Analysis of Winning in the 2019-2020 NFL Regular Season

Quang Nguyen, Frank D'Ambrosio, and Charles Hwang

1) Introduction

1.1) Topic

Football is certainly one of, if not the most popular sport in the United States; and it also plays an important role in American culture. The National Football League (NFL) is the top-tier football league in the country, and it is home to some of the most financially valuable and famous sports franchises and athletes in the world. In addition to being the most-watched sports league in the U.S., the NFL is also well-known for its level of intensity, competition quality, and overall balance. Just like every other sport, the main goal for every team, coach, and player in the NFL is to win the game. For this reason, we decided to choose winning in the NFL as the main topic for our study, with specific focus on the following variables: win percentage, point differential, playoff status, and coaching change.

1.2) Design

For this project, we used a dataset on the 2019-2020 NFL regular season, obtained from pro-football-reference.com. Our observational units are the 32 NFL franchises, and our datafile consists of the following attributes for each observation: team name; conference; division; number of wins, losses, and ties; whether a team made the playoffs; total points scored; total yards gained; total touchdown scored; total points, yards, and TD allowed; and whether a new coach was hired before the season. From these initial variables, we calculated each team's win percentage from how many wins, losses, and ties they had in 2019; as well as the differential statistics for point, yardage, and touchdown. Based on our data, the questions/relationships that we would like to investigate are:

(1) Is point differential a helpful predictor for win percentage?
(2) Is there a difference in net point for playoff and non-playoff teams?
(3) Is having a new coach a good predictor for whether a team made the playoffs?

2) Analysis

2.1) Win Percentage and Point Differential

First, we would like to look at the relationship between win percentage and point differential. We suspected a correlation between these two variables and decided to use the method of linear regression. We fitted a single linear regression model with win percentage as the response variable and point differential as the explanatory variable. Before the modeling stage, we looked at some visual and numerical summaries of the relationship between win percentage and point differential. We observed a strong, positive, and linear relationship between win percentage and point differential from the scatterplot of those two variables with a correlation coefficient of 0.877 (Figures 1 and 2). In context, this means that a greater point differential is strongly associated with a higher win percentage for NFL teams.

Fig. 1: Correlation Between Win Percentage and Point Differential

The CORR Procedure

2 Variables:	winpct pd

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
winpct	32	50.00000	19.81199	1600	12.50000	87.50000
pd	32	0	107.69761	0	-188.0000	249.0000

Pearson Correlation Coefficients, N = 32
Prob > |r| under H0: Rho=0

	winpct	pd
winpct	1.00000	0.87710 <.0001
pd	0.87710 <.0001	1.00000

Figure 1: Correlation between Win Percentage and Point Differential

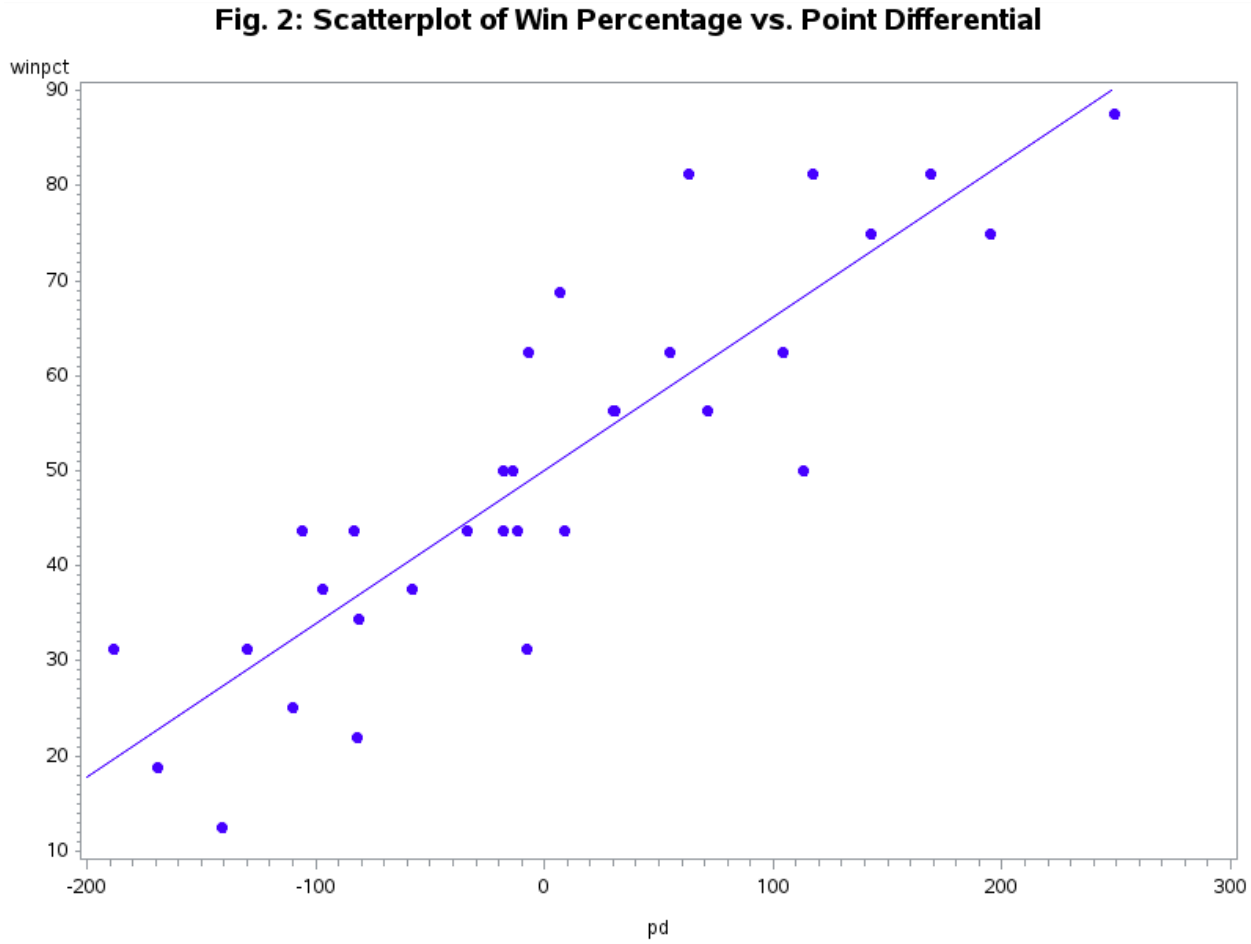

Figure 2: Scatterplot of Win Percentage vs Point Differential

We then fitted a regression model for win percentage and point differential. Before evaluating the model's coefficients, we assessed the assumptions for this single linear regression model (Figure 3). In terms of linearity, the residual vs. predicted plot shows no mathematical function pattern, so the errors are independent. We also confirmed the constant variance condition here, since from the residual vs. predicted plot, there does not seem to be a fanning pattern or much of a difference in variances from left to right. Lastly, for normality, the percent vs. residual histogram and quantile-quantile plots show that the errors are normally distributed.

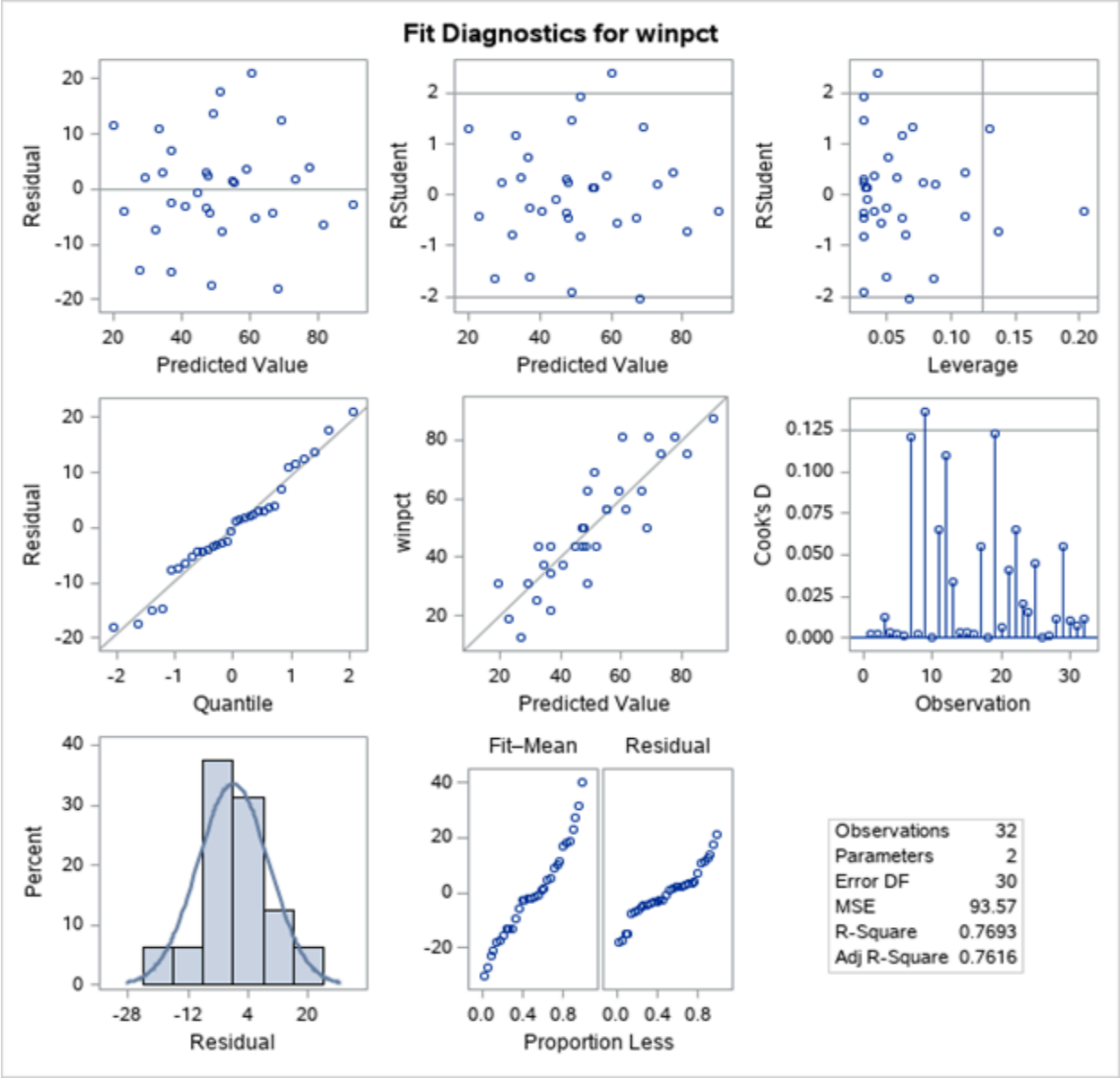

Figure 3: Diagnostic Plots for Linear Model of Win Percentage vs. Point Differential

Fig. 4: Linear Model of Win Percentage and Point Differential

The REG Procedure

Model: MODEL1

Dependent Variable: winpct

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9360.86890	9360.86890	100.04	<.0001
Error	30	2807.09985	93.56999		
Corrected Total	31	12168			

Root MSE	9.67316	R-Square	0.7693
Dependent Mean	50.00000	Adj R-Sq	0.7616
Coeff Var	19.34632		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	50.00000	1.70999	29.24	<.0001	46.50774	53.49226
pd	1	0.16135	0.01613	10.00	<.0001	0.12841	0.19430

Figure 4: Summary for Linear Model of Win Percentage vs. Point Differential

After assessing the conditions, we went on to analyze the summary table for our linear regression model (Figure 4). Our fitted equation for predicting win percentage from point differential is:

$$\widehat{WinPct} = 50 + 0.161 \cdot PD.$$

With t = 10.00 and p-value <.0001, we have sufficient evidence to conclude that our linear model is appropriate at significance level $\alpha = 0.05$, and thus, win percentage and point differential are linearly related. SAS output also gives us a correlation of determination (R^2) of 0.7693; hence, 76.93% of the variability in win percentage can be explained by the linear relationship between win percentage and point differential. The slope of our equation is 0.161, which indicates that every extra point in point differential is associated with 0.161% of higher win percentage. We also obtained a 95% confidence interval for the slope, which is (0.128, 0.194). In context, we are 95% confident that as point differential increases by 1, win percent increases by 0.128% to 0.194%. In addition, we evaluated the intercept for our model and a 95% confidence interval for it. Teams with a point differential of 0 would have a 50% win percentage, as illustrated by an intercept term of 50. In other words, those teams would win as many games

as they would lose, with an even number of ties. The 95% confidence interval for the intercept is (46.51, 53.49); thus, we are 95% confident that the winning percentage for teams with a point differential of 0 is between 46.51% and 53.49%.

After analyzing the model's coefficients, we looked at what this model would predict for NFL teams. We used the Chicago Bears as our case of investigation (Figure 5, row 6). The Bears had a point differential of -18 in 2018-19, and this model predicts a 47.1% win percentage for them. Chicago actually went 8-8 last year (50% win percentage), so our model's prediction for the Bears' winning percentage is 2.9% lower than their actual value. For all teams with a point differential of -18, we are 95% confident that their mean win percentage is between 43.55% and 50.64%; and for any given NFL team with a point differential of -18, we are 95% sure that their winning percentage is between 27.03% and 67.17%.

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	34.4	36.9306	2.1521	32.5354	41.3257	16.6923	57.1688	-2.5556
2	43.8	47.0957	1.7345	43.5534	50.6379	27.0254	67.1660	-3.3457
3	87.5	90.1764	4.3656	81.2605	99.0922	68.5024	111.8503	-2.6764
4	62.5	58.8743	1.9265	54.9399	62.8087	38.7311	79.0175	3.6257
5	31.3	29.0244	2.7059	23.4982	34.5506	8.5108	49.5380	2.2256
6	50.0	47.0957	1.7345	43.5534	50.6379	27.0254	67.1660	2.9043
7	12.5	27.2495	2.8457	21.4379	33.0611	6.6572	47.8419	-14.7495
8	37.5	40.6417	1.9492	36.6608	44.6225	20.4893	60.7940	-3.1417
9	50.0	68.2326	2.4994	63.1282	73.3371	47.8286	88.6367	-18.2326
10	43.8	44.5141	1.7958	40.8466	48.1816	24.4213	64.6068	-0.7641
11	21.9	36.7692	2.1619	32.3540	41.1844	16.5266	57.0118	-14.8942
12	81.3	60.1651	1.9892	56.1026	64.2276	39.9965	80.3337	21.0849
13	62.5	48.8705	1.7137	45.3707	52.3704	28.8077	68.9334	13.6295
14	43.8	48.0638	1.7209	44.5492	51.5784	27.9984	68.1292	-4.3138
15	37.5	34.3490	2.3179	29.6152	39.0827	14.0345	54.6634	3.1510
16	75.0	73.0732	2.8715	67.2088	78.9376	52.4659	93.6805	1.9268
17	31.3	48.7092	1.7149	45.2070	52.2114	28.6459	68.7725	-17.4592
18	56.3	54.8405	1.7772	51.2111	58.4700	34.7547	74.9264	1.4095
19	31.3	19.6660	3.4816	12.5556	26.7765	-1.3298	40.6619	11.5840
20	62.5	66.7805	2.3956	61.8881	71.6729	46.4285	87.1325	-4.2805

21	75.0	81.4634	3.5804	74.1512	88.7756	60.3983	102.5285	-6.4634
22	81.3	68.8780	2.5468	63.6767	74.0794	48.4496	89.3065	12.3720
23	25.0	32.2514	2.4643	27.2186	37.2842	11.8652	52.6376	-7.2514
24	43.8	36.6079	2.1718	32.1724	41.0433	16.3609	56.8549	7.1421
25	43.8	32.8968	2.4183	27.9580	37.8356	12.5336	53.2600	10.8532
26	56.3	55.0019	1.7816	51.3633	58.6404	34.9144	75.0894	1.2481
27	50.0	47.7411	1.7248	44.2185	51.2637	27.6743	67.8079	2.2589
28	81.3	77.2683	3.2182	70.6959	83.8407	56.4485	98.0881	3.9817
29	68.8	51.1295	1.7137	47.6296	54.6293	31.0666	71.1923	17.6205
30	43.8	51.4522	1.7161	47.9473	54.9570	31.3884	71.5159	-7.7022
31	56.3	61.4559	2.0581	57.2526	65.6592	41.2585	81.6533	-5.2059
32	18.8	22.7317	3.2182	16.1593	29.3041	1.9119	43.5515	-3.9817

Sum of Residuals	0
Sum of Squared Residuals	2807.09985
Predicted Residual SS (PRESS)	3156.88467

Figure 5: Win Percentage Estimations for NFL Teams

2.2) Point Differential and Playoff Appearance

The second relationship that we wanted to look at was whether point differential differ for teams that made the 2019-2020 NFL postseason and those that did not. We started off with a boxplot (Figure 6) as well as some basic descriptive statistics for point differential (PD), broken down by whether a team made the 2019-2020 postseason (Figure 7). Overall, it is obvious that playoff teams averaged a higher point differential (+99.75) than non-playoff teams (-59.85). We used a directional t-test to verify this difference.

Fig. 6: Boxplot for Point Differential by Playoffs Appearance

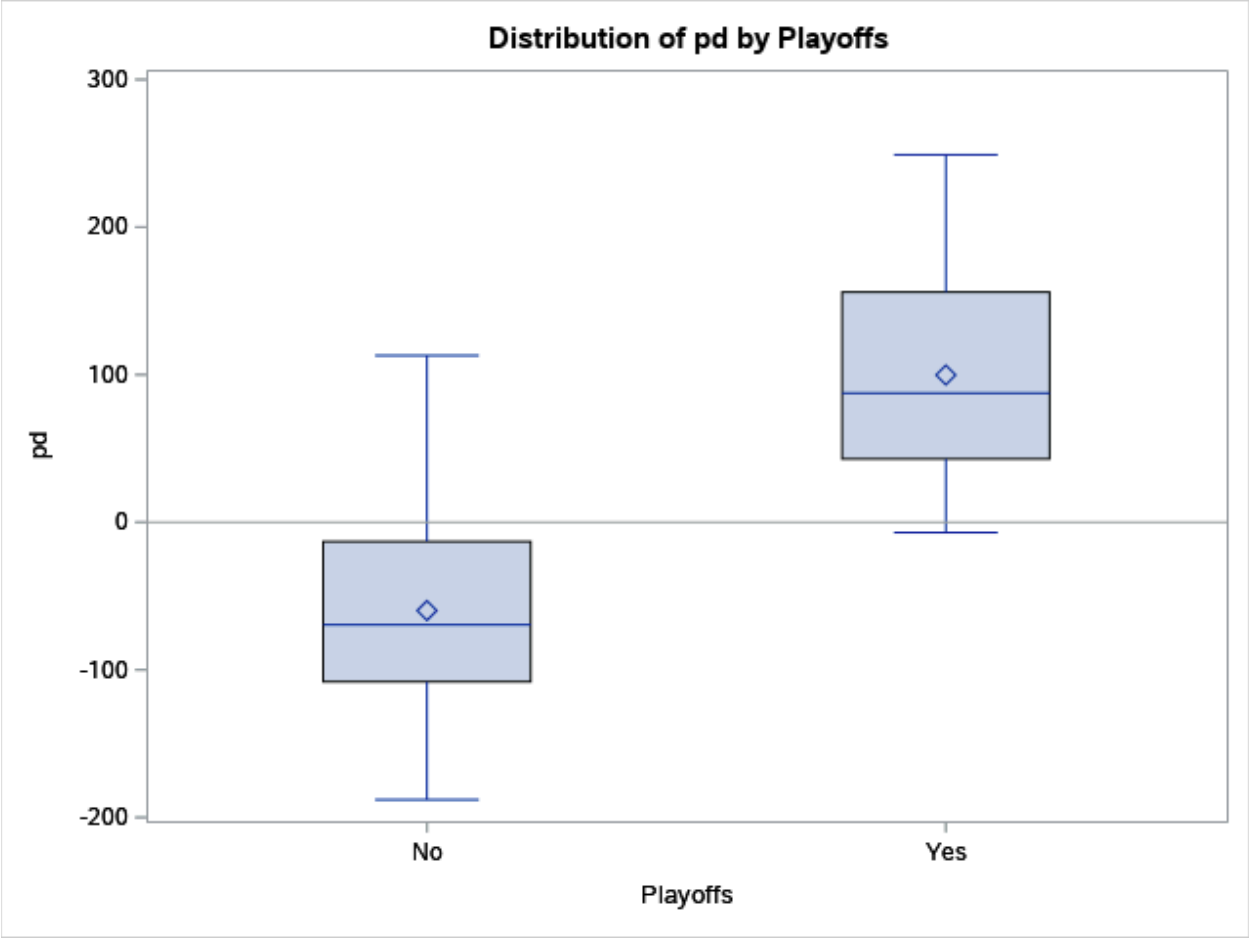

Figure 6: Boxplot of Point Differential by Playoff Appearance

Fig. 7: Descriptive Statistics for Points Differential by Playoffs Appearance

The MEANS Procedure

Analysis Variable : pd						
Playoffs	**N Obs**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
No	20	-59.850	73.0128	-69.500	-188.000	113.000
Yes	12	99.750	78.1806	87.500	-7.000	249.000

Figure 7: Descriptive Statistics for Point Differential by Playoff Appearance

The null and alternative hypotheses for our t-test are: H_0: mean point differential is the same for non-playoff and playoff teams; and H_A: mean point differential for non-playoff teams is less than mean point differential for playoff teams. Regarding our model assumptions, the

distribution of point differential looks right-skewed (Figure 8), so our sample size of 32 is large enough for us to appropriately carry out the t-test. For this test, we used the pooled variance method, due to the difference in variances not being significant (F = 1.15, p-value = 0.7639). SAS output (Figure 9) gave us a test statistic of t = -5.83 (on 30 df) and p-value <.0001. Thus at 5% significance level, we have evidence that the mean point differential for non-playoff teams is less than the mean point differential for playoff teams. We also have a 95% confidence interval for the difference, (-∞, -113.2), meaning we are 95% confident that the difference between the average point differential for non-playoff teams and playoff teams in the 2019-2020 NFL season is at most -113.2 points.

Fig. 8: Histogram of Point Differential

The UNIVARIATE Procedure

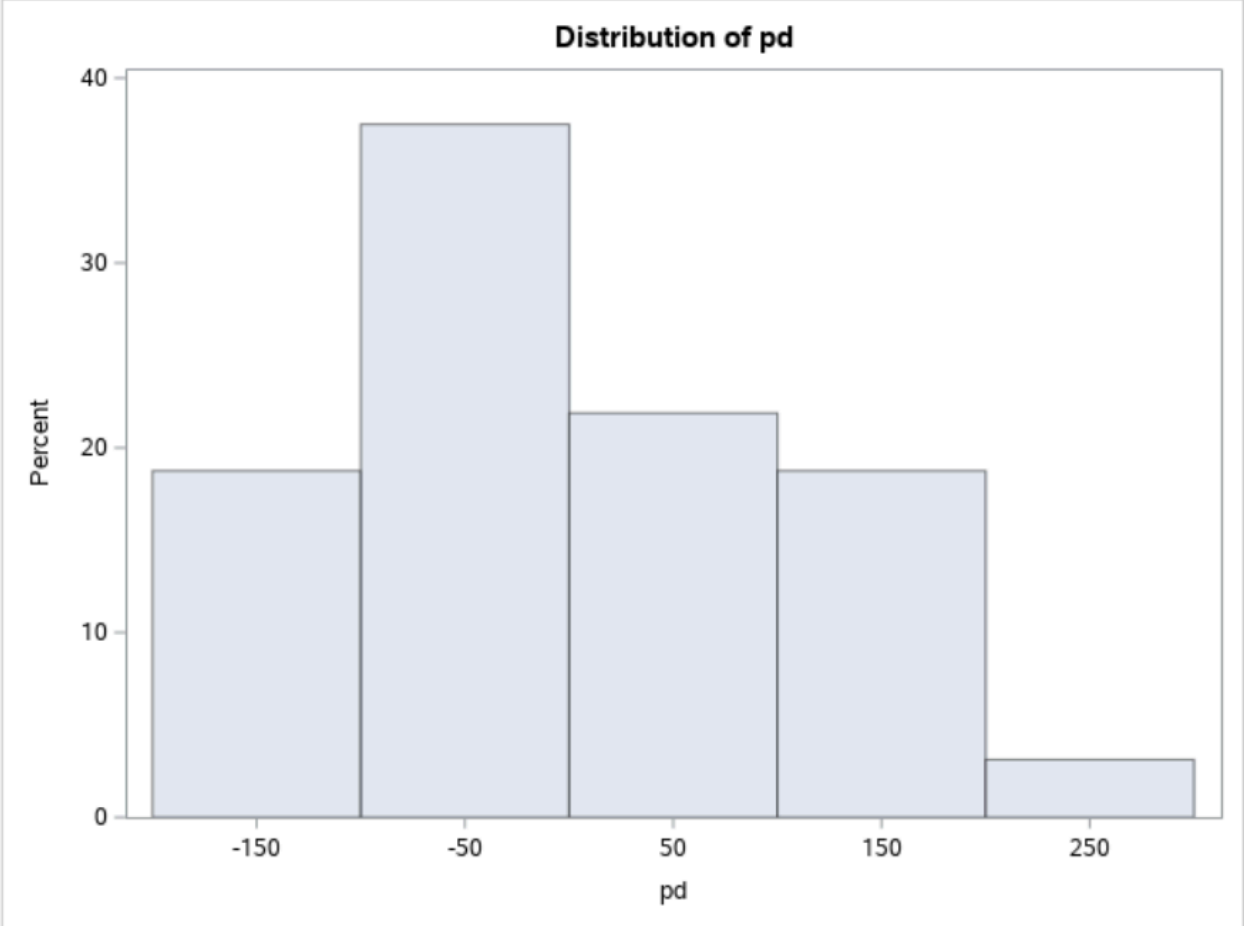

Figure 8: Histogram of Point Differential

Fig. 9: Two-Sample T-Test of Point Differential for Playoff and Non-Playoff Teams

The TTEST Procedure

Variable: pd

Playoffs	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
No		20	-59.850	73.0128	16.3262	-188.0	113.0
Yes		12	99.750	78.1806	22.5688	-7.0000	249.0
Diff (1-2)	Pooled		-159.6	74.9490	27.3675		
Diff (1-2)	Satterthwaite		-159.6		27.8549		

Playoffs	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
No		-59.850	-94.0210	-25.6790	73.0128	55.5255	106.6
Yes		99.750	50.0764	149.4	78.1806	55.3828	132.7
Diff (1-2)	Pooled	-159.6	-Infty	-113.2	74.9490	59.8927	100.2
Diff (1-2)	Satterthwaite	-159.6	-Infty	-111.8			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	30	-5.83	<.0001
Satterthwaite	Unequal	22.032	-5.73	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	11	19	1.15	0.7639

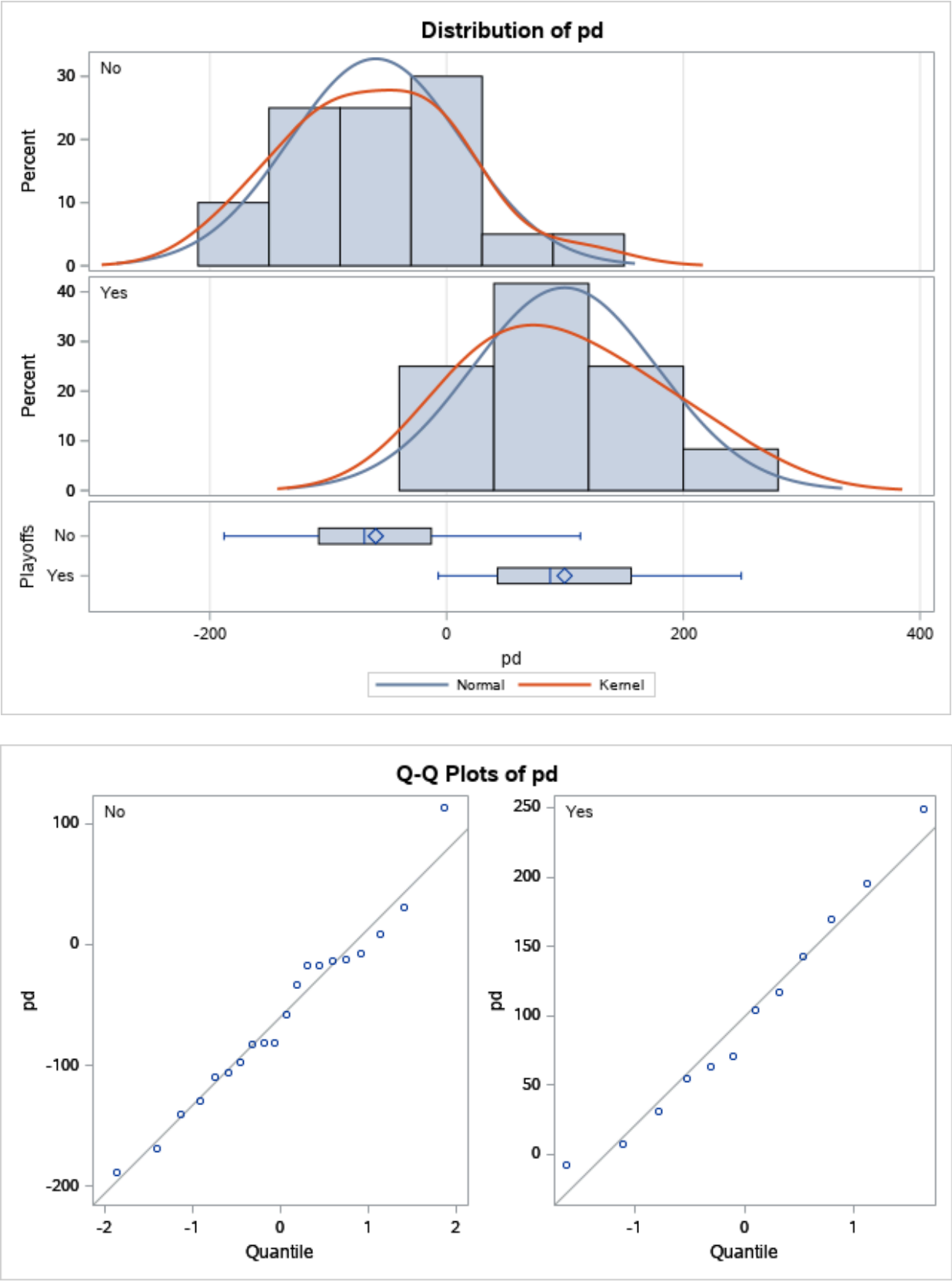

Figure 9: T-test of Point Differential for Playoff and Non-Playoff Teams

2.3) Playoff Appearance and Coaching Change

The last question that we were interested in examining was on whether or not there is an association between teams' playoff status and whether they had a coaching change before the start of last season. As usual, we began with visual and numerical summaries of our variables of interest. Looking at the 2x2 contingency table and mosaic plot (Figure 10), 8 NFL franchises out of 32 total (25%) started the season with a new head coach, and out of those 8 teams, 7 ended up missing the playoffs (with the Green Bay Packers being the only exception). Hence the probability of a team making the playoffs, given that they had a new coach, is 1/8.

Fig. 10: Contingency Table and Mosaic Plot for Coaching Change and Playoffs Appearance

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Playoffs by NewCoach		
	NewCoach(NewCoach)		
Playoffs(Playoffs)	Yes	No	Total
No	7 21.88 35.00 87.50	13 40.63 65.00 54.17	20 62.50
Yes	1 3.13 8.33 12.50	11 34.38 91.67 45.83	12 37.50
Total	8 25.00	24 75.00	32 100.00

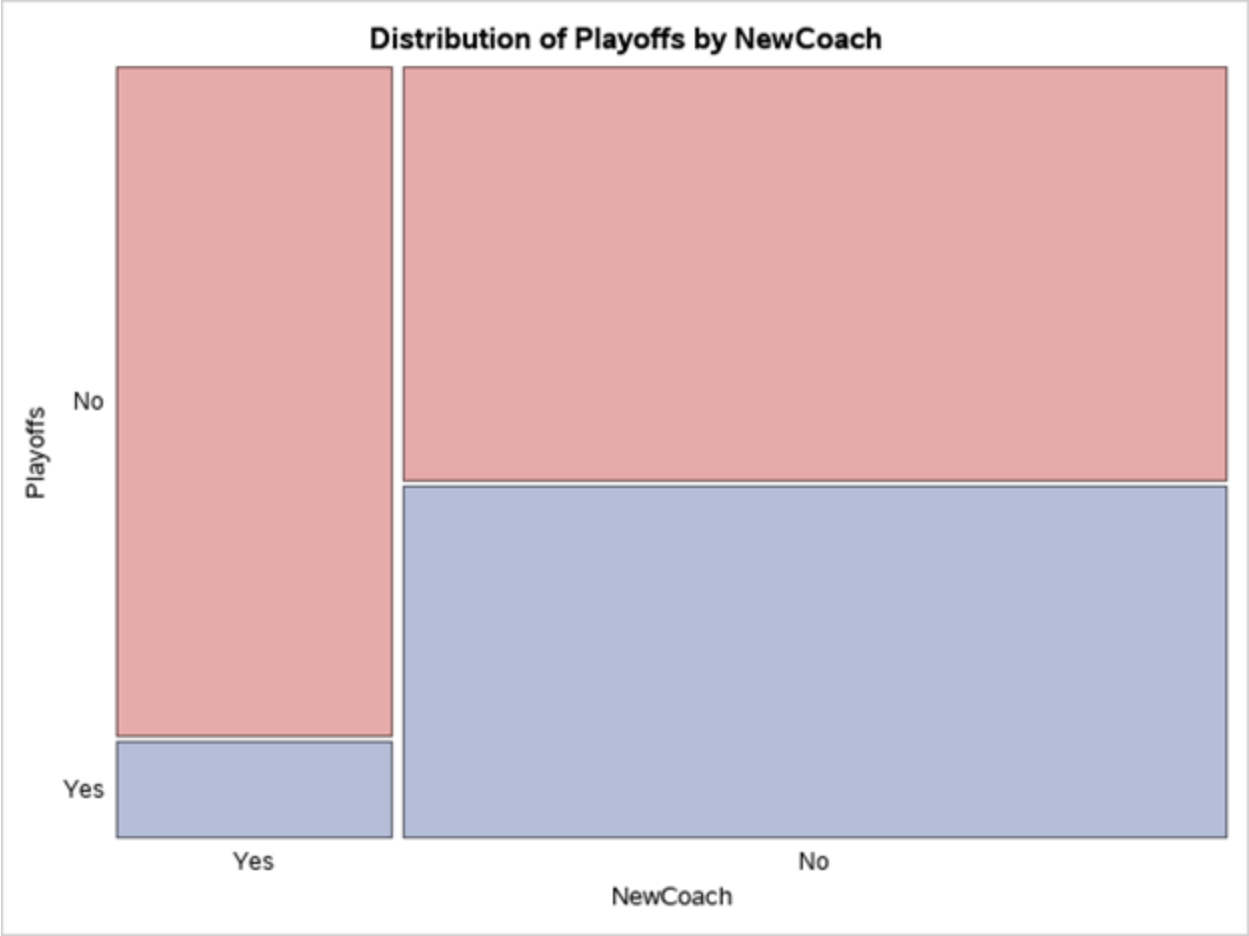

Figure 10: Contingency Table and Mosaic Plot for Playoff Appearance and New Coach

To answer our last question, we conducted a Chi-square test for our two variables - coaching change and playoff appearance. In terms of our Chi-square test conditions, we actually had a concern here, since Green Bay was the one team that made the playoffs and also had a new coach, but we need at least 5 observations in each cell so that the expected count condition for the Chi-square test is not violated. Therefore, instead of using Karl Pearson's test for categorical data, we decided to use Fisher's Exact Test, which is more friendly toward small expected numbers. With a two-sided p-value of 0.2045 (Figure 11), we do not have evidence to conclude that an association between playoff status and coaching change exists.

Fig. 11: Chi-Square and Fisher's Exact Tests for Coaching Change and Playoffs Appearance

The FREQ Procedure

Frequency
Expected
Deviation
Percent

Table of Playoffs by NewCoach			
	NewCoach(NewCoach)		
Playoffs(Playoffs)	Yes	No	Total
No	7 5 2 21.88	13 15 -2 40.63	20 62.50
Yes	1 3 -2 3.13	11 9 2 34.38	12 37.50
Total	8 25.00	24 75.00	32 100.00

Statistics for Table of Playoffs by NewCoach

Statistic	DF	Value	Prob
Chi-Square	1	2.8444	0.0917
Likelihood Ratio Chi-Square	1	3.2075	0.0733
Continuity Adj. Chi-Square	1	1.6000	0.2059
Mantel-Haenszel Chi-Square	1	2.7556	0.0969
Phi Coefficient		0.2981	
Contingency Coefficient		0.2857	
Cramer's V		0.2981	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	7
Left-sided Pr <= F	0.9880
Right-sided Pr >= F	0.1004
Table Probability (P)	0.0884
Two-sided Pr <= P	0.2045

Sample Size = 32

Figure 11: Chi-square and Fisher's Exact Tests for Playoff Appearance and Coaching Change

3) Conclusion and Discussion

In summary, we examined the topic of winning in the NFL and looked at multiple relationships between different variables in the game of football. We were able to verify the relationship between point differential and win percentage in the 2019 NFL season, as well as the fact that point differential for playoff teams is higher than non-playoff teams. However, we did not have sufficient evidence for an association between coaching change and whether a team made the postseason last year.

There are some improvements we could make and some other topics we could explore in the future. In terms of model selection, we could try to determine the best subset of predictors for win percentage, using methods like forward selection, backward elimination, or stepwise regression. Additionally, we could use a larger dataset containing data from multiple NFL seasons rather than a single season. This would allow us to have more observations and more importantly to compare how these team statistics vary across different seasons.

Moreover, we did not use the variables for yards gained or yards allowed, as they only recorded offensive yards and do not account for defensive or special teams yardage (interception returns, fumble returns, kick returns, etc.). Similarly, we did not use the variables for touchdowns scored and touchdowns allowed, as we felt these were simply proxy variables for points scored and points allowed. However, future analyses could create variables for yardage differential and touchdown differential and test whether these variables are significant in a model, perhaps with touchdown differential being separate from point differential. The correlation (or lack thereof) between yardage differential and record would be one possible analysis. As several studies have shown that yardage differential can be misleading in determining the outcome of a game (teams trailing by more than one possession late in a game tend to gain a disproportionate amount of yardage as such the defense allows, and thus the losing team sometimes has more yards than the winning team), we anticipate the correlation between the two variables is weak and very well may be negative. Analyzing the individual variables for yards gained/touchdowns scored and yards/touchdowns allowed can also help in

finding out a team's general play style or strategy (offensive vs. defensive or aggressive vs. conservative).

```
proc import out=nfl2019 datafile="/home/u49451619/nfl2019.xlsx" dbms=xlsx;
        getnames=yes;
run;

* Calculate point differential, td differential, yard differential, win pct;
data nfl;
set nfl2019;
winpct = 100*(wins + 0.5*ties)/(wins + losses + ties);
pd = ps - pa;
yd = yg - ya;
tdd = tds - tda;
run;

* Linear Regression: Win Percentage and Point Differential;
proc corr data = nfl;
title "Fig. 1: Correlation between Win Percentage and Point Differential";
var winpct pd;
run;

symbol1 value=dot color=blue I=R;
proc gplot data = nfl;
title "Fig. 2: Scatterplot of Win Percentage vs. Point Differential";
plot winpct*pd;
run;

proc reg data=nfl;
title "Fig. 4: Linear Model of Win Percentage vs. Point Differential";
model winpct = pd / clb cli clm;
run;

* T-test: PD and Playoffs Appearance;
proc sort data=nfl;
by playoffs;
proc boxplot data=nfl;
title "Fig. 6: Boxplot for Point Differential by Playoffs Appearance";
plot pd*playoffs /vref=0;
run;

proc means data=nfl mean std median min max;
title "Fig. 7: Descriptive Statistics for Point Differential by Playoffs Appearance";
class playoffs;
var pd;
run;
```

```
proc univariate data=nfl noprint;
title "Fig. 8: Histogram of Point Differential";
histogram pd;
run;

proc ttest data=nfl alpha=0.05 sides=l;
title "Fig. 9: Two-Sample T-Test of Point Differential for Playoff and Non-Playoff Teams";
class playoffs;
var pd;
run;

* Chi-square test: Coaching Change and Playoffs Appearance;
proc freq data=nfl order=data;
title "Fig. 10: Contingency Table and Mosaic Plot for Coaching Change and Playoffs Appearance";
table playoffs*newcoach / plots= mosaicplot;
run;

proc freq data=nfl order=data;
title "Fig. 11: Chi-Square and Fisher's Exact Tests for Coaching Change and Playoffs Appearance";
table playoffs*newcoach / chisq expected deviation norow nocol nocum;
run;
```

Presentation Link:
https://docs.google.com/presentation/d/1Pxhb8C6ltQYU2ty2nQrot7vzotuqqf-kv1OOPkq4t5U/edit?usp=sharing