Second Life Automotive

Data Analysis and Visualization Business Report

Feature Extraction & Quantitative Analysis

Contents / Agenda

Executive Summary

With the automobile market experiencing significant changes due to market conditions, globalization, cost pressure, industry volatility, and technology, we are shifting your focus to a less volatile sector of the market, vintage cars.

In order to successfully shift predominately into the vintage car market, we needed to look at our past sales of vintage cars to gain insight into what features were attributes of the cars we sold.

The data analyzed in this report provides key insights from 398 entries surrounding the top 7 features of the vintage cars sold across SecondLife's outlets in the US.

To gain insight on a feature level, I performed analysis on the features alone and in relationship to one another.

- Performed Univariate Analysis to determine the standalone observations of the features.
- Performed Bivariate Analysis for an understanding of how each key feature relates to one another as a whole.

To gain insight on the correlated relationship of the features, I reduced dimension to prevent features with larger values from dominating the analysis.

- Performed PCA Technique to identify how many components held the majority of the data and to what degree they held the greatest variance.
- Performed t-SNE Technique to identify the similarity between the components.

Used Box Plots to understand the order of significance for the 7 key features within the newly identified condensed groupings.

It was found that 3 Attributes comprise 90% of the data

1. Engine
   - Cylinders
   - Displacement
   - Horsepower
2. Model Year
3. Acceleration

It is my recommendation for SecondLife to procure inventory of 4 types of cars in order to market to the consumer groups who purchase in these groups

- Vintage muscle car of any model year, with the focus being on the engine power
- Classic vintage model muscle cars
- Vintage cruisers
- Newer model vintage cars which offer good gas mileage in relation to the model year

After procurement, in order to bring consumer awareness to the new line of Vintage offerings at SecondLife, I suggest launching a marketing campaign to past purchasers.

- A good strategy would be to host a Car Show.
- One key aspect of the marketing strategy would be to invite past customers to bring their Vintage cars to SecondLIfe's Car Show.

My recommendations will provide SecondLife with a well-rounded selection of carefully procured inventory, surely suitable to fit the desires of Vintage car lovers. Utilizing your hot market of past purchasers, by hosting a car show, will increase awareness of SecondLife's new position in the Vintage car market.

Business Problem Overview and Solution Approach

With the automobile market experiencing significant changes due to market conditions, globalization, cost pressure, industry volatility, and technology, we are shifting your focus to a less volatile sector of the market, vintage cars.

We need to take an in-depth look at all of our past sales of vintage cars, from all of our outlets across the U.S. This will provide us with a better understanding of what key commonalities were present in the varying groups of vintage cars. We will also gain insight into the consumer groups who previously purchased, so we can strategically target our audience more effectively.

Using Univariate Analysis, I have examined the standalone features to clearly distinguish which ones stand out above the others. I performed Bivariate Analysis to examine how each key feature relates to the others as a whole. This also gave insight into the varying combination of feature differences which shows a relationship to the consumer groups that purchased them.

Based on the data feedback from the Principal Component Analysis, which grouped the combination of features, I checked which features were dominate in these 3 categories. I then used a box plot comparison method to visually see the order of importance of features within each grouping. This provides a second key insight into the preferences of the consumer groups who purchased these cars. We want to gain an understanding of our target consumer group in order to effectively market to them.

Data Overview

There were 398 entries across 8 columns of data.

The column 'car name' contained 305 unique entries out of 398. This is not numerical data and I do not feel this would add value to our analysis, so this column was dropped.

We were left with 7 categories to analyze

The 7 column categories are

1. mpg float64
2. cylinders int64
3. displacement float64
4. horsepower object converted
5. weight int64
6. acceleration float64
7. model year int64
8. car name object

There weren't any missing data values in any of the categories.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year
count	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000
mean	23.514573	5.454774	193.425879	104.304020	2970.424623	15.568090	76.010050
std	7.815984	1.701004	104.269838	38.222625	846.841774	2.757689	3.697627
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000
25%	17.500000	4.000000	104.250000	76.000000	2223.750000	13.825000	73.000000
50%	23.000000	4.000000	148.500000	93.500000	2803.500000	15.500000	76.000000
75%	29.000000	8.000000	262.000000	125.000000	3608.000000	17.175000	79.000000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000	82.000000

Observations

If mpg standard deviation is low in comparison to the mean = high mpg (good gas mileage).

A higher number of cylinders in an engine is associated with improved performance and smoother operation

- avg = 5.45 with 50% falling below a 4 cylinder

Higher mean displacement generally correlates with higher power output and torque.

- Max at 455 and 50% below 148 with a min of 68
- Lower power output preferred.

A lower standard deviation suggests more consistent engine performance.

It can affect the predictability of fuel efficiency and power delivery.

- Standard deviation is about 1/2 of the average.

Average horsepower is 104.3 with standard deviation at 38 and 50% below 93.5.

- Horsepower preference is wide spread.

Weight has an average of 2970 with 50% below 2803.

- Lower weight number is preferred.

Acceleration average is 15.56 with a max of 24.

- High acceleration number is preferred.

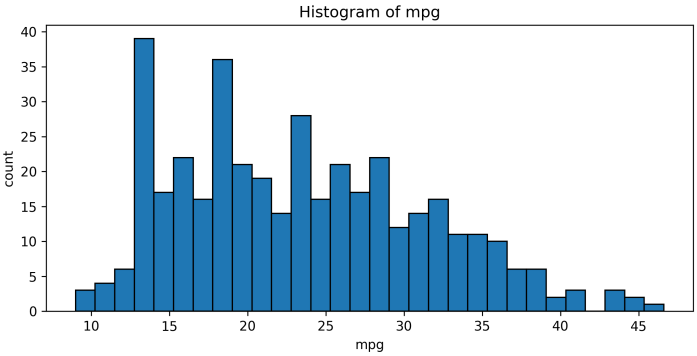
Model year average is 76 with max at 82 and min at 70.

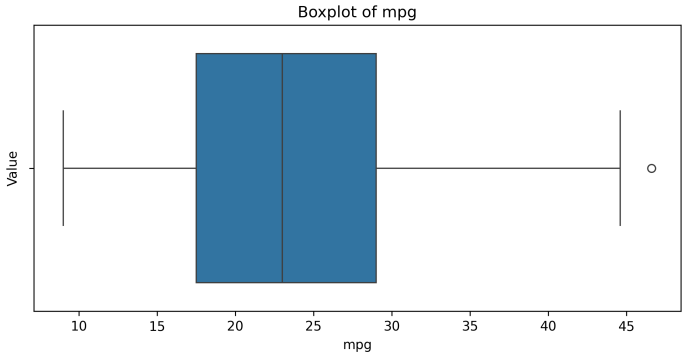
- Older model year is preferred by the majority.

EDA and Data Preprocessing

Univariate Analysis

- 398 data entries across 7 categories
- Used Histograms and Box Plots to gather 2 perspectives on the same data for each category

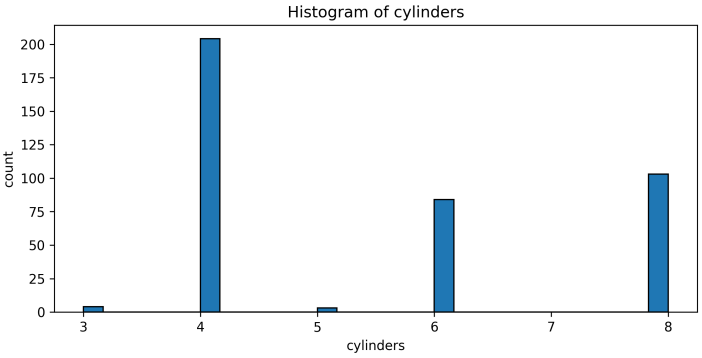


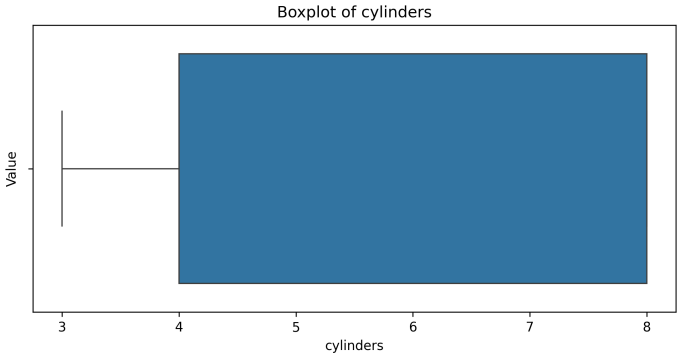


MPG

histogram: skewed to left (lower mgp) largest amount of data in the 13 and 18 mpg (13 falls outside boxplot 25% but holds lgst amt of data)

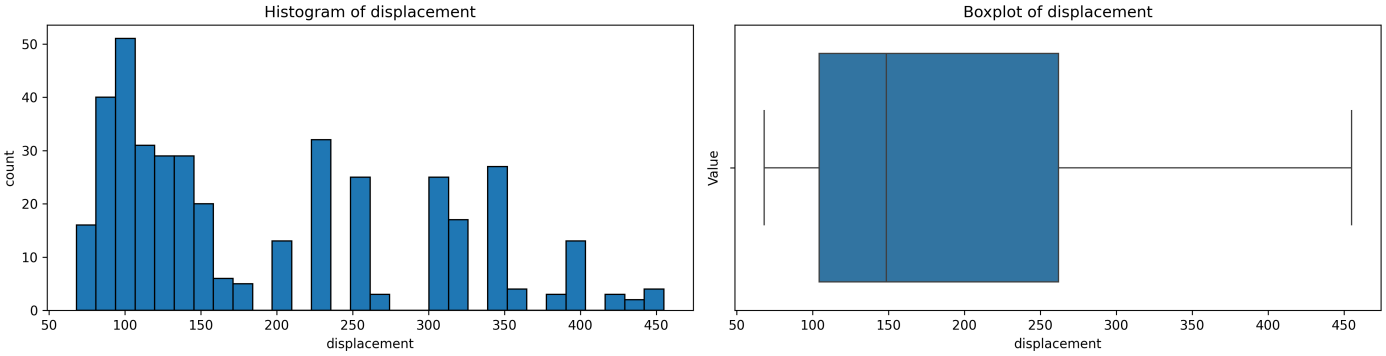
boxplot 25-75%: mpg = 17 - 29 / 23 avg





CYLINDERS

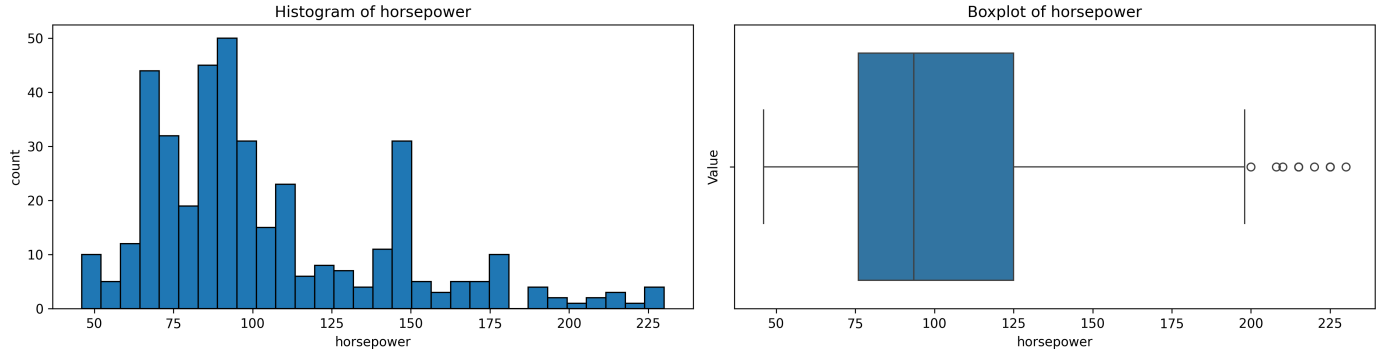
histogram: 4 most of data



boxplot: 4-8 range heavy

DISPLACEMENT

histogram: 100 most of data / 125-175, 225-275, 320-340 had good amount of data - consider as secondary categories of market

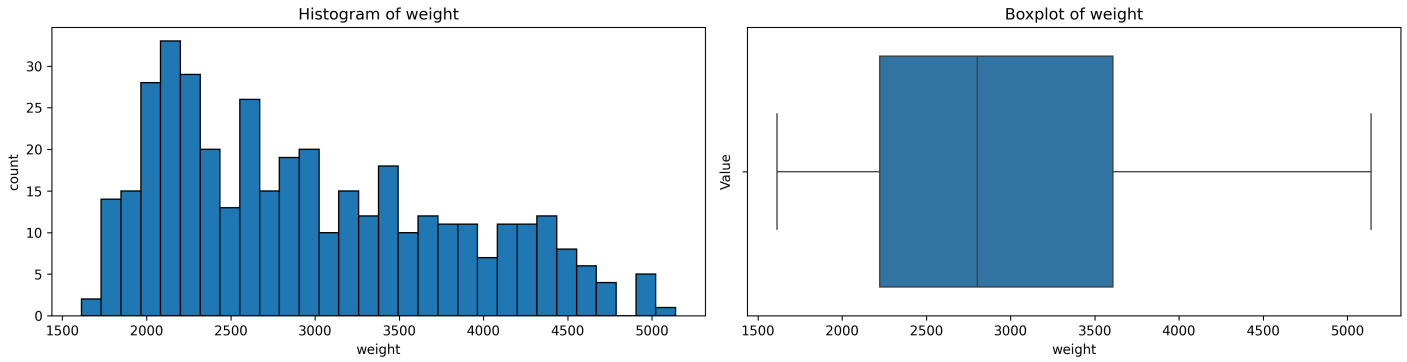


boxplot: 100-225 median lower / lower end skewed

HORSEPOWER

histogram:65 & 85 held alot of data w/ 90 having most - 145 is an outlier with fair amount of data / other outliers are insignificant numbers

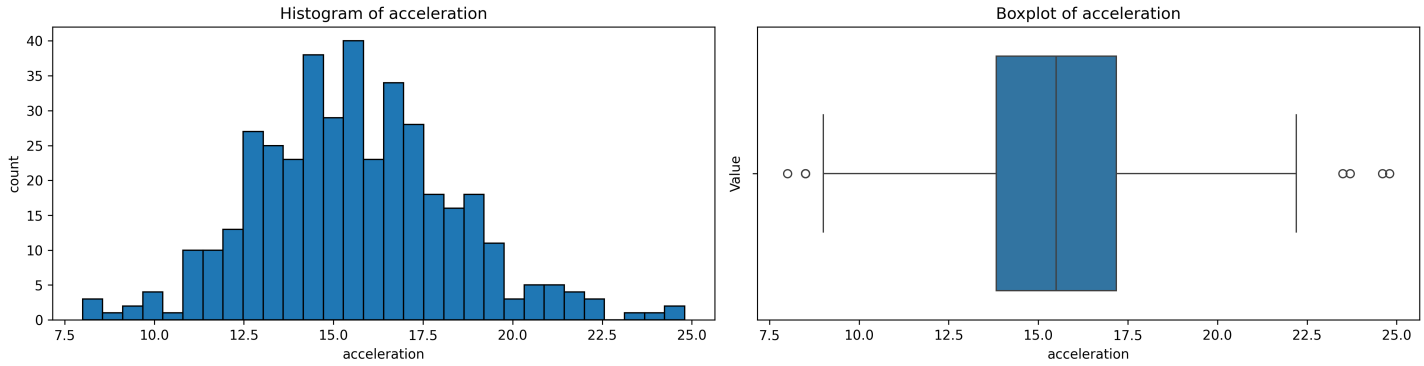
boxplot:75-125 lower end skewed / median lower



WEIGHT

histogram:2200 held most of data

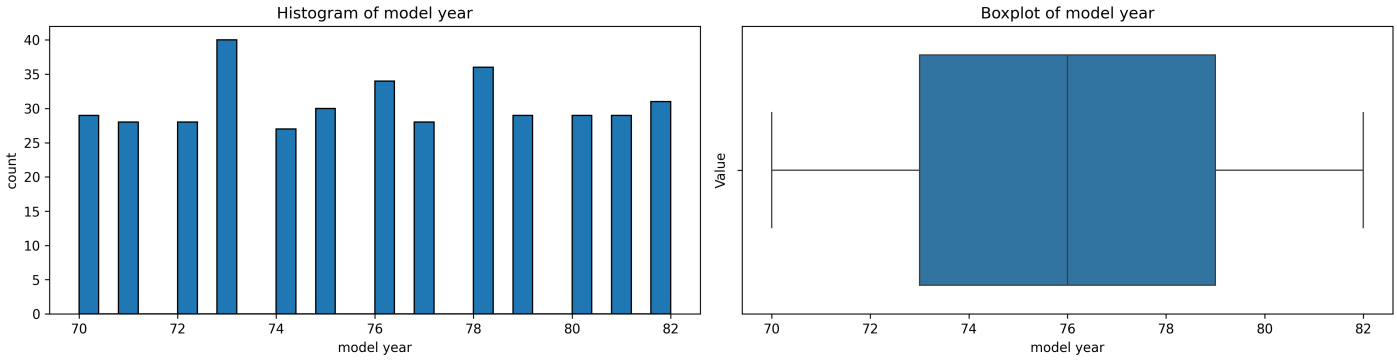
boxplot: 2200 - 3600 lower end skew / avg on lower



ACCELERATION

histogram: bulk of data falls in this center range

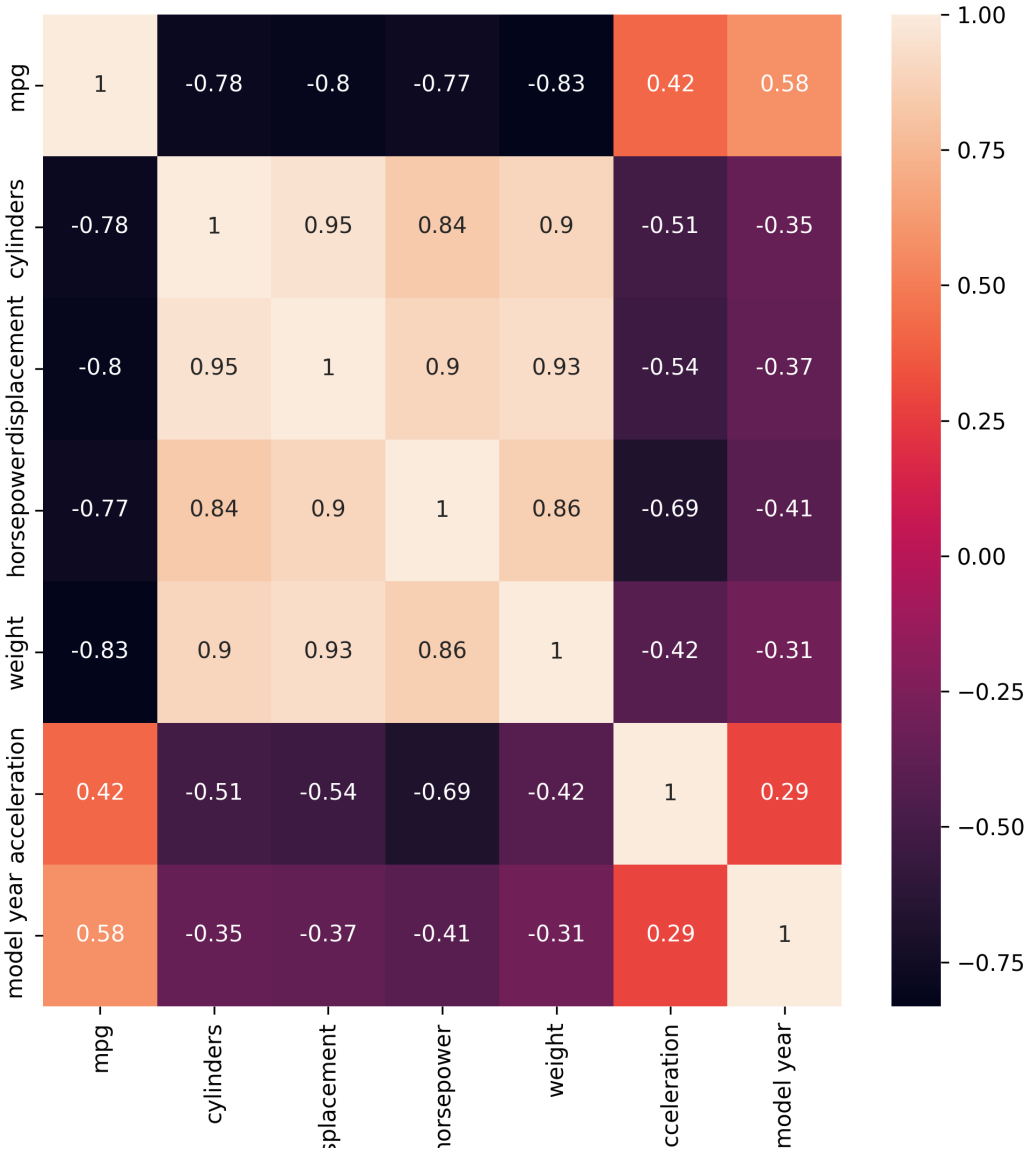
boxplot: balanced 14-17 range / avg in middle



MODEL YEAR

histogram: evenly distributed 73 most / 78 & 76 equally next

boxplot: completely symmetrical with 76 as average


mpg
cylinders
displacement
horsepower
weight
acceleration
model year
1
-0.78
-0.8
-0.77
-0.83
0.42
0.58
-0.78
1
0.95
0.84
0.9
-0.51
-0.35
-0.8
0.95
1
0.9
0.93
-0.54
-0.37
-0.77
0.84
0.9
1
0.86
-0.69
-0.41
-0.83
0.9
0.93
0.86
1
-0.42
-0.31
0.42
-0.51
-0.54
-0.69
-0.42
1
0.29
0.58
-0.35
-0.37
-0.41
-0.31
0.29
1
mpg
cylinders
displacement
horsepower
weight
acceleration
model year
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Bivariate Analysis

The variable mpg has a strong negative correlation with cylinders, displacement, horsepower, and weight.

- More power = more weight = less mpg

Horsepower and acceleration are negatively correlated.

- Higher the horsepower = less time to peak acceleration

The variable weight has a strong positive correlation with horsepower, displacement, and cylinders.

- Increase in horsepower, displacement, and cylinders = increase in weight

Model year is positively correlated with mpg.

- Newer model = more mpg

Dimensionality Reduction

Performed a standard scaling method for each column and value in each column.

- Each column is transformed to have a mean of 0 and a standard deviation of 1.

This prevents features with larger values from dominating the analysis.

index	mpg	cylinders	displacement	horsepower	weight	acceleration	model year
0	-0.71	1.5	1.09	0.67	0.63	-1.3	-1.63
1	-1.09	1.5	1.5	1.59	0.85	-1.48	-1.63
2	-0.71	1.5	1.2	1.2	0.55	-1.66	-1.63
3	-0.96	1.5	1.06	1.2	0.55	-1.3	-1.63
4	-0.83	1.5	1.04	0.94	0.57	-1.84	-1.63

I then applied the PCA technique to identify the groups of greatest variance.

The number of PCs that explain at least 90% variance: 3

	PC1 Loadings	PC2 Loadings	PC3 Loadings
acceleration	-0.284741	-0.024603	**0.892732**
weight	0.414089	0.221482	0.279512
cylinders	**0.417246**	0.192019	0.141058
displacement	**0.429315**	0.177656	0.103334
model year	-0.229556	**0.910532**	-0.017572
horsepower	**0.422181**	0.090036	-0.167806
mpg	-0.397637	0.211502	-0.256632

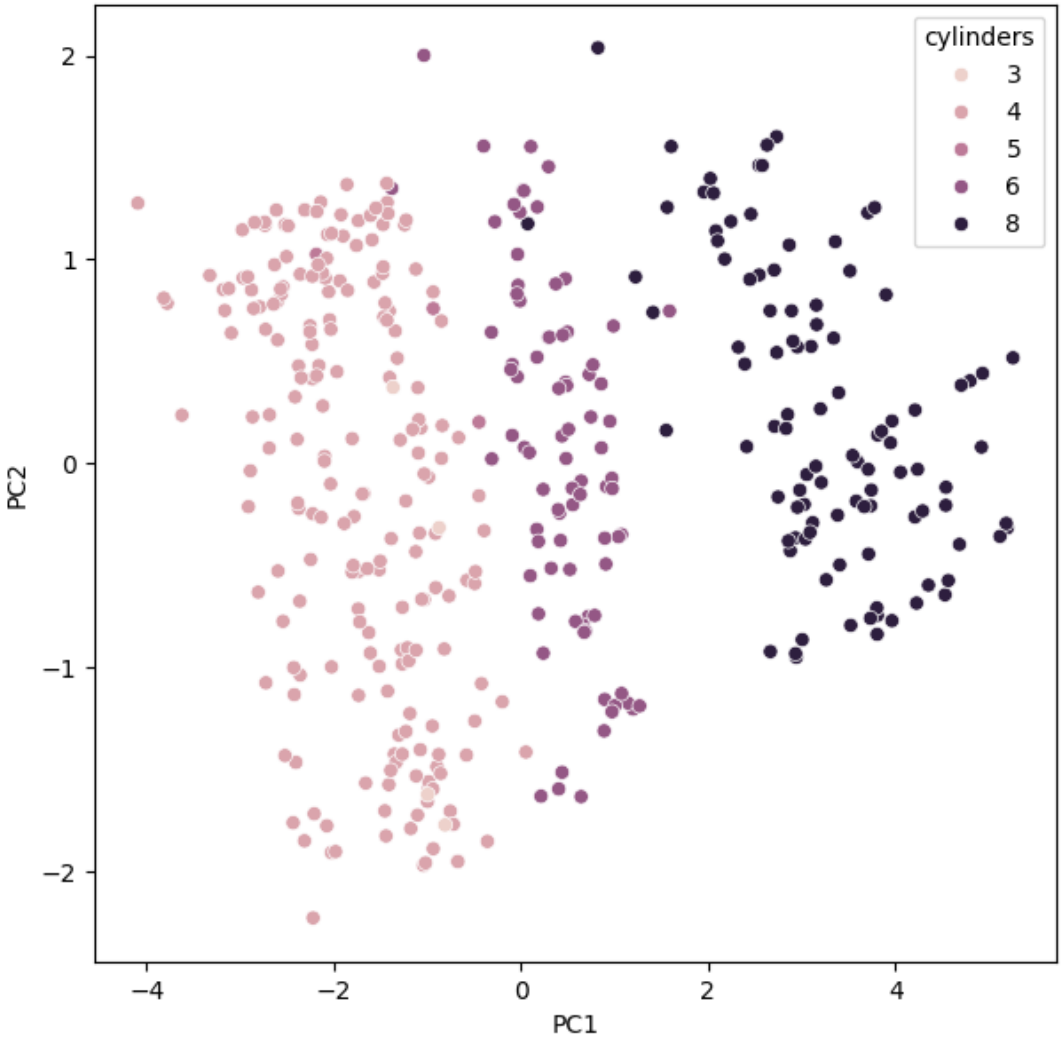
Looking at the data above, we can conclude:
PC1 represents the **Engine**
PC2 represents the **Model Year**
PC3 represents **Acceleration**

In the PC chart below, the positive numbers have a positive impact and the negative numbers have a negative impact.
The higher the number, the more positive the impact.

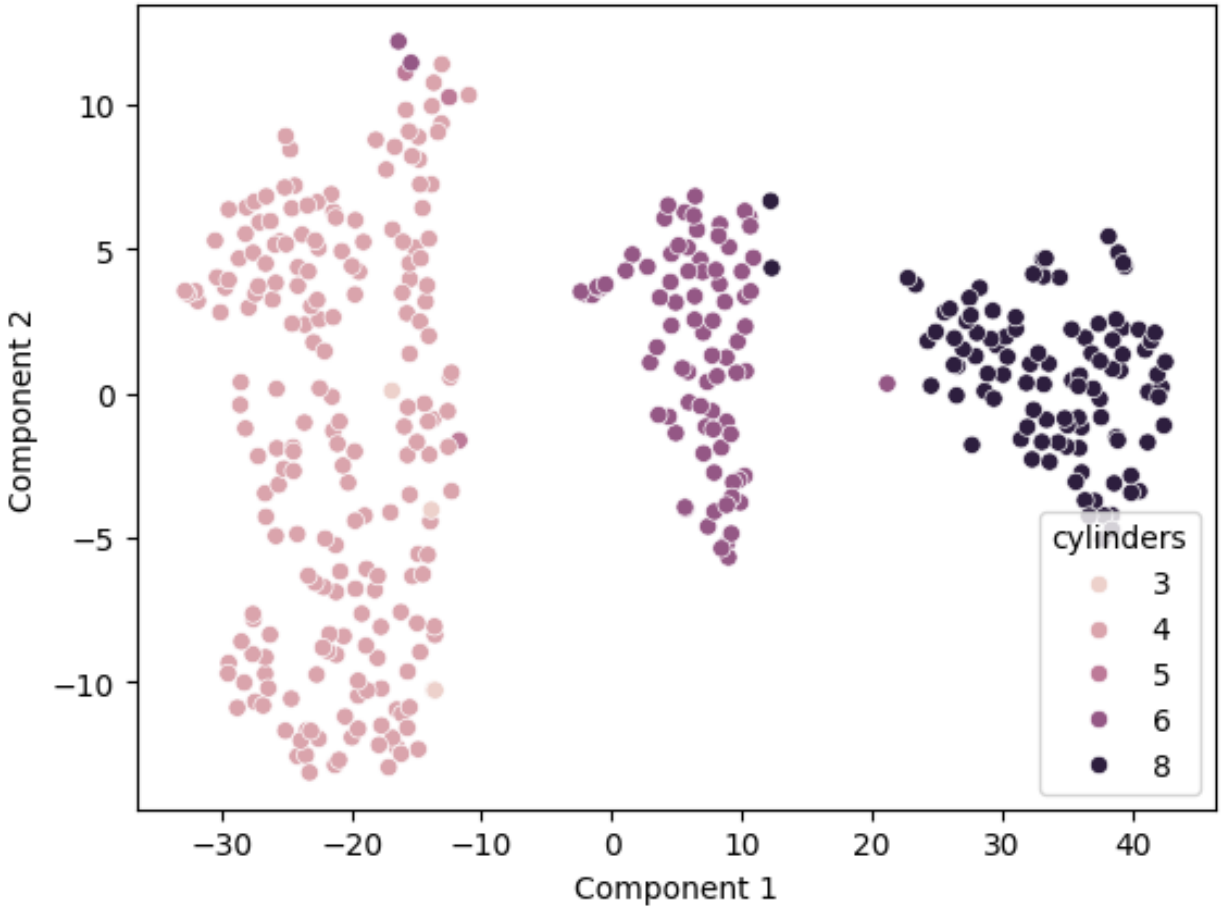
index	PC1	PC2	PC3
mpg	-0.4	0.21	-0.26
cylinders	0.42	0.19	0.14
displacement	0.43	0.18	0.1
horsepower	0.42	0.09	-0.17
weight	0.41	0.22	0.28
acceleration	-0.28	-0.02	0.89
model year	-0.23	0.91	-0.02

high mpg = low PC1
high cylinder, displacement, horsepower, weight = high PC1
high model year = high pc2
high acceleration = high pc3

For PC1, cylinders was one of our highly correlated features. Here we can see the 7 features in the data set track in a positive relationship to the increasing number of cylinders. We have a few features which track a positive correlation and are similar and a grouping we can potentially consider outliers and appear to not be similar to the cylinder feature, yet they still maintain the positive upward correlation trend relative to the cylinder count.

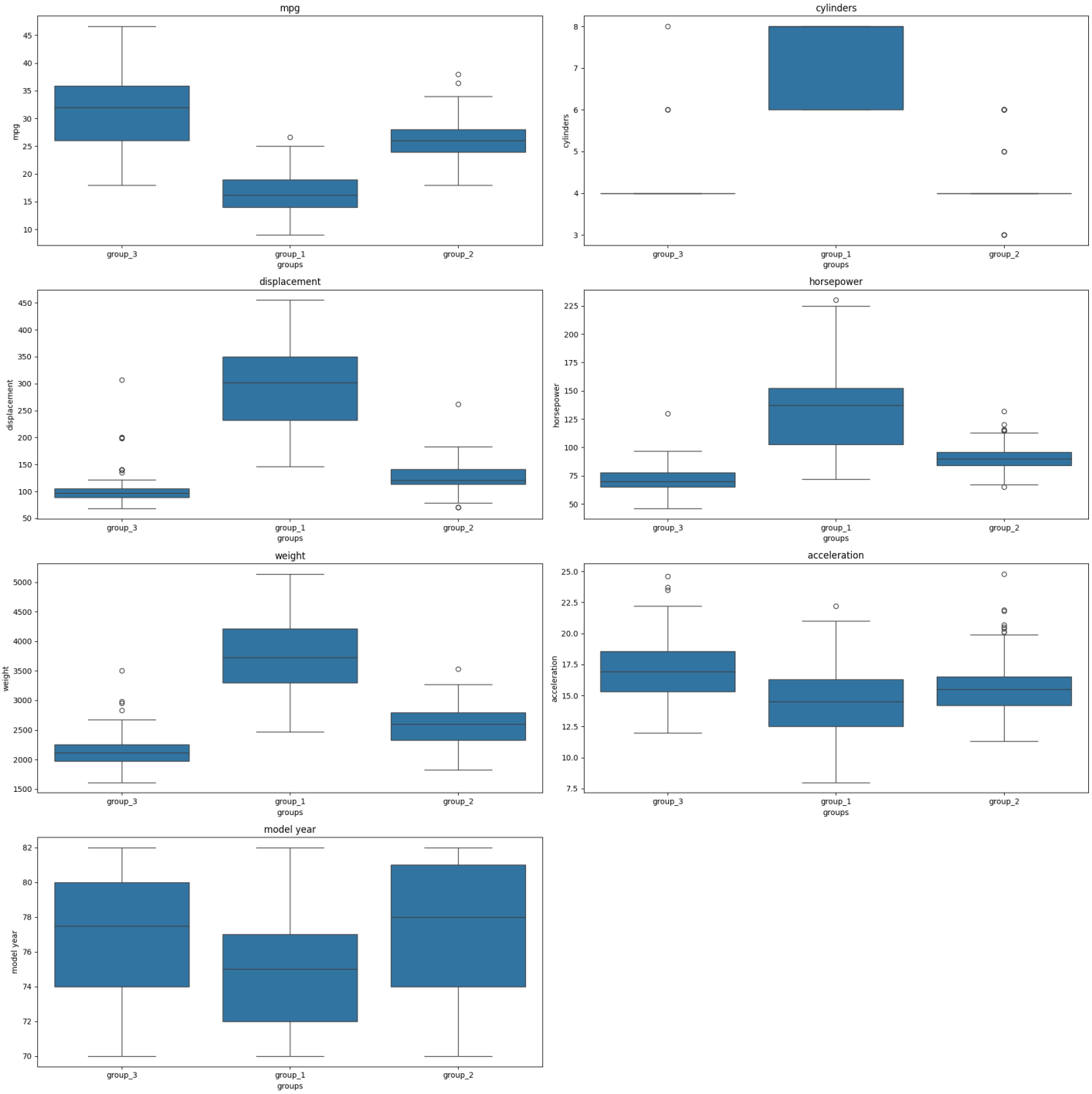

cylinders
3
4
5
6
8
PC2
PC1
2
1
0
−1
−2
−4
−2
0
2
4

This t-SNE scatter plot is also looking at the number of cylinders. We can see that with the 3 & 4 cylinder grouping, there are more similar features than with the other number of cylinders.



The boxplots below show the significance of features within three groups.

Gp 1 greatest = cylinders, displacement, horsepower, *weight
Gp 1 middle = na
Gp 1 lowest = mpg, model year, **acceleration



- *Given the high correlation weight has with the other features of great importance to group 1, it is logical to conclude weight is a byproduct measurement of the other features in the grouping and not a feature of importance related to a purchasing decision.
- **Acceleration is low for group 1 as a byproduct measurement of the inverse relationship to the features of great importance for group 1. It is logical to conclude this is not a feature of less importance related to a purchasing decision.
    - **Group 1 Consumer**
        - **primary decision factor = powerful engine features even if it means a heavier car (also = less time to peak acceleration)**
        - **not a decision factor = mpg or model year**
    - **Vintage muscle car (heavy focus on the muscle) of any model year**

Gp 2 greatest = model yr
Gp 2 middle = mpg, displacement, horsepower, weight, acceleration
Gp 2 lowest = na

- Mpg has an inverse relationship to the other features of moderate concern to this consumer so it is a logical to conclude this consumer group will fall into 1 of 2 subcategories.
    - **Group 2 Consumer A**
        - **primary decision factor = model year**
        - **secondary decision factor = low mpg with a powerful engine**
    - **Classic vintage model muscle cars**

    - **Group 2 Consumer B**
        - **primary decision factor = model year**
        - **secondary decision factor = more mpg with a less powerful engine**
    - **Vintage cruisers**
    - **Classic vintage model muscle cars**

- **Newer model vintage cars with good gas mileage in relation to the model year**

Gp 3 greatest = mpg, *acceleration
Gp 3 middle = model year
Gp 3 lowest = displacement, horsepower, weight

- Given this consumer group is least concerned with engine power, it is logical to conclude their high concern with mpg is their preference for a good mpg rating, which is inverse to a powerful engine.
- *We see acceleration as having high importance due to the inverse relationship to the engine features. This is not a not a feature of greater importance related to a purchasing decision, but is more reflective of the inverse relationship with the engine features.
    - **Group 3 Consumer**
        - **primary decision factor = high mpg rating**
        - **secondary decision factor = model year**
        - **not a decision factor = engine power**
    - **Newer model vintage cars with good gas mileage in relation to the model year**

Conclusions and Recommendations

SecondLife desires to shift predominately into the vintage car market. An analysis of past sales of vintage cars was done in order to gain insight into what features were dominate attributes of the cars sold.

It was found that 3 Attributes comprise 90% of the data. Those are the Engine (which includes cylinder, displacement, and horsepower features) the Model Year, and Acceleration.

It is my recommendation for SecondLife to procure inventory of 4 types of cars in order to market to the consumer groups who purchase in these groups

- Vintage muscle car of any model year, with the focus being on the engine power
- Classic vintage model muscle cars
- Vintage cruisers
- Newer model vintage cars which offer good gas mileage in relation to the model year

After procurement, in order to bring consumer awareness to the new line of Vintage offerings at SecondLife, I suggest launching a marketing campaign to past purchasers.

- A good strategy would be to host a Car Show.
- One key aspect of the marketing strategy would be to invite past customers to bring their Vintage cars to SecondLIfe's Car Show.

My recommendations will provide SecondLife with a well-rounded selection of carefully procured inventory, surely suitable to fit the desires of Vintage car lovers. Utilizing your hot market of past purchasers, by hosting a car show, will increase awareness of SecondLife's new position in the Vintage car market.