College Football Games 2000-2018
Exploratory Analysis

Clayton Catanzarite, ccatanzarite@bellarmine.edu

I. INTRODUCTION

This data set is a list of college football games between the year 2000 and 2018. This dataset includes data such as the date of the game, which teams are playing, their opponent, each team's record, multiple columns to include weather data, etc.

https://www.kaggle.com/jeffgallini/college-football-attendance-2000-to-2018

II. 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. It was difficult to determine if certain variables were categorical or continuous in some instances.

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%

III. Data Set Summary Statistics

Here is all 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 attending than what the capacity allowed, 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*

**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*

**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*

**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*

**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*

**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

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

Table 9: Proportions for Tailgating

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

Table 10: Proportions for Opponent Rank

**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

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

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

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

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

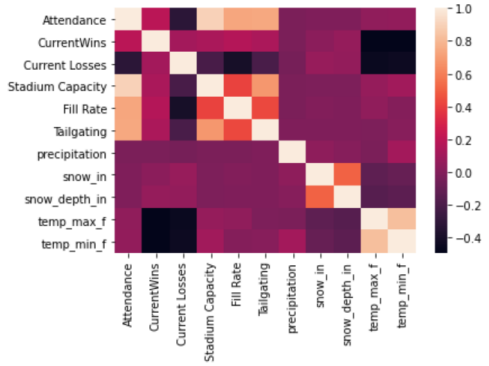
Category	*Frequency*	*Proportion (%)*
1	*287*	*4.30%*
6	*273*	*4.09%*
30	*253*	*3.79%*
3	*251*	*3.76%*
8	*251*	*3.76%*

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 categorical including Year, Month, Day, and New Coach.

	Attendance	CurrentWins	Current Losses	Stadium Capacity	Fill Rate	Tailgating	precipitation	snow_in	snow_depth_in	temp_max_f	temp_min_f
Attendance	1.000000	0.193411	-0.318375	0.899150	0.736451	0.739330	-0.019183	-0.010320	-0.011297	0.046539	0.054595
CurrentWins	0.193411	1.000000	0.113141	0.134485	0.171554	0.141993	-0.022694	0.033066	0.063576	-0.490596	-0.486005
Current Losses	-0.318375	0.113141	1.000000	-0.209049	-0.399002	-0.201371	-0.038956	0.070013	0.060223	-0.451237	-0.439433
Stadium Capacity	0.899150	0.134485	-0.209049	1.000000	0.412737	0.681427	-0.024220	-0.017017	-0.015403	0.068018	0.108162
Fill Rate	0.736451	0.171554	-0.399002	0.412737	1.000000	0.430187	-0.022863	0.000019	-0.008200	0.047763	0.006804
Tailgating	0.739330	0.141993	-0.201371	0.681427	0.430187	1.000000	-0.007432	-0.009955	-0.007608	-0.016288	0.020884
precipitation	-0.019183	-0.022694	-0.038956	-0.024220	-0.022863	-0.007432	1.000000	0.040167	0.013654	-0.022923	0.111581
snow_in	-0.010320	0.033066	0.070013	-0.017017	0.000019	-0.009955	0.040167	1.000000	0.508124	-0.124852	-0.097667
snow_depth_in	-0.011297	0.063576	0.060223	-0.015403	-0.008200	-0.007608	0.013654	0.508124	1.000000	-0.154523	-0.130962
temp_max_f	0.046539	-0.490596	-0.451237	0.068018	0.047763	-0.016288	-0.022923	-0.124852	-0.154523	1.000000	0.828600
temp_min_f	0.054595	-0.486005	-0.439433	0.108162	0.006804	0.020884	0.111581	-0.097667	-0.130962	0.828600	1.000000

Here is the heatmap of this correlation Matrix:



IV. DATA SET GRAPHICAL EXPLORATION

In this section, I will be creating and exploring different graphs from my data. This includes distribution graphs, scatterplots, pairwise plots, bar charts, and other plots as well. This also includes me breaking down the main data set into other sub datasets as well as breaking it down by conference, rank, etc.

A. 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.

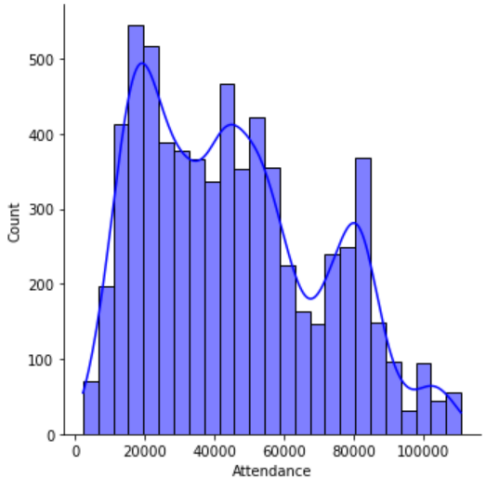


Figure 1,Attendance Distribution

Attendance of the Power 5 conferences compared

For **Figure 2** I compared the Power 5 conferences attendance 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.

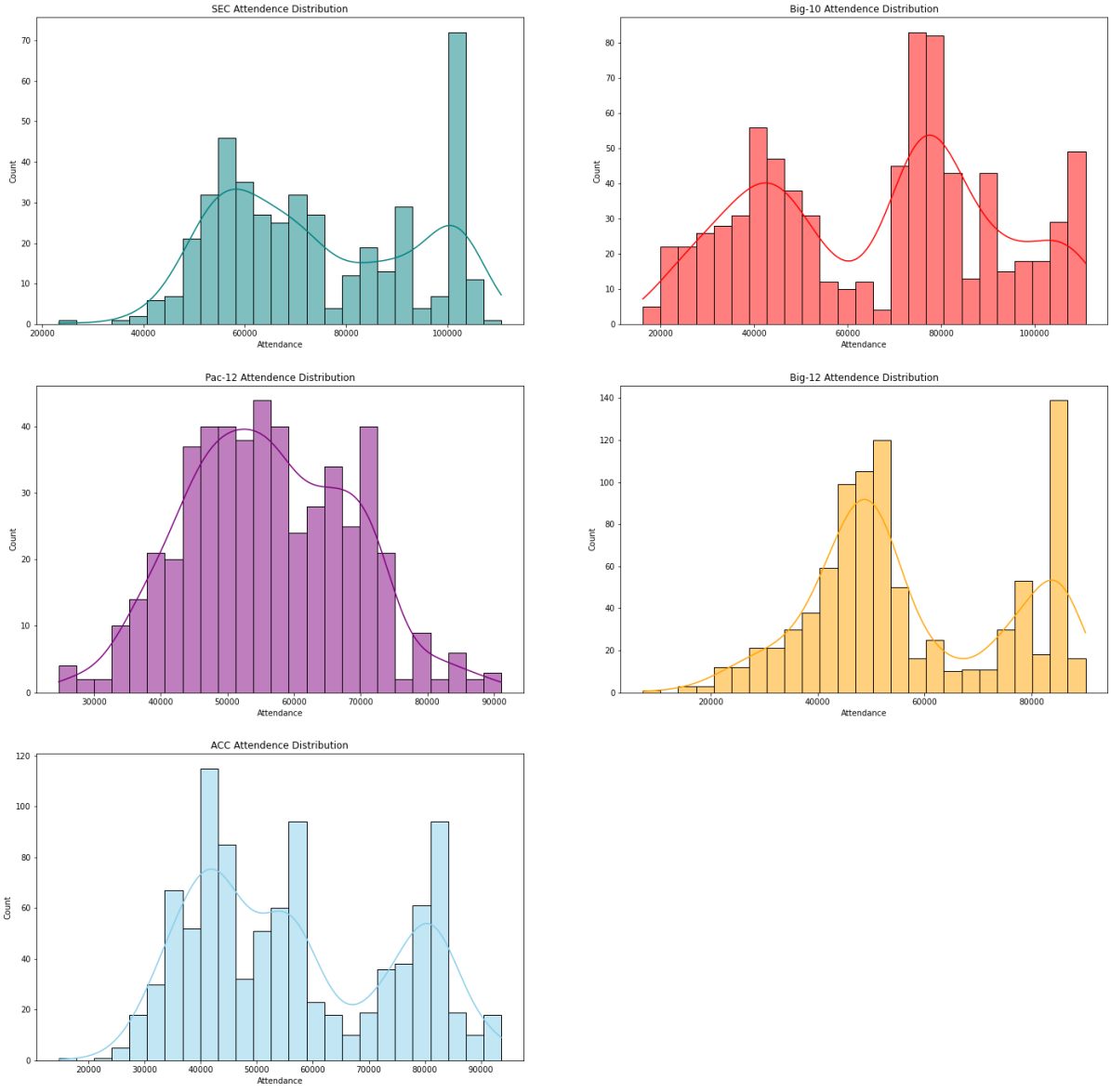


Figure 2, Power 5 Conferences Attendance Distribution

B. Scatter Plots / 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.

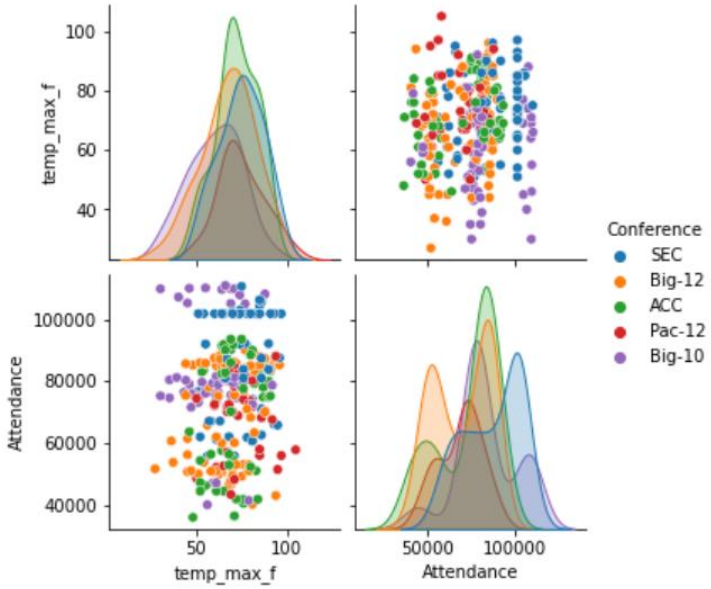


Figure 3, Ranked Power 5 games Max Temperature vs Attendance

The second Pairwise Plot that I created was using the highest correlation from the correlation matrix. This was the correlation between the stadium capacity and the attendance. I used the Ranked Power 5 games sub dataset again because the correlation was a little harder to see in the original dataset. The result of this pairwise plot was that you could see that as the stadium capacity went up, so did the attendance. This would make sense because the games can sell more tickets with a higher stadium capacity. This pairwise plot can be seen in **Figure 4.**

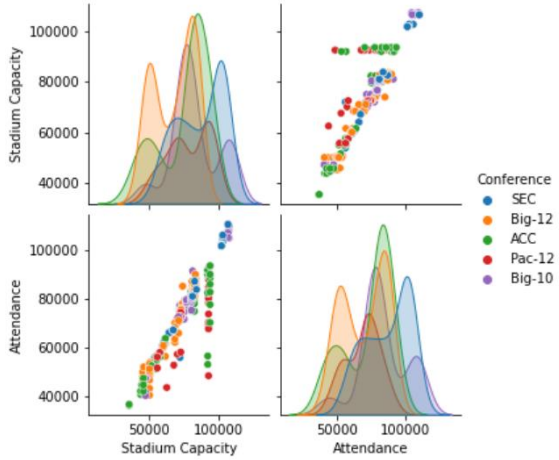


Figure 4, Ranked Power 5 games Stadium Capacity vs Attendance

C. Bar charts (categorical variables)

The first bar chart that I created was using the Power 5 ranked games subset. I was able to create a bar chart that compares what channel a Power 5 Ranked game was on. **Figure 5** shows that the channel that has the most Power 5 ranked games is ABC, followed by ESPN, CBS, and FOX. Everything after this has less than 10 games per year that are a Power 5 game and ranked.

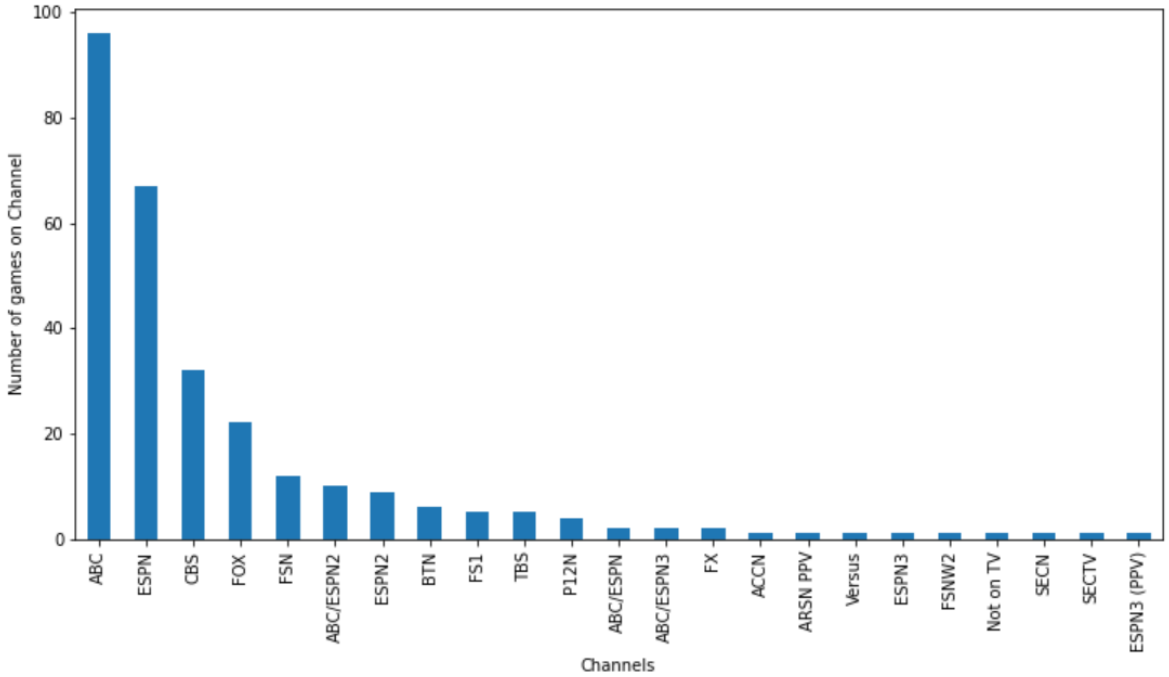


Figure 5, Number of Power 5 ranked games per channel

The next bar chart that I created was dividing up what time each Power 5 ranked game was on. The time listed is what time the kickoff of the game was. I kept the times in chronological order so you can see how the number of important games would progress throughout the days. This can all be seen in **Figure 6**.

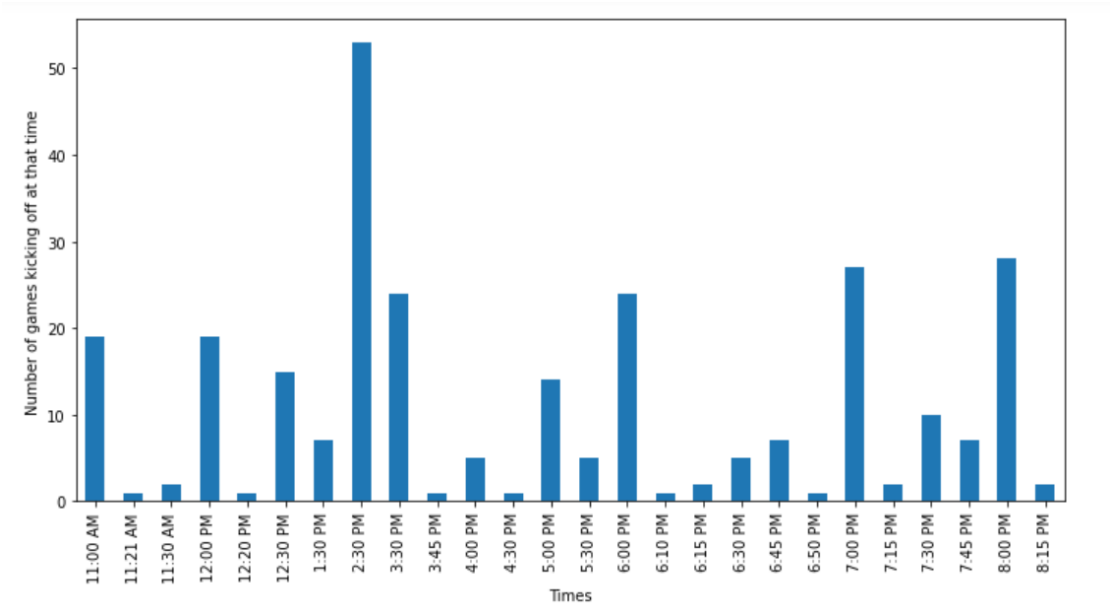


Figure 6, Count of kickoff times in power 5 ranked games

D. Other Plots

I have decided to recreate the times chart but this time with a line chart. The line chart on **Figure 7** shows the relationship between the time throughout the day and how many games are played throughout that time.

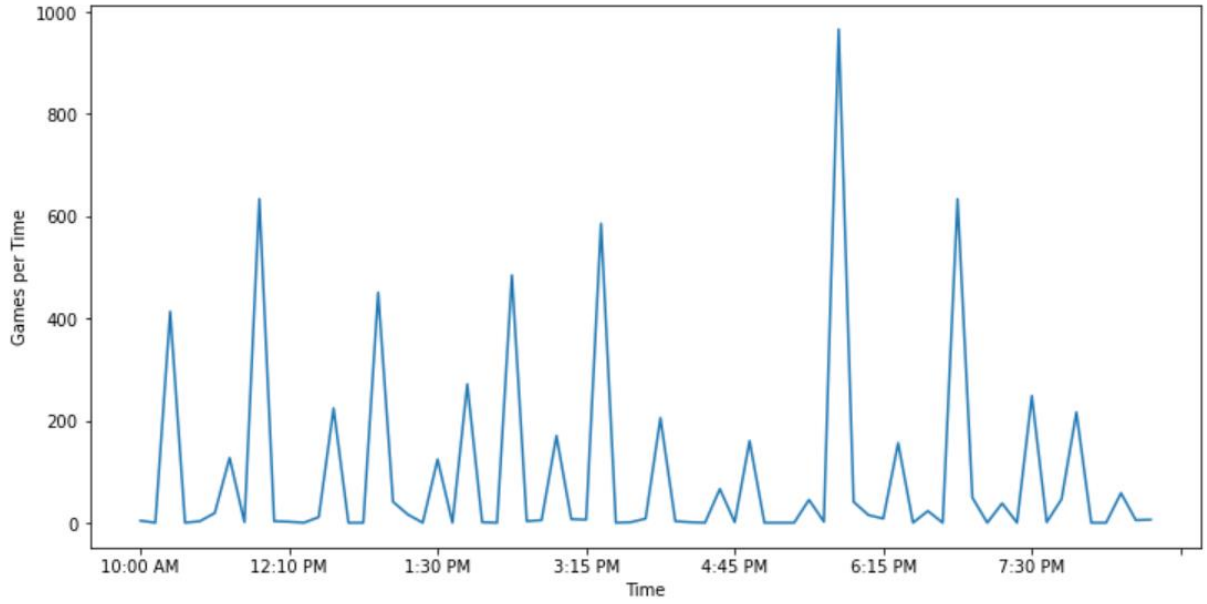


Figure 7, Line graph of games for the times per day

V. SUMMARY OF FINDINGS

This dataset gave me some interesting results that I was not expecting. For example, I was not expecting the fill rate to be above 1.0 in some cases. These are cases where the attendance was larger than the stadium capacity. Some of the other results were not too shocking. You can see in the time graphs that the bigger corporations have figured out what the best channels and what the best times to start the game would be. The weather did not play too much of a concern regarding attendance as I thought it would. I thought if it was hotter or colder it would affect the attendance, but it did not.

Overall, I think I learned some valuable information from this dataset. It was tough finding a full dataset with all the specifications that I needed. This a very interesting topic for me as a fan of college football and it was fun exploring this data from it.