Project Progress

What constitutes "Home field advantage" in the NFL?

Chau Nguyen

December 2, 2020

Background & Motivation

I want to study "home field advantage" in the National Football League (NFL). When football analysts and pundits discuss "home field advantage", it usually is a combination of a few things:

Background & Motivation

I want to study "home field advantage" in the National Football League (NFL). When football analysts and pundits discuss "home field advantage", it usually is a combination of a few things:

- Distance traveled: The Home team does not have to travel to play the game (there are rare exceptions). This helps with their ability to rest, practice, and game plan.

Background & Motivation

I want to study "home field advantage" in the National Football League (NFL). When football analysts and pundits discuss "home field advantage", it usually is a combination of a few things:

- Distance traveled: The Home team does not have to travel to play the game (there are rare exceptions). This helps with their ability to rest, practice, and game plan.
- The Away team is at an even more disadvantage if they had to travel across timezones for the match up.

Background & Motivation

I want to study "home field advantage" in the National Football League (NFL). When football analysts and pundits discuss "home field advantage", it usually is a combination of a few things:

- Distance traveled: The Home team does not have to travel to play the game (there are rare exceptions). This helps with their ability to rest, practice, and game plan.
- The Away team is at an even more disadvantage if they had to travel across timezones for the match up.
- With fans in the stand, the Home crowd is usually silent during the Home team's Offensive snaps (so that the Quarterback can "read" the opposing Defense and make adjustments to the play before the snap).

Background & Motivation

I want to study "home field advantage" in the National Football League (NFL). When football analysts and pundits discuss "home field advantage", it usually is a combination of a few things:

- Distance traveled: The Home team does not have to travel to play the game (there are rare exceptions). This helps with their ability to rest, practice, and game plan.
- The Away team is at an even more disadvantage if they had to travel across timezones for the match up.
- With fans in the stand, the Home crowd is usually silent during the Home team's Offensive snaps (so that the Quarterback can "read" the opposing Defense and make adjustments to the play before the snap).
- On the other hand, during the Away team's Offensive snaps, the Home crowd is encouraged to get loud - therefore the opposing QB can't do the same.

Background & Motivation

The situation is different in 2020. Many stadiums are required to limit fan attendance to a very low number to mitigate covid risks. Players have commented on how eerily quiet and different the game day atmosphere has been this year.
With that in mind, I wanted to see if I could build a model that quantifies "Home field advantage" in a normal year.

- Do Home teams really win more?

If the model does well between the training and test datasets, I want to use it on game outcomes for the 2020 Season and see what comes from there.

Background & Motivation

The situation is different in 2020. Many stadiums are required to limit fan attendance to a very low number to mitigate covid risks. Players have commented on how eerily quiet and different the game day atmosphere has been this year.
With that in mind, I wanted to see if I could build a model that quantifies "Home field advantage" in a normal year.

- Do Home teams really win more?
- What is it about the Home Stadium that helps a team win?

If the model does well between the training and test datasets, I want to use it on game outcomes for the 2020 Season and see what comes from there.

Background & Motivation

The situation is different in 2020. Many stadiums are required to limit fan attendance to a very low number to mitigate covid risks. Players have commented on how eerily quiet and different the game day atmosphere has been this year.
With that in mind, I wanted to see if I could build a model that quantifies "Home field advantage" in a normal year.

- Do Home teams really win more?
- What is it about the Home Stadium that helps a team win?
- Is it the Stadium itself? The Home fans in attendance?

If the model does well between the training and test datasets, I want to use it on game outcomes for the 2020 Season and see what comes from there.

Background & Motivation

The situation is different in 2020. Many stadiums are required to limit fan attendance to a very low number to mitigate covid risks. Players have commented on how eerily quiet and different the game day atmosphere has been this year.
With that in mind, I wanted to see if I could build a model that quantifies "Home field advantage" in a normal year.

- Do Home teams really win more?
- What is it about the Home Stadium that helps a team win?
- Is it the Stadium itself? The Home fans in attendance?
- Or is it the lack of travel that helps?

If the model does well between the training and test datasets, I want to use it on game outcomes for the 2020 Season and see what comes from there.

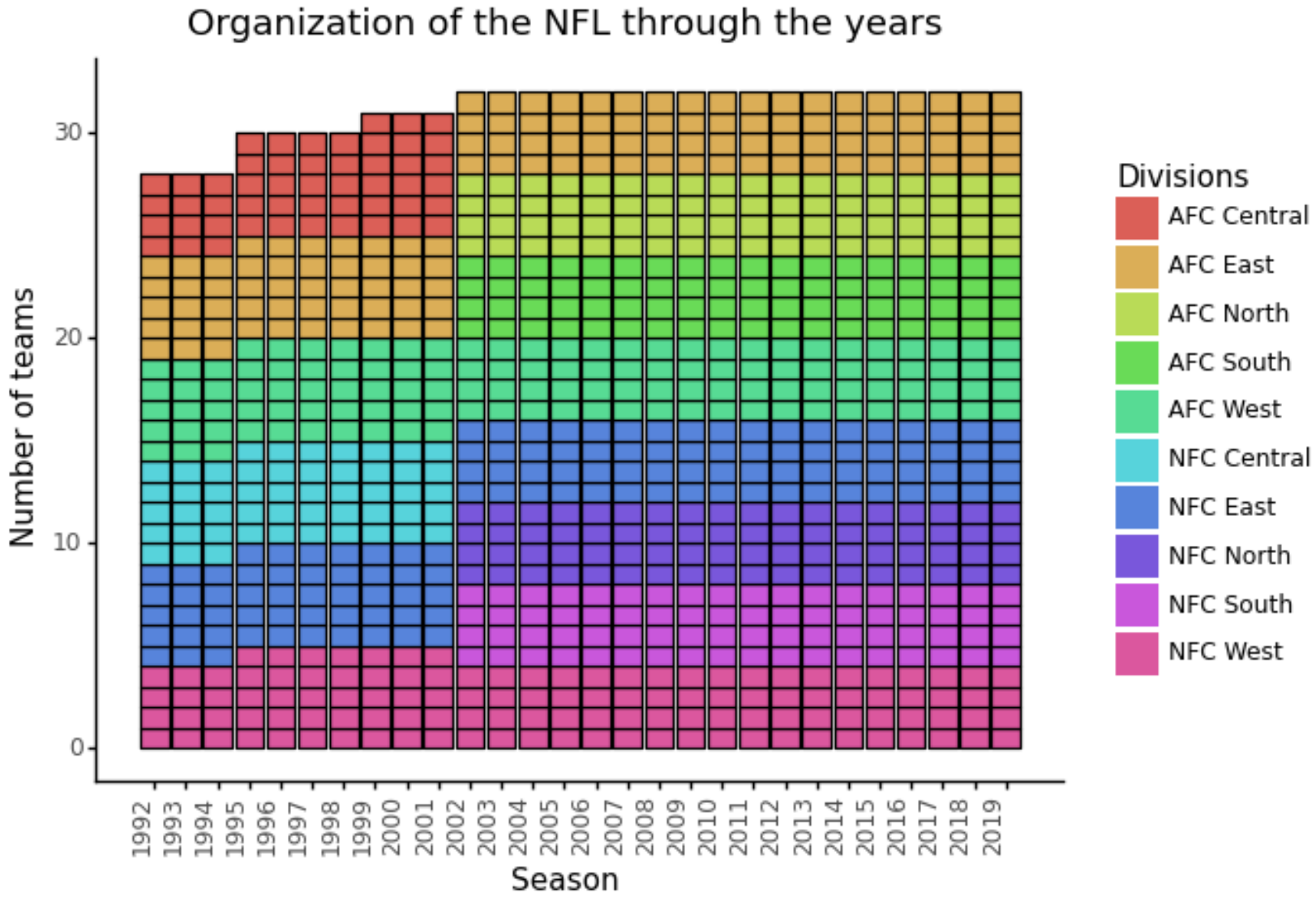
Overview of the NFL

- Currently, there are 32 teams in the NFL, divided into 2 conferences: the National Football Conference (NFC) and the American Football Conference (AFC). Each conference is divided into 4 divisions: North, South, East, West - each with 4 teams.

Overview of the NFL

- Currently, there are 32 teams in the NFL, divided into 2 conferences: the National Football Conference (NFC) and the American Football Conference (AFC). Each conference is divided into 4 divisions: North, South, East, West - each with 4 teams.
- This has not always been the case before the "realignment" in 2002, the landscape of the NFL looked quite different.

Overview of the NFL



Methods and approaches I considered using

- Pro-Football-Reference.com has attendance data for every NFL game starting in 1992 - this was my starting point.

Methods and approaches I considered using

- Pro-Football-Reference.com has attendance data for every NFL game starting in 1992 - this was my starting point.
- I wanted to practice webscraping and data wrangling with Python, so I set out to write scrapers for the data I needed - mostly from Wikipedia and Pro-Football-Reference.

Methods and approaches I considered using

- Pro-Football-Reference.com has attendance data for every NFL game starting in 1992 - this was my starting point.
- I wanted to practice webscraping and data wrangling with Python, so I set out to write scrapers for the data I needed - mostly from Wikipedia and Pro-Football-Reference.
- I would calculate the "distance traveled" for each team using the distance measured between their Home stadium and the stadium they're playing at.

Methods and approaches I considered using

- Pro-Football-Reference.com has attendance data for every NFL game starting in 1992 - this was my starting point.
- I wanted to practice webscraping and data wrangling with Python, so I set out to write scrapers for the data I needed - mostly from Wikipedia and Pro-Football-Reference.
- I would calculate the "distance traveled" for each team using the distance measured between their Home stadium and the stadium they're playing at.
- I can also calculate the amount of time each team has had to rest inbetween games.

Methods and approaches I considered using

- Pro-Football-Reference.com has attendance data for every NFL game starting in 1992 - this was my starting point.
- I wanted to practice webscraping and data wrangling with Python, so I set out to write scrapers for the data I needed - mostly from Wikipedia and Pro-Football-Reference.
- I would calculate the "distance traveled" for each team using the distance measured between their Home stadium and the stadium they're playing at.
- I can also calculate the amount of time each team has had to rest inbetween games.
- I want to format my dataset into a dyadic panel and use Machine Learning to study the difference between each pair.

Methods and tools used to date, and rationale for their use

- I have written many scrapers for Wikipedia using BeautifulSoup, because a lot of the time the read_html function in pandas does not give me what I need.

Methods and tools used to date, and rationale for their use

- I have written many scrapers for Wikipedia using BeautifulSoup, because a lot of the time the read_html function in pandas does not give me what I need.
- Even then, because of the data challenges above, I still needed to go in and clean up the scraped data manually - this was the most time consuming part.

Methods and tools used to date, and rationale for their use

- I have written many scrapers for Wikipedia using BeautifulSoup, because a lot of the time the read_html function in pandas does not give me what I need.
- Even then, because of the data challenges above, I still needed to go in and clean up the scraped data manually - this was the most time consuming part.
- Why?

Challenge 1: Team Names, Stadiums and Game Locations

- NFL franchises move cities and rename themselves over the years.

Challenge 1: Team Names, Stadiums and Game Locations

- NFL franchises move cities and rename themselves over the years.
  - Cleveland Rams (1936 - 1945, LA Rams (1946-1994), St. Louis Rams (1995 - 2015), LA Rams (2016 - present).
  - The team previously known as "Washington Redskins" is being called "Washington Football Team" temporarily beginning 2020 until the franchise finds a new name.

Challenge 1: Team Names, Stadiums and Game Locations

- NFL franchises move cities and rename themselves over the years.
  - Cleveland Rams (1936 - 1945, LA Rams (1946-1994), St. Louis Rams (1995 - 2015), LA Rams (2016 - present).
  - The team previously known as "Washington Redskins" is being called "Washington Football Team" temporarily beginning 2020 until the franchise finds a new name.
- Unique games such as the NFL International Series played in London and Mexico City every year, or the Bill Toronto Series.

Challenge 1: Team Names, Stadiums and Game Locations

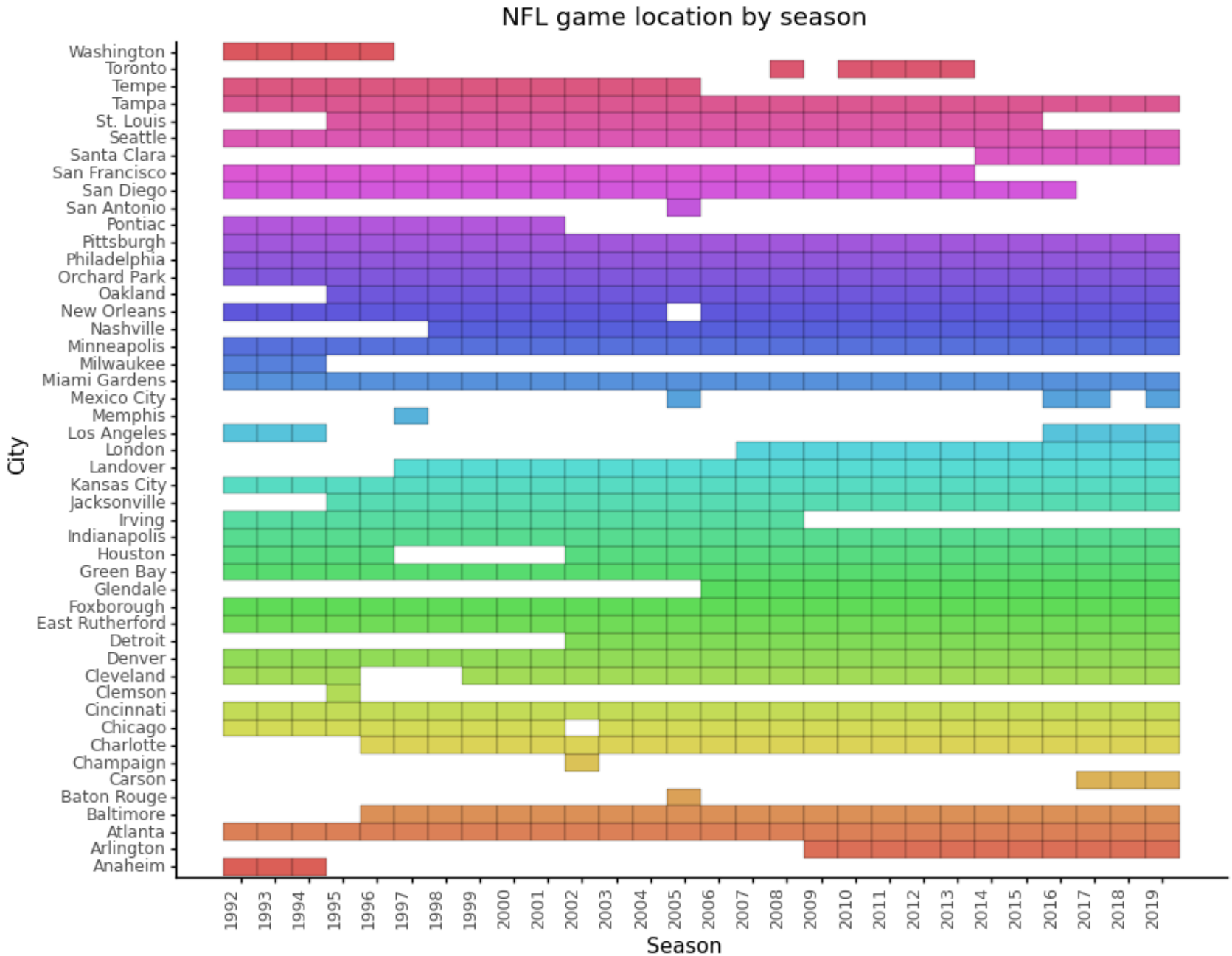
- NFL franchises move cities and rename themselves over the years.
  - Cleveland Rams (1936 - 1945, LA Rams (1946-1994), St. Louis Rams (1995 - 2015), LA Rams (2016 - present).
  - The team previously known as "Washington Redskins" is being called "Washington Football Team" temporarily beginning 2020 until the franchise finds a new name.
- Unique games such as the NFL International Series played in London and Mexico City every year, or the Bill Toronto Series.
- Games being moved because of "Act of God" events

Challenge 1: Team Names, Stadiums and Game Locations

- NFL franchises move cities and rename themselves over the years.
  - Cleveland Rams (1936 - 1945, LA Rams (1946-1994), St. Louis Rams (1995 - 2015), LA Rams (2016 - present).
  - The team previously known as "Washington Redskins" is being called "Washington Football Team" temporarily beginning 2020 until the franchise finds a new name.
- Unique games such as the NFL International Series played in London and Mexico City every year, or the Bill Toronto Series.
- Games being moved because of "Act of God" events
  - The New Orleans Saints played their 2005 season on the road due to Hurricane Katrina - their "Home" games were played in nearby stadiums, but not the Superdome.

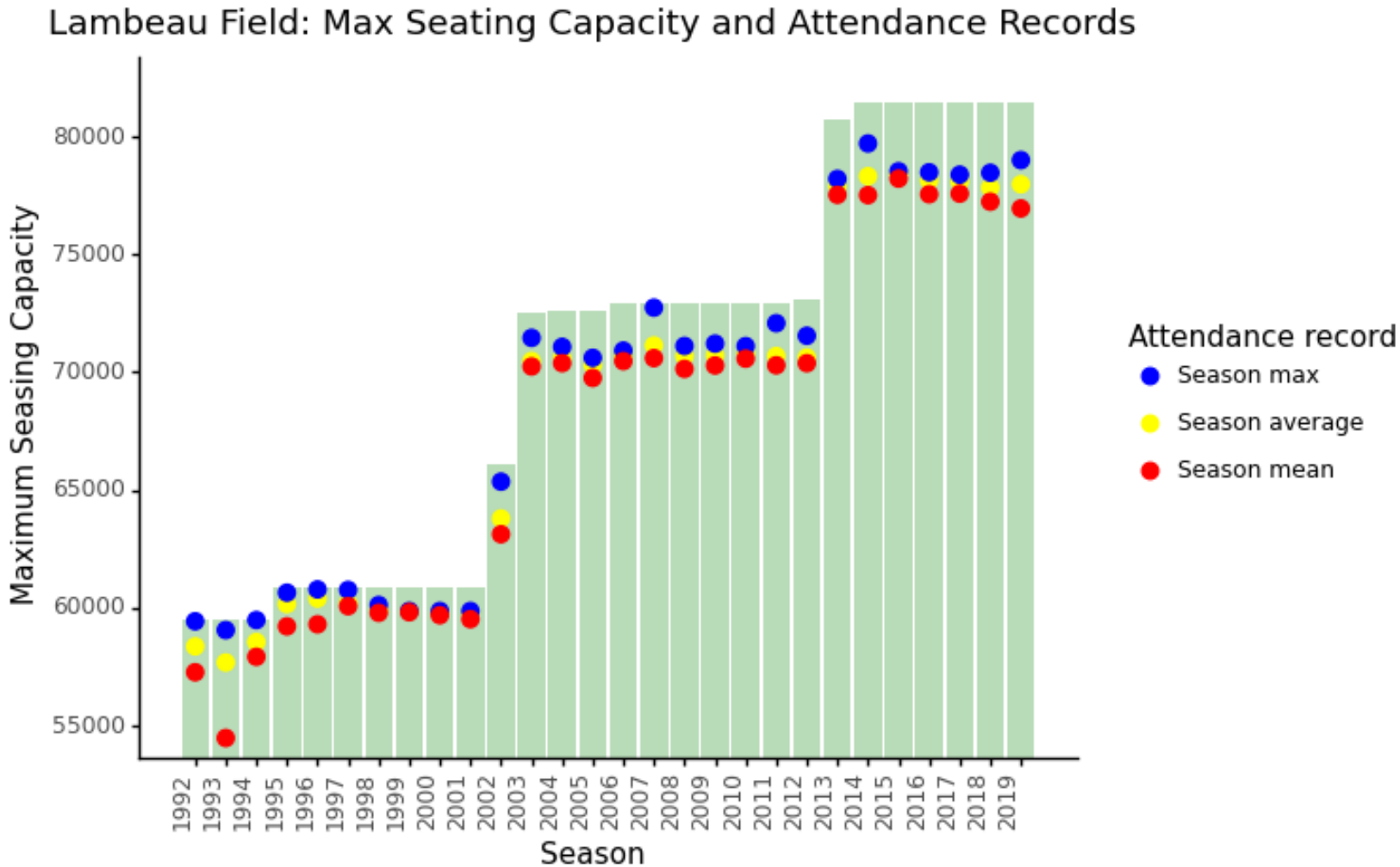
Challenge 1: Team Names, Stadiums and Game Locations

Challenge 1: Team Names, Stadiums and Game Locations



Challenge 2: Even in the same stadium, maximum capacity is not static

Challenge 2: Even in the same stadium, maximum capacity is not static

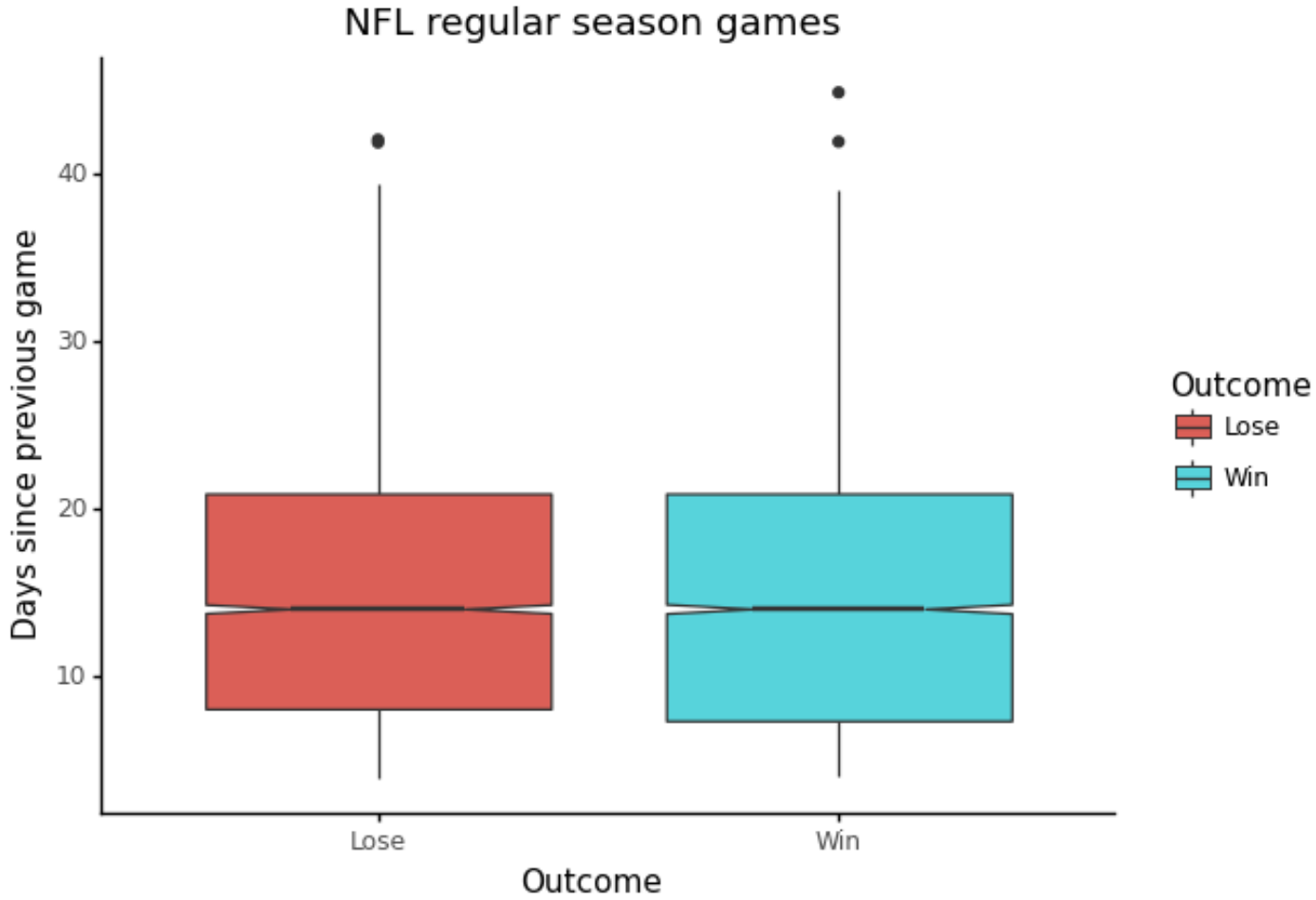


Preliminary results and conclusions

- I do not have any preliminary results yet, because I'm still not 100% happy with my data (as I'm making this presentation, I saw a few things that needed improving).

Preliminary results and conclusions

- I do not have any preliminary results yet, because I'm still not 100% happy with my data (as I'm making this presentation, I saw a few things that needed improving).
  - For example, take a look at this graph



Where I am right now

- I do not have preliminary results yet, because I'm still not 100% happy with my data (as I'm making this presentation, I saw a few things that needed improving).

Where I am right now

- I do not have preliminary results yet, because I'm still not 100% happy with my data (as I'm making this presentation, I saw a few things that needed improving).
  - If you are not familiar with the NFL - teams play each other every weekend. Then why is the average number of days between the losing team and the winning team around 14 days?

Where I am right now

- I do not have preliminary results yet, because I'm still not 100% happy with my data (as I'm making this presentation, I saw a few things that needed improving).
  - If you are not familiar with the NFL - teams play each other every weekend. Then why is the average number of days between the losing team and the winning team around 14 days?
  - This has to do with how I cleaned my data before turning it into a dyad. I will have to revisit this and fix it.

Where I am right now

- I have data for 7,292 games across 28 NFL seasons from 1992 to 2019.

Where I am right now

- I have data for 7,292 games across 28 NFL seasons from 1992 to 2019.
- This is a snippet of my (dyad) data at the moment (ignore the incorrect Time_rest_days).

	Season	Week	Team_A	Team_B	Outcome	Miles_traveled	Time_rest_days	Capacity	attendance	Surface	Same_surface	Rivalry
13784	2018	10	Pittsburgh Steelers	Carolina Panthers	Win	0.00	11.31	68400.0	62881.0	Grass	Yes	No
13785	2018	10	Carolina Panthers	Pittsburgh Steelers	Lose	363.60	18.31	68400.0	62881.0	Grass	Yes	No
13786	2018	10	Buffalo Bills	New York Jets	Win	277.93	21.00	82500.0	77982.0	Turf	Yes	Division
13787	2018	10	New York Jets	Buffalo Bills	Lose	0.00	21.00	82500.0	77982.0	Turf	Yes	Division
13788	2018	10	New England Patriots	Tennessee Titans	Lose	924.59	12.70	69143.0	69363.0	Grass	No	Conference
13789	2018	10	Tennessee Titans	New England Patriots	Win	0.00	27.86	69143.0	69363.0	Grass	Yes	Conference
13790	2018	10	Tampa Bay Buccaneers	Washington Redskins	Lose	0.00	21.00	65618.0	52667.0	Grass	Yes	Conference
13791	2018	10	Washington Redskins	Tampa Bay Buccaneers	Win	821.85	14.00	65618.0	52667.0	Grass	Yes	Conference
13792	2018	10	Kansas City Chiefs	Arizona Cardinals	Win	0.00	14.00	76416.0	76712.0	Grass	Yes	No
13793	2018	10	Arizona Cardinals	Kansas City Chiefs	Lose	1058.86	28.00	76416.0	76712.0	Grass	Yes	No

Where I am right now

- I have data for 7,292 games across 28 NFL seasons from 1992 to 2019.
- This is a snippet of my (dyad) data at the moment (ignore the incorrect Time_rest_days).

	Season	Week	Team_A	Team_B	Outcome	Miles_traveled	Time_rest_days	Capacity	attendance	Surface	Same_surface	Rivalry
13784	2018	10	Pittsburgh Steelers	Carolina Panthers	Win	0.00	11.31	68400.0	62881.0	Grass	Yes	No
13785	2018	10	Carolina Panthers	Pittsburgh Steelers	Lose	363.60	18.31	68400.0	62881.0	Grass	Yes	No
13786	2018	10	Buffalo Bills	New York Jets	Win	277.93	21.00	82500.0	77982.0	Turf	Yes	Division
13787	2018	10	New York Jets	Buffalo Bills	Lose	0.00	21.00	82500.0	77982.0	Turf	Yes	Division
13788	2018	10	New England Patriots	Tennessee Titans	Lose	924.59	12.70	69143.0	69363.0	Grass	No	Conference
13789	2018	10	Tennessee Titans	New England Patriots	Win	0.00	27.86	69143.0	69363.0	Grass	Yes	Conference
13790	2018	10	Tampa Bay Buccaneers	Washington Redskins	Lose	0.00	21.00	65618.0	52667.0	Grass	Yes	Conference
13791	2018	10	Washington Redskins	Tampa Bay Buccaneers	Win	821.85	14.00	65618.0	52667.0	Grass	Yes	Conference
13792	2018	10	Kansas City Chiefs	Arizona Cardinals	Win	0.00	14.00	76416.0	76712.0	Grass	Yes	No
13793	2018	10	Arizona Cardinals	Kansas City Chiefs	Lose	1058.86	28.00	76416.0	76712.0	Grass	Yes	No

- I am not entirely confident in studying dyadic data.

Lessons learned so far

- Webscraping is not easy, even from a prety standardized site like Wikipedia.

Lessons learned so far

- Webscraping is not easy, even from a prety standardized site like Wikipedia.
- I learned a lot about parsing HTML data (which was one of my original goals when I started this project).

Lessons learned so far

- Webscraping is not easy, even from a prety standardized site like Wikipedia.
- I learned a lot about parsing HTML data (which was one of my original goals when I started this project).
- Instead of Markdown, I am using pure LaTeX to make this presentation.

Mitigation plans

- I still need to do more cleaning to my dataset - but I think I'm almost there.

Mitigation plans

- I still need to do more cleaning to my dataset - but I think I'm almost there.
- I do not think I will have enough time to scrape weather/ dome/ stadium size data as Professor Dunford suggested in my project proposal.

Mitigation plans

- I still need to do more cleaning to my dataset - but I think I'm almost there.
- I do not think I will have enough time to scrape weather/ dome/ stadium size data as Professor Dunford suggested in my project proposal.
- My priorities right now are to finish up the very last bits of cleaning fixes to the data, and move onto the machine learning part.

Mitigation plans

- I still need to do more cleaning to my dataset - but I think I'm almost there.
- I do not think I will have enough time to scrape weather/ dome/ stadium size data as Professor Dunford suggested in my project proposal.
- My priorities right now are to finish up the very last bits of cleaning fixes to the data, and move onto the machine learning part.
- I think it's realistic to think that I will not be able to complete the 2020 season part of this project - but if things go well I do want to keep playing with the models even after the semester ends.

Thank you!

Thank you for taking your time to listen to my presentation. Any feedback is greatly appreciated!