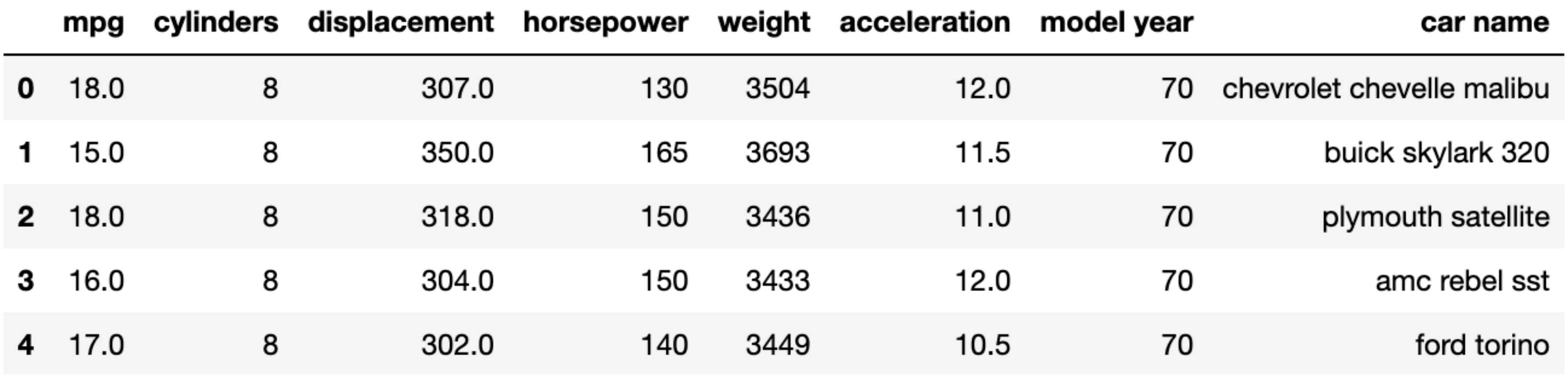
Vintage Car Segmentation

Dimensionality Reduction (PCA, tSNE)

‘Femi Bolarinwa

Data Snapshot

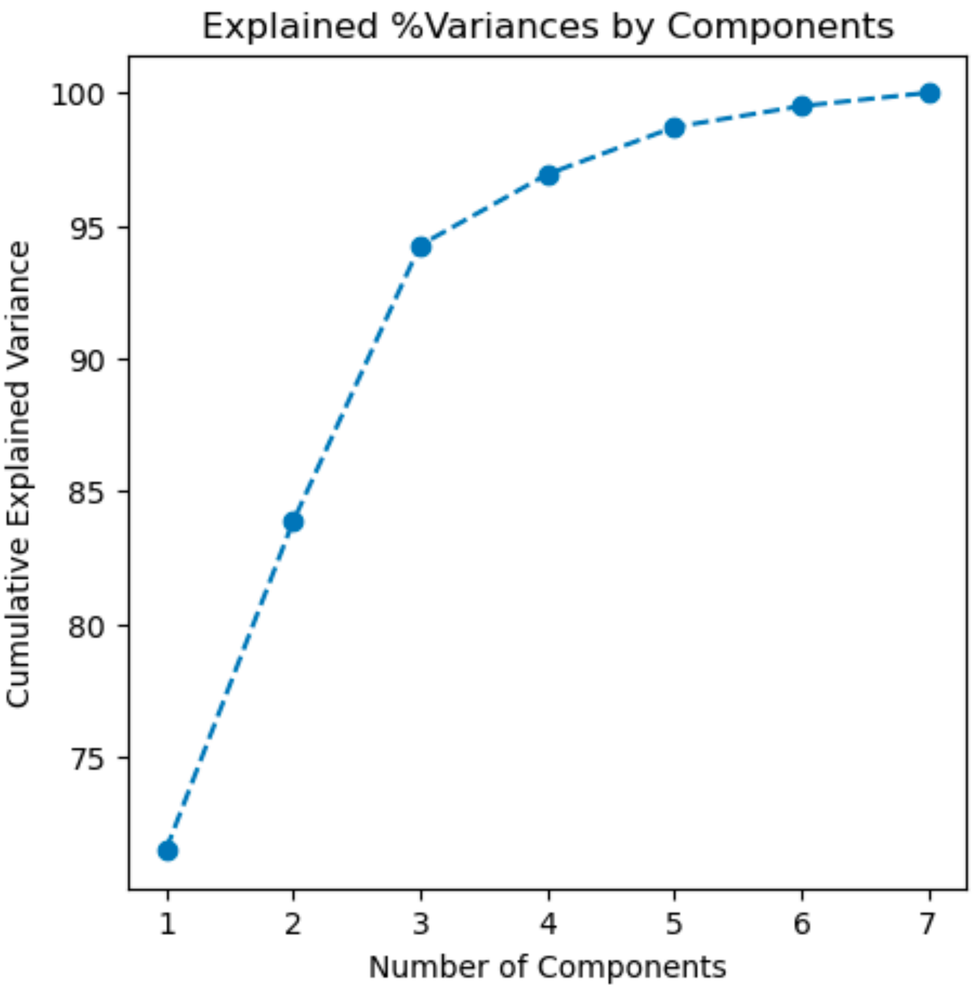


	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	car name
0	18.0	8	307.0	130	3504	12.0	70	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	ford torino

PCA

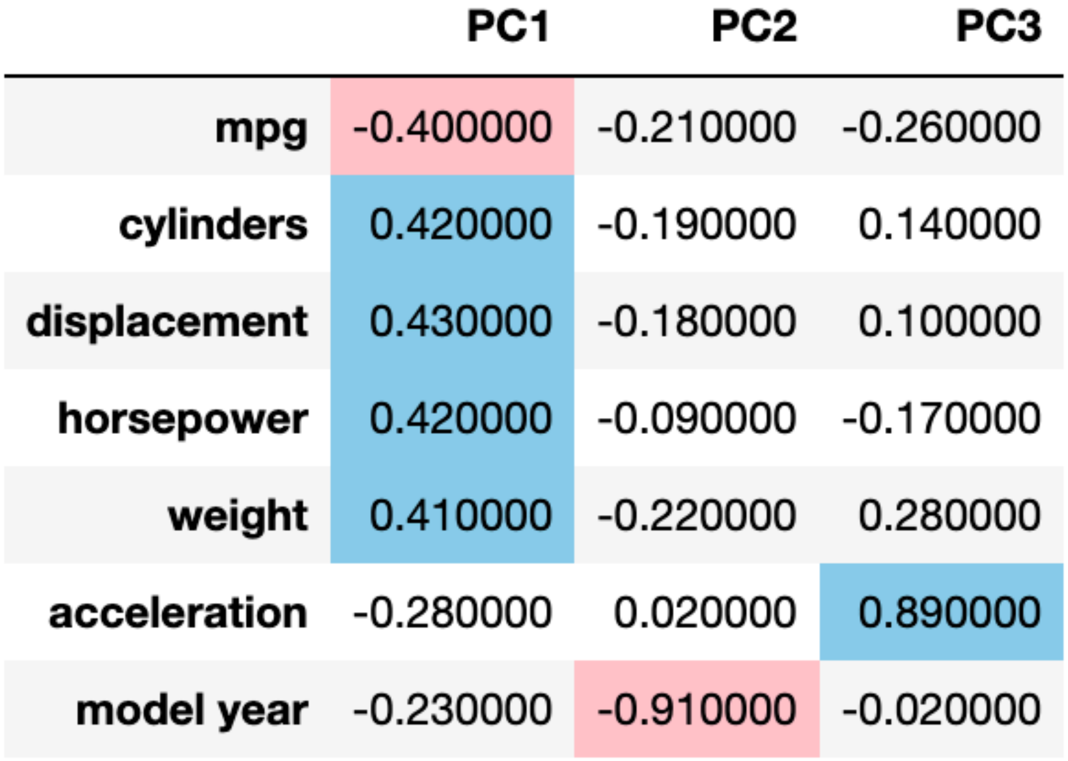
Principal Components

- Each PC represents an eigenvector of the covariance/correlation matrix of the dataset.
- The explained variances of the principal components are the corresponding eigenvalues of the eigenvectors.
- PC1 captures or explains the most variance (about 70%) in the data set. PC7 explains the least (0.5%).
- 3 of 7 components captures (or explains) about 95% of variances in the dataset.
- That means 57% dimensionality reduction with only 5% loss in explained variance.



Coefficients of PCs

- Each principal component (PC) is a linear combination of the features (columns) of the original dataset.
- The features are weighted by the coefficients shown for the first three PCs
- Some features have more effect on the principal components than others as highlighted in colors.



	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

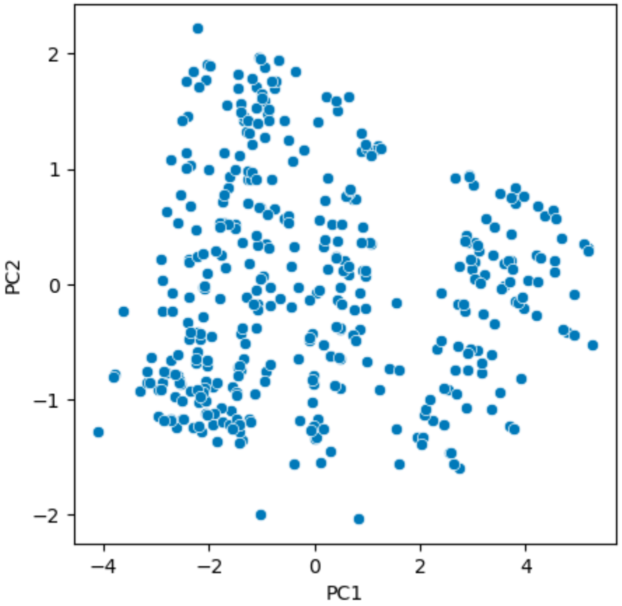
Visualizing PC Plane

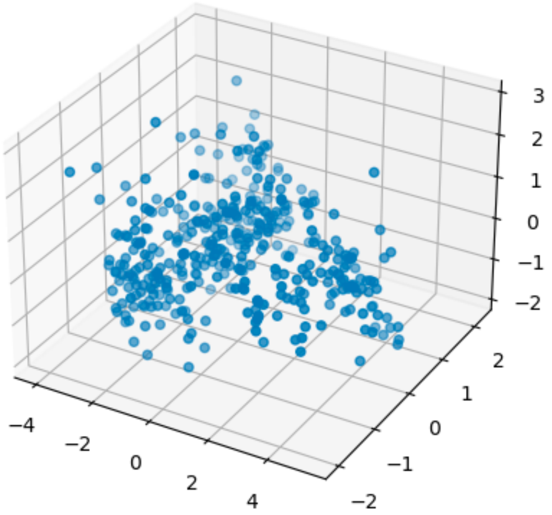
2D & 3D

- Pair of PC1 and PC2 captures the most variance based on the principal component analysis.

- But no apparent cluster or pattern either in 2 or 3D.

- A more powerful dimensionality reduction technique - TSNE (t-distributed stochastic neighbour embedding) might help.

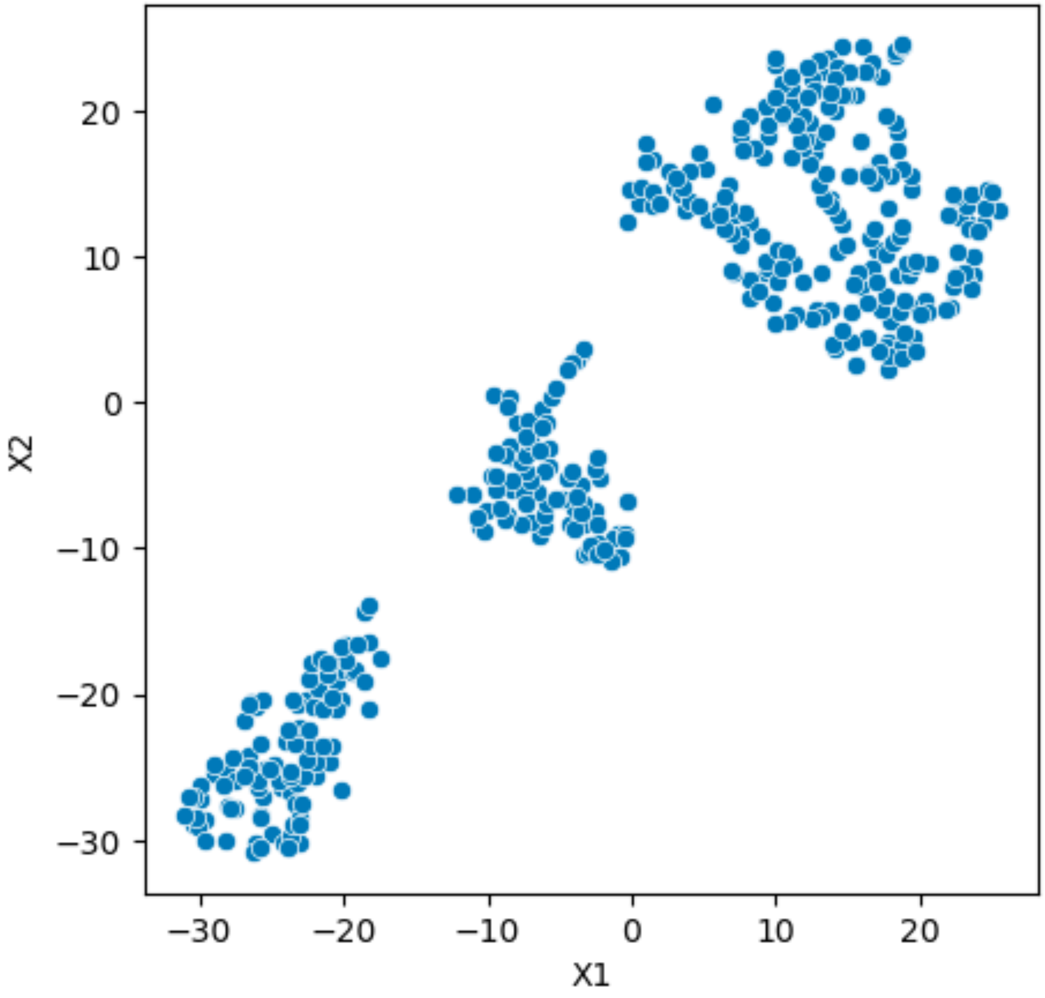




tSNE

Non-linear Embedding

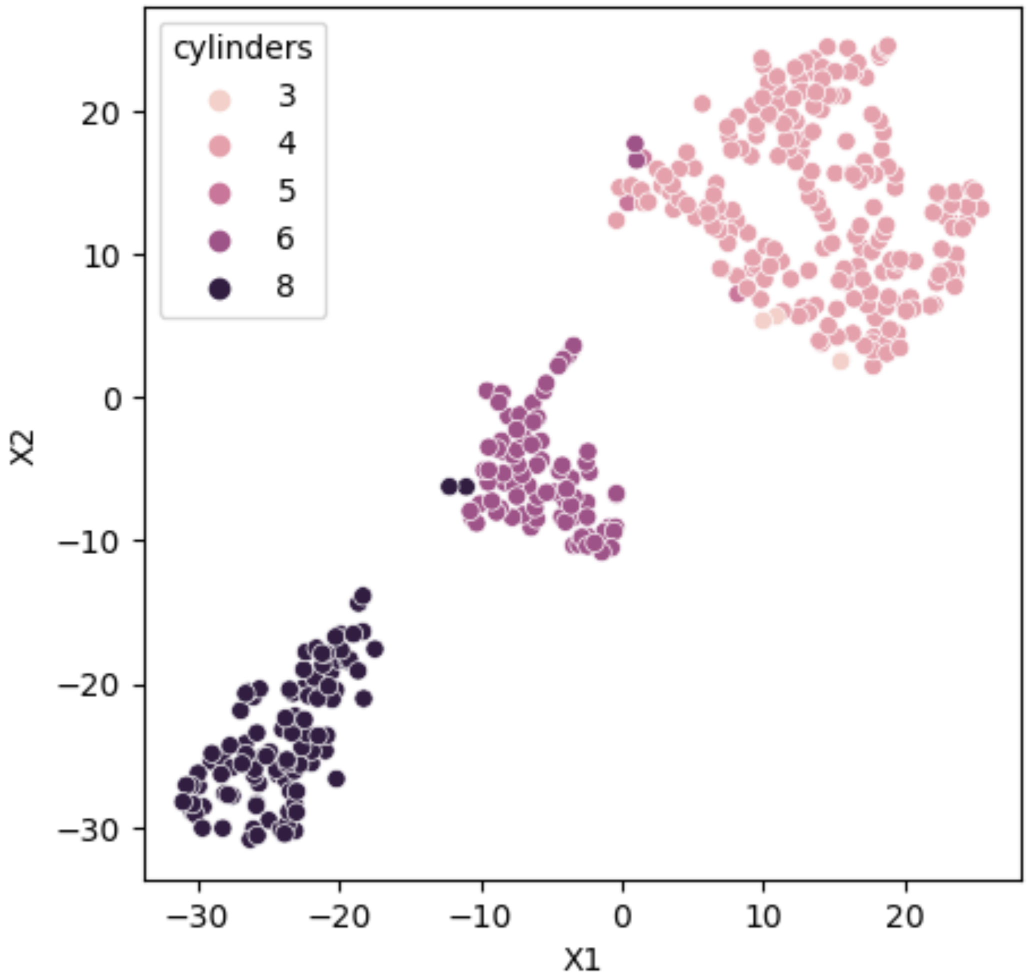
- tSNE provides clear patterns in the data but took more computation time.
- The are 3 groups of vintage cars in the data.
- I'll try to identify the peculiarity of each group or cluster. My first guess is number of cylinders.



Segmentation

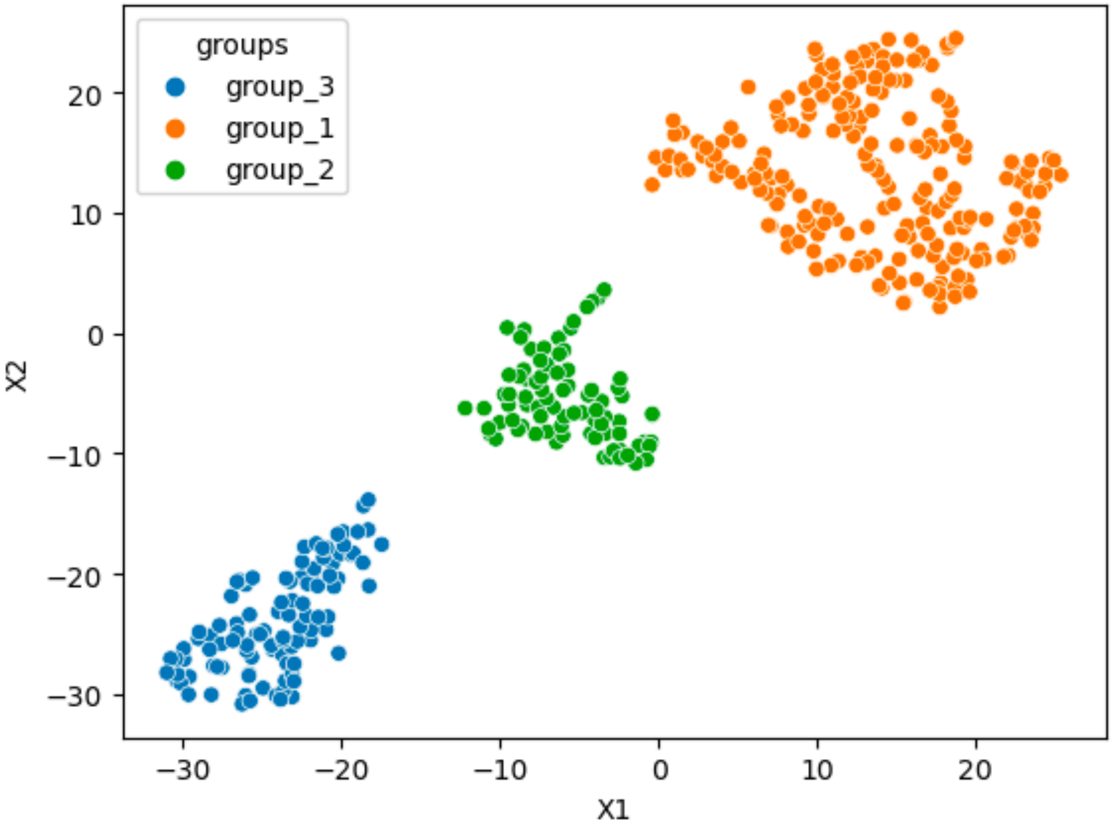
By Number of Cylinders

- Number of cylinders appears to be a major factor distinguishing the groups.
- But there is a slight overlap among the groups.
- I'll figure out other distinguishing features without using clustering algorithm (like k-mean, GMM, PAM, etc.) since clusters appear distinct enough.



Segmentation

- Grouping the 3 different clusters based on there location on the tsne-plane.
- I'll analyze the peculiarity of each cluster with respect to features in the dataset

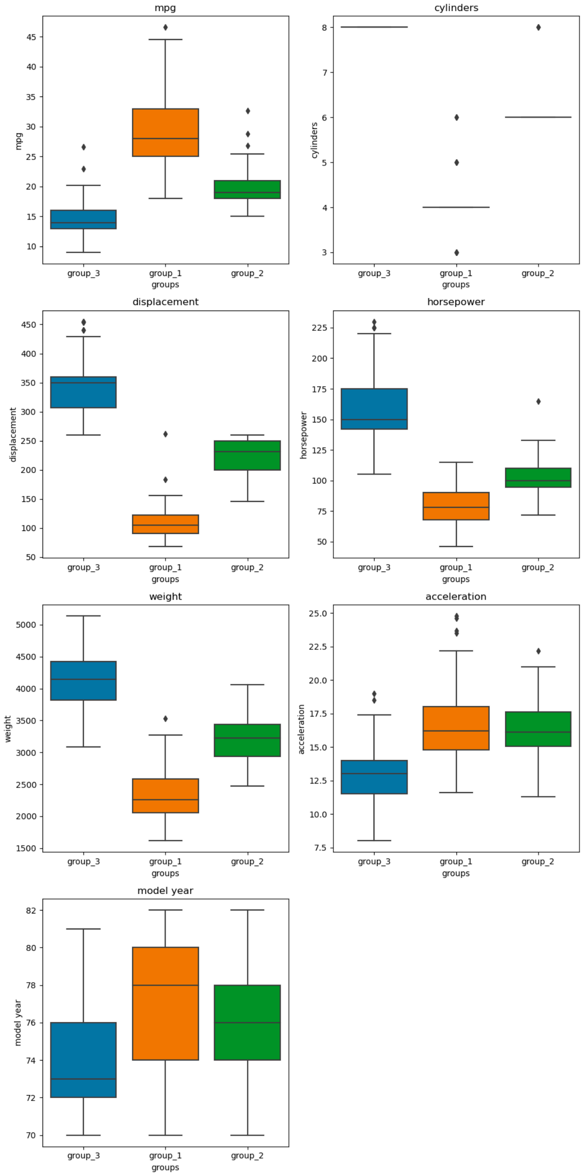


Segment Profiling

- Group 1: vintage cars with 4 cylinders, high mpg, small engine size, low horsepower, light weight, good acceleration - **LEISURE CARS.**

- Group 2: vintage cars with 6 cylinders, moderate mpg, moderate engine size, moderate horsepower, moderate weight, good acceleration - **UTILITY CARS.**

- Group 3: vintage cars with 8 cylinders, low mpg, large engine size, high horsepower, large weight, low acceleration - **HEAVY DUTY CARS.**



Business Insight and Recommendation

- Ads and promotions could be better targeted. Young people are more likely to be interested in Leisure cars, families - utility cars.
- Outlet manager can adjust the composition of there dealership shops based on the demography of the area or clientele. This also applies to new outlets.