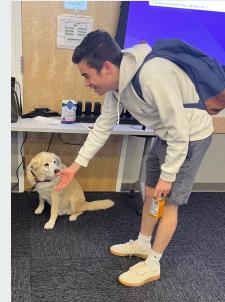




Sports Betting: How Have the Books Done Over Time?

By: Bofan (Will) Chen, Joe Gyorda, Anton Hung, Sean Pietrowicz, and Kevin Rouse (TeamDougFans)
With help from Carly, Doug, and Shrey





Overview

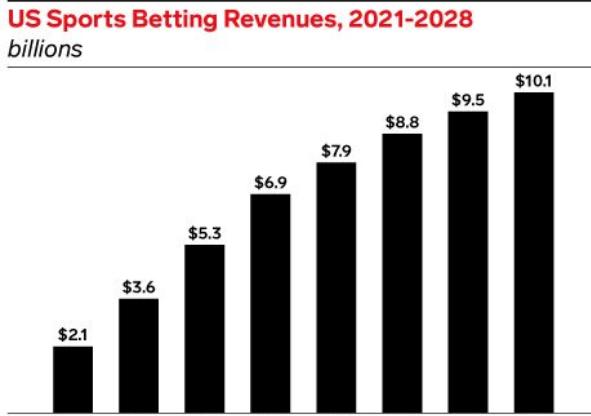
- **Introduction and Motivation**
- **Data Wrangling Methods**
 - Data Acquisition
 - Excel Data Cleaning
 - Tidyverse Data Cleaning
- **Preliminary Analyses**
 - Word Cloud and Exploratory Data Analysis
 - Missing Value Analysis
 - The effect of weather conditions on the accuracy of the betting spread
- **Advanced Modeling**
 - Predictive modeling on the accuracy of the betting spread
 - Network model of team's performance against spread
- **Web Scraping**

Introduction and Motivation (We already miss Doug)





Introduction - Background



266156

eMarketer | InsiderIntelligence.com



<https://www.insiderintelligence.com/content/sports-gambling-opportunity-marketers>

<https://www.techradar.com/news/fanduel-vs-draftkings>



Brief Overview of Sports Betting

FOOTBALL WEEK 1	SPREAD	MONEY	TOTAL
Lovable Losers	+7 -110	+250	Ov 46 -110
FanDuel Favorites	-7 -110	-300	Un 46 -110

- Here, the FanDuel Favorites are a better team than the Lovable Losers
- In order to make betting on the game fair, a spread is created (certain amount of points are added to worse team's score or are subtracted from the better team's score)



Introduction - What will we be analyzing?

Aim 1: How has the accuracy of the spreads set by the books changed over time?

- This will allow us to determine if the methods the bookmakers use are getting better and harder to compete with

Aim 2: Determine possible sources of volatility in spread accuracy

- Does certain weather, team matchups, grass types, etc. affect the accuracy of the spreads
- Can these factors be used to the bettors benefit to try and beat the books

Data Wrangling Methods





Methods - Data Acquisition

- Data is from kaggle.com
- Contains 3 Different csv files

Spread Data

- Includes date of game, home and away team, home and away score, favored team, and spread, weather, etc.

Team Data

- Includes team name, abbreviation, division, and conference

Stadium Data

- Includes stadium name and other information about the stadium such as grass type, climate, etc.



Methods - Excel Data Cleaning

- Raw data did not include results of the game relative to the spread, and a series of VLOOKUPs and IF statements were needed to determine the result of the game relative to the spread
- Added a column which was actual result of game - predicted result(spread)
- Did same procedure for the over/under in game versus predicted
- Merged stadium and team data using VLOOKUPs

FOOTBALL WEEK 1	SPREAD	MONEY	TOTAL
Lovable Losers	+7 -110	+250	Ov 46 -110
FanDuel Favorites	-7 -110	-300	Un 46 -110



Sheet Name	Sheet Description	Sheet Notes
Raw Spreads	Raw data of the spreads of the games and the results	
Raw Stadiums	Raw data of the stadiums the games were played at	
Raw teams	Raw data of NFL teams	
Adding Abbreviations	Add abbreviations from Raw teams to spread data	Used VLOOKUP
Difference Favored	Found Difference between the favored team and the other team	Filtered data to only include games where there was a spread Used IF to determine whether to subtract away-home or home-away using abbreviations from previous table
Comparing spread	Found difference between actual score and spread	**A positive difference means the favored team outperformed the spread, a negative difference means that the favored team underperformed against the spread, a difference of 0 means that the spread was correct*
Comparing Over Under	Found Difference betweeen actual total and predicted	** A positive difference means that more points were scored than predicted. A negative difference means that less points were scored than predicted
Merged Stadium	Merged Stadium Data with Raw spreads	Used a series of vlookups to add the stadium data to the spreads sheet
Merged Teams	Merged Division and conference of the home and away teams	Used VLOOKUPS using the Raw Teams Sheet Note that the data includes pre and post 2002, so different VLOOKUPS were used for these dates



Methods - Tidyverse Data Cleaning

Adding Abbreviations: Select columns, **Rename column** and Save the dataframe as df1, Do a **left join** for spreads by df1

Difference Favored: Use **ifelse function** to create new variable difference_favored_minus_notfavored

Comparing Spread: Take the absolute value of spread_favorite, and then difference favored minus this absolute value

Comparing Over Under: score_home+score_away-over_under_line

Merge Stadium: **Rename column** stadium in table stadium, Merge spread by doing a **left join** with stadium

Merge Teams: Select columns in table teams, Rename column, Do a **left join**

Preliminary Analyses:

I. Word Cloud and Exploratory Data Analysis



Word Cloud - Team



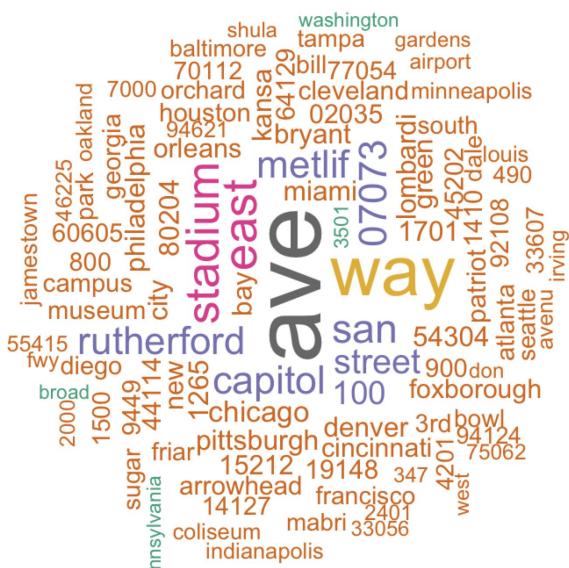
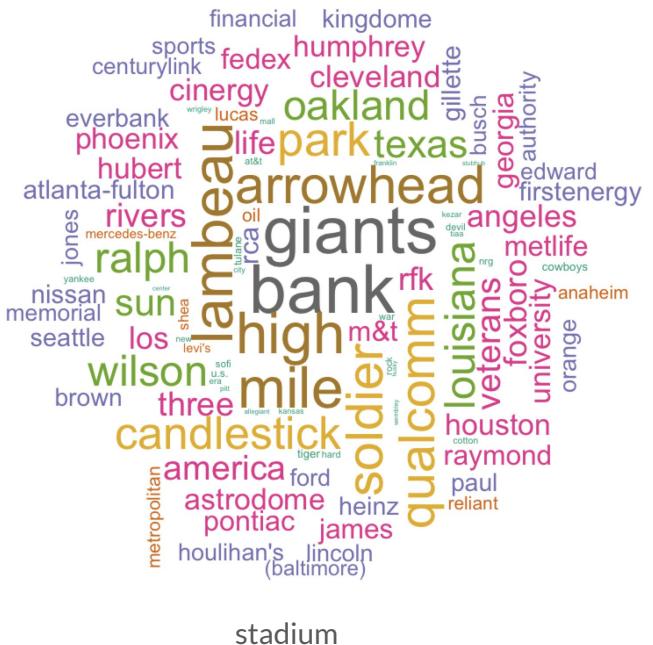
team home



team away



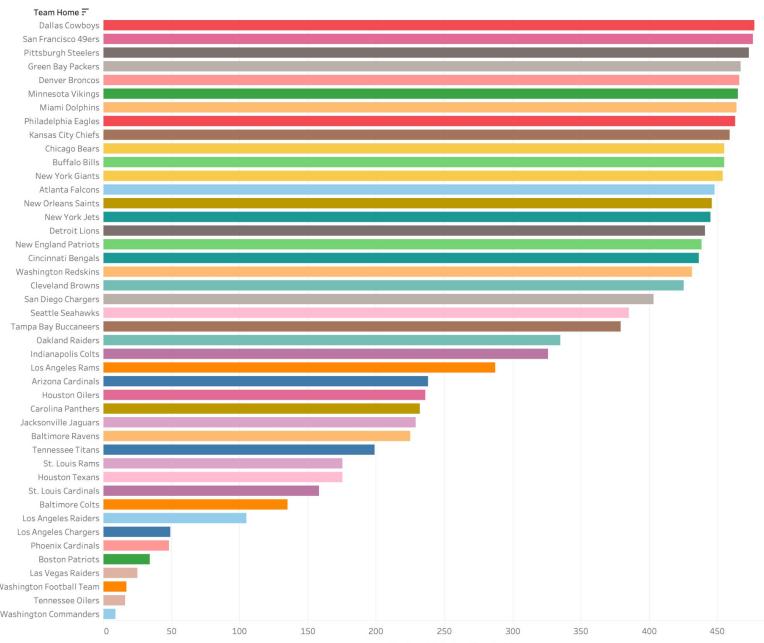
Word Cloud - Stadium



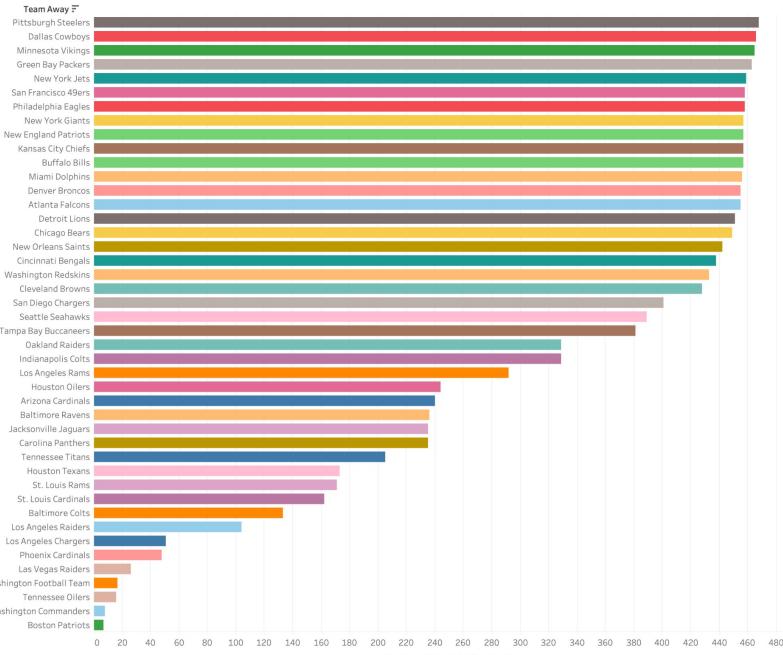


Exploratory Data Analysis - Team

Team Home Order by Number of Events



Team Away Order by Number of Events





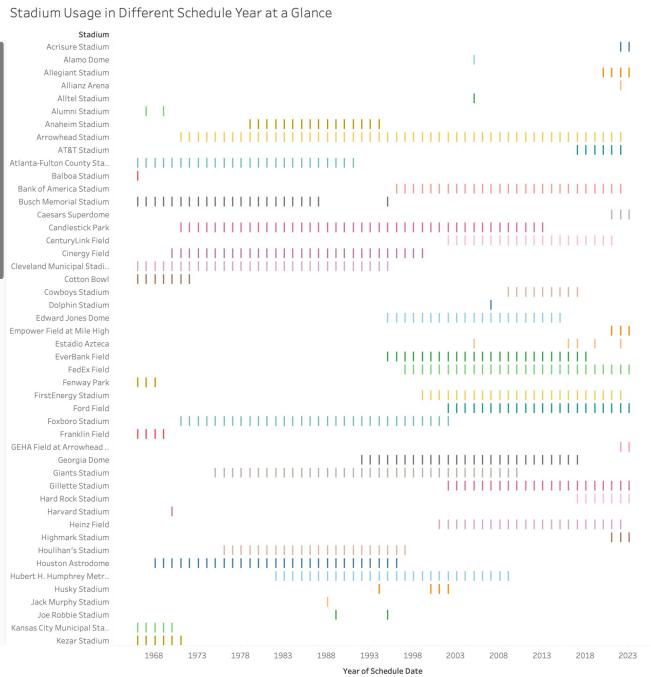
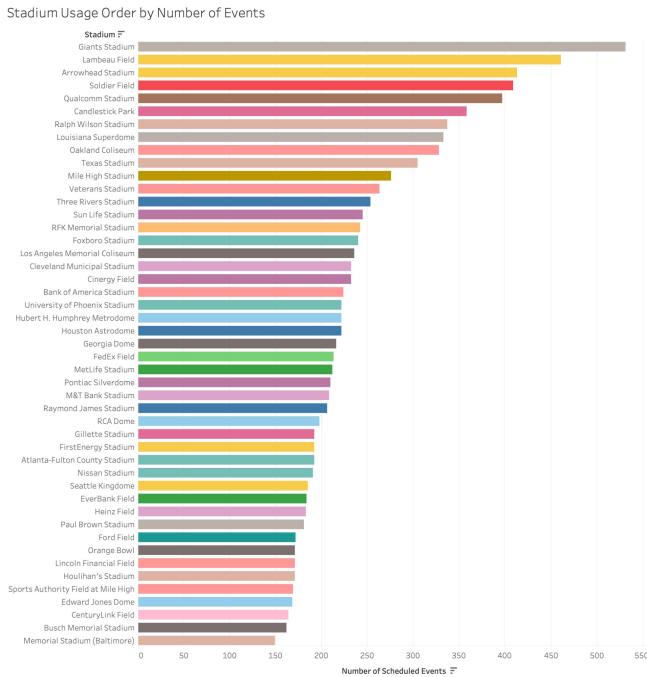
Exploratory Data Analysis - Team

Team Home vs Team Away

		Team Away																																						
Team Home		Arizona Cardinals	Atlanta Falcons	Baltimore Ravens	Baltimore Ravens	Boston Patriots	Buffalo Bills	Carolina Panthers	Chicago Bears	Cincinnati Bengals	Cleveland Browns	Dallas Cowboys	Denver Broncos	Detroit Lions	Green Bay Packers	Houston Oilers	Indianapolis Colts	Jacksonville Jaguars	Kansas City Chiefs	Las Vegas Raiders	Los Angeles Chargers	Los Angeles Rams	Miami Dolphins	Minnesota Vikings	New England Patriots	New Orleans Saints	New York Giants	New York Jets	Oakland Raiders	Philadelphia Eagles	Pittsburgh Steelers	San Diego Chargers	Seattle Seahawks	St. Louis Rams	St. Louis Rams	Tampa Bay Buccaneers	Tennessee Titans	Tennessee Titans	Washington Redskins	Washington Redskins
Arizona Cardinals		7	3	4	8	5	3	4	14	4	11	7	3	3	2	5	1	7	4	8	4	7	13	3	4	14	3	22	22	15	6	1	2	1	12					
Atlanta Falcons		5	4	6	28	15	7	8	15	8	17	17	5	2	4	4	5	1	29	5	15	9	54	13	7	4	18	3	9	4	40	11	7	11	28	3	1	12		
Baltimore Colts		3	1	13	4	6	5	4	3	4	5	4	3	4	5	4	4	5	14	3	13	2	14	14	4	3	5	2	4	4	2	4	2	1	12	2	1	4		
Baltimore Ravens		4	3	7	3	3	27	24	4	8	3	3	1	6	10	12	7	2	6	1	6	5	5	4	6	7	3	27	4	3	3	3	3	2	9	2				
Boston Patriots		1	5	1	1	2	3	3	3	2	4	4	4	4	4	4	4	4	1	4	1	1	5	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4		
Buffalo Bills		2	7	14	4	1	5	7	15	11	5	20	6	7	14	5	23	9	18	2	6	5	59	9	56	5	7	56	10	8	1	16	10	5	5	2	3	6	6	
Cincinnati Bengals		12	29	14	4	3	5	5	4	3	8	4	7	9	2	3	4	3	1	3	6	4	28	5	5	3	6	4	3	12	8	10	23	3	1	1	8			
Chicago Bears		5	14	4	4	4	7	6	7	5	12	7	57	57	4	4	5	4	7	1	2	14	6	55	10	19	13	7	5	17	11	5	5	29	2	1	15			
Cleveland Browns		5	8	3	27	1	18	5	9	49	6	16	6	6	20	6	10	13	15	1	2	5	4	13	8	13	7	12	10	5	1	54	17	8	12	2	3	8		
Dallas Cowboys		14	15	3	2	12	4	8	6	51	10	14	7	9	25	6	9	13	2	1	6	17	6	14	8	10	9	15	14	6	8	5	25	1	1	5				
Denver Broncos		4	8	3	8	8	4	9	17	17	7	7	7	7	10	6	10	8	57	2	3	4	7	1	2	10	7	57	6	12	12	9	4	10	2	1	1	5		
Detroit Lions		9	22	4	3	6	4	56	7	8	14	7	58	2	3	4	4	7	1	2	14	10	13	6	15	57	4	3	29	11	20	4	14	1	3	6	25	54		
Green Bay Packers		4	18	3	4	7	8	56	8	6	15	8	57	3	2	5	3	6	2	17	7	59	6	13	13	8	7	16	4	5	23	16	4	6	28	1	4	9		
Houston Oilers		1	4	2	1	10	1	3	28	26	4	12	5	3	2	14	3	2	1	2	14	5	4	12	2	7	4	2	12	8	3	1	28	10	5	6	2	4	2	
Houston Texans		2	3	6	6	6	4	2	7	7	3	3	2	3	22	21	9	2	1	5	3	7	2	3	4	6	3	3	2	3	2	2	21	1	2	2				
Indians Polts		3	5	7	21	4	5	13	10	5	13	5	5	5	21	22	9	1	1	2	24	4	26	5	4	24	5	5	1	8	14	5	5	1	3	6	22	6		
Jacksonville Jaguars		4	4	11	9	3	4	11	9	3	7	4	2	21	22	7	1	2	1	8	2	5	4	4	6	3	4	14	5	4	5	1	4	2	5	4				
Kansas City Chiefs		4	5	4	1	23	4	6	17	15	6	56	7	8	16	5	11	7	3	6	14	4	12	7	10	5	8	13	41	5	21	51	8	27	2	3	6	8		
Las Vegas Raiders		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
Los Angeles Chargers		1	1	1	1	1	1	1	2	1	6	1	1	1	1	2	1	6	3	1	2	2	2	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1		
Los Angeles Raiders		2	3	2	5	5	1	14	1	2	2	2	3	11	3	2	2	2	3	1	2	2	3	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
Los Angeles Rams		9	30	5	1	4	2	13	5	4	14	5	13	13	3	1	2	1	3	1	2	4	13	5	27	15	5	2	5	4	37	9	5	1	6	1	10			
Miami Dolphins		3	9	15	12	1	59	4	8	11	9	5	9	6	9	12	6	23	3	18	2	5	4	57	5	5	58	16	8	1	16	14	6	10	2	3	5	1	9	
Minnesota Vikings		10	17	5	2	6	10	19	8	56	6	5	10	19	8	56	55	5	2	5	6	2	14	7	6	18	13	6	7	11	1	8	11	2	2	5	1	7		
New England Patriots		4	12	12	11	51	4	5	12	12	8	18	7	7	9	7	30	9	13	1	2	3	4	53	7	9	7	54	5	6	16	11	8	11	2	2	5	1	7	
New Orleans Saints		6	94	1	3	29	13	7	11	16	5	15	14	5	3	6	3	7	1	1	2	28	8	19	6	15	6	5	19	1	10	6	51	7	6	10	3	14		
New York Giants		13	15	4	4	6	8	11	4	9	54	8	16	16	3	3	5	4	7	1	1	12	15	15	5	18	7	7	5	59	6	10	6	16	10	21	4	10		
New York Jets		5	14	6	6	1	57	3	6	17	10	7	12	6	6	10	5	23	11	15	1	1	1	1	56	6	56	6	15	6	15	9	8	11	2	3	8	4		
Oakland Raiders		3	5	1	5	3	6	12	11	6	45	6	3	6	5	6	43	3	2	16	18	5	10	4	4	16	5	12	38	4	4	15	6	3	6	6	6			
Philadelphia Eagles		14	19	4	3	6	7	13	9	8	57	7	10	13	3	3	6	3	4	1	2	10	7	17	9	16	57	6	5	6	6	10	5	13	9	20	7	12		
Phoenix Cardinals		2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Pittsburgh Steelers		2	8	5	31	13	4	8	53	57	10	15	10	7	28	4	12	13	16	2	1	5	13	9	23	7	8	12	13	9	18	8	12	5	3	2	10			
San Diego Chargers		4	6	3	7	13	3	7	17	11	6	51	5	6	11	3	10	4	51	13	2	18	4	12	6	5	15	39	6	14	7	7	2	5	6	6				
San Francisco 49ers		22	41	5	4	8	11	21	9	8	18	19	4	3	5	2	7	1	3	37	9	20	6	38	20	6	5	16	2	8	7	23	5	21	16	3	1	1	14	
Seattle Seahawks		22	10	2	3	1	9	7	8	8	11	20	29	10	8	5	2	6	4	26	1	13	5	11	8	11	10	13	10	2	8	26	26	1	7	2	3	10		
St. Louis Rams		15	10	3	2	10	8	3	3	3	2	6	1	2	3	3	3	3	3	6	3	5	2	6	21	1	1	6	1	1	5	1	2	2	3	6	8			
Tampa Bay Buccaneers		4	3	4	10	22	32	5	6	7	4	29	30	1	3	7	3	9	1	12	7	27	5	28	15	5	3	9	2	7	5	10	8	6	6	1	2	13		
Tennessee Oilers		2	1	1	1	1	1	1	2	1	1	2	1	2	1	1	1	2	1	1	5	2	4	4	4	7	8	3	10	3	3	2	4	4	4	1	1	1	1	
Washington Commanders		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
Washington Football Team		13	14	3	4	7	8	13	5	7	56	6	18	10	4	3	5	3	5	4	14	4	13	54	7	4	54	6	5	6	15	10	20	6	10	3	3	2	2	2

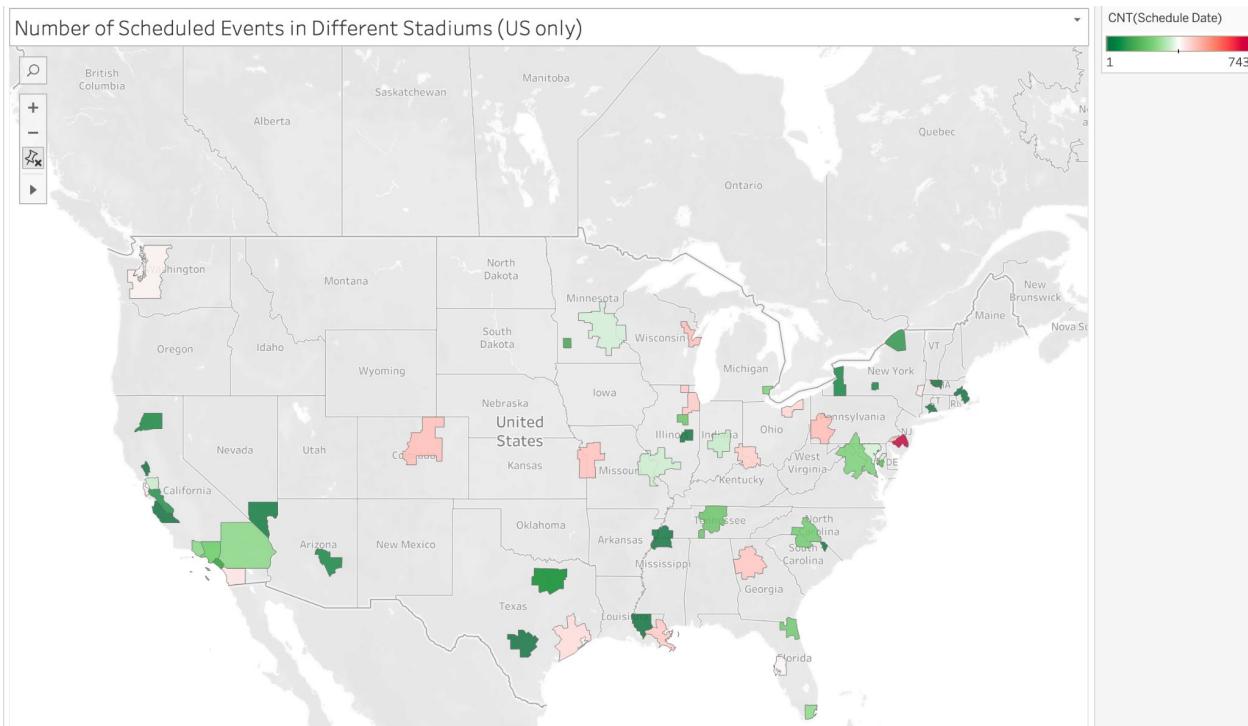


Exploratory Data Analysis - Stadium





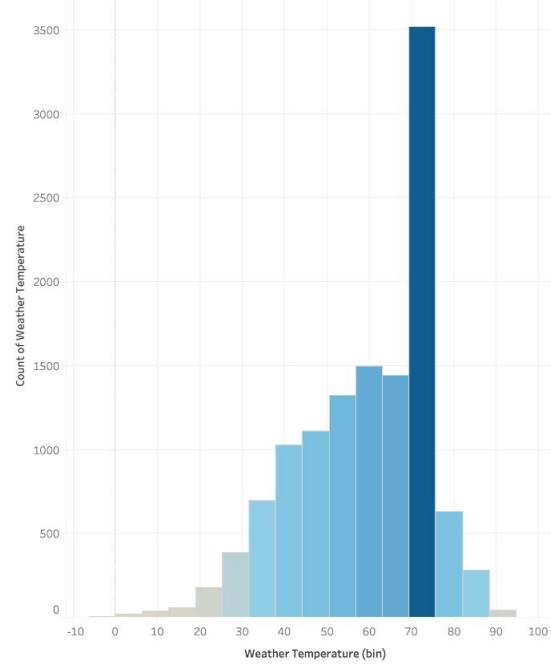
Exploratory Data Analysis - Stadium



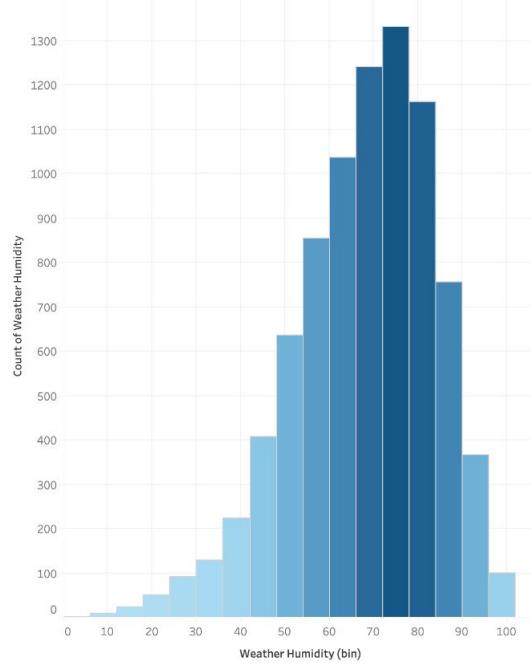


Exploratory Data Analysis - Weather

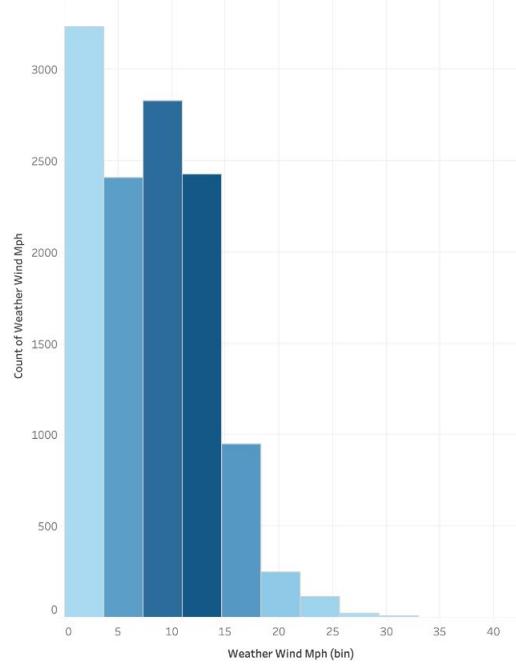
Distribution of Temperature



Distribution of Humidity



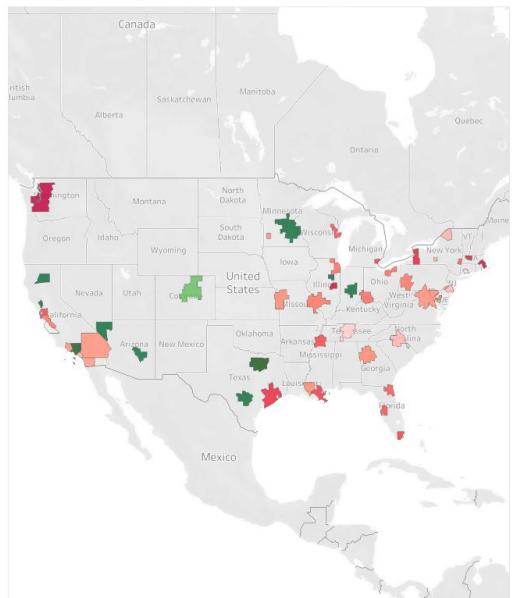
Distribution of Wind Speed



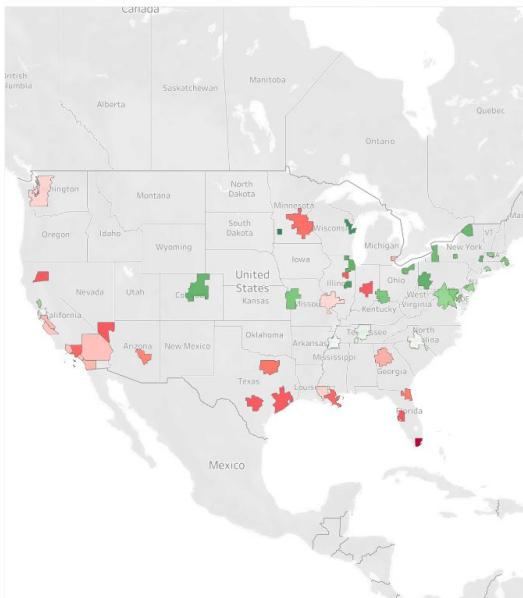


Exploratory Data Analysis - Weather

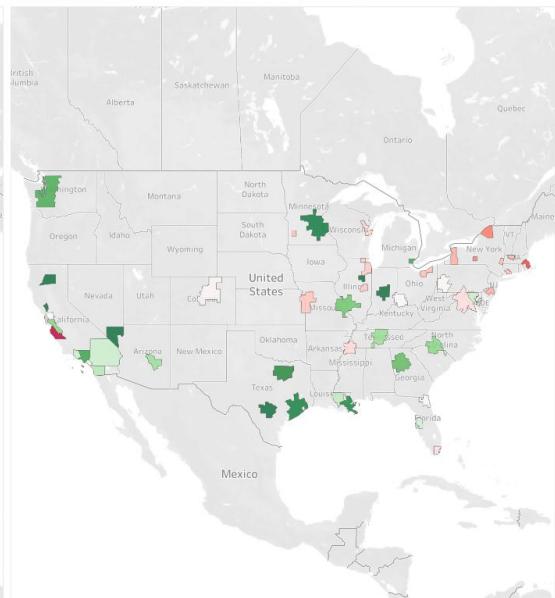
Average Humidity in Each Stadium (US Only)



Average Temperature in Each Stadium (US Only)

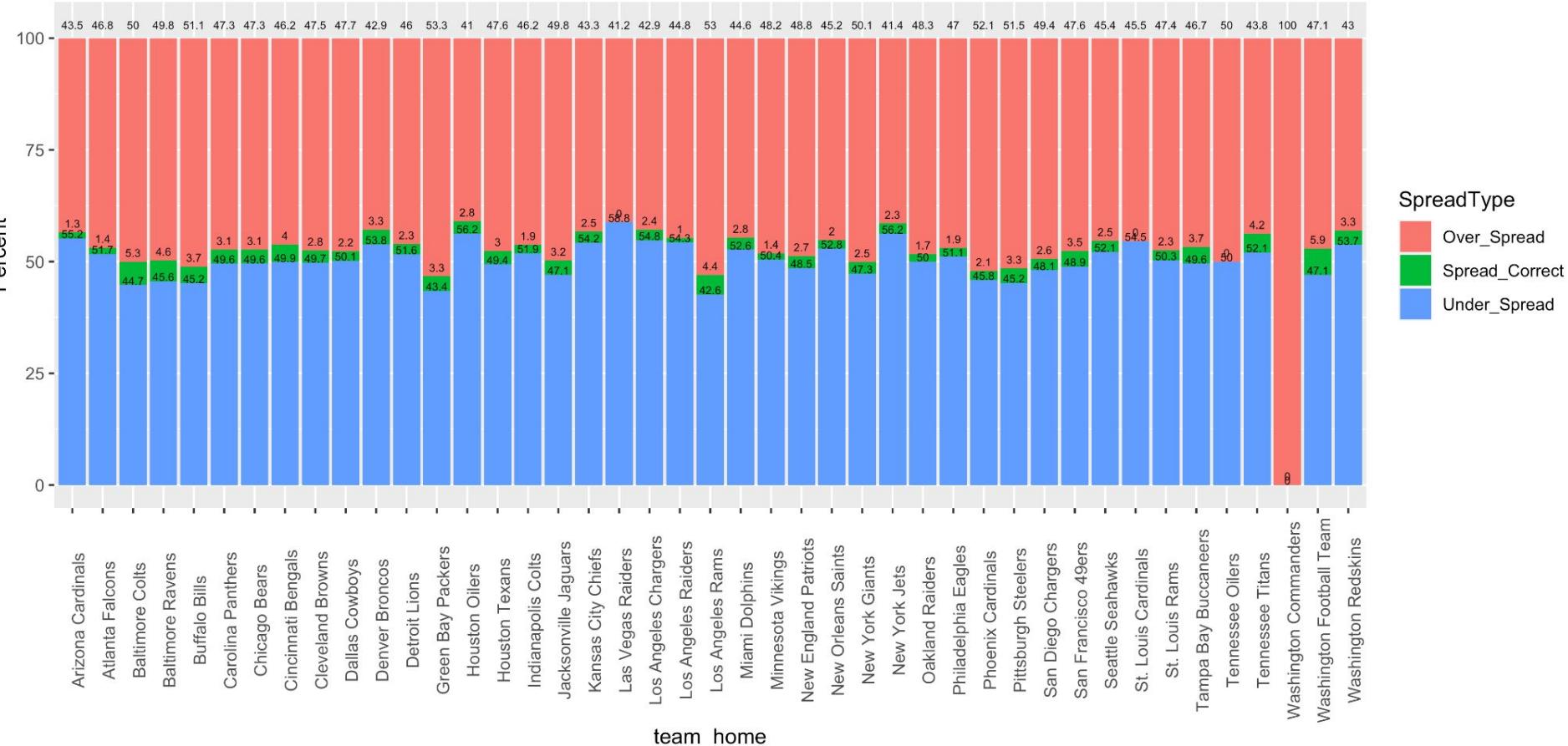


Average Wind Speed in Each Stadium (US Only)





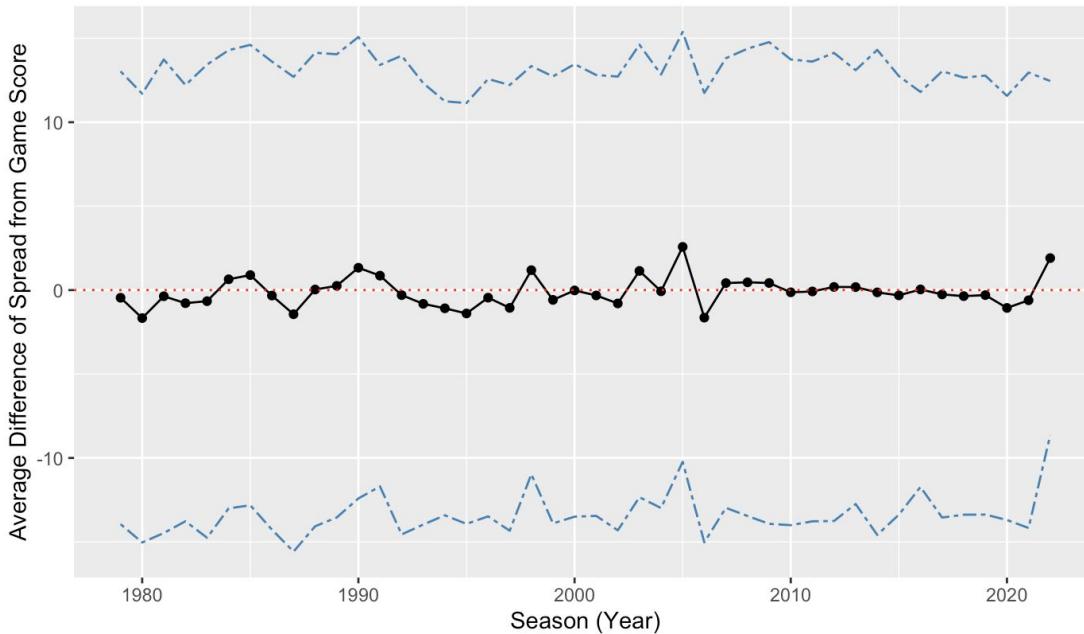
Accuracy of Spread Across NFL History for Each NFL Team





Variation in Spread Accuracy Over Time

Average Accuracy of the Spread per Year for the Entire NFL





Preliminary Analyses:

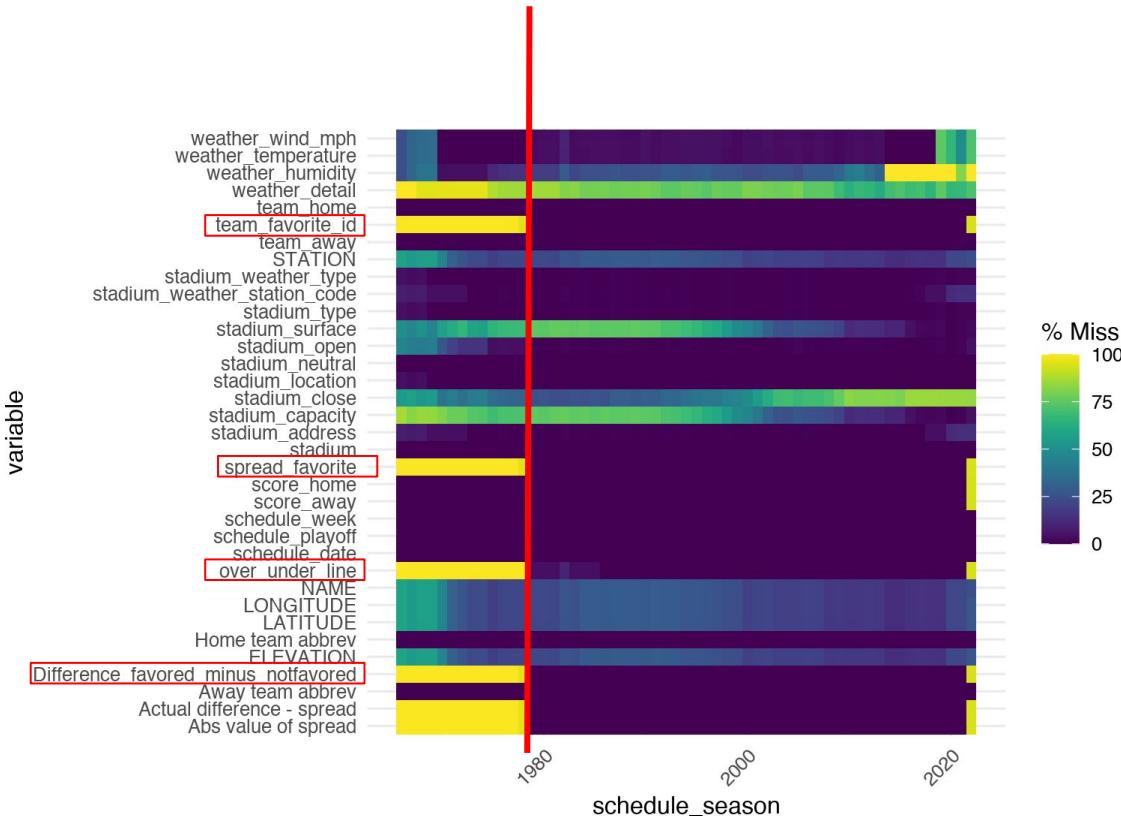
ii. The effect of weather conditions on the accuracy of the betting spread



Missing Values

Betting-related data missing pre-1978

- Favored team
- Spread
- Over/under line



Weather and Football



Go Bills!!!!



Variables related to weather conditions

- Temperature
- Wind (mph)
- Humidity
- Whether or not the stadium has a dome (roof)

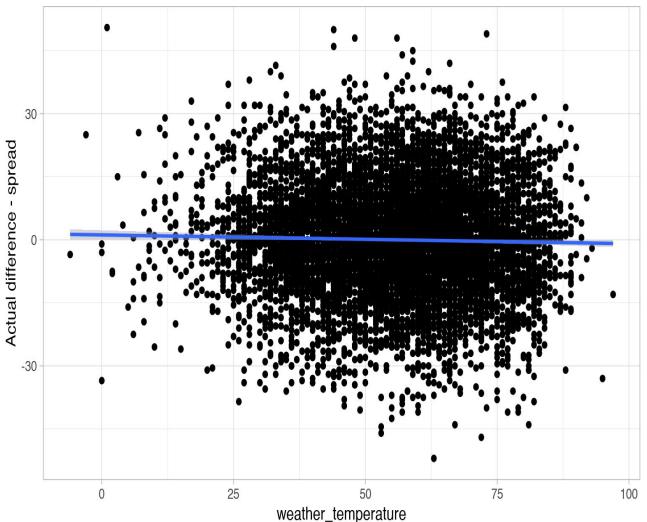
Outcome variable

- Our measurement of the accuracy of the spread

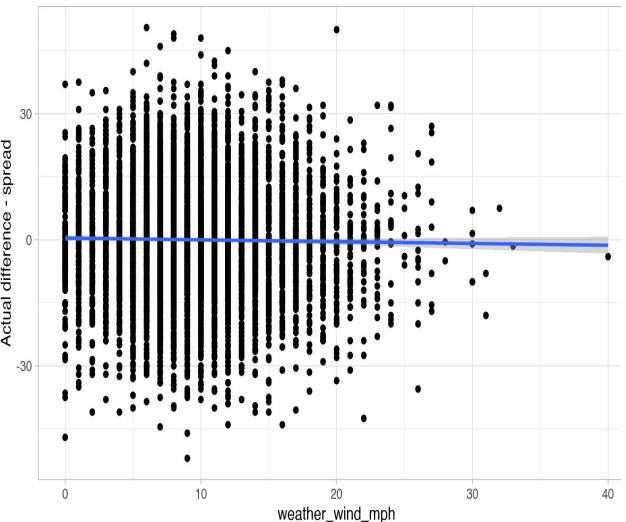
Continuous variables can be visualized in a scatter plot with regression analysis



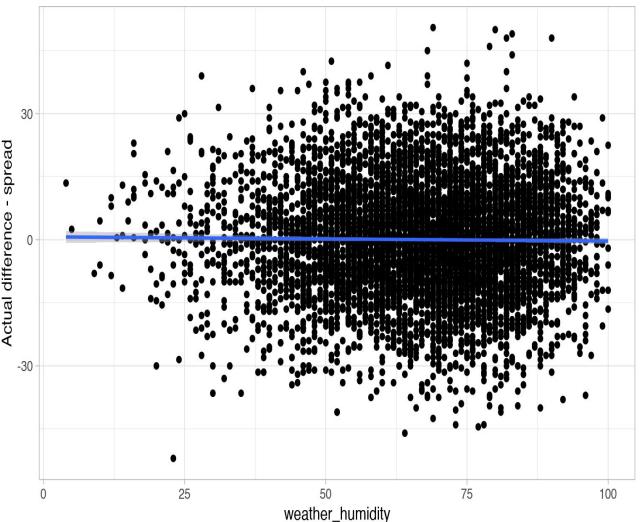
Examining the correlation between weather variables and the accuracy of the spread



Temperature:
 $r = -0.02477375$



Wind:
 $r = -0.01495849$



Humidity:
 $r = -0.01162643$

The spread accuracy is not affected by any of these variables



Categorizing weather conditions as “Ideal”, “ok” or “Poor” for playing football

Poor weather (n=1120):

- Wind speed > 12 mph AND
- Temperature < 45 OR Temperature > 67.5

Ideal weather (n=3536)

- Wind speed < 6 mph AND
- Temperature between 45-67.5

OR

- Domed stadium

Ok weather (n=5276)

- Anything else

> summary(non_domed_stadiums\$weather_temperature)						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-6.00	45.00	57.00	55.83	67.50	97.00	799
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.000	6.000	9.000	9.538	12.000	40.000	810





Poor weather:

- Wind > 12 mph
- Temperature < 45 or > 67.5

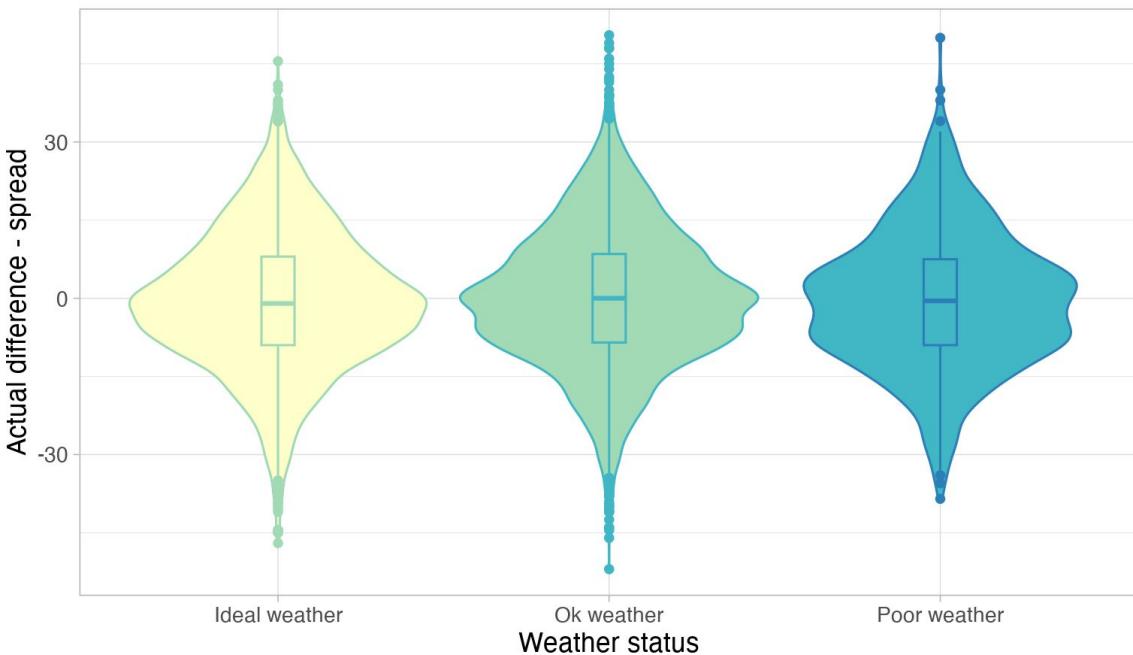
Ideal weather

- Wind < 6 mph
- Temperature between 45-67.5

OR

- Domed stadium

Accuracy of the spread vs weather type at the stadium





Summary of the effect of weather

No observed association between the weather conditions of the game and the accuracy of the betting spread.

- The betting spreads that are provided have been well adjusted to account for the effect of weather on the game result

Limitations:

- Arbitrary cutoff values for categorization of weather variables
- Many other weather conditions were not considered (rain, snow)



Doug Break

...
Anyone gonna do a
question? Bork bork



Advanced Modeling

I. Predictive modeling on the accuracy of the betting spread

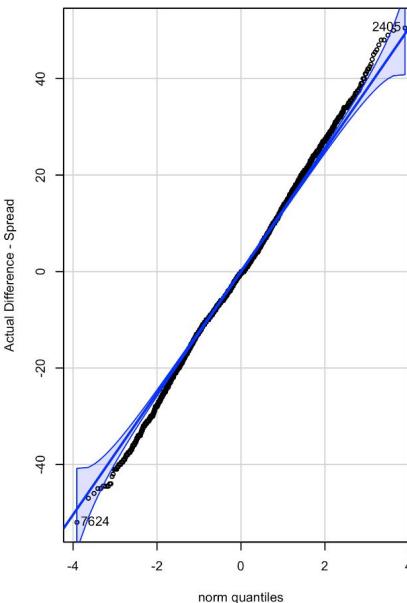
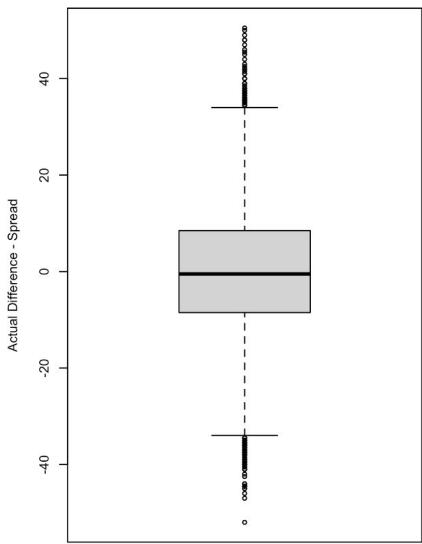
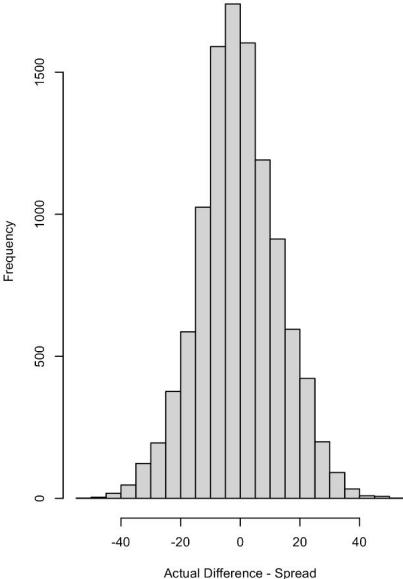


Assessing Influence of Weather and Stadium Conditions on Spread Accuracy

- **Intuition:** Can we predict spread accuracy (the difference between an NFL game spread and the actual difference in game score) with a simple linear model?
- However, **an important assumption of regression is violated:** the observations (e.g., individual games) **are not independent** of one another!
 - Games are “nested” within teams and season, and the effects of weather/stadium on spread may be different for a given team or year
- To address the violation of independence, we can perform a **linear mixed-effects model**, which allows for the inclusion of both **fixed effects** (e.g., regular regression terms) and **random effects** (e.g., regression terms that we allow to vary across groups)
- Note: Only complete cases were used (all NAs dropped)



Assessing Normality in Outcome





Linear Mixed-Effects Model Formula

Accuracy_of_Spread = temperature + wind_speed + humidity + weather_type + stadium_design + stadium_weather + stadium_playing_surface + stadium_elevation + (1|schedule_year) + (schedule_year|team)

- The orange terms are **fixed/main effects** - these are the same as we'd see in a normal regression
- The green terms are **random effects**
 - We have a different (random) intercepts of the model for each year that games were played in, and we also have different slopes for each NFL team that played games.



Model Performance and Conclusions

Random effects:

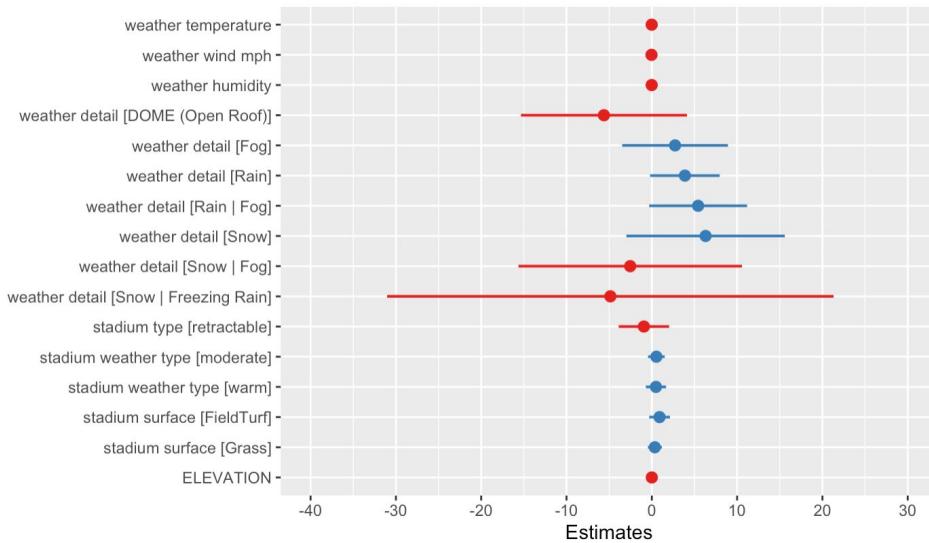
Groups	Name	Variance	Std.Dev.	Corr
schedule_season	(Intercept)	2.430e-06	0.001559	
team_favorite_id	(Intercept)	1.987e+02	14.095487	
	schedule_season	5.619e-05	0.007496	-1.00
Residual		1.773e+02	13.317210	

Number of obs: 6080, groups: schedule_season, 43; team_favorite_id, 33

- $R^2 = 0.0036$ (marginal), 0.0079 (conditional)
- No signal! Weather and stadium conditions do not explain any substantial variance in spread accuracy
- Model is probably a poor fit, but there is also so much noise in the data

Fixed effects:

Predicting Actual Difference - Spread



Advanced Modeling

II. Network model of team's performance against spread

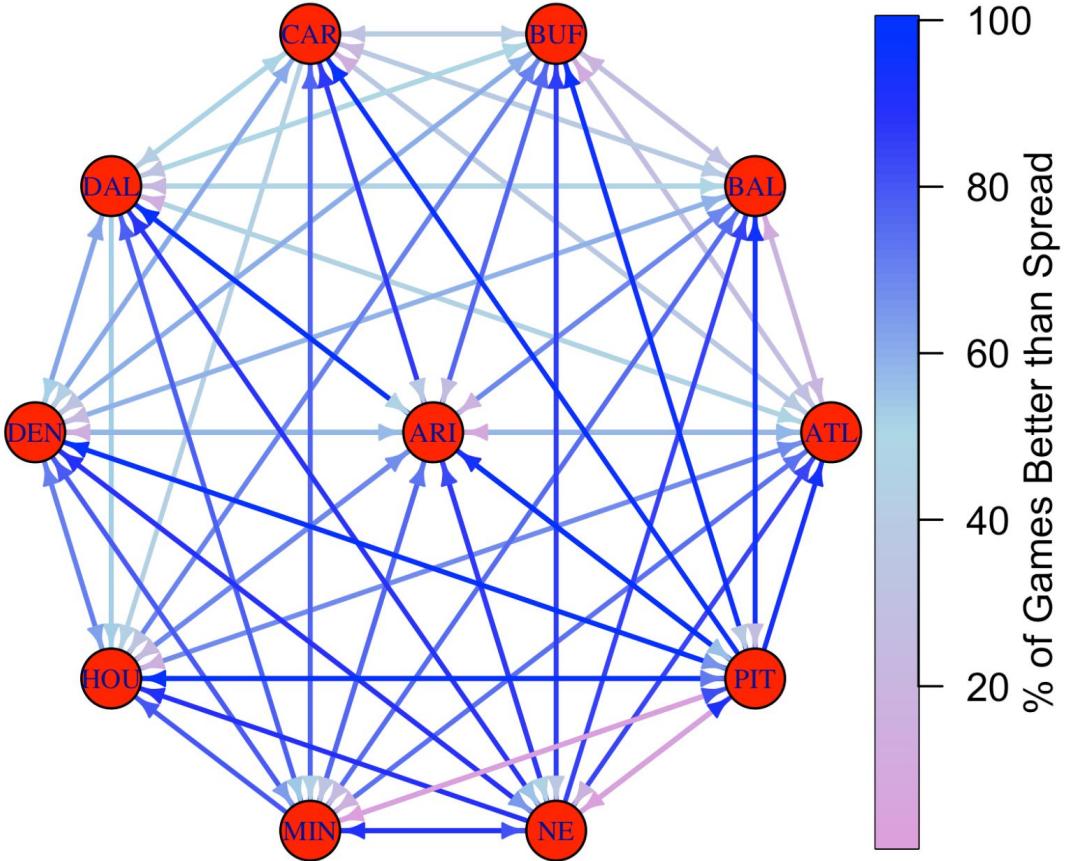


Assessing Historical Team Performance Against Spread with a Network - Intuition

- A social network is a representation of relationships between social actors or nodes, with analyses of these networks focusing on the strength of these relationships (represented as edges between nodes)
- **Nodes:** NFL Teams (32)
- **Edges:** Directed - games played between teams where one team was favored over the other
- **Edge weights:** Directed, the proportion of games when the outgoing team was favored that the spread was equaled or exceeded (e.g., actual game score difference \geq spread)
- This representation will allow us to easily visualize the performance of NFL teams against the spread against each other team



Network Model





Conclusions from Network Model

- Only three teams had instances where they've never been favored to win against another team historically (Denver → Arizona; Philadelphia → Houston; Green Bay → Jacksonville)
- The majority of teams (21) equal or exceed the spread >50% of the time on average, with Baltimore (61.3%), Green Bay (59.5%), and Buffalo (55.8%) performing the best
 - Based only on historical data, these teams might be the best to bet on
- Atlanta (46.3%), Las Vegas (47.1%), and New York (48.3%) are the worst on average when it comes to equalling or exceeding the spread
 - Might be best to bet against these teams!

Web Scraping Supplementary Weather Data





Scraping Weather Data for Missing Values

- Sourced weather data from <http://nflweather.com/> which had temperature, wind and forecast details from 2011-2021

NFLWeather™ Forecast Week 1															Updated 09/12 @ 10:00 PM EST	
					P1			P2			P3					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		16	17	18												
Game				Time (ET)				Forecast				Wind				
►		@		Final:	31	-	10					DOME		7m W	<button>Details</button>	
►		@		Final:	27	-	26					DOME		2m SSW	<button>Details</button>	
►		@		Final:	10	-	19					68f Rain		9m NE	<button>Details</button>	
►		@		Final:	23	-	20					75f Mostly Cloudy		4m S	<button>Details</button>	
►		@		Final:	38	-	35					DOME		8m SSE	<button>Details</button>	
►		@		Final:	7	-	20					90f Humid and Partly Cloudy		10m ESE	<button>Details</button>	
►		@		Final:	24	-	9					74f Light Rain		3m SSW	<button>Details</button>	

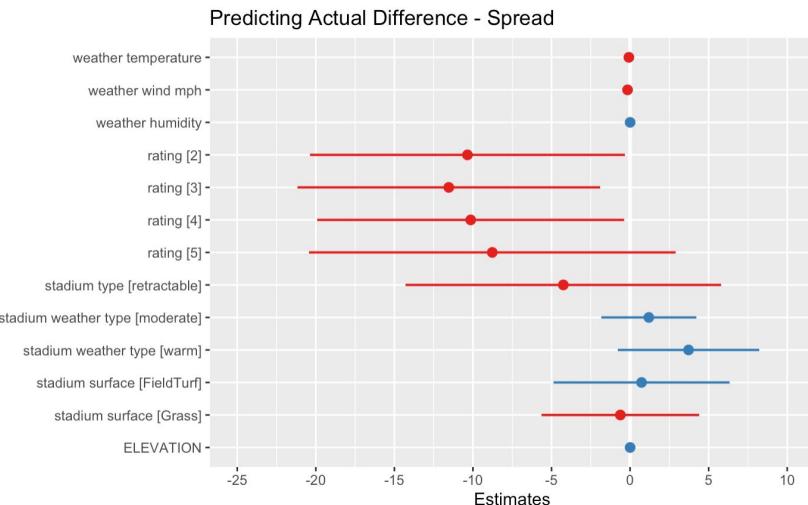
weather_temperature	weather_wind_mph	weather_humidity	weather_detail
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
72	0	NA	DOME
NA	NA	NA	NA
72	0	NA	DOME
72	0	NA	DOME
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
72	0	NA	DOME
72	0	NA	DOME
NA	NA	NA	NA
NA	NA	NA	NA
72	0	NA	DOME
NA	NA	NA	NA
NA	NA	NA	NA

wind	temperature	weather	descriptor	rating
10	40	Overcast	ok	3
5	56	Clear	great	5
5	48	Mostly Cloudy	ok	3
9	68	Clear	great	5
4	33	Overcast	ok	3
0	71	DOME	great	5
0	71	DOME	great	5
5	47	Overcast	ok	3
7	55	Clear	great	5
5	75	Clear	great	5
9	62	Clear	great	5
3	75	Mostly Cloudy	ok	3
6	37	Drizzle	ok	3
8	42	Overcast	ok	3
10	52	Partly Cloudy	good	4
0	71	DOME	great	5
0	71	DOME	great	5
4	85	Clear	great	5
7	60	Overcast	ok	3
0	71	DOME	great	5
5	68	Clear	great	5
8	42	Overcast	ok	3



Re-Running the Linear Mixed-Effects Model

- Before: $R^2 = 0.0036$ (marginal), 0.0079 (conditional)
- After: $R^2 = 0.0289$ (marginal), 0.1265 (conditional)
 - Still not good, but reflects small improvement



Conclusions, Next Steps, Limitations





Closing Remarks

Conclusions in relation to our aims

- Despite great game-to-game variation in the accuracy of the spread, on average, the spread seems to be fairly consistent with true game scores and has been throughout NFL history
- Weather and stadium data do not seem to relate to the accuracy of the spread
- Historically, might be best to bet on Baltimore and not on Atlanta

Limitations

- Lots of missing weather data – we chose not to impute
- The NIH would not approve our grant proposal, so we could not see this project to its full potential (just kidding)

Future work

- We did not examine data related to over/under
- We could rerun the present analyses with more expansive weather data for all years (were unable to due to time constraints)
- We could repeat these analyses on another professional sport, should data be available
- Kevin will use these models and see a therapist about his gambling addiction

Than

