School of Information Studies
Syracuse University

THE EXTRA POINT

IST 652 FINAL PROJECT

Professor Landowski
Amanda Norwood

TABLE OF CONTENTS

BACKGROUND

Our final project is based on NFL statistics and we analyzed datasets from Github and Kaggle. The goal is to provide various stakeholders such as NFL Owners, General Managers, Scouts, or Players with analysis to help them make an informed decision for their organization.

DATA SOURCES

NFL_GAMES

This primary data set contains information on each NFL game played between 1997 to present. There are 6,409 rows and 45 columns. It contains information about each game played and the conditions it was played in.

NFL_STATS

The second data set contains statistics on NFL games since 2019 to present. There are 895 rows and 21 columns. This dataset is smaller than the primary but it contains important data such as total yards, passing yards, and pushing yards.

NFL_TEAMS

The third data set pulls in the team names, 3 letter team code, the conference, and the division. It contains 32 rows and 4 columns.

DATA IMPORTATION

Python programming language was used for this project. We used the read_csv command within pandas to read the csv's into data frames.

DATA CLEANSING

NFL_GAMES

- The column names were simple and easy to decipher, therefore they were not renamed.

- A few relevant columns we retained and renamed:
    a. GAME_ID
    b. SEASON
    c. GAME_TYPE
    d. WEEK
    e. AWAY_TEAM
    f. AWAY_SCORE
    g. HOME_TEAM
    h. HOME_SCORE
    i. RESULT
    j. TOTAL
    k. ROOF
    l. SURFACE
    m. TEMP
    n. WIND
    o. AWAY_QB_NAME
    p. HOME_QB_NAME
- Columns removed due to irrelevance:
    a. OLD_GAME_ID
    b. GAMETIME
    c. AWAY_QB_ID
    d. HOME_QB_ID
    e. GSIS
    f. NFL_DETAIL_ID
    g. PFR
    h. PFF
    i. ESPN
    j. AWAY_MONEYLINE
    k. HOME_MONEYLINE
    l. SPREAD_LINE
    m. AWAY_SPREAD_ODDS
    n. HOME_SPREAD_ODDS
    o. TOTAL_LINE
    p. UNDER_ODDS
    q. OVER_ODDS
    r. REFEREE
- Columns added for analysis:
    a. INDOOR_OUTDOOR: this column is a boolean field based off the ‘roof’ column in the dataset. It categorized indoor stations as 0 and outdoor stadiums as 1.
    b. WL: a ‘W’ or ‘L’ for win or loss based on the ‘result’ column
- We also removed ‘null’ results, which means the game has yet to be played.

NFL_STATS

- Relevant columns we retained and renamed:
    a. W/L
    b. AWAY_TEAM

c. AWAY_SCORE
d. HOME_TEAM
e. HOME_SCORE
f. TOTYARDS
g. O_YARDS (OPPONENT)
h. PASSYDS
i. O_PASSYDS
j. RUSHYDS
k. O_RUSHYDS
l. YEAR

NFL_TEAMS

- Cleansing was not performed on this dataset as it had proper headings and no missing data.
- Columns:
    a. NAME
    b. ABBREVIATION
    c. CONFERENCE
    d. DIVISION

NFL_GAMES_SURFACE

- The dataset is based on the games NFL_GAMES dataset explained above.
- Data cleansing was performed to remove null values from SURFACE and renamed various turf types to a single value 'Turf' for analysis purposes

NFL_STANDINGS

- Relevant Columns from this dataset are given below:
    a. Season
    b. Conf
    c. Division
    d. Team
    e. Wins
    f. Losses
    g. Ties
    h. Pct
    i. Div_rank
    j. Scored
    k. Allowed
    l. Net
    m. Sov
    n. Sos
    o. Seed
    p. Playoff
- As part of cleansing, removed null values from Playoff(i.e. teams who did not qualify for the playoff)

FIELD GOAL STATISTICS

- Relevant Columns from this dataset
    a. Player Id
    b. Name
    c. Year
    d. Team
    e. Kicks Blocked
    f. Longest FG Made
    g. FGs Made
    h. FGs Attempted
    i. FG Percentage
    j. FGs Made 20-29 Yards
    k. FGs Attempted 20-29 Yards
    l. FG Percentage 20-29 Yards
    m. FGs Made 30-39 Yards
    n. FGs Attempted 30-39 Yards
    o. FG Percentage 30-39 Yards
    p. FGs Made 40-49 Yards
    q. FGs Attempted 40-49 Yards
    r. FG Percentage 40-49 Yards
    s. FGs Made 50+ Yards
    t. FGs Attempted 50+ Yards
    u. FG Percentage 50+ Yards
    v. Extra Points Attempted
    w. Extra Points Made
    x. Percentage of Extra Points Made
    y. Extra Points Blocked
- As part of Data Cleansing, removed null values and scoped is updated to include seasons from 2001. (Excluded old data starting from 1933)

DATA PREPARATION

- NFL_GAMES used the abbreviated team name and NFL_STATS used the full team name. We used the NFL_TEAMS in order to merge and join the two data sets.
    a. Merging the NFL_GAMES and NFL_STATS allowed us to see the result of the game and also the offensive statistics of the game. However, NFL_STATS was limited to games from 2019 to present.
- Indoor stadiums report their temperature and wind as 'NaN'-- so we filled in that data as 0.
- NFL_GAMES shows score results by home and away team. We created dataframes for home and away teams. Then merge them together so we can further analyze how each term performs when they are away and while at home.
- NFL_GAMES and NFL_STANDINGS were merged together to get a holistic view of the games and playoff statistics

LIBRARIES USED FOR ANALYSIS

- Pandas
- Matplotlib
- Numpy
- Seaborn

BUSINESS QUESTIONS

By doing extensive analysis on the NFL data, we would like to provide the game lovers with statistics and information that potentially impact game outcome and decision making in various areas.

- ➔ Stadium Type
- ➔ Surface Type
- ➔ Environmental Factors such as wind and temperature
- ➔ Special Teams Training Needs
- ➔ Game Predictions

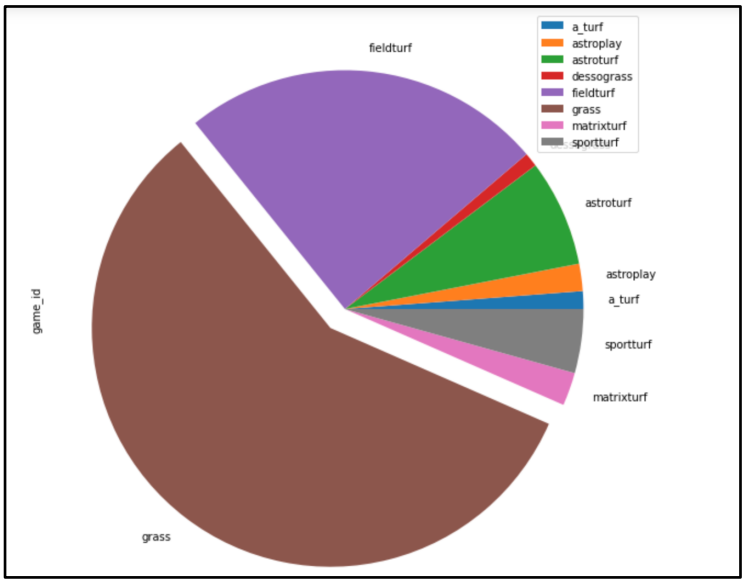
These were addressed by exploring the below questions:

- How do playoff teams compare to each other in terms of wins, losses, score, points allowed, and net points?
- Sample Career Statistics of players/teams(eg: field goals attempted vs scored)
- What is the significance of home field advantage
- How does temperature and wind speeds impact the game score?
- How do turf fields compare to grass fields?
- Is there home field advantage?
- Do teams perform in different roof types?
- How do home teams with an indoor stadium perform in outdoor stadiums?
- How do home teams score at home?
- Who are the top passing quarterbacks?

ANALYSIS & RESULTS

SURFACE TYPE: TURF/GRASS

- The main purpose of this analysis is to understand the impact of the ground surface on a team's performance and will that be the deciding factor when a new stadium needs to be constructed?
- For this analysis , we used the NFL GAMES dataset(metadata is explained in the above section.
- Some statistics derived from the analysis are given below:
    a. Overall games played on each surfaces:



- This shows that a significant number of games were played on grass. To make the analysis easier we divided the surface types broadly into two (Grass and Turf). Turf represents all varieties of non-grass surfaces.
- Observation: The percentage win on each surface was very close which led to a conclusion that there is no significant impact on a Team's performance with the surface type.
- Some additional Data Exploration is done by looking at the top performers on each of these surfaces.
- Top Grass Performers list is given below:

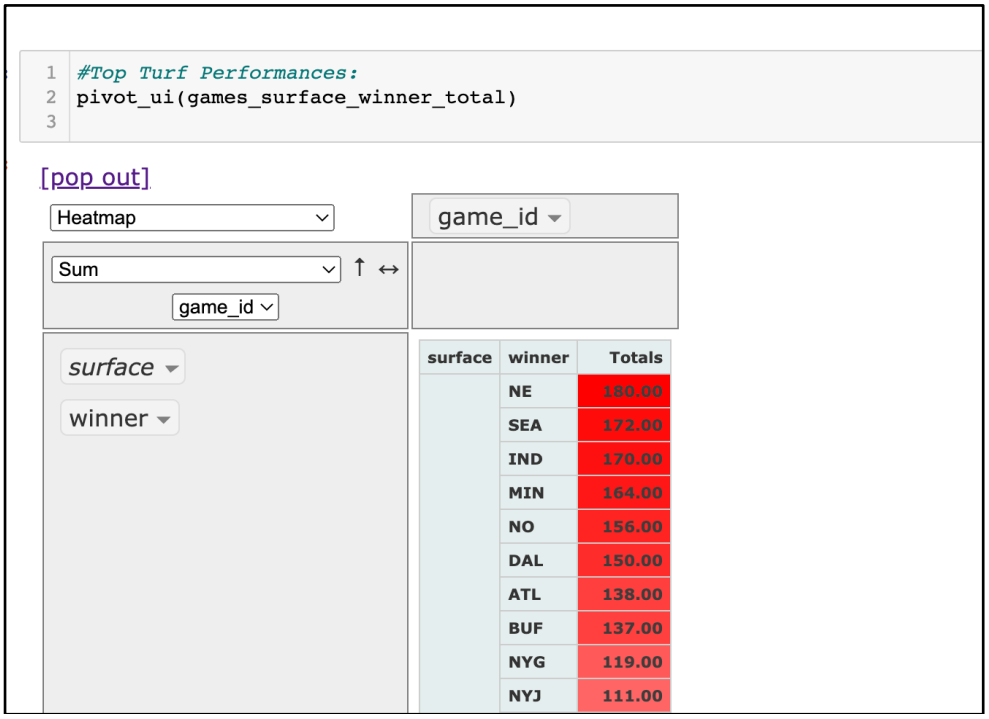
```
# Grass Top Performers
pivot_ui(games_surface_winner_total)
```

[pop out]

Heatmap | game_id | Sum | game_id | surface | winner

surface	winner	Totals
	GB	204.00
	KC	182.00
	DEN	181.00
	TEN	173.00
	PHI	156.00
	PIT	156.00
	CAR	148.00
	MIA	148.00
	SF	147.00
	TB	143.00
	CHI	140.00
	ARI	135.00

- Top Turf Teams are given below:



```
#Top Turf Performances:
pivot_ui(games_surface_winner_total)
```

[pop out]

Heatmap | game_id

Sum | game_id

surface | winner

surface	winner	Totals
	NE	180.00
	SEA	172.00
	IND	170.00
	MIN	164.00
	NO	156.00
	DAL	150.00
	ATL	138.00
	BUF	137.00
	NYG	119.00
	NYJ	111.00

- Additional Research Scope:
- Though the surface type has no direct impact on the performances, surface type can really contribute to player safety. As an additional research, we would like to obtain the injury report and analyze it further to arrive at an informative decision on which surface type to promote for new/existing stadiums

PLAYOFF TEAMS/HOME FIELD ADVANTAGE DURING PLAYOFFS

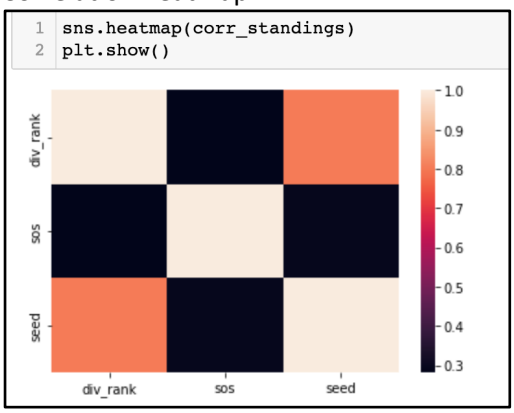
- For this analysis, we used the NFL STANDINGS file from season 2002 through 2019 with some data cleansing to remove the team details when the teams did not reach the playoffs.
- The main purpose of this analysis is to understand
    - How the regular season wins impact the playoff seed/div_rank.
        - To understand the relation between regular season wins and seed, we used a correlation study. Based on the correlation table presented below:

```
#correlation between seed, netpoints, divisional rank, and win rate
corr_standings=standings_df2[['div_rank','sos','seed']].corr()
corr_standings
```

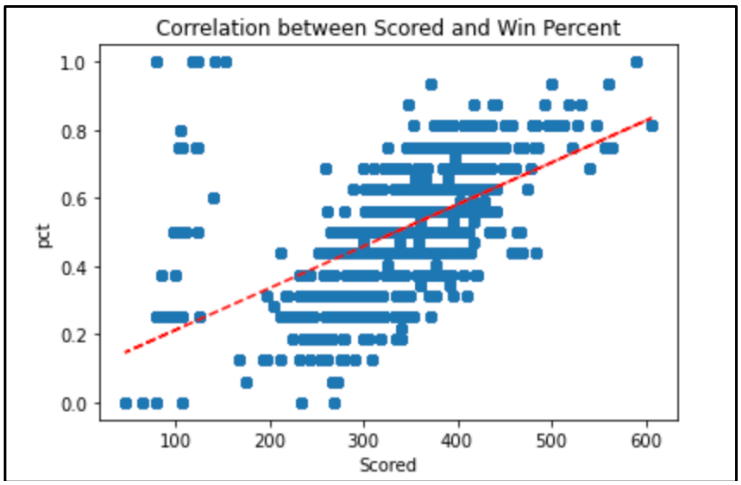
	div_rank	sos	seed
div_rank	1.000000	0.282809	0.803621
sos	0.282809	1.000000	0.291761
seed	0.803621	0.291761	1.000000

Correlation heatmap:

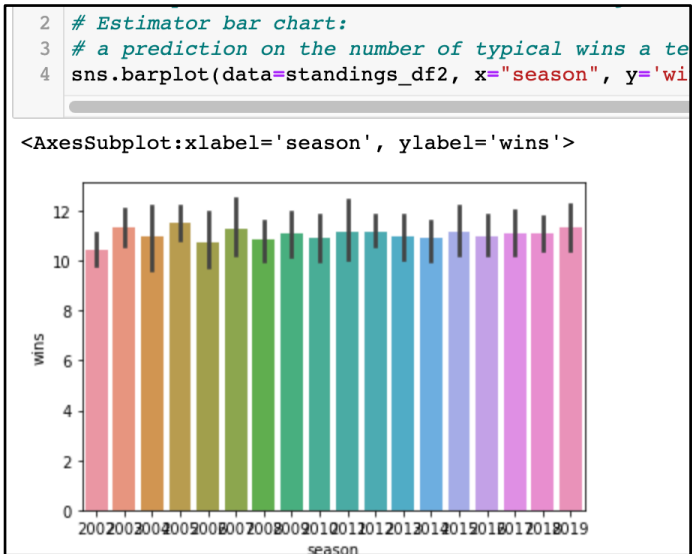
```
sns.heatmap(corr_standings)
plt.show()
```



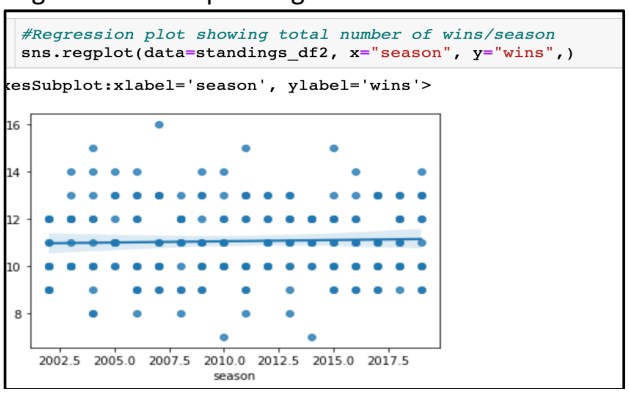
- Correlation represents the direction of movement of values/relation between the two variables. A positive value indicates that direction of movement is positive. In this scenario, SOS represents the win rate of a team when they played in the regular season amongst the group. The positive value indicates, as the win rate goes up the seed also goes up.
- Observations:
    - Positive Correlation between seed and win rate(strength of schedule). Higher wins will improve the seed/division rank. As the higher seeds enter the playoffs, the teams will receive a home field advantage. However, the superbowl is not impacted by this as the venue will be decided well in advance of a season.
- The approximate number of wins needed by a team to secure a place in playoffs
    - By exploring the estimator bar chart and regression charts, we figured out that, if a team secures approximately 11 or 12 regular season wins, it could secure a playoff spot.
- The details are as shown below:
- Correlation between scores and wins. The higher the scores(more touchdowns!!) the chances of winning are higher.



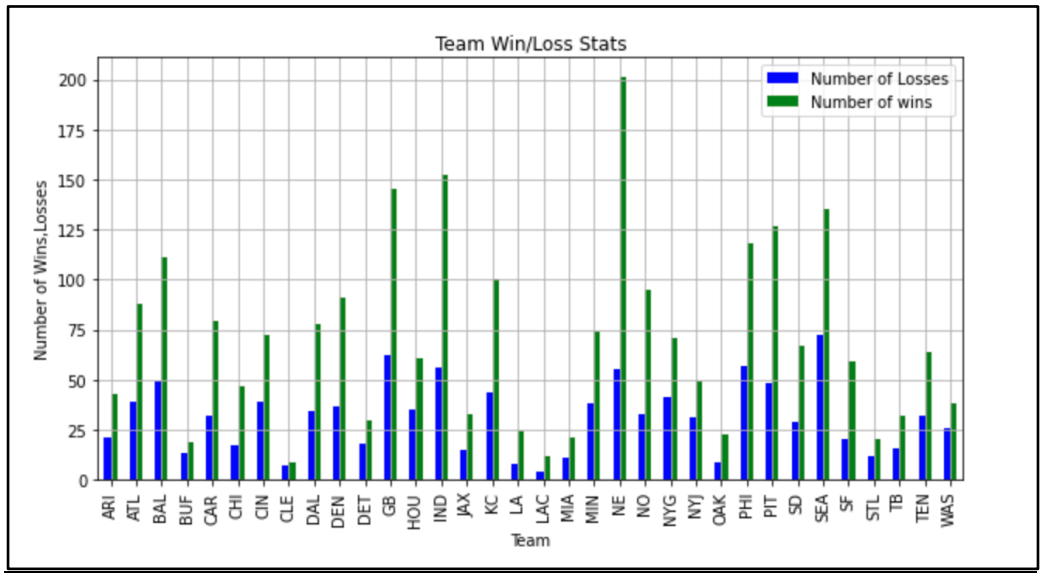
Predictor chart for the number of wins needed in regular seasons



Regression chart pointing to a similar result:



<u>Overall Team performance during the Regular Season:</u>



- Additional Research Scope:
    - To understand home field advantages in depth, we should include the crowd details present in the stadium(whether it was full attendance) and the percentage of fans from both the teams.

SPECIAL TEAMS(FIELD GOAL KICKERS)

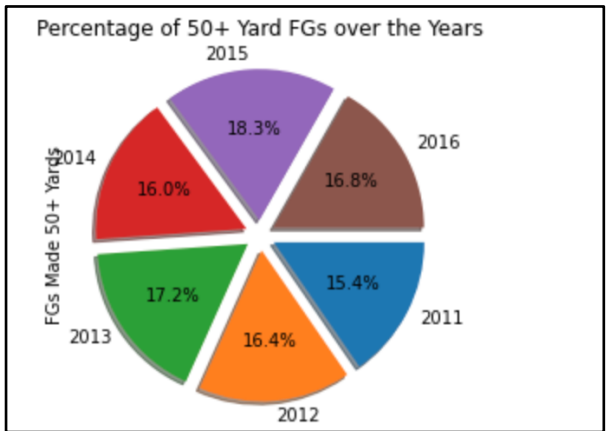
- Main purpose of analyzing field goals data is to understand the contribution of field goal kicking teams to the overall team performances. Field Goals are considered as the last attempt to score from a possession. By analyzing the trend we aim to see the number of longest goals attempted.
- The dataset is sourced from Kaggle and contains data from 1933 through 2016. For our analysis we used a subset of data from the year 2001 through 2016.
- Who scored the Longest Field Goal(from the data in scope)?

```
# Who made the longest Field Goal??
longest_FG_details=(kickers_df2[kickers_df2['Longest FG Made'] == kickers_df2['Longest FG Made'].max()])

print ('Team:',longest_FG_details['Team'].to_string(index=False),'\n',
       'Year:',longest_FG_details['Year'].to_string(index=False),'\n',
       'Player Name:',longest_FG_details['Name'].to_string(index=False),'\n',
       'Total Distance:',longest_FG_details['Longest FG Made'].to_string(index=False))
```

```
Team: Denver Broncos
 Year: 2013
 Player Name: Prater, Matt
 Total Distance: 64
```

Based on our data, Matt Prater kicked a 64 yards Field Goals for Broncos in 2013.

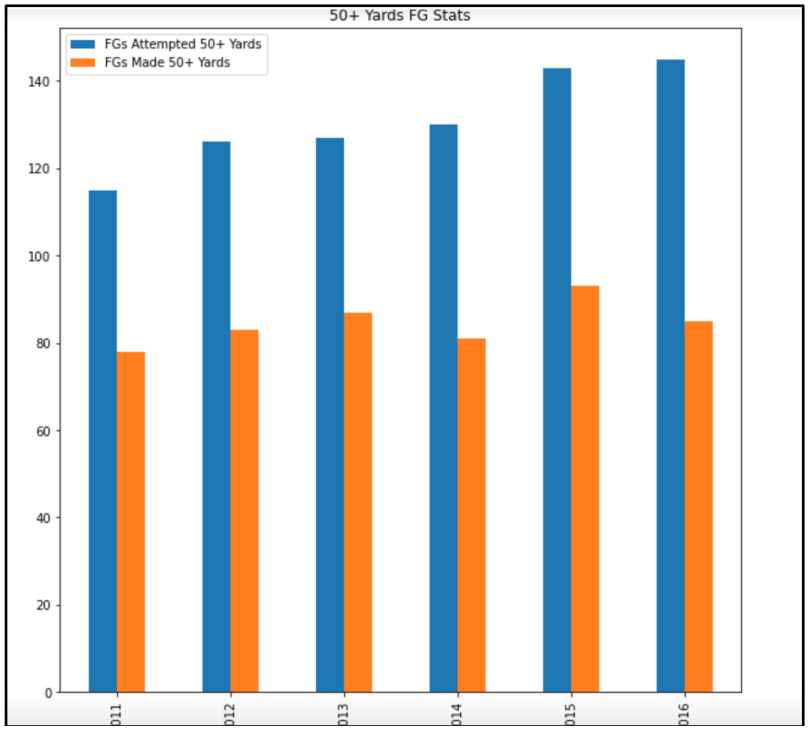
Pie chart showing the % of longest field goals(50+ yards) per season



Observation:

From the above 2015 had a little higher percentage of 50+ yard FGs

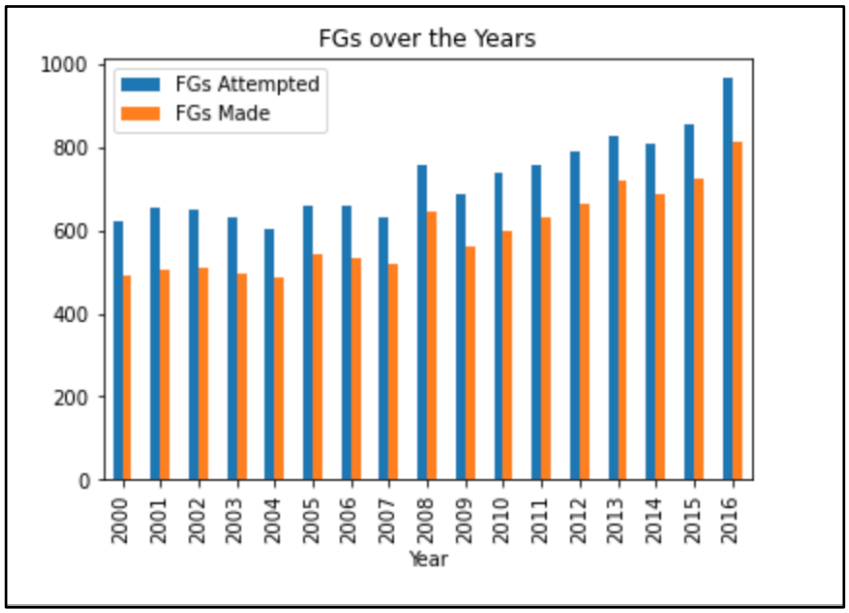
Chart comparing 50+ Yards FGs attempted vs Made



Observation:

The number of attempts shows an upward trend which in turn translates to the need of special teams to score more.

Overall Trend in Field Goals Attempted vs Made



- Observation:
  - As we can see from the above chart, an upward trend of total number of field goals attempted/made.
- Additional Research Scope:
  - The research shows an increasing attempt for field goals which can be supported by additional training to the team. Also this research can be enhanced to include other speciality teams performances over the years and their impact on overall performance of the team during a season.

EFFECTS OF WINDS ON THE HOME TEAMS

Here we will analyze the effects of weather, specifically the temperature and wind conditions. From the data it was determined that there are 31 venues. Out of 31 venues, 23 are open stadiums and the remaining 8 are fixed domes or retractable roofs. It is common intuition/assumption that weather conditions likely provide home field advantages. We will use the data to quantify and provide insights to the correlations.

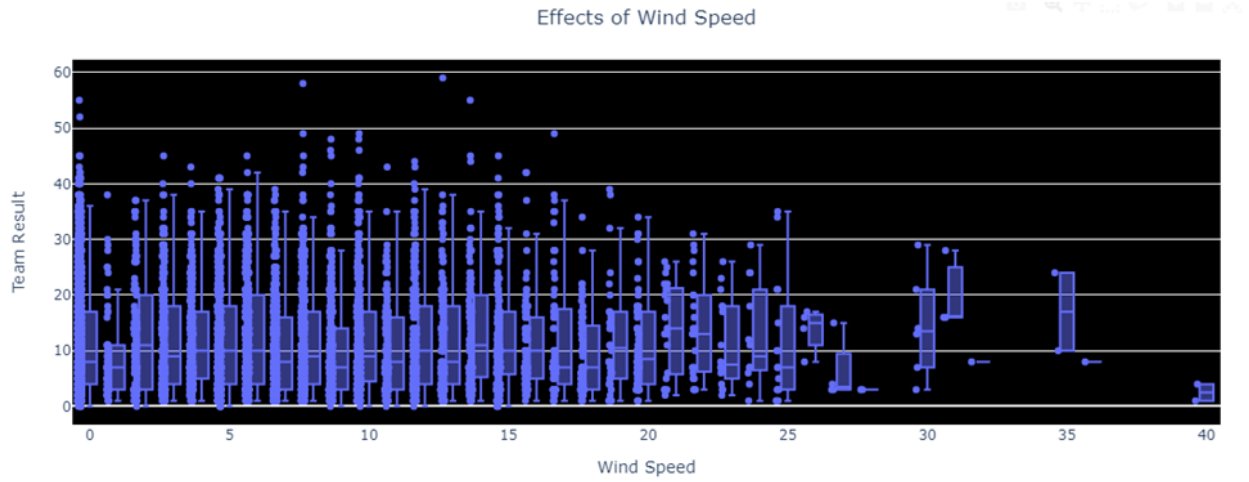
First, we begin by data cleanup efforts. Here we take the wind column and fill zeros in places where data might not be available (like closed stadiums)

```
games['wind'] = games['wind'].fillna(0)  # Apply to 1 column
#games = games.fillna(0)                 # this would apply to whole dataset

games.head(2)
```

Now we take a look at the scoring. First we will use the results column. Results column is the score of each team. Taking an absolute value of the result provides consistency in the charting. Following is the bar chart that shows results from each team versus the wind speed (in mph). As you can see there isn't a consistent breakout of data as

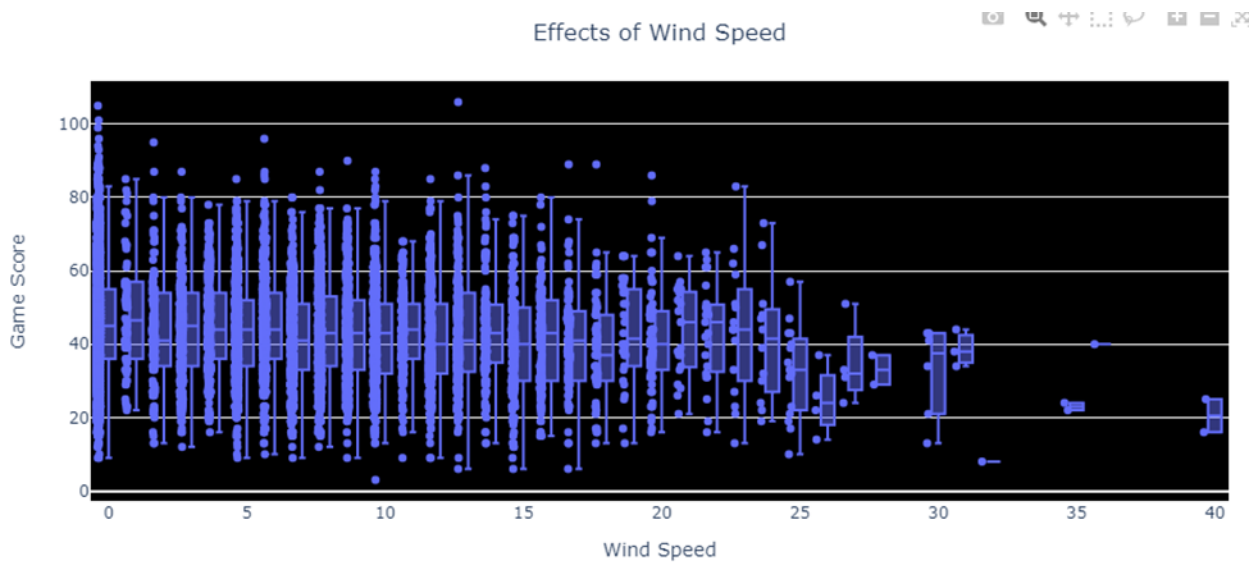
wind speeds increase. There is some fluctuation of data past wind speeds of 25 mph, however that may be contributed to not having enough data points.

```
# Effects of wind speeds - Team Scores per game.
# Box plot showing all points to see outliers
# This takes care of the outliers for wind       x= games.where(games.wind.abs() < 50)['wind']

games['abs_result'] = games.result.abs()
fig=px.box(games, x=games.where(games.wind.abs() < 50)['wind'], y='abs_result', hover_data=games.columns, points='all')
fig.update_layout(title={"text": "Effects of Wind Speed", "x": 0.5})
fig.update_layout(yaxis_title="Team Result")
fig.update_layout(xaxis_title="Wind Speed")
fig.show()
```



Secondly, we will now use the total score of each game (prior section looked at total score by each team). Superimposing the two charts together does not reveal any significant differences.

```
# Effects of wind speeds - Game Score per game.
# Box plot showing all points to see outliers
# This takes care of the outliers for wind       x= games.where(games.wind.abs() < 50)['wind']


fig=px.box(games, x=games.where(games.wind.abs() < 50)['wind'], y='total', hover_data=games.columns, points='all', width=1000)
# x is adjusted to remove outlier wind conditions

fig.update_layout({'plot_bgcolor': 'black','paper_bgcolor': 'white',})   #  Set background color

fig.update_layout(title={"text": "Effects of Wind Speed", "x": 0.5})    # Set Title
fig.update_layout(yaxis_title="Game Score")
fig.update_layout(xaxis_title="Wind Speed")
fig.show()
```



Now we will take the analysis further by categorizing the wind speeds and displaying on charts with colors. Following lines of code uses function to come up with a new variables that categorizes

1) Home versus Away win/loss.

2) Wind speed category (low, medium, high etc.)

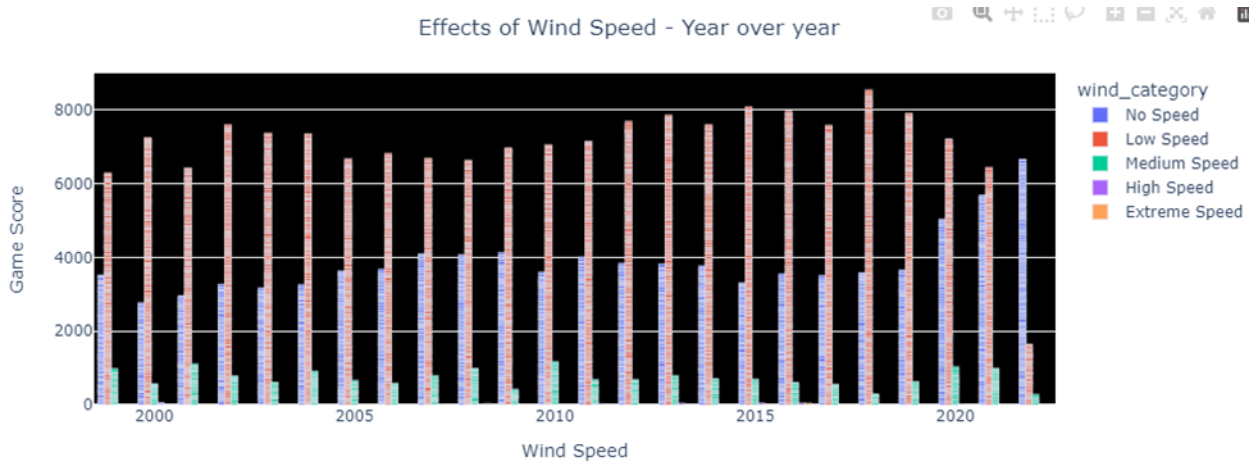
```
# Now we determine which team won based on results column
# (-) negative result shows home win lost
# (+) Positive result shows home team won
# This will add a new column to the df specifying win/loss for home team

games['result'].isna().sum()        # Look to see how many are nulls in the result column

def who_won(x):
    x = int(x)
    if x < 0 :
        return 'Away_team'
    elif x > 0 :
        return 'Home_team'
    elif x == 0:
        return 'Tie'

# Here we use (not = isna) and use function to determine if home team won/lost
games['who_won'] = games[~games['result'].isna()].result.apply(who_won)

games.head(1)
```

```
# Now we will categorize the wind speeds
# and create a variable (column) called wind_category

def wind_category(x):
    if x == 0 or x==None:
        return 'No Speed'
    elif x > 0 and x <= 15:
        return 'Low Speed'
    elif x > 15 and x <= 30:
        return 'Medium Speed'
    elif x > 30 and x <= 45:
        return 'High Speed'
    elif x > 45 :
        return 'Extreme Speed'

games['wind_category'] = games['wind'].apply(wind_category)
##verify the def worked


games = games.convert_dtypes()
#print(games['wind'].dtypes)
```

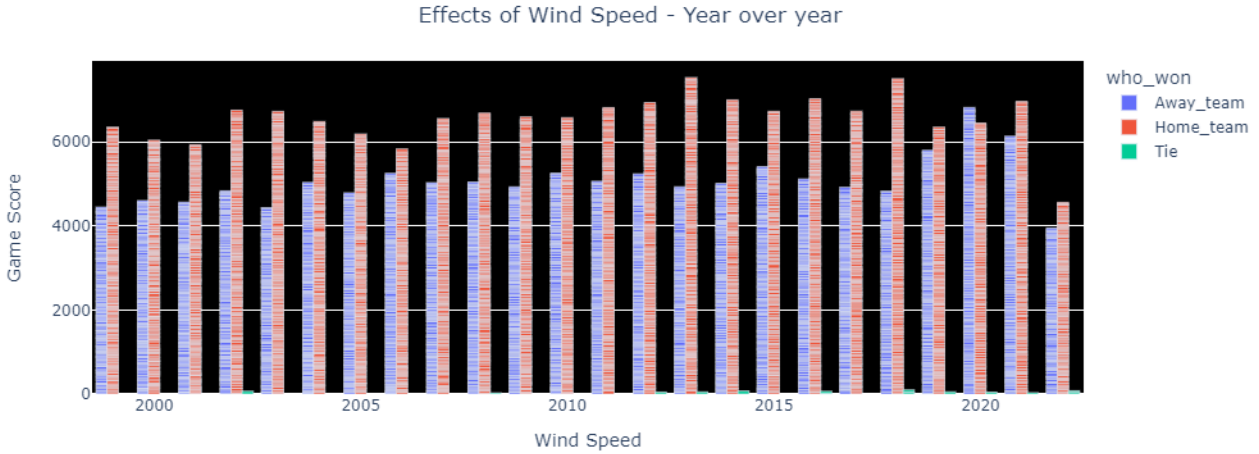
We also need to be mindful of the trends on a year over year basis. While overall mean & median type statistical methods may not prove any final conclusions towards home field advantage, here we will show time variance of data.

1) The data reveals in recent seasons/years, there seems to be significantly higher scoring than in the past.

2) The low speed conditions have lowered scoring

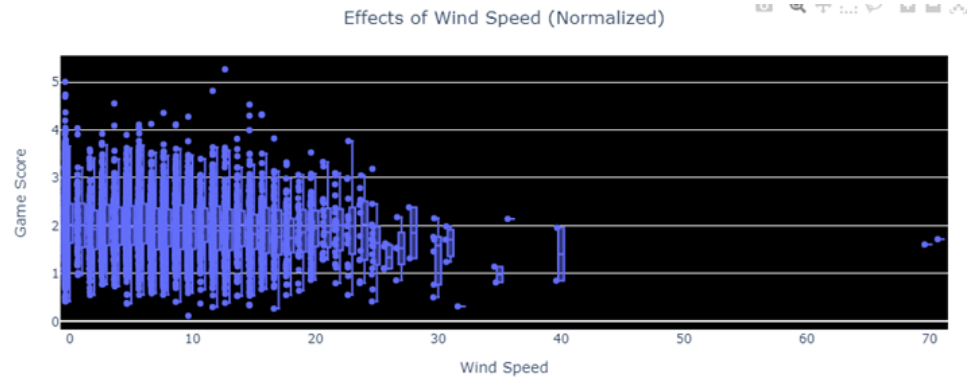


Reviewing the same data and setting the colors to be a home/away team, it shows that in recent years, there has been a significant shift in the game. The away team has been scoring higher. Year 2020 was the first year where away teams scored more than the home teams when all the scoring was combined.



We will now try to normalize our data to see if there are any revelations within the data trends. This set of code will build a list of teams, then build a function which gets the mean score for that season, the data output is determined by team, season, and mean score for each team in the NFL. Finally the data will be merged back to original complete df and normalized home and away scoring is displayed.

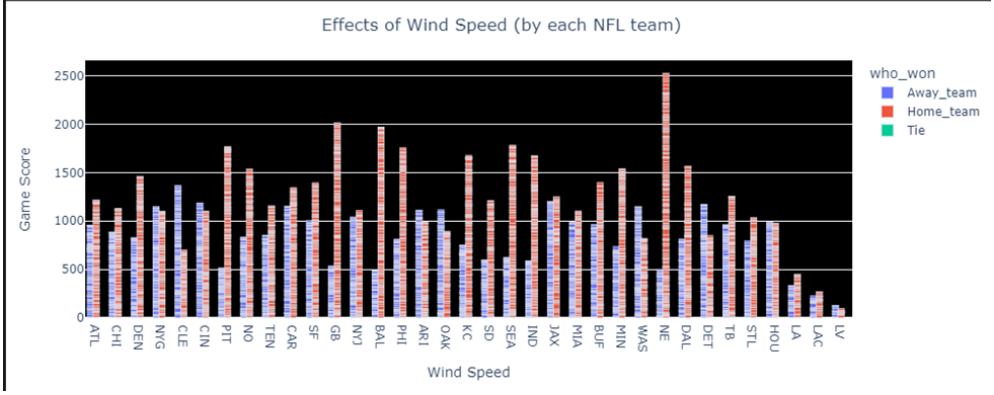
Once normalization is performed, it is charted similarly as shown previously. There does not seem to be any variation in the trend in comparison to the previous box plot.



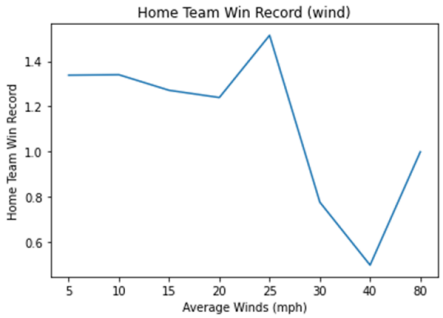
As you can see from the chart below, there are some teams that are skewing the data. New England and Green Bay have much higher scoring on home field.

```
tgames['abs_result'] = tgames.result.abs()
descending_order = tgames['home_team'].value_counts().sort_values(ascending=False).index
fig = px.bar(tgames, x="home_team", y=tgames['abs_result'],color='who_won', barmode='group',height=400, width= 1000)
fig.update_layout({'plot_bgcolor': 'black','paper_bgcolor': 'white',})   #  Set background color
fig.update_layout(title={"text": "Effects of Wind Speed (by each NFL team)", "x": 0.5})
fig.update_layout(yaxis_title="Game Score")
fig.update_layout(xaxis_title="Wind Speed")
fig.show()
```

✓ 0.1s

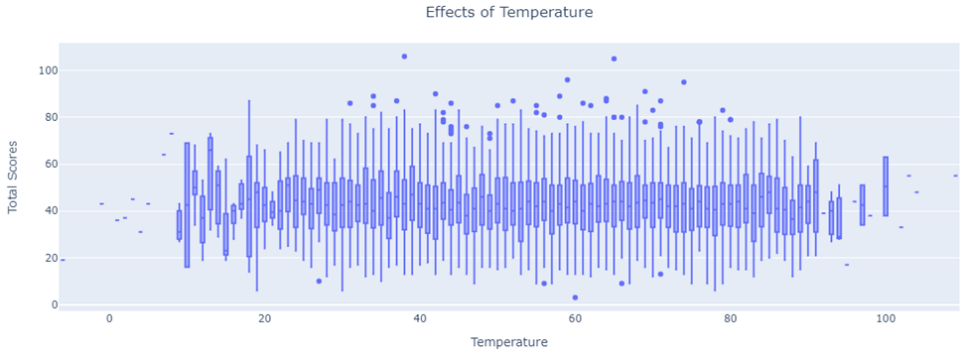


Key takeaways can be taken from this chart as well. Wind speeds higher than ~30 mph, the home team advantage diminishes. When the wind speeds are between 0 and ~30, the home field has considerable advantage as you can see from winning records. The peak happens around 25mph wind speeds where home field wins the most games.

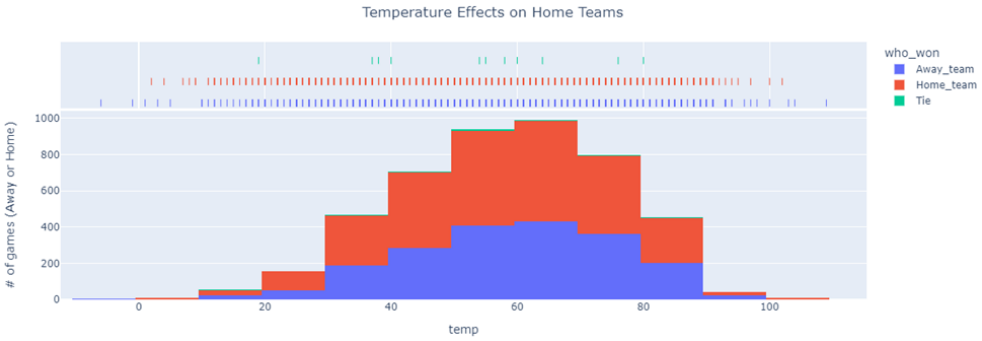


TEMPERATURE EFFECTS ON THE GAME OUTCOME

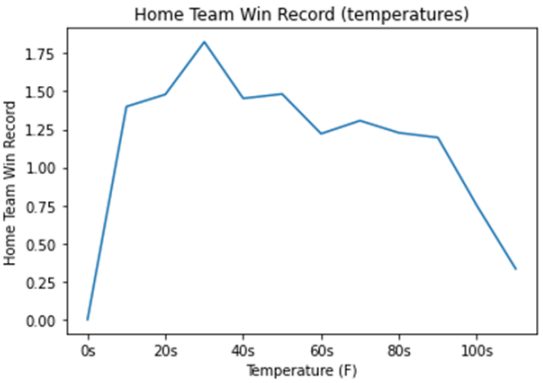
Similar to the wind speed analysis on the effects to the games, temperature was another factor that was analyzed. Box plot was created to visualize the dataset variations. As you can see, temperatures of less than 25 degrees Fahrenheit and greater than 90 degrees, there seems to be inconsistencies when comparing to the rest of the data set. Looking at the total scores of each game, it does not clearly show if there is an increase or decrease in scoring.



Further looking at the home/away teams versus temperature, the home team shows to have advantage when temperature is above 20 degrees and less than 90 degrees.



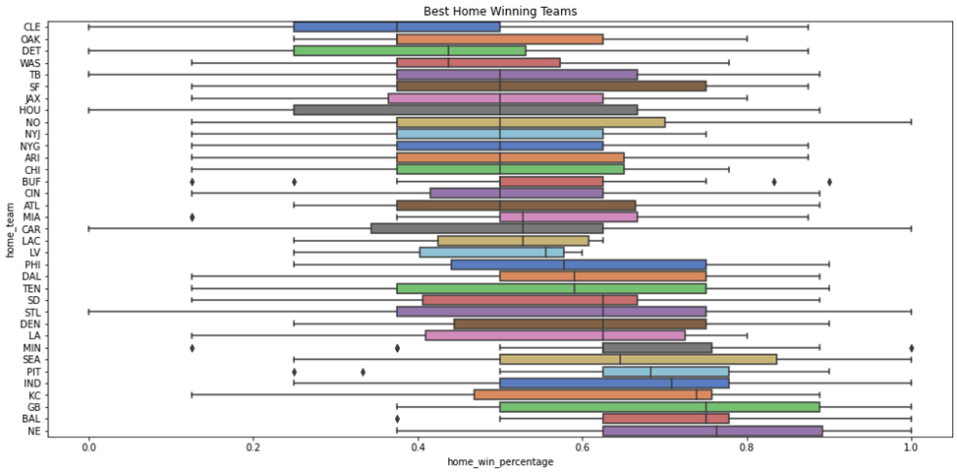
Taking into account winning records, this chart depicts that there is home field advantage if the temperature remains between 20 – 90. Outside those bounds, the home team shows disadvantageous winning records.



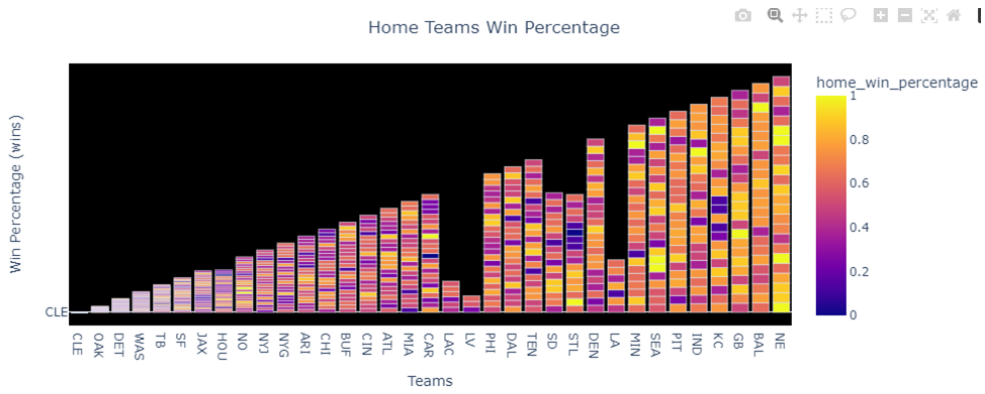
HOME FIELD ADVANTAGE

Analyzing the data, 19 out of 31 NFL teams have home field advantage year over year with a record of 0.5 or better.

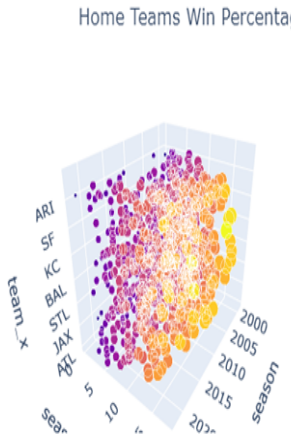
On the contrary, there are few teams as well that have disadvantage (i.e. CLE, OAK, DET, WAS) on playing at home field. Following chart shows the home team win percentage for each team sorted by median win percentage

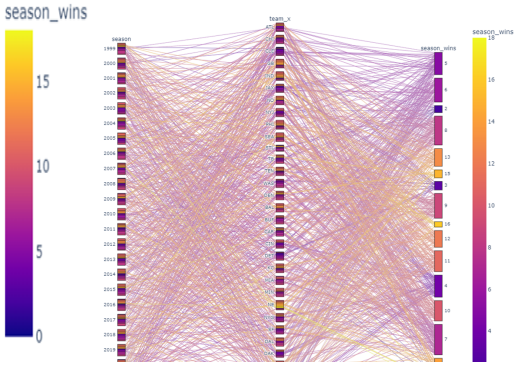


The highest winning percentage goes to NE, BAL, GB since 1999.



The following charts are meant for visual analysis as it contains interactive tools for visualizations when individual team/ season/ wins are selected.





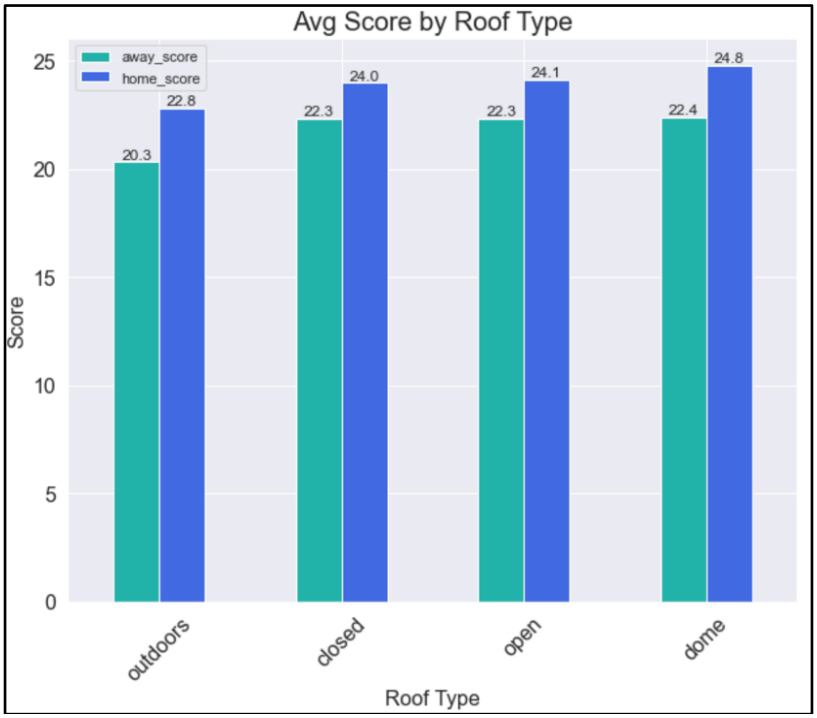
We also looked at a model called Pythagorean Expectations which is a sports analytics formula Pythagorean expectation was devised by Bill James to estimate the percentage of games a baseball team "should" have won based on the number of runs they scored and allowed. Comparing a team's actual and Pythagorean winning percentage can be used to make predictions and evaluate which teams are over-performing and under-performing. The name comes from the formula's resemblance to the Pythagorean theorem. Here are the prediction results devised for the year 2022 (current year).

Projected Wins for AFC Teams	
Team	**Wins**
BUF	12.28
KC	11.64
CIN	10.93
BAL	10.83
MIA	10.09
NYJ	8.65
TEN	8.26
NE	8.23
LAC	7.68
JAX	6.50
LV	6.47
CLE	6.36
PIT	6.05
IND	4.90
DEN	4.55
HOU	1.83

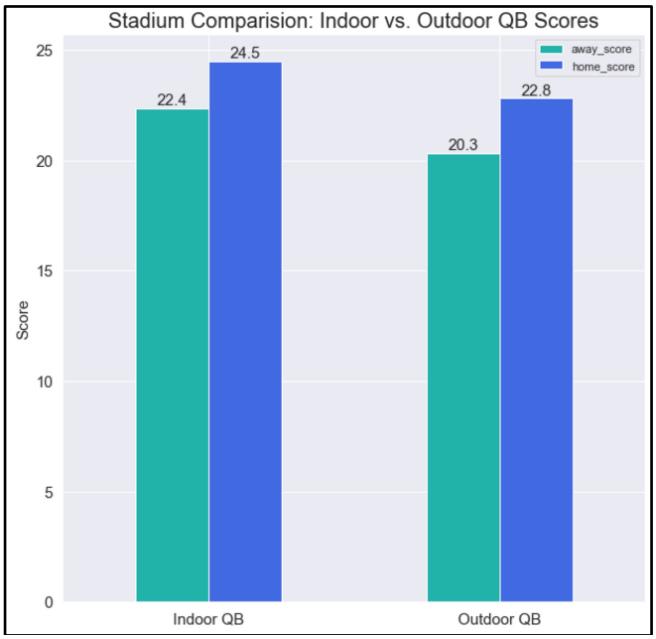
NFC Teams Projection	
Team	**Wins**
PHI	14.26
DAL	12.26
MIN	11.49
SF	10.88
SEA	9.12
WAS	8.48
NYG	8.29
TB	7.98
DET	7.51
ATL	6.35
GB	6.25
CAR	5.65
ARI	5.51
NO	5.29
LA	4.97
CHI	4.12

PERFORMANCE BASED ON ROOF TYPES

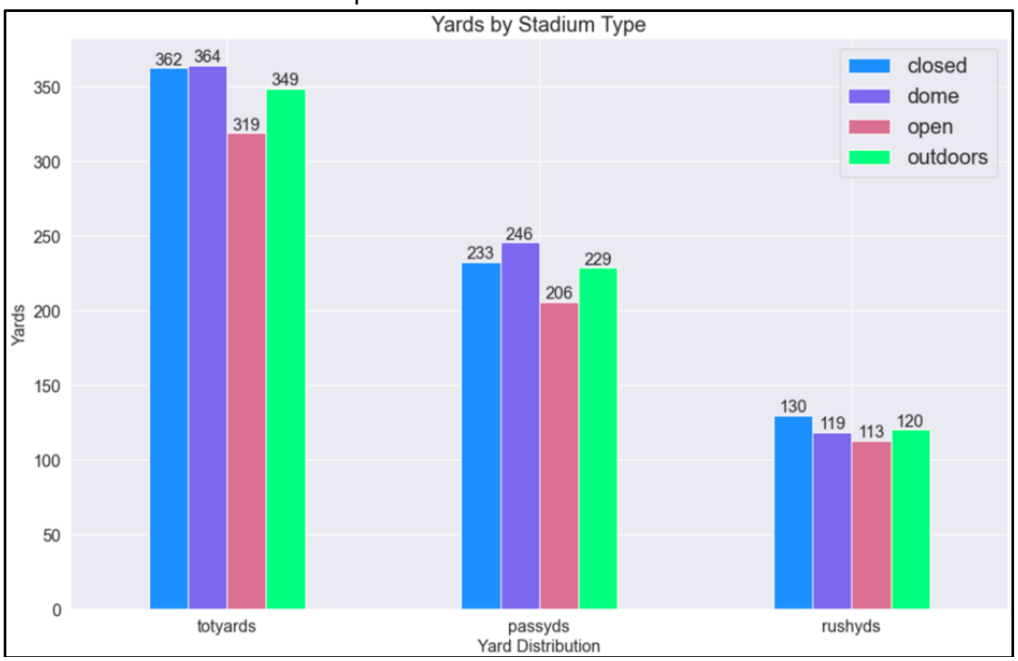
- **Description**
    a. NFL games are played in 4 stadium types: outdoors, closed, open, & dome. A dome has a permanent fixed roof and it is considered 100% indoors. Closed and open are retractable roofs. Open can be considered sunny with no wind impact and closed is considered equivalent to a dome stadium by being enclosed. Outdoors are completely in the elements with no roofs and sustain all weather conditions.
    b. For certain charts, we will classify closed, open, and dome stadiums as “INDOOR” and outdoors will be classified as “OUTDOOR” as the dataset reports 0 mph wind.
    c. We will evaluate home and away team performances in the various stadium types.
- **Result #1: Data from 1999 to present**



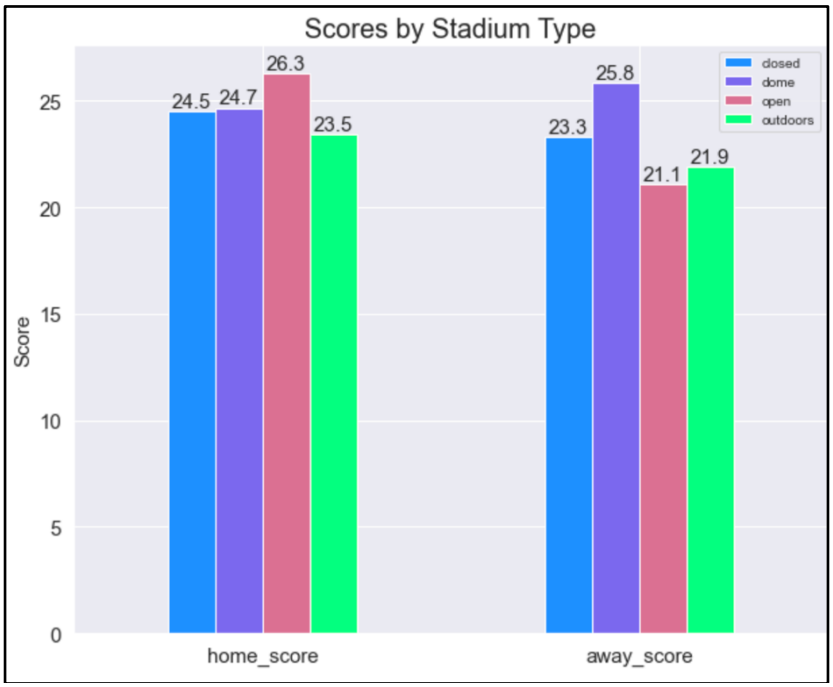
- **Analysis #1:** Outdoors reports the lowest home and away scores out of all the categories with an average home score of 22.8 and average away score of 20.3. Outdoor scores also have the largest point differential. Dome stadiums report the highest average out of all categories with an average home score of 24.8 and average away score of 22.4. Closed and open stadiums have similar results and do not vary heavily in point difference from the average dome scores. The outdoor stadiums vary from the others by at least 2 points.
    a. Based on the charts, it is possible that home teams have scored an average of 2 points more than the away team. The competition is more stiff in indoor stadiums due to the environmental elements eliminated, thus allowing teams to score more.
- **Result #2:** Data from 1999 to present



- **Analysis #2:** The data shows that when quarterbacks play indoors, their teams perform slightly better when both at home and away. This data confirms performance by quarterbacks.
- **Result #3:** Data from 2019 to present



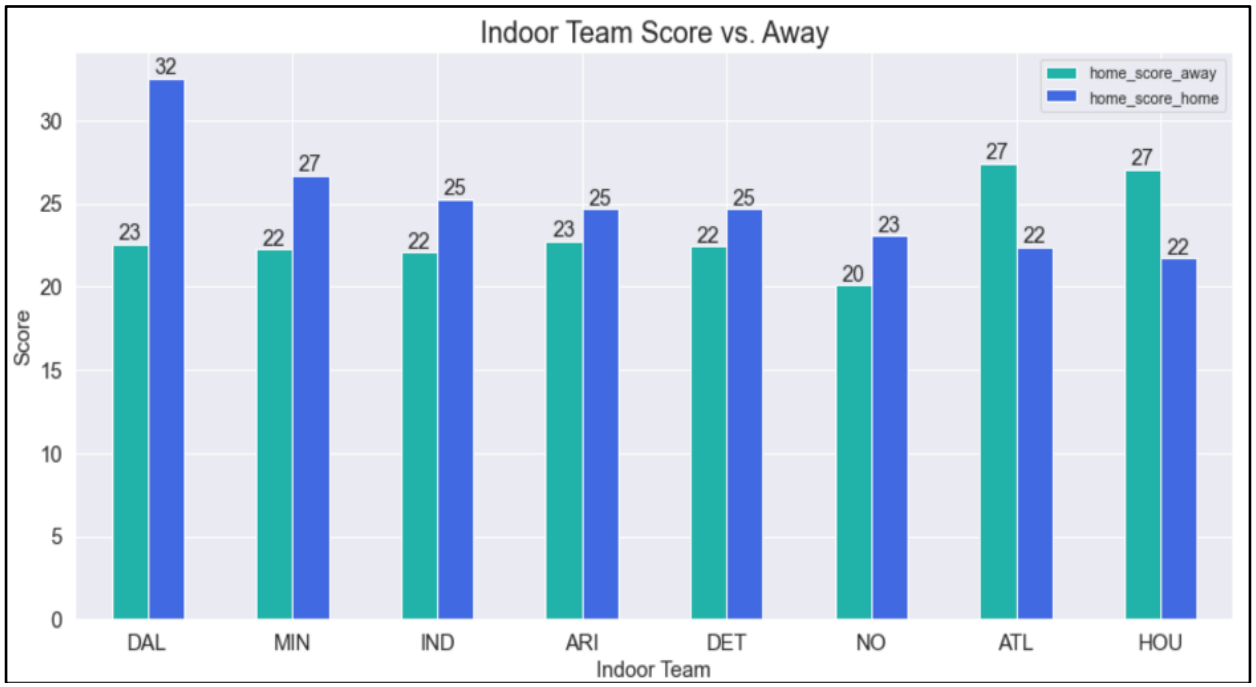
- **Analysis #4:** Game statistics from 2019 until present is analyzed in the chart above. Open stadiums with their roofs retracted have the lowest yards across the board. For outdoor stadiums, the yardage is competitive to closed and dome stadiums.
    a. Outdoor stadiums rank #3 for total yards, #3 for passing yards, and #2 for rushing yards. The two prior charts show outdoor performance not yielding as many points as outdoor stadiums, but they produce a fair amount of yards compared to the other stadium types.
- **Result #5:** Data from 2019 to present



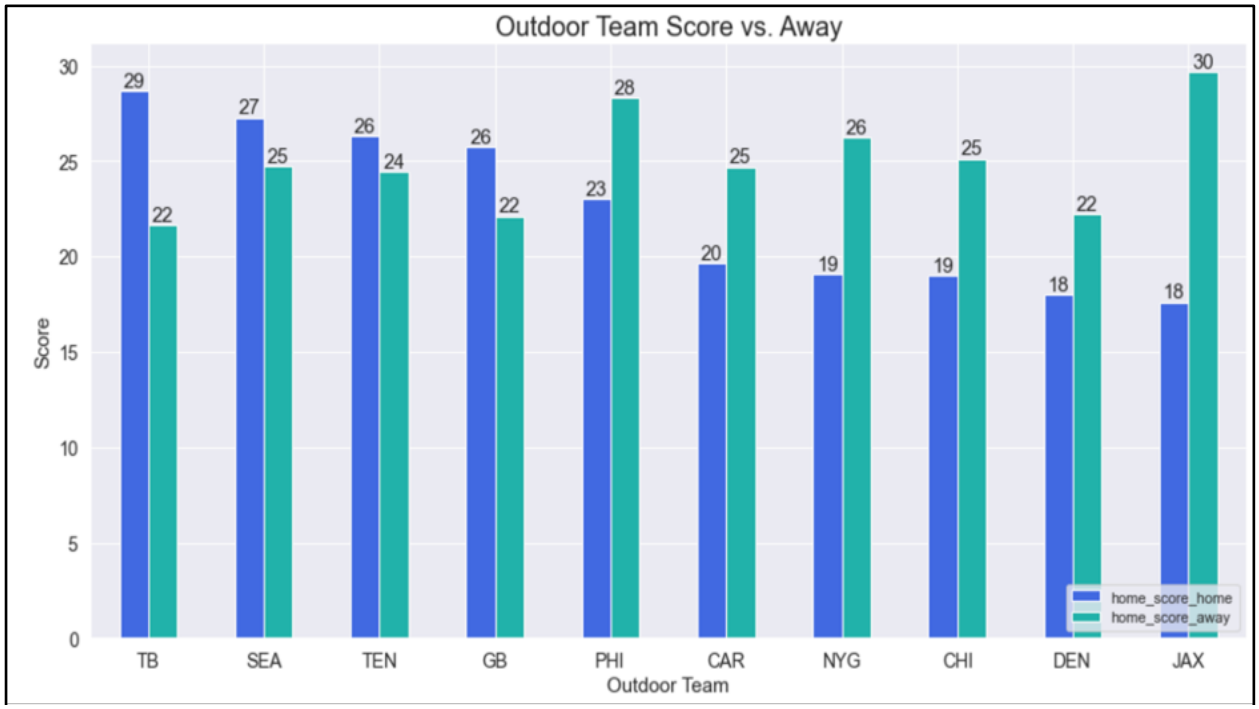
- **Analysis #5:** This chart is comparable to the home and away to Result #1 for this question. The home score reports outdoors as the lowest– however outdoor score is not the lowest when it comes to away games. We see a difference in performance in the past few years compared to data from 1999 to present. Narrowing down the years to 2019-presents helps us compare the latest trends.
    a. As we have seen in result #4, outdoor stadiums do not perform the lowest yards in any category. However, we do see that it is the lowest in the home category and second lowest in the away category. We can conclude that yards do not translate to higher scores.
    b. Another interesting comparison would be the dome stadium score results. Home teams with dome stadiums have not performed as well at home with the away team scoring a higher average with a differential of 1.1 points.
    c. Closed stadiums have the most competitive range between home and away with a differential of 1.2 points.

HOW DO HOME TEAMS WITH AN INDOOR STADIUM PERFORM IN OUTDOOR STADIUMS?

- **Result #1:**



- **Analysis #1:** Teams that have played 98% of their home games indoors were analyzed. Teams that recently moved into an indoor stadium were not included in this analysis.
    a. There are 8 teams with their away and home performance analyzed. 75% of teams with home indoor stadiums hold a strong advantage over their opponent,
    b. The Dallas Cowboys perform well over average when at home with a high average of 32 points. However, their away game performance is average compared to other teams.
- **Result #2:**

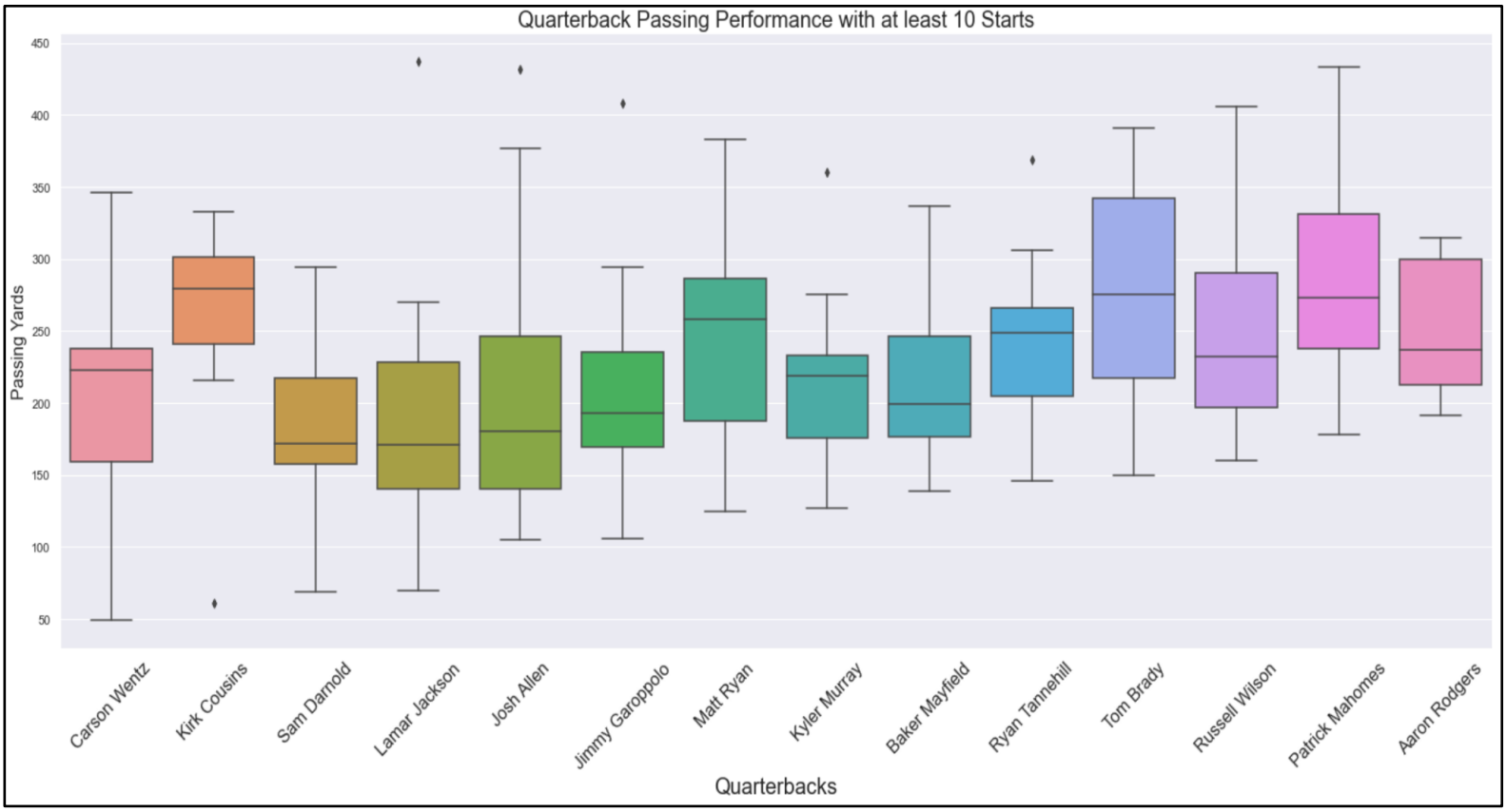


- **Analysis #2:** Teams that have an outdoor stadium and play at least 25% of their away games in indoor stadiums are analyzed.

a. There are 10 teams analyzed and the results are mixed between 40/60 ratio in performances. The data does not conclusively decide whether teams with outdoor stadiums perform better or away.
b. It is interesting to note Philadelphia Eagles and the New York Giants perform better when they are away considering the Dallas Cowboys are in their division with an indoor stadium and each set of teams meet twice a year.

TOP PASSING QUARTERBACKS WITH AT LEAST 10 STARTS

- Result #1: Data from 2019 to present

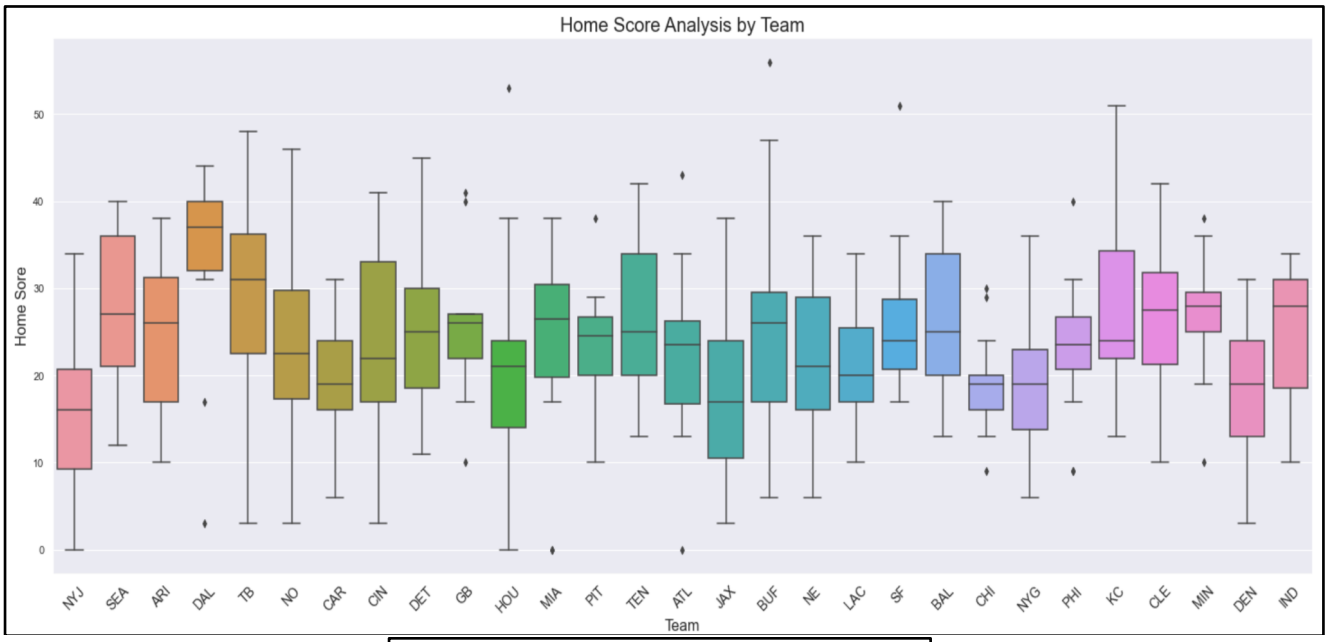


	passyds							
	count	mean	std	min	25%	50%	75%	max
home_qb_name								
Aaron Rodgers	11.0	250.636364	48.108778	191.0	212.50	237.0	299.5	315.0
Baker Mayfield	15.0	216.666667	56.910540	139.0	176.50	199.0	246.5	337.0
Carson Wentz	11.0	200.545455	82.780872	49.0	159.50	223.0	238.0	346.0
Jimmy Garoppolo	13.0	210.615385	79.496896	106.0	169.00	193.0	235.0	408.0
Josh Allen	19.0	212.736842	94.709991	105.0	140.00	180.0	246.0	432.0
Kirk Cousins	14.0	261.642857	68.043659	61.0	241.00	279.5	301.5	333.0
Kyler Murray	11.0	215.000000	65.966658	127.0	176.00	219.0	233.0	360.0
Lamar Jackson	17.0	189.411765	83.727280	70.0	140.00	171.0	228.0	437.0
Matt Ryan	15.0	243.266667	74.726613	125.0	187.50	258.0	286.5	383.0
Patrick Mahomes	22.0	290.136364	76.647800	178.0	238.00	273.0	331.5	433.0
Russell Wilson	14.0	255.142857	75.171452	160.0	196.50	232.5	290.0	406.0
Ryan Tannehill	16.0	237.312500	55.745516	146.0	204.75	248.5	266.0	369.0
Sam Darnold	12.0	187.666667	62.214195	69.0	157.25	172.0	217.0	294.0
Tom Brady	21.0	277.476190	73.065463	150.0	217.00	275.0	342.0	391.0

- **Analysis #1:** Quarterback passing performance with the highest averages and at least 10 starts were analyzed by boxplots. With a boxplot, we can see the minimum, maximum, average, 25% quartile, and 75% quartile in passing yards.
    a. We can assume the smaller the box, the more consistent their passing performance, regardless if it is low or high. It does not mean they perform well– just more consistent.
        i. We can see Aaron Rodgers has one of the smallest boxes with 0 outliers. His standard deviation is also +/- 48 yards, which is incredibly consistent and it is also the lowest standard deviation out of all the quarterbacks.
        ii. Apart from Kirk Cousins 1 outlier, he is also a consistent quarterback with an even tighter box compared to Aaron Rodgers. I believe his outlier is skewing his standard deviation of 68 yards. But overwise, he performs consistently.
        iii. Lamar Jackson and Josh Allen are potentially the most inconsistent in the passing game with a standard deviation of +/- 85 yards (Jackson) and +/- 94 yards (Allen).
    b. With some simple statistics, we can see that Tom Brady and Patrick Mahomes have the highest medians. But Mahomes has a higher average of 290 yards while Tom Brady's average is 277. But Tom Brady has the highest 75% quartile.
    c. Interesting to note that Tom Brady and Jimmy Garoppolo have close to perfect boxes with zero outliers.
    d. Quarterbacks with larger boxes have a wide range of passing performance

HOW DO TEAMS SCORE AT HOME?

- Result #1: Data from 2019 to present



	home_score							
	count	mean	std	min	25%	50%	75%	max
home_team								
ARI	12.0	24.666667	9.403417	10.0	17.00	26.0	31.25	38.0
ATL	16.0	22.375000	9.450750	0.0	16.75	23.5	26.25	43.0
BAL	19.0	26.789474	7.997075	13.0	20.00	25.0	34.00	40.0
BUF	20.0	25.600000	12.613276	6.0	17.00	26.0	29.50	56.0
CAR	17.0	19.647059	6.836859	6.0	16.00	19.0	24.00	31.0
CHI	17.0	19.000000	5.408327	9.0	16.00	19.0	20.00	30.0
CIN	13.0	23.461538	11.236958	3.0	17.00	22.0	33.00	41.0
CLE	14.0	26.500000	9.732973	10.0	21.25	27.5	31.75	42.0
DAL	10.0	32.500000	12.886255	3.0	32.00	37.0	40.00	44.0
DEN	13.0	18.000000	8.573214	3.0	13.00	19.0	24.00	31.0
DET	15.0	24.666667	9.099974	11.0	18.50	25.0	30.00	45.0
GB	11.0	25.727273	8.934102	10.0	22.00	26.0	27.00	41.0
HOU	14.0	21.714286	12.808823	0.0	14.00	21.0	24.00	53.0
IND	15.0	25.266667	8.013085	10.0	18.50	28.0	31.00	34.0
JAX	15.0	17.600000	9.796501	3.0	10.50	17.0	24.00	38.0
KC	22.0	28.181818	10.026804	13.0	22.00	24.0	34.25	51.0
LAC	14.0	21.357143	7.121011	10.0	17.00	20.0	25.50	34.0
MIA	18.0	24.000000	10.643971	0.0	19.75	26.5	30.50	38.0
MIN	15.0	26.666667	6.893751	10.0	25.00	28.0	29.50	38.0
NE	17.0	21.588235	8.775383	6.0	16.00	21.0	29.00	36.0
NO	18.0	23.055556	11.362851	3.0	17.25	22.5	29.75	46.0
NYG	12.0	19.083333	8.106769	6.0	13.75	19.0	23.00	36.0
NYJ	14.0	15.071429	9.817287	0.0	9.25	16.0	20.75	34.0
PHI	12.0	23.000000	8.686458	9.0	20.75	23.5	26.75	40.0
PIT	18.0	22.833333	6.887756	10.0	20.00	24.5	26.75	38.0
SEA	15.0	27.266667	9.414781	12.0	21.00	27.0	36.00	40.0
SF	16.0	26.250000	8.551803	17.0	20.75	24.0	28.75	51.0
TB	22.0	28.681818	10.580244	3.0	22.50	31.0	36.25	48.0
TEN	17.0	26.294118	9.292233	13.0	20.00	25.0	34.00	42.0

- **Analysis #1:** Home scores are analyzed by the team in this analysis. The Dallas Cowboys absolutely blew all other teams out of the water with their home performance, despite their two outliers. Their average home score is 32.5 points with the next highest average at 28.7 points for the Tampa Bay Bucs.

a. The New York Jets have the lowest scores in all categories. It is not fun to be dominated by the Patriots for 20 years, just be to defeated by Josh Allen as the new AFC East dominator.
b. Despite Aaron Rodgers being the most consistent thrower, Green Bay does not show consistency in scoring.

CONCLUSION

Based on our analysis on indoor and outdoor stadiums, we can conclude that teams perform better overall in an indoor type of stadium. We infer this is due to the environmental elements removed from the game. Weather does not seem to impact a team's ability to gain passing or rushing yards, however there is enough data to show that it does impact the scoring. Despite the yards gained in outdoor games, it does not convert to points. Perhaps the game plan for the red zone should be adjusted for outdoor games.

In addition to indoor games, teams with an indoor stadium tend to perform better when at home than away. However, there is no supporting data that shows that teams with an outdoor stadium perform better or worse when they are away. Away teams that travel to indoor games should prepare to face a tougher game.

Based on the surface type analysis, it was concluded that the type of surface is not really contributing to the scoring. Based on the top performance analysis, teams tend to perform well on surfaces same as their home field surfaces. Green Bay and Kansas City Chiefs can perform well on grass while NE Patriots really do well on Turf. This could be because of their respective home field surfaces and the more practice they get to do on those surfaces.

We also see an upward trend in the demand of field goals, and this can suggest the improved training given to specialty teams.

As for the temperature conditions, there seems to exist a slight correlation in the extreme temperature ranges. Generally, the home team is favored in the temperature range of 20-90 degrees. In addition, wind speeds lower than 30 mph tend to favor the home field teams.

NEXT STEPS…

We had a rich dataset for the NFL games and its statistics, but our team would have greatly benefited from additional data regarding the offense, defense, and special teams to supplement our game datasets to further expand on our analysis.

The datasets provide a wide range of offensive statistics but it does not include defensive statistics such as interceptions and fumbles. It would have been nice to include defensive statistics and the interceptions and passing/rushing touchdowns for the quarterbacks. It would have been interesting to merge that type of data and compare the quarterback passing touchdown and interception ratio by season.

To enhance the surface type analysis, player injury data needs to be incorporated into the scope. Recent studies conducted by the NFL show that player injury rates have come down on Turf surfaces compared to previous years. This could be due to the quality of turf being improved over the years.

Higher seeded playoff team gets the home field advantage. But to quantify the advantage, the crowd details (capacity vs. attendance) along with the percentage of fan crowd for each of the team has to be taken into account.

In addition to the capacity and attendance, the decibel measurement for each game would be a great addition to determine if noise impacts the game as well. This could also help determine if the level of noise impacts away teams.

As for the prediction model in the presentation, the next step is to determine its accuracy in predicting the number of wins per team for the 2022-2023 season.