LAB 2

Christina Hartnett

I am a avid viewer of NASCAR and I found it very interesting that there is almost no visualizations or analysis done on NASCAR driver data. With many other sports there is an endless amount of visualizations available so I thought it would be important to visualize NASCAR Data. I thought this data would be interesting to see if there was any strong correlation to being in the top 15 to winning a race. I also thought an interesting aspect was how overall driver rating could be influenced by the other values. These should all somewhat correlation because they are all taken from the same races but in NASCAR style racing being a great driver and having a good average start does not always equate to winning because of unpredictable things such as crashes or car parts breaking. I was very interested to see that almost most of the variance can be described using two principal components. This is very good for our biplot and other visualization.

While looking at the biplot, it can easily be noticed that the most directly correlated attribute to PC1 is Driver Rating. This makes a lot of sense because Driver Rating would be based off of all the other attributes and explain their variance very well. The closest attribute to PC2 is Pass Difference, this is the difference between Driver passes and how many times they were passed by others. This attribute is not very significant overall because we can see it is a short line and not directly correlated to either principal component.

PCA has some disadvantages because we standardized our data to make it easily interpreted and comparable. In PCA all categorical data needs to be converted into numerical before PCA can be used. PCA also does not preserve similarity relationships as well as MDS.

The scatterplot matrix only shows us each relationship individually, so this does not tell us how they all interact over 4 dimensions. It only shows us the 2 attributes dimensions. This makes it very hard to see multivariate relations. Therefore, we also use Biplots. It does help us easily visualize the most important attributes depending on the % variance chosen.

The downfalls of a biplot are that the data needs to be described in only two principal components. This means that if we have more than two principal components that are important, we will be losing all of this variance. It can be useful in showing us which attributes are important and how correlated different attributes may be.

MDS can help us determine when data values are truly similar. If they are not similar, they will not be plotted near each other, which is very different than other dimension reduction techniques. One downfall to using MDS is that the axes are hard to interpret because they are not really axes.

One issue with parallel coordinates is that there are many different ordering options, especially as the number of attributes gets larger. We should always try to find a meaningful ordering, such as correlated value ordering, to make our scatterplot more comprehensive. A parallel coordinate plot can help us visualize all of our attribute relationships on one plot, which is more useful than the scatterplot matrix in that sense.

This data set was joined from https://www.nascar.com/stats/2021/1/box-score and https://frcs.pro/nascar/cup/drivers/point-standings/2021/36 to find out more information on individual drivers.

LAB 2
Christina Hartnett

The attributes are listed below:

Year

Driver Name

# Fastest Laps
Number of laps where the driver had the fastest speed on the lap.

% Laps in Top 15
The sum of the Laps Run in The Top 15 divided by Total Laps Completed.

% Laps Led
The sum of the Laps Led divided by Total Laps Completed.

% Quality Passes
The sum of the Quality Passes divided by Green Flag Passes.

Average Finish
The sum of the driver's finishing positions divided by the number of finishes.

Average Position
The sum of the drivers position each lap divided by the laps run in the race.

Average Start
The sum of the driver's starting positions divided by the number of starts.

Driver Rating
Formula combining the following categories: Win, Finish, Top-15 Finish, Average Running Position While on Lead Lap, Average Speed Under Green, Fastest Lap, Led Most Laps, Lead-Lap Finish. Maximum: 150 points per race

Green Flag Passed
Drivers being passed the most during green flag conditions in the race.

Green Flag Passes
Drivers with the most passes during green flag conditions in the race.

Laps in Top 15
Number of laps completed while running in a Top 15 position.

Laps Led
The sum of the Laps Led in the race.

LAB 2
Christina Hartnett

Pass Diff
The sum of Green Flag Passes minus Green Times Passed.

Points
Refers to the points awarded to the driver for finishing positions during the stages and race.

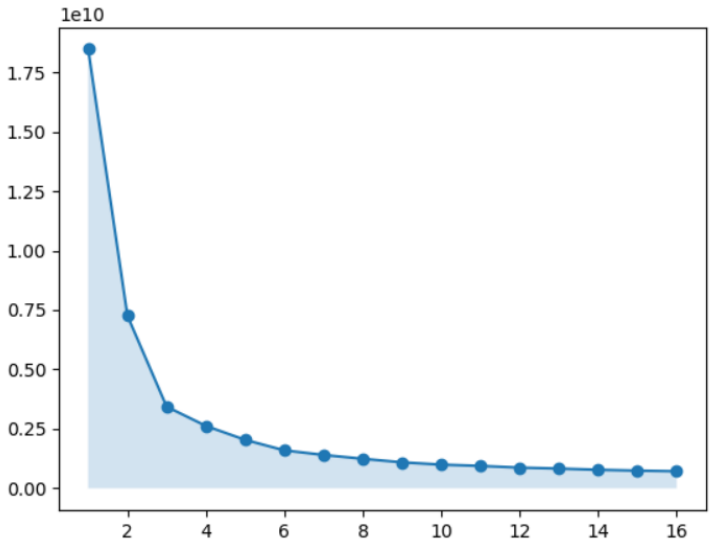
Quality Passes
Passing a car running in the Top 15 while under green flag conditions.

Total Laps
The sum of the laps completed by a driver.

Wins
The sum of a drivers' victories.

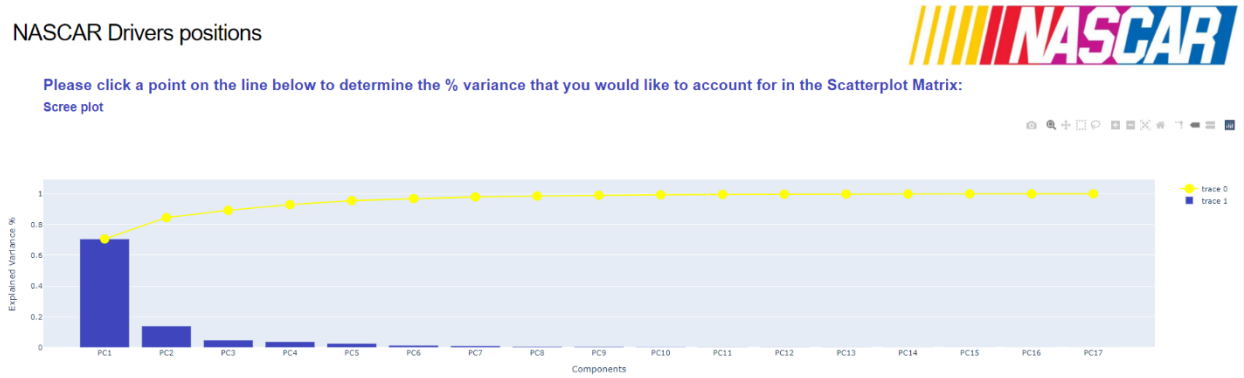
Car
The driver’s car manufacturer.

This is the elbow plot to decide the number of clusters (K) that are necessary. This uses the MSE on the Y-axis and the number of clusters on the X-axis.



LAB 2
Christina Hartnett

# Page Layout:




This is the screeplot that displays the variance and the cumulative variance of each principal component. This is an interactive plot where the user can click the cumulative line plot to decide how much variance they would like to account for in the Scatter Matrix and the Table.

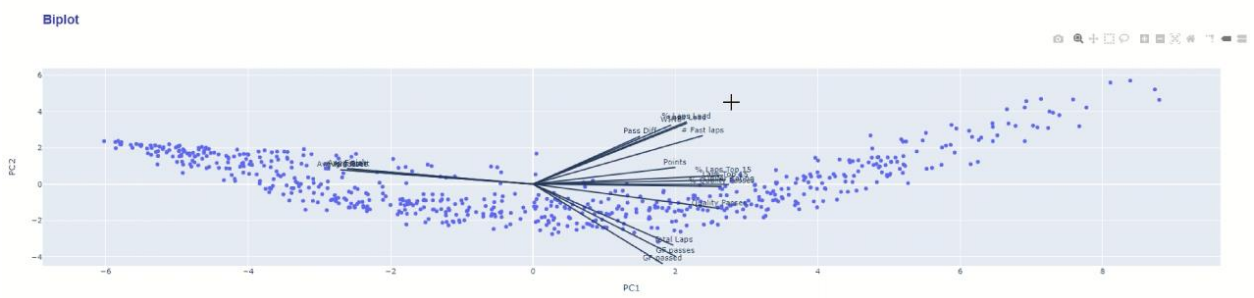

This is the scatter matrix showing the correlation between the top four attributes based on the user input for cumulative variance. We will calculate the loadings between the attributes and the principal components and use the sum of square loading to find the top four attributes.

Table of highest PCA Loadings

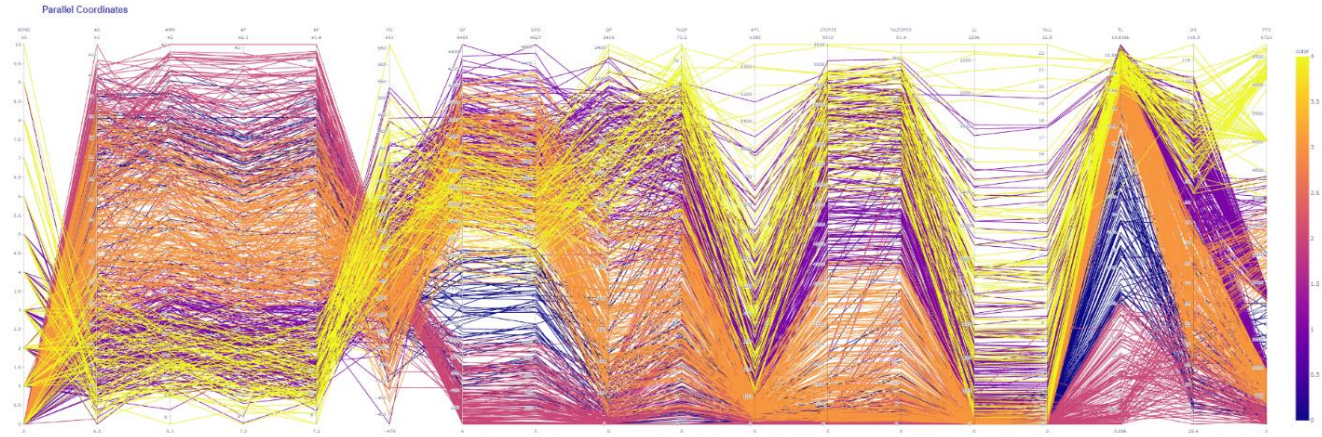
| PC1 | PC2 | PC3 | SSS | Attributes |
|---|---|---|---|---|
| 0.18216542391116897 | -0.43946079912492086 | 0.288828001905120 8 | 0.30973165032075445 | GF passed |
| 0.19707988519134606 | -0.3368048122401312 | 0.3576570306543145 | 0.2801965142716054 | Total Laps |
| 0.21683335750082272 | 0.336662353562238 | 0.3173629824153591 | 0.26107750783871647 | Laps Lead |
| 0.19395221145285285 | 0.3257556431409297 | 0.34207393122899304 | 0.26074877379207073 | WINS |

This is a table of the top four attributes using the user interaction, in this case the user picked PC3 for how much variance will be explained. These are the loadings between the top 4 attributes and the principal component.

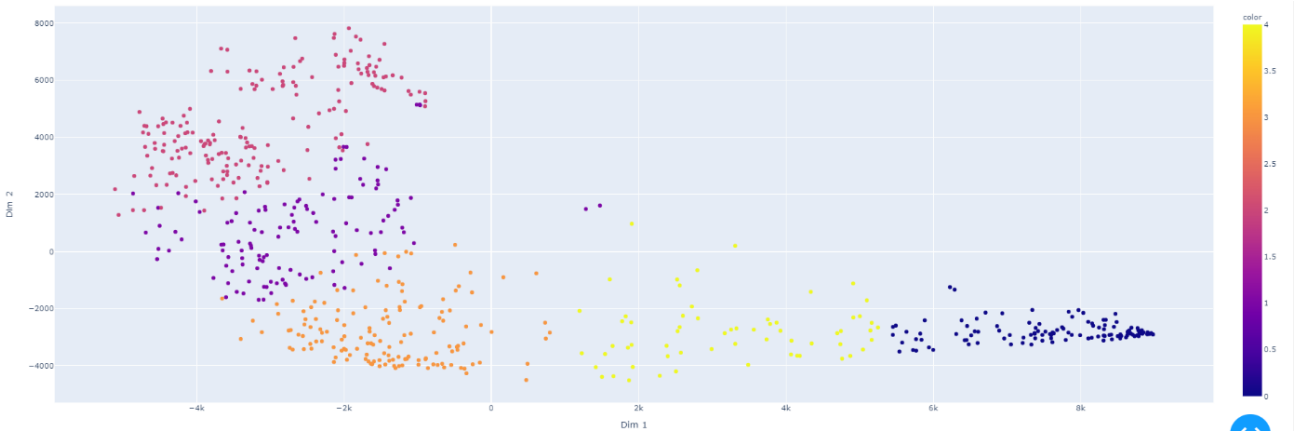
LAB 2
Christina Hartnett




This is the Biplot which plots the data in terms of the top two principal components. It also plots the attributes as lines to help understand the importance of the attributes to these principal components. The more correlated the attribute the closer it will be to the principal component.

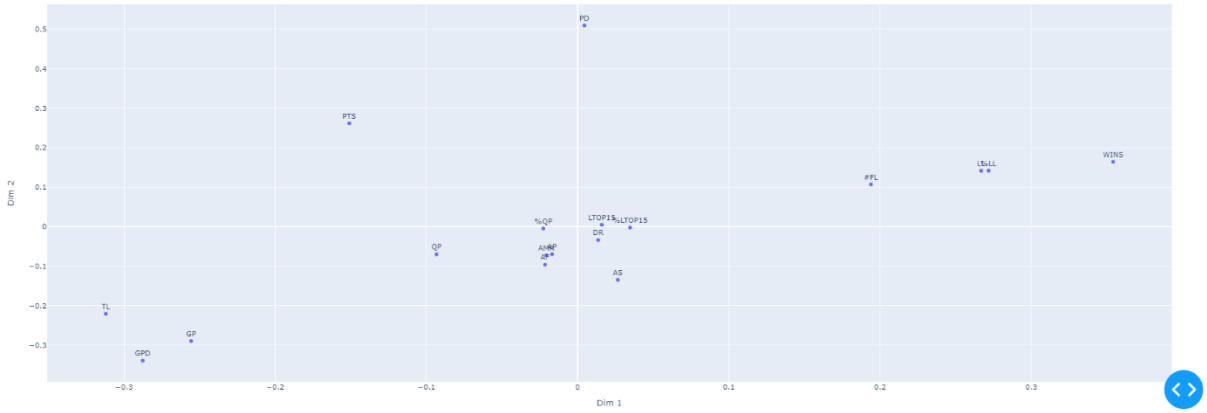



This is the interactive parallel coordinates plot that allows the user to decide the ordering of the attributes. I have also applied K-means clustering to this plot using K=5.



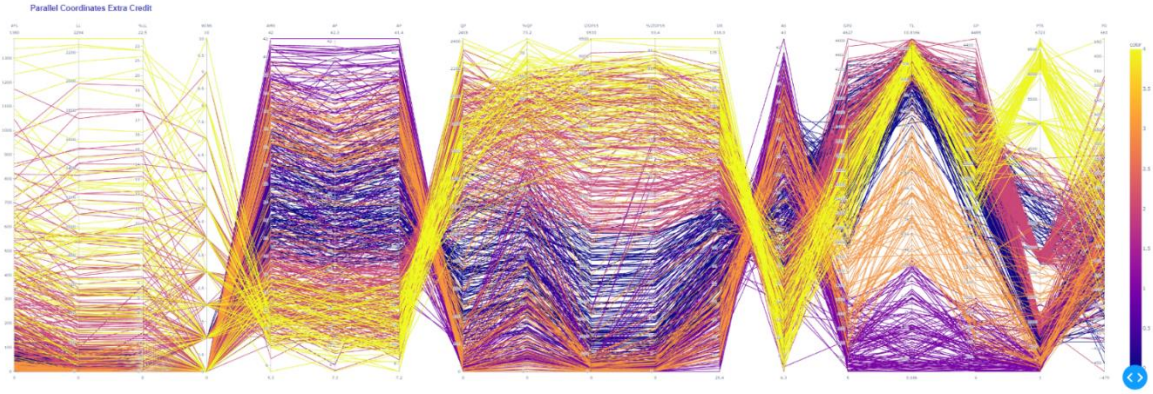

This is the MDS plot of the data points. We used the Euclidean distance to calculate the distance between these data points over higher dimensions. We hope to minimize the difference between the real distance and the dimension reduction distance (we would prefer the distances to be equal) to preserve similarity relationships. K-means clustering was preformed using K=5.

LAB 2
Christina Hartnett

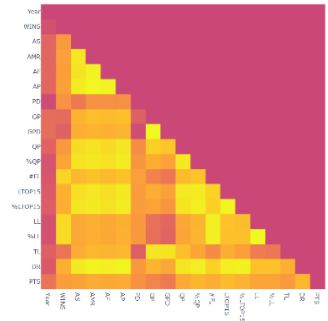



The MDS plot of the attributes uses 1- correlation to determine the true distance between these attributes. The plot seems to show that there are four groups of attributes.

## Extra credit:




We can see using the correlation matrix and grouping correlations greater than the absolute value of .7 that below are the attributes that can be grouped:



['# Fast laps', 'Laps Lead', '% Laps Lead', 'WINS'], ['Avg Mid Race', 'Avg Finish', 'Avg Pos', 'Quality Passes', '% Quality Passes', 'Laps Top 15', '% Laps Top 15', 'Driver Rating', 'Avg Start'], ['GF passed', 'Quality Passes', 'Total Laps', 'GF passes'], [‘Points’, ‘Pass Diff’]. The MDS plot of attributes also shows that us that these are the groups of the most similar attributes.

LAB 2
Christina Hartnett

Resources:

https://www.reneshbedre.com/blog/principal-component-analysis.html

https://github.com/Coding-with-Adam/Dash-by-Plotly/blob/master/Dash%20Components/Graph/dash-graph.py

https://github.com/plotly/dash-recipes/blob/master/dash-scattergl-multi-dropdown-bug.py

https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html#:~:text=PCA%20loadings%20are%20the%20coefficients,components%20(PCs)%20are%20constructed.