**College Football Games 2000-2018**

**Exploratory Analysis**

Clayton Catanzarite, [ccatanzarite@bellarmine.edu](mailto:ccatanzarite@bellarmine.edu)

1. **INTRODUCTION**

This data set is a list of college football games between the year 2000 and 2018. This dataset includes data including the date of the game, which teams are playing including home and away, each teams record, multiple columns to include weather data, etc.

[**https://www.kaggle.com/jeffgallini/college-football-attendance-2000-to-2018**](https://www.kaggle.com/jeffgallini/college-football-attendance-2000-to-2018)

1. **DATA SET DESCRIPTION**

This data set contains 6672 samples with 25 columns. A complete listing is shown in **Table 1.** There are multiple different data types in this set. While this set has a large amount of discrete data, it also has its fair share of continuous data as well. Most of the weather columns are continuous and there is no missing data in this set.

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Date | Interval/object | 0% |
| Team | Nominal/object | 0% |
| Time | Ratio/object | 0% |
| Opponent | Nominal/object | 0% |
| Rank | Ordinal/object | 0% |
| Site | Nominal/object | 0% |
| TV | Nominal/object | 0% |
| Result | Ratio/object | 0% |
| Attendance | Ratio/int64 | 0% |
| CurrentWins | Ratio/int64 | 0% |
| Current Loses | Ratio/int64 | 0% |
| Stadium Capacity | Ratio/int64 | 0% |
| Fill Rate | Ratio/float64 | 0% |
| New Coach | Nominal/bool | 0% |
| Tailgating | Nominal/bool | 0% |
| PRCP | Ratio/float64 | 0% |
| SNOW | Ratio/float64 | 0% |
| SNWD | Ratio/float64 | 0% |
| TMAX | Interval/int64 | 0% |
| TMIN | Interval/int64 | 0% |
| Opponent\_Rank | Ordinal/object | 0% |
| Conference | Nominal/object | 0% |
| Year | Interval/int64 | 0% |
| Month | Interval/int64 | 0% |
| Day | Interval/int64 | 0% |

1. **Data Set Summary Statistics**

Here is all of the statistics of the continuous variables from my data set. In **Table 2,** you can see the statistics for the variables. The most surprising post of these summary statistics is the relationship of the attendance and the stadium capacity. The max attendance to a game was 110,889 people while the maximum capacity was 107,282. This means some of the games had more people then what the capacity was which I found interesting.

**Table 2: Summary Statistics for College Football Games (2000 to 2018)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| *Attendance* | *6672* | *45311.5* | *25185.7* | *2267* | *23301* | *42527.5* | *62358* | *110889* |
| *Current Wins* | *6672* | *2.87* | *2.42* | *0* | *1* | *2* | *4* | *12* |
| *Current Loses* | *6672* | *2.34* | *2.23* | *0* | *0* | *2* | *4* | *11* |
| *Stadium Capacity* | *6672* | *54567.47* | *21755.52* | *17000* | *36000* | *52180* | *71799* | *107282* |
| *Fill Rate* | *6672* | *0.79* | *0.22* | *0.06* | *0.64* | *0.84* | *0.99* | *1.4* |
| *PRCP* | *6672* | *0.09* | *0.33* | *0* | *0* | *0* | *0.01* | *6.45* |
| *SNOW* | *6672* | *0.009* | *0.19* | *0.0* | *0.0* | *0.0* | *0.0* | *8.2* |
| *SNWD* | *6672* | *0.015* | *0.255* | *0.0* | *0.0* | *0.0* | *0.0* | *7.2* |
| *TMAX* | *6672* | *71.92* | *14.45* | *19* | *62* | *73* | *83* | *111* |
| *TMIN* | *6672* | *50.14* | *14.17* | *0* | *40* | *51* | *61* | *103* |

Table 4: Proportions for Team\* (n=yyy)

*\*Top 5 most frequent, there is 63 different Teams*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Hawaii* | *132* | *1.97%* |
| *Nebraska* | *130* | *1.94%* |
| *Penn State* | *130* | *1.94%* |
| *Arkansas* | *129* | *1.93%* |
| *Michigan State* | *129* | *1.93%* |

Table 5: Proportions for Opponent\* (n=yyy)

*\*Top 5 most frequent, there is 1365 different opponents*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Iowa State* | *48* | *.710%* |
| *Kansas* | *47* | *.704%* |
| *Akron* | *46* | *.689%* |
| *Colorado* | *45* | *.674%* |
| *Purdue* | *45* | *.674%* |

Table 6: Proportions for Rank\* (n=yyy)

*\*Top 5 most frequent, there is 26 different rankings*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *NR (not ranked)* | *5213* | *78.13%* |
| *24* | *77* | *1.15%* |
| *19* | *73* | *1.09%* |
| *1* | *73* | *1.09%* |
| *17* | *65* | *.974%* |

Table 7: Proportions for Site\* (n=yyy)

*\*Top 5 most frequent, there is 446 different site of the game*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Aloha Stadium Honolulu,HI* | *114* | *1.71%* |
| *Rose Bowl Pasadena, CA* | *105* | *1.57%* |
| *Camp Randall Stadium Madison, WI* | *103* | *1.54%* |
| *Beaver Stadium University Park, PA* | *102* | *1.52%* |
| *Jack Trice Stadium Ames, IA* | *114* | *1.50%* |

Table 3: Proportions for TV\* (n=yyy)

*\*Top 5 most frequent, there is 245 different tv channels the game was on*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Not on TV* | *1317* | *19.73%* |
| *ESPN* | *508* | *7.61%* |
| *ESPN2* | *460* | *6.89%* |
| *ABC* | *453* | *6.79%* |
| *FSN* | *403* | *6.04%* |

Table 8: Proportions for New Coach (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *True* | *1093* | *16.38* |
| *False* | *5579* | *83.61* |

Table 9: Proportions for Tailgating (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *True* | *1431* | *21.44%* |
| *Fales* | *5241* | *78.55%* |

Table 10: Proportions for Opponent Rank (n=yyy)

*\*Top 5 most frequent, there is 26 different rankings for the opponent*

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *NR (not ranked)* | *5468* | *81.95%* |
| *1* | *79* | *1.18%* |
| *2* | *68* | *1.01%* |
| *21* | *57* | *0.85%* |
| *10* | *52* | *0.77%* |

Table 12: Proportions for Conference (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *ACC* | *957* | *14.34%* |
| *Big-12* | *903* | *13.53%* |
| *Big-10* | *782* | *11.72%* |
| *Mid-American* | *711* | *10.65%* |
| *CUSA* | *537* | *8.04%* |
| *Pac-12* | *508* | *7.61%* |
| *SEC* | *434* | *6.50%* |
| *MWC* | *414* | *6.20%* |
| *WAC* | *396* | *5.94%* |
| *Sun Belt* | *392* | *5.87%* |
| *AAC* | *273* | *4.09%* |
| *Independent* | *193* | *2.89%* |
| *FCS* | *91* | *1.36%* |
| *Big East* | *81* | *1.21%* |

Table 13: Proportions for Year (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *2008* | *378* | *5.67%* |
| *2011* | *373* | *5.59%* |
| *2009* | *371* | *5.56%* |
| *2016* | *370* | *5.55%* |
| *2010* | *370* | *5.55%* |
| *2007* | *366* | *5.49%* |
| *2015* | *366* | *5.49%* |
| *2012* | *366* | *5.49%* |
| *2014* | *363* | *5.44%* |
| *2013* | *363* | *5.44%* |
| *2006* | *359* | *5.38%* |
| *2018* | *355* | *5.32%* |
| *2017* | *354* | *5.31%* |
| *2002* | *352* | *5.28%* |
| *2003* | *351* | *5.26%* |
| *2005* | *317* | *4.75%* |
| *2004* | *309* | *4.63%* |
| *2001* | *297* | *4.45%* |
| *2000* | *292* | *4.38%* |

Table 14: Proportions for Month (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *9* | *2316* | *34.71%* |
| *10* | *2080* | *31.18%* |
| *11* | *1941* | *29.09%* |
| *8* | *220* | *3.30%* |
| *12* | *110* | *1.65%* |
| *4* | *4* | *0.06%* |
| *1* | *1* | *0.01%* |

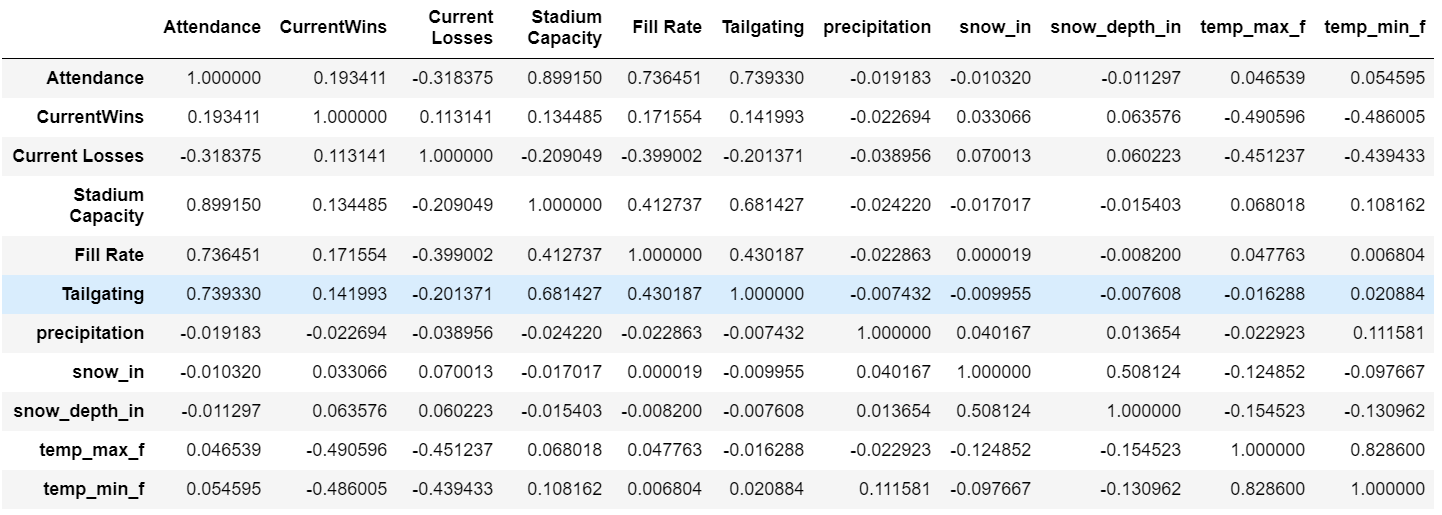
Table 15: Proportions for Day (n=yyy)

*\*Top 5 most frequent, there is 31 different days of the month*

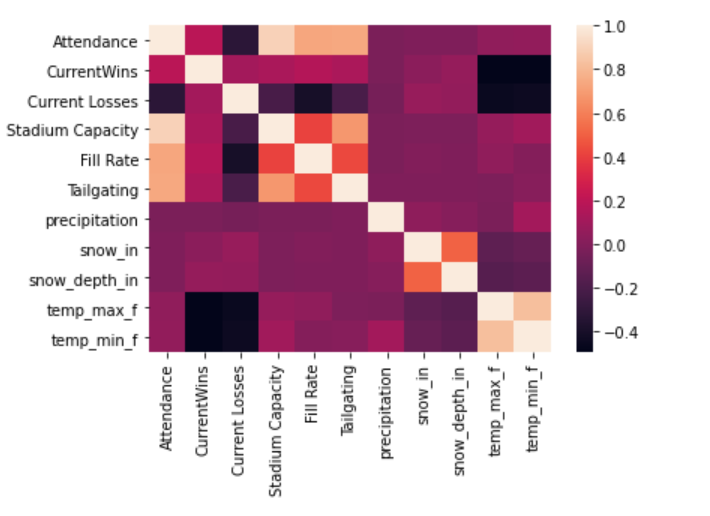
|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *1* | *287* | *4.30%* |
| *6* | *273* | *4.09%* |
| *30* | *253* | *3.79%* |
| *3* | *251* | *3.76%* |
| *8* | *251* | *3.76%* |

Table 16: Correlation Table/Tables

My table was too large to include as an actual table in Microsoft word. I included a picture of the table from my Notebook instead. I also had to drop some columns that Pandas considered continuous but were actually categorical including Year, Month, Day, and New Coach.



Here is the heatmap of this correlation Matrix:



1. **DATA SET GRAPHICAL EXPLORATION**

In this section, I will be creating and exploring different graphs that I create from my data. This includes distribution graphs, scatterplots, pairwise plots, barcharts, and other plots as well.

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 1. *Distributions*

**Attendance Distribution**

I created a chart to show the distribution for attendance at these college football games show in **Figure 1**. I expected the attendance to be a normal distribution with a bell curve. Instead, the distribution plot is skewed slightly to the left. It also has a lack of kurtosis to it as well.

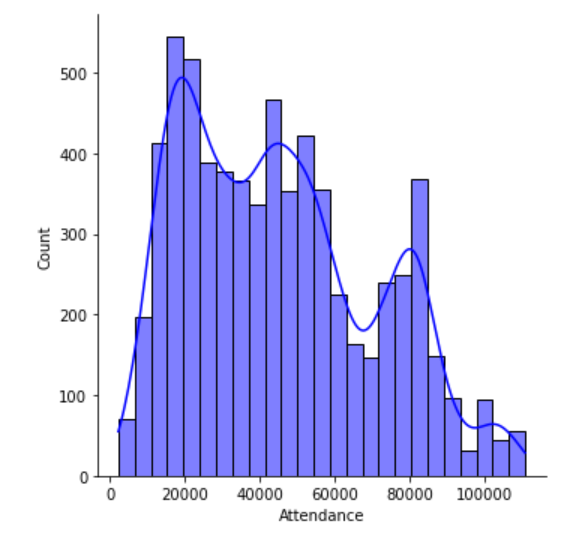


Figure ,Attendance Distribution

**Attendance of the Power 5 conferences compared**

For **Figure 2** I compared the Power 5 conferences attendances distributions. This image came out a little unclear when I exported it as a .png image. There is a clearer version in the Notebook. Each conference had a unique distribution. The SEC and the Big-10 were the only conferences with attendances over 100,000. The Pac-12 had the most normal distribution but not nearly as many fans as the SEC or Big-10.

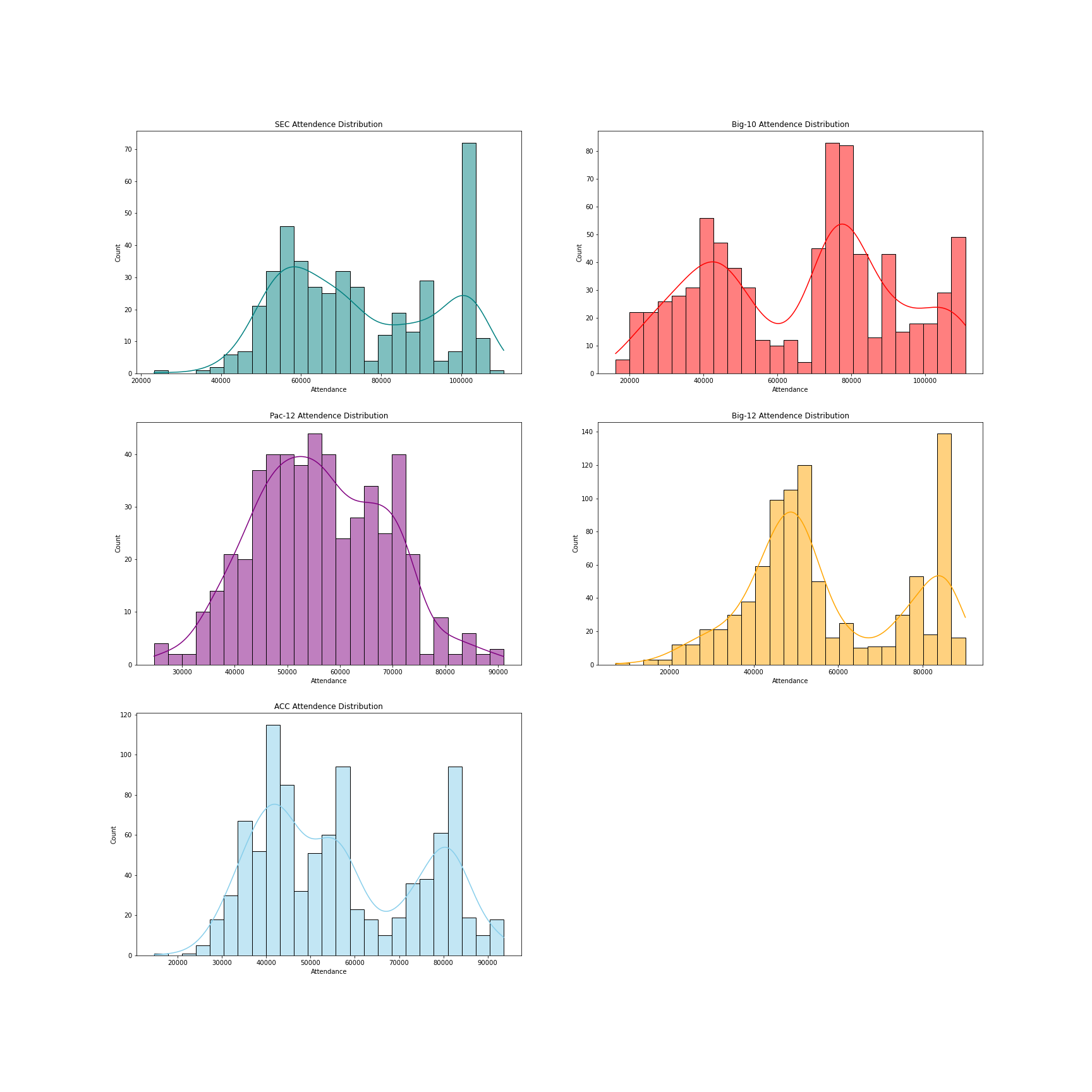
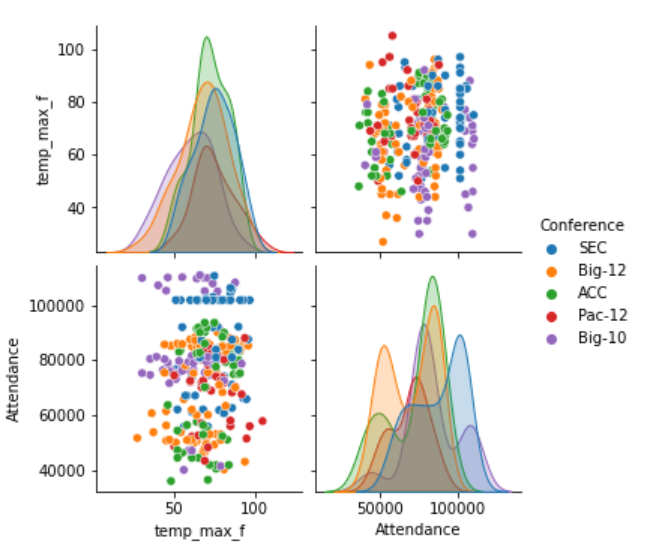


Figure , Power 5 Conferences Attendance Distribution

* 1. *ScatterPlots / Pairwise Plots*

The first Pairwise Plot that I created was finding the connection between attendance at Power 5 conference games and what the max temperature is for those games. I only included games that had two ranked opponents as well This can be found in **Figure 3.** The results show that the attendance stays consistent. There are slightly more fans when the max temperature is a little bit higher



* 1. *Barcharts (categorical variables)*
  2. *Other Plots*

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.