Analyzing MSRP Trends: A Data-Driven Study of Vehicle Pricing from 2007 to 2017

Unit 1

Avinash Bunga

Master of Science in Information Systems and Business Analytics

Park University

CIS607HOF1P2024 Applied Business Forecasting

Professor: Dr. Abdelmonaem Jornaz

Aug 18, 2024

Analyzing MSRP Trends: A Data-Driven Study of Vehicle Pricing from 2007 to 2017

Introduction

In the automotive industry, the Manufacturer's Suggested Retail Price (MSRP) has a significant impact on consumer purchase decisions, brand positioning, and overall market competition. Understanding what drives MSRP and how prices have changed over time can provide significant information to manufacturers, dealerships, and consumers (Shih, n.d.; Taylor, n.d.).

This research analyzes MSRP trends from 2007 to 2017, focusing on three car sizes: compact, midsize, and large. The purpose of investigating how MSRP fluctuates among different categories is to identify trends that may represent broader market dynamics, technology developments, and changes in customer preferences during this time span.

The analysis uses a dataset containing numerous automotive parameters such as vehicle size, engine specifications, fuel efficiency, and MSRP for each model. The goal is to compute essential descriptive statistics, develop visuals to compare pricing across vehicle sizes and investigate relevant questions for predictive analytics.

This study aims to provide a detailed summary of how MSRP has changed over the last decade and the reasons that may have caused those changes. The findings of this report may serve as the foundation for future research on pricing strategies and market trends in the automotive industry (InfinitiResearch, 2018; Shih, n.d.; Taylor, n.d.).

Data Collection and Preparation

Dataset Description

The dataset used in this analysis is titled **"Car Features and MSRP"** and was sourced from Kaggle. This dataset provides a comprehensive view of various automobile features and their corresponding Manufacturer's Suggested Retail Price (MSRP). It includes

data for a wide range of vehicles manufactured between 1990 and 2017, offering a broad perspective on the automotive market over nearly three decades (Encora, 2023; Shih, n.d.).

Key Features of the Dataset:

- **MSRP (Manufacturer's Suggested Retail Price):** This is the primary variable of interest in the analysis. MSRP represents the price at which the manufacturer recommends that the vehicle be sold. It serves as a critical reference point for both dealerships and consumers.
- **Vehicle Size:** The dataset categorizes vehicles into different size segments, including Compact, Midsize, and Large. This variable is essential for understanding how vehicle size influences pricing strategies.
- **Year:** The dataset includes the manufacturing year for each vehicle, allowing for a temporal analysis of MSRP trends over time.
- **Engine Specifications**
    - **Engine HP:** Horsepower of the vehicle's engine, which is often a key factor in determining vehicle performance and, consequently, its price.
    - **Engine Cylinders:** The number of cylinders in the vehicle's engine, which can affect both performance and fuel efficiency.
- **Fuel Efficiency**
    - **City MPG (Miles Per Gallon):** Fuel efficiency measured in miles per gallon during city driving conditions.
    - **Highway MPG (Miles Per Gallon):** Fuel efficiency measured in miles per gallon during highway driving conditions.
- **Additional Features**
    - **Transmission Type:** The type of transmission (e.g., automatic or manual), which can influence both vehicle performance and consumer preference.

- **Driven Wheels:** Information on whether the vehicle is front-wheel drive, rear-wheel drive, or all-wheel drive.
- **Market Category:** Broad categories such as Luxury, Performance, or Economy, providing context for the vehicle's target market (Encora, 2023; Shih, n.d.).

Checking for Missing Values

I used Python to check for any missing values to ensure the dataset was complete and ready for analysis. Identifying and addressing missing values is crucial because complete data can lead to accurate analysis results. Below is the Python script that I used to check for missing values in the dataset.

Python Script

```
import pandas as pd
# Load the dataset
file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1 Assignment/Car Features and MSRP.csv'
car_data = pd.read_csv(file_path)
# Check for missing values in the dataset
missing_values = car_data.isnull().sum()
print("Checking for missing values...\n")
print(missing_values)
# Display rows with missing values
missing_data = car_data[car_data.isnull().any(axis=1)]
print("Rows with missing values:\n")
print(missing_data)
```

This script loads the dataset and checks for missing values across all columns. The isnull().sum() function is used to count the number of missing entries in each column. After identifying the columns with missing values, the script displays the specific rows where data is missing (Datagy, 2022; MachineLearningTutorials, 2023).

Figure 1

Output

The script revealed that several rows in the dataset had missing values. Below is a screenshot showing the output of the script, which lists the rows containing missing data.

```
(myenv) avinash@avinashs-MacBook-Pro PY % python3 check_missing_values.py
Checking for missing values...

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      3742
Vehicle Size            0
Vehicle Style           0
highway MPG             0
city mpg                0
Popularity              0
MSRP                    0
dtype: int64
Rows with missing values:

         Make     Model  Year  ... city mpg  Popularity   MSRP
87     Nissan     200SX  1996  ...       26        2009   2000
88     Nissan     200SX  1996  ...       26        2009   2000
91     Nissan     200SX  1997  ...       25        2009   2000
92     Nissan     200SX  1997  ...       25        2009   2000
93     Nissan     200SX  1998  ...       25        2009   2000
...       ...       ...   ...  ...      ...         ...    ...
11794  Subaru        XT  1991  ...       16         640   2000
11809  Toyota  Yaris iA  2017  ...       30        2031  15950
11810  Toyota  Yaris iA  2017  ...       32        2031  17050
11867     GMC     Yukon  2015  ...       15         549  64520
11868     GMC     Yukon  2015  ...       14         549  67520

[3830 rows x 16 columns]
(myenv) avinash@avinashs-MacBook-Pro PY %
```

As seen in Figure 1, the dataset had 3,830 rows with missing values spread across 16 columns. These missing values were distributed among various features, such as Engine HP, Engine Cylinders, and Market Category.

Handling Missing Data

After identifying the rows with missing data, the next step was to decide how to handle these incomplete entries. Given that missing data can skew analysis results, it was crucial to address this issue before proceeding further.

Python Script

Below is the Python script used to drop rows with missing data and to check the impact on the dataset:

```
import pandas as pd
#Load the Data
file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1 Assignment/Car Features and MSRP.csv'
car_data = pd.read_csv(file_path)
# Step 1: Check the total number of rows before dropping missing data
total_rows_before = car_data.shape[0]
print(f"Total number of rows before dropping missing data: {total_rows_before}")
# Step 2: Drop rows with missing data
car_data_cleaned = car_data.dropna()
# Step 3: Check the total number of rows after dropping missing data
total_rows_after = car_data_cleaned.shape[0]
print(f"Total number of rows after dropping missing data: {total_rows_after}")
# Save the cleaned data to a new CSV file
```

```
car_data_cleaned.to_csv('/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/Car_Features_and_MSRP_Cleaned.csv', index=False)
```

This script first checks the total number of rows in the dataset before and after dropping the rows with missing values. The dropna() function is used to remove any rows that contain missing data (Datagy, 2022; MachineLearningTutorials, 2023).

Figure 2

Output

```
(myenv) avinash@avinashs-MacBook-Pro PY % python3 drop_missing_data.py
Total number of rows before dropping missing data: 11914
Total number of rows after dropping missing data: 8084
(myenv) avinash@avinashs-MacBook-Pro PY %
```

As seen in Figure 2, The output shows that the dataset originally contained 11,914 rows, but after dropping the rows with missing data, 8,084 rows remained.

A new file named **"Car_Features_and_MSRP_Cleaned.csv"** was created, containing only the complete data. This cleaned dataset is now ready for further analysis.

Rationale for Dropping Missing Data:

The decision to drop rows with missing data was based on the consideration that the affected rows could introduce bias or inaccuracies into the analysis. While the removal of approximately 32% of the data might seem significant, the remaining 8,084 rows provided a sufficient dataset for conducting robust statistical analysis (Encora, 2023; Shih, n.d.).

Analyzing Data Entries by Year

To ensure the dataset provided a sufficient and relevant number of data points for analysis, I examined the distribution of entries by year. Understanding the timeline of the data is crucial for making informed decisions about which years to include in the analysis, as this can impact the validity and relevance of the results (Encora, 2023; Shih, n.d.).

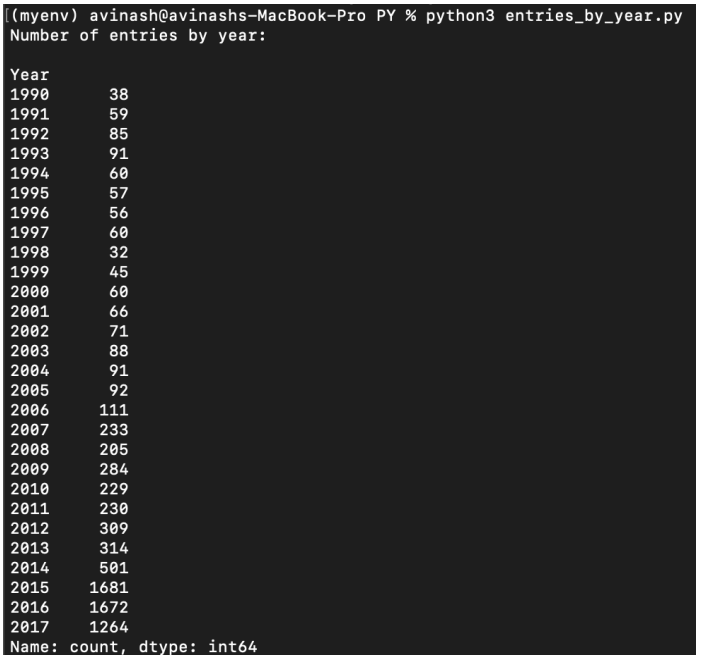
Python Script

```
import pandas as pd

# Load the cleaned data

file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1

Assignment/Car_Features_and_MSRP_Cleaned.csv'

car_data_cleaned = pd.read_csv(file_path)

# Step 1: Group the data by Year and count the number of entries for each year

entries_by_year = car_data_cleaned['Year'].value_counts().sort_index()

# Print the number of entries for each year

print("Number of entries by year:\n")

print(entries_by_year)
```

This script loads the cleaned dataset and then counts the number of entries available for each year. The results are sorted by year to provide a clear view of how the data is distributed over time (Solomon, n.d.; SparkByExamples, n.d.).

Figure 3

Output



Decision to Focus on 2007 to 2017

As seen in Figure 3, the dataset originally included data from 1990 to 2017, the number of entries before 2007 was relatively low, making it less reliable for analysis. For instance, there were only 38 entries in 1990, with similarly sparse data in the years that followed. This limited data availability in the earlier years could lead to biases if included in the analysis.

After carefully examining the number of entries per year, I decided to focus on the period from 2007 to 2017. This timeframe provided a much more substantial number of data points each year, ensuring a robust and reliable analysis. Additionally, by limiting the analysis to the most recent 10 years, the findings would be more relevant to current market trends, avoiding the potential biases introduced by older, less representative data (Encora, 2023; Shih, n.d.).

Filtering the Dataset for Analysis

To focus the analysis on the most relevant and recent data, I decided to filter the dataset to include only entries from 2007 to 2017. This step was essential to ensure that the analysis reflects current market trends and avoids the potential biases of older data.

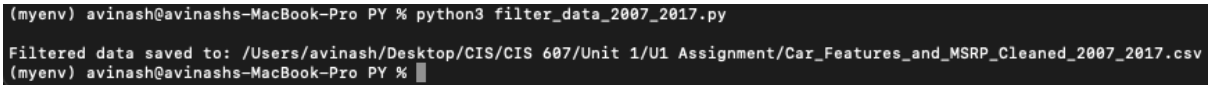
Python Script

```
import pandas as pd
# Load the cleaned dataset
file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/Car_Features_and_MSRP_Cleaned.csv'
car_data_cleaned = pd.read_csv(file_path)
# Filter the data to include only the years 2007-2017
filtered_data = car_data_cleaned[(car_data_cleaned['Year'] >= 2007) &
(car_data_cleaned['Year'] <= 2017)]
# Save the filtered dataset to a new file
filtered_file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/Car_Features_and_MSRP_Cleaned_2007_2017.csv'
filtered_data.to_csv(filtered_file_path, index=False)
print(f"Filtered data saved to: {filtered_file_path}")
```

This script first loads the cleaned dataset and then filters it to include only the rows where the Year is between 2007 and 2017. The filtered dataset is then saved to a new file, ensuring that all subsequent analysis focuses solely on this time period (Datagy, n.d.; Tunalı, 2021).

Figure 4

Output



As seen in Figure 4, the filtered data was successfully saved to a new file named "Car_Features_and_MSRP_Cleaned_2007_2017.csv", This new file contains only the data from 2007 to 2017 and is the basis for the subsequent analysis.

Categorical Variable: Vehicle Size

In this analysis, the categorical variable selected for segmentation is **"Vehicle Size".** This variable classifies vehicles into three main categories: Compact, Midsize, and Large.

- **Compact Vehicles:** These are generally smaller vehicles that are designed for city driving and efficiency. They tend to have lower MSRPs and are popular among budget-conscious consumers.
- **Midsize Vehicles:** This category represents a middle ground between compact and large vehicles. Midsize vehicles often offer a balance of space, comfort, and fuel efficiency, making them popular choices for families and professionals.
- **Large Vehicles:** Large vehicles, including SUVs and trucks, are known for powerful engines, spacious interiors, and higher price points. They are often chosen for their capacity and performance, particularly in regions where larger vehicles are favored.

Reason for Selection: The Vehicle Size variable was chosen because it is a crucial factor influencing the MSRP of a vehicle. Different vehicle sizes cater to different market segments, and understanding how MSRP varies across these categories can provide insights into pricing strategies and consumer preferences within the automotive industry (Encora, 2023; Learn Statistics Easily, 2024; Shih, n.d.).

Descriptive Statistics for MSRP by Vehicle Size (2007-2017)

To gain insights into how the Manufacturer's Suggested Retail Price (MSRP) varies across different vehicle sizes, I calculated key descriptive statistics for each category (Compact, Midsize, and Large) using the filtered dataset from 2007 to 2017.

Python Script

```
import pandas as pd
# Load the new dataset for 2007-2017
file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/Car_Features_and_MSRP_Cleaned_2007_2017.csv'
car_data_filtered = pd.read_csv(file_path)
# Segment the data by Vehicle Size and calculate descriptive statistics for MSRP
msrp_by_vehicle_size = car_data_filtered.groupby('Vehicle Size')['MSRP'].agg(['count',
'mean', 'std', 'min', 'median', 'max'])
# Print the descriptive statistics
print("Descriptive statistics for MSRP by Vehicle Size (2007-2017):\n")
print(msrp_by_vehicle_size)
```

This script loads the filtered dataset and then segments the data by the "Vehicle Size" category. It calculates the count, mean, standard deviation, min, median, and max values for the MSRP within each vehicle size category (Bobbitt, 2022; Shane, n.d.).

Figure 5

Output

```
(myenv) avinash@avinashs-MacBook-Pro PY % python3 msrp_stats_2007_2017.py
Descriptive statistics for MSRP by Vehicle Size (2007-2017):

              count          mean           std    min   median      max
Vehicle Size
Compact        2435  47913.400000  84284.627001  11965  28700.0  2065902
Large          1709  69175.887068  81306.540869  15840  45860.0  1382750
Midsize        2778  49974.735781  49146.339232  18995  38577.5   548800
```

Interpretation from Figure 5

- **Compact Vehicles:** The average MSRP for Compact vehicles is $47,913.40, with a wide range from $11,965 to over $2,065,902, indicating a diverse market segment with both budget and luxury models.
- **Large Vehicles:** Large vehicles have the highest average MSRP at $69,175.89, with prices ranging from $15,840 to $1,382,750. This suggests that larger vehicles tend to be more expensive, likely due to higher manufacturing costs and added features.
- **Midsize Vehicles:** Midsize vehicles have an average MSRP of $49,974.74, with a median price of $38,577.50. The prices for Midsize vehicles are more concentrated, with a smaller range compared to Compact and Large vehicles.

Box Plot of MSRP by Vehicle Size (2007-2017)

To better understand how the Manufacturer's Suggested Retail Price (MSRP) varies across different vehicle sizes, I created a box plot using Python. This plot provides a visual representation of how MSRP values differ among Compact, Midsize, and Large vehicles during the period from 2007 to 2017.

Python Script

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

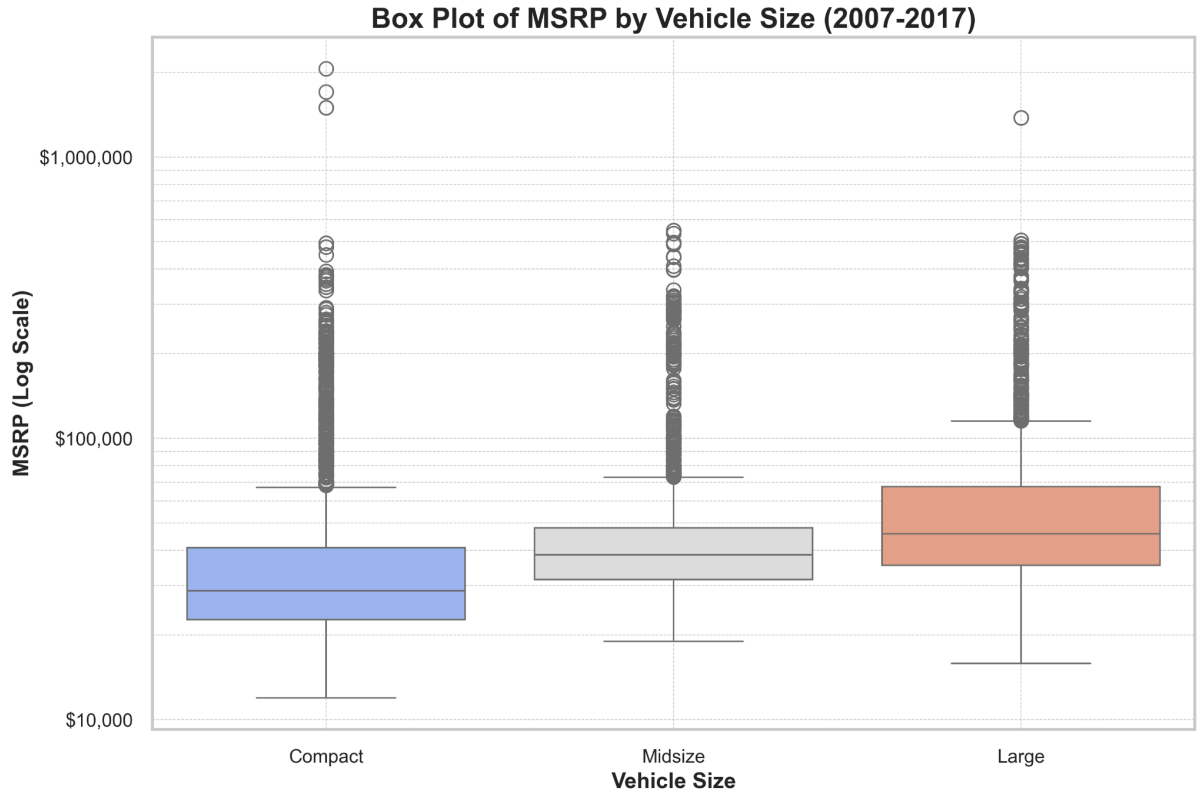
```
import matplotlib.ticker as mticker

# Load the final dataset for analysis

file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1

Assignment/Car_Features_and_MSRP_Cleaned_2007_2017.csv'

car_data_filtered = pd.read_csv(file_path)

# Set the style and context for the plot

sns.set(style="whitegrid", context="talk")

# Create a box plot for MSRP by Vehicle Size

plt.figure(figsize=(12, 8))

sns.boxplot(x='Vehicle Size', y='MSRP', data=car_data_filtered, palette="coolwarm")

# Customize the plot

plt.title('Box Plot of MSRP by Vehicle Size (2007-2017)', fontsize=18, weight='bold')

plt.xlabel('Vehicle Size', fontsize=14, weight='bold')

plt.ylabel('MSRP (Log Scale)', fontsize=14, weight='bold')

# Format the y-axis to show dollar values in a readable format

plt.yscale('log')  # Log scale for better visualization of the spread

plt.gca().yaxis.set_major_formatter(mticker.FuncFormatter(lambda x, _: f'${int(x):,}'))

# Customize ticks for better readability

plt.xticks(fontsize=12)

plt.yticks(fontsize=12)

plt.grid(True, which="both", ls="--", linewidth=0.5)

# Save the plot to the same directory

output_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1

Assignment/box_plot_msrp_vehicle_size.png'

plt.savefig(output_path, bbox_inches='tight', dpi=300)
```

```
# Show the plot

plt.show()

print(f"Box plot saved to: {output_path}")
```

This script uses Seaborn and Matplotlib to create a high-quality box plot, and I saved the plot as a PNG file in the given directory (Datacamp, 2024; McCullum, n.d.).

Figure 6

Output



Interpretation from Figure 6

- **Compact Vehicles:** The median MSRP for compact vehicles is the lowest among the categories, but the plot reveals significant outliers. This suggests that some luxury compact models have much higher prices.

- **Midsize Vehicles:** The midsize category shows a slightly higher median MSRP, with a more consistent distribution of prices compared to compact vehicles.
- **Large Vehicles:** Large vehicles have the highest median MSRP, reflecting their position as more expensive options in the market. The spread of prices is wider, with several outliers on the higher end, indicating a range of premium models within this category.

Histogram of MSRP (2007-2017)

To further explore the distribution of Manufacturer's Suggested Retail Price (MSRP) for vehicles between 2007 and 2017, I generated a histogram. This visualization helps to understand how the prices are distributed across different ranges.

Python Script

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.ticker as mticker
import numpy as np
# Load the final dataset for analysis
file_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/Car_Features_and_MSRP_Cleaned_2007_2017.csv'
car_data_filtered = pd.read_csv(file_path)
# Set the style and context for the plot
sns.set(style="whitegrid", context="talk")
# Create a color palette with a temperature range
palette = sns.color_palette("coolwarm", as_cmap=True)
# Create a histogram for MSRP with the color palette and log scale on X-axis
```

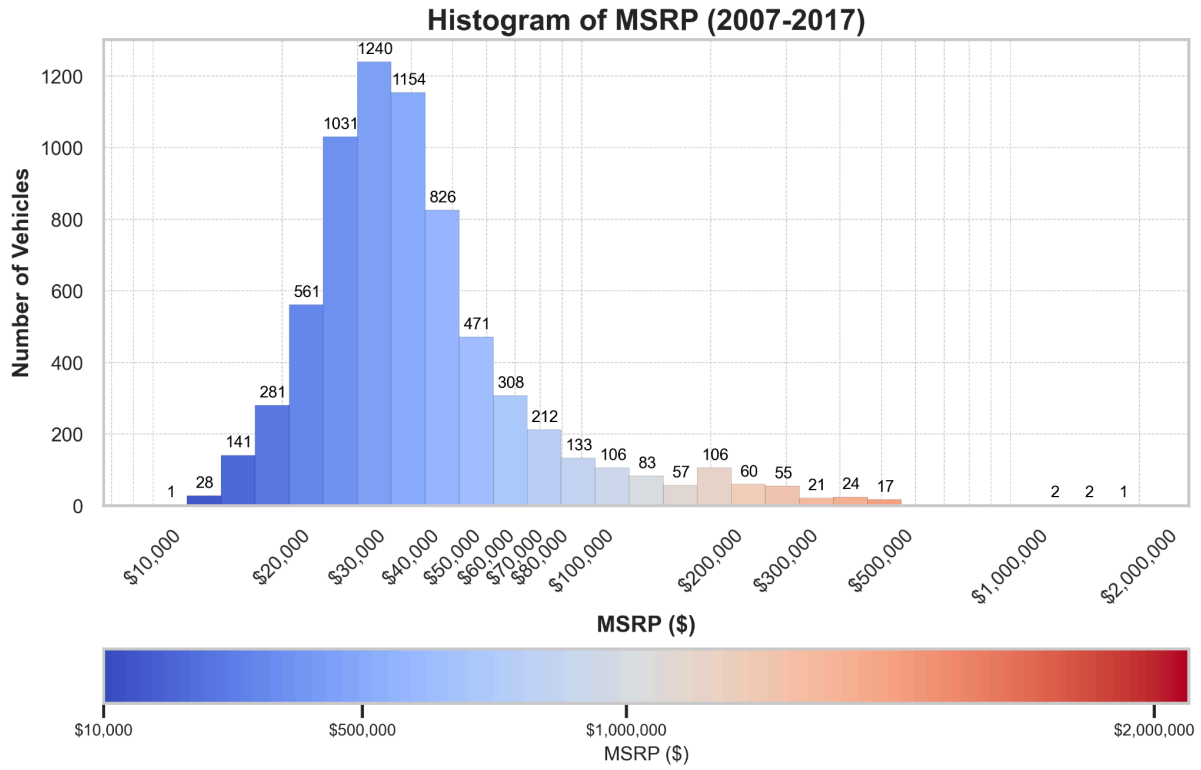
```
plt.figure(figsize=(12, 8))
hist = sns.histplot(car_data_filtered['MSRP'], bins=np.logspace(np.log10(10000),
np.log10(2000000), 30), kde=False, edgecolor='black')
# Annotate each bar with the count value
for p in hist.patches:
  height = p.get_height()
  if height > 0:  # Only annotate bars with counts greater than 0
    hist.annotate(f'{int(height)}', xy=(p.get_x() + p.get_width() / 2, height),
                  xytext=(0, 5), textcoords='offset points', ha='center', fontsize=10, color='black')
# Customize the plot
plt.title('Histogram of MSRP (2007-2017)', fontsize=18, weight='bold')
plt.xlabel('MSRP ($)', fontsize=14, weight='bold')
plt.ylabel('Number of Vehicles', fontsize=14, weight='bold')
# Apply a logarithmic scale to the X-axis
plt.xscale('log')
# Format the X-axis to show specific dollar values at log scale
plt.gca().xaxis.set_major_formatter(mticker.FuncFormatter(lambda x, _: f'${int(x):,}'))
# Add more specific tick marks for readability, including the highest value
plt.gca().set_xticks([10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 100000,
200000, 300000, 500000, 1000000, 2000000])
# Apply the color gradient to the bars
for i, patch in enumerate(hist.patches):
  patch.set_facecolor(palette(i / len(hist.patches)))
# Add a color bar legend with $ values and reduce its size for better readability
```

```
sm = plt.cm.ScalarMappable(cmap=palette,
norm=plt.Normalize(vmin=car_data_filtered['MSRP'].min(),
vmax=car_data_filtered['MSRP'].max()))
sm.set_array([])
cbar = plt.colorbar(sm, ax=hist.axes, orientation='horizontal', pad=0.2)
cbar.set_label('MSRP ($)', fontsize=12)
cbar.set_ticks([10000, 500000, 1000000, 2000000])
cbar.ax.set_xticklabels(['$10,000', '$500,000', '$1,000,000', '$2,000,000'], fontsize=10)
# Customize ticks for better readability
plt.xticks(fontsize=12, rotation=45)
plt.yticks(fontsize=12)
plt.grid(True, which="both", ls="--", linewidth=0.5)
# Save the histogram to the same directory
output_path = '/Users/avinash/Desktop/CIS/CIS 607/Unit 1/U1
Assignment/histogram_msrp_final_detailed.png'
plt.savefig(output_path, bbox_inches='tight', dpi=300)
# Show the plot
plt.show()
print(f"Histogram saved to: {output_path}")
```

(Datagy, 2023; Solomon, n.d.).

Figure 7

Output



Interpretation from Figure 7

- **Distribution of Prices:** The histogram clearly shows that the majority of vehicles have an MSRP between $20,000 and $40,000. The distribution skews towards the lower price range, with fewer vehicles in the higher price brackets.
- **Logarithmic Scale:** The logarithmic scale on the X-axis was used to effectively visualize the wide range of prices, from $10,000 to $2,000,000.
- **Color Gradient:** The bars are color-coded with a temperature range, where cooler colors represent lower prices and warmer colors represent higher prices. This gradient provides a visual representation of the distribution across different price ranges.
- **Annotations:** Each bar is annotated with the exact number of vehicles in that price range, which helps to understand the concentration of vehicles in each price bracket quickly.

- **Legend:** A color bar legend is included below the histogram to help interpret the color gradient in terms of MSRP values, enhancing the readability of the visualization.

Overall Data Analysis Summary

In the previous sections, I have already explained what each output tells us about the data. In this section, I am putting everything together to discuss what these findings mean in a broader context.

1. **Descriptive Statistics Overview**:
    - **MSRP by Vehicle Size**: I found that larger vehicles generally have a higher average MSRP, with an average of $69,175 compared to $49,974 for midsize vehicles and $47,913 for compact vehicles. The standard deviation is also higher for large vehicles ($81,306), meaning there is more variation in their prices. This suggests that in the “Large” category, prices can range from more affordable to very expensive, likely due to a mix of standard and luxury models.
    - **Price Range and Market Segmentation**: There are some high-priced outliers in the data, especially in the compact and midsize categories. These outliers indicate that there are a few very expensive vehicles in these categories, likely high-performance or luxury models. For example, in the compact category, prices go as high as $2,065,902.
2. **Visualizations Overview**:
    - **Box Plot of MSRP by Vehicle Size**: The box plot confirms what I saw in the descriptive statistics. Larger vehicles not only have higher median prices, but they also have more outliers. These outliers likely represent luxury models.

- **Histogram of MSRP**: The histogram shows that most vehicles are priced on the lower end, with the majority under $50,000. There is a sharp drop-off as prices increase, indicating that while there are some high-priced vehicles, they are rare. The highest number of vehicles (1,240) are priced around $30,000, highlighting a significant concentration in this price range.

3. **Market Implications**:
    - The data suggests that from 2007 to 2017, the automobile market was mostly focused on lower-priced vehicles, with fewer vehicles falling into the higher price ranges. This could mean that there was a strong demand for more affordable cars, with fewer luxury models being sold. However, the presence of high-priced outliers in all categories shows that there was still a market for premium vehicles, even though they were less common.
    - For car manufacturers and marketers, this information can be helpful when deciding on pricing strategies. I have shown that while most consumers were buying cars priced under $50,000, there was still some demand for higher-priced models, especially in the large vehicle category.

This summary ties together the key findings and offers a clearer picture of the trends in vehicle pricing during the 2007 to 2017 period (Datacamp, 2024; Encora, 2023; Shih, n.d.; Solomon, n.d.).

Discussion on Predictive Analytics Questions Applicable

In this section, I will focus on four key predictive analytics questions that are specifically tailored to the "Car Features and MSRP" dataset from 2007 to 2017. These questions can provide valuable insights into vehicle pricing, market trends, and consumer behavior.

1. **Predicting MSRP Based on Vehicle Features**
    - **Header Name:** "MSRP" Prediction by "Vehicle Size", "Engine HP", "Engine Cylinders"
    - **Question:** How can I predict the MSRP of a vehicle based on features like "Vehicle Size", "Engine HP", and "Engine Cylinders"?
    - **Relation to Dataset:** The dataset includes detailed information on various vehicle features, such as "Engine HP", "Engine Cylinders", and "Vehicle Size". By analyzing this data, I can create predictive models to estimate the MSRP for new vehicles based on these characteristics. This helps manufacturers set competitive prices that reflect the value of these features and align with market trends.
2. **Identifying Factors That Influence High MSRP Vehicles**
    - **Header Name:** High "MSRP" Analysis by "Make" and "Engine HP"
    - **Question:** What specific vehicle features and brands (from the "Make" column) are most strongly correlated with higher MSRPs?
    - **Relation to Dataset:** The "Make" column in the dataset allows for an analysis of whether certain brands consistently command higher prices. By examining features such as "Engine HP" and "Engine Cylinders", I can identify the factors that most influence a vehicle's MSRP. Understanding these factors can guide manufacturers in designing high-value vehicles that justify premium pricing and cater to consumer expectations.
3. **Market Trend Prediction for Electric Vehicles**
    - **Header Name:** "Engine Fuel Type" Analysis for Predicting Electric Vehicle Demand

- **Question:** How will the demand for "Electric" vehicles change in the next 5 to 10 years based on past trends?
- **Relation to Dataset:** The dataset includes vehicles with different "Engine Fuel Type", including "Electric". By analyzing historical trends in electric vehicle MSRPs and their market share, I can predict future demand for electric vehicles. This is particularly relevant as the automotive industry shifts towards greener technologies. Predicting this trend can help manufacturers decide on the level of investment needed in electric vehicle production.

4. **Customer Segmentation for Targeted Marketing**
    - **Header Name:** Customer Segmentation by "MSRP" Range and "Popularity"
    - **Question:** Can I segment customers based on their likelihood to purchase vehicles in different "MSRP" ranges?
    - **Relation to Dataset:** By using the "MSRP" and "Popularity" columns, I can identify patterns in how different market segments respond to vehicle pricing. This segmentation could be based on vehicle popularity or pricing trends, helping to tailor marketing strategies to specific customer groups. This allows for more effective targeting of consumers who tend to be interested in specific types of vehicles, thereby improving marketing efficiency and sales performance (Cote, 2021; Podolean, 2023; Pdmautomotive, 2024).

Conclusion

In this analysis, I have thoroughly examined the "Car Features and MSRP" dataset, focusing on the years 2007 to 2017. The primary goal was to understand how various vehicle features, such as vehicle size, engine horsepower, and engine type, influence the Manufacturer's Suggested Retail Price (MSRP). Through detailed data cleaning, statistical analysis, and visualizations, I uncovered significant insights into the pricing strategies within the automotive industry during this period.

By segmenting the data based on vehicle size, I identified clear trends in pricing, with larger vehicles generally commanding higher MSRPs due to their features and market positioning. The analysis also highlighted the existence of high-priced outliers across different vehicle sizes, reflecting the presence of luxury and performance models in the market.

The visualizations provided further clarity, showing that while most vehicles are priced in the lower to mid range, there is still a notable market for premium vehicles, particularly in the large vehicle segment. The predictive analytics questions discussed offer a roadmap for future research, focusing on predicting vehicle prices, understanding the factors driving high MSRPs, forecasting demand for electric vehicles, and segmenting customers for targeted marketing.

Overall, this analysis offers valuable insights to guide manufacturers, marketers, and industry stakeholders in making informed decisions about vehicle pricing, product development, and market strategies. The findings from this study shed light on past trends and provide a foundation for future predictive models that can help anticipate market shifts and consumer preferences in the changing automobile industry (Encora, 2023; Shih, n.d.).

References

Bobbitt, Z. (2022, August 30). *Pandas: How to Use describe() by Group*. Statology. https://www.statology.org/pandas-groupby-describe/

Cote, C. (2021, October 26). *What Is Predictive Analytics? 5 Examples*. Harvard Business School. https://online.hbs.edu/blog/post/predictive-analytics

Datagy (n.d.). *All the Ways to Filter Pandas Dataframes*. Retrieved May 31, 2020, from https://datagy.io/filter-pandas/

Datagy (2022, September 7). *Pandas dropna(): Drop Missing Records and Columns in DataFrames*. https://datagy.io/pandas-dropna/

Datagy (2023, January 25). *Seaborn histplot – Creating Histograms in Seaborn*. https://datagy.io/seaborn-histplot/

Datacamp (2024, July). *Python Boxplots: A Comprehensive Guide for Beginners*. https://www.datacamp.com/tutorial/python-boxplots

Encora (2023, October 24). *Data Analytics in the Automotive Industry: An Overview*. https://www.encora.com/insights/guide-to-data-analytics-in-the-automotive-industry

InfinitiResearch (2018, December 21). *Competitive Pricing Strategy for Automotive Manufacturers.* https://www.infinitiresearch.com/casestudy/competitive-pricing-automotive-manufacturers/

Learn Statistics Easily (2024, January 6). *Categorical Variable: A Comprehensive Guide for Data Scientists*. Statisticseasily. https://statisticseasily.com/categorical-variable/

McCullum, N. (n.d.). *How To Create Boxplots in Python Using Matplotlib*. Nickmccullum. Retrieved August 18, 2024, from https://www.nickmccullum.com/python-visualization/boxplot/

MachineLearningTutorials (2023, August 23). *Pandas isnull() Function Explained (With Examples)*. https://machinelearningtutorials.org/pandas-isnull-function-explained-with-examples/

Podolean, I. (2023, August 21). *Predictive Analytics in the Automotive Industry: Opportunities and Challenges*. Oneest. https://oneest.com/predictive-analytics-in-the-automotive-industry-opportunities-and-challenges

Pdmautomotive (2024, June 19). *Leveraging Data Analytics to Forecast Aftermarket Trends*. https://pdmautomotive.com/leveraging-data-analytics-to-forecast-aftermarket-trends/

Shih, J. (n.d.). *Car Features and MSRP*. Kaggle. Retrieved August 18, 2024, from https://www.kaggle.com/datasets/CooperUnion/cardataset?resource=download

Shane (n.d.). *Use Pandas Groupby to Group and Summarise DataFrames*. Retrieved August 18, 2024, from https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas/

Solomon, B. (n.d.). *Python Histogram Plotting: NumPy, Matplotlib, pandas & Seaborn*. Realpython. Retrieved August 18, 2024, from https://realpython.com/python-histograms/

Solomon, B. (n.d.). *Pandas GroupBy: Your Guide to Grouping Data in Python*. RealPython. Retrieved August 18, 2024, from https://realpython.com/pandas-groupby/

SparkByExamples (n.d.). *Pandas groupby() and count() with Examples*. Retrieved August 18, 2024, from https://sparkbyexamples.com/pandas/pandas-groupby-count-examples/

Taylor, C. (n.d.). *Role of MSRP in Pricing Decisions*. Chron. Retrieved August 18, 2024, from https://smallbusiness.chron.com/role-msrp-pricing-decisions-39487.html

Tunalı, Y. A. (2021, September 30). *How to Filter Rows and Select Columns in a Python Data Frame With Pandas*. LearnPython. https://learnpython.com/blog/filter-rows-select-in-pandas/