Connor Ehn, Gavin Manley, Shane Johnson

DS 150 Capstone Project Report

1. The question you explored

We have two main questions. How do down number, field position, conversion distance, time, and other factors determine how an NFL team calls plays? When an NFL team is calling a play, each of these factors is important. If you need 20 yards on a 3rd down, running the ball is likely not going to get you all the way. We wanted to see if we could notice patterns in certain situations and see if other factors were more important than others.

Our other question was, can we reliably predict when a team would call a play based on our data? We wanted to see if through the patterns we were able to find in the previous question, we could accurately determine what the average NFL team would do in that scenario. Some scenarios (like the previously mentioned 3rd & 20) would likely result in most teams calling a deep pass play, so there are standards across the league.

2. Data collection process and answering questions

Our data collection was very simple. We found a Kaggle post with play-by-play data of every season since 2017. We wanted a full season, so we chose the most recent full year of 2024. This contained a large CSV of over 55,000 rows, one for each play by every team every week of the season (preseason and playoffs included). We cut out all the non-regular season games and polished the data into a new CSV, which is what we used for our code. We still had a very large amount of data to work with (~45,000 rows), so we thought that this would be very good to visualize. We created line charts, bar charts, and heatmaps with it.

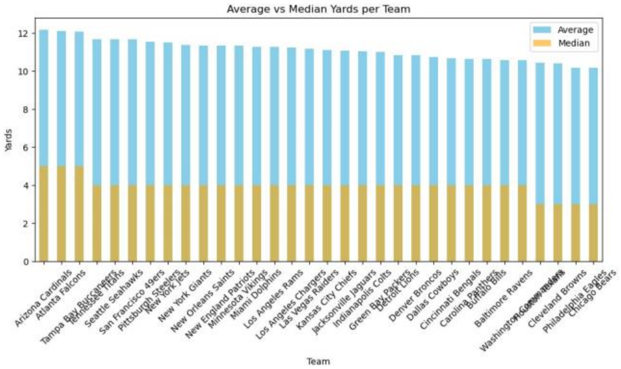
For the first question, we found multiple different sources of play-by-play data for the entire NFL. Our most important source was from Kaggle and contained a CSV of every play for the entire 2024 season. This is what we used to create the main league predictor.

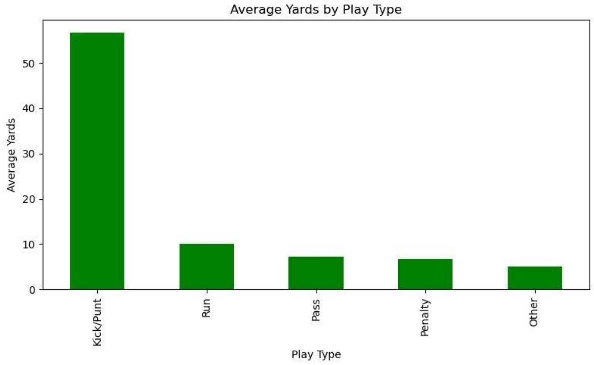
For the second question, we used the same sources as question 1, but we looked at the trends from the play-by-play data to distinguish what teams may do in certain scenarios. We

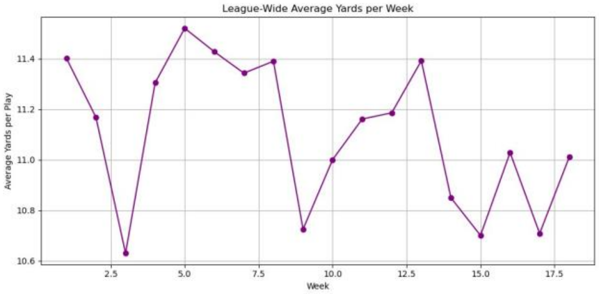
Kaggle source: https://www.kaggle.com/datasets/keonim/nfl-game-scores-dataset-2017-2023

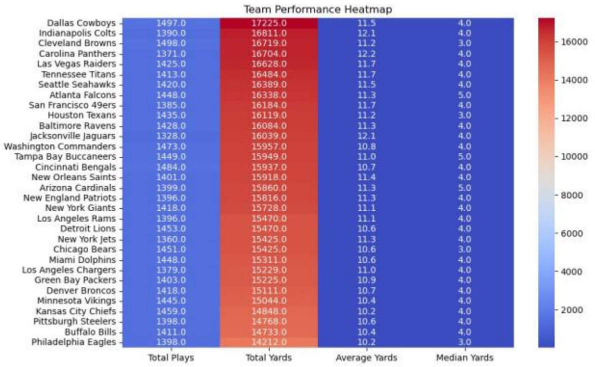
3. Key findings with visualizations

We believe that there is definitely a way to predict plays. Our model was quite simple, taking only 5 inputs, and was made to function for the entire league. There are some things that are very difficult to get play by play data on that are extremely important to the play call. Formation, the teams' dependencies on certain players, and other factors would be crucial. Even with our simple model, we still had almost 64% accuracy, which is significantly better than guessing.

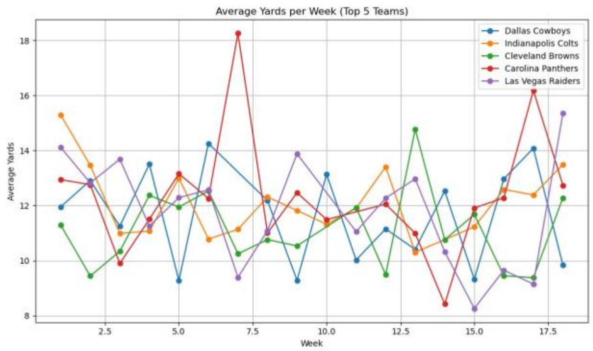







Team Performance Heatmap

	Total Plays	Total Yards	Average Yards	Median Yards
Dallas Cowboys	1497.0	17225.0	11.5	4.0
Indianapolis Colts	1390.0	16811.0	12.1	4.0
Cleveland Browns	1498.0	16719.0	11.2	3.0
Carolina Panthers	1371.0	16704.0	12.2	4.0
Las Vegas Raiders	1425.0	16628.0	11.7	4.0
Tennessee Titans	1413.0	16484.0	11.7	4.0
Seattle Seahawks	1420.0	16389.0	11.5	4.0
Atlanta Falcons	1448.0	16338.0	11.3	5.0
San Francisco 49ers	1385.0	16184.0	11.7	4.0
Houston Texans	1435.0	16119.0	11.2	3.0
Baltimore Ravens	1428.0	16084.0	11.3	4.0
Jacksonville Jaguars	1328.0	16039.0	12.1	4.0
Washington Commanders	1473.0	15957.0	10.8	4.0
Tampa Bay Buccaneers	1449.0	15949.0	11.0	5.0
Cincinnati Bengals	1484.0	15937.0	10.7	4.0
New Orleans Saints	1401.0	15918.0	11.4	4.0
Arizona Cardinals	1399.0	15860.0	11.3	5.0
New England Patriots	1396.0	15816.0	11.3	4.0
New York Giants	1418.0	15728.0	11.1	4.0
Los Angeles Rams	1396.0	15470.0	11.1	4.0
Detroit Lions	1453.0	15470.0	10.6	4.0
New York Jets	1360.0	15425.0	11.3	4.0
Chicago Bears	1451.0	15425.0	10.6	3.0
Miami Dolphins	1448.0	15311.0	10.6	4.0
Los Angeles Chargers	1379.0	15229.0	11.0	4.0
Green Bay Packers	1403.0	15225.0	10.9	4.0
Denver Broncos	1418.0	15111.0	10.7	4.0
Minnesota Vikings	1445.0	15044.0	10.4	4.0
Kansas City Chiefs	1459.0	14848.0	10.2	4.0
Pittsburgh Steelers	1398.0	14768.0	10.6	4.0
Buffalo Bills	1411.0	14733.0	10.4	4.0
Philadelphia Eagles	1398.0	14212.0	10.2	3.0



4. Methodology

We first gathered our data from our source in the form of a large CSV. We cleaned it, removing all unnecessary rows for the non-regular season games. We then created our first Python file and wrote a formula to read the CSV into it as a DataFrame. We then reformatted the DataFrame so that it would work for both our visualizations and play predictor. We created 5 different visualizations with matplotlib and seaborn. For the predictor, we first categorized each play to create the initial pass and run ratio for each team. We then created each condition with its respective weighted score before finally creating the function that would take the inputs and return the prediction.

5. Conclusions, challenges faced, and future work

Overall we were satisfied with how the project came out for how much time we had to do it. The entire job of an NFL defense is to predict a play and stop it, and being able to do it two thirds of the time is not a bad success rate. There are plenty of ways to improve upon this idea, whether that be through simple weight adjustments or advanced statistics. Overall, I think we proved our initial question: it is possible to reliably predict NFL plays with a Python program.

The most difficult part of the project was determining how important each factor we considered was. Since our model operated by altering a preset score, certain conditions gave higher alterations than others. It is very difficult to accurately weigh these conditions, and the ones we settled upon could be very wrong without our knowing.

Any future work with this model should be an attempt to make it more accurate. Adding more conditions like formation, more accurate time considerations (quarter, score, etc.), and star players would not be overly complex and would likely improve the model greatly. While we have predicted plays at a more accurate level than just guessing at 50-50, we think there is much more that can be divulged from this data.