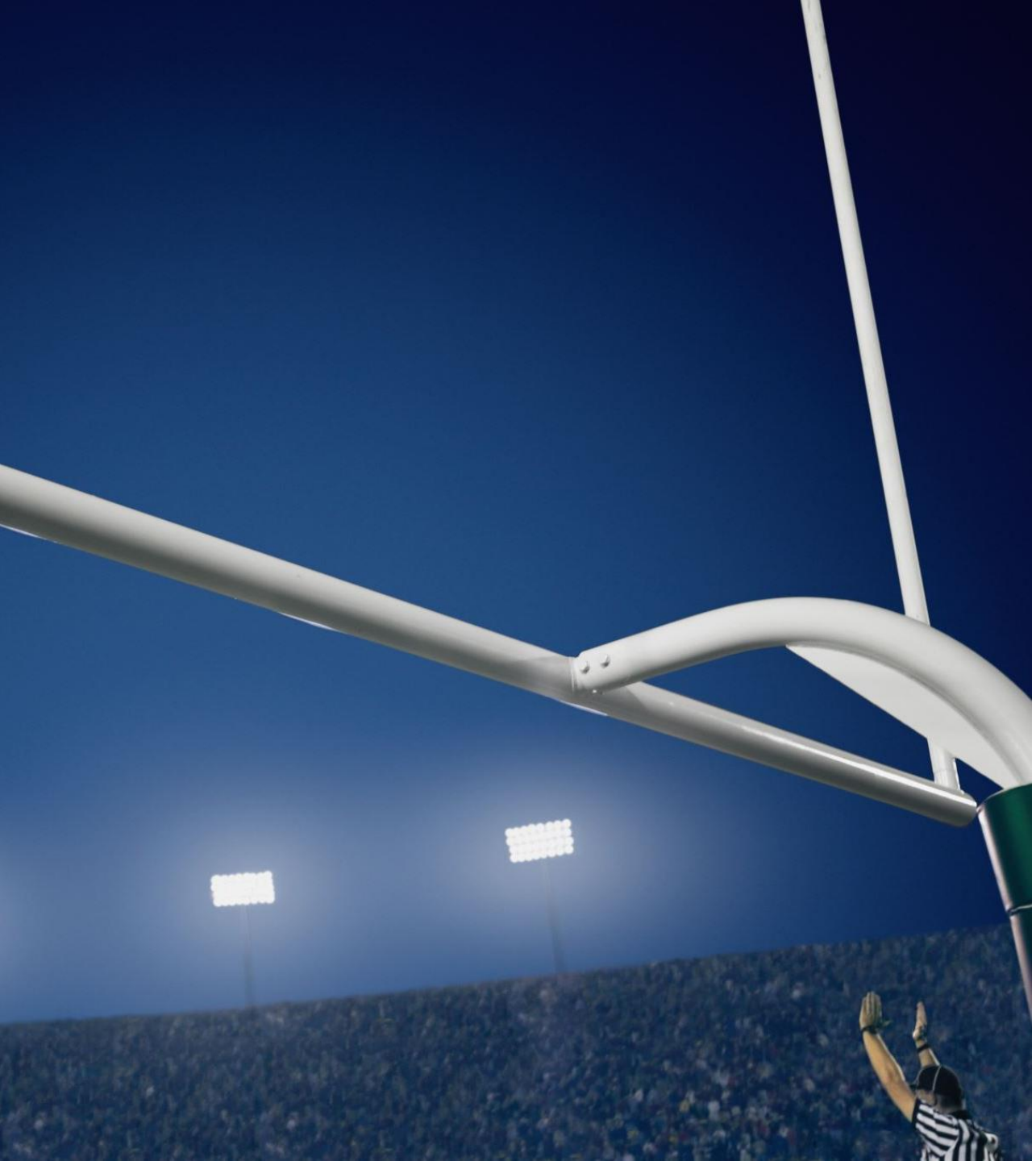


College Football EDA

DAN WILTSE

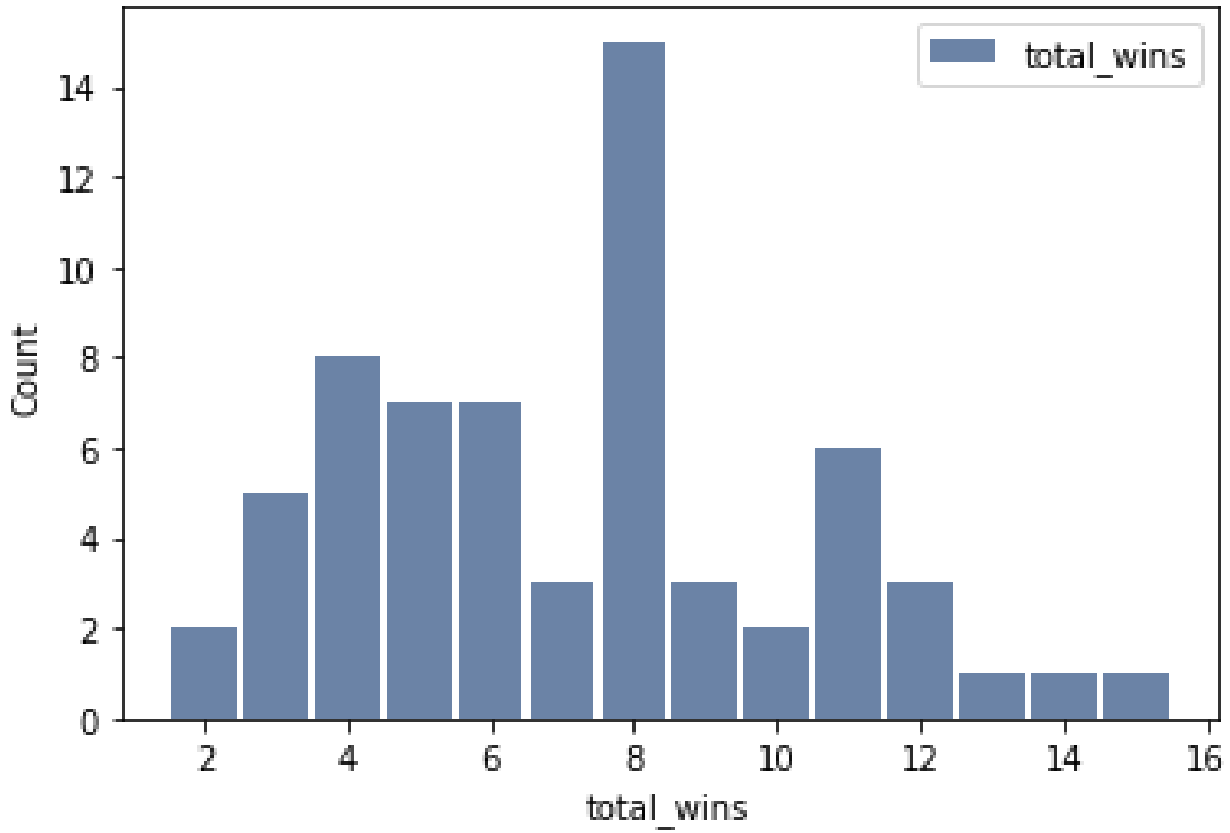
05/30/2020

WHAT FACTORS CONTRIBUTE TO COLLEGE FOOTBALL TEAMS WINNING MORE THAN OTHERS?

Statistical question

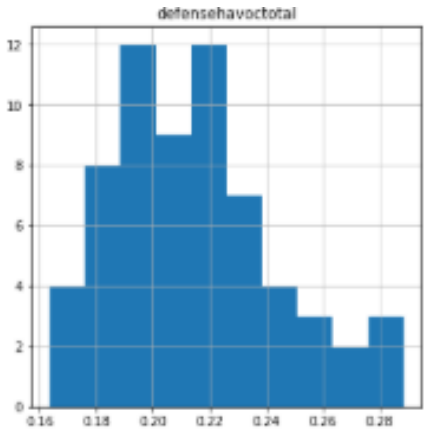
Variable	Description
Wins	Total wins during the 2019 college football season
PF	Points scored by the team during the season
PA	Points scored by opposing team during the season
Five_year_avg_recruit_rank	Average recruiting class rank across previous five recruiting classes (2015-2019)
OffensesuccessRate	efficiency metric that determines the success of a play. Successful plays meet one of the following criteria: • the offense scored • 1st downs which gain at least 50% of the yards to go • 2nd downs which gain at least 70% of the yards to go • 3rd and 4th downs which gain at least 100% of the yards to go
DefensesuccessRate	Same criteria for offense success rate above, only for team's defense
Defensehavoctotal	percentage of plays in with the defense recorded a TFL, forced a fumble, intercepted a pass or broke up a pass

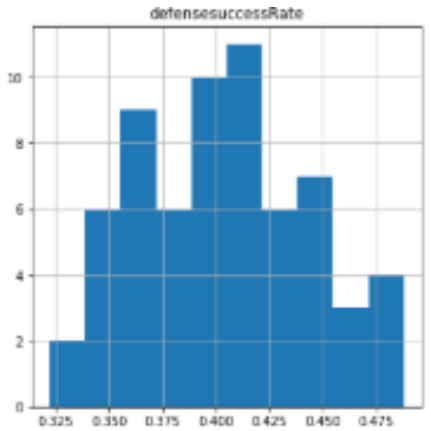
Variables Used

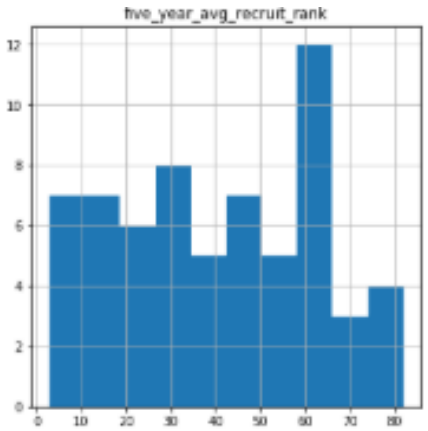


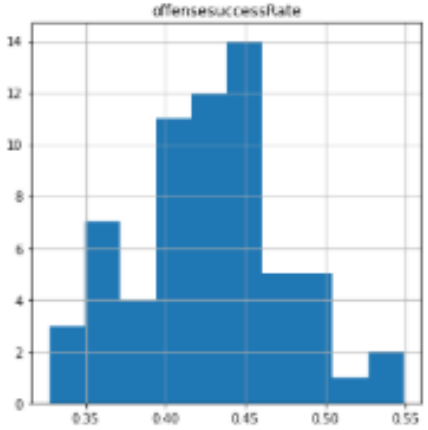
Histogram

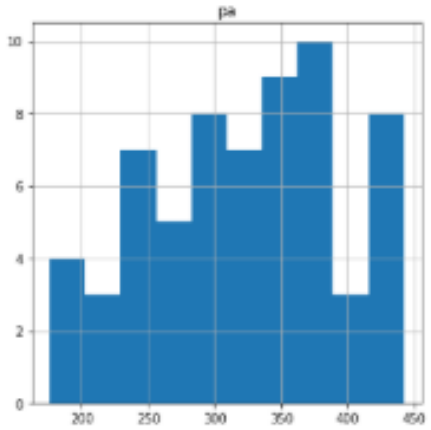
Most common win total is 8

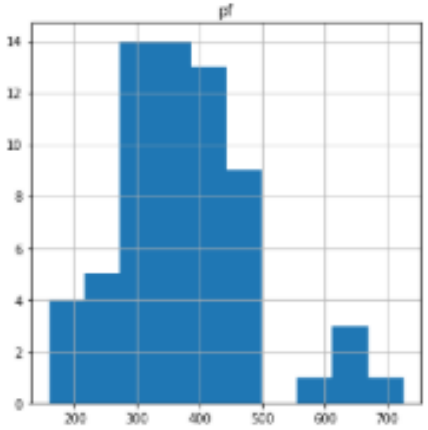












Histograms

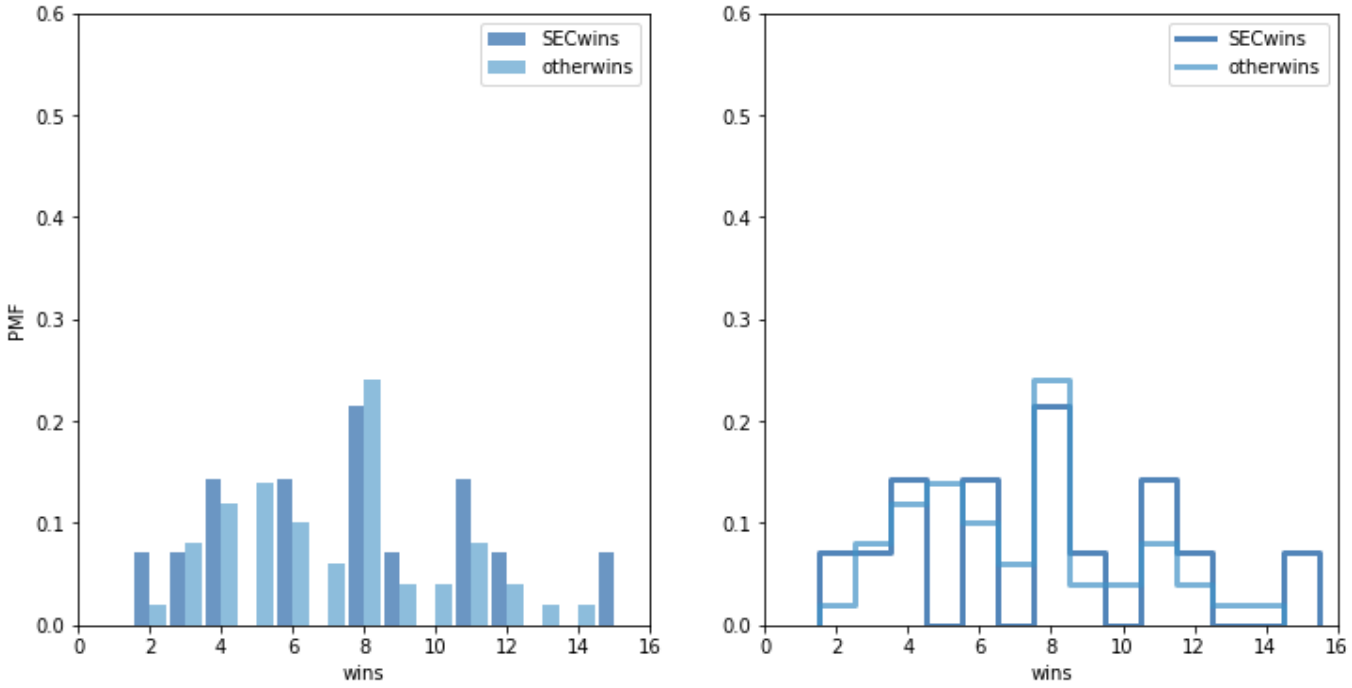
- None of the variables had normal distributions due to certain teams having extremely strong offenses or defense (or extremely weak defenses and offenses)

- Kept in variables as they impacted win totals

- Ex – Points For – there were 5 teams that scored more than 500 points, all but one of them made the playoffs

Descriptive Statistics

	wins	pf	pa	five_year_avg_recruit_rank	offensesuccessRate	defensesuccessRate	defensehavoctotal
count	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000	64.000000
mean	7.187500	378.281250	321.250000	41.003125	0.426212	0.404381	0.214900
std	3.085424	111.801194	71.501582	22.335181	0.046824	0.040480	0.028833
min	2.000000	159.000000	176.000000	2.800000	0.327422	0.322093	0.163415
25%	5.000000	303.750000	275.000000	20.700000	0.398367	0.370277	0.194158
50%	7.500000	365.000000	324.500000	41.900000	0.427156	0.404907	0.211306
75%	9.000000	431.250000	378.500000	61.250000	0.450596	0.433590	0.230357
max	15.000000	726.000000	442.000000	82.000000	0.549020	0.488082	0.288483

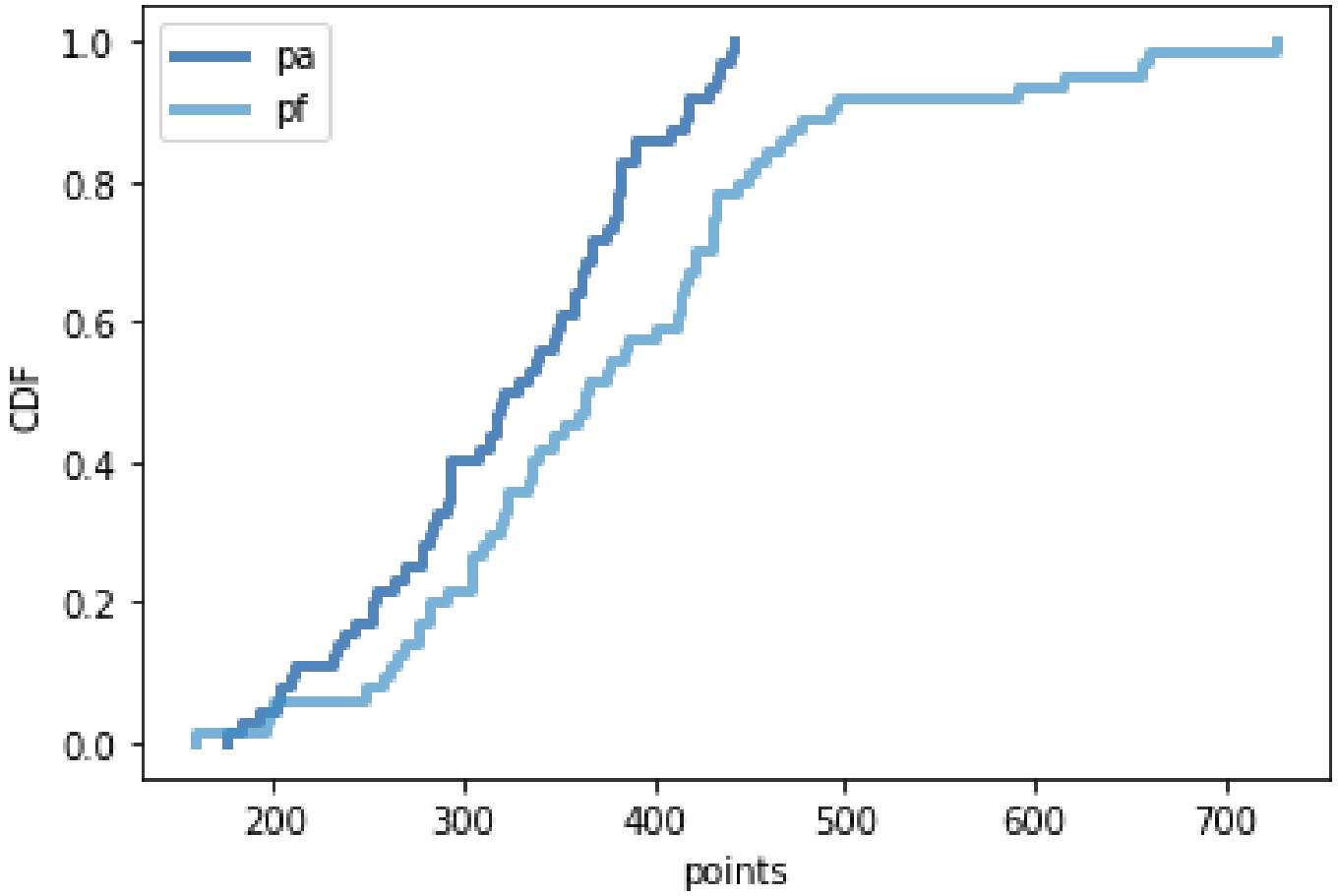


PMF – SEC wins vs all other conferences

- Comparing SEC wins versus all other conferences

- SEC is skewed right compared to other conferences due to two highly successful teams (Alabama and LSU) but both conferences had most teams with 8 wins

- SEC also had slightly higher percent of PMF in lowest portion of PMF as well



CDF

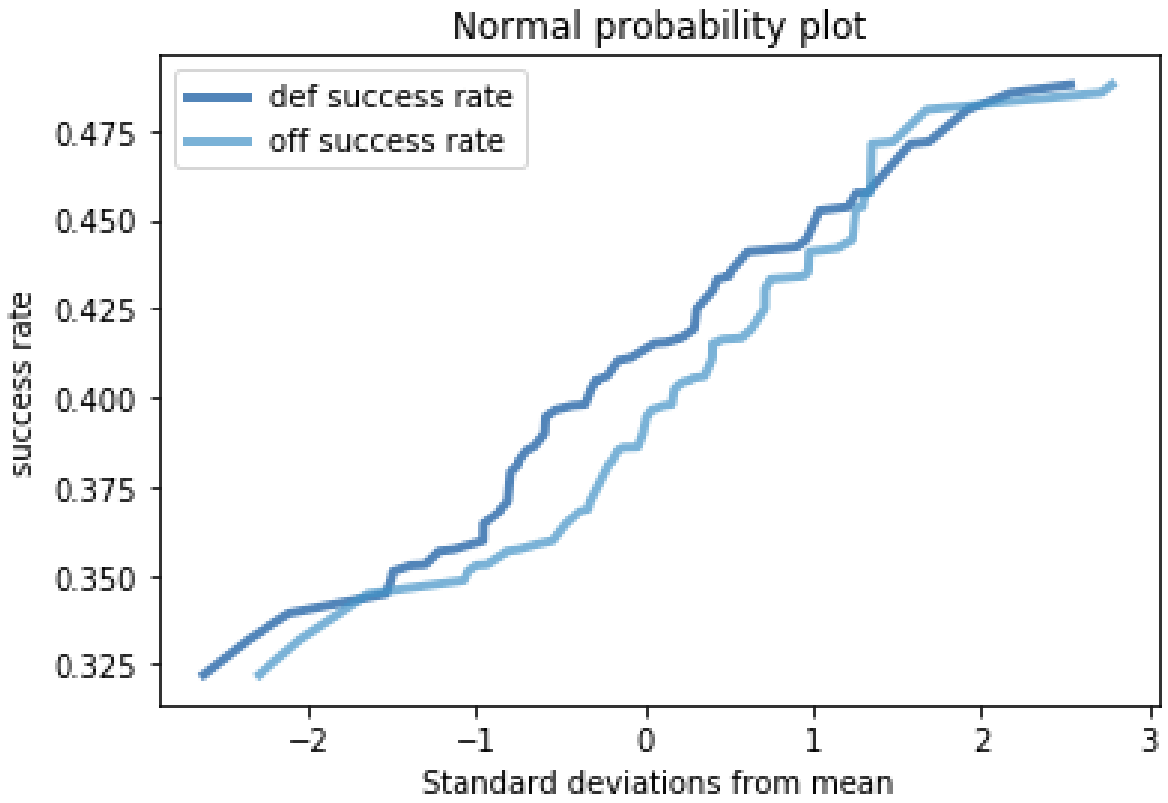
Comparison of Points For and Points Against

- Longer tail on points for than points given up, showing bigger variability in offensive points

Analytical Distribution

Normal Probability Plot

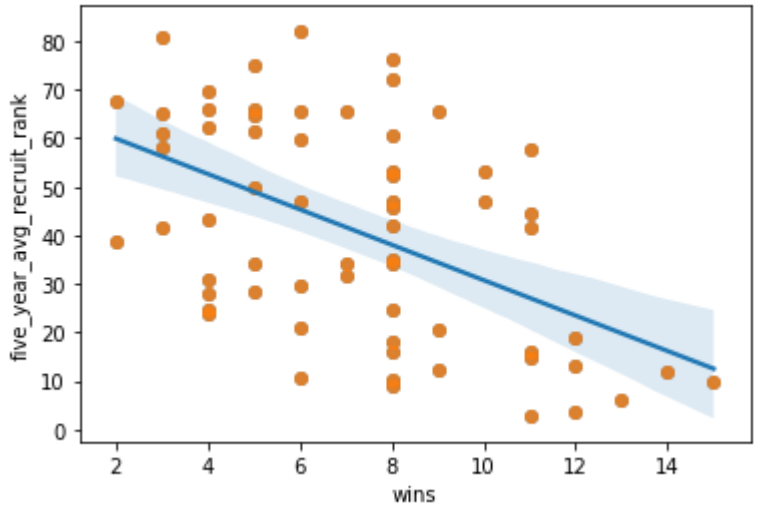
Similar trending of standard deviations across offensive and defensive success rate



Correlations

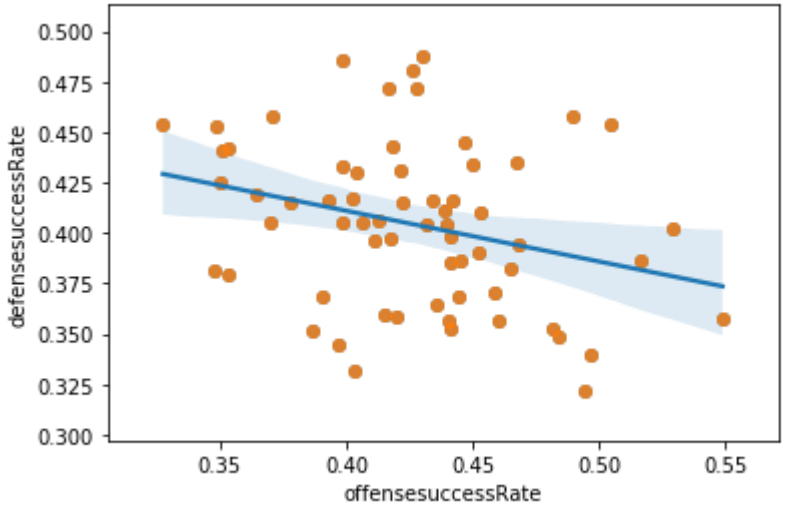
Wins by Recruiting Ranking

Better recruiting ranking related to more wins



Offense vs Defense Success Rate

Appears to be slight relationship between offense and defense success rate



Hypothesis Test

Correlation can be small but still be statistically significant. Correlation between two metrics is -.504, which means the lower your rank (1 is best rank), the more likely to win more games.

Hypothesis test below shows a relationship between the two variables as well, pvalue = 0.0.

Doesn't mean relationship is significant, just means that it's unlikely that the effect occurred by chance

```
In [23]: #Correlation between wins and recruiting ranking
         r = np.corrcoef(df1.wins, df1.five_year_avg_recruit_rank)
         r

Out[23]: array([[ 1.       , -0.5040221],
                [-0.5040221,  1.       ]])

In [24]: #Hypothesis Test
         cleaned = df1.dropna(subset=['offensesuccessRate', 'defensesuccessRate'])
         data = cleaned.wins.values, cleaned.five_year_avg_recruit_rank.values
         ht = CorrelationPermute(data)
         pvalue = ht.PValue()
         pvalue

Out[24]: 0.0
```

Regression Analysis – All Metrics

Looked at both offensive and defensive metrics combined

Points Against and Points for were significant variables

None of the other variables were significant

Adjusted R Squared - .85

OLS Regression Results

Dep. Variable:	wins	R-squared:	0.864
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	60.34
Date:	Sat, 30 May 2020	Prob (F-statistic):	6.89e-23
Time:	09:31:35	Log-Likelihood:	-98.578
No. Observations:	64	AIC:	211.2
Df Residuals:	57	BIC:	226.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.7125	4.034	0.672	0.504	-5.365	10.790
pf	0.0185	0.003	5.541	0.000	0.012	0.025
pa	-0.0178	0.004	-4.625	0.000	-0.026	-0.010
five_year_avg_recruit_rank	0.0041	0.008	0.489	0.627	-0.013	0.021
offensesuccessRate	1.7164	7.467	0.230	0.819	-13.237	16.670
defensesuccessRate	2.2928	8.251	0.278	0.782	-14.230	18.816
defensehavoctotal	6.4433	7.676	0.839	0.405	-8.927	21.814

Omnibus:	4.503	Durbin-Watson:	1.890
Prob(Omnibus):	0.105	Jarque-Bera (JB):	2.155
Skew:	-0.115	Prob(JB):	0.340
Kurtosis:	2.131	Cond. No.	3.55e+04

Regression Analysis – Defense Only

Looked at defense only metrics to see if defense wins championships (or predicts wins)

Points Against and Recruiting Ranking were significant variables

Defense Success Rate and Defense Havoc Total were not statistically significant

Adjusted R Squared: .512

OLS Regression Results

Dep. Variable:	wins	R-squared:	0.543
Model:	OLS	Adj. R-squared:	0.512
Method:	Least Squares	F-statistic:	17.55
Date:	Thu, 28 May 2020	Prob (F-statistic):	1.54e-09
Time:	22:09:32	Log-Likelihood:	-137.33
No. Observations:	64	AIC:	284.7
Df Residuals:	59	BIC:	295.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	14.4458	6.687	2.160	0.035	1.064	27.827
pa	-0.0154	0.007	-2.317	0.024	-0.029	-0.002
five_year_avg_recruit_rank	-0.0294	0.014	-2.108	0.039	-0.057	-0.001
defensesuccessRate	-12.3210	13.641	-0.903	0.370	-39.617	14.975
defensehavoctotal	17.9950	13.704	1.313	0.194	-9.426	45.416

Omnibus:	0.642	Durbin-Watson:	1.758
Prob(Omnibus):	0.725	Jarque-Bera (JB):	0.255
Skew:	0.136	Prob(JB):	0.880
Kurtosis:	3.148	Cond. No.	2.27e+04

Summary

- Several Factors contribute to wins in college football

- As expected, the amount of points scored (or given up) had highest impact on winning

- Recruiting ranking also highly correlated but was only significant in defense-only regression model

- Offense and defense success rate didn’t appear to have a big impact on predicting wins in the regression model

- Further analysis needed to see if there are conference difference impacts in these metrics, versus looking at all teams across all conferences.