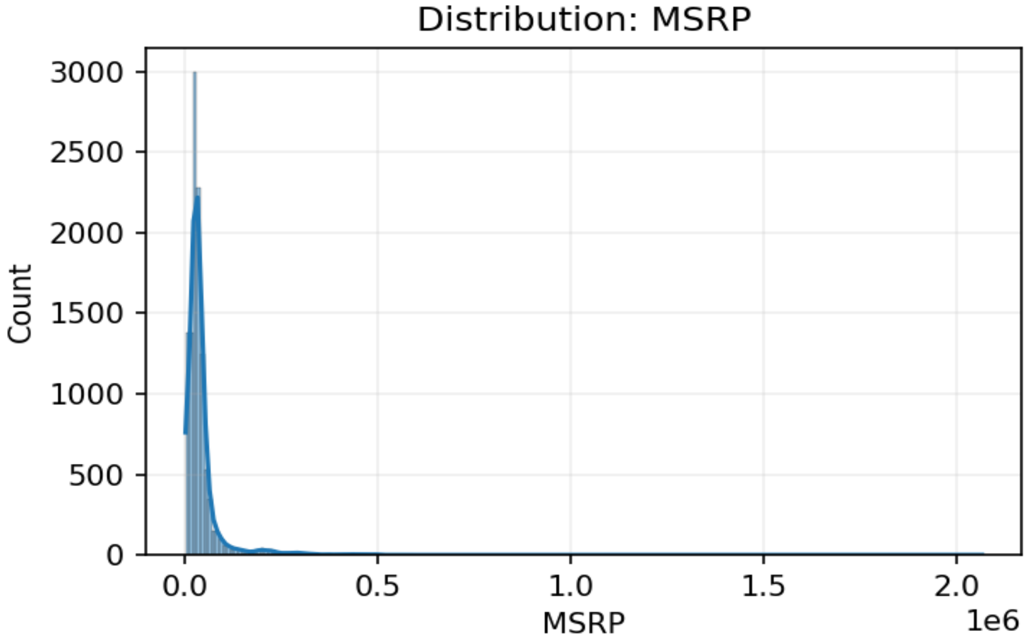
## Correlation Matrix

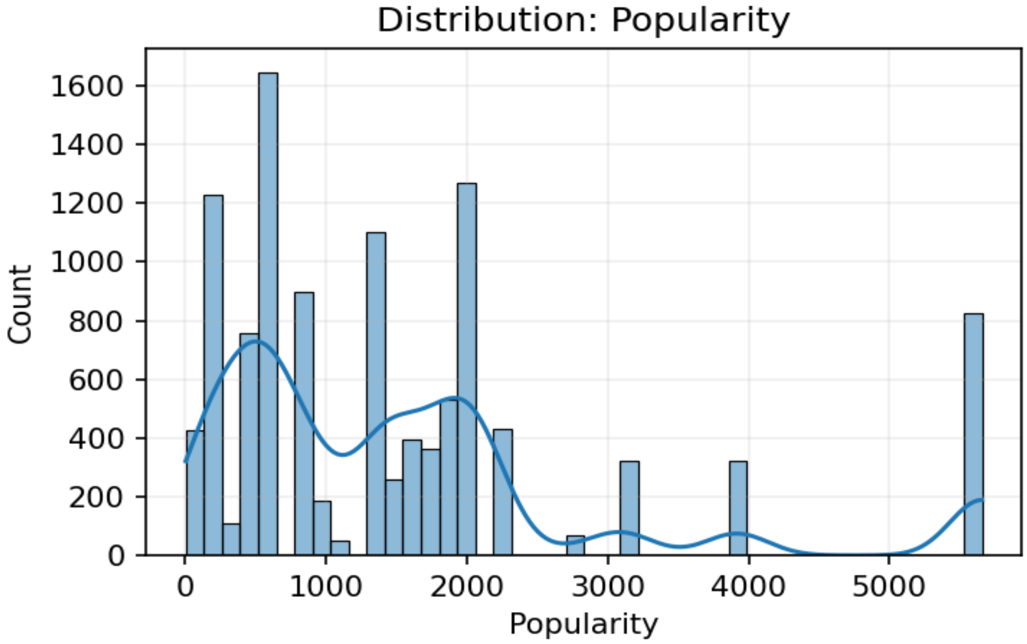
| feature | Year | Engine HP | Engine Cylinders | Number of Doors | highway MPG | city mpg | Popularity | MSRP |
|---|---|---|---|---|---|---|---|---|
| Year | 1.0 | 0.34 | -0.03 | 0.25 | 0.24 | 0.19 | 0.09 | 0.21 |
| Engine HP | 0.34 | 1.0 | 0.77 | -0.13 | -0.36 | -0.36 | 0.04 | 0.66 |
| Engine Cylinders | -0.03 | 0.77 | 1.0 | -0.15 | -0.6 | -0.56 | 0.04 | 0.54 |
| Number of Doors | 0.25 | -0.13 | -0.15 | 1.0 | 0.12 | 0.12 | -0.06 | -0.14 |
| highway MPG | 0.24 | -0.36 | -0.6 | 0.12 | 1.0 | 0.89 | -0.02 | -0.17 |
| city mpg | 0.19 | -0.36 | -0.56 | 0.12 | 0.89 | 1.0 | -0.0 | -0.16 |
| Popularity | 0.09 | 0.04 | 0.04 | -0.06 | -0.02 | -0.0 | 1.0 | -0.05 |
| MSRP | 0.21 | 0.66 | 0.54 | -0.14 | -0.17 | -0.16 | -0.05 | 1.0 |

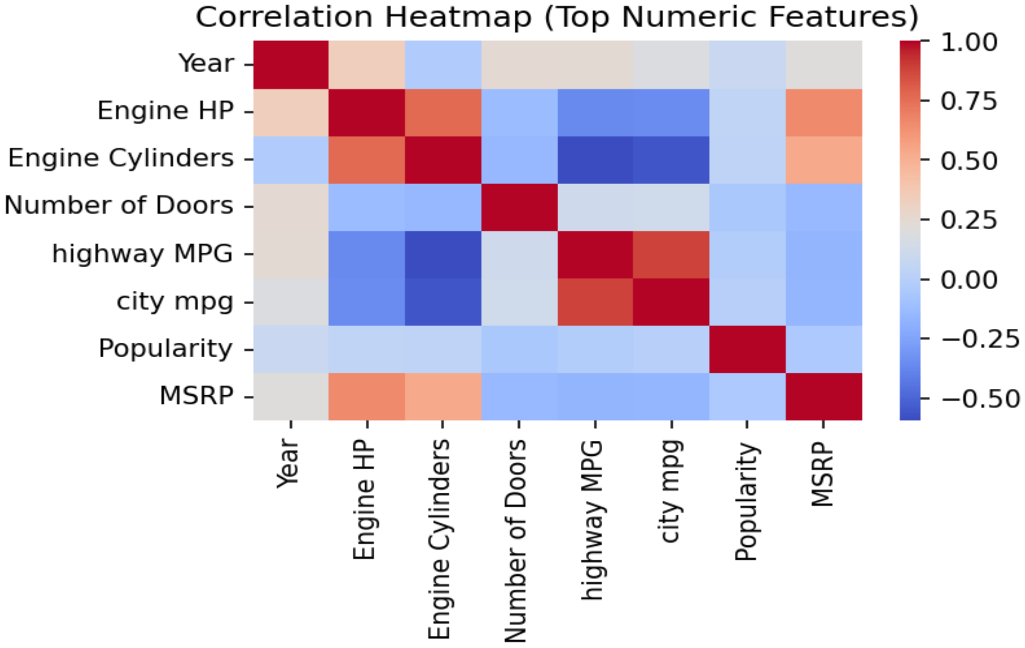
### *Top Correlations*

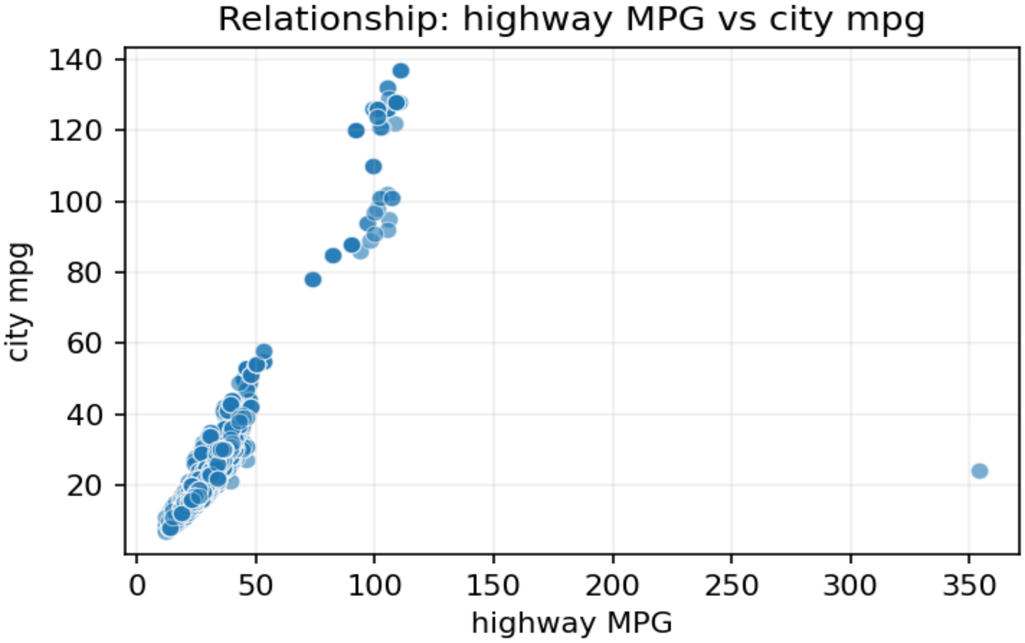
| feature_1 | feature_2 | abs_corr |
|---|---|---|
| highway MPG | city mpg | 0.89 |
| Engine HP | Engine Cylinders | 0.77 |
| Engine HP | MSRP | 0.66 |
| highway MPG | Engine Cylinders | 0.6 |
| Engine Cylinders | city mpg | 0.56 |
| MSRP | Engine Cylinders | 0.54 |
| highway MPG | Engine HP | 0.36 |
| city mpg | Engine HP | 0.36 |
| Engine HP | Year | 0.34 |
| Year | Number of Doors | 0.25 |

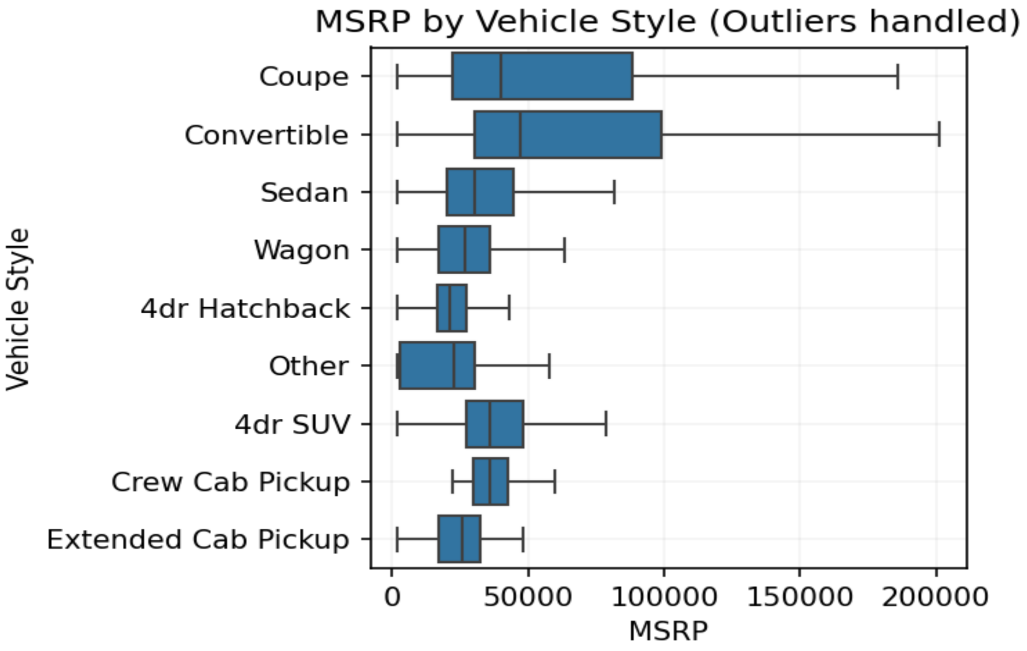
## Visual Analysis

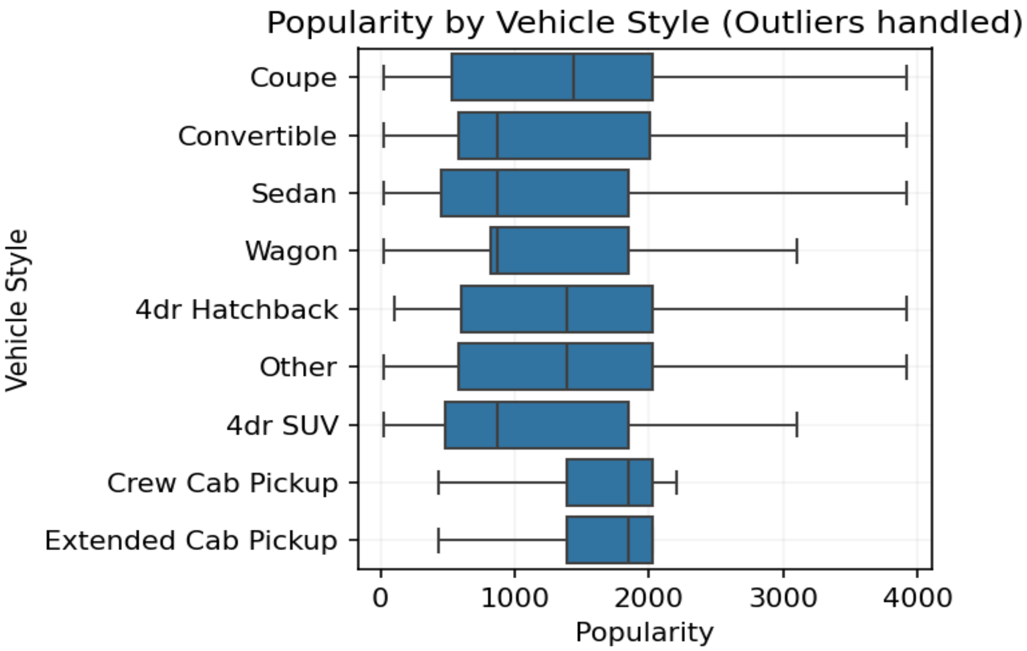

Distribution: MSRP
Count
3000
2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
1e6
MSRP


Distribution: Popularity
Count
1600
1400
1200
1000
800
600
400
200
0
0
1000
2000
3000
4000
5000
Popularity


Correlation Heatmap (Top Numeric Features)
Year
Engine HP
Engine Cylinders
Number of Doors
highway MPG
city mpg
Popularity
MSRP
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50


Relationship: highway MPG vs city mpg
city mpg
140
120
100
80
60
40
20
0
50
100
150
200
250
300
350
highway MPG





## 4. AI Narrative Insights

## Exploratory Data Analysis (EDA) Report

### *Executive Summary*

The dataset contains 11199 rows and 16 columns after cleaning. The top correlations indicate strong relationships between MPG, Engine HP, and Engine Cylinders. Feature engineering involved dropping high-cardinality columns 'Model' and 'Market Category'. Missing values were filled using mode or median, depending on the column.

### *Introduction*

This Exploratory Data Analysis (EDA) report aims to provide an understanding of the relationships between key variables in the dataset. The analysis involves data cleaning, feature engineering, and statistical analysis to identify patterns and correlations.

## *Data Overview*

The dataset contains 11199 rows and 16 columns. The top correlations indicate strong relationships between:

- Highway MPG and city MPG (0.89)
- Engine HP and Engine Cylinders (0.77)
- Engine HP and MSRP (0.66)
- Highway MPG and Engine Cylinders (0.6)
- Engine Cylinders and city MPG (0.56)

## *Data Loading and Cleaning*

The dataset was cleaned by removing 715 duplicate rows. Missing values were filled using the following methods:

- Engine Fuel Type: 3 missing values filled using mode
- Engine HP: 69 missing values filled using median
- Engine Cylinders: 30 missing values filled using median
- Number of Doors: 6 missing values filled using median
- Market Category: 3376 missing values filled using mode

## *Feature Engineering*

Feature engineering involved dropping high-cardinality columns:

- Model (too many unique categories)
- Market Category (too many unique categories)

## *Correlation Analysis*

The top correlations indicate strong relationships between:

- Highway MPG and city MPG (0.89)
- Engine HP and Engine Cylinders (0.77)
- Engine HP and MSRP (0.66)
- Highway MPG and Engine Cylinders (0.6)
- Engine Cylinders and city MPG (0.56)

## *Conclusions and Recommendations*

Based on the analysis, the following conclusions can be drawn:

- There are strong relationships between MPG, Engine HP, and Engine Cylinders.
- The feature engineering process involved dropping high-cardinality columns to improve model performance.
- The missing values analysis indicates that the dataset is mostly complete, with some missing values filled using mode or median.

## *Next Steps*

Future analysis should focus on:

- Building a predictive model using the engineered features.
- Exploring the relationships between other variables in the dataset.

- Investigating the impact of the dropped high-cardinality columns on model performance.