**QMBE 1320 Final Project Write Up Part 1**

Kyle Mortenson and Per Waage

**What data did you choose and why?**

We decided to choose the NFL data set because that's the one we thought we would be most interested in as well as most familiar with. We both have a good understanding of sports leagues and how they operate as far as the corresponding stats that are provided with the data. We also looked at the data summary statistics between all of the data sets and it was made clear that the NFL data would be very workable. The main item we looked at was the missing values in the data and the NFL did have some missing values however it wasn’t the data set with the most. We also decided to look at the correlation between different variables and it had stated that wins and losses were highly correlated to some of the variables listed. This we both agreed upon would lead to a much easier time to try and figure out why some things lead to either a win or a loss. This will help us determine what we want to do with our data and how we will come up with the questions we want to ask to help us better sort our data.

**2-3 Problem Statements:**

1. We want to increase the attendance for games by figuring out what the key influences are for greater attendance numbers.
2. We want to find out how the result of a win is affected by the winning teams turnovers as well as the losing teams turnover
3. We want to find out how point differential changes if a team makes the playoffs or not.

We chose these questions because we feel there would be a high cause and effect relationship at play here. Meaning that we believe these would give us results based on our past knowledge of how the NFL works that would best correlate to each other. These stats we chose to pick out usually lead to higher win percentages which in turn leads to the variables that we picked out to ask questions about.

**What is your plan to analyze these problem statements? How will this analysis help the consumer or business?**

To analyze our first problem statement we plan to make 2 different pivot tables showing the attendance on a weekly basis (one pivot table for weeks 1-8, and one for weeks 9-17) and identify the changes in attendance from the first half of the season to the second half. After observing the pivot tables of attendance for each half of the season, we were able to find a decrease in attendance in the second half (weeks 9-17) for teams that did not have a lot of wins. We then plan to make two separate bar charts with one showing the win totals of each team and the other showing the attendance of each team from the first half of the season to the second half.

For our second problem statement we plan to make a bar chart showing two separate X variables (Turnovers in win, Turnovers in loss) and a Y variable of the amount of games in the 2000 season to find trends in the data that the bar chart shows to try to answer our question on if a win or loss is affected by a team's turnovers.

For the third problem statement we plan to make a dummy statistic of teams that made the playoffs and teams that missed the playoffs by using either 1 or 0. We then will make a scatter chart showing the point differential of each team and if they made the playoffs (1) or if they missed the playoffs (0) and try to find a correlation between the two variables.

**3. How Did You Clean Your Data?**

We cleaned our data by making a small sample size of games that occurred in the year 2000. This sample size made it easier for us to focus on each team and the 17 games they played in the 2000 season instead of using the whole data set and looking at their attendance for 19 years. By using a smaller sample size, we were still able to find a trend in the attendance when looking at the win totals throughout the two halves of the season for each team. We also looked and saw that the data set had some teams that came into the league during the time period that the data set was collecting data, so in turn we decided to get rid of those teams. Houston and Los Angeles were those teams. We also had to look at the data and interpret differences between the New York Teams and Los Angeles teams because when we had sorted them they came up summed together. So we made sure to have each of their data separate from each other.

**4. Descriptive Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Descriptive Statistics** |  |  |  |  |
| **In the year (2000)** | **Mean** | **Median** | **Mode** | **STDEV** | **MIN** | **MAX** |
| **Attendance Wks 1-8** | **66123.97** | **65588** | **#N/A** | **5082** | **55154** | **77299** |
| **Attendance Wks 9-17** | **65023.48** | **64783** | **#N/A** | **4508** | **51791** | **74094** |
| **Wins** | **8** | **9** | **11** | **3.152** | **1** | **13** |
| **Turnovers In Loss** | **2.515385** | **2** | **2** | **1.461** | **0** | **7** |
| **Turnovers In Win** | **1.27027** | **1** | **1** | **1.167** | **0** | **7** |
| **Point Differential** | **0** | **12** | **0** | **116.7** | **-258** | **180** |

**5. Visualize Data:**

a. Understanding the Dataset a. What are the key variables in the dataset?

The key variables in this dataset include the attendance number per game, how many wins a team had in a given season, the teams, turnovers in a game for a winning and losing team, the year, point differential, and if a team made the playoffs or not.

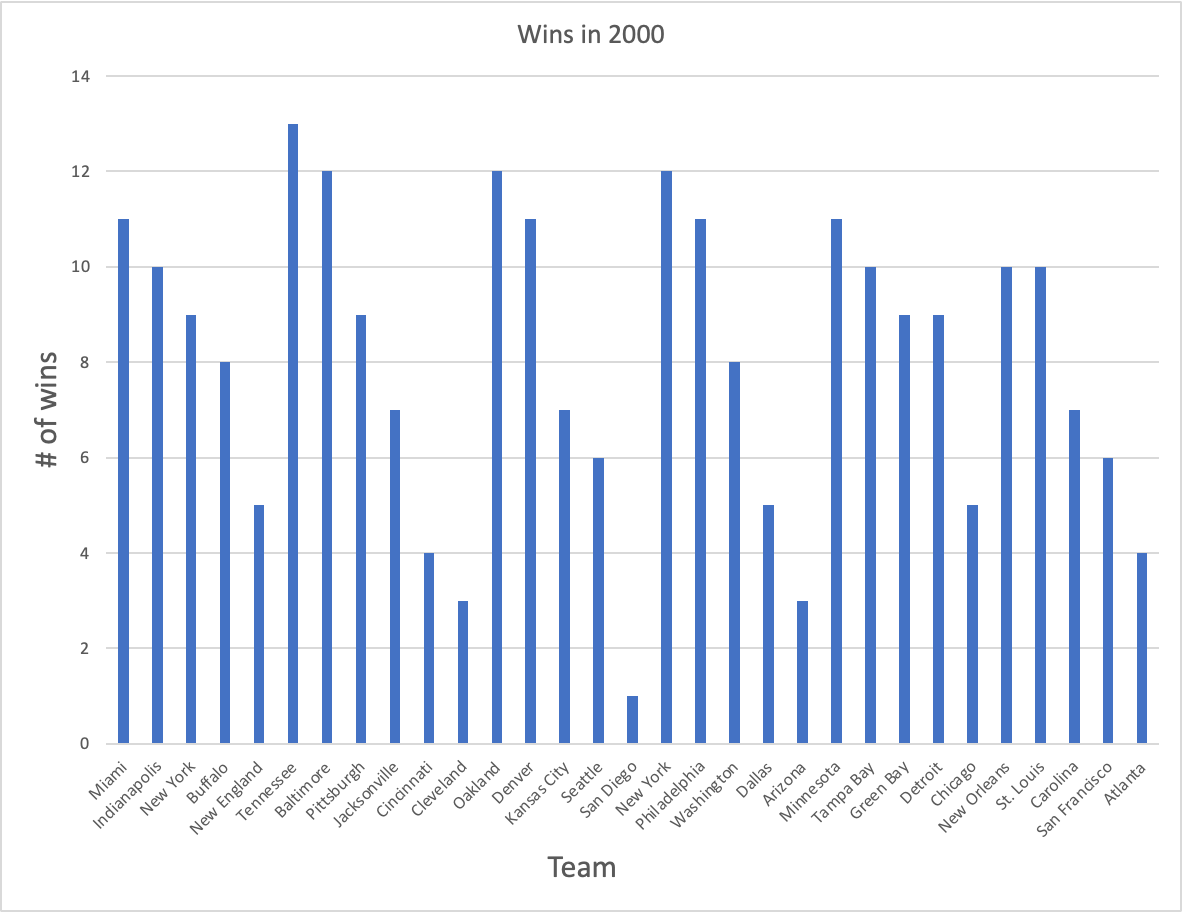
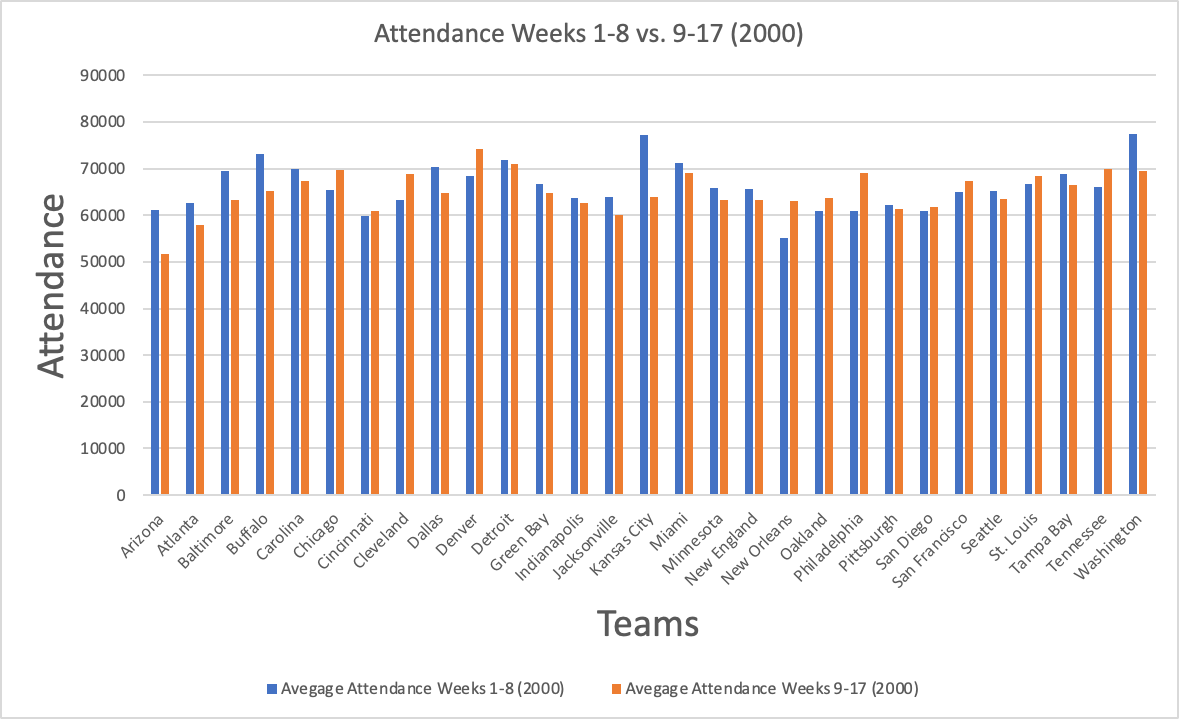
b. What type of variables (e.g., categorical, numerical, datetime) do I have?

Attendance (numerical), Wins (numerical), Teams (categorical), Turnovers (numerical), Year (datetime), Point Differential (numerical), Playoffs (categorical).

c. Are there any missing values, and how prevalent are they?

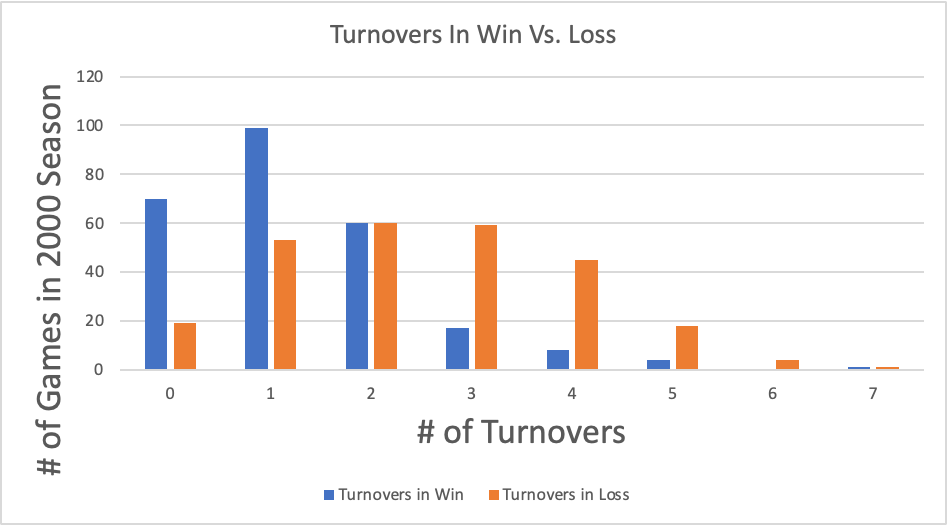
There are missing values for some teams during certain years. They are not prevalent as we decided it was best to just get rid of some of the values because some teams were added and taken away throughout the years but also wouldn't affect our data too much.

|  |  |  |
| --- | --- | --- |
|  | **Average Attendance Weeks 1-8 (2000)** | **Average Attendance Weeks 9-17 (2000)** |
| **Arizona** | **61114** | **51791** |
| **Atlanta** | **62597** | **57975** |
| **Baltimore** | **69562** | **63234** |
| **Buffalo** | **73011** | **65278** |
| **Carolina** | **69835** | **67371** |
| **Chicago** | **65358** | **69726** |
| **Cincinnati** | **59858** | **60935** |
| **Cleveland** | **63241** | **68901** |
| **Dallas** | **70274** | **64839** |
| **Denver** | **68409** | **74094** |
| **Detroit** | **71782** | **70939** |
| **Green Bay** | **66650** | **64783** |
| **Indianapolis** | **63610** | **62526** |
| **Jacksonville** | **63854** | **60140** |
| **Kansas City** | **77117** | **63939** |
| **Miami** | **71099** | **69021** |
| **Minnesota** | **65738** | **63232** |
| **New England** | **65588** | **63235** |
| **New Orleans** | **55154** | **62969** |
| **Oakland** | **60825** | **63652** |
| **Philadelphia** | **60991** | **68990** |
| **Pittsburgh** | **62087** | **61380** |
| **San Diego** | **60992** | **61657** |
| **San Francisco** | **64993** | **67251** |
| **Seattle** | **65091** | **63544** |
| **St. Louis** | **66671** | **68296** |
| **Tampa Bay** | **68808** | **66563** |
| **Tennessee** | **65987** | **69929** |
| **Washington** | **77299** | **69491** |



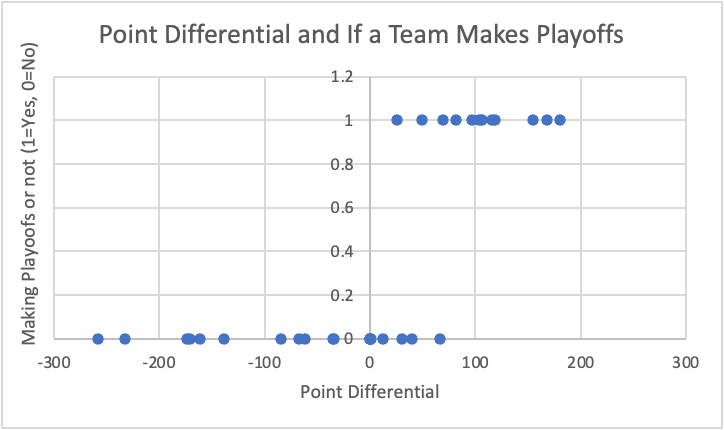
Bar Charts Showing The # of Wins For Each Team In The Season of 2000 and the attendance for the first half and second half of the 2000 season.

We decided to make these graphs to see if there was a correlation between an increase in attendance in the second half of the season for the teams with high win totals as well as a decrease in attendance in the second half for the teams with low win totals. By making the first chart, we were able to visualize which teams had success and which teams struggled in the season of 2000. After examining the chart showing the win totals of each team, we were able to see that the few teams with the lowest win totals were San Diego, Cleveland, Cincinnati, Arizona, and Atlanta. The teams with the highest win totals were Tennessee, Baltimore, and Oakland. We then made a chart to show the change in attendance totals from the first half of the season (week 1-8) to the second half of the season (week 9-17). By making this chart we were able to visualize which teams had an increase or decrease in attendance in the second half of the season. After examining which teams had the biggest decrease we were able to look at our other chart and see what those teams' win totals were for the season. Pointing out a couple of teams that had a big increase in attendance, we noticed that Philadelphia was a team that had a significant increase in attendance in the second half of the season, when looking at their win total for the season they were amongst one of the more successful teams in the league with a win total of 11. Examining teams that showed a decrease in attendance for the second half of the season, Arizona was one of the more significant decreases in the league. Looking at the win total of Arizona, they were amongst some of the least successful teams in the league with a win total of 3 wins.

****

Bar chart showing the difference in the # of turnovers in a win vs. the # of turnovers in a loss for the amount of games in the 2000 season

We decided to make this chart to try to find a trend of the number of turnovers in a win vs the number of turnovers in a loss. This bar chart shows two different X variables and the side by side analysis of turnovers in a win as well as turnovers in a loss. There is a trend here showing how less turnovers correlates to the idea of a win. The turnovers in loss have a bell or normal shaped distribution while the turnovers in win visually look right skewed. By looking at this this seems to have a high correlation of the less turnovers you have in a game means to a higher likelihood of you winning the game as well as the more turnovers you commit in a game the less likelihood of the result of the game for that team to end up losing. This leads us to believe this would support our question we asked in how the less turnovers you have the chances of you winning is higher and is something we want to explore more in depth.

****

Scatter Chart Showing the point differential of teams who made the playoffs (1) vs. the teams who did not make the playoffs (0).

By making this chart we wanted to see if point differential impacts if a team makes the playoffs or not. To make this graph we had to make a dummy statistic for the teams who made the playoffs and the ones who didn’t with “1” representing making the playoffs and “0” representing a failure to make the playoffs. Looking at the chart you are able to identify that every team that made the playoffs were able to have a positive point differential and most of the teams that failed to make the playoffs held a negative point differential. After looking at the chart, we can conclude that point differential does in fact have an impact on the chances of making the playoffs or not.

**Summary Statistics:**

1. What are the key takeaways?

Key takeaways is that there are a lot of different variables that can be correlated with each other and there is lots of analysis that can be done on multiple variables between each other. Another takeaway is that wins can be very predictable depending on what happens during the actual game itself.

2. What are the primary trends and patterns observed?

There were many patterns but the biggest one that sticks out to us is the one with turnovers and how they lead to a team winning or not winning. This was most evident to us in a bar chart we had created as it was made very clear that more turnovers leads to less overall wins. Another trend we observed was the point differential leading to a team making the playoffs or not. It appeared that there was a hard cut off in point differential in terms of making the playoffs. There was no significant team that had a negative point differential that actually ended up making the playoffs. The scatter chart clearly shows this as all the teams with the higher point differential appeared on the graph towards the top meaning they made the playoffs and vice versa.

3. Are there any unusual or unexpected findings?

Since the data is over the time from 2000 to 2019, it was unexpected to find that some teams did not exist or moved locations for the time that the data was recorded. This unexpected finding made us clean our data because it would have been difficult to find trends for the teams that got moved or added. Also, during our data visualization portion we were able to see that there were some teams that had very high win totals but experienced a decrease in attendance in the second half of the season. We also noticed that a handful of teams had fairly high point differentials but failed to reach the playoffs.

4. Are there any areas that require further investigation or refinement?

We can look at the attendance more between the first and second half of the season and see if we can find more variables that would affect the numbers to either rise or fall. Another thing we can look at for refinement is looking at teams that had a positive point differential and didn’t end up making the playoffs so just diving deeper into that and seeing if we can find something that would make that make sense. Lastly another are for improvement is to get a bigger sample size and figuring out how to make the data easier to work with so it would allow us to look at multiple seasons instead of just one in case there was a “fluke” season or something that happened that year to cause different outcomes that wasn't recorded to get more accurate results.