LOUISIANA STATE UNIVERSITY SHREVEPORT

MBA 744 – DATA VISUALIZATION

FINAL PROJECT WRITTEN REPORT

PREPARED FOR DR. LIND

BY GARRETT PHILLIPS

08/17/2022

**Data Sources**

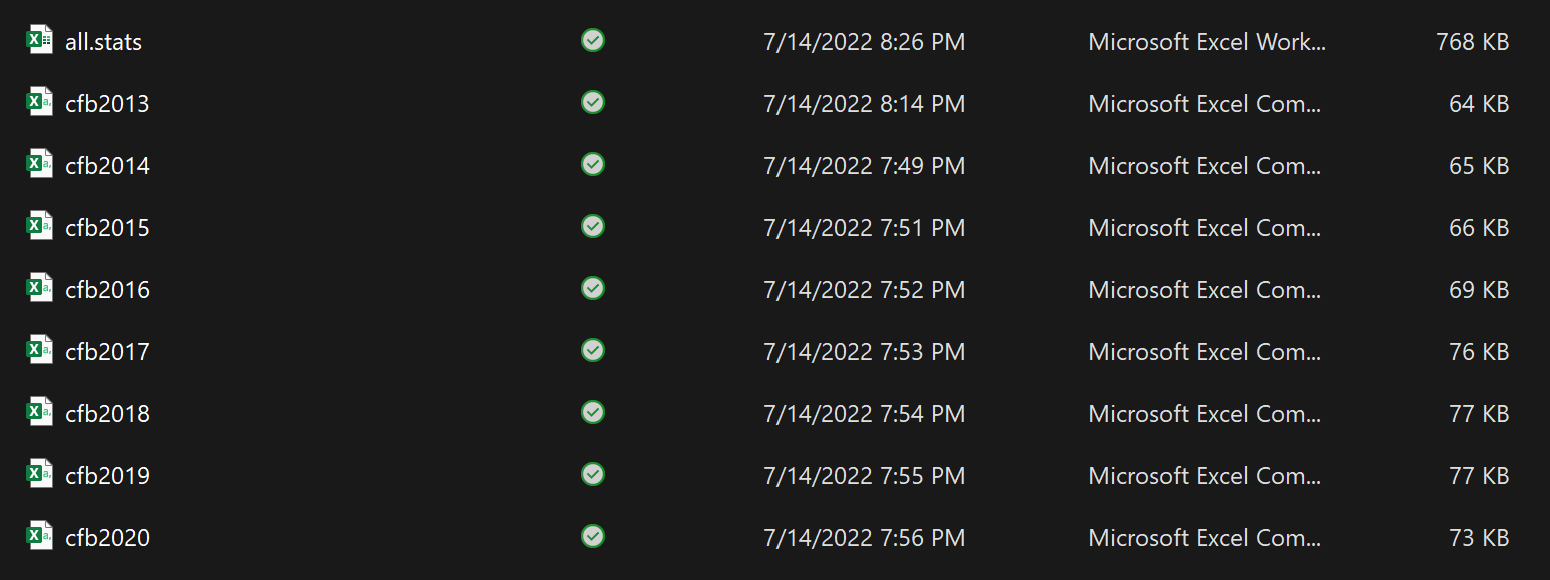
For my project I used three separate data sources. First, I used a dataset with college football attendance statistics. I found this set in Kaggle at <https://www.kaggle.com/datasets/jeffgallini/college-football-attendance-2000-to-2018>. The data was extracted from Wikipedia, and the author claims “…many of the data points have been collected from Wikipedia and therefore lack total integrity” (Gallini, 2019). This note will become important later in my explanation of a few KPI’s. Although the data is not 100% accurate, it is the most comprehensive dataset I could find regarding this matter. This file was a Microsoft Excel worksheet (.xlsx) and contained data about college football teams, spectator turnout, location of games played, and fill rate. This dataset contained valuable information about audience turnout. The information within this source allowed me to begin formulating KPI’s and how to go about shaping this project.

For my next dataset I drew from a comma separated variable file (.csv) regarding offensive statistics for each team. This file contained information such as touchdowns scored, total offensive yards, interceptions, sacks, etc. I found this dataset in Kaggle at <https://www.kaggle.com/datasets/braydenrogowski/college-football-offensive-stats-20102020>. These datapoints hold information for deeper analysis regarding how a team operates (i.e., passing team vs. running team, offense or defense heavy). I was able to gain insight through analyzing this data and finding trends within the other datasets.

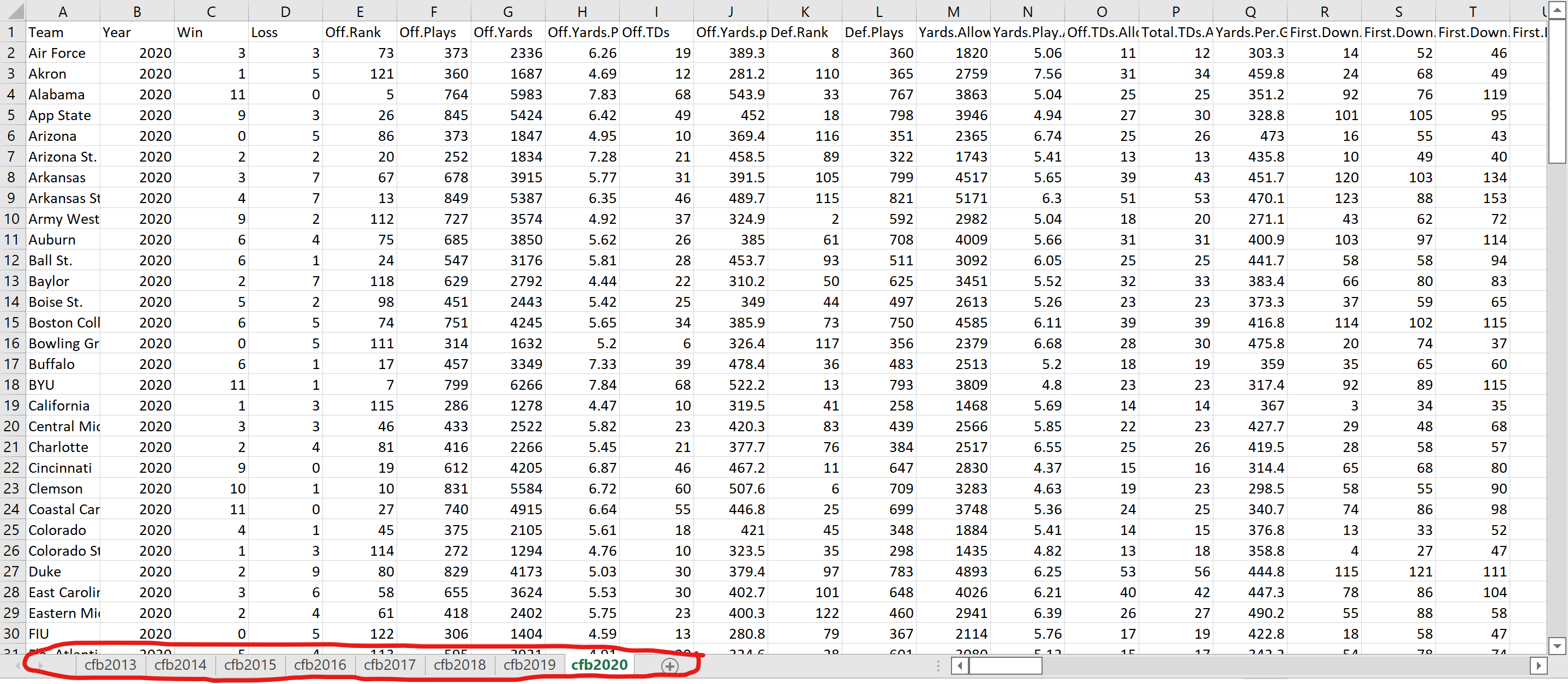
Finally, my last dataset contained information about overall team performance year by year. The dataset was found at <https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?select=cfb19.csv>. This data was formatted as separate years in .csv files, so I had to individually download each season I was analyzing (2013-2018). I copy and pasted each year onto the same sheet. The information housed in this set pertained to wins and losses, offensive and defensive ranks, and total touchdowns. This information gave an overview of each season’s performance by each team.

**Data Cleaning**

Due to the various formats of my datasets, I had to get creative when it came to making my data uniform. After considerable thought, I decided to clean my data using Microsoft Excel’s Power Query. However, before making all of my data uniform, I had to ensure the information was in as few places as possible. As I mentioned previously, the dataset regarding overall team performance was separated by year in .csv files. Here is a screenshot of how that data looked:



I downloaded each data set as a separate Excel csv file and created a .xlsx file containing all the data on separate sheets. The sheets combined looked like this:



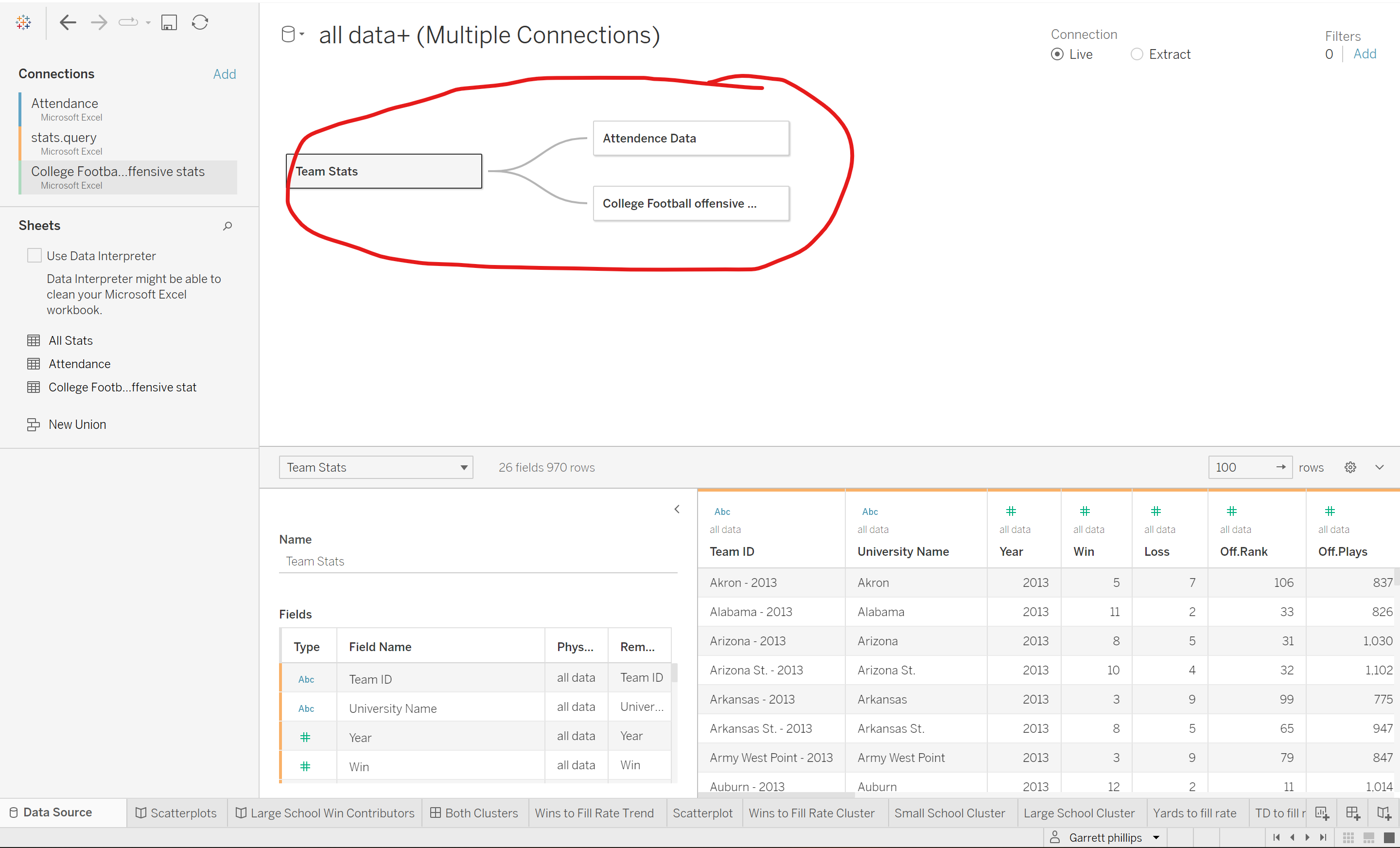
The red circle represents the separate files transferred to sheets. I then copy and pasted each sheet into one large sheet that housed all the information. My new file was prepped by creating a unique ID for each season. The key I created was “Team Name – Year”. The ID was created using a text join function combining the team name and year with “-“ being used as the delimiter. Here is the formula used to get the correct formatting:

=TEXTJOIN("- ", TRUE, [@[University Name]],[@Year])

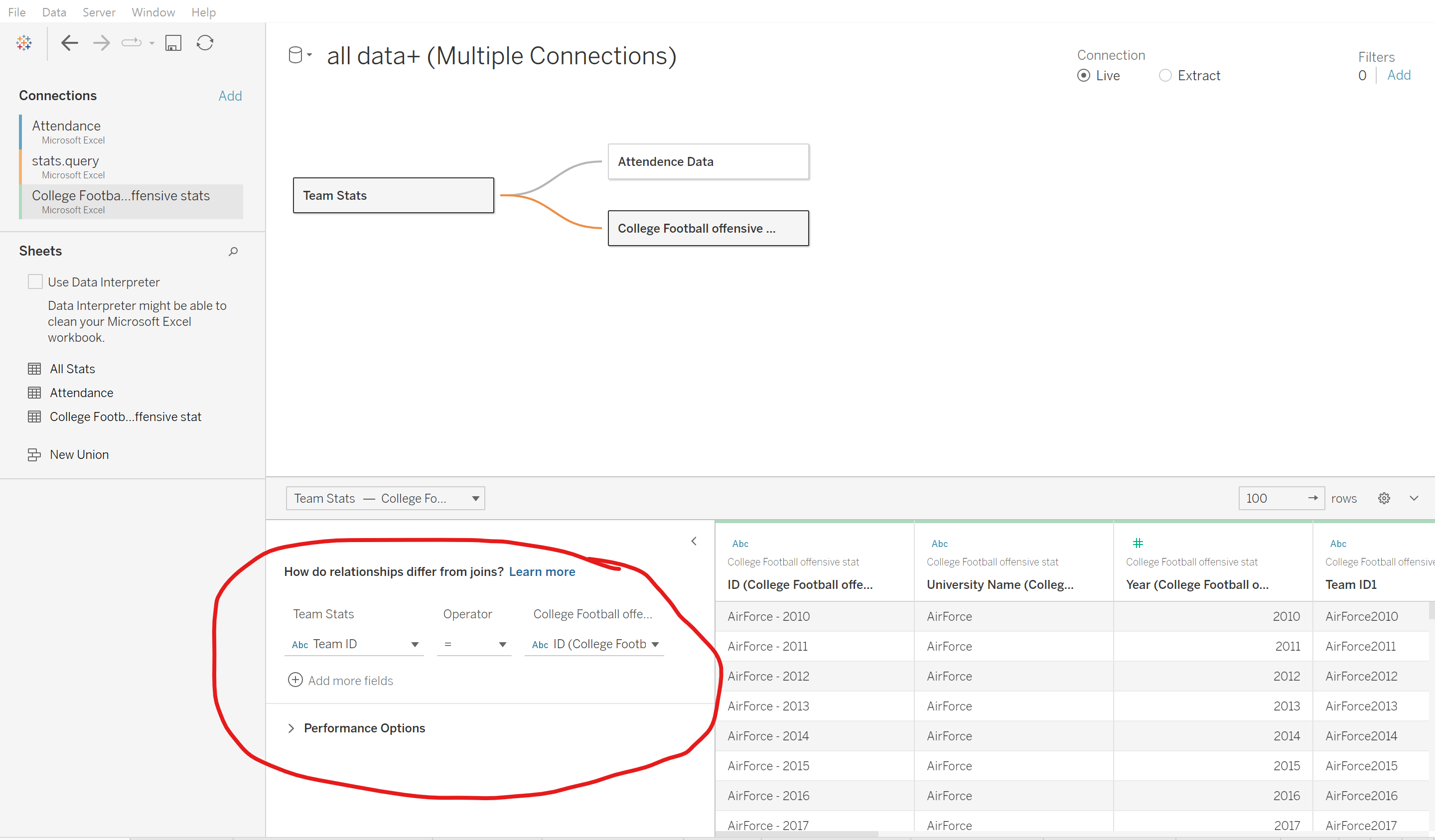
After creating the key, this dataset was ready to be loaded into Tableau. To ensure my tables could join, I did the same to the other two datasets. Once the key was created, my datasets were ready to be loaded in.

**Table Joins**

To load all three of my datasets into Tableau and have them joined correctly, I linked the attendance and in depth offensive stats to the overall team data. That join looks like this:



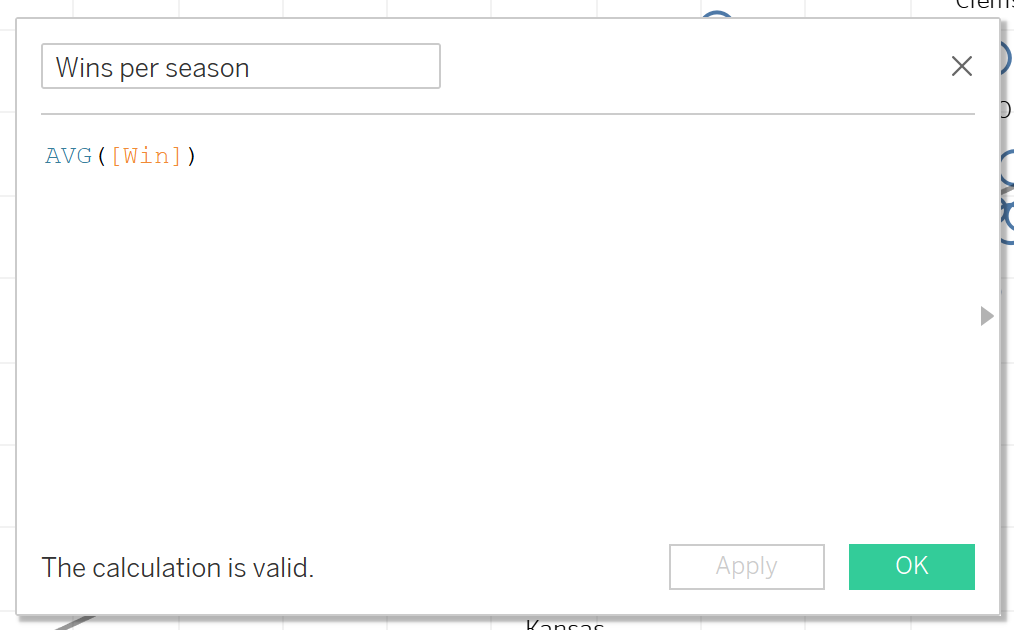
Once I had dragged them into the connections, I linked them using the unique team ID I had created with the TEXTJOIN function.



Now that each connection was made and reviewed, I was able to begin working with my datasets.

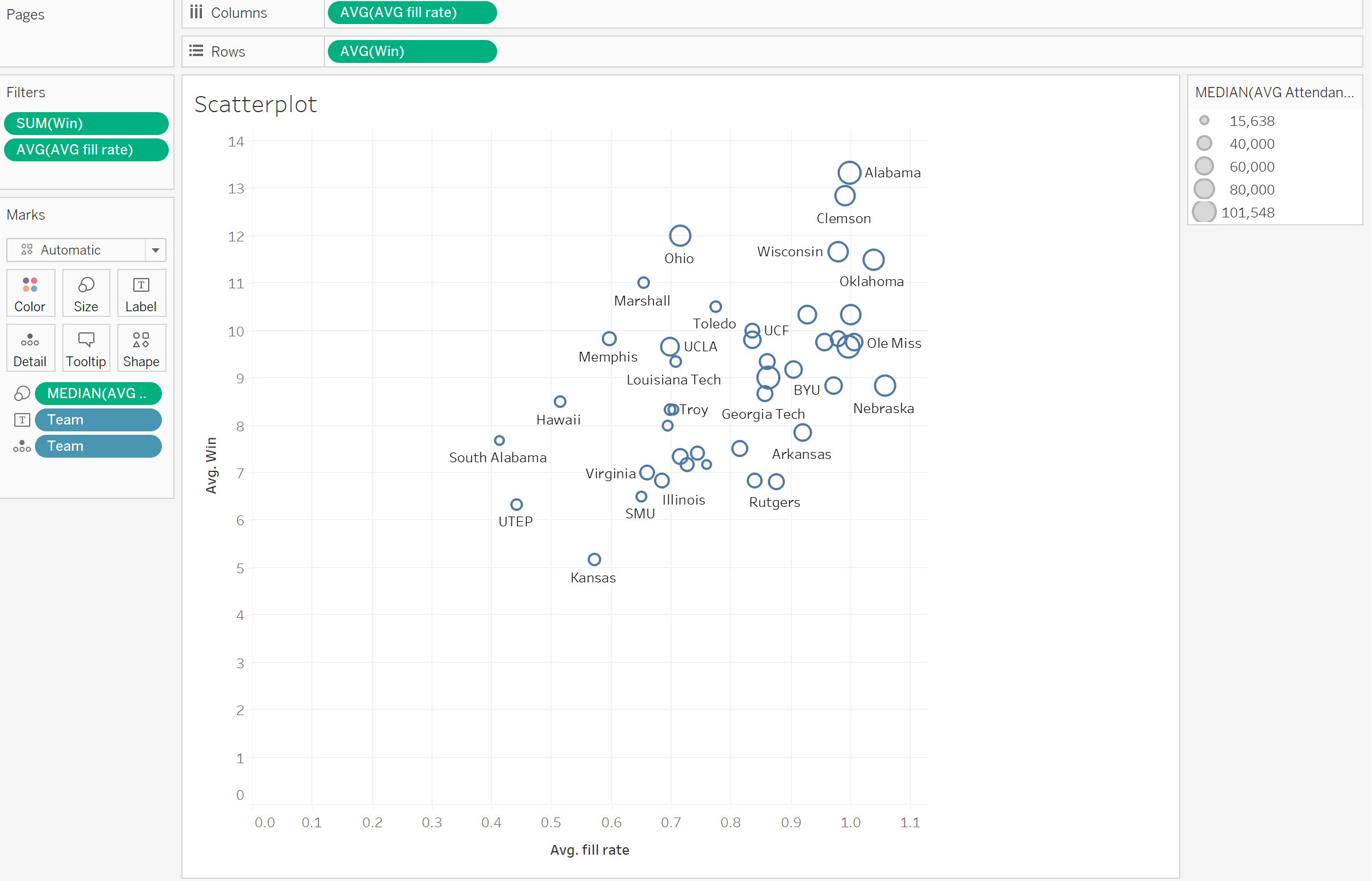
**Key Performance Indicators**

After my data had been correctly loaded in and linked, I could begin examining my desired KPI’s. Because my project is centered around examining the relationship between team performance and ticket sales, three KPI’s immediately stood out. The first and most obvious KPI for my data story is that of fill rate. Unlike total attendance, fill rate accounts for stadium size which shows relative ticket sales for each team. For example, UCLA selling 336,000 tickets per game in 2017 is much more impressive than Georgia States average sale of 233,000 the same year, but when we account for fill rate, Georgia State had a 84.8% fill rate for that season compared to UCLA’s 60.5%. Because of these discrepancies, fill rate is a much more accurate representation of filled seats. As stated earlier, the attendance data originated from Wikipedia. This was the most exhaustive set of year over year attendance statistics I could find publicly, so the results must be viewed with a grain of salt.

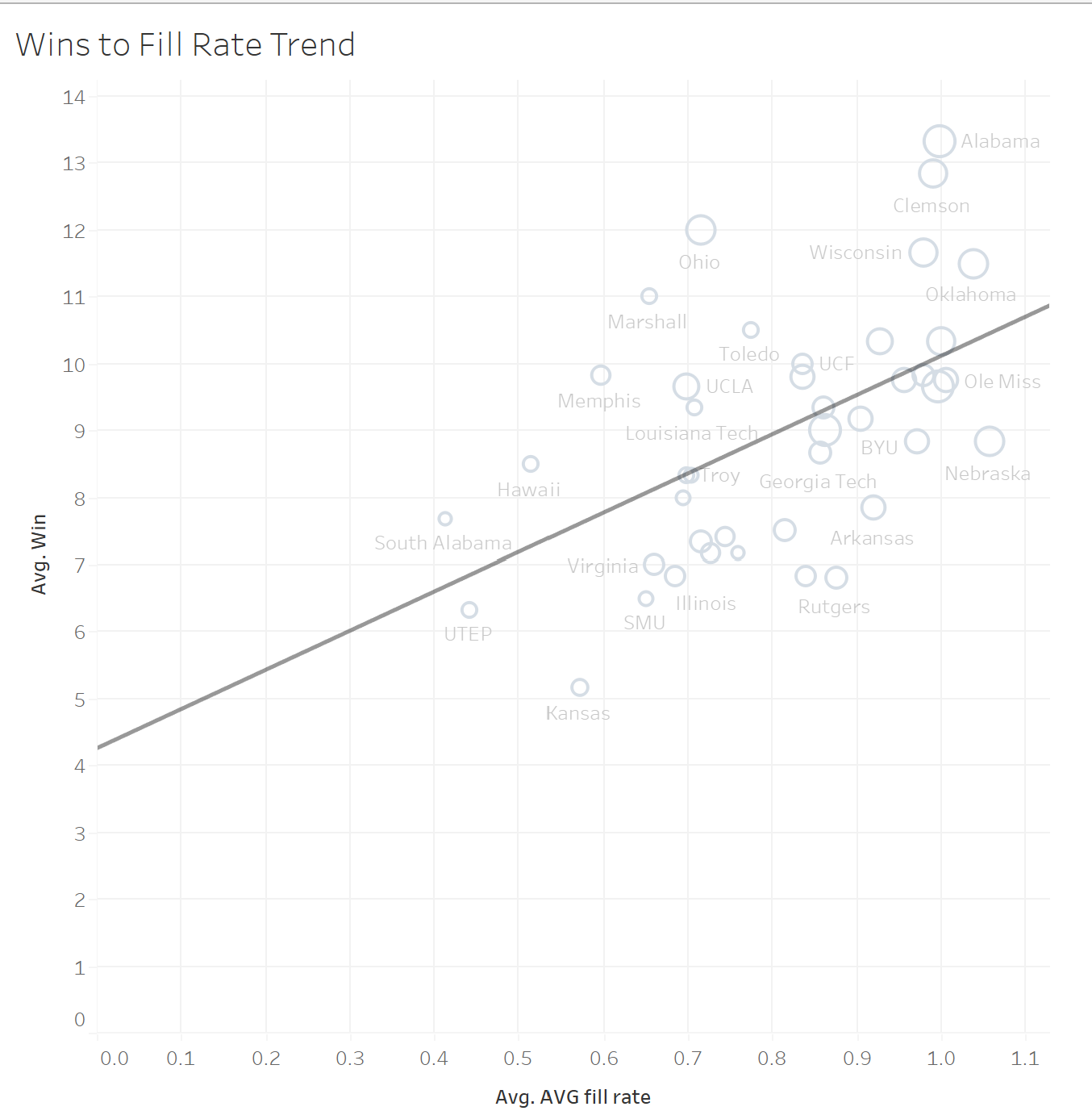
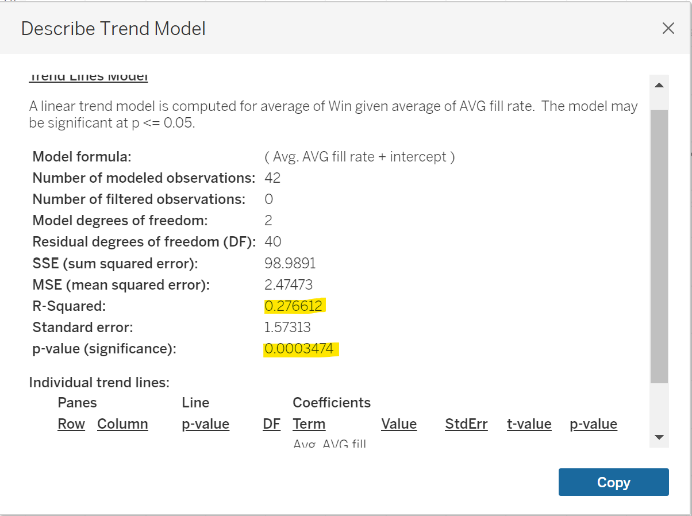
The next important metric is average wins. By studying the relationship between wins and fill rate, I uncovered some key insights. I used average wins for each team to create most of my views. The purpose of this project is to examine how various factors affect crowd turnout, and wins per season seems to be an obvious indicator. The wins per season were already data points in the “Team Stats” sheet, so I simply created a value for average wins per season. 

After I had created this value, I considered my third and final KPI. I decided to use average yards per season and average touchdowns per season in the same way to highlight their effects on fill rate. These indicators were not very insgihtful in and of themselves, but when tied to total crowd size and fill rate, I was able to track various trends in clusters I was able to identify.

**Trends**

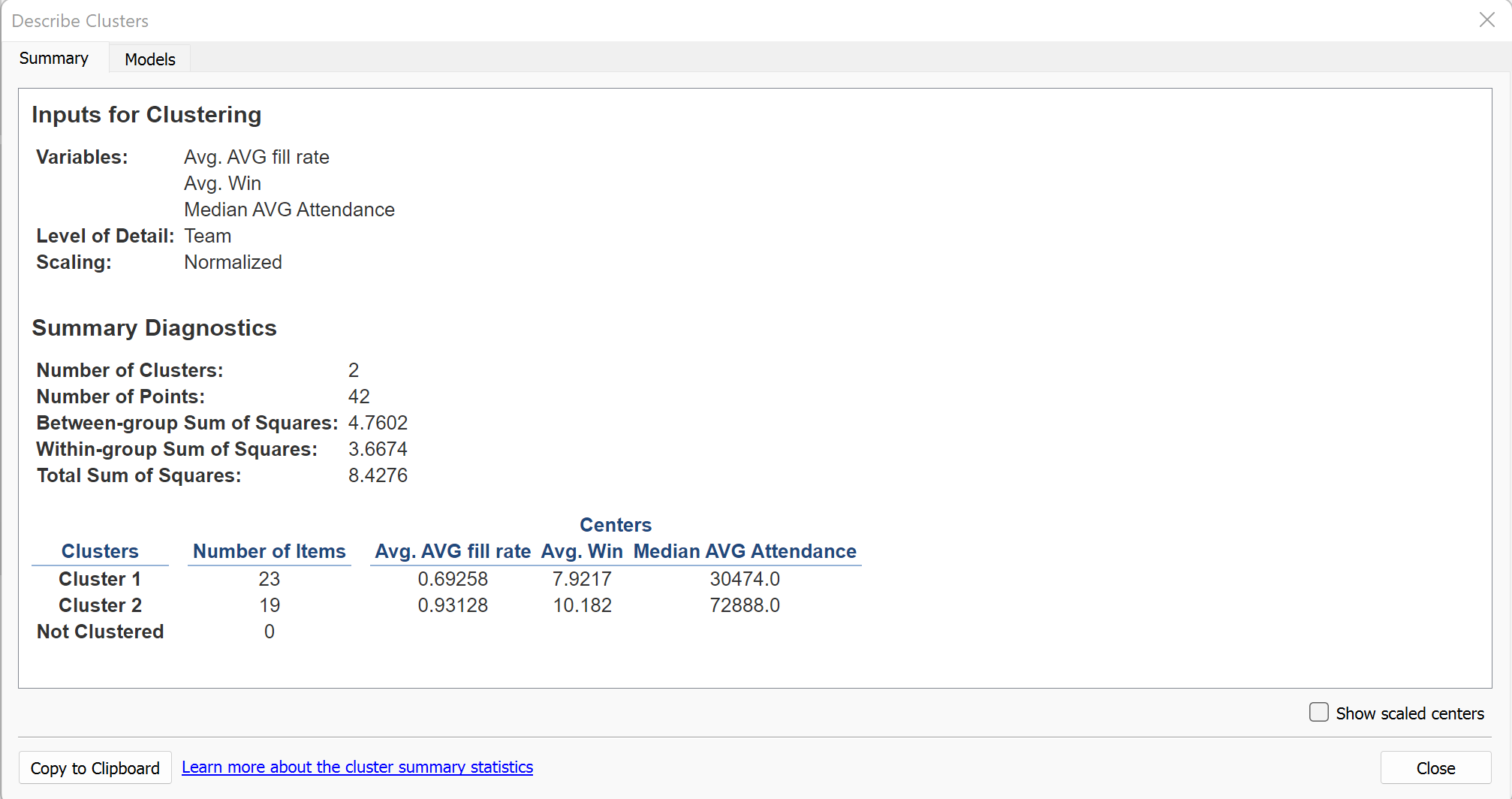
Initially I began to investigate the trends within wins to fill rate. I placed each team on a scatterplot where I placed average wins on the y axis and average fill rate on the x axis. 

One thing that immediately stood out to me were the teams with fill rates over 1.00. I began to investigate this issue and found that the stadium capacity was slightly incorrect. I formatted the data best I could and continued to work with this scatter plot. I could immediately see the trend in the data, and placed a trend line into the graph for the next view.

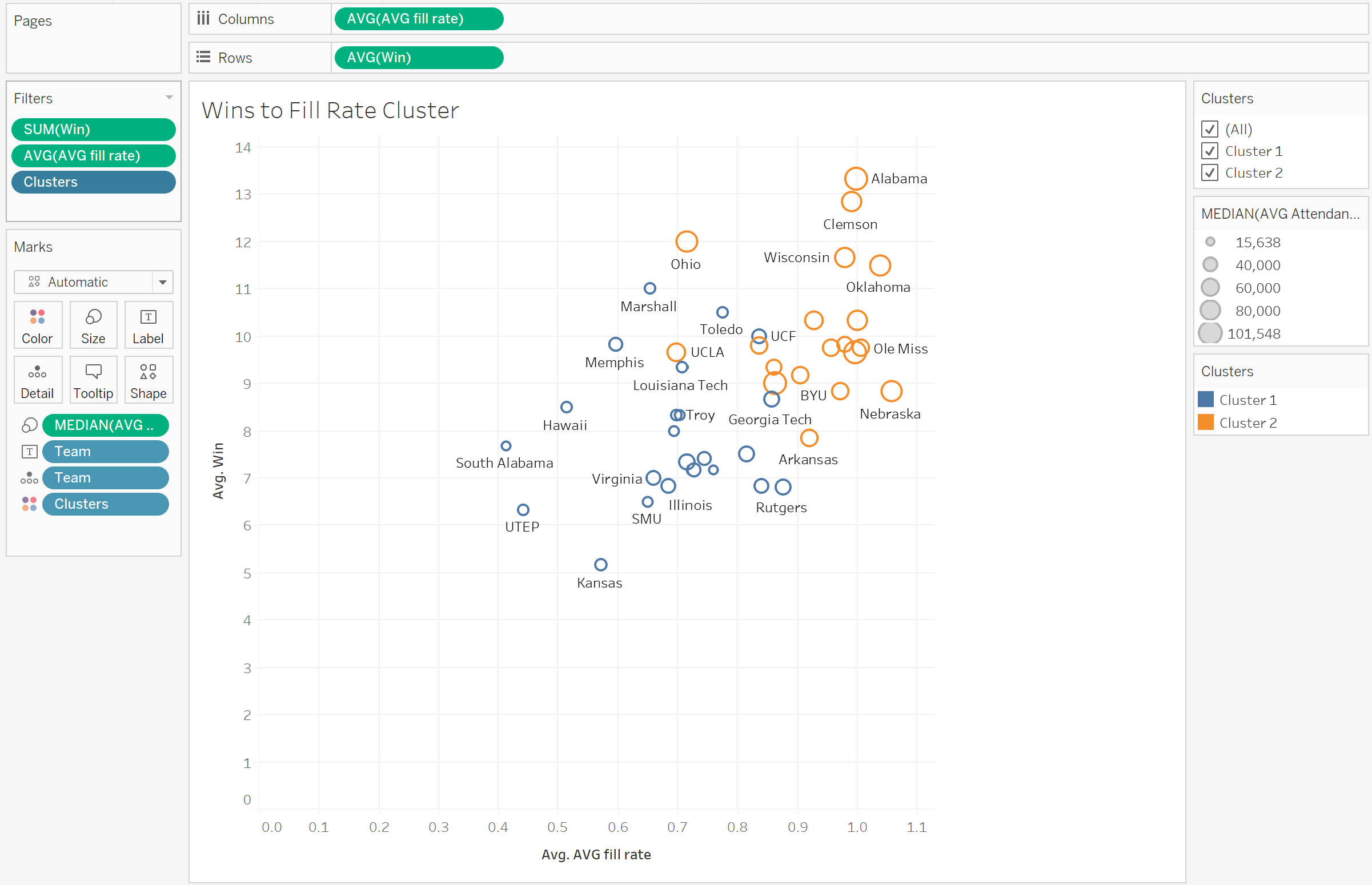


Through the trend information, I found that the R-Squared value was 0.2766 with a p-value of 0.00035, so I knew this trend highlighted a statistically significant relationship between the two variables. The R-Squared value allowed me to interpret the accuracy of the prediction. Although I would have liked to see a much higher r-sqaured value, the value means that the trend line can predict about fill rate from wins at about 27%.

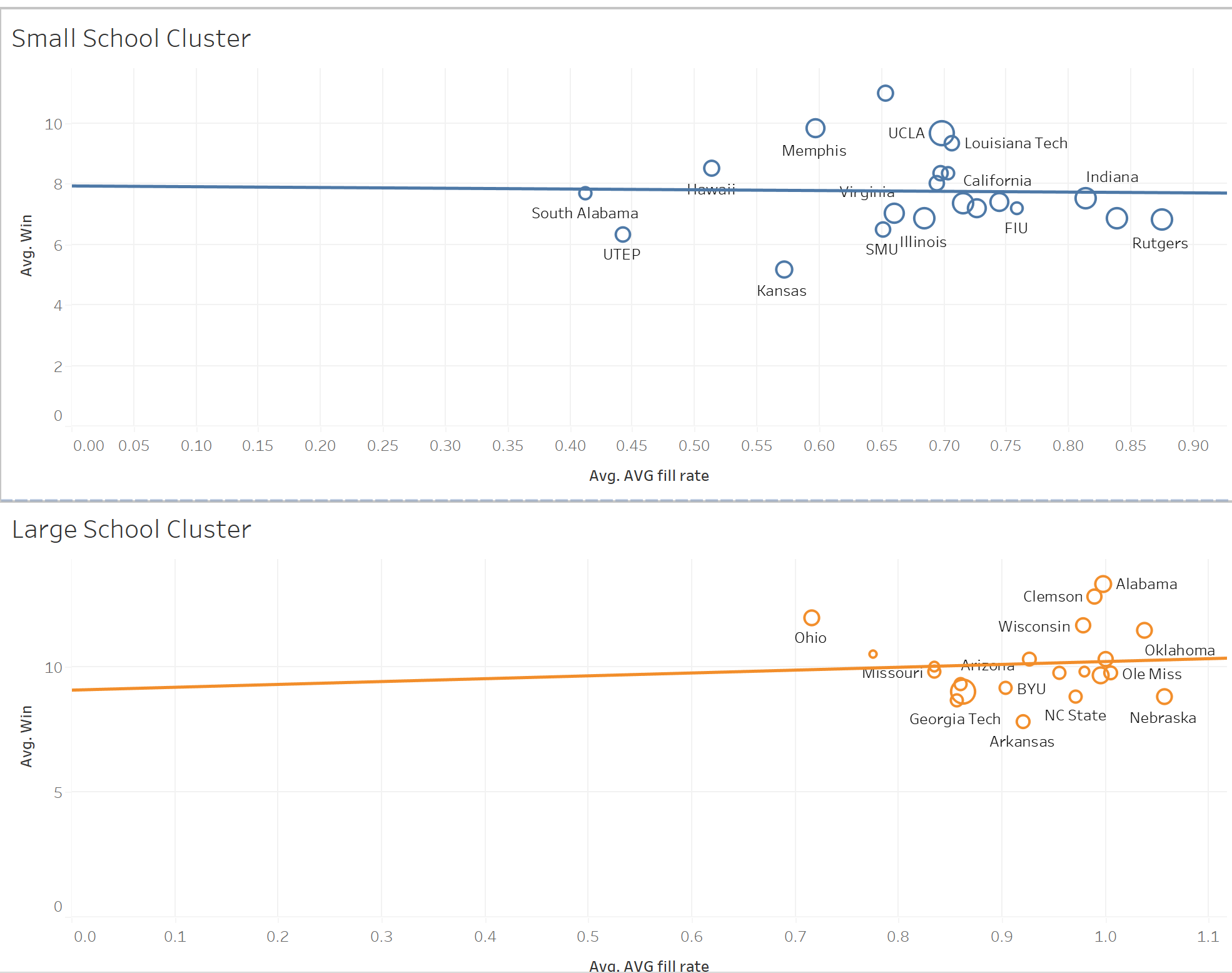
Next I utilized the cluster feature within Tableau. I bagan by dragging the cluster feature in the analytics tab and dragging to the satterchart. The analysis took into consideration both average fill rate, average attendence, and average wins for each respective university. The statistics regarding this analysis is shown below.



The two groups this analysis created are shown in the following image.



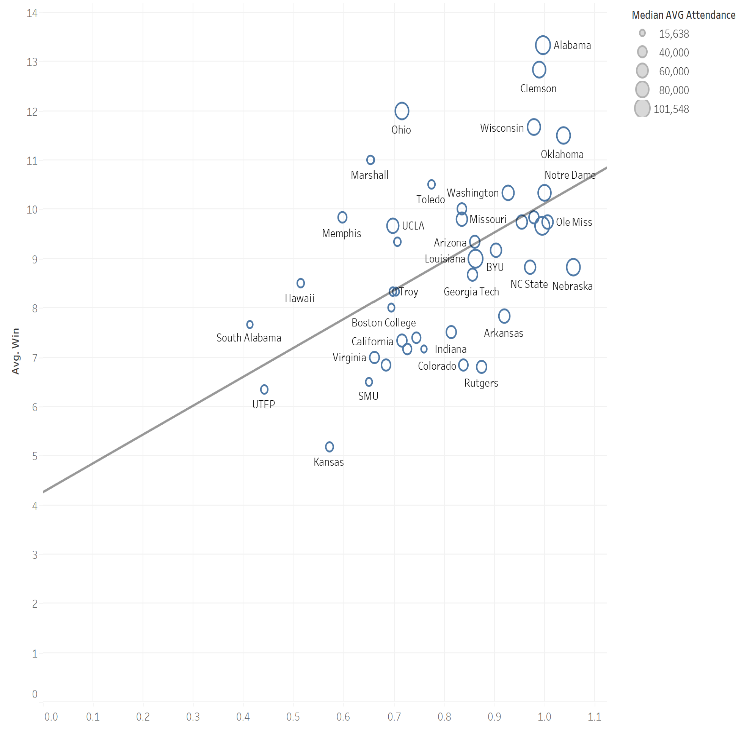
Once I had the two clusters, I created two tend lines for each cluster and placed them in the dashboard shown below.

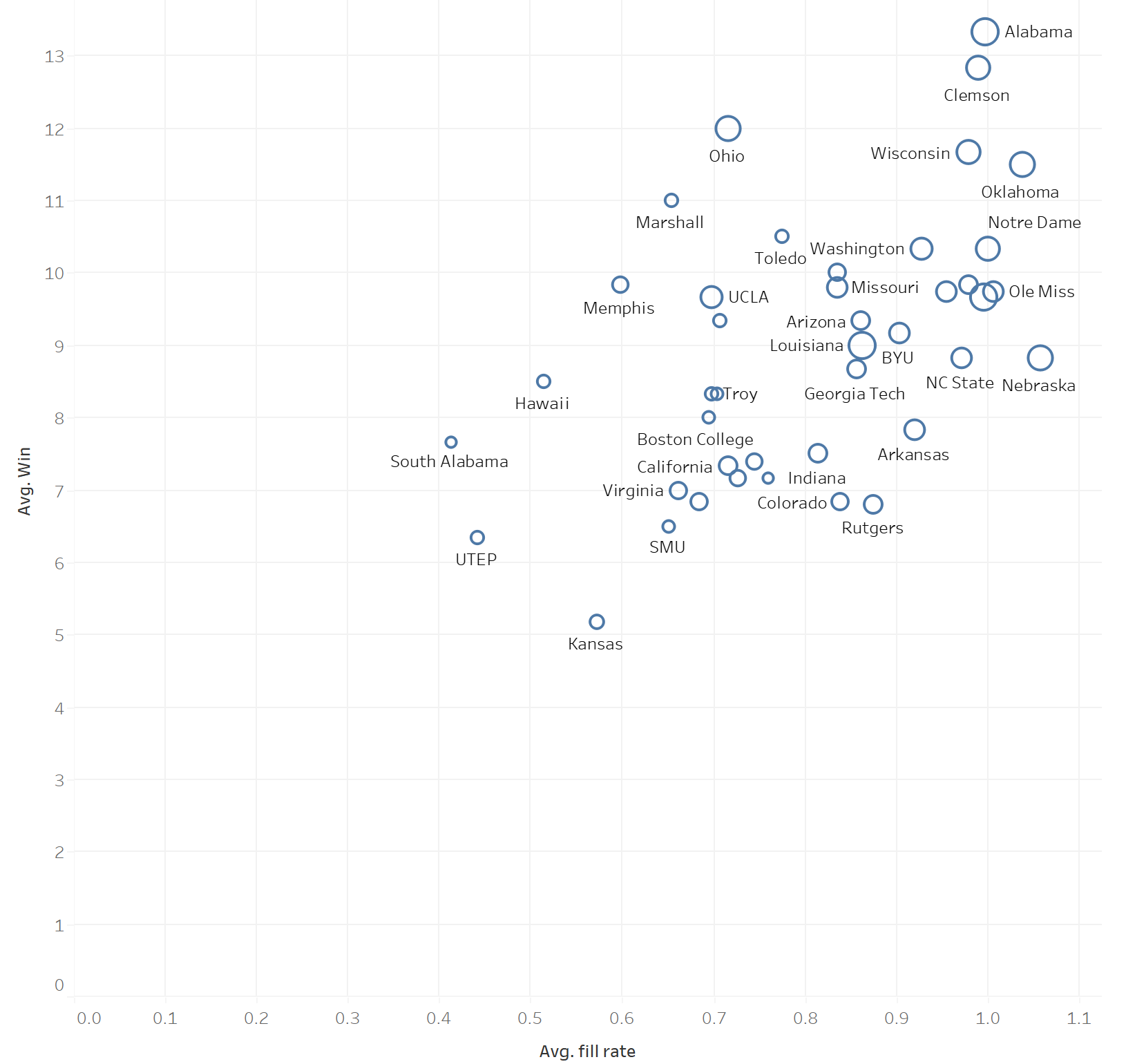


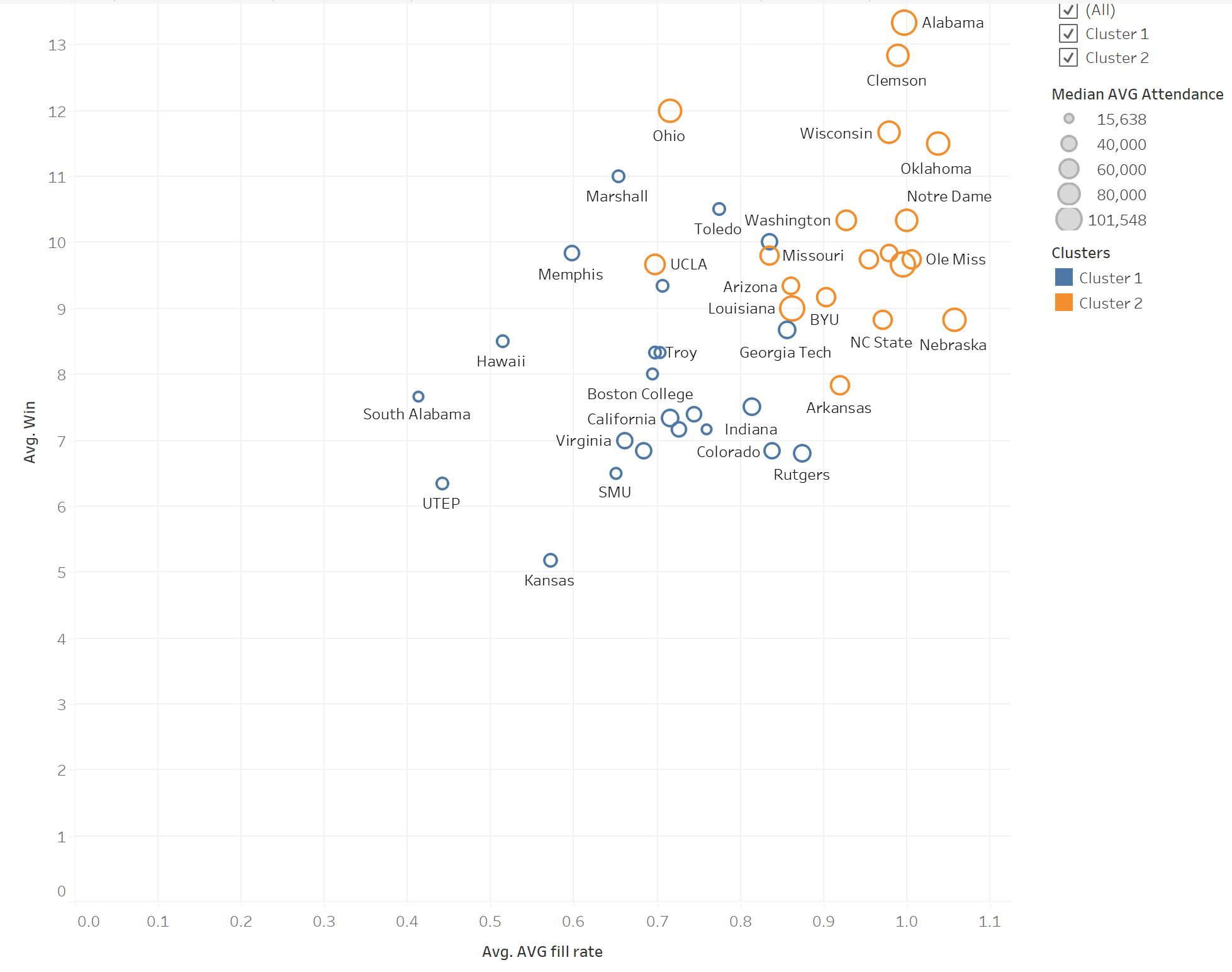
I noticed a greater relationship within the alrger school cluster, which I’ll touch on later. Thorugh these two analysis, I was able to gain a greater understanding of the relationship of wins and fill rate between the two groups.

**Graph Selections**

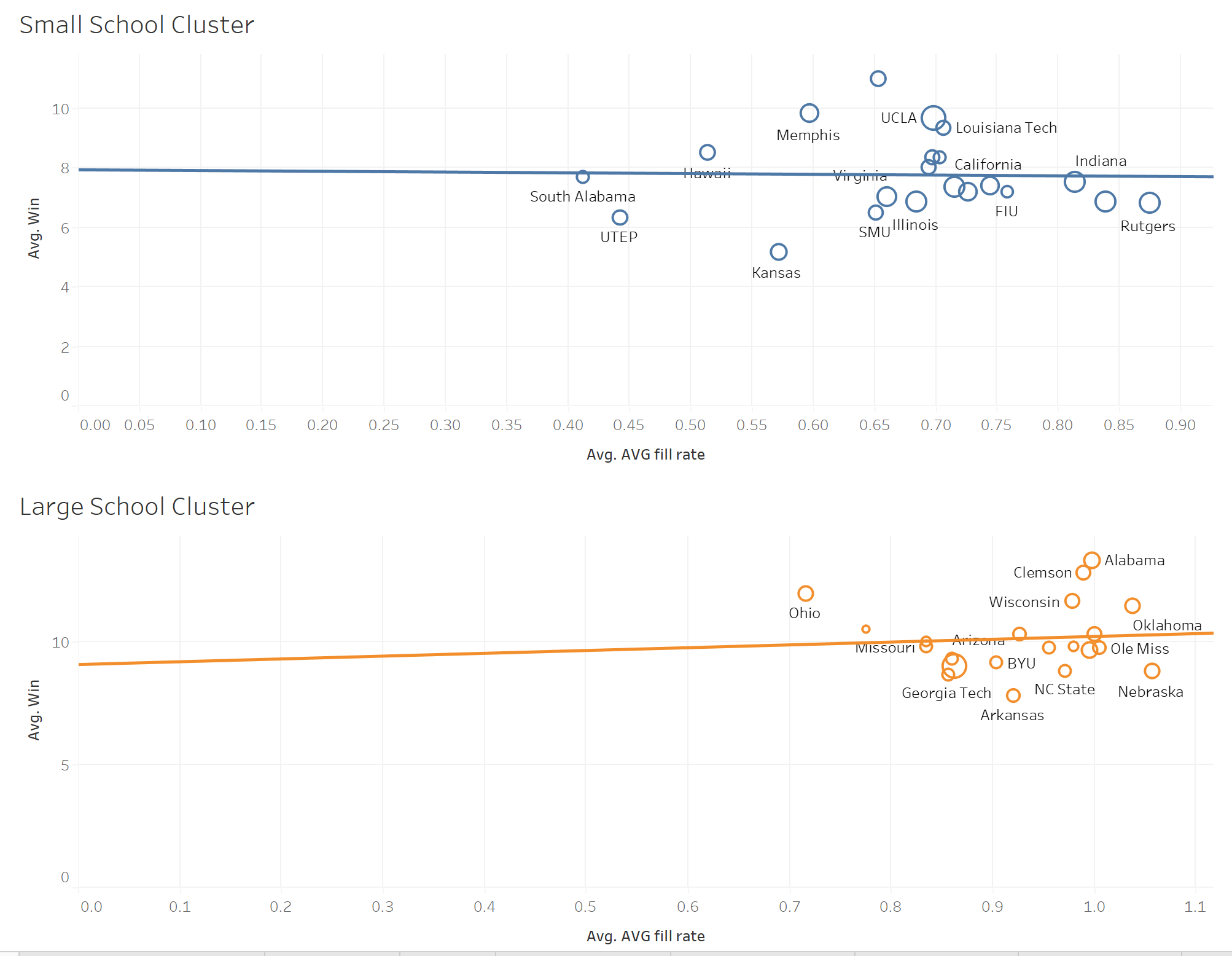
To show the various relationships within my data, I decided to begin at a generic level and drill down to the specifics of each group. There are many ways to present the data, but I decided to introduce the subject by familiarizing my audience with the big picture goals. The first portion of my presentation will be composed of highlighting my key performance indicators and explaining the formulas used to come to my conclusions. By doing this, it will allow the audience to understand the metrics used and how they are relevant to my discoveries.

The first views I used in my report are that of a story. Through a five-slide presentation, I touch on a few key areas. In my first view I introduced my audience to the subject matter and explain the correlation between average fill rate and the average amount of wins each team records. I used a scatter plot with my y axis representing average wins and an x axis representing average fill rate. Upon introducing the subject, I highlight teams the audience is familiar with (i.e., Alabama, and Clemson) as well as smaller teams the audience may not be familiar with (i.e., UTEP and South Alabama). Because the opening is an overview, my audience will quickly understand the purpose of my presentation and the relationships between my chosen variables. Once the audience is familiar with the metrics, I progress to my next slide in which I highlight the trend of wins to fill rate. This view incorporates Tableau’s analytic capabilities and retrieves statistics such as p values and R squared values. I touch on the statistical significance and explain the correlation in the data. Below is the first set of views in my presentation.



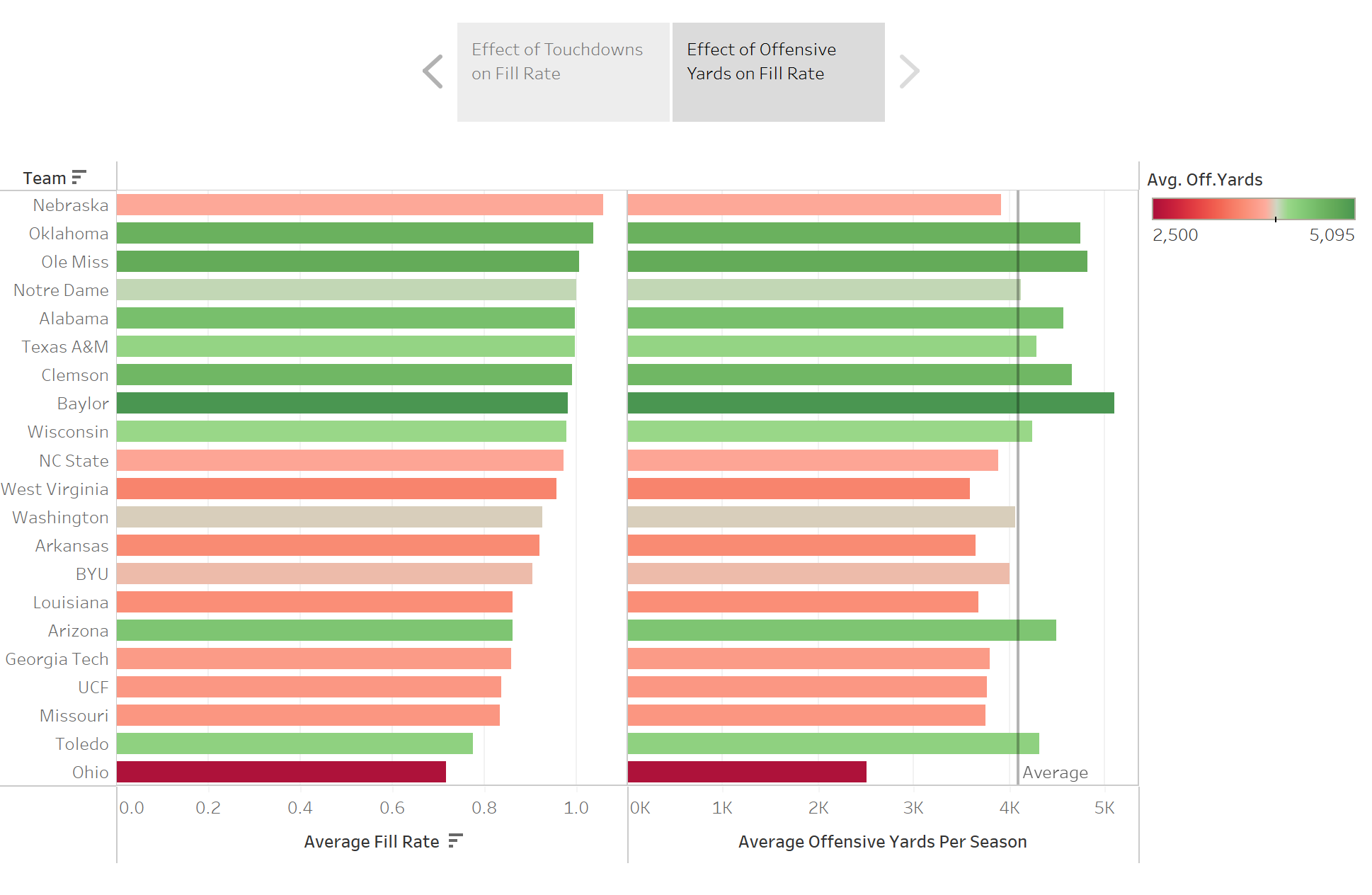


After my viewers understand the trends in the data, I move to my next iteration of the scatterplot. In the next view, I utilized clustering in my graph to highlight two distinct groups within my data as shown above. Once the two groups have been explained thoroughly, I broke down the trends within the two groups into a single dashboard within the story. Originally, I had planned to show the two groups in separate views, but after reviewing the story I noticed that the extra space within each view was excessive. I believed putting both in a dashboard would bombard my audience with information, but after reviewing the dashboard I believe this is a more effective way to communicate the findings.



After showing the overall trends and groups within my dataset, I dove deeper into the drivers for wins. Because I will explain to my audience that only one cluster is affected by wins, the next series of views were filtered to only show that cluster. Similar to my previous approach, I began at a broad level and drill down to specifics. In my first two visuals I will show the top fill rate teams ranked, and their offensive performance. This will allow my viewers to understand the offensive drivers behind filling stadiums. To create this view, I color coded the bars to show a green to red divergent representative of each team’s performance I was analyzing. I chose this color scheme to highlight the distance from the mean value for each subset. This will allow the audience to quickly identify a key trend within the cluster, as divergent color schemes enhance the viewers ability to immediately identify positive and negative values (Kashyap, 2020). I also added an average line so the viewer could easily create a mental benchmark of the cluster performance. Those views are shown below.





**Conclusion**

Through my analysis, I was able to gain a better understanding of the data and how the relationships between the different data points. In my findings, I found two distinct clusters with different qualities. In the cluster regarding smaller schools, wins seem to drive ticket sales less than amoung the cluster of larger schools. Through this insight, it is appearant that wins have an effect on crowd turnout, however, the degree to which it interacts with this variable depends on the school community. While some schools will see consistent turnout regardless of team performance, others will have a drastic drop in ticket sales if the team does not perform well.

Larger schools rely more heavily on wins to fill seats due to the expectations of fans. For a school with consistent seasons (i.e., Clemson and Alabama), fan bases expect to win more games than they lose each year. If these teams were to fall below these expectations, their ticket sales would suffer. On the other hand, teams like Rutgers, SMU, and Indiana can expect consistent fan bases regardless of the amount of losses that season. One thing to note, however, is that if the teams with more loyal fan bases were to begin winning more games their attendence begins to rise rapidly. For example, for the 2021 football season, SMU recorded 8 wins, and their avergae fill rate was significantly higher than average.

**Recommendations**

I have two separate recommendations dependent on the cluster each team resides in. For larger schools, I would recommend continuing to invest heavily in the recruitment of offensive superstars. By accumulating a good recruitment class, teams will be able to have more high-powered offenses which leads to offensive yards, touchdowns, and wins. In addition to talent acquisition, I would encourage these teams to increase ticket prices during good seasons to increase profits.

For smaller schools, I would advise athletic departments to focus on the overall fan experience. Because smaller schools have a lower relationship between wins and ticket sales, I would invite alumni and community members to attend games based on the gameday experience. Additionally, I would advise these schools to charge less for alcoholic beverages so that fans are more intoxicated, thus less invested in the team’s performance. I believe by pairing these two initaitives, smaller schools would be better positioned to fill seats and keep fans happy.

**References**

Gallini, J. (2019, December 9). *College football games (2000 to 2018)*. Kaggle. Retrieved August 17, 2022, from https://www.kaggle.com/datasets/jeffgallini/college-football-attendance-2000-to-2018