Vintage Automobile Market Analysis Report

Data-Driven Insights and Recommendations for SecondLife.

February 16th, 2025.

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA and Data Preprocessing
- Dimensionality Reduction
- Appendix

Executive Summary

This report provides an in-depth analysis of the vintage cars sold by SecondLife. The findings in this report indicate distinct segmentation within the market based on engine size, weight, fuel efficiency, and performance characteristics. Two primary customer segments emerged during the data analysis: collectors and enthusiasts favoring high-performance vehicles and economy-conscious buyers seeking fuel-efficient classic cars. Data-driven marketing strategies are recommended to optimize targeting and sales conversions based on the findings detailed in this report.

Business Problem Overview and Solution Approach

Business Problem

SecondLife aims to optimize its sales strategy for vintage cars by identifying key vehicle groupings based on performance, fuel efficiency, and historical value. The challenge is to categorize vehicles effectively to segment the company's vintage inventory in a way that enables efficient targeted marketing to different customer groups in order to achieve the business objective to maximize sales and improve customer engagement.

Proposed Solution Approach

To use statistical techniques such as Univariate, and Bivariate Analysis, Principal Component Analysis (PCA), and machine learning algorithms like t-Distributed Stochastic Neighbor Embedding (t-SNE) clustering, to obtain insights that will be useful in guiding targeted marketing strategies and pricing optimizations for the company.

Data Overview

The dataset used for this analysis consists of 398 observations and 8 variables, capturing key attributes of vintage cars. The primary features include miles per gallon (mpg), number of cylinders, engine displacement, horsepower, vehicle weight, acceleration, and model year.

Additionally, the car name column was removed from the analysis due to its large number of unique values and limited analytical relevance, and was found not to be a determining factor affecting sales. The analysis indicated that performance factors such as fuel efficiency and acceleration have a significantly greater impact on sales, making the car name unnecessary for this study.

Exploratory Data Analysis and Data Preprocessing

The exploratory data analysis (EDA) revealed several key insights about vintage cars. For instance, fuel efficiency (mpg) appears to vary significantly across vehicles, ranging from 9 to 46.6 mpg, with a median of 23 mpg as shown in the table, histogram and boxplot below. Let us look at the variables that were considered relevant for this report.

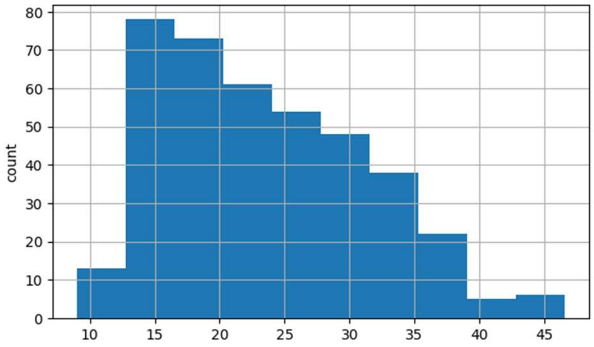
Summary Statistics Table

	MPG	**Cylinders**	**Displacement (in inches)**	**Weight**	**Acceleration**	**Model year**
Count	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000
Mean	23.514573	5.454774	193.425879	2970.424623	15.568090	76.010050
Std	7.815984	1.701004	104.269838	846.841774	2.757689	3.697627
Min	9.000000	3.000000	68.000000	1613.000000	8.000000	70.000000
25%	17.500000	4.000000	104.250000	2223.750000	13.825000	73.000000
50%	23.000000	4.000000	148.500000	2803.500000	15.500000	76.000000
75%	29.000000	8.000000	262.000000	3608.000000	17.175000	79.000000
Max	46.600000	8.000000	455.000000	5140.000000	24.800000	82.000000

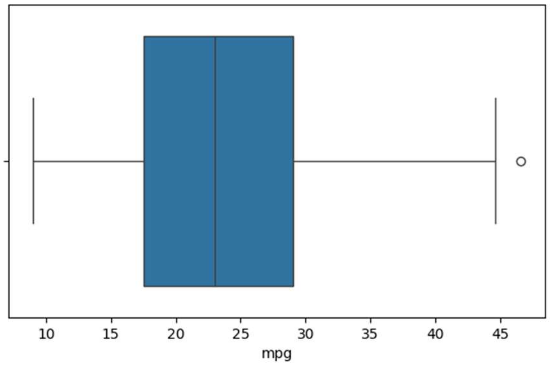
1. MPG/Fuel Efficiency

The MPG distribution is slightly right-skewed, meaning more cars have lower fuel efficiency, but a few have high mpg while the boxplot shows some outliers at the high end (high-mpg cars). The data also suggests there is a diverse mix of economy and performance-focused cars are present within the company's vintage inventory.

MPG/Fuel Efficiency Histogram



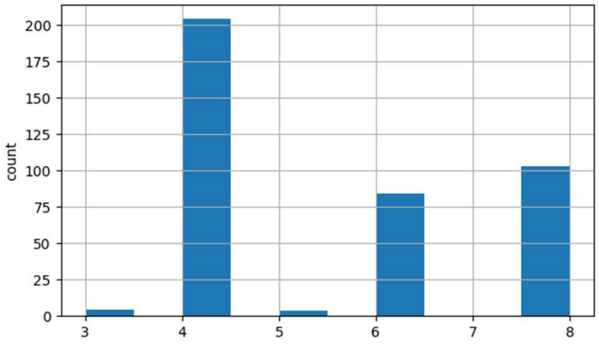
MPG/Fuel Efficiency Boxplot



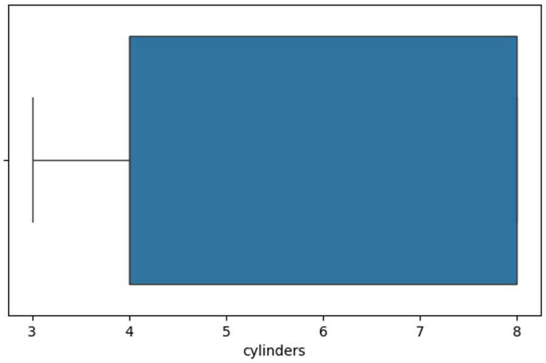
The histogram shows three main peaks at 4, 6, and 8 cylinders, confirming three dominant engine types. The number of cylinders showed that most cars had either 4 cylinders (economy vehicles) or 8 cylinders (muscle cars), making this an important feature for segmentation.

2. Number of Cylinders

Number of Cylinders Histogram



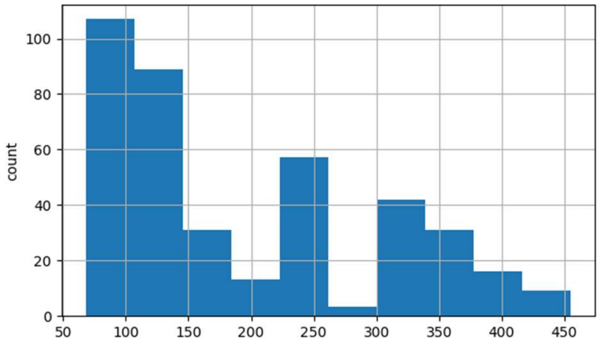
Number of Cylinders Boxplot



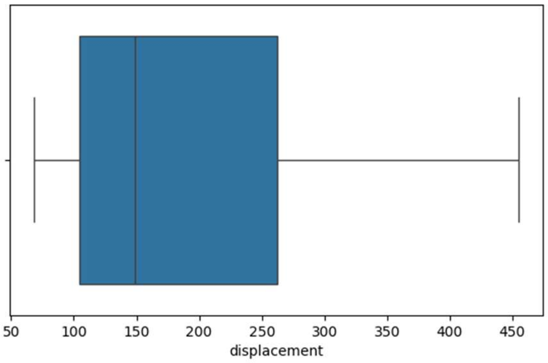
3. Engine Displacement

Engine displacement ranged from 68 to 455 cubic inches, with most cars clustering around 148-262 cubic inches, reflecting a mix of small, fuel-efficient engines and larger high-performance engines. High-displacement cars (300+ cubic inches) can be categorized as muscle cars, and are ideal for targeting collectors. On the other hand, low-displacement cars, typically more compact, appear to be more fuel-efficient models, which makes them suitable for economy-focused buyers.

Engine Displacement/Size Histogram



Engine Displacement/Size Boxplot

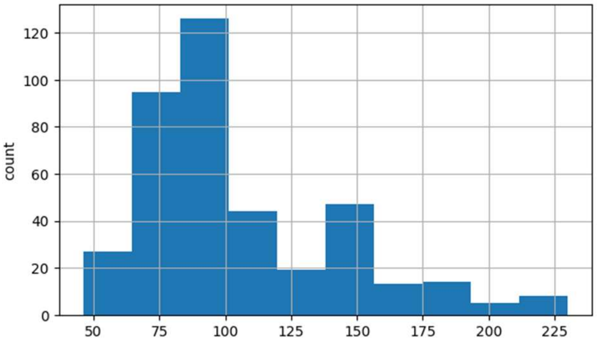


4. Horsepower

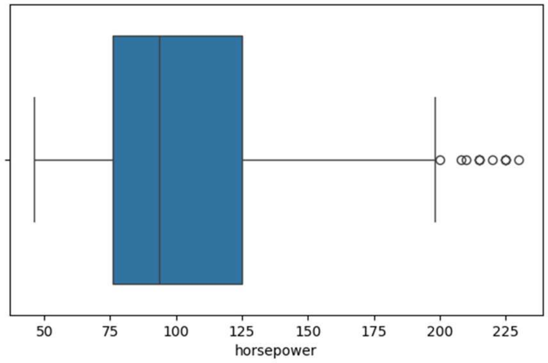
During data preprocessing, six missing values were found in the horsepower column, which were imputed using the mean to maintain dataset integrity. Outliers were identified in displacement, weight, and horsepower; however, they were retained as they provided valuable insights for clustering and segmentation.

The horsepower distribution in the dataset is highly right-skewed, indicating that most vintage cars have lower horsepower, while a few have significantly high values. The histogram reveals this uneven distribution, showing a concentration of cars with lower horsepower and a small number of high-performance vehicles.

Horsepower Histogram



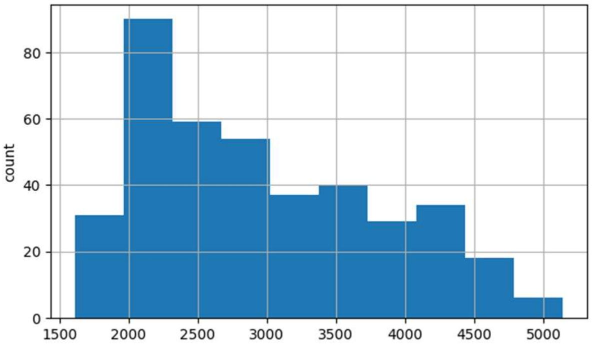
Horsepower Boxplot



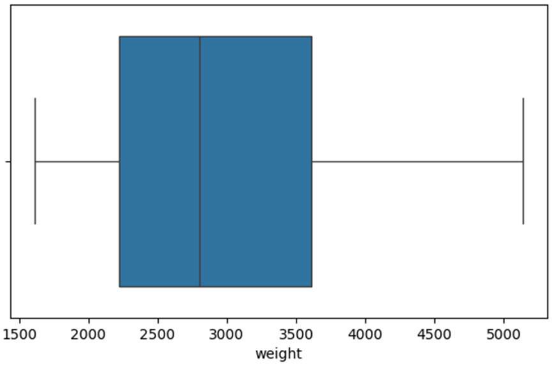
The boxplot further confirms the presence of high-end outliers, which are likely muscle cars equipped with V8 engines. These high-horsepower vehicles stand out from the rest of the dataset and could be classified as premium collectibles. From a business perspective, this insight is valuable, as segmenting cars based on horsepower can enhance targeted marketing strategies for the business. High-horsepower cars can be marketed to performance enthusiasts and collectors, while lower-horsepower vehicles may appeal to everyday vintage car buyers looking for fuel efficiency and practicality.

5. Weight

Vehicle Weight Histogram



Vehicle Weight Boxplot

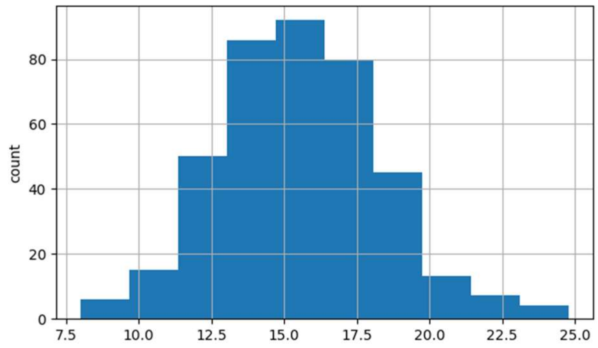


Vehicle weight was found to have a strong correlation with both acceleration and mpg, where heavier cars exhibited lower fuel efficiency and slower acceleration both of which makes these heavy vehicles less appealing to consumers. Thus, heavy cars might need a different marketing approach such as marketing them as muscle cars, or classic luxury vehicles.

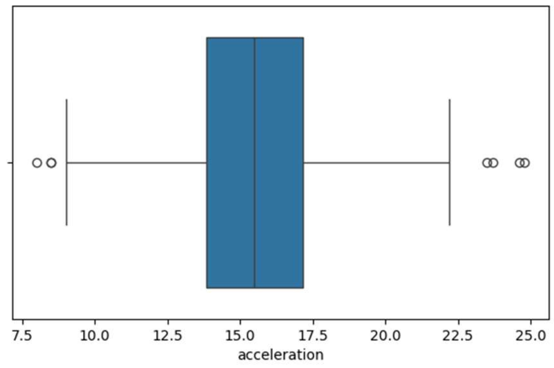
6. Acceleration

The acceleration times suggested that lighter cars with smaller engines tended to be faster, while heavier vehicles took longer to reach 60 mph. The data suggests that most cars have acceleration times between 10-18 seconds. However, the absence of major outliers suggests that acceleration isn't a strong differentiator for the purpose of this report.

Acceleration Histogram

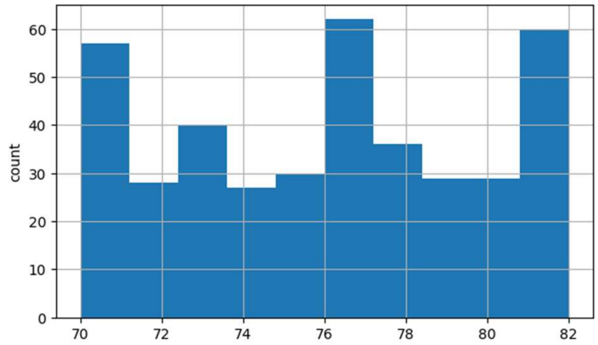


Acceleration Boxplot

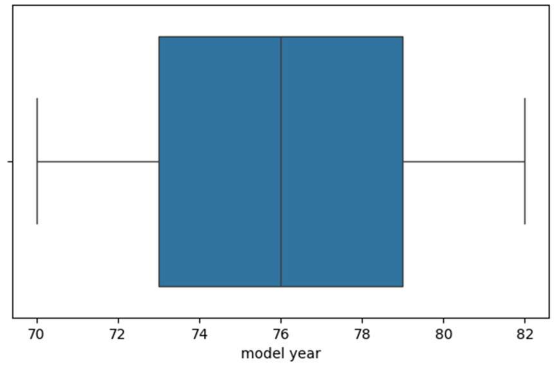


7. Model Year

Vehicle Model Year Histogram



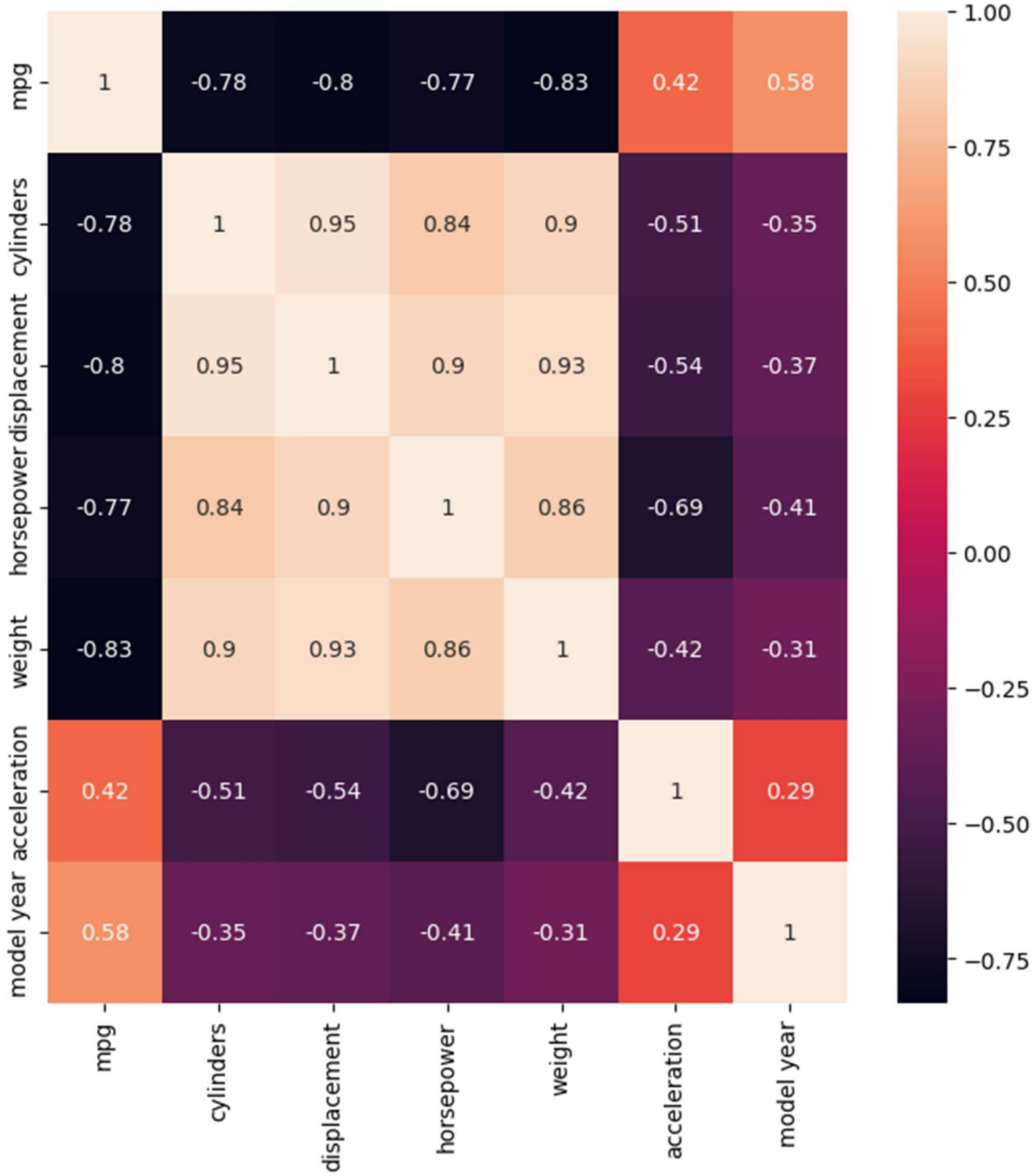
Vehicle Model Year Boxplot



Model years ranged from 1970 to 1982, with cars from before 1975 considered to be collectibles, making them particularly valuable for vintage car enthusiasts.

Next, let us examen in further detail the correlation between variables. Below is the heatmap showing this correlation.

Variable Correlation Heatmap



The correlation heatmap provides valuable insights into the relationships between different variables in the dataset, highlighting how certain car attributes influence one another. The correlation between the variables (vehicle characteristics) is as follows:

Fuel Efficiency (mpg) and Engine/Weight Factors

The variable mpg (miles per gallon) shows a strong negative correlation with cylinders (-0.78), displacement (-0.8), horsepower (-0.77), and weight (-0.83). This means that as any of these variables increase, fuel efficiency decreases. High-performance and muscle cars consume more fuel, while compact, fuel-efficient vehicles achieve better mileage.

Horsepower and Acceleration

Horsepower and acceleration have a negative correlation (-0.69), indicating that cars with higher horsepower tend to have lower acceleration times (meaning they accelerate faster). This relationship is expected because more powerful engines allow vehicles to reach higher speeds in a shorter time. However, the correlation is not perfect, suggesting that other factors such as vehicle weight and transmission efficiency also influence acceleration. Acceleration is inversely related to horsepower. This indicates that high performance cars accelerate faster.

Weight and Engine Characteristics

The variable weight is strongly positively correlated with horsepower (0.86), displacement (0.93), and cylinders (0.95). This means that heavier cars typically have larger engines with more cylinders and higher horsepower. This is consistent with the fact that muscle cars and luxury vintage vehicles tend to have both a larger body and a more powerful engine, while lighter, economy-focused cars tend to have smaller engines with lower horsepower and fewer cylinders.

Model Year and Fuel Efficiency

Model year is positively correlated with mpg (0.58), meaning that newer vintage cars tend to have better fuel efficiency compared to older ones. This suggests that vehicles produced later in the vintage range (closer to the 1980s) may be more desirable for customers looking for efficiency.

The strong correlation between weight, horsepower, and displacement highlights that these features should be key factors when segmenting and marketing vintage cars to different buyer personas. From these discoveries, two clear groups emerge:

1. Performance-oriented cars (higher displacement, more cylinders, more horsepower, and greater weight) tend to have lower fuel efficiency.
2. Economy-oriented cars (lower displacement, fewer cylinders, lower horsepower, and lighter weight) have higher fuel efficiency.

Scaling the Data

When scaling the data using StandardScaler, this standardizes all numerical features so that they have a mean of approximately 0 and a standard deviation of 1. This process was used to convert raw values into z-scores during the data analysis to determine how many standard deviations each value was from the mean. This is crucial for machine learning models, especially PCA and t-SNE, to ensure that all features contribute equally to the analysis rather than being dominated by large-valued attributes. Doing this, removes scale differences between variables, making them more comparable.

Feature Scale Table

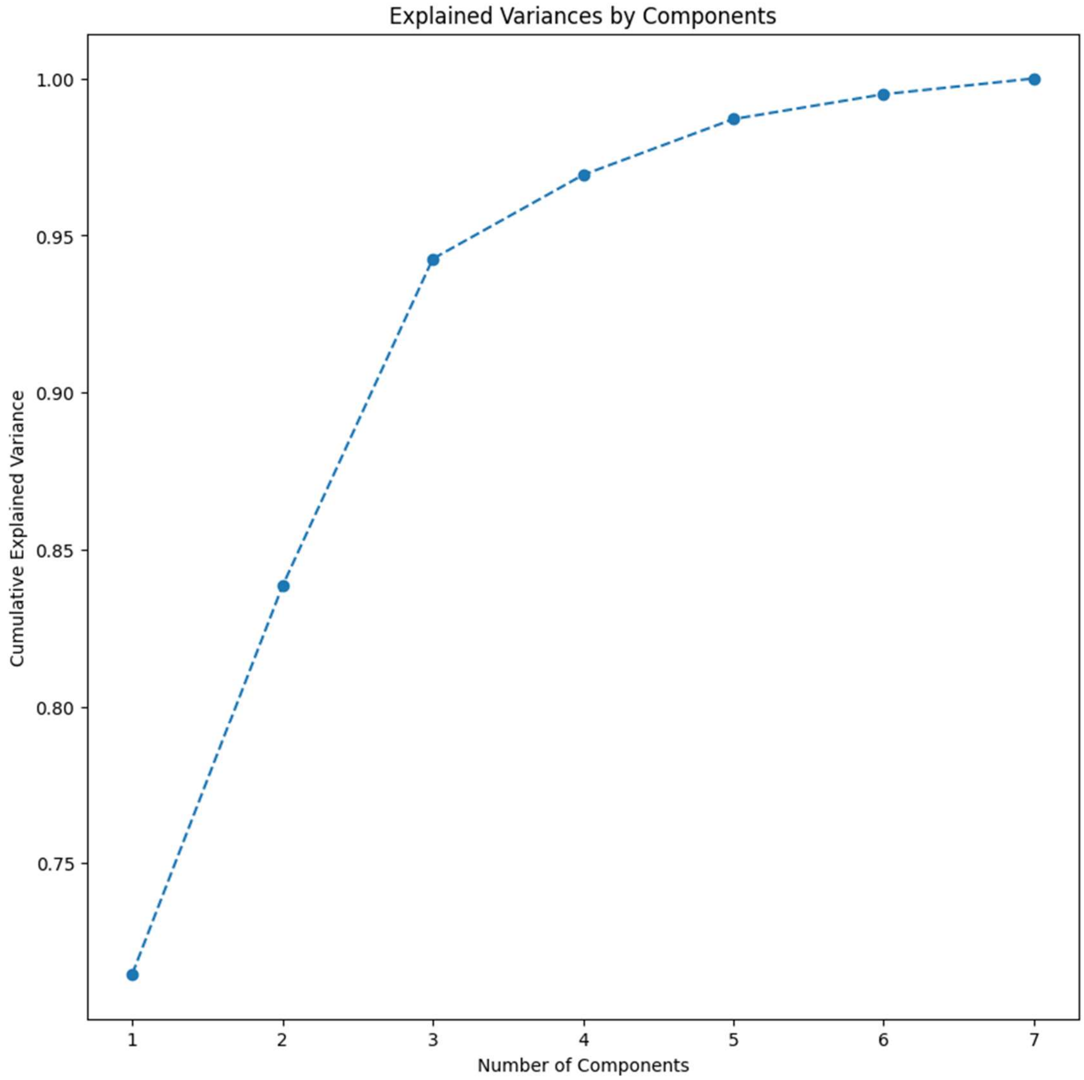
	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year
0	-0.706439	1.498191	1.090604	0.673118	0.630870	-1.295498	-1.627426
1	-1.090751	1.498191	1.503514	1.589958	0.854333	-1.477038	-1.627426
2	-0.706439	1.498191	1.196232	1.197027	0.550470	-1.658577	-1.627426
3	-0.962647	1.498191	1.061796	1.197027	0.546923	-1.295498	-1.627426
4	-0.834543	1.498191	1.042591	0.935072	0.565841	-1.840117	-1.627426

For example, in the table above, Model Year values of ~ -1.627, indicate that most of the dataset consists of later model years, meaning cars from 1970-1973 are less frequent than those from later years. This insight can help with marketing strategies, as earlier models may be more collectible, while later models may appeal to nostalgic buyers.

Dimensionality Reduction

PCA

PCA was used to streamline the analysis and enhance segmentation. PCA is a linear dimensionality reduction technique used to simplify complex datasets while retaining as much variance as possible. It transforms high-dimensional data into a coordinate system where the most significant variations are captured by the principal components. PCA determined that 3 principal components can effectively represent 90% of the variance in the data. Instead of working with 7 features, PCA reduced the dataset to just 3 components while still preserving most of the information. Since only 3 principal components are needed, the data can be visualized in 2D or 3D without much loss of information as observed in the graph below.



In the table below, each column (PC1, PC2, PC3) represents a principal component. Each row (MPG, Cylinders, etc.) shows how much that feature contributes to the respective principal component. Higher absolute values (closer to ±1) indicate a stronger influence.

PCA Loadings (Eigenvectors)

	PC1	PC2	PC3
MPG	-0.400000	0.210000	-0.260000
Cylinders	0.420000	0.190000	0.140000
Displacement	0.430000	0.180000	0.100000
Horsepower	0.420000	0.090000	-0.170000
Weight	0.410000	0.220000	0.280000
Acceleration	-0.280000	-0.020000	0.890000
Model Year	-0.230000	0.910000	-0.020000

The principal component loadings in the table above reveal how each original feature contributes to the first three principal components (PC1, PC2, and PC3). Color coding indicates the following:

- Pink (≤ -0.40): Strong negative correlation with the component.
- Blue (≥ 0.40): Strong positive correlation with the component.

PC1 represents a trade-off between engine power and fuel efficiency, meaning it can separate muscle cars from fuel-efficient economy cars. PC2 mainly captures the model year effect. PC2 can help group cars by model year and identify collectability trends. Older cars may have more vintage appeal, while newer ones might be more practical for daily driving. PC3 represents acceleration performance. PC3 can help distinguish performance cars from slower classic models. Faster cars might appeal to sports car enthusiasts. Slower but more powerful cars could be classic luxury or muscle cars.

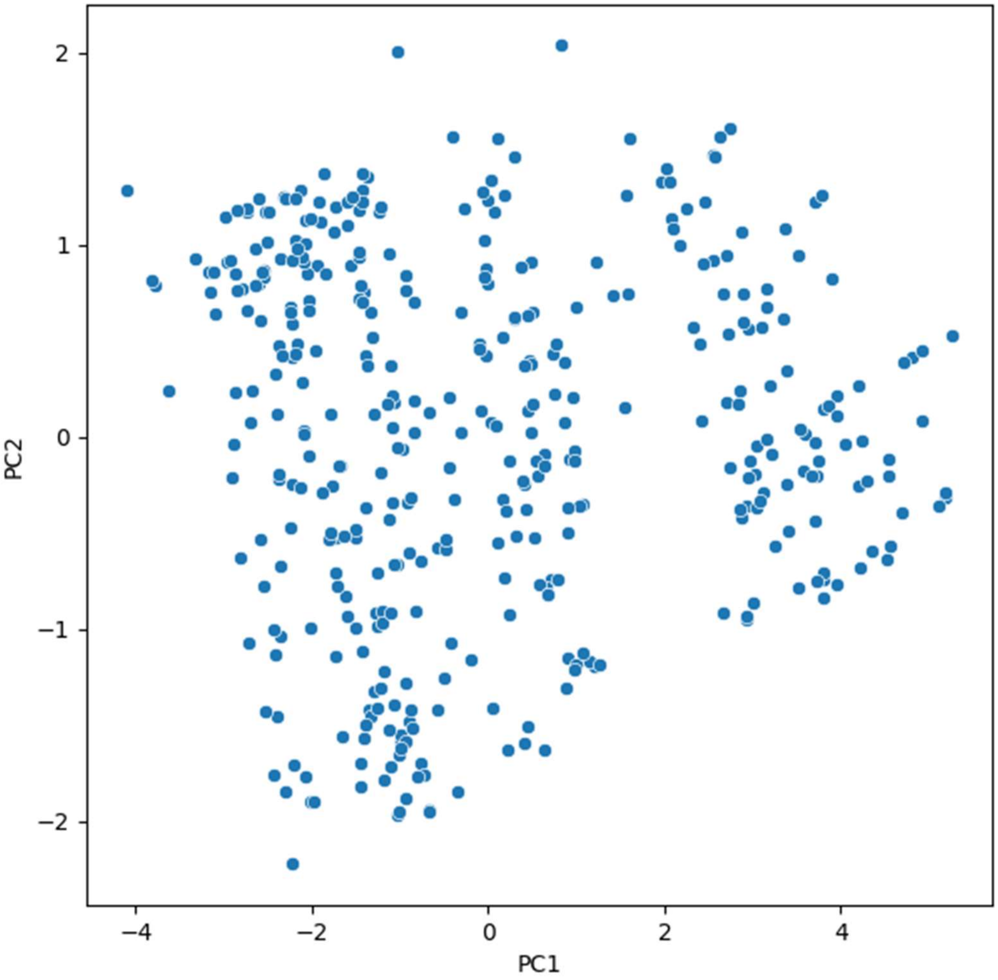
This information is useful for clustering cars into marketing categories:

1. **Muscle Cars:** High PC1, Low MPG, High HP
2. **Fuel-Efficient Cars**: Low PC1, High MPG, Low HP

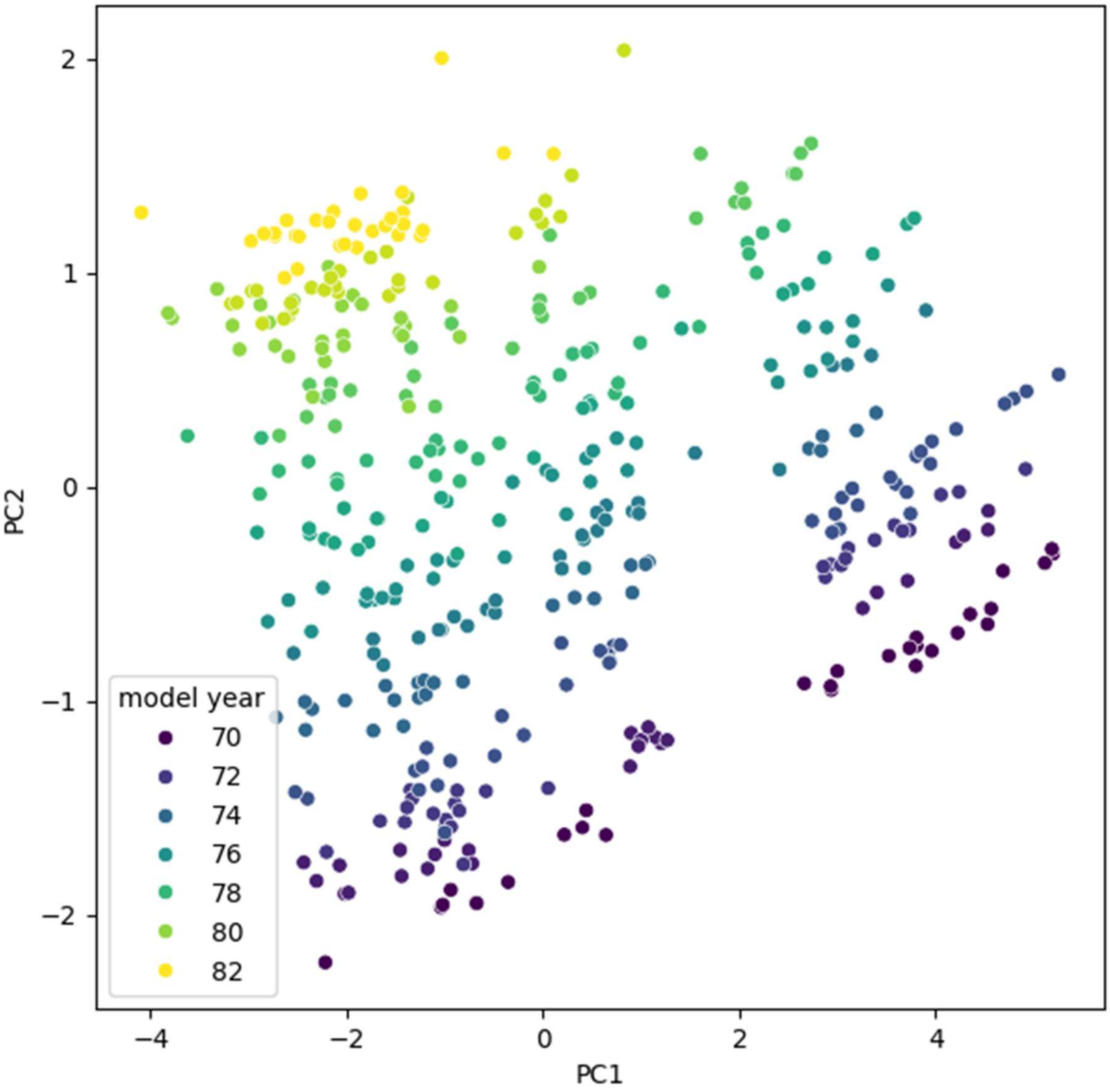
3. **Performance Cars:** High PC3, Fast Acceleration
4. **Luxury Heavy Cars:** High PC1, High Weight, Slow Acceleration

We can also visualize the data in 2 dimensions using the first two principal components.

PC1 and PC2 Scatterplot



PC1 (x-axis) captures the most significant variance in the data, while PC2 (y-axis) captures the second-most significant variance. The spread of points along these axes indicates how well these two components represent variations in the dataset. The y-axis (PC2) represents the second-largest variance, which is less significant than PC1. The scatterplot shows some natural grouping of points, which indicates different car categories. Color-coding the points by a variable will make it easier to see clearer patterns. When adding a hue feature such as the model year to the scatterplot we get the following:



The scatterplot shows the distribution of vehicles in a lower-dimensional space, colored by model year. The color gradient (from dark purple for older cars in 1970 to bright yellow for newer cars in 1982) suggests a clear temporal trend. Older cars (model year 70-74, dark purple to blue) are concentrated on one side of the plot (negative PC1 values). Newer cars (model year 78-82, green to yellow) are on the opposite side (positive PC1 values). This suggests that over time, vehicle attributes (such as weight, acceleration, horsepower, and fuel efficiency) changed systematically, reflecting trends like the shift towards lighter, more fuel-efficient cars.

t-SNE

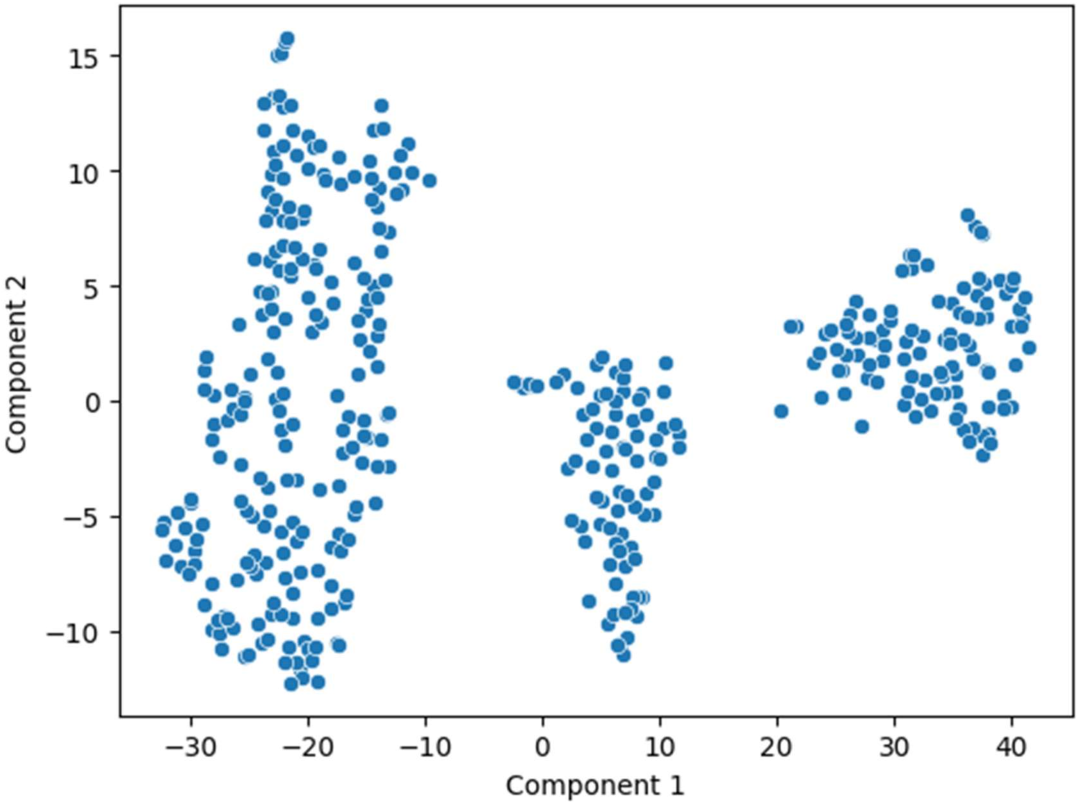
The next step involved applying t-SNE, a non-linear dimensionality reduction technique used to visualize high-dimensional data. t-SNE preserves local structure while reducing dimensions by computing pairwise similarities in both high and low-dimensional spaces.

t-SNE transformed the data into to main components. These components are the two-dimensional representations (2D) of the original high-dimensional data. The t-SNE component values are shown in the table below. The numbers in Component 1 and Component 2 represent the new coordinates of each car after applying t-SNE. Instead of 7 features, each car is now represented by just two values, making it easier to visualize and identify clusters in a scatterplot.

t-SNE Components

	Component 1	Component 2
0	37.486866	-2.327360
1	38.000546	-0.265707
2	38.038231	-1.401425
3	37.438309	-1.503876
4	38.138500	-1.882064

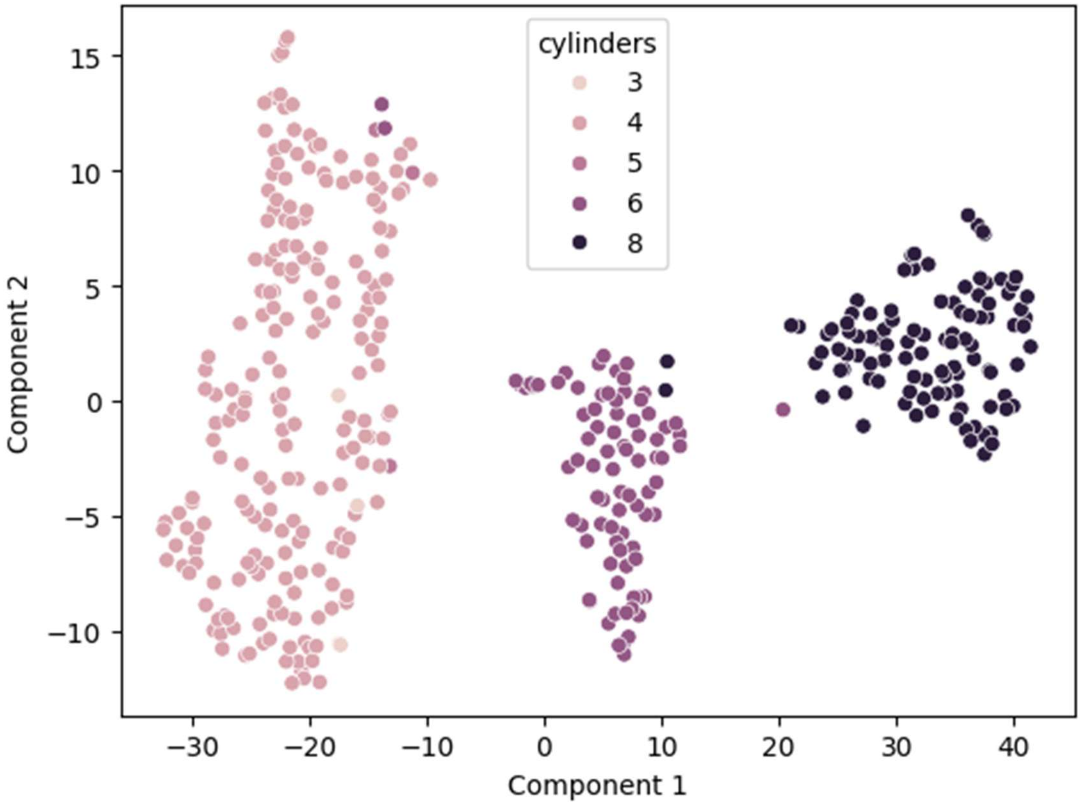
t-SNE Scatter Plot



The t-SNE scatter plot above shows how t-SNE preserves local relationships, meaning that cars that are similar in attributes are placed close together, and cars that are different appear farther apart resulting in further refined segmentation.

When applying t-SNE solely on one of the features, the result is a cluster separation based on that particular feature as seen in the scatterplot below.

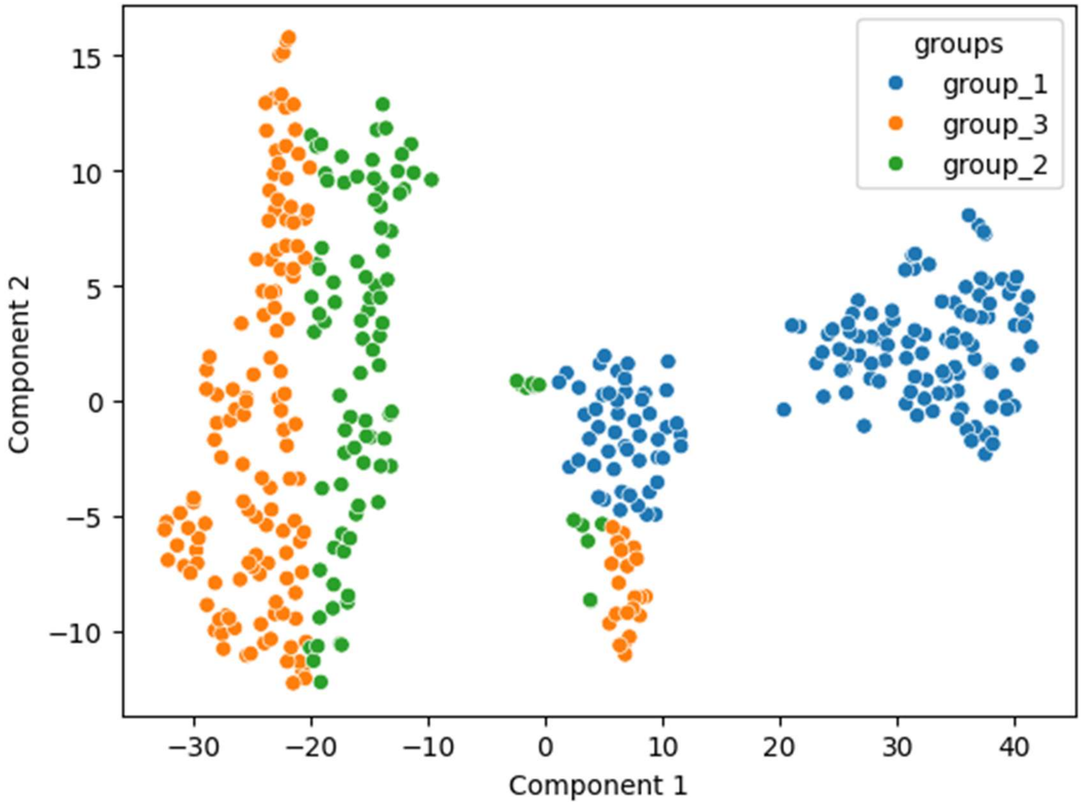
t-SNE Cylinders Scatterplot



The scatterplot shows 3 clear groups based on cylinder count. Four-cylinder cars form a tight cluster, meaning they are similar in design such as lightweight bodies and high fuel efficiency. In contrast, six and eight-cylinder cars create separate groups due to differences in weight, horsepower, and fuel consumption. A smooth color transition within clusters suggests that six-cylinder cars share traits with both four-cylinder and eight-cylinder cars. Lastly, the eight-cylinder group is more spread out, showing a wider range in performance, sizes, and weight. Since the clusters are clearly separated, it means cylinder count strongly affects a car’s characteristics.

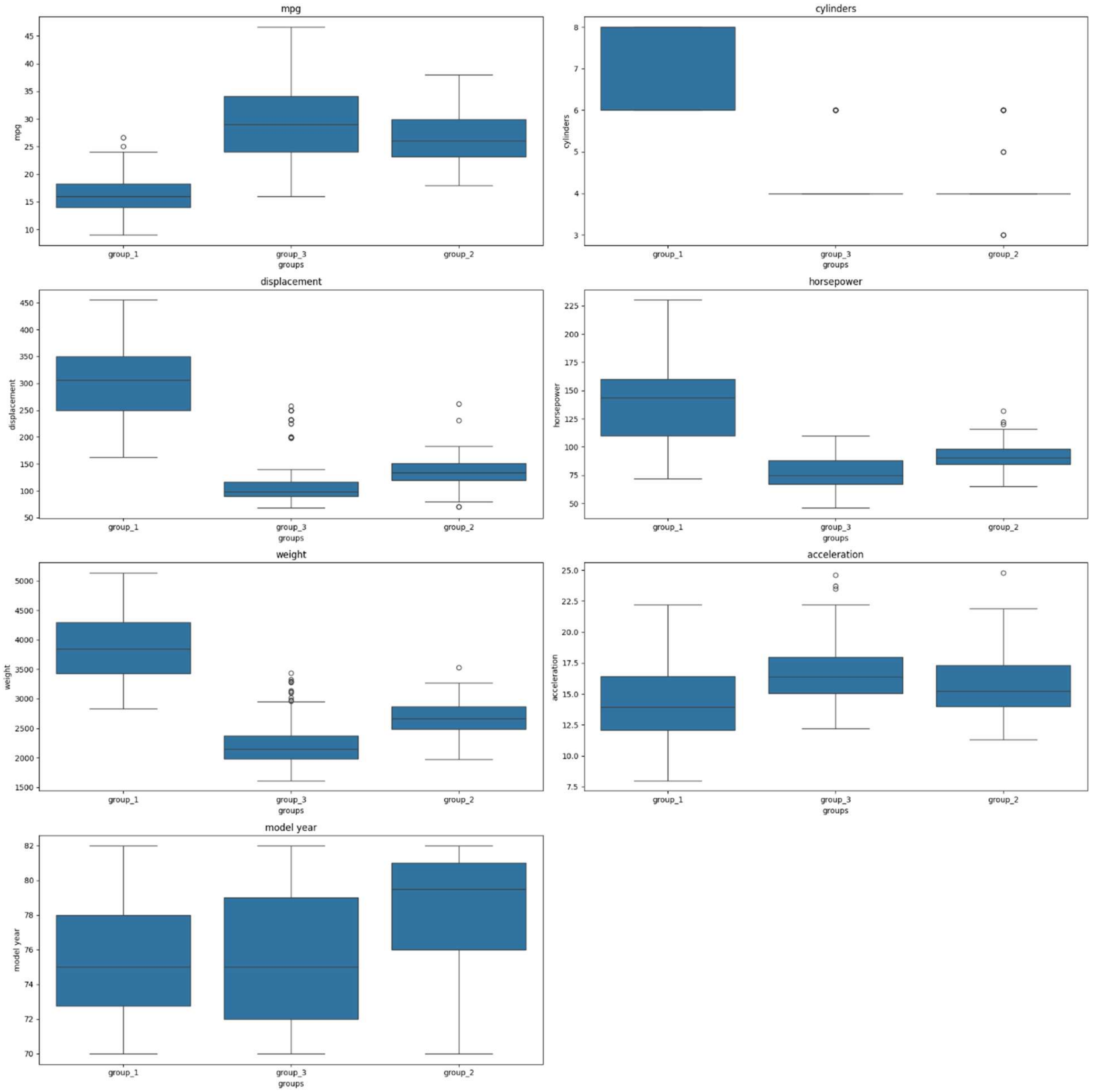
When assigning points to the 3 different groups, results in the following scatterplot:

t-SNE Groups Scatterplot



This clustering provides additional insights into the dataset's structure and vehicle classifications. At a glance, 3 distinct groups can be seen without any major overlap where Group 1 (blue) represents high-performance, heavy cars, group 2 (green) represents lightweight, fuel-efficient cars and group 3 (orange) represents mid-range vehicles. The insights from this scatterplot can be more easily identified by plotting each group according to its features into a boxplot as seen in the boxplot below.

t-SNE Group Feature Boxplots



From these boxplots, the key findings are as follows:

1. For MPG, group 1 has the lowest MPG, indicating these vehicles are fuel-inefficient. group 2 has moderate MPG, and group 3 has the highest MPG, representing the most fuel-efficient vehicles.

2. For cylinders, group 1 consists mainly of 8-cylinder cars, indicating larger and more powerful engines. Groups 2 and 3 contain mostly 4-cylinder vehicles, meaning smaller and more efficient engines.

3. For Displacement, group 1 consists mainly of 8-cylinder cars, indicating larger and more powerful engines. groups 2 and 3 contain mostly 4-cylinder vehicles, meaning smaller and more efficient engines.

4. For Horsepower, group 1 has significantly higher displacement, aligning with larger engine sizes. Groups 2 and 3 have much smaller displacement, consistent with compact cars.

5. For Weight, group 1 includes the heaviest vehicles, which aligns with their high displacement and horsepower. Groups 2 and 3 are much lighter, reinforcing their classification as smaller and more efficient cars.

6. Acceleration is fairly similar across all groups, but Groups 2 and 3 show slightly better acceleration, possibly due to their lighter weight.

7. For Year model, group 1 consists of a mix of older and newer vehicles. Groups 2 and 3 mostly contain newer cars, indicating a shift towards lighter and more fuel-efficient designs in later years.

Conclusion

By refining segmentation through PCA and t-SNE clustering, three distinct groups emerged:

1. **High-performance muscle cars** – V8 engines, high horsepower, and heavy weight.
2. **Fuel-efficient lightweight classics** – 4-cylinder engines, high MPG, and compact design.
3. **Balanced mid-range cars** – Moderate fuel economy and weight.

By utilizing the insights obtained from the t-SNE clustering, the company can tailor its marketing and product strategies to better align with consumer preferences and behaviors. This level of segmentation enables targeted messaging. For example, instead of a one-size-fits-all campaign, the company can tailor ads, promotions, and product positioning to each group's priorities. Optimizing pricing and financing through customized financing plans based on the target audience's buying power can increase conversion rates. Understanding demand for specific types of vehicles helps the company allocate production and dealership stock more efficiently by effectively managing inventory. By aligning features with customer needs, the company can increase customer satisfaction, brand loyalty and repeat purchases.

Appendix

Key Visualizations and Insights

PCA Scatterplot: Model Year vs Principal Components

- The PCA scatterplot highlights how vintage car features changed over time.
- Older cars (1970-1974) cluster together, while newer models (1978-1982) shift toward higher fuel efficiency and lower displacement.
- PC1 captures differences in engine power and efficiency, while PC2 reflects model year variations.

t-SNE Scatterplot: Vehicle Clusters

- Three clear groups emerged from t-SNE clustering:
    1. High-performance muscle cars – V8 engines, high horsepower, and heavier weight.
    2. Fuel-efficient lightweight classics – 4-cylinder engines, high MPG, and compact design.
    3. Balanced mid-range cars – Moderate fuel economy and weight.
- The grouping allows the company to refine marketing strategies based on consumer preferences.

Boxplots of Key Features by Cluster

- MPG: Group 1 has the lowest MPG, indicating fuel-inefficient muscle cars, while Group 3 has the highest MPG, representing economy vehicles.

- Cylinders: Group 1 consists mainly of 8-cylinder cars, whereas Groups 2 and 3 contain mostly 4-cylinder vehicles.
- Horsepower and Weight: Group 1 cars are significantly heavier and more powerful than those in Groups 2 and 3.

Technical Details

PCA Implementation

- PCA reduced the dataset from 7 features to 3 components, preserving 90% of the variance.
- Visualization in 2D and 3D revealed natural car groupings, helping optimize segmentation.

PCA Code Snippet:

```
# Defining the number of principal components to generate
n = data_scaled.shape[1]
# Finding principal components for the data
# Apply the PCA algorithm with random_state = 1
pca = PCA(n_components=n, random_state=1) # Code completed here.
# Fit and transform the pca function on scaled data
data_pca1 = pd.DataFrame(pca.fit_transform(data_scaled)) # Code completed here.
# The percentage of variance explained by each principal component
exp_var = pca.explained_variance_ratio_
```

t-SNE Implementation

- t-SNE revealed distinct clusters of similar vehicles, reinforcing PCA findings.
- The clustering enables precise customer targeting based on vehicle characteristics.

t-SNE Code Snippet:

```
# Apply the t-SNE algorithm with random_state = 1
tsne = TSNE(random_state=1) # Complete the code
# Fit and transform t-SNE function on the scaled data
data_tsne = tsne.fit_transform(data_scaled) # Complete
```

Additional Supporting Information

- The dataset contained 398 vintage cars with key attributes such as MPG, weight, horsepower, cylinders, and model year.
- StandardScaler was used to normalize numerical values before applying PCA and t-SNE.
- Outliers in horsepower, weight, and displacement were retained for better segmentation insights.
- Correlation analysis showed that weight, horsepower, and displacement strongly influence a vehicle's performance and efficiency.