Predicting the 2020 Superbowl Champion

Nebe Samuel + Ray LaForte

What Data Did We Select?

- For this project we decided to utilize an NFL dataset in hopes of predicting the NFL champion for the 2020 season. Although we are unable to truly answer this question, as the NFL season is still underway, we pulled in data from current NFL rankings to gage if our Superbowl predictions are on track, based off the season so far.
- The dataset that we used came from “pro-football-reference.com”, and was a combination of a couple of analytical breakdowns listed on the site. We were ultimately drawn to this site, because it is a reputable reference in football analytics, amongst the football community.
- The final csv file that was utilized was a dataset containing NFL data across several categories from 2017 to the current point of the 2020 season. As we’ll discuss later we used the historical stats from the dataset to ultimately train the models we built in hopes of gaging whether our 2020 predictions are correct based on the current standings.
- Snippets of Data tables from PFR that were used to get to our final dataset are shown below:

Team Offense Share & more ▾ Glossary Toggle Per-Game Stats

				Tot Yds & TO						Passing						Rushing				Penalties				
Rk	Tm	G	PF	Yds	Ply	Y/P	TO	FL	1stD	Cmp	Att	Yds	TD	Int	NY/A	1stD	Att	Yds	TD	Y/A	1stD	Pen	Yds	1stPy	Sc%	TO%	EXP
1	Green Bay Packers	11	349	4322	697	6.2	9	5	248	261	381	2981	33	4	7.6	152	304	1341	9	4.4	74	58	506	22	47.8	7.0	161.11
2	Kansas City Chiefs	11	348	4684	725	6.5	8	6	275	298	432	3460	30	2	7.7	179	278	1224	10	4.4	73	76	655	23	52.3	7.3	192.86
3	Tampa Bay Buccaneers	12	344	4353	768	5.7	16	5	262	307	475	3197	28	11	6.5	169	277	1156	11	4.2	60	67	583	33	44.9	11.0	95.09
4	Seattle Seahawks	11	341	4301	704	6.1	14	4	255	278	393	3012	31	10	7.0	158	276	1289	11	4.7	75	61	475	22	46.7	11.5	125.03
5	New Orleans Saints	11	326	4026	718	5.6	11	7	253	257	353	2544	18	4	6.8	136	346	1482	19	4.3	100	62	712	17	47.5	8.3	112.84
6	Tennessee Titans	11	324	4243	709	6.0	5	1	260	218	336	2503	23	4	7.1	138	357	1740	15	4.9	95	62	566	27	47.4	4.4	165.77
7	Pittsburgh Steelers	11	317	3812	730	5.2	11	5	235	294	435	2722	25	6	6.1	146	285	1090	10	3.8	63	60	497	26	42.1	7.1	93.18
8	Arizona Cardinals	11	304	4441	745	6.0	13	4	280	264	387	2726	19	9	6.7	150	340	1715	18	5.0	104	85	684	26	45.3	11.1	121.78
9	Indianapolis Colts	11	302	4070	722	5.6	12	3	250	265	399	2922	16	9	7.1	147	311	1148	12	3.7	79	67	666	24	43.6	10.3	79.64
10	Buffalo Bills	11	299	4097	695	5.9	16	8	261	272	395	2949	24	8	7.1	148	277	1148	10	4.1	87	77	725	26	47.8	13.0	117.18
11	Atlanta Falcons	11	295	4123	762	5.4	11	3	255	277	428	2985	17	8	6.5	165	306	1138	12	3.7	64	64	565	26	47.6	8.9	99.54
12	Minnesota Vikings	11	292	4258	662	6.4	19	8	244	218	320	2655	23	11	7.8	132	321	1603	14	5.0	93	63	478	19	39.2	15.0	93.75
13	Las Vegas Raiders	11	292	3937	694	5.7	15	11	237	244	353	2555	19	4	6.9	131	322	1382	13	4.3	84	65	582	22	48.7	13.3	95.72
14	Miami Dolphins	11	284	3431	664	5.2	13	5	224	234	351	2383	18	8	6.4	133	289	1048	9	3.6	65	49	410	26	42.5	10.8	31.06
15	Baltimore Ravens	11	282	3658	679	5.4	14	6	208	186	297	1924	16	8	5.9	103	355	1734	12	4.9	92	72	715	13	39.2	11.7	40.31

Passing Offense Share & more ▾ Glossary Toggle Per-Game Stats

Rk	Tm	G	Cmp	Att	Cmp%	Yds	TD	TD%	Int	Int%	Lng	Y/A	AY/A	Y/C	Y/G	Rate	Sk	Yds	NY/A	ANY/A	Sk%	4QC	GWD	EXP
1	Kansas City Chiefs	11	298	432	69.0	3460	30	6.9	2	0.5	75	8.2	9.4	11.9	314.5	115.0	15	84	7.7	8.9	3.4	2	2	188.32
2	Tampa Bay Buccaneers	12	307	475	64.6	3197	28	5.9	11	2.3	50	6.9	7.1	10.7	266.4	94.9	16	103	6.5	6.6	3.3	2	2	113.26
3	Los Angeles Chargers	11	286	434	65.9	3087	23	5.3	7	1.6	72	7.4	7.8	11.3	280.6	98.9	24	140	6.7	7.1	5.2	1	1	108.50
4	Houston Texans	11	250	364	68.7	3050	24	6.6	5	1.4	77	8.8	9.5	12.8	277.3	112.2	28	151	7.8	8.4	7.1			126.45
5	Seattle Seahawks	11	278	393	70.7	3012	31	7.9	10	2.5	62	8.2	8.6	11.6	273.8	110.8	35	204	7.0	7.4	8.2	2	2	116.92
6	Atlanta Falcons	11	277	428	64.7	2985	17	4.0	8	1.9	63	7.4	7.3	11.4	271.4	92.3	28	178	6.5	6.5	6.1			119.40
7	Carolina Panthers	12	285	410	69.5	2981	15	3.7	10	2.4	75	7.6	7.2	10.9	248.4	93.8	22	139	6.9	6.6	5.1			85.17
8	Green Bay Packers	11	261	381	68.5	2981	33	8.7	4	1.0	78	8.1	9.4	11.9	271.0	117.6	12	119	7.6	8.8	3.1	1	2	155.71
9	Buffalo Bills	11	272	395	68.9	2949	24	6.1	8	2.0	49	7.8	8.1	11.3	268.1	103.8	23	133	7.1	7.3	5.5	2	3	129.89
10	Dallas Cowboys	11	295	457	64.6	2948	16	3.5	10	2.2	58	7.0	6.7	10.8	268.0	87.6	31	251	6.0	5.8	6.4	3	3	2.35
11	Indianapolis Colts	11	265	399	66.4	2922	16	4.0	9	2.3	55	7.5	7.3	11.3	265.6	92.7	12	73	7.1	6.9	2.9	2	2	106.36
12	Los Angeles Rams	11	271	403	67.2	2910	16	4.0	10	2.5	56	7.5	7.2	11.1	264.5	92.3	15	111	7.0	6.7	3.6	0	1	56.08
13	San Francisco 49ers	11	256	384	66.7	2760	14	3.6	12	3.1	76	7.7	7.1	11.6	250.9	89.0	27	210	6.7	6.1	6.6	1	1	75.70

Team Defense Share & more ▾ Glossary Toggle Per-Game Stats

					Tot Yds & TO						Passing						Rushing				Penalties				
Rk	Tm	G	PF	Yds	Ply	Y/P	TO	FL	1stD	Cmp	Att	Yds	TD	Int	NY/A	1stD	Att	Yds	TD	Y/A	1stD	Pen	Yds	1stPy	Sc%	TO%	EXP
1	Pittsburgh Steelers	11	188	3288	670	4.9	23	7	187	198	363	2125	16	16	5.3	109	266	1163	6	4.4	60	72	752	18	26.0	17.3	53.60
2	Miami Dolphins	11	205	4066	710	5.7	19	8	237	240	385	2637	13	11	6.4	130	300	1429	13	4.8	88	62	611	19	27.6	15.4	-9.95
3	Baltimore Ravens	11	214	3664	709	5.2	17	11	234	268	409	2436	15	6	5.6	147	273	1228	8	4.5	62	47	382	25	30.0	13.3	-9.80
4	Los Angeles Rams	11	215	3264	695	4.7	18	7	205	251	396	2239	11	11	5.2	124	265	1025	9	3.9	66	51	404	15	28.8	14.4	47.57
5	New Orleans Saints	11	225	3134	651	4.8	18	5	202	226	363	2291	20	13	5.8	124	255	843	5	3.3	44	48	396	34	35.9	12.8	-20.88
6	Kansas City Chiefs	11	238	3968	701	5.7	17	5	240	242	380	2558	17	12	6.4	131	301	1410	10	4.7	84	69	533	25	36.8	14.9	-60.76
7	Washington Football Team	11	243	3405	676	5.0	13	3	200	222	343	2141	16	10	5.6	113	297	1264	9	4.3	69	61	507	18	37.1	10.5	1.32
8	Chicago Bears	11	250	3794	726	5.2	11	5	233	235	379	2461	16	6	6.2	127	326	1333	7	4.1	75	67	616	31	38.3	8.6	-39.59
9	Indianapolis Colts	11	253	3430	666	5.2	17	5	217	218	346	2309	15	12	6.3	121	297	1121	12	3.8	72	63	565	24	34.8	13.9	-34.88

Question of Interest:

- Who will win the Superbowl this year?
    - What statistical metrics are good predictors of the team that will win the Superbowl?
    - Are preseason rankings and future odds an indication of who usually wins in the postseason?

Betting Odds at the Beginning of the Season + Our Picks:

Super Bowl 55 odds for 2021, from best to worst

- Kansas City Chiefs +650
- Baltimore Ravens +700
- New Orleans Saints +1000
- San Francisco 49ers +1000
- Tampa Bay Buccaneers +1200
- Dallas Cowboys +1600
- Green Bay Packers +2000
- Indianapolis Colts +2000
- New England Patriots +2000
- Philadelphia Eagles +2000
- Pittsburgh Steelers +2200
- Seattle Seahawks +2200
- Buffalo Bills +2500
- Minnesota Vikings +2500
- Tennessee Titans +2800
- Cleveland Browns +3000
- Chicago Bears +4000
- Arizona Cardinals +5000
- Atlanta Falcons +5000

Before doing any analysis we both made some predictions on who we thought would win the Superbowl this year.

Ray: liked the Chiefs to repeat

Nebe: liked the Cowboys to repeat…

Regardless, of the outcome we thought it'd be fun to make predictions before truly starting the project to see if either of us or our ml algos could accurately predict the 2020 champion, based on historical data.

The snippet to the left shows the odds for all of the football teams before the season begun, from the Vegas bookmakers. Examining line movement was of natural interest to both of us, as we're avid sports fans, so we decided to include a comparison of the lines alongside our predictions..

Pre-Processing + Feature Selection:

- Since our data source has data relating to statistics, there were no missing values we had to worry about which was nice. In regards to the data cleaning there were relatively few problems with having to manipulate the dataset, as the file we used ultimately ended up being a tidy csv file.

- Although machine learning was not necessarily in the scope of the project, due to the predictive nature of our question of interest we decided to selectively select variables to use in our dataset, before doing any analysis. Thus, we found no need to have to normalize our data.

Output of the top 5 records → head(data):

	Rank	Team	Year	Playoff Wins	Points/G	Total Yards	Passing Yards	Rushing Yards	Yards per Play	Yards per Pass Attempt	Pass/Rec TD	Rush TD	Def. Yards per Play	Def. Passing Yards	Def. Rushing Yards	Def. Score %	Def. Turnover %
0	1	Kansas City Chiefs	2019	3	5	6	5	23	2	5	5	14	14	8	26	11	24
1	2	Tennessee Titans	2019	3	10	12	21	3	4	1	8	3	13	24	12	6	17
2	3	San Fransisco 49ers	2019	2	2	4	13	2	5	2	10	1	2	1	17	3	27
3	4	Green Bay Packers	2019	1	15	18	17	15	18	17	15	9	22	14	23	12	26
4	5	Seattle Seahawks	2019	1	9	8	14	4	12	8	4	15	29	27	22	19	29

As seen from the output above, we can better see the names of the fields in our dataset by calling data.columns as shown below

```
In [5]:  1 data.columns

Out[5]: Index(['Rank', 'Team', 'Year', 'Playoff Wins', 'Points/G', 'Total Yards',
               'Passing Yards', 'Rushing Yards', 'Yards per Play',
               'Yards per Pass Attempt', 'Pass/Rec TD', 'Rush TD',
               'Def. Yards per Play', 'Def. Passing Yards', 'Def. Rushing Yards',
               'Def. Score %', 'Def. Turnover %'],
              dtype='object')
```

EDA

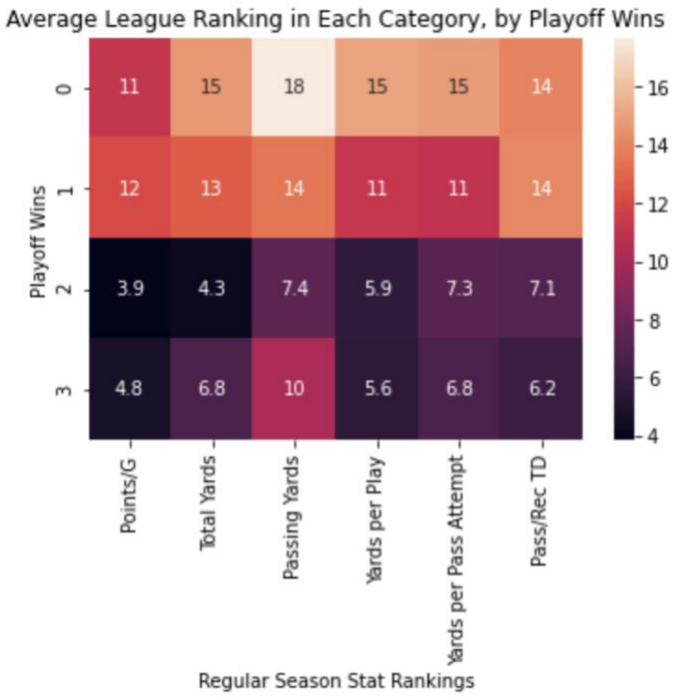
The first step of our Exploratory data analysis was determining what our dependent variable would be in our dataset. Since Champion wasn't explicitly labeled for any of the historical data used, we decided playoff wins would be our dependent variable, as it most resembled who won the Superbowl. From there we went on to run a bunch of basic statistical tests to get basic measures such as the correlation between variables in hopes of assessing if there is multicollinearity between any of the predictors.

The correlation matrix that was generated from our analysis is shown below as is a snippet of the code used to generate it:

```
In [24]:
#finding variables that have a Pearson's Correlation Coefficient of at least medium strength with number of playoff wins
corr = data.corr().abs()
corr = corr.loc[corr['Playoff Wins']>.25]
corr.index

Out[24]: Index(['Playoff Wins', 'Points/G', 'Total Yards', 'Passing Yards',
       'Yards per Play', 'Yards per Pass Attempt', 'Pass/Rec TD'],
      dtype='object')
```

```
In [25]:
#Relationship between the varibales with playoff wins
variables = list(corr.index)
corr_df = data[variables].groupby('Playoff Wins').mean()
corr_df
plt.title("Average League Ranking in Each Category, by Playoff Wins")
sns.heatmap(data=corr_df,annot=True)
plt.xlabel("Regular Season Stat Rankings")
```



The code above shows what was used to output the heatmap shown to the left. From the first blob of code, you can see that only the variables with a correlation to playoff wins above .25 were included in the visual. The output of the corr.index functions returns all variables who met the threshold set above.

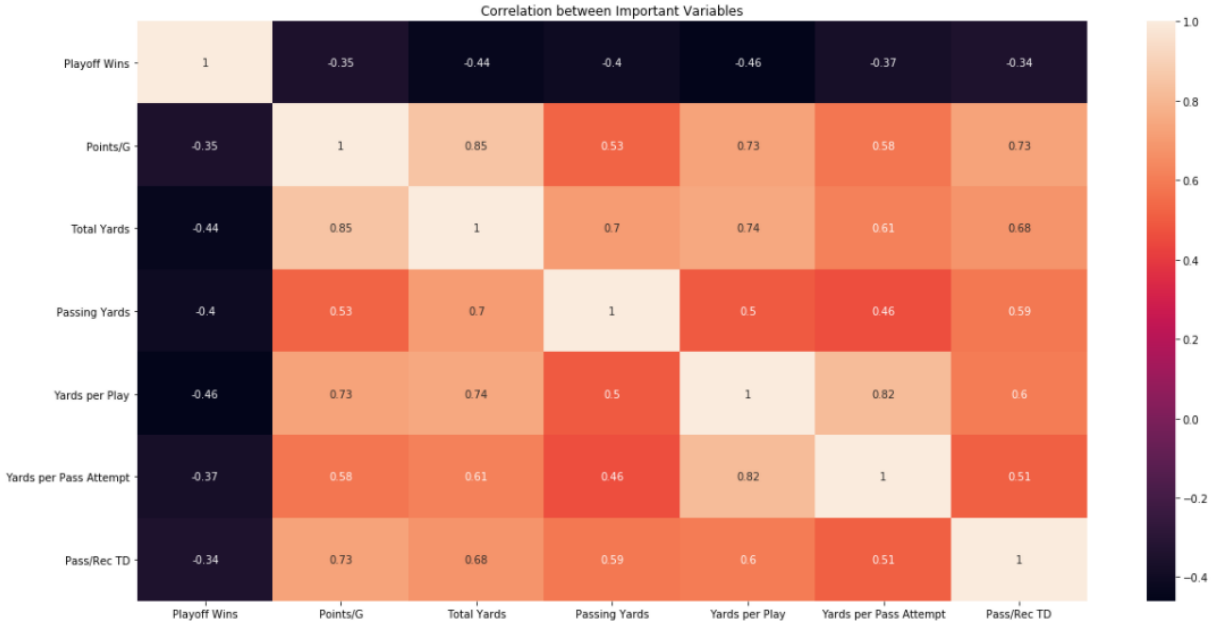
From there the next blob of code groups our data by playoff wins and computes the average ranking across all variables previously selected and outputs a heatmap.

The results of the heatmap show that teams that rank highest in points per game, or score the most points in the league are more likely to win 3 playoff games.

\

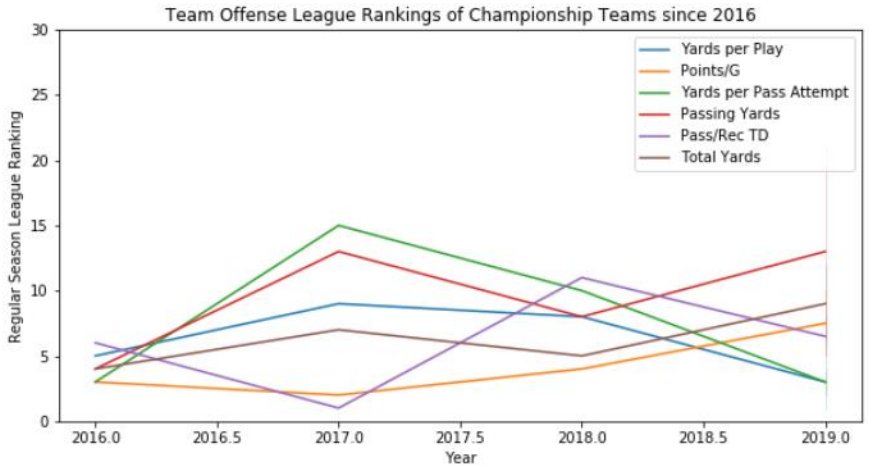
Next we created a correlation matrix to check for multicollinearity between our predictor variables. The output and code of the correlation matrix are shown on the next page.

```python
#Check for Multicollinearity between the important variables
plt.figure(figsize=(20,10))
plt.title("Correlation between Important Variables")
sns.heatmap(data=data[variables].corr(), annot=True)
```



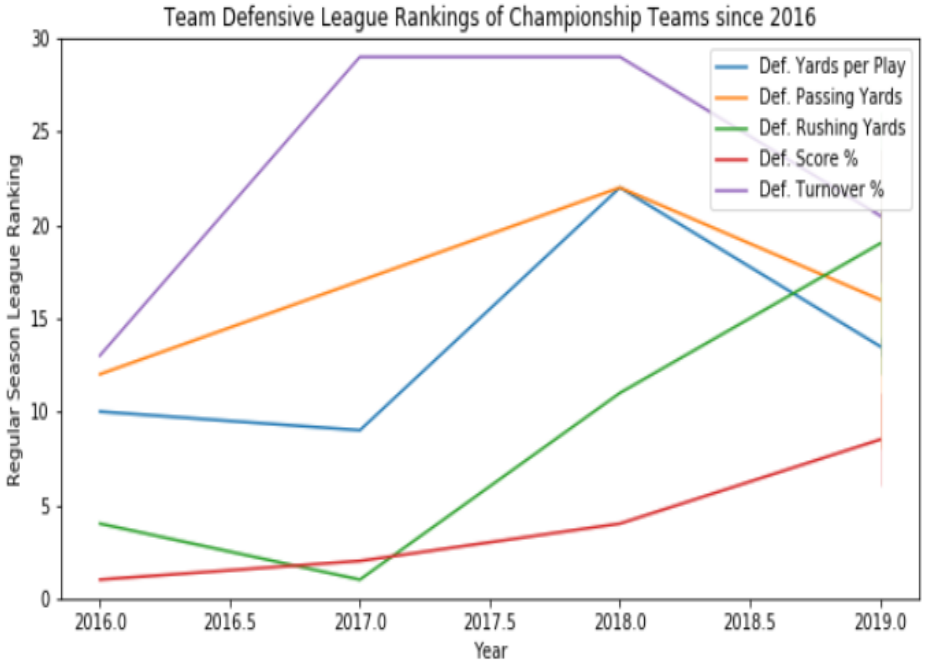
The output of our correlation matrix shows all negative values for the predictors, due to the ordinal nature of our dependent variable. After fully understanding this we went on to interpret the results as Yards per Play being the strongest predictor of playoff wins due to it's correlation coefficient being the strongest.

Next, we did a few visualizations in hopes of seeing if we could see any patterns or irregular trends in our data visually. The first visualization that was created was done so in hopes of gaging whether there has been a shift in how to be successful in the NFL over the years. This graph is extremely interesting because, like how the NBA has transformed due to the immergence of the three ball, it can be used to gauge if different requirements are now expected of Superbowl champions..



The output to the left confirms that there hasn't been any drastic measures that influenced Superbowl champions over the years.

Here is another visualization that like the previous visual, analyzes defensive metrics for Superbowl champions from 2016 to 2019.



Seeing as the passing yards increased in 2018, it's rather intuitive that Defensive Score % increases (as there are more interceptions and turnovers to be had in the air rather than on the ground).

Models

After getting rid of the rank variable from our dataset, we split up the data in testing and training groups. We assigned the 2016-2019 historical data to our training group, in hopes of using that to predict our 2020 champion.

Our linear regression model predicted this for 2020:

	Team	Playoff Wins
1	Kansas City Chiefs	1.645193
17	Minnesota Vikings	1.485775
3	Green Bay Packers	1.388634
5	Seattle Seahawks	1.360795
12	Arizona Cardinals	1.355259
8	Los Angeles Rams	1.181668
11	Tampa Bay Buccaneers	1.167420
4	Tennessee Titans	1.141462
6	Buffalo Bills	1.119272
7	Cleveland Browns	0.865149
15	Las Vegas Raiders	0.856895
10	Indianapolis Colts	0.696026
2	New Orleans Saints	0.668625
20	Atlanta Falcons	0.650341
16	New England Patriots	0.582319
14	Baltimore Ravens	0.089705
18	Washington Football Team	0.087780
13	New York Giants	0.064951
0	Pittsburgh Steelers	-0.040491
9	Miami Dolphins	-0.049553
19	Chicago Bears	-0.067424

Since it wouldn't make sense to use Mean Square Error as a performance metric for our data set, we used current NFL standings to gauge if our predictions were correct.

The current NFL standings in both divisions are:

NFC Playoff Standings Share & more ▾ Glossary

Tm	W	L	T	Position	Reason
New Orleans Saints (1)	9	2	0	South Champion	
Seattle Seahawks (2)	8	3	0	West Champion	strength of victory
Green Bay Packers (3)	8	3	0	North Champion	
New York Giants (4)	4	7	0	East Champion	head-to-head record
Los Angeles Rams (5)	7	4	0	Wild Card #1	
Tampa Bay Buccaneers (6)	7	5	0	Wild Card #2	
Arizona Cardinals (7)	6	5	0	Wild Card #3	
Minnesota Vikings	5	6	0		conference win percentage
Chicago Bears	5	6	0		conference win percentage
San Francisco 49ers	5	6	0		
Detroit Lions	4	7	0		head-to-head record
Washington Football Team	4	7	0		conference win percentage

AFC Playoff Standings Share & more ▾ Glossary

Tm	W	L	T	Position	Reason
Pittsburgh Steelers (1)	11	0	0	North Champion	
Kansas City Chiefs (2)	10	1	0	West Champion	
Tennessee Titans (3)	8	3	0	South Champion	head-to-head record
Buffalo Bills (4)	8	3	0	East Champion	
Cleveland Browns (5)	8	3	0	Wild Card #1	
Miami Dolphins (6)	7	4	0	Wild Card #2	conference win percentage
Indianapolis Colts (7)	7	4	0	Wild Card #3	
Las Vegas Raiders	6	5	0		conference win percentage
Baltimore Ravens	6	5	0		
New England Patriots	5	6	0		

As noted during our presentations, our regression model wasn't to shabby in comparison to the current live rankings. But it was noted that the Steelers were noticeably outperforming there prediction that came from our model. After investigating further we believe the Steelers could be outliers in our model due to different factors and variables such as injuries and different stats not being included in the model.

We also ran a random forest prediction to predict the champion and got the output shown below:

```
rfr_predicted_wins_df = pred_data[['Team','Playoff Wins']]
i=0
while i<21:
    rfr_predicted_wins_df.at[i, 'Playoff Wins'] = rfr_predicted_wins[i]
    i+=1
rfr_predicted_wins_df.sort_values(by='Playoff Wins',ascending=False)
```

	Team	Playoff Wins
5	Seattle Seahawks	2.402857
1	Kansas City Chiefs	1.951429
3	Green Bay Packers	1.900000
17	Minnesota Vikings	1.671429
8	Los Angeles Rams	1.214286
6	Buffalo Bills	1.060000
11	Tampa Bay Buccaneers	1.045714
12	Arizona Cardinals	0.774286
2	New Orleans Saints	0.751429
15	Las Vegas Raiders	0.682857
16	New England Patriots	0.671429
7	Cleveland Browns	0.660000
13	New York Giants	0.634286
4	Tennessee Titans	0.582857
18	Washington Football Team	0.511429
19	Chicago Bears	0.468571
20	Atlanta Falcons	0.462857
0	Pittsburgh Steelers	0.360000
10	Indianapolis Colts	0.354286
9	Miami Dolphins	0.291429
14	Baltimore Ravens	0.120000

While this prediction is somewhat more accurate than the linear regression, it still was unable to recognize the Steelers impressive form, and furthermore odds to win the 2020 Superbowl.

Conclusion:

Both the regression and random forest predictions were relatively accurate in comparison to the current NFL playoff standings. This assignment taught us both a ton about working in python to answer research questions. In terms of the divvy up of work, we both came up with our research questions together, and worked collaboratively on: the code for this assignment, the project proposal, the presentation, and this project report.

What we learned?

This project allowed us to get our feet wet when it comes to doing data analysis and performing different tasks that we'd previously not learned in Python. In terms of the actual analysis we did, this project allowed us to see how many things actually influence predicting the Superbowl champion. The results of our models showed us that we desperately need some new variables to strengthen our model, and that the preseason standings/odds are a relatively good prediction of who will win the superbowl.

Sidenote: Seeing as gambling is an emerging industry across the United States, it makes sense that that the math wizards who set the gambling lines for futures such as who will win the superbowl, make odds that are so good at predicting what's really happening during the season.

Below is a snippet of the odds before the season vs the odds right now.

Super Bowl 55 odds for 2021, from best to worst

- Kansas City Chiefs +650
- Baltimore Ravens +700
- New Orleans Saints +1000
- San Francisco 49ers +1000
- Tampa Bay Buccaneers +1200
- Dallas Cowboys +1600
- Green Bay Packers +2000
- Indianapolis Colts +2000
- New England Patriots +2000
- Philadelphia Eagles +2000
- Pittsburgh Steelers +2200
- Seattle Seahawks +2200
- Buffalo Bills +2500
- Minnesota Vikings +2500
- Tennessee Titans +2800
- Cleveland Browns +3000
- Chicago Bears +4000
- Arizona Cardinals +5000
- Atlanta Falcons +5000

Pro Football Championship 55

Kansas City Chiefs	+280	New Orleans Saints	+550	Pittsburgh Steelers	+550
Seattle Seahawks	+900	Green Bay Packers	+1000	Los Angeles Rams	+1600
Tampa Bay Buccaneers	+1600	Tennessee Titans	+2000	Buffalo Bills	+2200
Baltimore Ravens	+2700	Indianapolis Colts	+3100	Cleveland Browns	+4600
Arizona Cardinals	+4800	Miami Dolphins	+5000	Las Vegas Raiders	+5500
Minnesota Vikings	+8000	San Francisco 49ers	+10000	Philadelphia Eagles	+12000
New England Patriots	+13000	New York Giants	+13000	Washington	+13000
Chicago Bears	+17000	Dallas Cowboys	+20000	Detroit Lions	+40000
Houston Texans	+40000	Atlanta Falcons	+50000	Carolina Panthers	+50000
Denver Broncos	+75000	Los Angeles Chargers	+100000	Cincinnati Bengals	+150000

The snippets above show that my Cowboys pick is very unlikely to happen and that Ray's still got a shot according to Vegas if he bet on his preseason pick the Chiefs.