Final Report

* Who?   
  + Now we will begin by introducing ourselves. Hi, my name is Camilo Riviere and I’m part of the Big Data Analytics program. I’m in this program to enhance my current skill set in Data Science. In addition to enhancing my current skill set I’m also improving my job-related skills and improving my chances of moving up the technical ladder at my workplace, Next Era Energy (Florida Power & Light).
  + Hi, my name is Armando Zapata and I’m also part of the Big Data Analytics program. I’m in this program to solidify my foundation on my skill-sets in Data Science. I’m looking forward to applying these skills in the investment banking industry that I’m currently working in. (Cross Keys Capital).
* What? What is our project?
  + Our project is a prescriptive analysis containing 2 models (Total score, and difference) of the NCAA (National Collegiate Athletic Association) Division I (NCAA Division I is the highest level of intercollegiate athletics sanctioned by the National Collegiate Athletic Association in the United States.) Men’s College Basketball. In this project we are predicting three things: the outcome (inferred from the score of each individual team), the overall score in the matchup (predicted by the total model), the score of each individual team in the matchup which is calculated by a difference model, and the implied probability of home team winning the game.
* Where? Where are some examples for practice use cases of our model?
  + Prescriptive models regarding predicting outcomes of a specific sporting event are in high demand in geographic locations such as Las Vegas where sports betting is a billion-dollar business (annually). Veteran Sports bettors and sports betting companies alike are actively looking for any kind of competitive edge and are willing to pay top dollar for highly accurate prediction models.
  + Outside of the sports industry. Large retail corporations such as Target or Wal-Mart may eventually garner interest in pre-determining whether to manufacture merchandise after any sports team wins the championship for their respective sport. Often, these retailers will have both sets of championship clothes manufactured for both the winning & losing team which inevitably creates waste because the clothes for the losing team is often discarded or sent overseas in donation to foreign countries such as Africa where there is a need for clothing.
* When? When can we use this model?
  + Because our stats are exclusively for Men’s Division I College Basketball our model is best used during both the regular seasons for College Basketball and the post season, or March Madness. Users can select the matchup for the day and proceed to make a prediction on the teams which are playing that day in the tournament.
* Why? Why was this project important to us?
  + Given that we are both into sports, and March Madness is right around the corner, we thought it would be a great idea to immerse ourselves into this environment and apply some of the skills we’ve learned thus far in the Big Data Analytics program.
  + This is an event that is perceived as a phenomenon, which galvanizes people from different races, religion, beliefs, and spans across the world. To put it in to context, this event is so prominent that even the former President Barack Obama on National Television (ESPN) filled out a tournament bracket.
  + Additionally, the Chairman and CEO of Berkshire Hathaway Warren Buffet, offered $1 million in perpetuity to anyone who predicted a perfect Sweet 16 (Last 16 teams left in the tournament).
  + Also, given the complexity and unpredictability of accurately predicting this event, even Kaggle and Google host annual competitions for those who attempt to create models that accurately predict the outcome all the way to the champion of the entire tournament.
  + Note – Include picture of tournament bracket.
* Why? Why is this project important to the class?
  + In this project we introduced the concept of productionizing and hosting your model as an endpoint/user application. This is a concept that has yet to be covered in our program. By introducing this concept, we intend to see students in the future will see the value of this and apply it to their future projects as this will eventually be a client requirement in a real-life job setting.
  + We are also going to introduce the use of advanced statistics in the form of efficiency statistics which these types of statistics are proven to be more effective when performing an analysis in contrast to traditional box score statistics which are typically utilized by sports television analyst (FG%, 3PFG%, Blocks, Assists, etc.). By introducing this concept, we intend to see students in the future utilize higher quality data which provide more insight into their area of research.

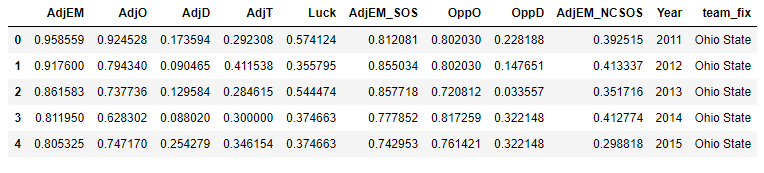
The Data –

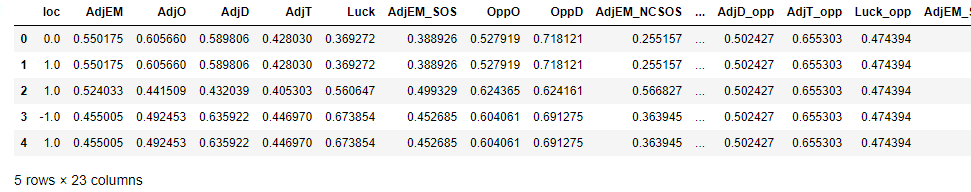
* The data in our model was pulled from kenpom.com, His (Ken Pomeroy’s) website includes his College Basketball Ratings, tempo-free statistics for every NCAA men's Division I basketball team, with archives dating back to the 2002 season, as well as a blog about current college basketball news. His work on tempo-based basketball statistics is compared by many to the work of Bill James in baseball.
* Our data is comprised of 12 independent variables (features) and 3 dependent variables (results):
  + **Adjusted Offensive Efficiency** - relative to number of possessions each team has – this takes efficiency of scoring into play, rather than just looking at points per game, since teams play at different speeds, and get different numbers of opportunities per game
  + **Adjusted Defensive Efficiency** - defensive efficiency relative to number of possessions each team has – same explanation as offense.
  + **Adjusted Tempo** - measures the speed at which teams play – purpose of including in the model is to see if faster or slower teams have an advantage.
  + **SOS Difference** - Strength of schedule measures the quality of competition for each team (KenPom.com, 2).
  + **Location** - Whether the winning team was playing at home, away or on a neutral court.
  + **Offensive Rebound Differential** – Average offensive rebound differential per game for a team, calculated the same way as average point differential.
  + **Defensive Rebound Differential** - Average defensive rebound differential per game for a team, calculated the same way as others in this section.
  + **Turnover Differential -** Average turnover differential per game for a team, calculated the same way as other in this section.
  + **Three Point Reliance -** Percentage of points scored from three pointers for a team.
  + **True Shooting Percentage -** True shooting percentage for a team. True shooting percentage is a more accurate version of shooting percentage, as it weights shots based on their value (ex. Three-pointers are more valuable).
  + **True Shooting Percentage Allowed -** True shooting percentage allowed for a team. Same as above, but what the team allows rather than shoots.
  + **Block Differential -** Average difference in blocks per game for team vs. opponents.

Data Processing –

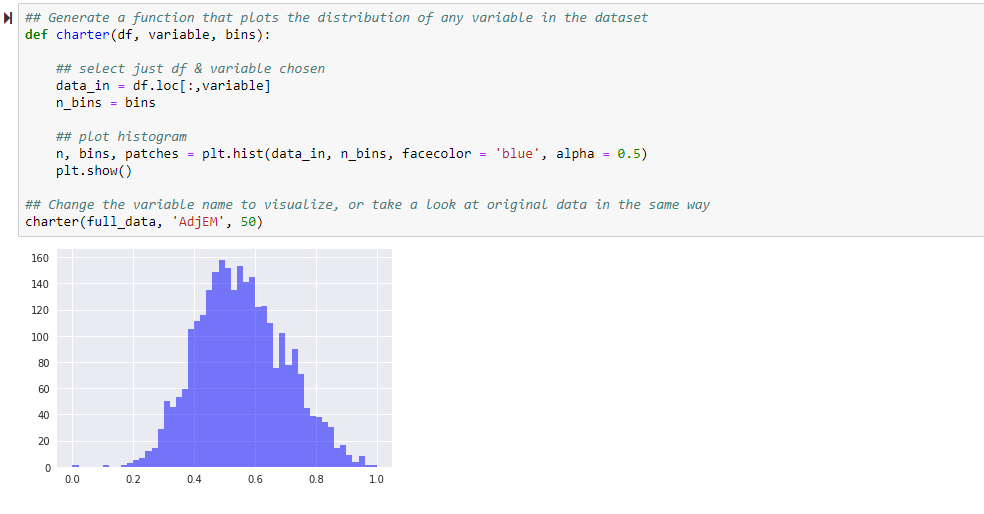
* One of the advantages we had in this project was that our data was already cleaned for us because we were able to find datasets on the web that did not require transformation. Best of all, the data we found was of high quality.

We have 2 distinct datasets. Kenpom & Historical game results dating back to 2002. Prior to merging both datasets we rescaled the Kenpom data from 0-1:



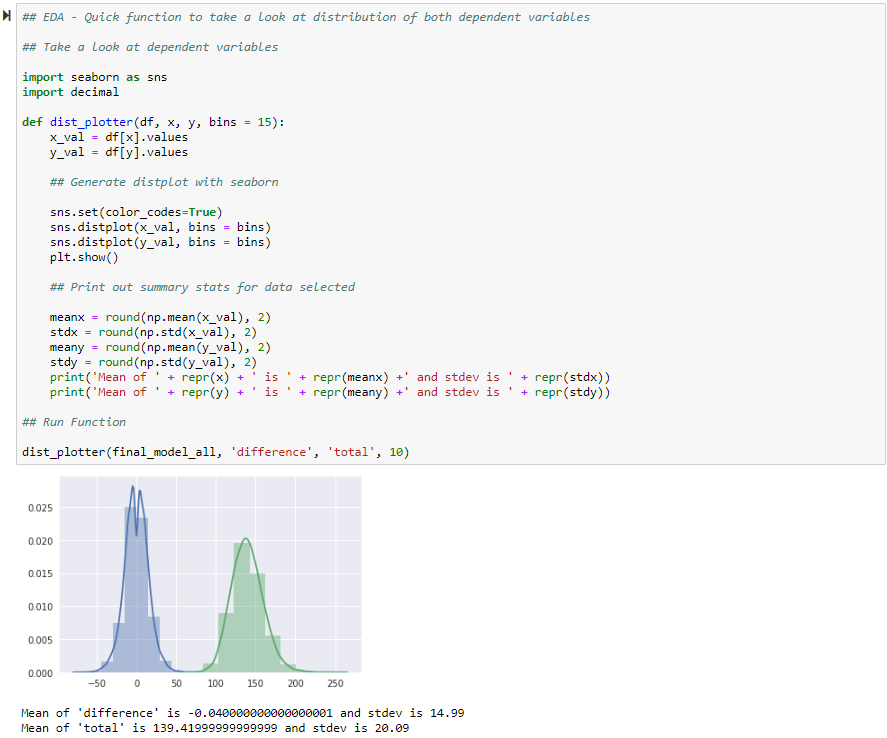
After rescaling our Kenpom data we proceeded to merge both datasets (Kenpom, and historical). You can see a snippet below of some of the added data. The table was too long to fully capture in this PowerPoint:  
  


Moving forward we wanted to see the distribution of each variable in our dataset. Below you can see the code we used to be able to visualize the distributions. In this graph we are looking specifically at Adjusted Efficiency Margin.



Exploratory Data Analysis –

The function beneath outlines two models against one another and is especially valuable for comparing the two target factors (total and difference). In view of the structure of the dataset, since a lion's share of the watched information contains an away and home group, it has sense that the normal effect is well over 0, as home groups are bound to beat away groups in a vacuum. As demonstrated in the remarks, the nonpartisan court pointer isn't totally dependable, as it is just catching March Madness diversions, and no other unbiased court events that happen over the span of the period.



Model Performance Evaluation –



