NFL - FIRST DOWN OR TOUCHDOWN PREDICTION ANALYSIS

BY BALA KANDIKONDA

INTRODUCTION

NFL Big Data Bowl

- The dataset contains play-by-play data broken into frames for all passing plays in Weeks 1-8 of the 2021 season

Goal

- Uncover findings that can help players and coaches make in-play adjustments to improve the success rate of their play

DATA

Pre-Processing

- All features were checked for inconsistent, missing, or NA values
- Created dummy variables for all Categorical features

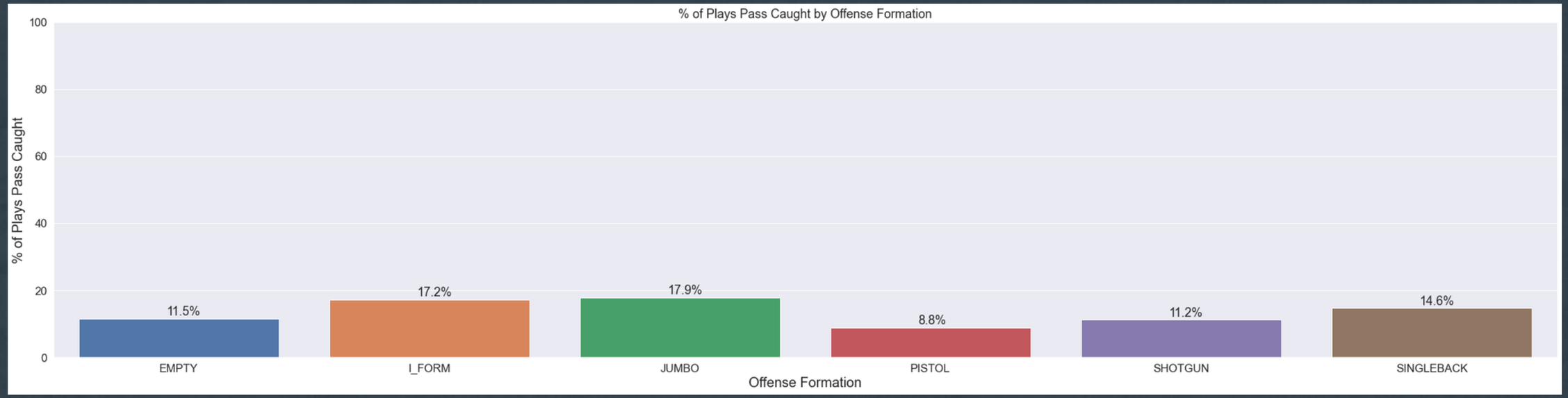
Metrics Created

- WR separation from closest defender
- WR distance from the football
- WR receiving route
- First Down or Score Achieved

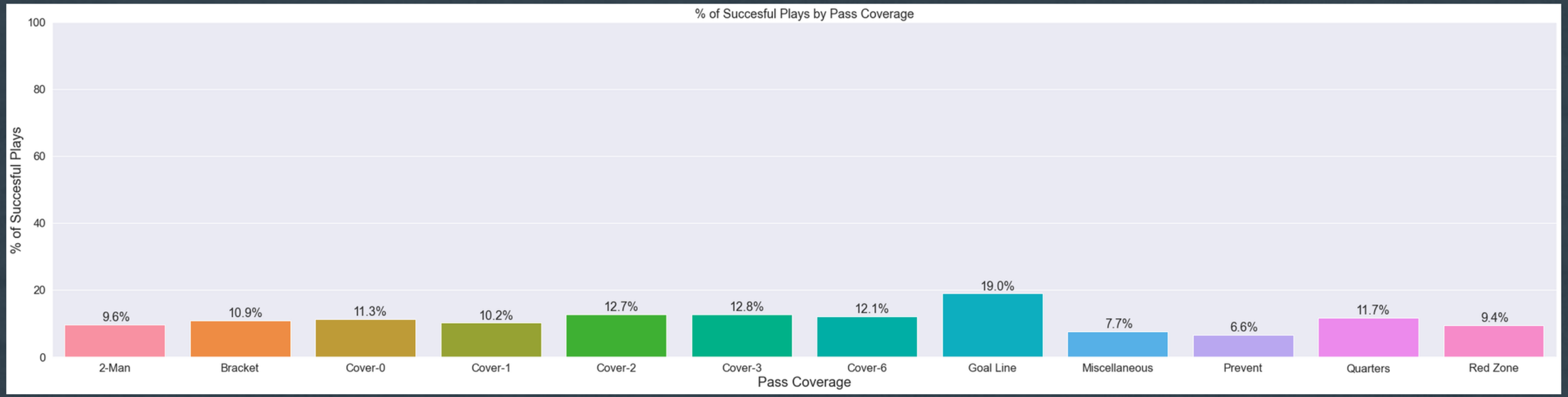
Final Dataset

- Grouped all relevant features by play so each row is the data from each play and used the 'First Down or Score Achieved' metric as the classifier

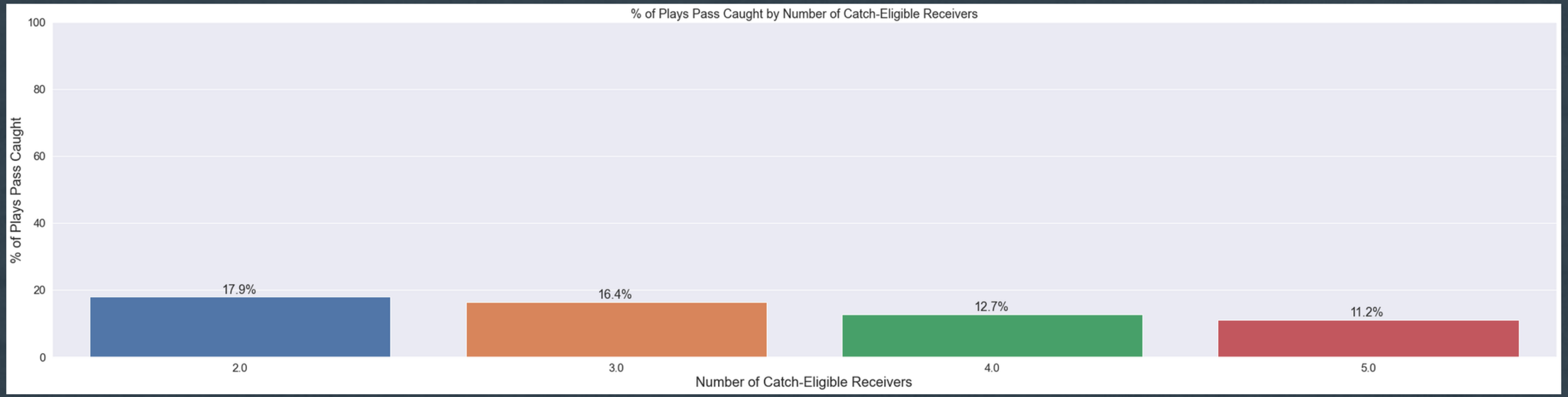
EDA: SUCCESS RATE BY OFFENSIVE FORMATION



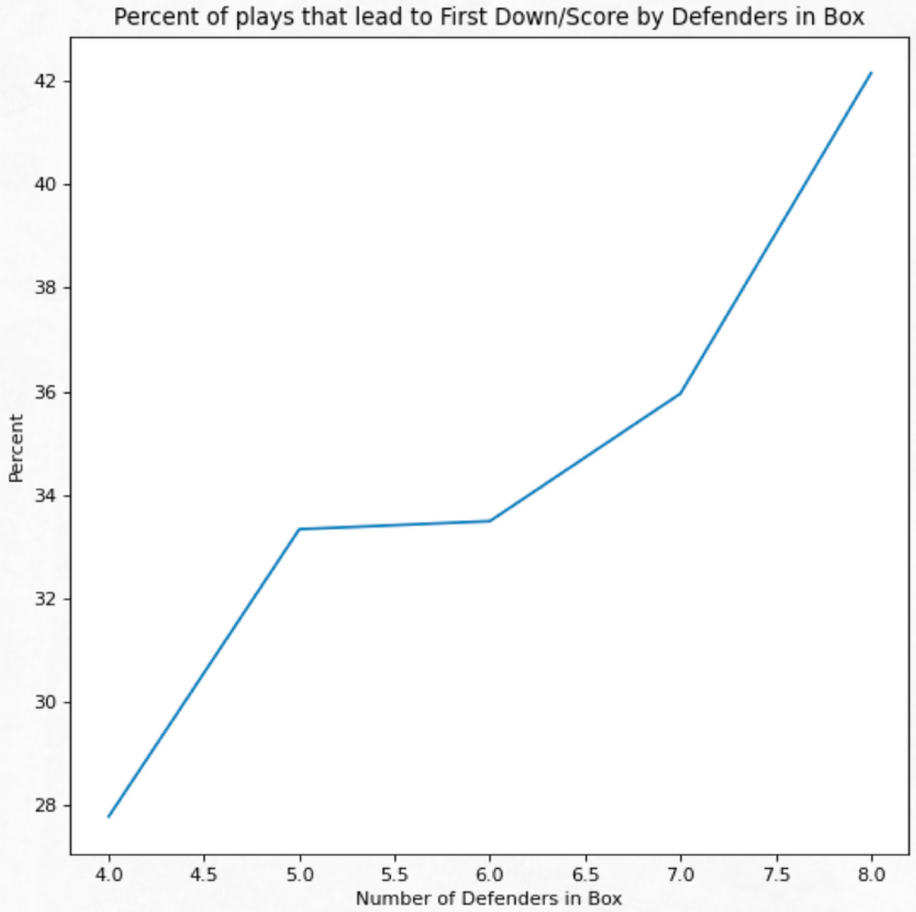
EDA: SUCCESS RATE BY PASS COVERAGE



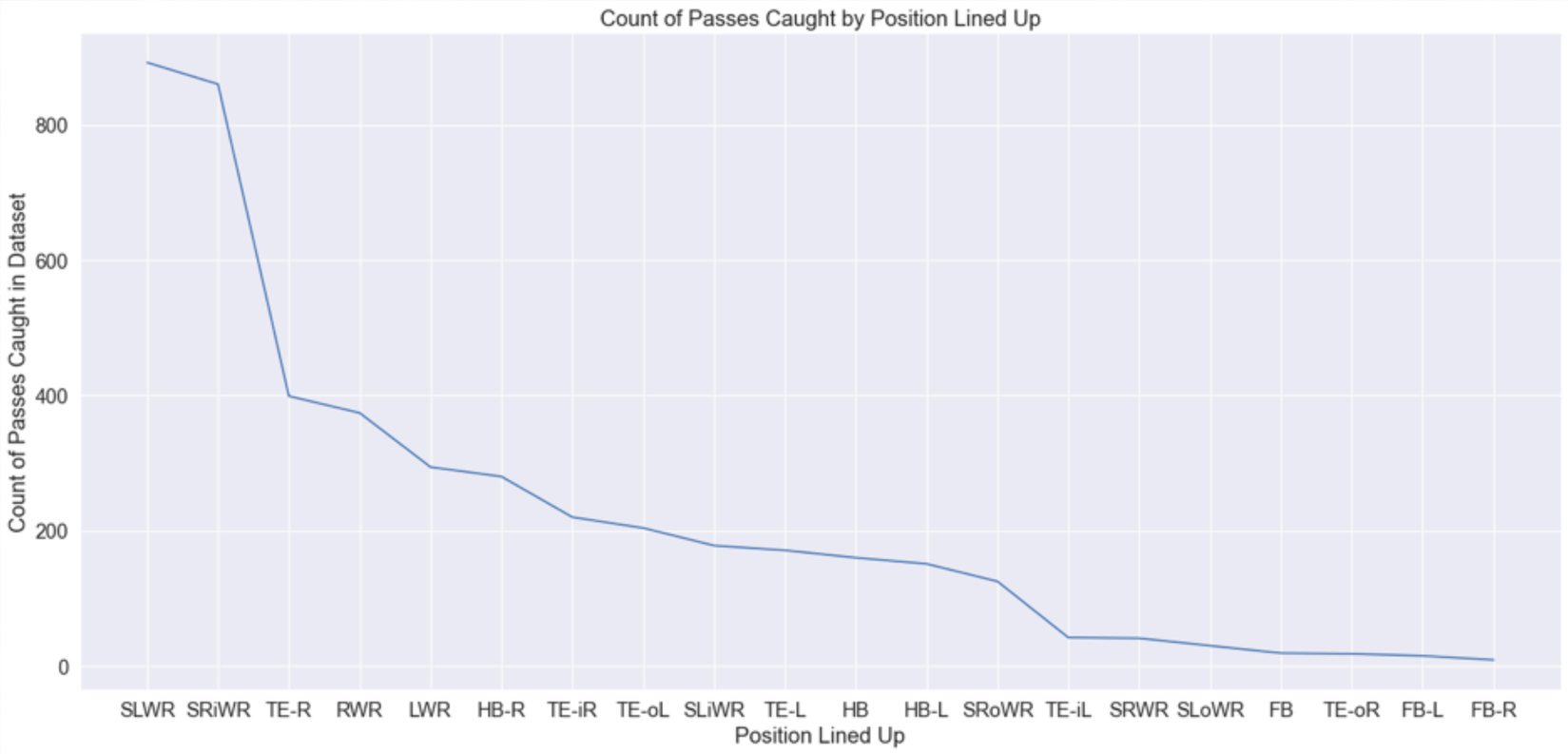
EDA: SUCCESS RATE BY RECEIVERS



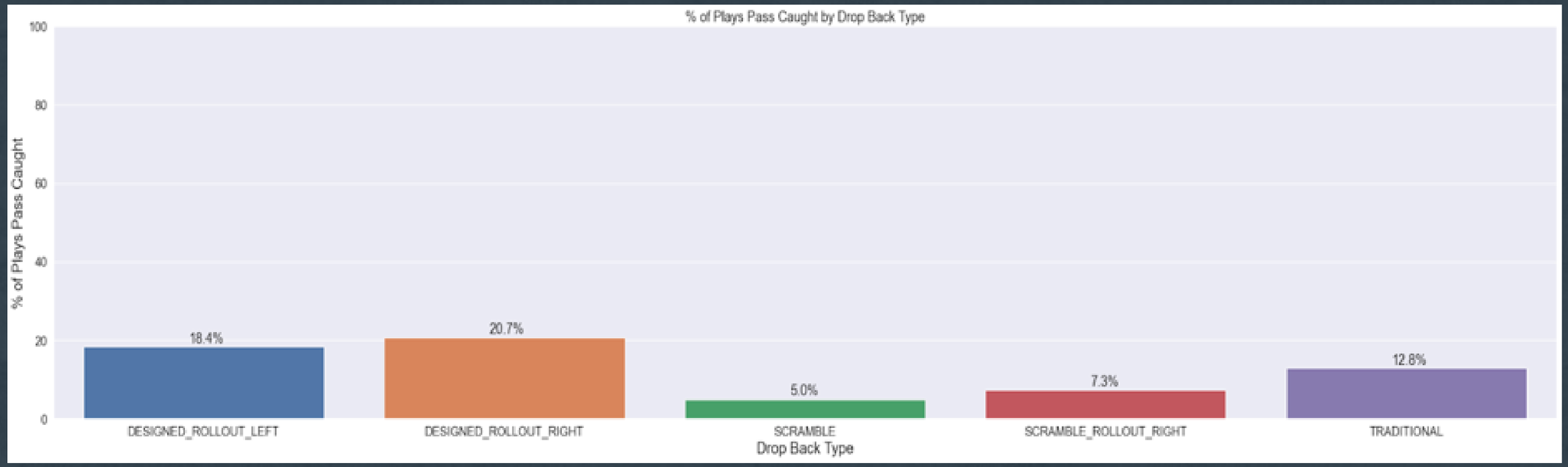
EDA: SUCCESS RATE BY DEFENDERS IN THE BOX



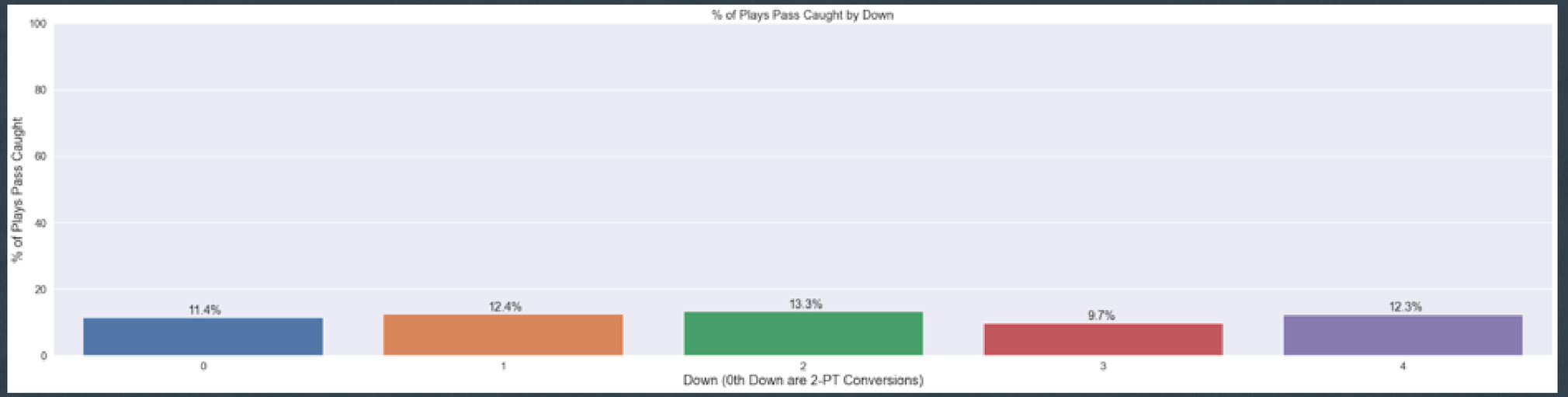
EDA: PASSES CAUGHT BY POSITION LINED UP



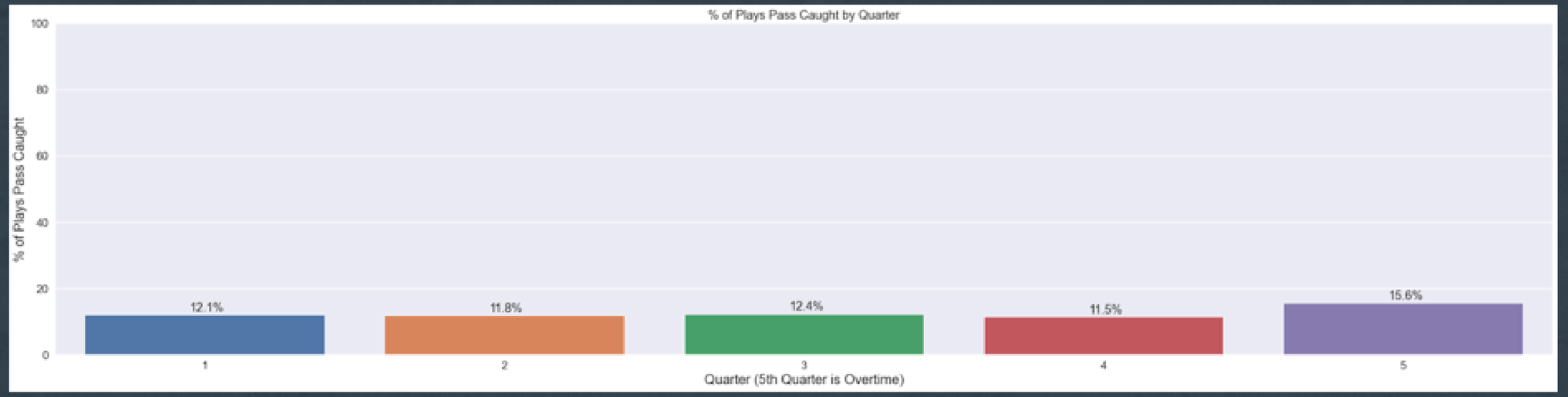
EDA: SUCCESS RATE BY DROP BACK TYPE



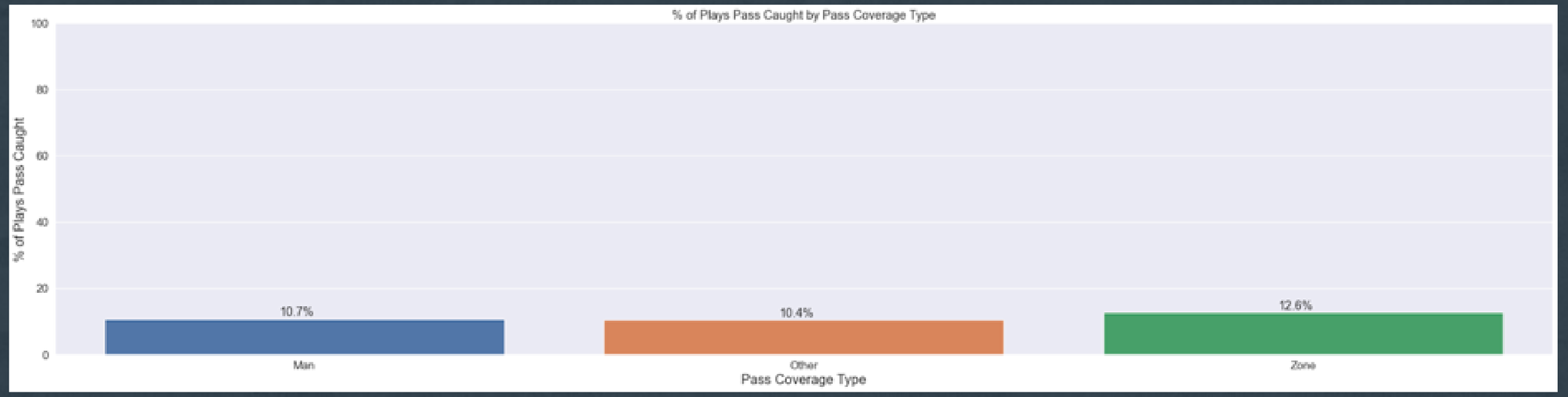
EDA: SUCCESS RATE BY DOWN



EDA: SUCCESS RATE BY QUARTER



EDA: SUCCESS RATE BY PASS COVERAGE TYPE



MODELING: CLASSIFICATION REPORTS

Classification Report for Logistic Regression model - 3 Receivers

	precision	recall	f1-score	support
0	0.58	0.97	0.72	112
1	0.40	0.02	0.05	82
accuracy			0.57	194
macro avg	0.49	0.50	0.39	194
weighted avg	0.50	0.57	0.44	194

Classification Report for Logistic Regression model - 4 Receivers

	precision	recall	f1-score	support
0	0.63	0.92	0.75	314
1	0.46	0.11	0.18	193
accuracy			0.61	507
macro avg	0.54	0.52	0.46	507
weighted avg	0.56	0.61	0.53	507

Classification Report for Logistic Regression model - 5 Receivers

	precision	recall	f1-score	support
0	0.67	1.00	0.80	1191
1	0.00	0.00	0.00	581
accuracy			0.67	1772
macro avg	0.34	0.50	0.40	1772
weighted avg	0.45	0.67	0.54	1772

Classification Report for Random Forest model - 3 Receivers

	precision	recall	f1-score	support
0	0.58	1.00	0.73	112
1	0.00	0.00	0.00	82
accuracy			0.58	194
macro avg	0.29	0.50	0.37	194
weighted avg	0.33	0.58	0.42	194

Classification Report for Random Forest model - 4 Receivers

	precision	recall	f1-score	support
0	0.62	1.00	0.77	314
1	1.00	0.01	0.02	193
accuracy			0.62	507
macro avg	0.81	0.51	0.39	507
weighted avg	0.77	0.62	0.48	507

Classification Report for Random Forest model - 5 Receivers

	precision	recall	f1-score	support
0	0.67	1.00	0.80	1191
1	0.50	0.01	0.02	581
accuracy			0.67	1772
macro avg	0.59	0.50	0.41	1772
weighted avg	0.62	0.67	0.55	1772

Classification Report for KNN model - 3 Receivers

	precision	recall	f1-score	support
0	0.59	0.90	0.71	112
1	0.52	0.15	0.23	82
accuracy			0.58	194
macro avg	0.56	0.52	0.47	194
weighted avg	0.56	0.58	0.51	194

Classification Report for KNN model - 4 Receivers

	precision	recall	f1-score	support
0	0.61	0.91	0.73	314
1	0.27	0.05	0.09	193
accuracy			0.59	507
macro avg	0.44	0.48	0.41	507
weighted avg	0.48	0.59	0.49	507

Classification Report for KNN model - 5 Receivers

	precision	recall	f1-score	support
0	0.67	0.99	0.80	1191
1	0.31	0.01	0.02	581
accuracy			0.67	1772
macro avg	0.49	0.50	0.41	1772
weighted avg	0.55	0.67	0.54	1772

CONCLUSION

Best Model

- Based on the metrics on the slide before, the Random Forest model performed the best overall out of all 3 models for all 3 receiver-sets
  - 3-Receivers
    - Highest accuracy and recall
  - 4-Receivers
    - Highest accuracy and precision
  - 5-Receivers
    - Highest precision and recall

Top Features

- No individual feature really stood out as an indicator of success
- This could be due to the volatility and the numerous moving parts of each play in the NFL
- There are a ton of interesting findings within the dataset to help teams and players succeed but the models were not accurate enough to make significant changes to the gameplay

FURTHER ANALYSIS

More Data

- Since the dataset had to be divided into 3 separate groups of plays, more data can help smooth out the variances between models
- Adding features that weren't included in the dataset such as player performance ratings can help take the analysis to the next step