Executive Summary

Project Proposal

American football is one of the most popular sports in the United States, and it is also a very complex sport. There are many different factors that can affect the outcome of a game, including the players, the coaches, the plays, and the weather.

Our project will focus on using data to analyze American football. We will use data from past games to identify trends and patterns. We will also be using team statistics and team data for past seasons and use the game and team data to predict the outcome of future games and for this current NFL season.

We are interested in this project because we are big fanatics of the NFL and football in general. Also, we believe that this data could be used to improve the understanding of the sports and its results.

Our data sources for the project include:

- Game statistics
- Team statistics
- Weather data

We will use this data to answer the following research questions:

- What are the factors that most affect the outcome of an American football game?
- How can data be used to predict the outcome of future games?

We believe that this project has the potential to make a significant contribution to the field of American football. We are excited to see what we can learn from the data.

Description of Data Sources

NFL Box Scores

- Dataset includes box scores for every NFL game from 1960 to 2017
- Box scores include statistics for each home and away team for each game
- Another csv file includes the weather data for each game but we will not be using this data because it is not complete

NFL Game Results and Teams

- There are 2 main datasets that we will be using from this source
- The first dataset includes game results for every NFL game from 1966 to 2023
  - It also includes descriptive info such as if it is a playoff game, if it is played at a neutral site, and some weather information if available

    - It also includes betting data for each game such as the spread, over/under, and money line for the games
- The second dataset includes information regarding the teams in the NFL
    - It includes the team name, the team abbreviation, the conference, and the division

NFL Team Stats

- Dataset includes team statistics for each team for each season from 2010 to 2021
- The dataset includes basic information such as win-loss percentage and which season
- The dataset also includes more in depth information regarding offensive and defensive statistics such as points scored, yards gained, and turnovers
- It also includes playoff data for each team for each season

Description of Observational Units

NFL Box Scores

- Observational unit in the box scores data set is each game played in the NFL from 1960 to 2017
- There is an emphasis on the statistics between each home and away team
- Data set was built by an individual contributor in 2017, and no additional information is provided in Kaggle about the data set
- The observational unit in the weather data set is each game played in the NFL from 1960 to 2013
    - Data after 2013 is not included in the data set

NFL Game Results and Teams

- The observational unit in game results dataset is each game played in the NFL from 1966 to 2023
- The observational unit in the team dataset is each team in the NFL (including teams that are no longer in the NFL or have changed their names, i.e. the San Diego Chargers or the Washington Football Team)
- Data set was built from publicly available NFL data, weather provided by the NOAA, and betting data from a variety of sources but cross referenced with Pro Football Reference.

NFL Team Stats

- The observational unit in the team stats dataset is each team in the NFL for each season from 2010 to 2021
    - So, the dataset includes data for each of the 32 teams for each of the seasons from 2010 to 2021

- This data was created using three different sources: pro-football-reference.com, covers.com, teamrankings.com
- The author used various CSV files and web scraping to combine 88 tables to create this data set

Description of Main Variables of Interest

NFL Box Scores

- In this dataset, the main variables of interest for us would be home and away teams' first downs, their net yards, the total plays, average gain, and time of possession
- We trimmed down to these variables because this data will be valuable when combining with the other datasets
- We deemed these variables important since they give a good summary of the game and how the statistics differed throughout the seasons when teams were at home or away

NFL Scores and Teams

- ScoresGame Resultset Game Results main variables of interest in this dataset are the home and away teams' scores, the home and away teams' yards, and the home and away teams' turnovers - The main variables of interest in this dataset are the home and away teams' scores, the betting favorite, the date of the game, and the weather
  - The score and date variables were taken from publicly available NFl data, the weather was taken from the NOAA, and the betting favorite was taken from a variety of sources but cross referenced with Pro Football Reference
- Team Dataset
  - The main variables of interest in this dataset are the team name, the team abbreviation, the conference, and the division
  - These variables are important because they will be used to merge the data with the other datasets
  - This data was taken from public NFL data

NFL Team Stats

- The main variables of interest in this dataset are the team name, the season, the win-loss percentage, the point differential, the points scored per game, the points allowed per game, the yards gained per game, the yards allowed per game, and the possession time per game
- This data was taken from 3 sources: pro-football-reference.com, covers.com, teamrankings.come

Description of Data Cleaning Process

NFL Box Scores

- We decided to only take into account seasons from 2010 and onwards since this would be more relevant to the current NFL season
- First we got the columns information and decided which variables were important to us
- Next, we dropped the columns that we did not need and converted the date column to a datetime object
- We also converted all the numeric columns to numeric data types
- We discarded the weather csv file in this data set

NFL Game Results and Teams

- Game Result Dataset
  - We decided to only take into account seasons from 2010 and onwards since this would be more relevant to the current NFL season (especially so the players, coaches, and teams would be more relevant and similar to the current NFL season)
  - We first decided which columns were important to us and dropped the columns that we did not need
  - There were some rows missing data for scores, so we dropped those rows
  - We also converted the date column to a datetime object
  - There were some rows missing data for the weather, but we decided to keep them as the scores for those games were still present and we could still use that data for our analysis
- Team Dataset
  - We decided which columns were important to us and dropped the columns that we did not need such as the team conference and division before 2002 (especially because we are only considering seasons from 2010 and onwards)
  - There were some teams that did not have a current division, so we dropped those rows (these were teams that existed before 2002)
  - We realized that the conference value for the Jets was incorrect, so we manually changed that to the correct value
  - We sorted the teams by their division and conference to make it easier to read
  - We decided to keep the teams that are no longer in the NFL because their team name still exists in the game results and scores datasets as the teams changed names or moved cities (i.e. the San Diego Chargers to the Los Angeles Chargers or the Washington Football Team to the Washington Commanders)
    - This provides a way to connect the past team data with more recent team data as the names and cities of the teams have changed over time

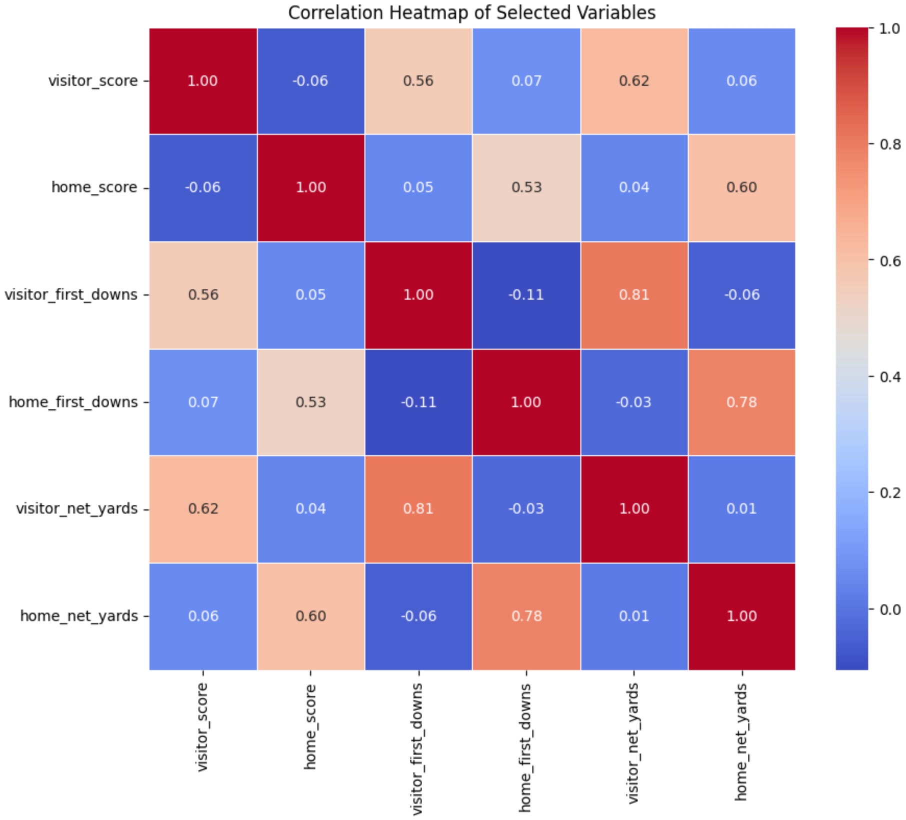
NFL Team Stats

- This data set only contains data from 2010 to 2021, so we did not need to filter out any data
- We decided which columns were important to us and dropped the columns that we did not need such as perc_punt_20, Turnover_perc, etc.
- There were no missing values in this data set, so we did not need to drop any rows

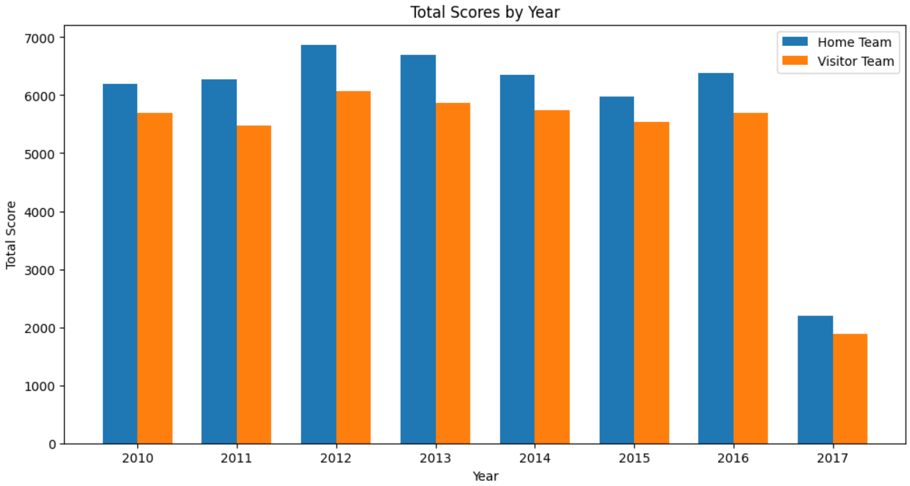
Visualizations

NFL Box Scores Visualizations

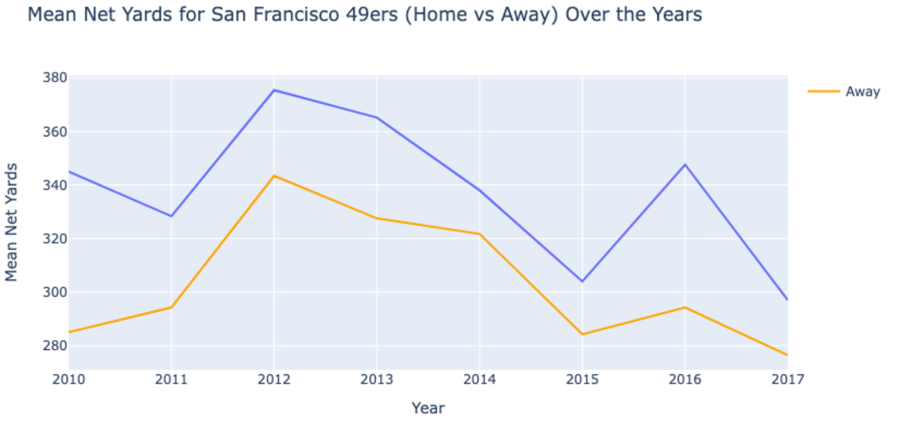
- Correlation Between Various Variables in the Box Scores Dataset
  - Created this for our purpose of seeing which variables are correlated with each other



- Total Scores by Year
  - Created this to see how the total scores have changed throughout the years

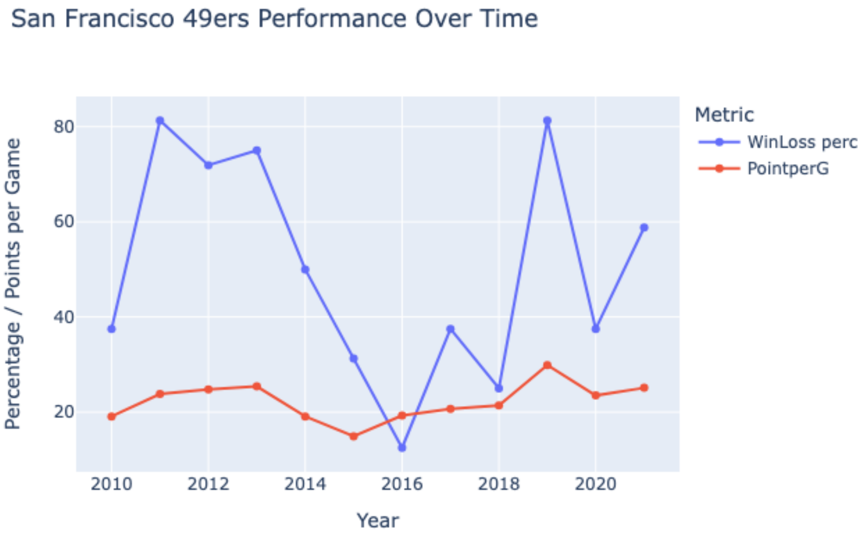


- Mean Net Yards by Year
  - Created this to see how the mean net yards have changed throughout the years for the 9ers

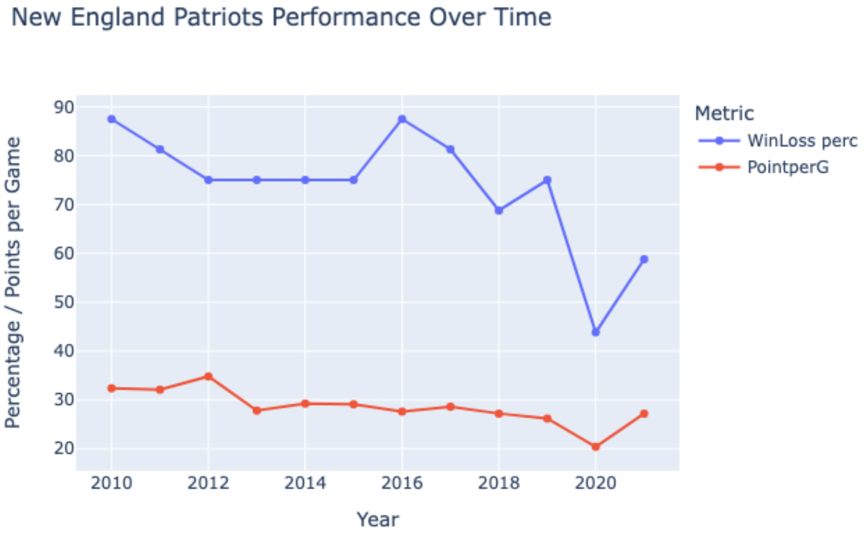


Team Stats Visualizations

- 9ers Performance Over Time
  - Shows the percentage / points per game over the seasons for the 9ers

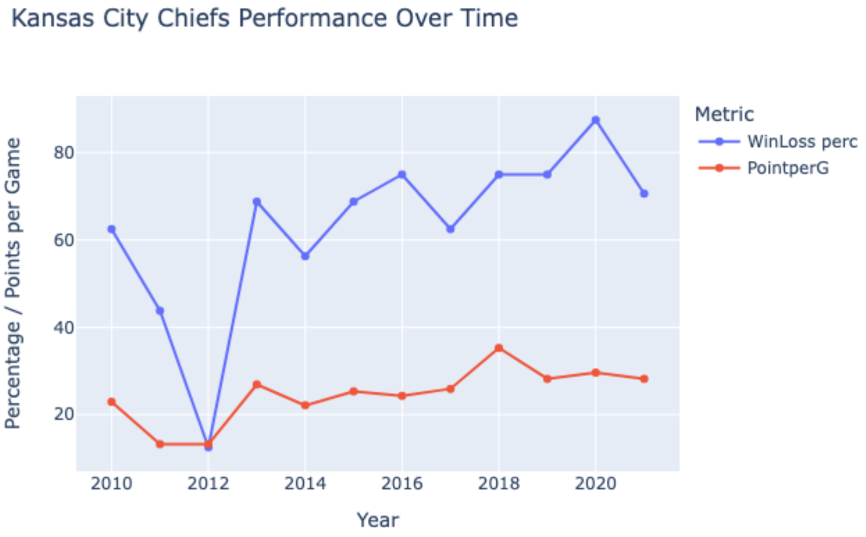


- Patriots Performance Over Time
  - Shows the percentage / points per game over the seasons for the Patriots

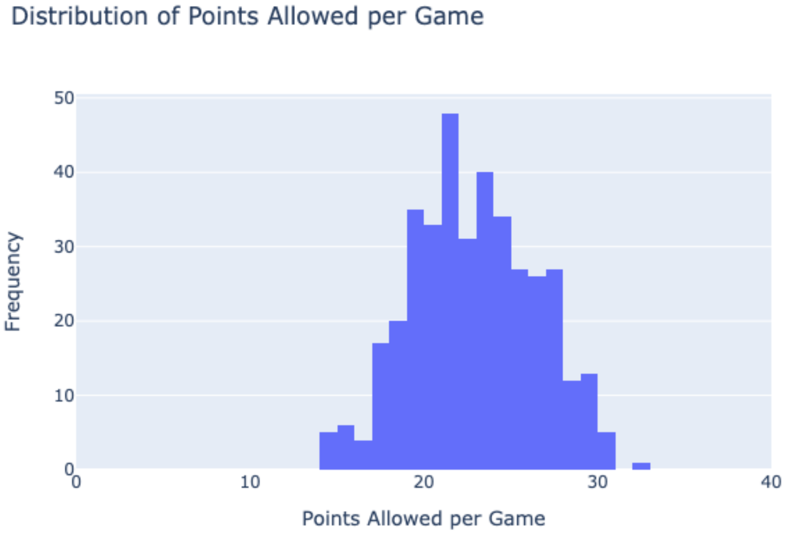


- Chiefs Performance Over Time
  - Shows the percentage / points per game over the seasons for the

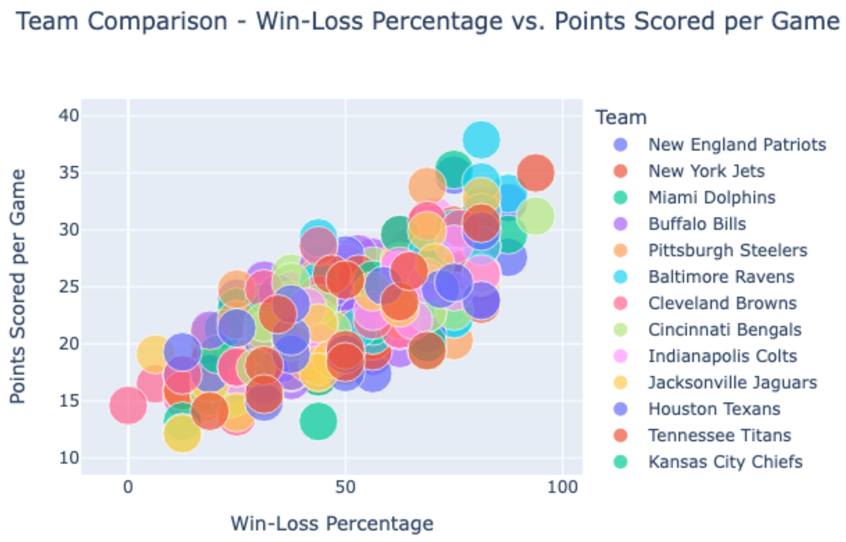
Chiefs



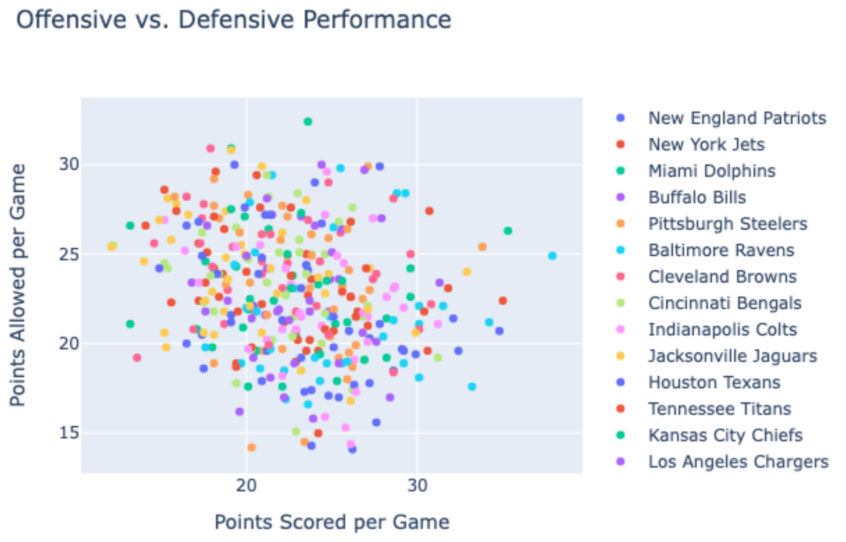
- Distribution of Points Allowed Per Game
  - Can demonstrate the defensive strength of a team



- Team Comparison of Win-Loss Percent vs Points Scored Per Game

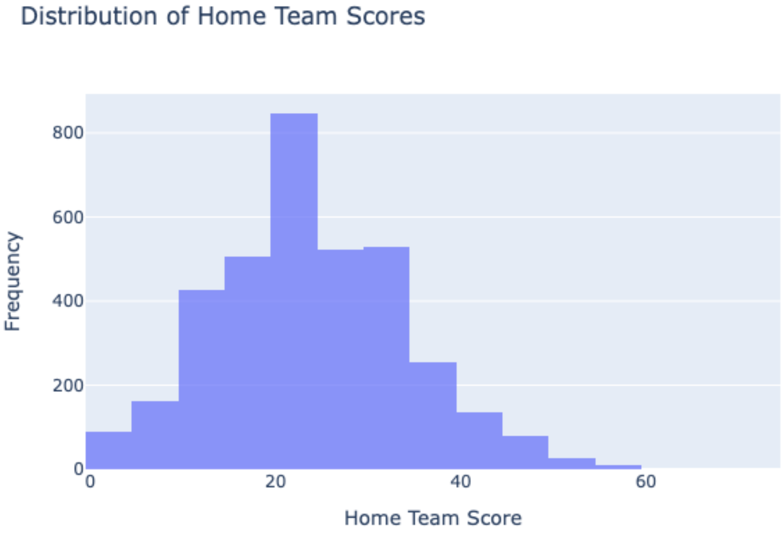


- Offensive vs Defensive Performance
  - Relationship between points allowed and points scored for each team

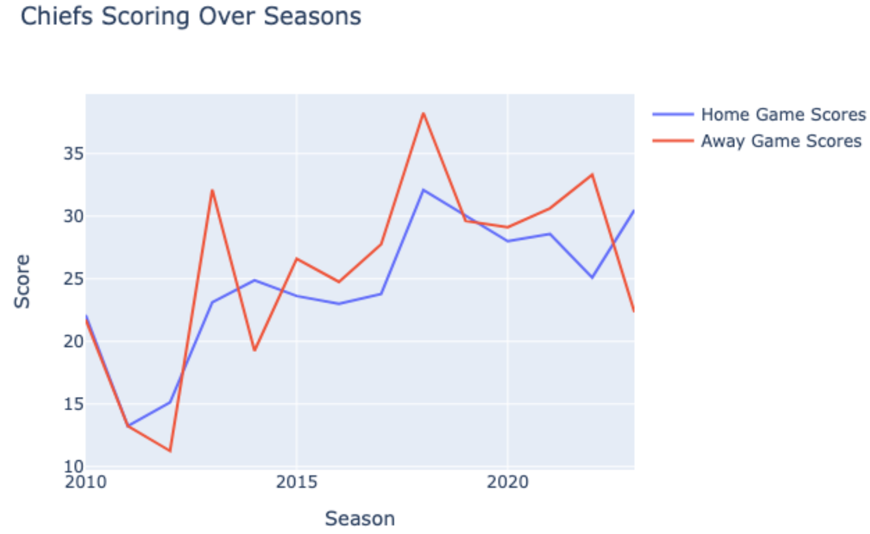


NFL Game Data Visualizations

- Distribution of Scores
  - Shows the distribution of scores for home teams



- Chiefs scoring over the Seasons for home and away games



Machine Learning

Setting up the Data

The first step in our ML process was to process all the cleaned data.

- We inspected all of our data and decided which teams names were missing/incorrect and fixed/added them in a dataframe
- Most importantly, we created a function called `get_team_id` which pulled the team id from the team dataframe applied it to the box scores and game data dataframes

```python
# function to get team id from city/team name
def get_team_id(city):
    # find the team name
    for team in nfl_teams["team_name"]:
        if city in team:
            return nfl_teams[nfl_teams["team_name"] == team]["team_id"].values[0]
        elif city == "NY Giants":
            return "NYG"
        elif city == "NY Jets":
            return "NYJ"
        elif city == "LA Rams":
            return "LAR"
        elif city == "LA Chargers":
            return "LAC"
```

- This ensured that we had a common key to join the dataframes on

Joining the Data

After the initial set up of the data was done, our next step was to join the box score dataset with the game data dataset.

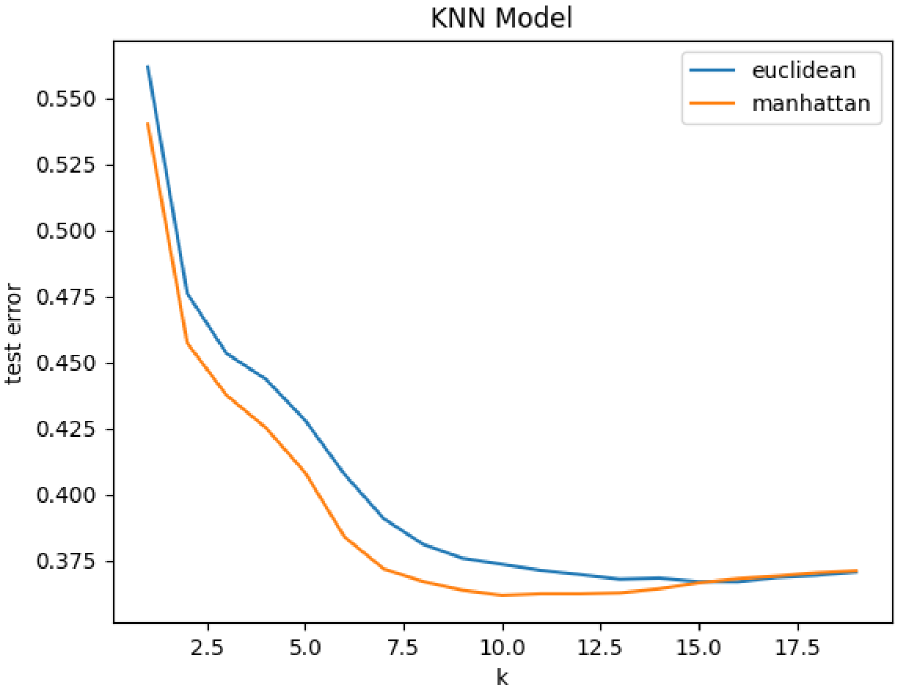
- we used inner join, so we only got the games that were present in both datasets
- we joined, so that we had more variables for each game that we could train our model on
- we then dropped repeating columns and columns we knew we were not going to use to train (i.e. duplicate score columns, duplicate date columns, and duplicate team names)

Training Model(s)

The next step was to train a model. The model that we chose to go with was a KNN Regressor, since we wanted to predict who would win a game given net yards, time of possession, and scores for both home and away teams.

- First we transformed the ids of the teams to become one hot encoded since they are categorical variables.

- Next we made a pipeline with a standard scaler and a KNN Regressor with 5 neighbors.
- To get the best k value, we used a grid search with cross validation.
    - For the scoring, we chose the negative mean squared error, since we wanted to minimize the error.
    - We used 10 folds for the cross validation, but we intend to model folds vs error to see if we can get a better k value.



- After getting the best k value of 10 with the manhattan distance metric as the best metric, we trained the model on the training data.
- We then used the model to predict the outcome of the test data.
- The predictions were float values, so we rounded them to the nearest integer.
    - From the predictions, the model predicted 340 results correctly out of 417 games.

Unfinished Thoughts

There are some pending thoughts that we are still yet to add / are also confused on.

- The current model uses KNN and we are using the model to predict the winner which is technically a categorical variable. So, we will try and use the Nearest Classifier Model and see if that works better.

  - A change we are thinking for the KNN model is to use it to predict the scores for each individual team first and then use those predicted scores variables to predict the winner (we will then compare this with the other models).
- We also are thinking of trying out an ensemble model using our other models and compare the predictions to the other models themselves.
- CONFUSION: We will do more research regarding this, but as of now we are confused on how to predict future games / games that do not have the existing variables that we built the model on. In order to use the model to predict data, we have to have the same variables in the test data that the train data was trained with. However, future games do not have values for the variables we trained on (i.e. time of possession of each team or the net yards for each team)