Group 13:

- Supra Khanal
- Claudia Otero
- Daniel Westerman
- Mihir Narayan

## **Overview:**

In this project, we want to analyze how the distance of a pass attempt (pass length) affects the likelihood of a completion in the NFL. We are going to use the 2025 NFL Big Data Bowl dataset (<u>https://www.kaggle.com/c/nfl-big-data-bowl-2025/data</u>) to predict whether a pass attempt will be a completion or an incompletion based on the pass length. We want to build a binary classification model that predicts whether a pass will be completed based solely on the throw distance. Our initial model will use the 'passLength' variable as the primary feature and the 'passResult' variable as the target.

The motivation behind this is to test the assumption that the farther a quarterback throws the ball, the less likely the pass is to be completed. We want to answer this overarching question: Does throw distance affect completion probability? Using this data, we can visualize the decline in completion probability, and identify distance thresholds where pass success rates drop.

One method we want to use is a logistic regression classifier, which will output a probability prediction for each distance. This will allow us to visualize how completion rates change as distance changes.

## Key Variables:

- Target: passResult (1 if complete, 0 otherwise)
- Feature: passLength (distance in yards beyond the line of scrimmage the balls travels)

## **Other Methods:**

- Clean and prepare data (drop NAs, recode passResult)
- Use matplotlib or seaborn to visualize the relationship between passLength and completion rate
- Explore non-linear models (decision trees) or add features

## Sample Code Snippet:

import pandas as pd
filepath = '/Users/skhanal/Desktop/plays.csv'
data = pd.read\_csv(filepath)

filtered\_data = data[['passLength', 'passResult']] cleaned = filtered\_data.dropna()