# Overview

Our Final Project was prepared by Kevin & Nina. The two utilized Microsoft Teams to work collaboratively on questions including searching for and selecting the datasets, writing the code, and producing the following paper.

# Contents

[Overview 1](#_Toc49944275)

[Contents 1](#_Toc49944276)

[Introduction to the data set 2](#_Toc49944277)

[Description of your data exploration and data cleaning steps: 2](#_Toc49944278)

[Questions: 3](#_Toc49944279)

[Should the Seahawks have run the ball or passed? 3](#_Toc49944280)

[On the goal line, should the Seahawks have run or passed? 5](#_Toc49944281)

[Would the direction of the running play have made a difference? 5](#_Toc49944282)

[What about the Seahawks specifically? Would the direction of the running play have made a difference? 6](#_Toc49944283)

[Did the Seahawks have a directional advantage running the ball within the 10-yard line? 7](#_Toc49944284)

[Did the Seahawks pass the ball in the right direction? 8](#_Toc49944285)

[What does all this mean? 9](#_Toc49944286)

[But how long would that have taken? Would they have had to call a timeout immediately or could they run another play? 9](#_Toc49944287)

[Sum it up: 10](#_Toc49944288)

[Summary: What was the general sentiment the day after the Super Bowl in tweets regarding Marshawn Lynch? 11](#_Toc49944289)

[Summary: Were the tweets following the Super Bowl surrounding Marshawn Lynch based primarily on opinion, emotion, judgement, or fact? 11](#_Toc49944290)

[Brief Description of the program: 12](#_Toc49944291)

[ADDITIONAL NFL PLAY ANALYSIS NOT RELATED TO THE SEAHAWKS SUPER BOWL 49 FINAL PLAY: 12](#_Toc49944292)

[What is the field goal percentage likelihood over distance? 12](#_Toc49944293)

[How much did moving back the extra point from a 20-yard attempt in 2015, to a 33-yard attempt after 2015 affect the success rate of the extra point? 13](#_Toc49944294)

[Is there a difference in the success rate of kicks if they were attempted in an indoor or outdoor stadium? 14](#_Toc49944295)

# Introduction to the data set

For our project, we selected a large data set with NFL play by play data. This data set is freely available on Kaggle.[[1]](#footnote-2) For the years 2009 – 2018, every play and its outcome on the game was recorded. This data has 255 columns and 449k rows. For our analysis we will not be considering data from the 2009 season. While large, this data set does not include any data about weather, whether the game was played indoor out outdoor, etc. We appended the data with a list of data frame that we created through extensive research regarding the NFL stadiums that includes information on whether the game was played indoor or outdoor.

In addition, we decided to use twitter data to answer questions related to the famous final play carried out by the Seattle Seahawks in Super Bowl XLIX. In the super bowl game Marshawn Lynch had 24 carries for 102 yards and a touchdown. On Seattle’s final offensive play of the game, they chose to pass the ball at the one-yard line instead of running it with Marshawn Lynch; Russel Wilson’s pass was intercepted which ultimately cost them the game.

We pulled tweets for the day immediately after the Super Bowl to see what the general sentiment was regarding one of the most talked about and controversial plays in NFL history, certainly Super Bowl history. To gather these tweets from 02-02-2015 to 02-03-2015 we imported a package called “GetOldTweets3.” Twitter’s official API has time constraint limitations that make pulling old tweets difficult, so using this package made it possible for us to pull older tweets and dump them into a JSON structured GOT3 object in Python.



# Description of your data exploration and data cleaning steps:

Since the NFL data set was so large, the most efficient way to analyze it in python was to import the csv as a pandas data frame. We imported supplemental data that described each team’s stadium as being either a domed stadium or an outdoor stadium into a pandas data frame.

Exploring our data, we realized that Jacksonville had some rows labeled as “JAX” and other labeled as “JAC”. We selected “JAC” as the abbreviation, replacing all “JAX” in the dataset with “JAC.” We confirmed the task was successful by taking the sum of a df.isin function.

Additionally, we knew that we only wanted to look at data from 2010 and beyond. So, we filtered and subset the entire dataset to show only the data with the dates that we were interested in.

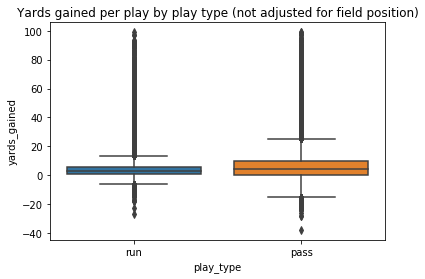
We needed to view a list of the 255 variables in the dataset so that we could choose what variables were going to be required in the four subsets that would be utilized to answer our questions. To do this we created a function to print the column names.

For our sentiment analysis, we only wanted to explore tweets that followed the super bowl by one day. To do that we set the start date at 02-02-2015 and the end date at 02-03-2015. We also wanted to narrow our search by querying only tweets that contained the name “Marshawn Lynch”. Our final criteria were to limit the number of tweets that we pulled to a sample size of 5,000. Once our tweets were pulled, we needed to convert our tweets into a list and into a data frame, from there we were able to process the polarity and subjectivity of the tweets in order to answer our questions regarding the sentiment of the super bowl play.

# Questions:

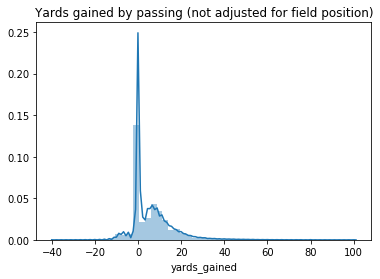
# Should the Seahawks have run the ball or passed?

To answer this question, we ran an analysis on the entire data set consisting of all NFL teams filtering on both pass plays and running plays. The deciding factor would be the comparison of yards gained on pass plays vs running plays. Our analysis resulted in the following Box Plot:

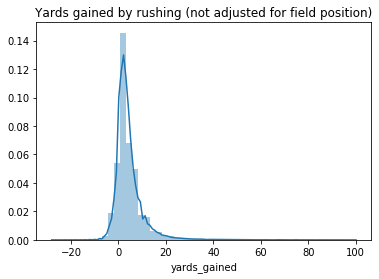


By examining the boxplot, we learned that passing is more volatile than running. Running has a reduced chance of a negative play. This is an old school way of thinking, but Darrell Royal once said, "Three things can happen on any passing play and two of them are bad."

The following histogram shows the passing volatility:



The following histogram represents the running plays:

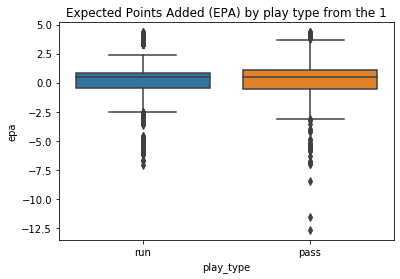


The spike at zero on both histograms represents the number of attempts that failed, meaning that zero yards were gained. In some instances, yards were even lost as the team was pushed back.

We then wondered, what this analysis would look like if it was performed on plays that were only on the goal line. What should a team do there and how is it different?

# On the goal line, should the Seahawks have run or passed?

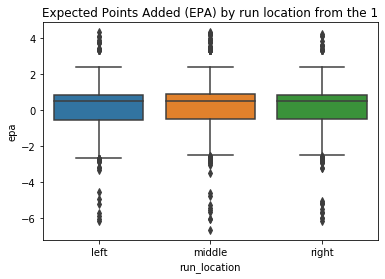
To determine this, we needed to subset our data frame to only plays that were on the one-yard line. We ran the same analysis as before which resulted in the following Box Plot:



Again, it looks like passing the ball is more volatile. Passing, because it requires the quarter back to be behind the line of scrimmage takes on the inherent risk of a negative play in the form of a sack if pass coverage breaks down passing does add more expected points per attempt but only marginally. Given that the Seahawks had at leas two plays to get the touchdown, it seems more likely that a run play followed by a timeout if unsuccessful would have been a better decision here.

# Would the direction of the running play have made a difference?

To answer this question the first step was to subset the data on the one-yard line that were only running plays. The following Box Plot shows the results:



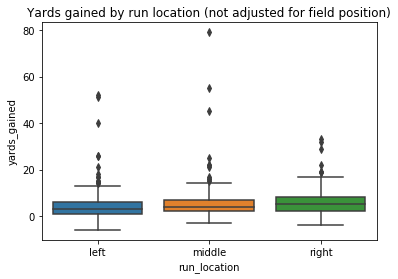
The Box Plot results indicate that there is no globally optimal direction to run the ball from the one.

# What about the Seahawks specifically? Would the direction of the running play have made a difference?

We can look at how the Seahawks ran the ball in 2015 overall but they only had a few attempts from the one all year so small sample size would preclude us from getting anything out of that. Just for fun we decided to see if they were better rushing to the left, right, or middle that year.

To begin we created a new data frame by sub-setting the data for the 2015 season, specifying the Seahawk team only, and filtering for only running plays.

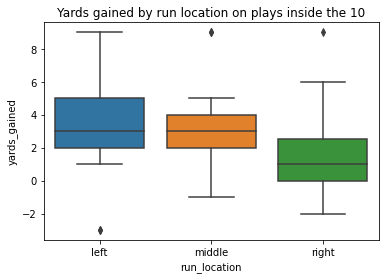
The following Box Plot shows how they did that year:



The Box Plot results show that Seahawks were better at running the ball to the right, only slightly. But now let's see how they did running the ball within the 10 yard line.

# Did the Seahawks have a directional advantage running the ball within the 10-yard line?

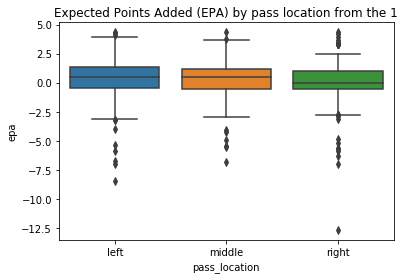
To answer this, we further subset the data to only include running plays within the 10 yard line. The following box plot displays our results:



The boxplot reveals that the Seahawks were significantly better at running left inside the 10-yard line. Given that we know that they passed it, did they at least pass it the correct direction?

# Did the Seahawks pass the ball in the most advantageous direction?

To answer this, we performed the same previous analysis, but subset the data by passing plays only. The following Box Plot shows our results:



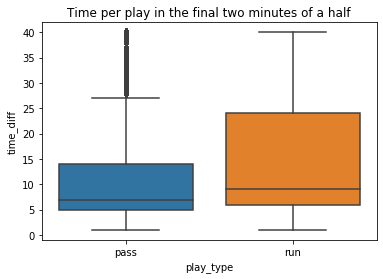
Passing to the left or middle is optimal if you were to pass it from the one-yard line. They threw the ball to the right/middle on a quick slant -- one of the least optimal throws.

# What does all this mean?

Ultimately, the Seahawks should have run the ball from the one-yard line rather than passing the ball.

# But how long would that have taken? Would they have had to call a timeout immediately or could they run another play?

To understand this, we needed to subset the data into a scenario that closely matched the scenario of the Seahawks in the final play of Super Bowl 49. The following Box Plot displays our results:



Our results show us that running the ball does take more time than passing because an incomplete (failed) pass stops the clock immediately. However, the median time for a run play in the final two minutes is only 10 seconds. They could have had the quarterback sneak the ball and had time for two more plays as the average quarterback sneak takes less than an average run to the running back and is lower risk as you don't take the ball backward before handing it off and going forward.

# Sum it up:

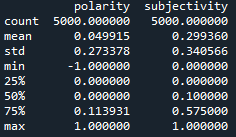
With 66 seconds left in the game the clock was stopped and the Seahawks had the ball on the five-yard line with one timeout remaining. They ran the ball to the left and got down to the one-yard line. The situation was that they were 2nd and goal from the one with 60 seconds left and the 40 second clock (play clock) ticking they ran the clock all the way down to 26 seconds when they snapped it. At this point there were two options: 1. Run, or 2. Pass. The play call was pass and it should have been run (preferably Quarter Back sneak). If they had run it and not gotten the ball in the end zone they then could have called their final timeout leaving them with 15 seconds and 2 downs (at most) to score with 15 seconds and 0 timeouts, there are two options: 1. Pass the ball in or 2. Run the ball likely against their pass defense.

The optimal decision is NOW to pass the ball being that an incomplete pass stops the clock it should have been a fade or other low percentage interception play instead of a slant. However then, if you still have not scored and it is 4th and 1, you give it to the best power back in football and let him work...but Pete Carroll is a clown in 2015 sports illustrated wrote an article referencing how NFL teams were treating analytics:

<https://www.espn.com/espn/feature/story/_/id/12331388/the-great-analytics-rankings>

The Seahawks were referenced as being "one foot in" despite the team’s owner being no other than Paul Allen, the co-founder of Microsoft so, let us see how the fans responded on twitter to this boneheaded call.

## Summary: What was the general sentiment the day after the Super Bowl in tweets regarding Marshawn Lynch?

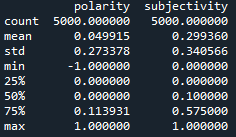


To analyze and determine the general sentiment of the tweets regarding Marshawn Lynch we used a package called “TextBlob.” We were able to pull various statistics to summarize polarity of the tweets. Polarity is a measure that ranges from [-1,1] where 1 means positive statement and -1 means a negative statement.

By examining the Polarity results from our twitter tweet pull we see that our average polarity score was .05. Being so close to zero yet still slightly above zero it seems that our average results would be considered neutral leaning very slightly towards the more positive side. Interestingly the 25 and 50 percent quartiles were both at zero also communicating neutral sentiment. The 75 percent quantile was leaning more towards positive sentiment at .11.

We were rather surprised with the results, from the media and news outlets, it had appeared that there would have been a much higher result in negative sentiment vs positive. The neutral mean could possibly be a bit deceiving being that 50% of tweets were on the extreme side of negative and 50% of tweets were on the extreme side of positive showing that there were equally strong opinions coming from either way. The results however do favor more positive sentiment specifically in the 75-percentile quadrant.

## Summary: Were the tweets following the Super Bowl surrounding Marshawn Lynch based primarily on opinion, emotion, judgement, or fact?



We used the same package “TextBlob” to analyze the subjectivity of the tweets regarding Marshawn Lynch for the 24-hour period following the super bowl in 2015. If the tweet is subjective that means that it is primarily based on or influenced primarily by a person's feelings, tastes, or opinions. If the tweet is objective that means it is a judgement where facts are represented. The scale of subjectivity is from [0,1] where zero is on the extreme end of subjective, 1 is on the extreme level of objective and 5 being a blend of the two or borderline.

Applying this to analysis to the Marshawn Lynch tweets we see that the average subjectivity is .3 meaning that on average the messages were more objective based on facts and judgement vs subjective based on opinions and feelings. The 50th percentile was .1 which was very objective. The 75th percentile then shifted to .58 which interestingly is slightly favoring a subjective message based more on opinion than fact.

Overall, this result aligned with our assumptions that the fans would we very worked up about the facts of the game and how poorly the final play of the Super Bowl was.

# Brief Description of the program:

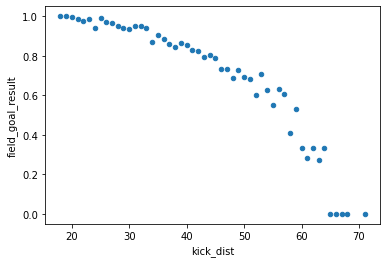
The program we created was specifically designed to answer the 4 questions outlined above. Essentially, it takes a very large dataset of NFL data containing 255 columns of variable data and subsets it to meet the need required to answer the identified questions. The program transforms and cleans the data, appends an original subset of stadium data, filters the data by question, performs pivots and aggregations, and provides several normalization formulas to convert the necessary variables into percentages to provide meaningful comparisons and conclusions. One chart was created to answer question two. This program can be adjusted to answer an endless amount of questions relating to the NFL dataset to provide the user of a very deep understanding of NFL play success and insight.

In addition the program transforms and cleans the twitter data. It can be expanded with additional analysis and coding to answer an endless amount of questions relating to the NFL tweet dataset to provide the user of a very deep understanding of the Marshawn Lynch Super Bowl final play fan sentiment.

ADDITIONAL NFL PLAY ANALYSIS NOT RELATED TO THE SEAHAWKS SUPER BOWL 49 FINAL PLAY:

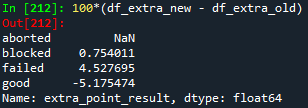
## What is the field goal percentage likelihood over distance?

To answer this question, we created a data subset containing the following fields: play type, field goal result, and kick distance. We were only interested in field goals, so we filtered the play type to show only the plays that were field goals. We replaced made field goals with the value one and any other value besides made with the value zero. This made determining likelihood of made kick an easy calculation. The data frame was pivoted to show mean success over distance.

Not a single attempt under the 20-yard line was outright missed, 0.4% were blocked though. 99.6% of kicks from this distance were made. If you are attempting a field goal under 40-yards, you have greater than a 80% chance of making the attempt! 

## How much did moving back the extra point from a 20-yard attempt in 2015, to a 33-yard attempt after 2015 affect the success rate of the extra point?

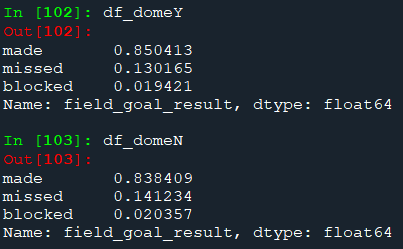
It was a big deal when the rules of the game changed in 2015 regarding the extra point attempts. To see the impact to of the new rule we needed to use the following variables from the NFL data set: game date, play type, and extra point result. First, we needed to see what the probability was of making the extra point under the old rule at the 20-yard line. To do that we filtered the game date for anything under 2015-04-01. From there we normalized the data to see the results. We repeated this step to examine the games that were played under the new rule by filtering for all games after 2015-04-01 and repeated the process of normalization. We took the new data and subtracted the old data to display the variance impact. In the following table we see that the new rules resulted in 5.2% less extra points being made, 4.5% increase in failed attempts and an 0.8% increase in blocked attempts. In conclusion, the new rule has negatively impacted extra points success rate.



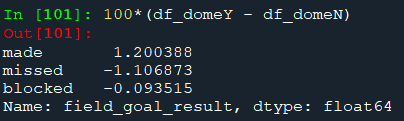
## Is there a difference in the success rate of kicks if they were attempted in an indoor or outdoor stadium?

Our original dataset did not contain information regarding the stadiums in which these games were played. We spent time searching for an appropriate data set that would describe weather or not the stadium was a dome or outdoor with little success. We then decided to research the various team stadiums and create our own dataset mapping the home team to their corresponding stadium, followed by identifying if it was a dome or not. After successfully merging the NFD data set with our original stadium data set, we needed to subset the data to include the following fields: Play type, goal result, and dome. The next step was to filter for only field goals under the play type variable. We then aggregated the data to get a visual interpretation like a pivot table. To obtain the results as a percentage we needed to apply to normalize function to the two separate groups identifying the results of kicks performed in a dome stadium and kicks performed not in a dome stadium. Once those results were identified, we subtracted the yes from the no and the following results were displayed. We learned that there did seem to be an increase in the resulting success of a kick being made in a dome vs not in a dome, but it was very minimal, 1.2%.

Pivot table view of aggregate dome data:



Results: (Dome – No Dome)



Program Output:

import os

import csv

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.feature\_extraction.text import CountVectorizer

import preprocessor as p

import re #regular expression

from textblob import TextBlob

import GetOldTweets3 as got

from nltk.tokenize import TweetTokenizer

pd.set\_option('display.max\_rows', 500)

pd.set\_option('display.max\_columns', 500)

pd.set\_option('display.width', 200)

os.chdir("C:\\Users\\16196\\Documents\\SYRACUSE\\IST 652\\Final\_Project")

df = pd.read\_csv("NFL.csv")

TEAMS = ('ARI', 'ATL', 'BAL', 'BUF', 'CAR', 'CHI', 'CIN', 'CLE', 'DAL', 'DEN', 'DET',

'GB', 'HOU', 'IND', 'JAC', 'KC', 'LAC', 'LAR', 'MIA', 'MIN', 'NE', 'NO',

'NYG', 'NYJ', 'OAK', 'PHI', 'PIT', 'SD', 'SEA', 'SF', 'STL', 'TB', 'TEN', 'WAS')

DOME = ('Yes', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No', 'Yes', 'No', 'Yes',

'Yes', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No', 'Yes', 'No', 'No', 'No', 'No',

'No', 'No', 'No', 'No', 'Yes', 'No', 'No', 'No')

dfStad = pd.DataFrame({'Team': TEAMS, 'Dome': DOME})

df = df.replace(to\_replace = "JAX", value = "JAC")

sum(df.isin(['JAX']).any())

Out[18]: 0

cols = df.columns.tolist()

for col in cols:

print(col)

play\_id

game\_id

home\_team

away\_team

posteam

posteam\_type

defteam

side\_of\_field

yardline\_100

game\_date

quarter\_seconds\_remaining

half\_seconds\_remaining

game\_seconds\_remaining

game\_half

quarter\_end

drive

sp

qtr

down

goal\_to\_go

time

yrdln

ydstogo

ydsnet

desc

play\_type

yards\_gained

shotgun

no\_huddle

qb\_dropback

qb\_kneel

qb\_spike

qb\_scramble

pass\_length

pass\_location

air\_yards

yards\_after\_catch

run\_location

run\_gap

field\_goal\_result

kick\_distance

extra\_point\_result

two\_point\_conv\_result

home\_timeouts\_remaining

away\_timeouts\_remaining

timeout

timeout\_team

td\_team

posteam\_timeouts\_remaining

defteam\_timeouts\_remaining

total\_home\_score

total\_away\_score

posteam\_score

defteam\_score

score\_differential

posteam\_score\_post

defteam\_score\_post

score\_differential\_post

no\_score\_prob

opp\_fg\_prob

opp\_safety\_prob

opp\_td\_prob

fg\_prob

safety\_prob

td\_prob

extra\_point\_prob

two\_point\_conversion\_prob

ep

epa

total\_home\_epa

total\_away\_epa

total\_home\_rush\_epa

total\_away\_rush\_epa

total\_home\_pass\_epa

total\_away\_pass\_epa

air\_epa

yac\_epa

comp\_air\_epa

comp\_yac\_epa

total\_home\_comp\_air\_epa

total\_away\_comp\_air\_epa

total\_home\_comp\_yac\_epa

total\_away\_comp\_yac\_epa

total\_home\_raw\_air\_epa

total\_away\_raw\_air\_epa

total\_home\_raw\_yac\_epa

total\_away\_raw\_yac\_epa

wp

def\_wp

home\_wp

away\_wp

wpa

home\_wp\_post

away\_wp\_post

total\_home\_rush\_wpa

total\_away\_rush\_wpa

total\_home\_pass\_wpa

total\_away\_pass\_wpa

air\_wpa

yac\_wpa

comp\_air\_wpa

comp\_yac\_wpa

total\_home\_comp\_air\_wpa

total\_away\_comp\_air\_wpa

total\_home\_comp\_yac\_wpa

total\_away\_comp\_yac\_wpa

total\_home\_raw\_air\_wpa

total\_away\_raw\_air\_wpa

total\_home\_raw\_yac\_wpa

total\_away\_raw\_yac\_wpa

punt\_blocked

first\_down\_rush

first\_down\_pass

first\_down\_penalty

third\_down\_converted

third\_down\_failed

fourth\_down\_converted

fourth\_down\_failed

incomplete\_pass

interception

punt\_inside\_twenty

punt\_in\_endzone

punt\_out\_of\_bounds

punt\_downed

punt\_fair\_catch

kickoff\_inside\_twenty

kickoff\_in\_endzone

kickoff\_out\_of\_bounds

kickoff\_downed

kickoff\_fair\_catch

fumble\_forced

fumble\_not\_forced

fumble\_out\_of\_bounds

solo\_tackle

safety

penalty

tackled\_for\_loss

fumble\_lost

own\_kickoff\_recovery

own\_kickoff\_recovery\_td

qb\_hit

rush\_attempt

pass\_attempt

sack

touchdown

pass\_touchdown

rush\_touchdown

return\_touchdown

extra\_point\_attempt

two\_point\_attempt

field\_goal\_attempt

kickoff\_attempt

punt\_attempt

fumble

complete\_pass

assist\_tackle

lateral\_reception

lateral\_rush

lateral\_return

lateral\_recovery

passer\_player\_id

passer\_player\_name

receiver\_player\_id

receiver\_player\_name

rusher\_player\_id

rusher\_player\_name

lateral\_receiver\_player\_id

lateral\_receiver\_player\_name

lateral\_rusher\_player\_id

lateral\_rusher\_player\_name

lateral\_sack\_player\_id

lateral\_sack\_player\_name

interception\_player\_id

interception\_player\_name

lateral\_interception\_player\_id

lateral\_interception\_player\_name

punt\_returner\_player\_id

punt\_returner\_player\_name

lateral\_punt\_returner\_player\_id

lateral\_punt\_returner\_player\_name

kickoff\_returner\_player\_name

kickoff\_returner\_player\_id

lateral\_kickoff\_returner\_player\_id

lateral\_kickoff\_returner\_player\_name

punter\_player\_id

punter\_player\_name

kicker\_player\_name

kicker\_player\_id

own\_kickoff\_recovery\_player\_id

own\_kickoff\_recovery\_player\_name

blocked\_player\_id

blocked\_player\_name

tackle\_for\_loss\_1\_player\_id

tackle\_for\_loss\_1\_player\_name

tackle\_for\_loss\_2\_player\_id

tackle\_for\_loss\_2\_player\_name

qb\_hit\_1\_player\_id

qb\_hit\_1\_player\_name

qb\_hit\_2\_player\_id

qb\_hit\_2\_player\_name

forced\_fumble\_player\_1\_team

forced\_fumble\_player\_1\_player\_id

forced\_fumble\_player\_1\_player\_name

forced\_fumble\_player\_2\_team

forced\_fumble\_player\_2\_player\_id

forced\_fumble\_player\_2\_player\_name

solo\_tackle\_1\_team

solo\_tackle\_2\_team

solo\_tackle\_1\_player\_id

solo\_tackle\_2\_player\_id

solo\_tackle\_1\_player\_name

solo\_tackle\_2\_player\_name

assist\_tackle\_1\_player\_id

assist\_tackle\_1\_player\_name

assist\_tackle\_1\_team

assist\_tackle\_2\_player\_id

assist\_tackle\_2\_player\_name

assist\_tackle\_2\_team

assist\_tackle\_3\_player\_id

assist\_tackle\_3\_player\_name

assist\_tackle\_3\_team

assist\_tackle\_4\_player\_id

assist\_tackle\_4\_player\_name

assist\_tackle\_4\_team

pass\_defense\_1\_player\_id

pass\_defense\_1\_player\_name

pass\_defense\_2\_player\_id

pass\_defense\_2\_player\_name

fumbled\_1\_team

fumbled\_1\_player\_id

fumbled\_1\_player\_name

fumbled\_2\_player\_id

fumbled\_2\_player\_name

fumbled\_2\_team

fumble\_recovery\_1\_team

fumble\_recovery\_1\_yards

fumble\_recovery\_1\_player\_id

fumble\_recovery\_1\_player\_name

fumble\_recovery\_2\_team

fumble\_recovery\_2\_yards

fumble\_recovery\_2\_player\_id

fumble\_recovery\_2\_player\_name

return\_team

return\_yards

penalty\_team

penalty\_player\_id

penalty\_player\_name

penalty\_yards

replay\_or\_challenge

replay\_or\_challenge\_result

penalty\_type

defensive\_two\_point\_attempt

defensive\_two\_point\_conv

defensive\_extra\_point\_attempt

defensive\_extra\_point\_conv

df\_toteff = df[['play\_type', 'yards\_gained', 'yardline\_100']]

df\_toteff = df\_toteff[(

(df\_toteff['play\_type']=='pass') |

(df\_toteff['play\_type']=='run')

)]

df\_toteff = df\_toteff.sort\_values(by='play\_type', ascending=False)

sns.boxplot(x = 'play\_type', y = 'yards\_gained', data = df\_toteff).set\_title(

'Yards gained per play by play type (not adjusted for field position)')

plt.show()

df\_toteffpass = df\_toteff[df\_toteff['play\_type']=='pass']

sns.distplot(df\_toteffpass['yards\_gained']).set\_title(

'Yards gained by passing (not adjusted for field position)')

plt.show()

df\_toteffrun = df\_toteff[df\_toteff['play\_type']=='run']

sns.distplot(df\_toteffrun['yards\_gained']).set\_title(

'Yards gained by rushing (not adjusted for field position)')

plt.show()

df\_goal = df[['play\_type', 'run\_location', 'pass\_location', 'yards\_gained', 'yardline\_100', 'epa']]

df\_goal = df\_goal[df\_goal['yardline\_100']==1]

df\_goal = df\_goal[(

(df\_goal['play\_type']=='pass') |

(df\_goal['play\_type']=='run')

)]

sns.boxplot(x = 'play\_type', y = 'epa', data = df\_goal).set\_title(

'Expected Points Added (EPA) by play type from the 1')

plt.show()

df\_goal\_run = df\_goal[df\_goal['play\_type']=='run']

df\_goal\_run = df\_goal\_run[['run\_location', 'yards\_gained', 'epa']]

df\_goal\_run = df\_goal\_run.sort\_values(by='run\_location', ascending=True)

# show the difference in expected points

sns.boxplot(x = 'run\_location', y = 'epa', data = df\_goal\_run).set\_title(

'Expected Points Added (EPA) by run location from the 1')

plt.show()

df\_sea\_run = df[['play\_type', 'run\_location', 'yards\_gained', 'posteam', 'game\_date', 'yardline\_100', 'epa']]

df\_sea\_run = df\_sea\_run[df\_sea\_run['play\_type']=='run']

df\_sea\_run = df\_sea\_run[df\_sea\_run['posteam']=='SEA']

df\_sea\_run = df\_sea\_run[(df\_sea\_run['game\_date'] > '2014-04-01')]

df\_sea\_run = df\_sea\_run[(df\_sea\_run['game\_date'] < '2015-04-01')]

df\_sea\_run = df\_sea\_run.sort\_values(by='run\_location', ascending=True)

# show how they did on the year

sns.boxplot(x = 'run\_location', y = 'yards\_gained', data = df\_sea\_run).set\_title(

'Yards gained by run location (not adjusted for field position)')

plt.show()

df\_sea\_run\_10 = df\_sea\_run[df\_sea\_run['yardline\_100'] <= 10]

# there were only 41 plays so we may wind up with a bad distribution

# show how they did on the year from inside the 10

sns.boxplot(x = 'run\_location', y = 'yards\_gained', data = df\_sea\_run).set\_title(

'Yards gained by run location on plays inside the 10')

plt.show()

# again, they were better inside the 10 going to the right

# given that we know that they passed it, did they at least pass it the correct direction?

# create a df for passing plays from the 1 with their outcome

df\_goal\_pass = df\_goal[df\_goal['play\_type']=='pass']

df\_goal\_pass = df\_goal\_pass[['pass\_location', 'yards\_gained', 'epa']]

df\_goal\_pass = df\_goal\_pass.sort\_values(by='pass\_location', ascending=True)

# show the difference in expected points

sns.boxplot(x = 'pass\_location', y = 'epa', data = df\_goal\_pass).set\_title(

'Expected Points Added (EPA) by pass location from the 1')

plt.show()

dfStad = pd.DataFrame({'Team': TEAMS, 'Dome': DOME})

df\_diff = pd.DataFrame({'half\_seconds\_remaining': df['half\_seconds\_remaining'], 'play\_type': df['play\_type'], 'time\_diff': df['game\_seconds\_remaining'].diff(periods=-1)})

# filter for plays in the final 2 minutes

# this scenario closely matches the seahawks scenario

df\_diff = df\_diff[df\_diff['half\_seconds\_remaining'] <= 120]

df\_diff = df\_diff[df\_diff['time\_diff'] > 0]

df\_diff = df\_diff[df\_diff['time\_diff'] <= 40]

df\_diff = df\_diff[(

(df\_diff['play\_type']=='pass') |

(df\_diff['play\_type']=='run')

)]

sns.boxplot(x = 'play\_type', y = 'time\_diff', data = df\_diff).set\_title(

'Time per play in the final two minutes of a half')

plt.show()

@author: kevin & Nina

"""

Out[114]: '\nCreated on Thu Aug 13 18:07:44 2020\n\n@author: kevin & Nina\n'

import os

from sklearn.feature\_extraction.text import CountVectorizer

import preprocessor as p

import pandas as pd

import re #regular expression

from textblob import TextBlob

import GetOldTweets3 as got

from nltk.tokenize import TweetTokenizer

os. getcwd()

# path\_kevin = ('C:\\Users\\kevin\\Syracuse University\\IST 652 - Scripting - General\\Project stuff')

path\_nina = ('C:\\Users\\16196\\Documents\\SYRACUSE\\IST 652\\Final\_Project')

# os.chdir(path\_kevin)

os.chdir(path\_nina)

text\_query = 'Marshawn Lynch'

since\_date = '2015-02-02'

until\_date = '2015-02-03'

count = 5000

tweetCriteria = got.manager.TweetCriteria().setQuerySearch(text\_query).setSince(since\_date).setUntil(until\_date).setMaxTweets(count)

# Creation of list that contains all tweets

tweets = got.manager.TweetManager.getTweets(tweetCriteria)

tweetlist = []

for tweet in tweets:

temp = tweet.text

tweetlist.append(temp)

# create the df

df = pd.DataFrame(tweetlist, columns = ['tweet'])

# write the tweets to csv before manipulating them

df.to\_csv('tweets.csv', encoding="utf8", index = False)

tweetlist = []

for tweet in tweets:

temp = tweet.text

temp = p.clean(temp)

# temp = clean\_tweets(temp)

blob = TextBlob(temp)

temp = TweetTokenizer().tokenize(temp)

temp = [w.lower() for w in temp]

Sentiment = blob.sentiment

polarity = Sentiment.polarity

subjectivity = Sentiment.subjectivity

tweetlist.append([temp, polarity, subjectivity])

df = pd.DataFrame(tweetlist, columns =['tweet', 'polarity', 'subjectivity'])

# describe the df

df[['polarity', 'subjectivity']].describe()

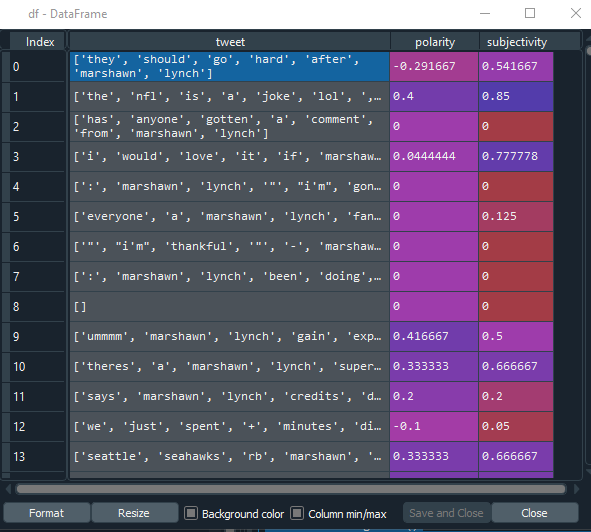
# write the data to csv2

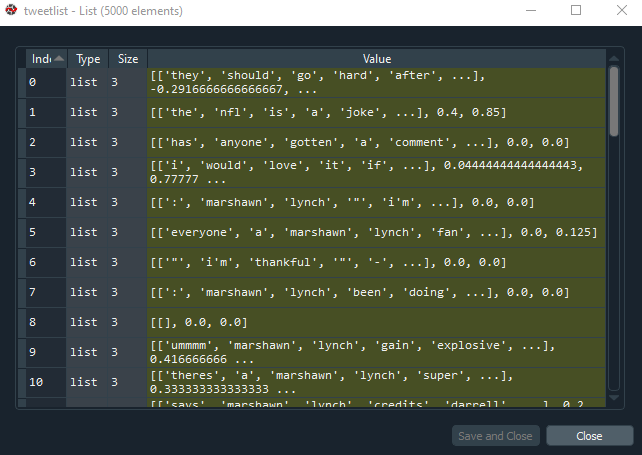
df.to\_csv('tweets2.csv', encoding="utf8", index = False)

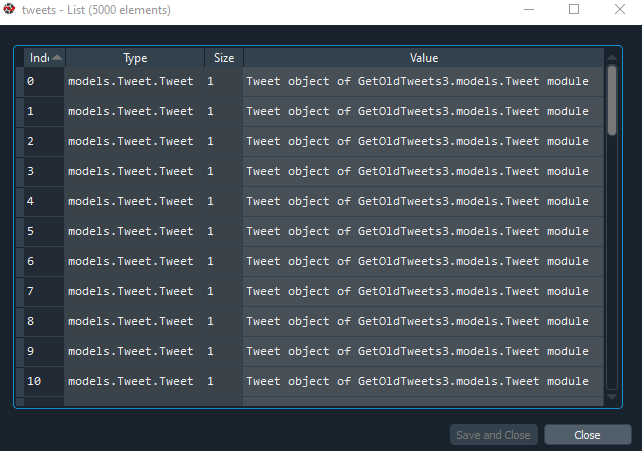
Traceback (most recent call last):

File "<ipython-input-121-a19a9d9f83d0>", line 7, in <module>

df.to\_csv('tweets2.csv', encoding="utf8", index = False)







1. <https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016> [↑](#footnote-ref-2)