Michael Griffin

Professor Sale

Data Analytics Bootcamp

January 19, 2018

**Data description:**

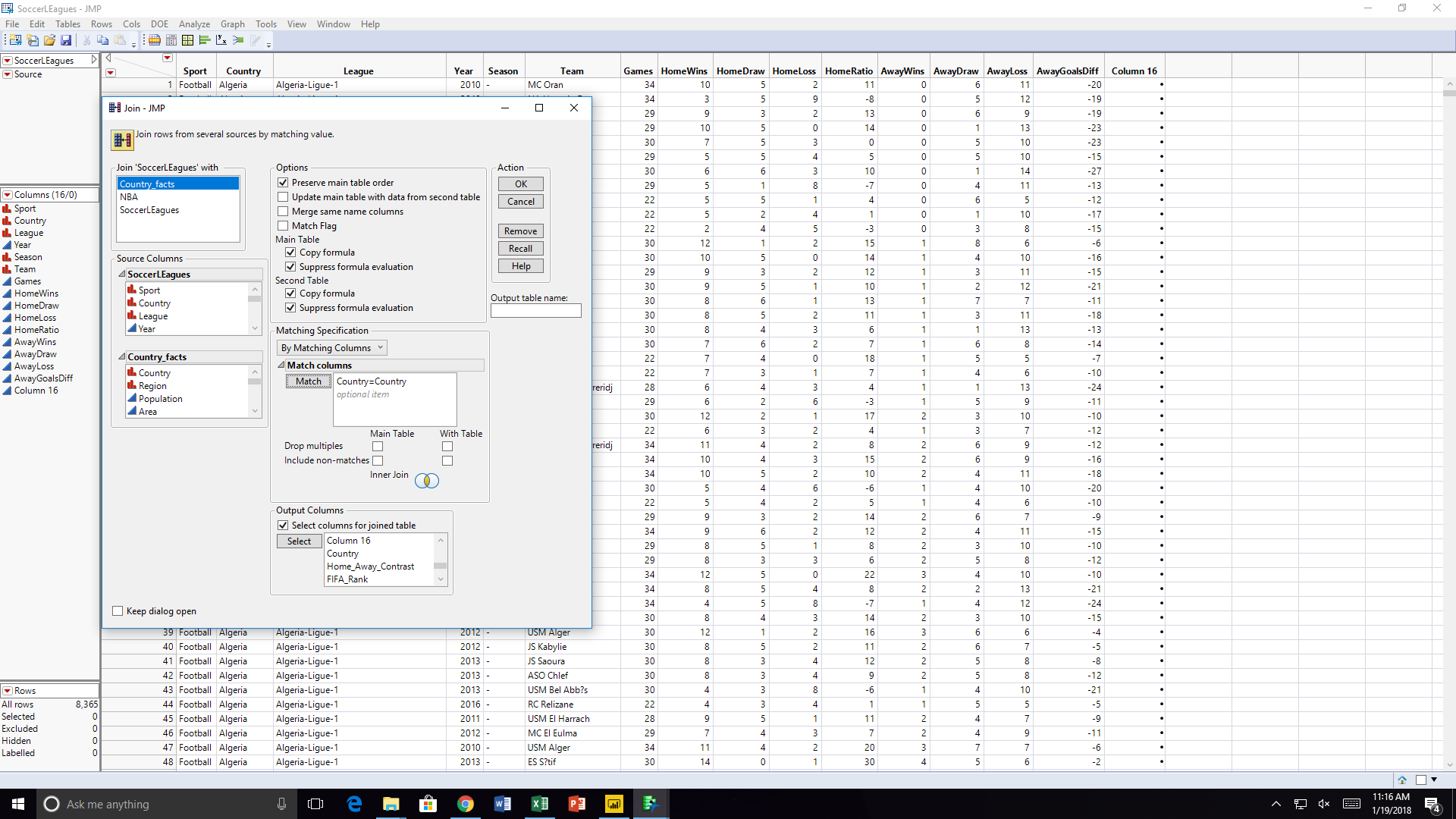
The dataset details eight years’ worth of worldwide soccer team’s basic data. An additional table included details regarding countries. There are 88 countries included in the data set and various teams from each country. Included variables from that table were climate, birthrate, deathrate, and arable (a rating). From the birth and death rate, I created a differential in an attempt to make a very basic figure to determine the countries’ overall health. Some of the metrics from the dataset are quantified ratings, such as “climate” and “arable” which are two of the variables that I will use in my multiple linear regression.

**Research Questions:**

1. Does a better climate result in a better home field differential?
2. Does a better birth/death rate (simplified version of an overall health metric) result in a better home field differential?
   1. I intend on using multiple regression to answer both of these questions.
3. Is there a natural grouping of the countries and teams in relation to rank and climate?

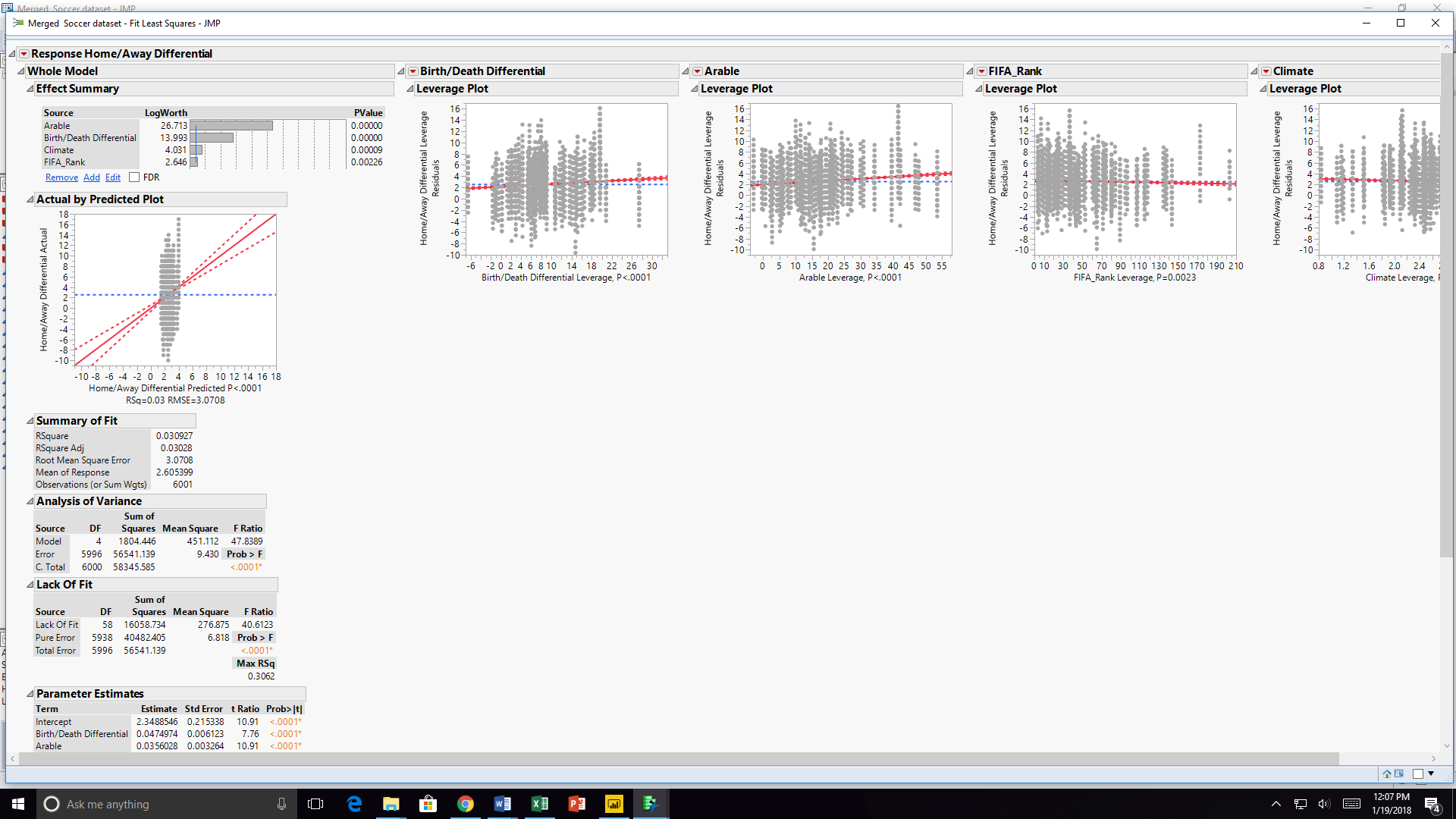
**Data preparations:**

I merged soccer data set to country facts on the “country” column. I then proceeded to make two calculated fields called home/away differential that calculates total home wins minus total away wins for each record. The data preparation is seen below.

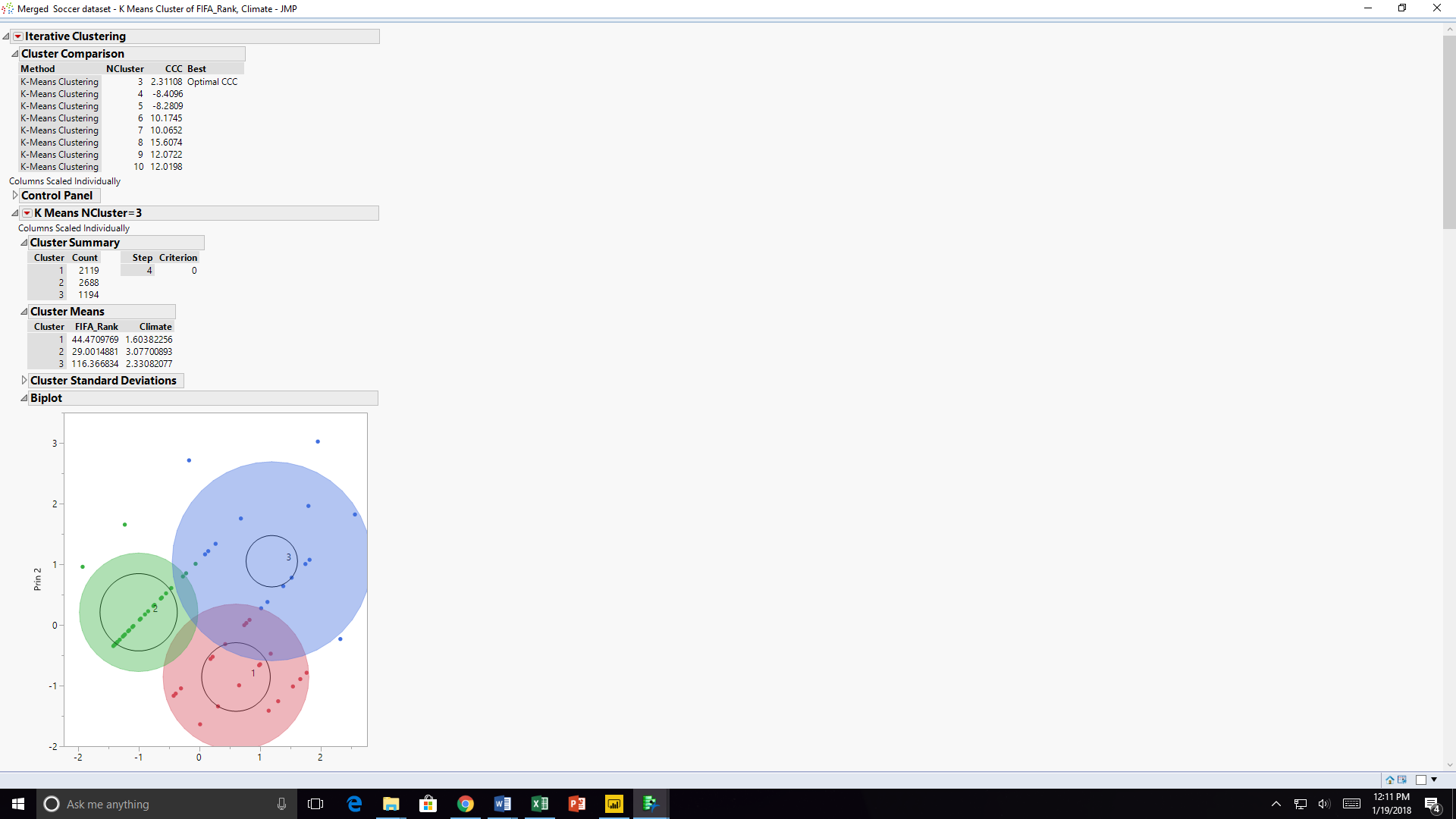


**Analysis:**

According to the multiple linear regression, there is almost no correlation between “arable”, “birth/death differential” “climate” and “FIFA\_Rank” with how strong a given team’s home field advantage is. Interestingly, the data still passes the f test, meaning that the regression is a better indicator than the mean is. However, I would say that the data is completely insignificant in terms of predicting how strong a team’s home-field advantage is. While in power BI, an interesting fact I noticed was that more developed countries tended to have lower birth rates. I found that to be counterintuitive and very interesting.



In this exercise, I clustered the data based upon kmeans by way of FIFA\_Rank and climate. The graphic and results are below. The data points have been superimposed onto the clusters, and show that the data does not cluster very well overall. It appears that there aren’t strong groupings of data that relate team rank and climate. If there was a relationship, then that could have provided further on why home-field advantage may have little correlation with our variables.



**Research Answers:**

1. There is no significant relationship between an area’s climate and the strength of its home-field advantage.
2. There is no significant relationship between an area’s birth/death differential and the strength of its home-field advantage.
3. FIFA\_Rank and climate do not cluster well. There appears to be very little relationship between the two.

**Power BI:**

Below, with the use of Power BI, I have created 4 visuals to futher understand the data. The visuals are all scatter plots, and show the following relationships: birthrate against deathrate by country, birthrate against home/away differential by country, and home/away contrast against FIFA rank. There are also two cards in the bottom left corner detailing the first country selected and the total number of countries selected.

