6302 Assignment 1 Principal Component Analysis
Austin Steel

For this project I am using the dataset *Elite Sports Cars in Data* from Kaggle. This dataset contains synthetic information on 5,000 sports cars. Each entry has 27 characteristics for the vehicle. The target variable for this analysis will be *Price.*

Preprocessing

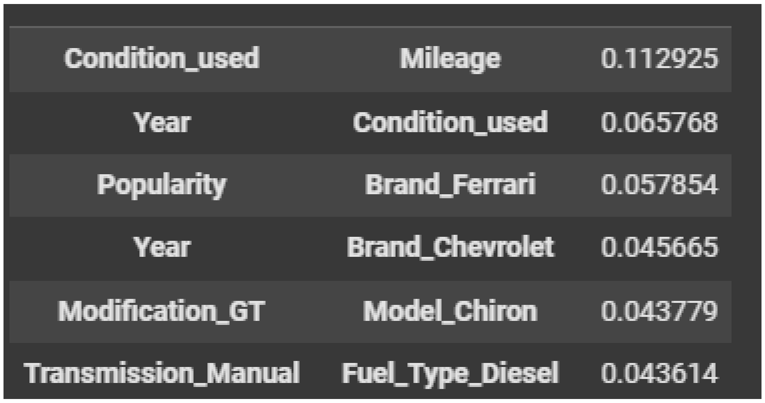
After importing the dataset and necessary packages, I displayed the first and last ten rows,then checked the data types, and checked for any missing values. I did not have any missing values, however the columns are both categorical and numerical. I checked each categorical column to see how many unique values there were. Given there was less than 50 total unique values, I decided it would be best to use one hat encoding to include these values in my analysis. The columns *Market_Demand,* and *Popularity* were values of *Low, Medium, High*. I thought these values would best be replaced with 1, 2, and 3 respectively. There were also two columns which seemed repative; *Log_Price,* and *Log_Mileage*. While these columns could be helpful for some analysi, they don't seem appropriate here, so I removed them. Now I have a dataset that is suitable for analysis.

Sampling

I sampled the data using three different methods. First I took a random sample, found by the program randomly selecting 150 entries. This sample would be good for most analysis, but may miss some trends which could be seen through other sampling. Second I took a stratified sample using the column *'Popularity'* as an index. There are three entries for *'Popularity'*, *'Low', 'Medium',* and *'High'*, which I have replaced with 1, 2, and 3. I choose this column as the index because it is the most relevant in whether the car is selling or not. A stratified sample like this could allow us to see trends among the different categories of Popularity. Lastly, I took a systematic random sample. This is just another method of sampling which could help remove any bias.

Correlation Matrix

Checking the correlations of the variables is a useful way to see if we have multicollinearity, and may need to remove some columns. I first dropped the target variable, *Price*, and then created a correlation matrix. I unstacked the correlation matrix, sorted them, removed the duplicates, and printed the top 10 correlations for evaluation. The highest correlation we see is between *Condition_used* and *Mileage.* This correlation makes sense, we would expect used cars to have a higher mileage than new cars. However, it is still very low at 0.113, and therefore not worth deleting a column over. Also, while used cars certainly have a higher mileage than new cars, there is valuable information in both entries. The next highest correlation we see is between *Year* and *Condition_Used* at 0.066. Again, we would expect the older the year on the car, the more likely the car is used, but this information is valuable in both columns and the correlation is very low, so we do not have multicollinearity. The third highest correlation we see is between *Popularity* and *Brand_Ferrari* of 0.058. This is also very low. Ultimately, I do not see any multicollinearity in my dataset, so I will not delete any columns.

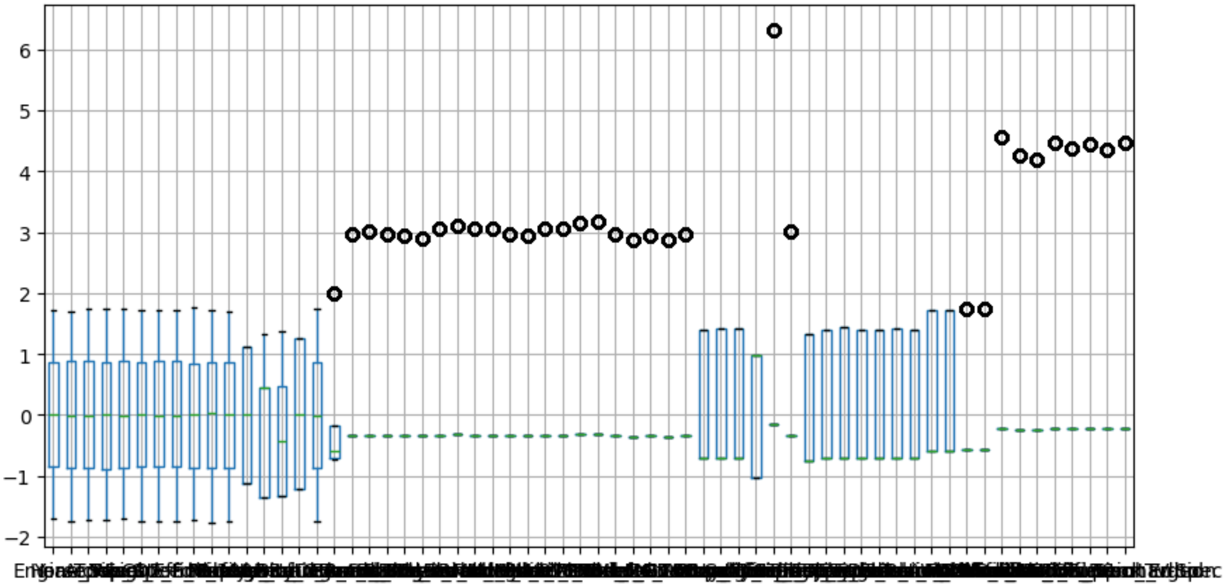


Condition_used	Mileage	0.112925
Year	Condition_used	0.065768
Popularity	Brand_Ferrari	0.057854
Year	Brand_Chevrolet	0.045665
Modification_GT	Model_Chiron	0.043779
Transmission_Manual	Fuel_Type_Diesel	0.043614

Normalization

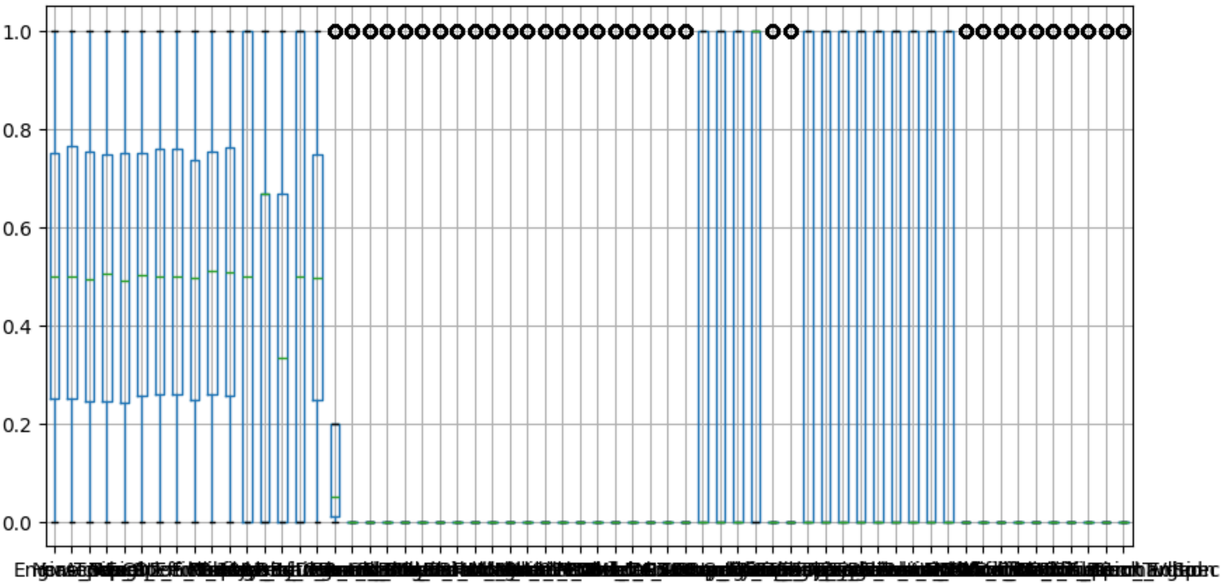
I used three different methods to normalize the dataset, Standardization, Min_Max Scaling, and Robust Scaling. Using boxplots, I assessed the distribution of each.

Standardized



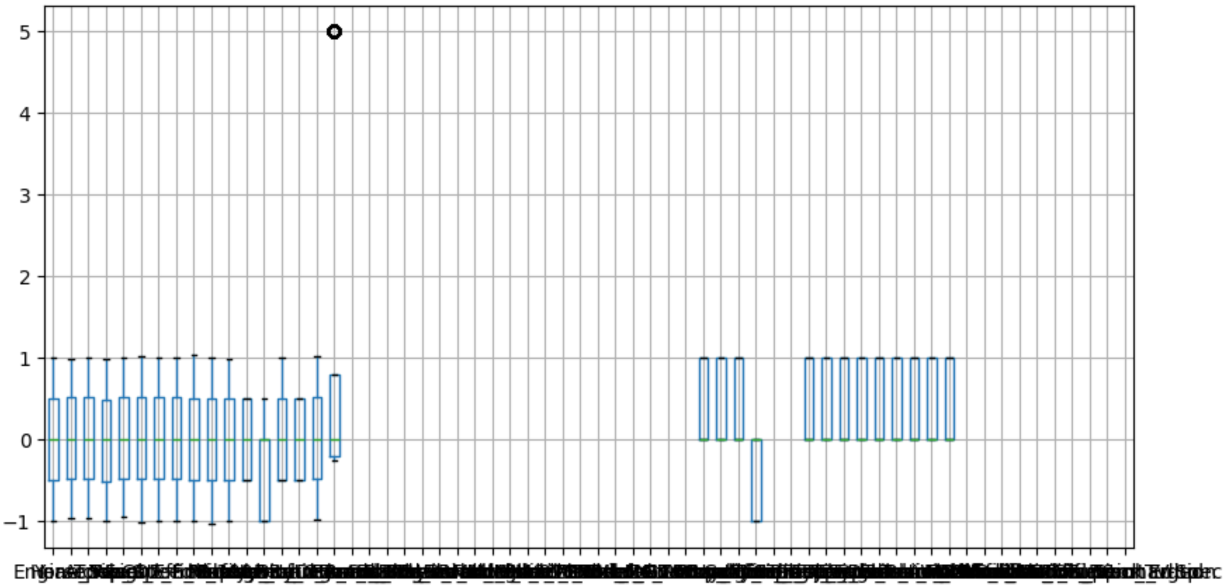
The Standardized dataset ranges from -2 to 6 and has several outliers on the high end.

Min_Max Scaling



The Min_Max scaled dataset ranges from 0 to 1, and also has several outliers on the high end.

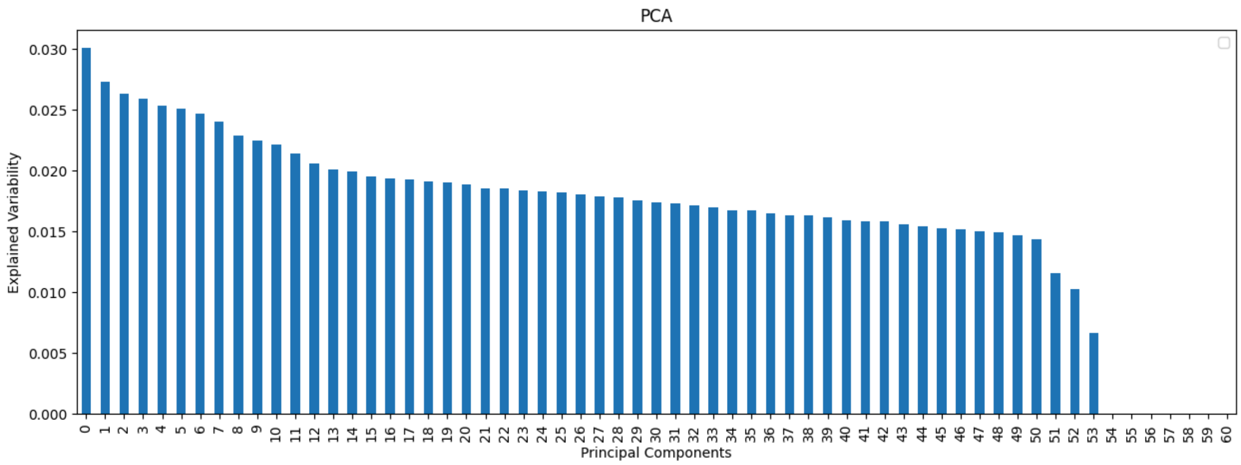
Robust Scaling



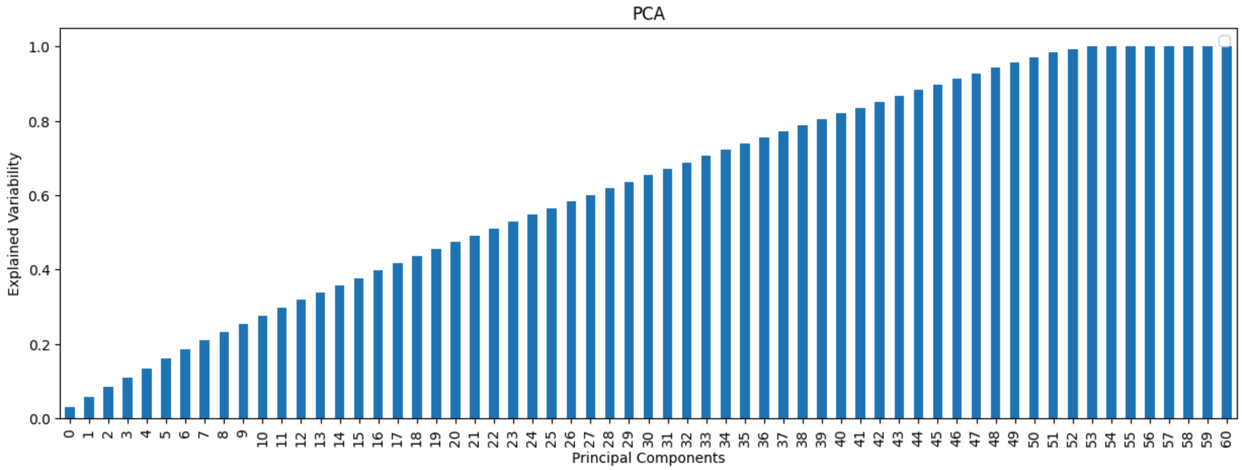
The Robust scaled dataset ranges from -1 to 5 and only has one outlier. Because this one has the least outliers I think it would be the best one to use.

Principal Component Analysis

From here I applied a Principal Component Analysis to the standardized dataset. Below is a bar graph representing the variance of each component.

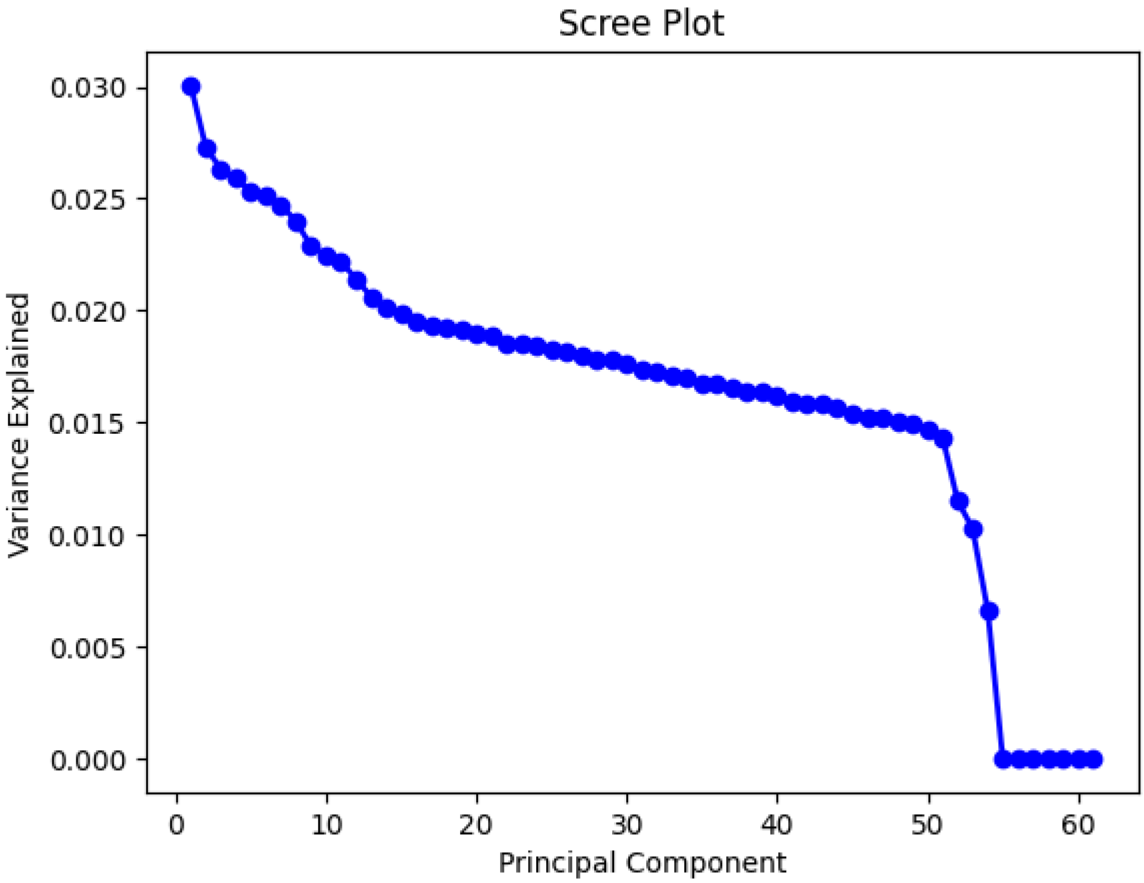


None of the components have a very large variance. To find out how many components we need to consider to explain at least 85% of the variance, I recreated the array using the cumulative summation formula, and graphed the cumulative variance as seen below.



To get to 85% we need to consider 43 of th principal components. These are: 'Year', 'Engine_Size', 'Horsepower', 'Torque', 'Weight', 'Top_Speed', 'Acceleration_0_100', 'Fuel_Efficiency', 'CO2_Emissions', 'Mileage', 'Popularity', 'Safety_Rating', 'Number_of_Owners', 'Market_Demand', 'Insurance_Cost', 'Production_Units', 'Brand_Aston Martin', 'Brand_BMW', 'Brand_Bugatti', 'Brand_Chevrolet', 'Brand_Ferrari', 'Brand_Ford', 'Brand_Lamborghini', 'Brand_McLaren', 'Brand_Nissan', 'Brand_Porsche', 'Model_488 GTB', 'Model_720S', 'Model_911 Turbo S', 'Model_Chiron', 'Model_Corvette Z06', 'Model_DBS', 'Model_GT-R', 'Model_Huracan', 'Model_M4 Competition', 'Model_Mustang GT', 'Country_Asia', 'Country_Europe', 'Country_USA', 'Condition_new', 'Condition_restored', 'Condition_salvage', 'Condition_used'.

This is almost all the variables except the ones related to Fuel Type, Drivetrain, Transmission, and Modifications. The Scree Plot below shows that several variables have no variance (those towards the left side of the plot).


Scree Plot
Variance Explained
Principal Component