**Project Overview**

The goal of this project is to create a KNN Classifier which will use historical weather data to predict Over/Unders at Rockies Home games.

**Data Collection**

The primary source for my Baseball Data will be <https://www.baseball-reference.com/>. I was able to build a “webscraper” in R which allowed me to pull data for every game that the Rockies have played. I then cleaned this data and put it into a form which could be better used for my analysis and then combined with my weather data.

The primary source for my weather data was <https://www.wunderground.com/>. Here I was able to use an API along with a paid account to extract weather data from the closest weather station to Coors Field. Unfortunately I was unable to combine to pull all of the historical weather data to correspond with each Rockies home game and was only able to pull 405 records which linked up with Rockies Home games.

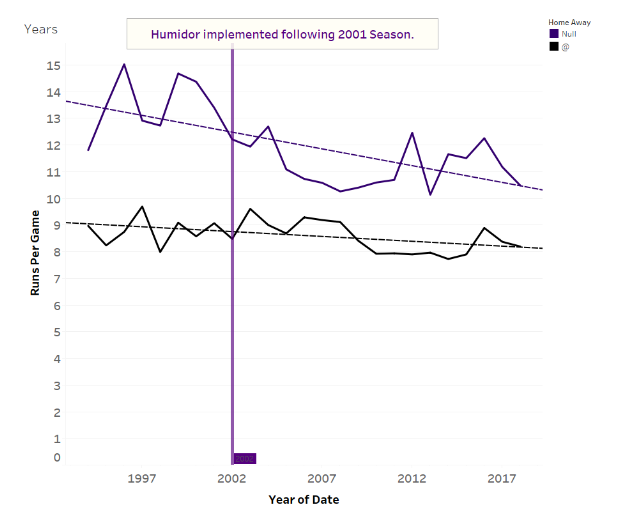
**Exploratory Data Analysis**

For my baseball data I did most of my exploratory analysis using Tableau as I am beginning to use that software more at work and wanted to get extra practice with it.

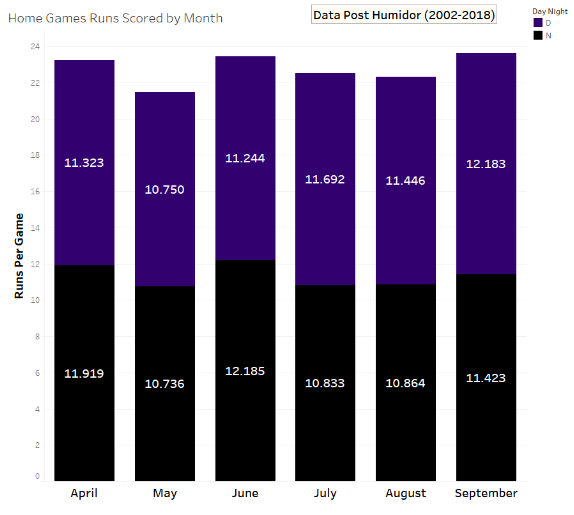
**Baseball Data**

**Findings**

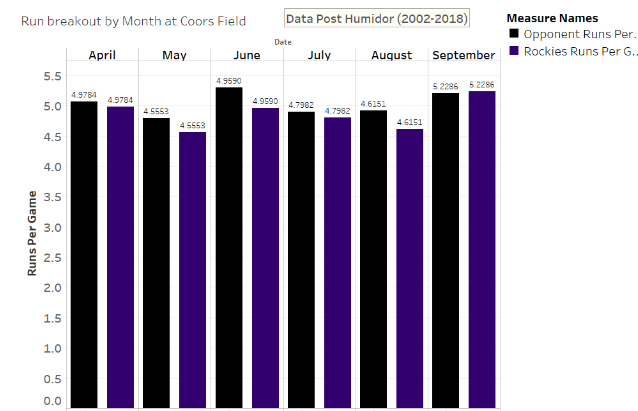
Clearly difference in total runs scored between home and away games. Additionally effects of the implementation of a Humidor in 2002 can be seen clearly in the data.



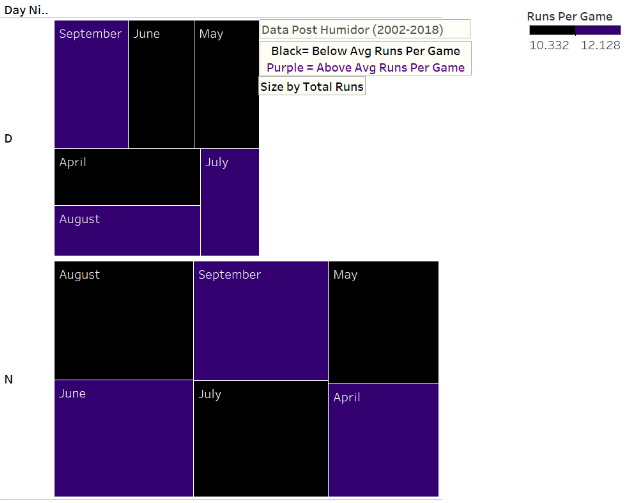
Additionally, June Night Games and September Day games showed more run production than any other month post humidor.



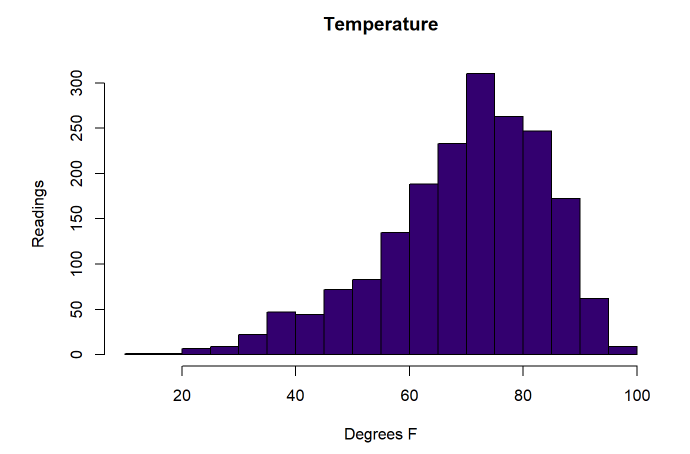
To show that it is not just a Rockies Pitching issue or Rockies having a potent offence over this time I plotted the Runs for and Runs Against

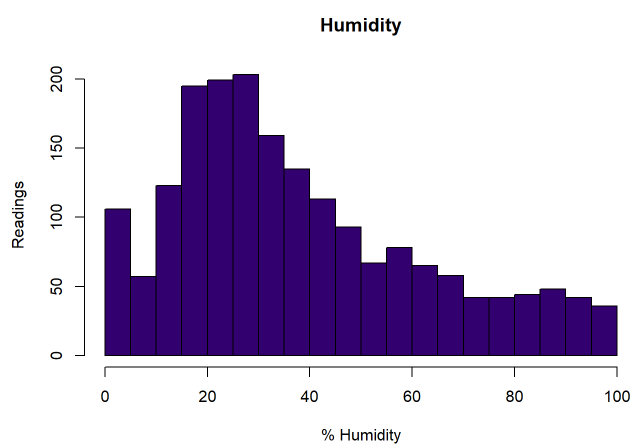


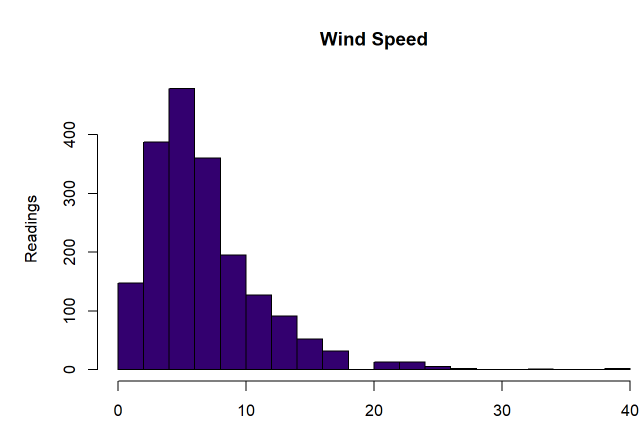
A visual I use frequently at work is the tree map which clearly shows months above the avg number of runs per game vs months with a below average total.

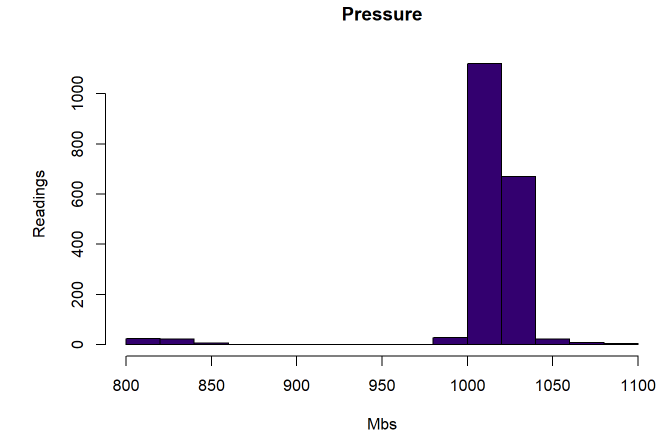


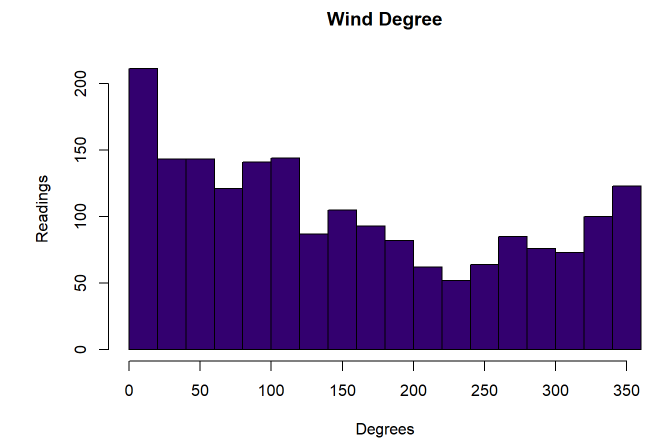
**Weather EDA**

****

****

****

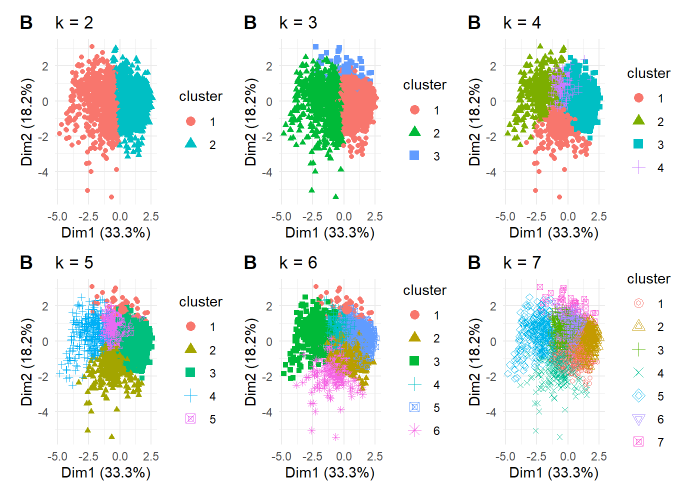
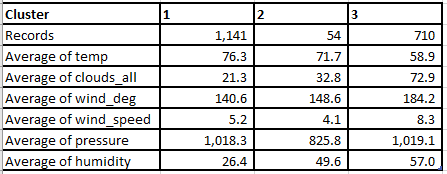
****

****

I ran some basic histograms and boxplots for my weather data, concluding that the distribution of these atmospheric conditions was not normal.

Clustering

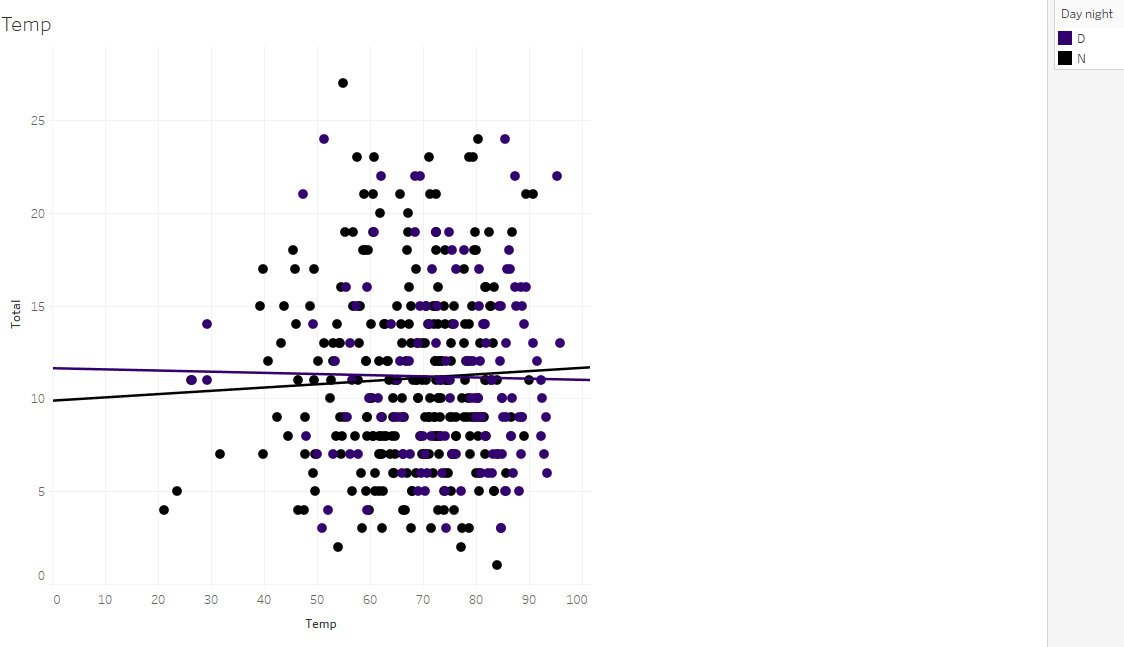
I performed clustering analysis on my weather data using a K value of 3. These results were not hugely conclusive especially when I looked at runs per game for each cluster.

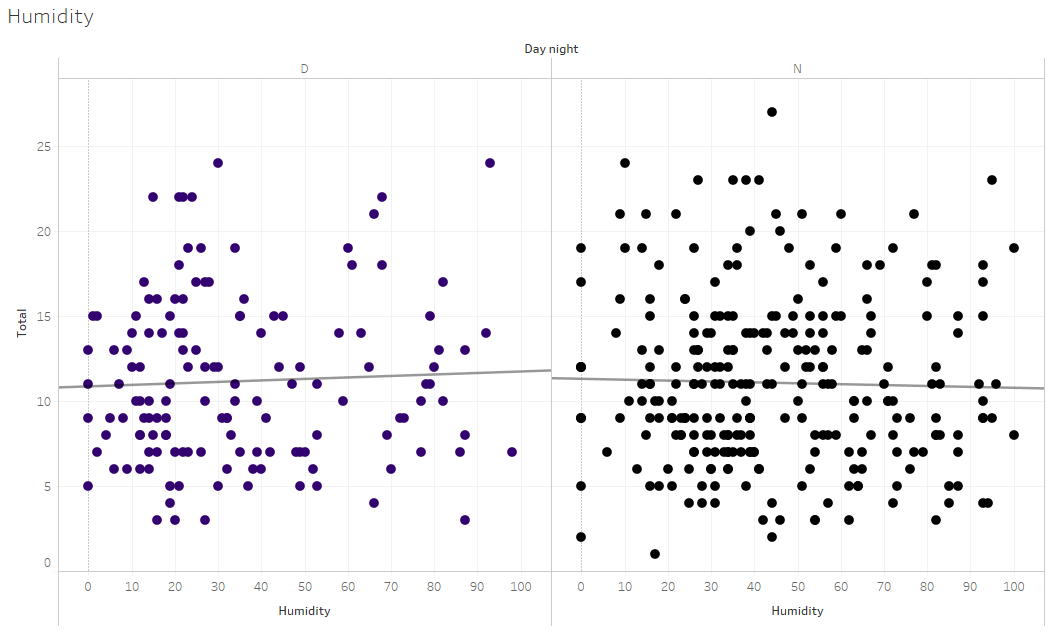
 

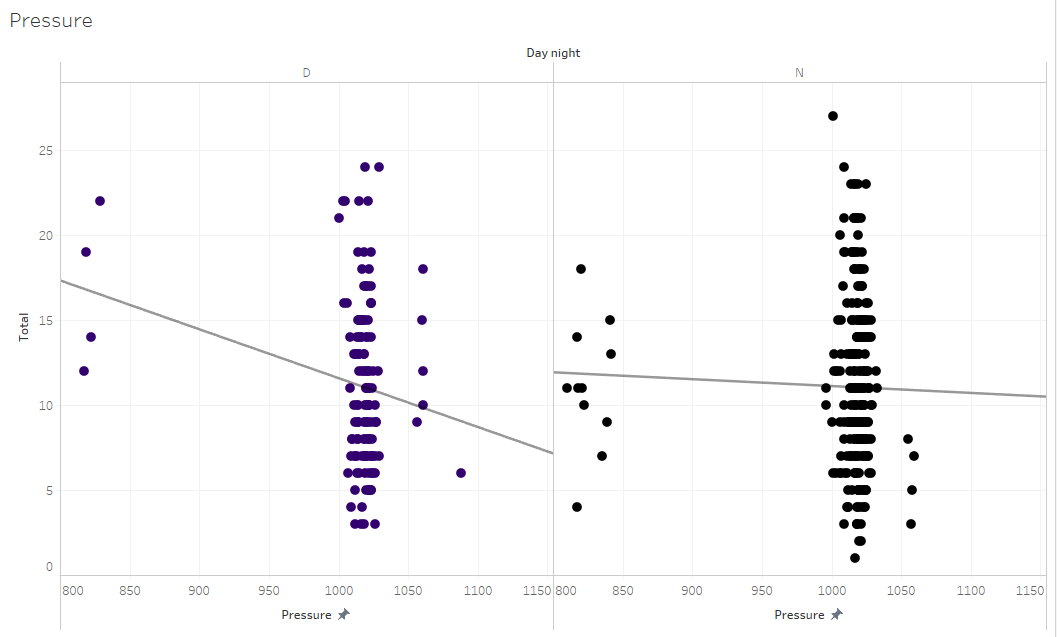
Cluster 1 shows Hot clear dayas with high presure and low humidity with a wind out of the Southeast. While Cluster 2 shows warm days with a SSE wind a little more humid with more cloud cover and low pressure. Cluster 3 shows cool humid days with a stronger wind coming out of the south.

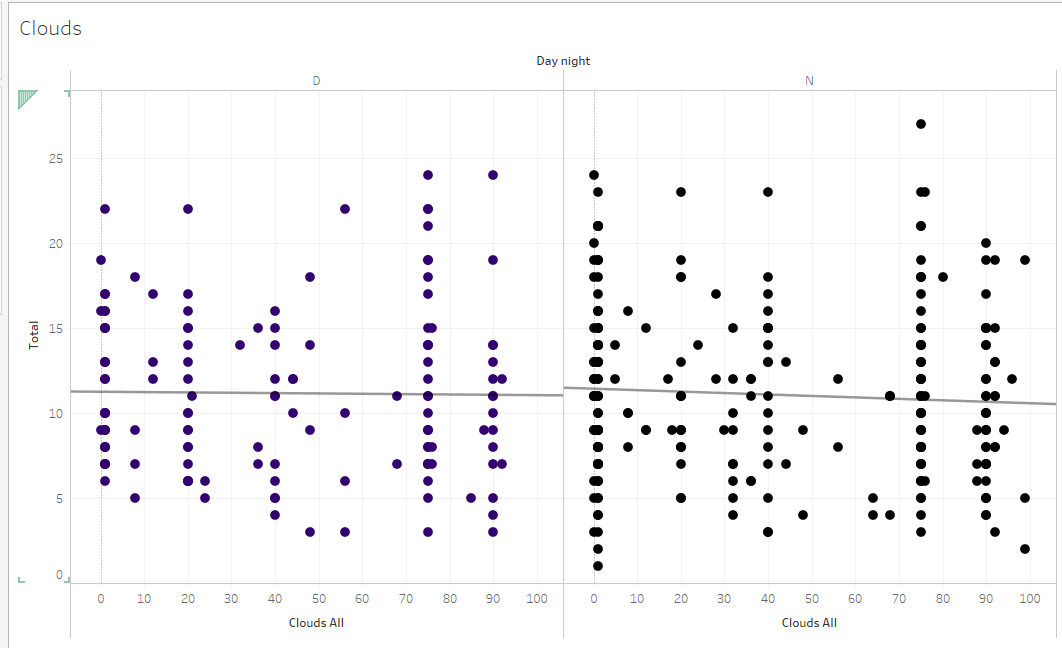
**Combined Data**

There was little conclusions which I could draw using individual measurements against runs. Additionally my correlations plots returned little in the way of clonclusive evidence of one way or another.









**Machine Learning Analysis**

I chose to use KNN Classsification for the machine learning portion of my project. I chose thise method as it does not require complex tuning of parameters and allowed me to set a variable for my line so that I can alter the line based on what is projected by Vegas for the game. In this process I scaled my variables so that no one measurement would have more of an impact on my analysis than the other. Then I divided my data set into training and testing data. I then ran these models for all values of k 1-50 which returned an accuracy of .6 with a k value of 9. Additioanlly, I ran K means cross validation on the training and enitre data set and returned and accuracy of 57% with the avergae value of k between 7 and 9.

**Conclusions**

In compelting this project I ran into many pitfalls. The first was my decision to use the closest weather station to Coors Field as oppose to using weather data from DIA which would have given me a far large set of data which definitely would have improved my analysis. Additioanlly, issues I ran into was not converting UTC (Universal Time Coordinate) to Mountain Time, this probably costr me a whole week because I could not figure out why my night games were consistently hotter than my day games. Another problem I encountered initially was years when the Rockies played in the play in wild card game through off my web scraper and it took me a while to figure how to set up my cleaning techniques to factor this in while cleaning the data. Another aspect fo the analysis that I may have to look into is scaling my wind direction data as the 360 degree scale does not seem to lend itself well for analysis. Finally, the results of my model were not super conclusive as it is a binary classification problem with an accuracy of 57-60%, I would like a higher accuracy, but I guess that is why they keep building skyscrapers in the desert.

**Future Analysis**

Though my project did not bring with it the conclusive results I had hoped for, I did learn A TON! I will be building upon this experience for future projects especially the techniques I used to acquire my data. For my next project, I will try to remind myself to think BIG and not get wrapped up trying to find the small solve all solutions.