Tej Seth
SI 330
97765649

NFL Home Field Advantage_Effects

I. Motivation

The overarching theme of this project was to analyze the effects of home field advantage in the NFL in the 2017 to 2021 seasons. The first goal was to see if there was a difference between home and away performance for a team's offense in a singular season. The next goal was to see if having a bigger attendance led to a bigger difference in home and away splits. Another goal was to see if there was a significant difference in the way defenses performed between home and away and if attendance made a difference there as well. My initial hypothesis was that home-field advantage in today's league doesn't impact the offensive side of the ball, but because the crowd usually tries to make an impact on the defensive side, some effects of home field advantage would be heightened.

II. Data Sources

nflfastpy: The nflfastpy is a python package for easily loading NFL play by play data via the nflfastR package that is maintained by Ben Baldwin, Tan Ho and Sebastin Case. nflfastpy is available on PyPl and supports Python 3.6+. The package can be loaded by running "pip install nflfastpy" and "import nflfastpy" can load the package before running "nflfastpy.load_pbp_data(2021)" to load play-by-play data into a pandas dataframe. The important variables that were used were:

- **"home_team"**: the team that was at home
- **"posteam"**: the team with possession
- **"epa"**: the expected points added on each play - an advanced metric that calculates the expected points based on yardline, down, distance and time remaining and then assesses if a singular play increases or decreases the offense's chance of scoring
- **"season"**: what season the play took place between 2017 and 2021

Additionally from the nflfastpy package, a dataset called "teams_logos_colors" could be loaded in that had a team's primary and secondary colors as well as their logo for plotting.

ESPN NFL Attendance: The attendance data hosted on ESPN's website under "ESPN NFL Attendance" had to be scraped using read_html with just the first table being selected due to there being multiple tables on the site. The years were scraped individually from 2017 to 2021 and then binded together. For example, the table from 2021 was from http://www.espn.com/nfl/attendance/_/year/2021. The most important variables from the web-scraped tables that became pandas dataframes were:

- **"TEAM"**: The city name of the team
- **"AVG"**: The average attendance for a team's home games that season. Average attendance was used instead of "TOTAL" because in past seasons some teams played a home or away game in London and in 2021 there were 17 games instead of 16 so some of the totals would be skewed based on teams having one more or less home game.

III. Data Manipulation

To manipulate the data, the first thing I needed to do was load in packages such as pandas, nflfastpy and numpy as well as the packages needed for plotting such as pyplot from

matplotlib and seaborn. The next step was loading in the play-by-play data for each season from 2017 to 2021 and then using concat to put all these seasons together. The play-by-play data has every regular season and postseason play on offense, defense and special teams for each team and is a very column-rich but row-poor dataset compared to other sports or datasets in general. The same process was repeated for scraping the attendance data from ESPN's website and then binding those together as well.

The first problem that I ran into was the "posteam" column in the play-by-play data didn't match up with the team names from the ESPN attendance records. This is because ESPN didn't use team abbreviations but instead used the city that the team plays in. A function had to be created that converts each city into a team abbreviation and the apply method had to be used on the attendance dataset. The Los Angeles teams, the Chargers and Rams, had to be thrown out of the analysis because with both listed as just "Los Angeles" in the attendance dataset, there wasn't a way of knowing which Los Angeles team was which. Once all that was processed, the play-by-play data and the attendance data could be joined by team abbreviation.

The play-by-play dataset needed to be manipulated to make summary statistics for each team in each season. Since Expected Points Added (EPA) is available in the dataset and is shown to be a predictive statistic, it was the one chosen. The dataset was filtered to just include passes and runs, removing special teams from the equation and also any rows that had NaN values for EPA were also removed. The dataset was then manipulated to create a new column that indicates if the home team has possession or not. An offense's amount of plays at home and away as well as their home EPA/play and away EPA/play were created in a home dataset and an away dataset using "groupby" and "agg" and then were joined using "merge" on team abbreviation and season so that home EPA/play and away EPA/play could be compared. If home field advantage was an actual thing, we'd expect most teams to have a higher home EPA/play then away EPA/play and the results of that will be shown in Section IV. To make a label for each team in each season, the two columns were added together with "astype(str).str.zfill(1)" to make it into a string.

Now that I had offensive EPA/play at home and on the road as well as the difference between those two values and it was joined with attendance records, we could make plots with the data. "sort_values" was used to sort from highest difference in home EPA/play and away EPA/play to lowest and the seasons that involved smaller attendance because of COVID (2020 and 2021) were filtered out. "plt.style.use(fivethirtyeight)" was used to make the graph a gray background and "plt.bar" was used to make the bar graph. In every plot, "plt.text" was used for the title and subtitle. This is because the function "plt.suptitle" was putting the subtitle above the title so the problem was solved by using plt.text and changing the x and y parameters so that it lined up in the middle and was above the plot. Next, a boxplot was created by taking each team abbreviation and using "pivot" to make each abbreviation its own column and then sorting the index using the mean of each column. Using the "teams_colors_logos" dataset, each team had their primary color line up with them using the palette argument of "sns.boxplot". After that, two scatter plots were created with one including the seasons with lower attendance because of COVID and the others just having 2017 to 2019 with both teams colors and labels. An issue that I ran into was when I tried to plot the team logos, the way that the nflfastpy package used to recommend to plot images was currently outdated on the most recent version of python and the "Pillow" package didn't install on my jupyter notebook so team colors had to be used instead of team logos. Everything listed above that was done on the offensive side of the ball was then repeated on the defensive side of the ball.

The last plot that I wanted to create was taking quarterbacks from 2017 to 2021 and seeing how strong the relationship between performance at home and performance on the road. The first step with this was getting a list of quarterbacks with at least 1,000 passes in that timeframe so passes in the regular season were filtered and then "groupby" was used with "agg" counting the amount of plays which was then converted into a list using "tolist" so that I could take the original play-by-play dataset and filtered just when those player ID's were involved. Then, following a similar process to the offense and defense the quarterbacks had their home and away splits created and were joined using "merge" as well as a dataset created that had the team that each quarterback threw the most passes so that their primary color could be joined using the team abbreviation. Lastly, the plot was made using "axvline" and "axhline" to show the averages at both the home and away. All the plot results will be shown below.

IV. Analysis and Visualization

As mentioned in Section III, if home field advantage still existed in the NFL we'd expect most teams to have a higher EPA/play at home than on the road on offense. However, on offense the average difference between home EPA/play and away EPA/play for each team in each season is -0.02 with only 39% of teams performing better at home than on the road. This is reflected in Figure 1.

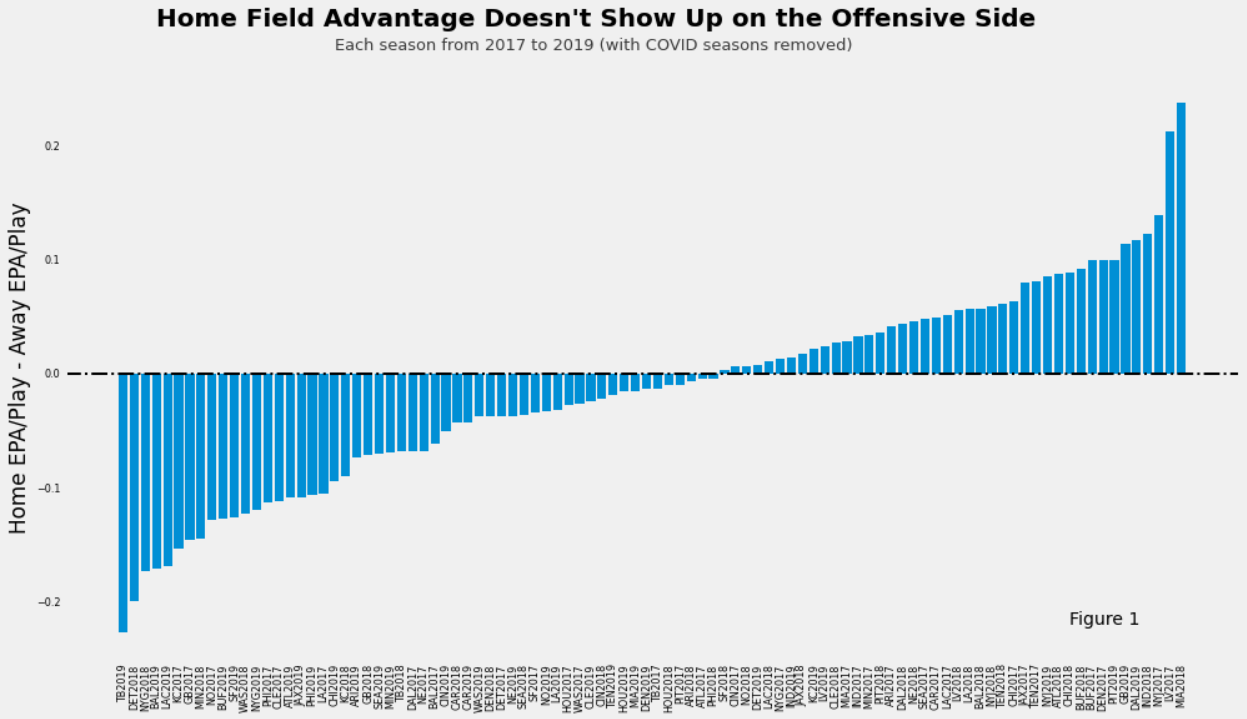
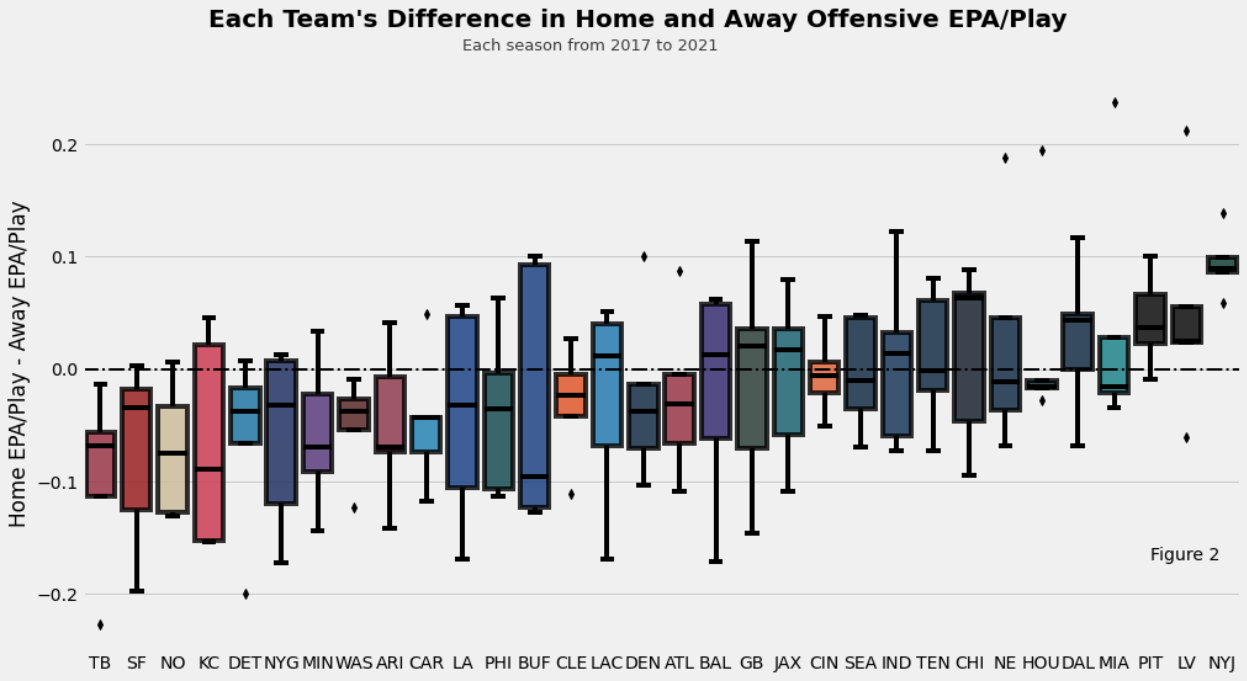
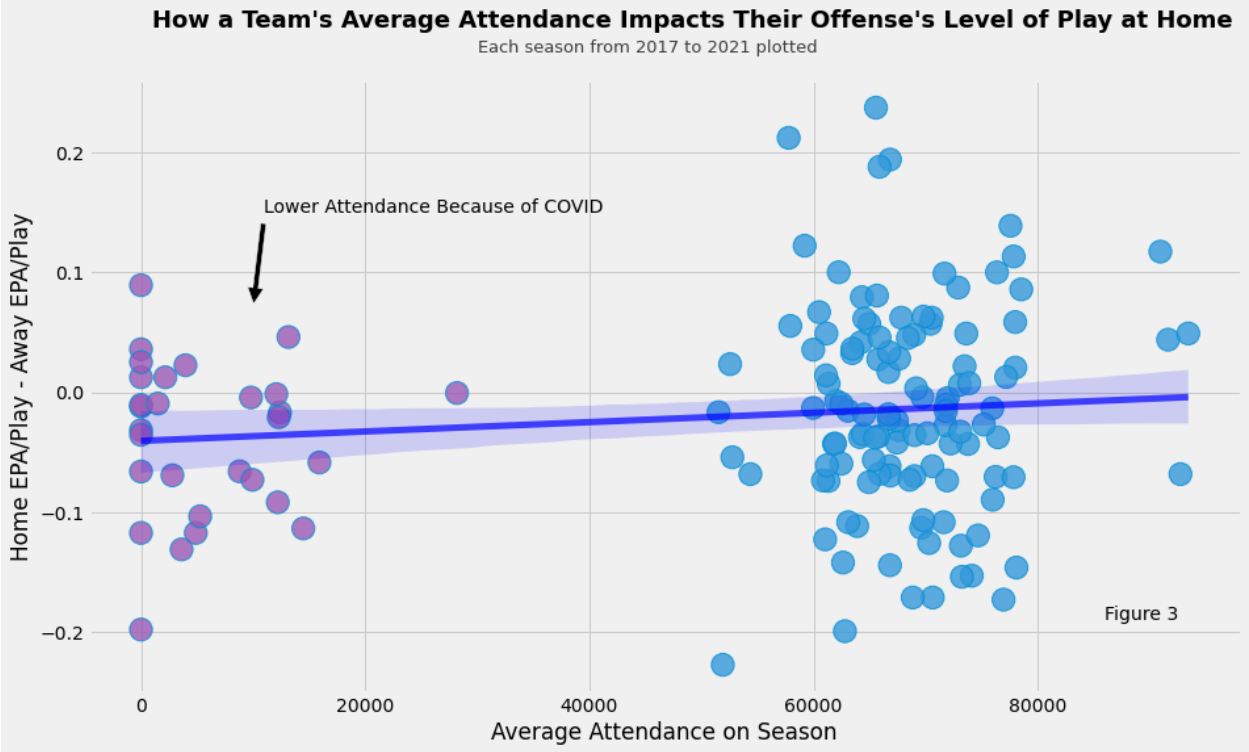


Figure 1 shows the difference between each team's home offensive EPA/play and away offensive EPA/play for each season from 2017 to 2019. If the team is above the line then they were better at home and if they were below then they were better on the road. If anything, this would suggest offenses actually are slightly better on the road but with only eight game sample sizes for each this can probably be chalked up to being noise. Increasing the sample

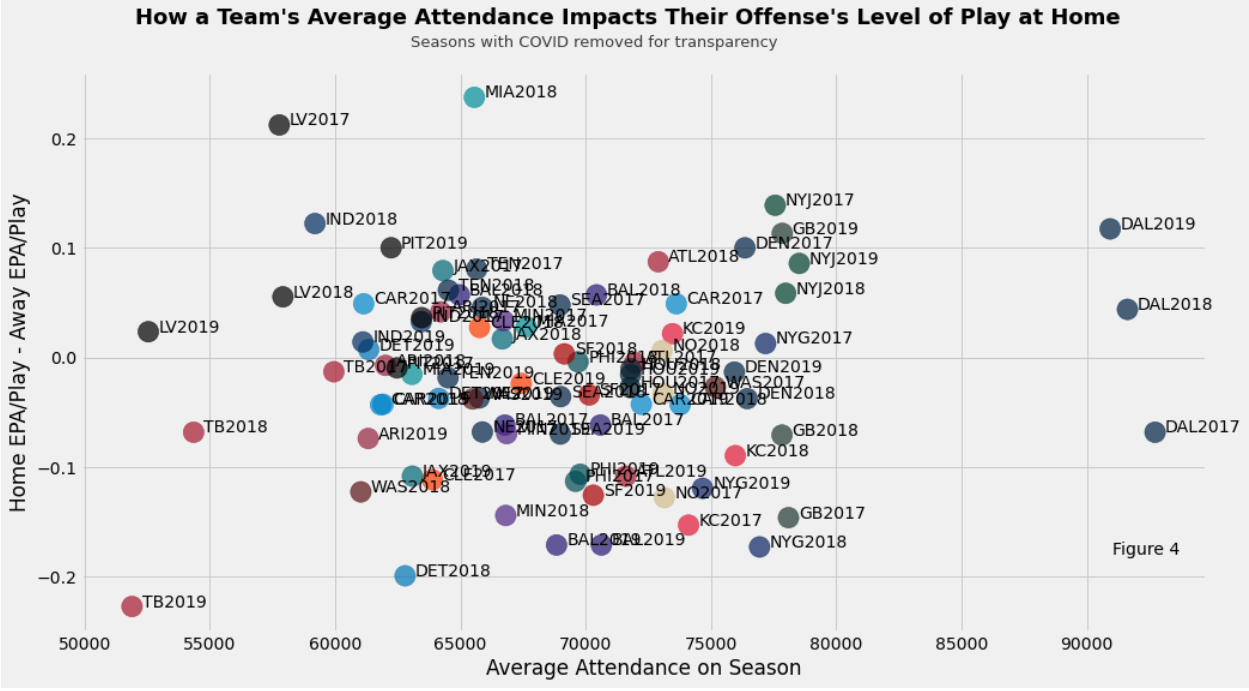
size, we can look at a boxplot of the difference in home performance and away performance for each team to see if each team's home field advantage on offense is stable year-to-year in Figure 2.



While some teams like the Washington Commanders, Houston Texans and New York Jets have very small interquartile ranges, the majority of teams have their 75th percentile have the 0.0 dotted line and their 25th percentile below meaning that home field advantage doesn't stay too consistent for teams between 2017 and 2021. Another argument can be made that having more attendance will increase a team's home-field advantage because there's more fans there to cheer them on and intermediate the referees, however Figure 3 disproves that.



Even removing the seasons where attendance was lower or none because of COVID, there is no relationship from 2017 to 2019 as evidenced in Figure 4.



Despite all of this evidence that playing at home doesn't help the offense perform, any football fan would anecdotally mention that home-field advantage seems to have a bigger impact on the defense because that's when the fans get loud and try to create havoc on the opposing team's offense. However the average difference between defensive EPA/play at home vs. defensive EPA/play on the road is -0.01 with only 46% of defenses having better performance at home than on the road. This is shown in Figure 5.

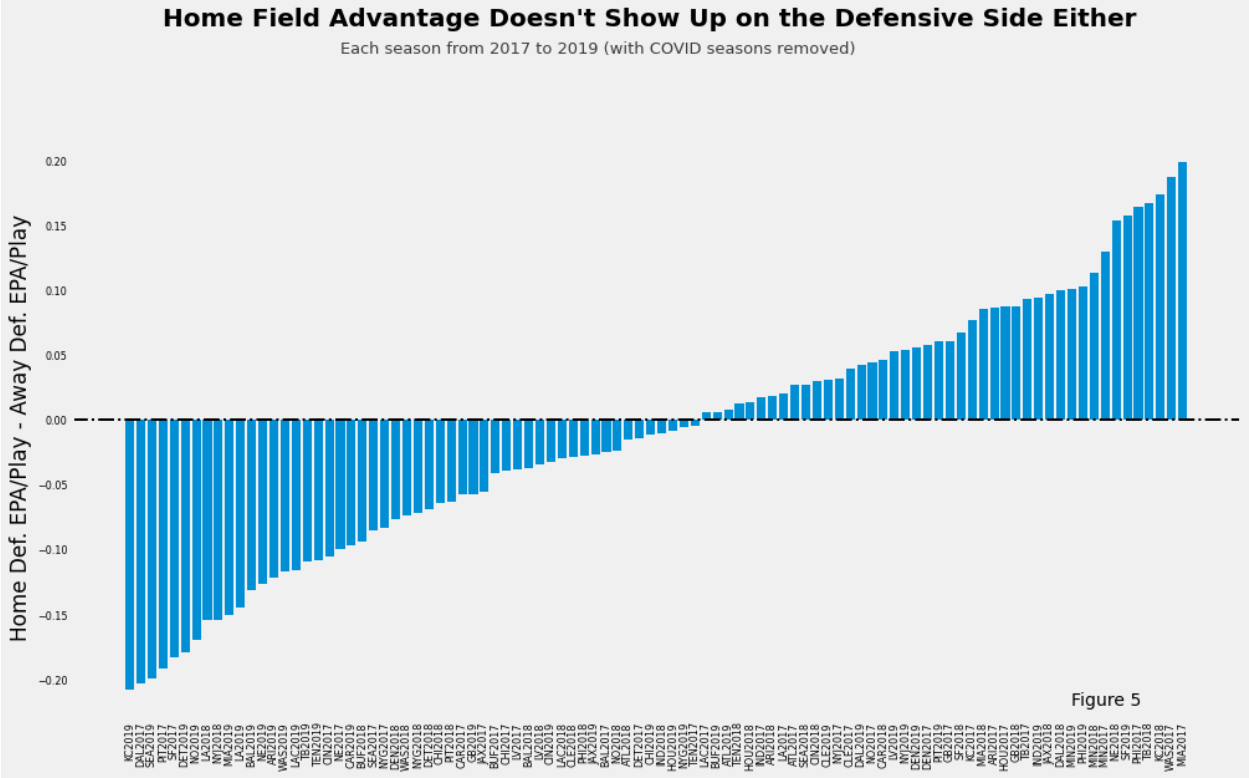
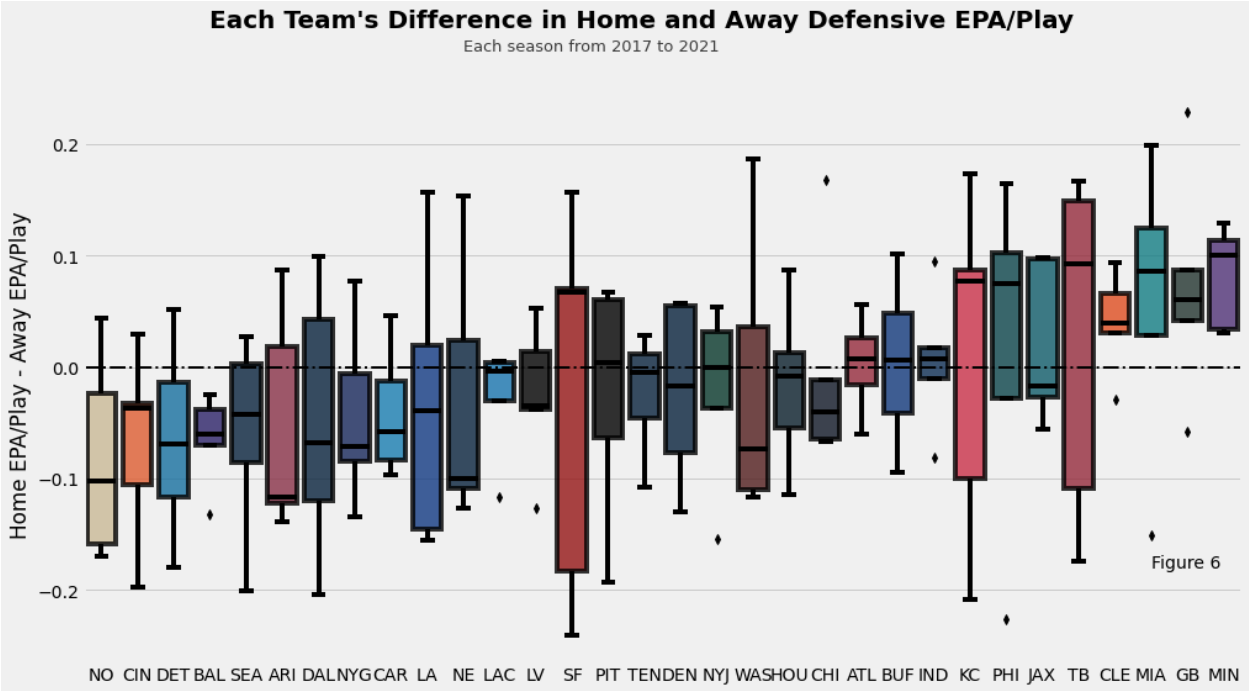
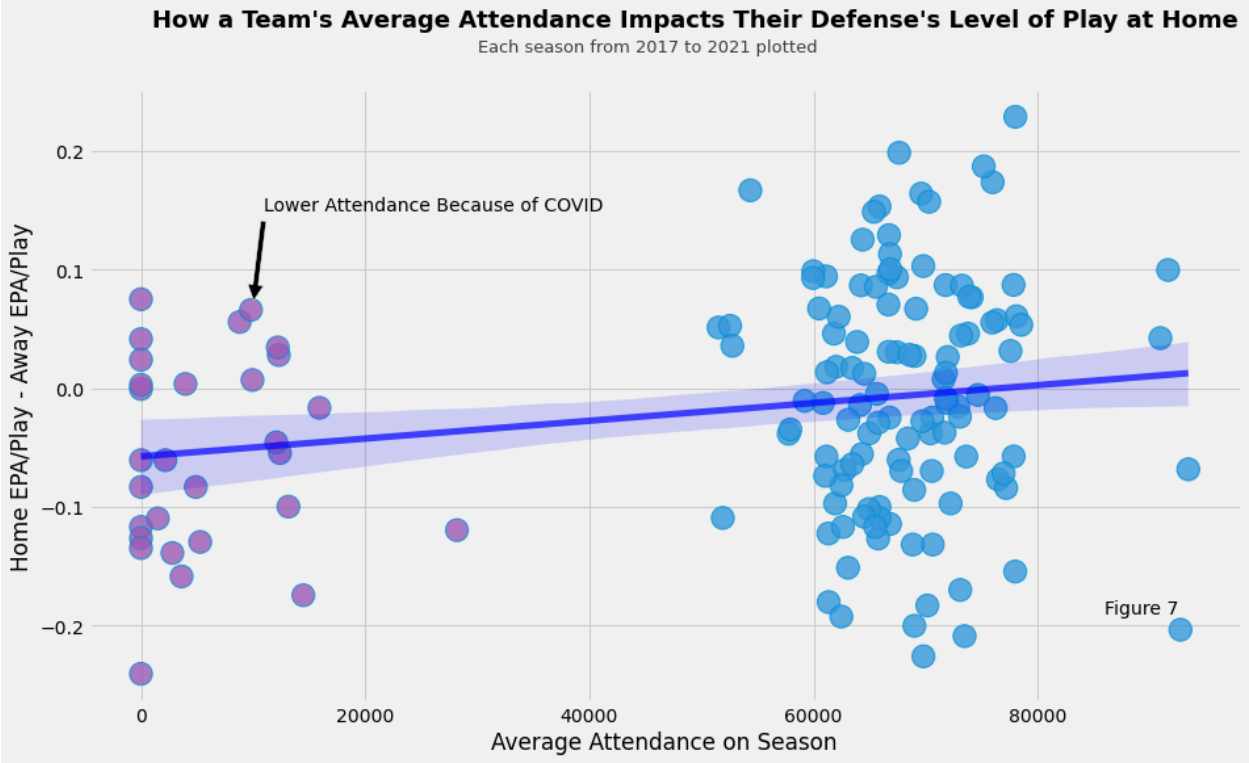


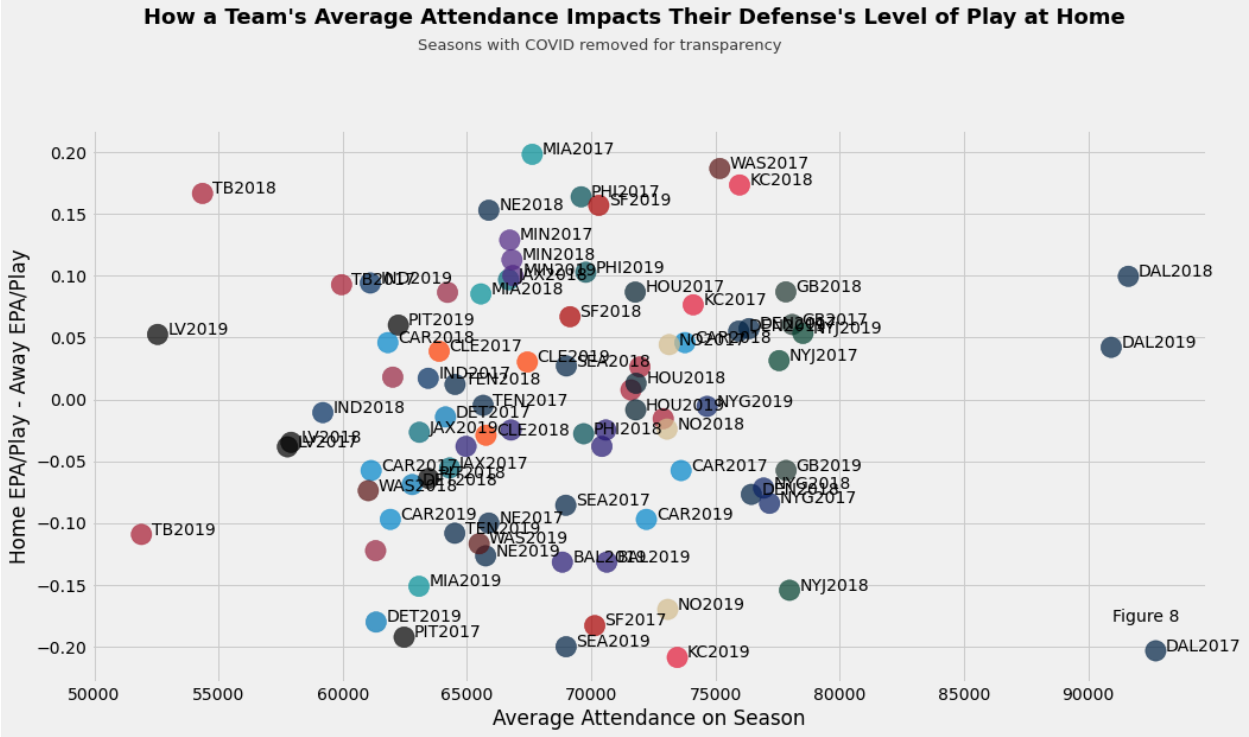
Figure 5

Figure 6 has the same boxplot results as Figure 2 did on the offensive side of the box and even shows less consistency in regards to home-field effects on defense.

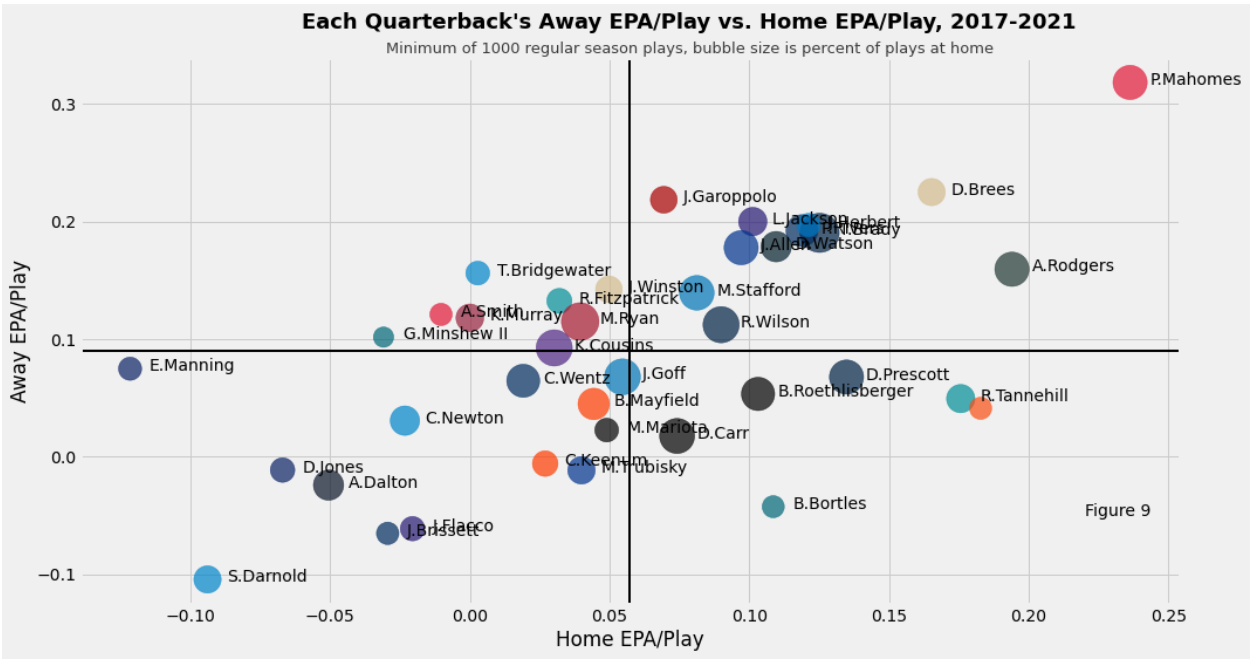


When looking at how attendance can impact how a defense performs at home compared to on the road, there does appear to be more of a relationship than on the offensive side. This goes back to our anecdotal theory that the fans make more noise and seem to have more impact when their defense is on the field so more fans being there should hurt the other team's offense more as seen in Figures 7 and 8.





Lastly, we can analyze how quarterbacks perform at home vs. on the road to see if there is a strong correlation between the two or if the majority of them do better at home. Figure 9 shows that most quarterbacks are either both below-average at home or on the road or above-average with some exceptions.



V. References

Code: https://github.com/tejseth/si-330-home-field

nflfastpy: https://github.com/fantasydatapros/nflfastpy

ESPN NFL Attendance: http://www.espn.com/nfl/attendance/_/year/2021