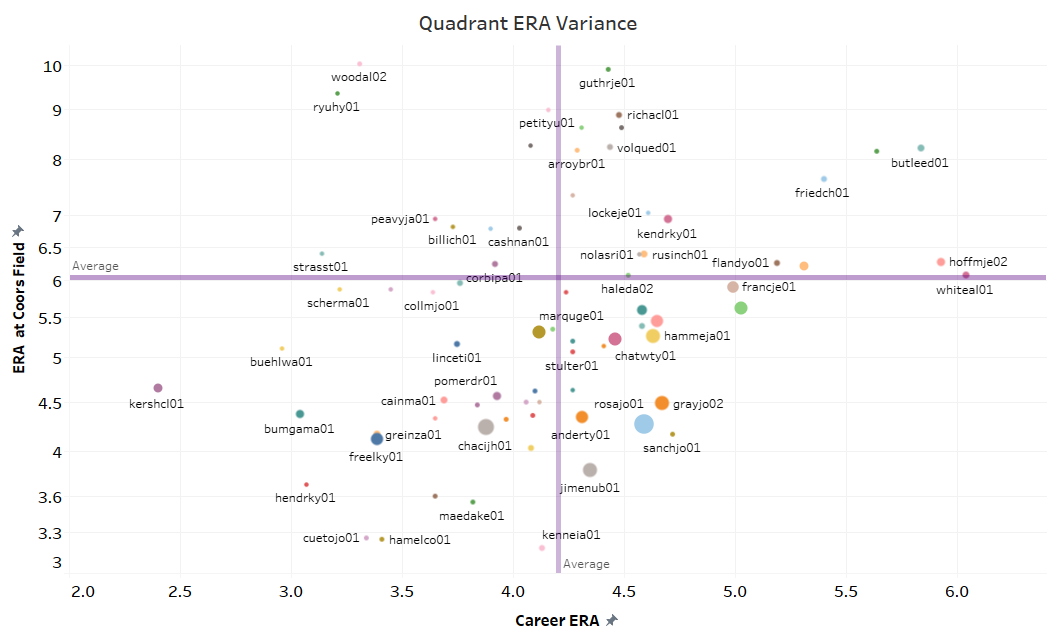
**Do you like making money?**

**What if I told you there was a way to dig further into traditional statistics to forecast over under totals?**

**Picking Pitchers**

**Not all pitchers are alike, some throw hard, some throw soft, some are accurate, some have just plain nasty stuff, some give up few runs, while others give up many runs. Success for career generally translates to equaling levels of success across ballparks, but as any pitcher will tell you, Coors Field is a different beast.**

Pitchers better than average for career yet struggle at Coors

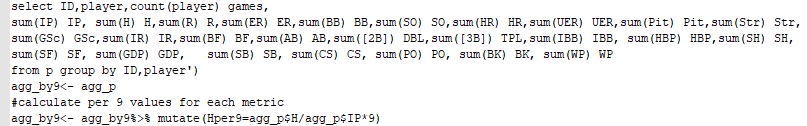


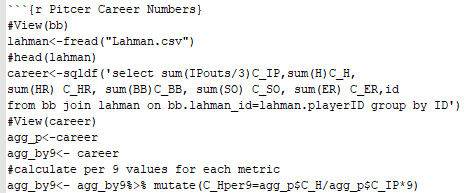
Pitchers better than average at Coors yet struggle during their careers

**Why is this the case?**

**How can we determine this?**

1. **Get Data**
2. **Explore Data**
3. **Analyze Data**
4. **Determine Results**
5. **Extract, Transform, and Load Data**
   1. Extract**:** Data comes from 4 sources
      1. Game by Game Data
         1. BaseballReference.com
            1. Read html tables
      2. Pitcher Career Stats Data
         1. BaseballReference.com
            1. Read html tables
      3. Pitcher Career Outcomes Data
         1. Brooksbaseball.net
            1. Read html tables
      4. Pitcher Career Pitch Description Data
         1. Brooksbaseball.net
            1. Read html tables
   2. Transform
      1. Game by Game Data
         1. Columns are renamed and values aggregated to per 9 rates

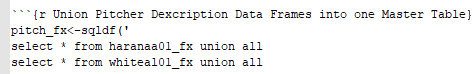


* + 1. Career Stats
       1. IDs are cross walked for later joins and values aggregated to per 9 rates 
    2. Outcomes Data
       1. Data is read into tables
       2. Pitcher name added to each data framed and then unioned

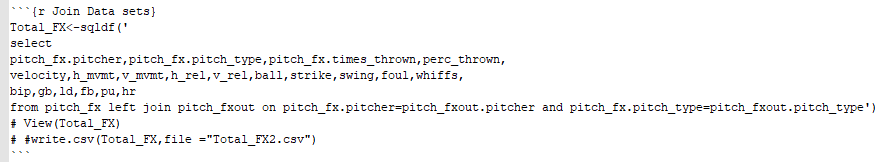




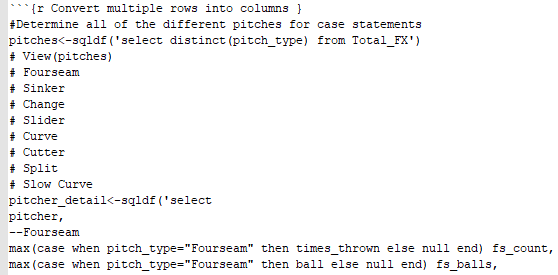
* + 1. Descritpion Data
       1. Data is read into tables
       2. Pitcher name added to each data framed and then unioned 

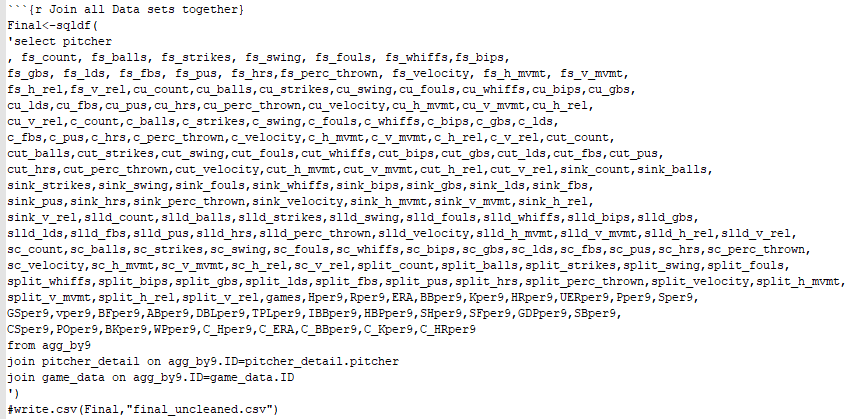
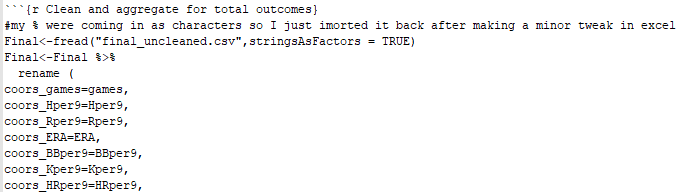


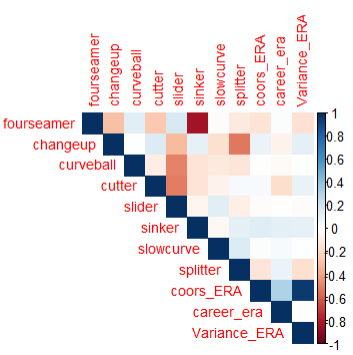
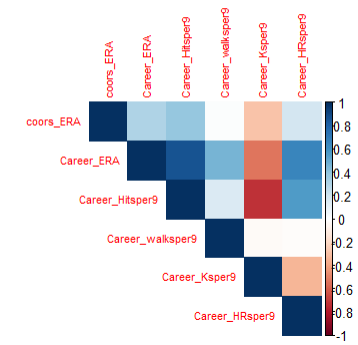
* + 1. Outcomes and Description tables Joined and Manipulated
       1. Joined

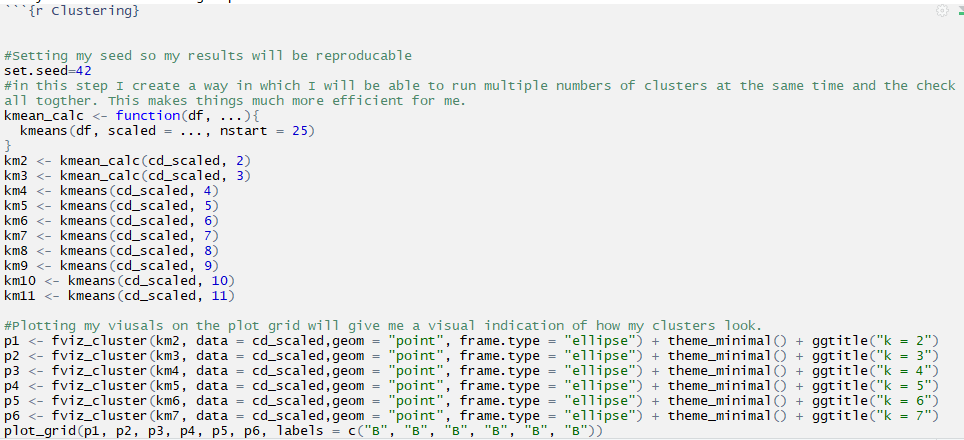


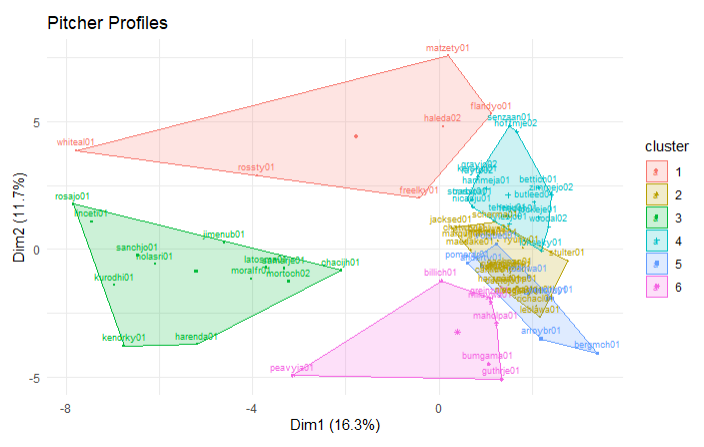
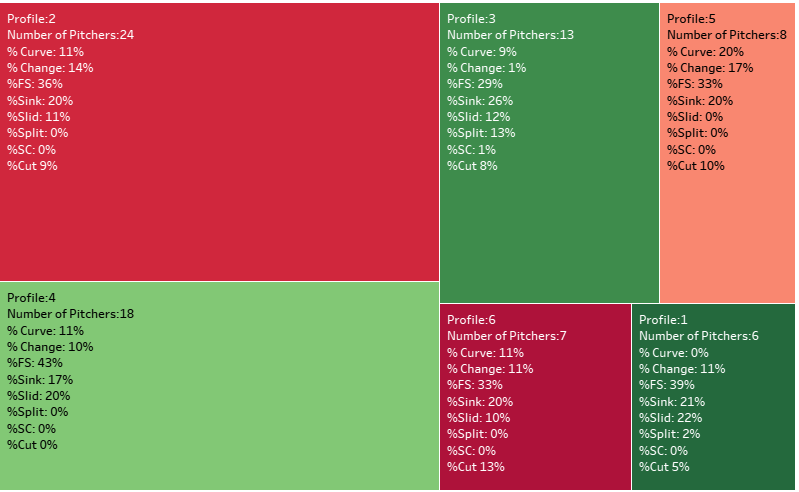
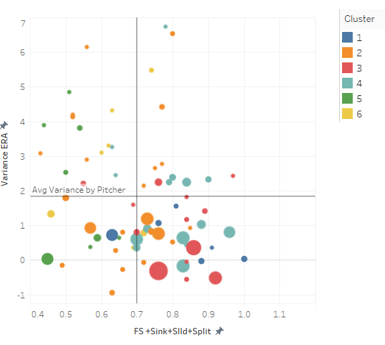
* + - 1. Manipulated to create pitcher as primary key
         1. Done for each pitch

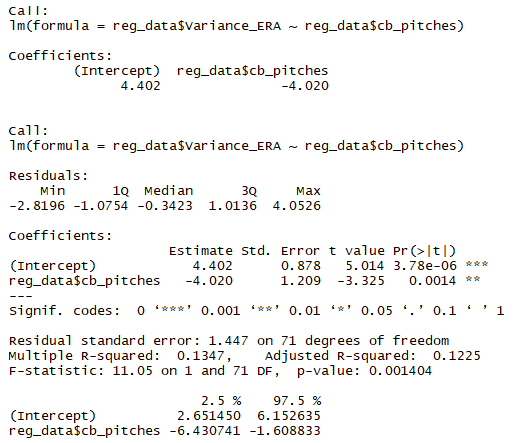


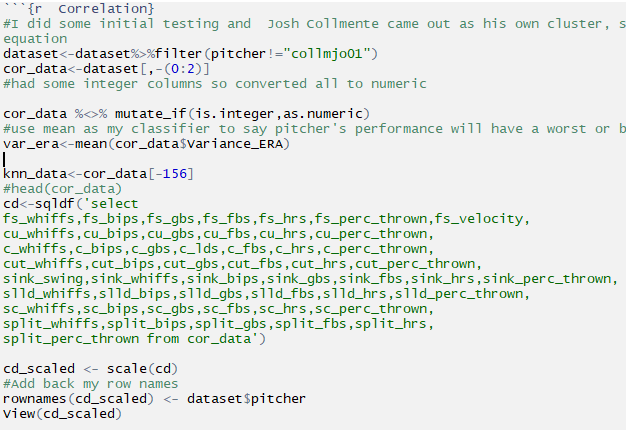
* + 1. Join all Dataset and clean up column names
       1. Join Data Sets 
       2. Clean up data 

1. **Explore Data**
   1. Looking at the various factors and how they correlate to Variance ERA (Coors-Career)
      1. Pitch Selection 
      2. Pitch Outcomes 
      3. Begin Clustering to try and Group Pitchers

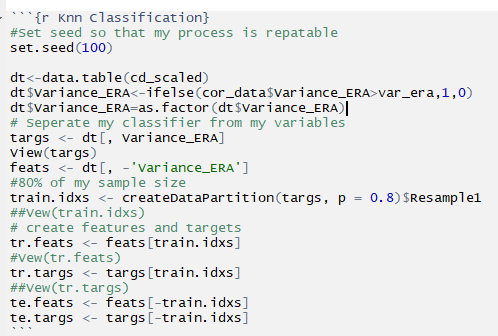


* + 1. Visualize and plot clusters 
    2. Analyze Clusters 
    3. Review Results
       1. Pitchers with a higher percentage of Fourseamer, Slider, Splitter, and Sinker generally have better than average ERAs at Coors
       2. Graph Findings 
       3. Determined many Rockies pitchers fit this profile and this can be seen by the size of the bubbles which indicate games pitched
       4. Guided my analysis to explore the relationship between pitch selection and ERA

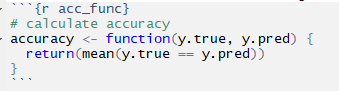
1. **Analyze Data** 
   1. Linear Regression 
      1. Review
         1. Low p-Value indicates that there is a relationship between the combined metric of four seamers, splitters, sinkers, and splitters
         2. R-squared values of 13.5% tells me my regreession only explains 13.47% of the variation on my data
         3. Not ideal, but it is clear that there is a relationship just not necessarily a linear one
   2. Knn Classicifcation
      1. Prep Data



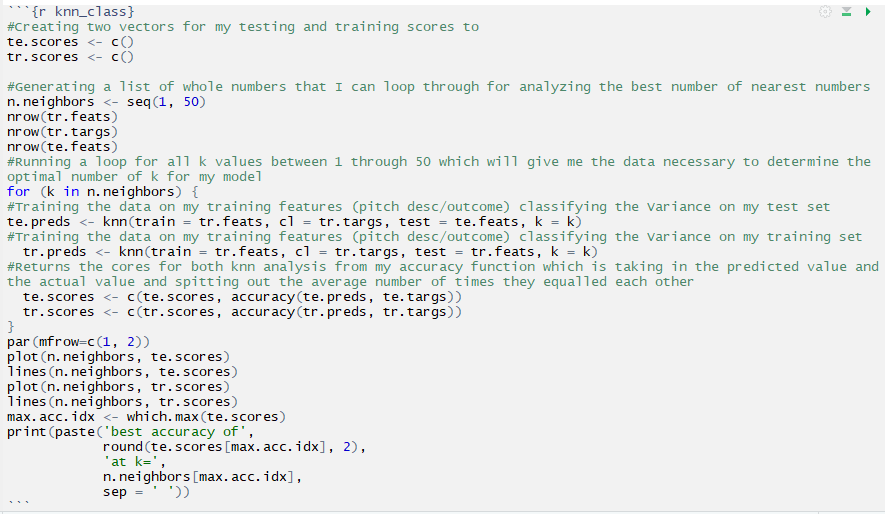
* + 1. Split Data into Training and Testing Data

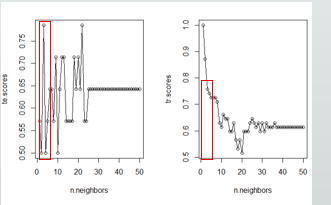


* + 1. Create mechanism to test

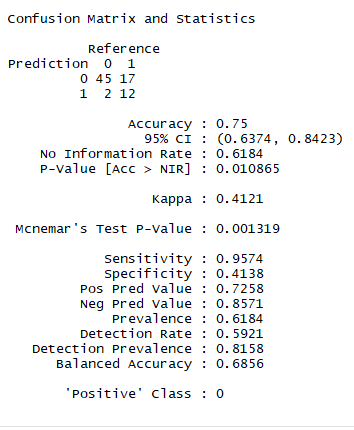


* + 1. Test for Optimal Values of K

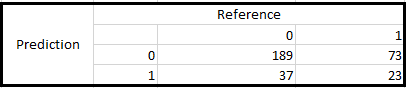




* + 1. Optimal Value of k is 3
    2. Test Accuracy on entire data set



* + 1. Test Data set on any pitcher with at least one start at Coors



1. **Determine Results**

For pitchers with at least 5 games pitched at Coors which I used to build my model, my classifier was 75% accurate with a tendency to under predict the variance, when I applied this to any pitcher with at least 1 game pitched at Coors my accuracy dropped to 66%. Moving forward I will build a model based on all pitchers and not limit it to a minimum number of starts as I believe my data became skewed towards Rockies pitchers. Luckily the infrastructure of data engineering is completed so further analysis will be pretty straight forward.