Negativity in NFL Fan Bases

The Data Cavaliers
Group Leader: Adina Kugler
Team: John Hope and Adam Kippenhan
DS 4002
September 14, 2022

Restate Hypothesis

NFL fan bases in the Northeast region of the United States exhibit a more negative average sentiment than fan bases elsewhere in the country.

Executive Summary

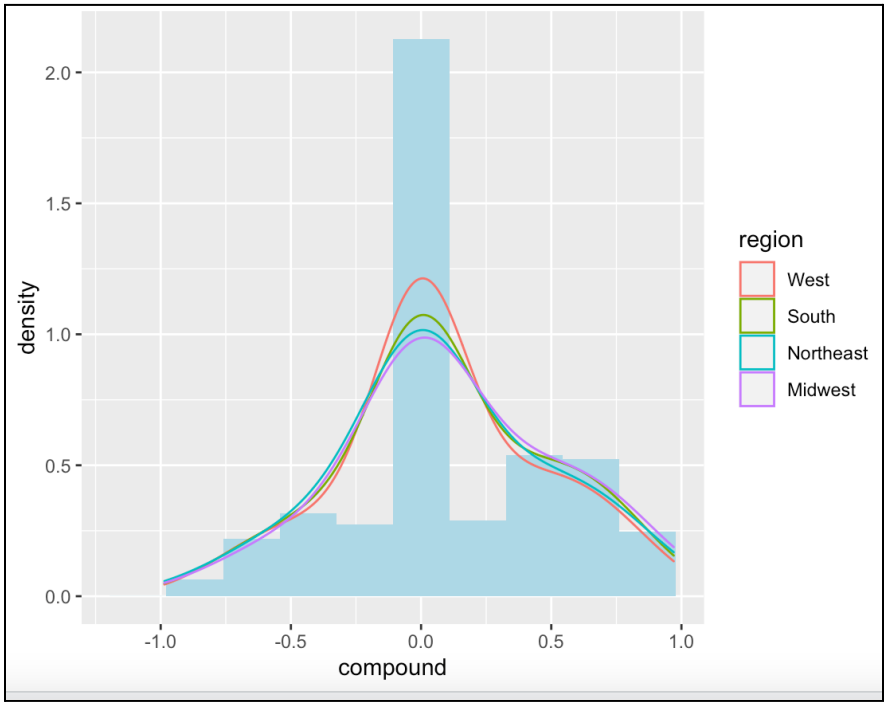
This document contains our data summary and plan for how we will analyze the data. We discuss the finding and establishment of our data with some summary statistics and visualizations and lay out a plan for how we will analyze that data. This plan will dictate how we will move forward in the manipulation and analysis of our data and the basis on which we will build our conclusions.

Data Discovery Findings

The dataset being used for this project was collected through the Twitter API. In Python, the Twitter API was used to pull tweets containing each team's specific hashtag, and then specify which team the hashtag was in support of, and calculate the sentiment scores of each tweet. 1000 tweets were pulled for each teams hashtag, so originally 32,000 tweets were recorded, after removing duplicate tweets (retweets), roughly 13,000 original tweets remained in the dataset. The tweets pulled were the most recent tweets found, so reflect recent sentiments expressed on the app by users. The following GitHub link contains the dataset being used, the Python file used to create it, as well the data dictionary explaining each of the variables: https://github.com/john-hope/DS-4002/tree/main/Project%201

Our hypothesis was considered throughout the data exploration phase. We contemplated dividing the teams based on NFL regions instead of USA regions. This would have changed our question to be the NFC EAST and AFC WEST regions have the most negativity. However, we wanted to stay true to our prediction about sentiment as a whole in the US in terms of regions, rather than the arbitrary lines the NFL drew. Regions were ultimately decided using the official census regions: West, South, Northeast, and Midwest[1].

Our exploratory data analysis consisted of generating different visuals and metrics to get an initial understanding of the difference in sentiments between region's fanbases. Our first question that guided us in the EDA was do the different regions have a difference in distribution of sentiment scores? In other words, could some regions have more polarized sentiments (binomial distribution), or more neutral sentiments (normal distribution)? This led us to create a histogram and accompanying density lines of the compound sentiment scores, seen below:



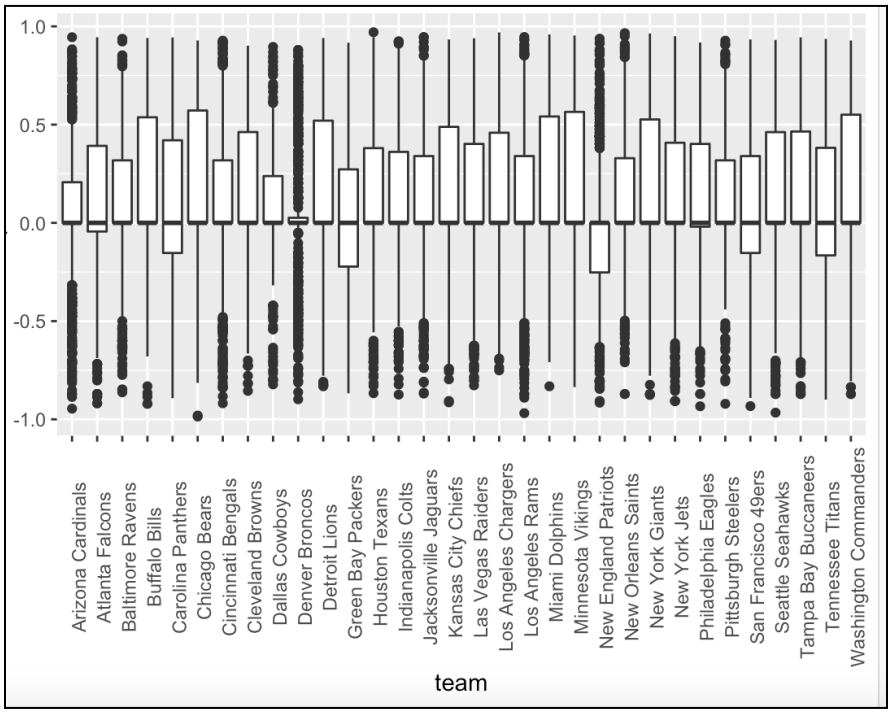
In the graph, the histogram bins represent the overall (country-wide) data, while each of the density lines represent each of the different regions. From the graphic we can see that the distribution of compound sentiment scores is roughly normal for all the regions and overall. Most tweets overall and in each of the regions are in the center, which represents more neutral compound sentiments, with less negative compound and positive compound scores on the outskirts, which represents a normal distribution.

Next, we wanted to have more metrics to be able to compare differences. A question we had was do the regions have different average, min, or max, sentiment scores? The following table below shows the five number summaries for each of the regions:

```
$West
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.96850  0.00000  0.00000  0.08027  0.36120  0.96940

$South
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.91860  0.00000  0.00000  0.09722  0.40190  0.97030

$Northeast
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.93330  0.00000  0.00000  0.09028  0.39870  0.96380

$Midwest
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.9857  0.0000  0.0000  0.1166  0.4404  0.9538
```

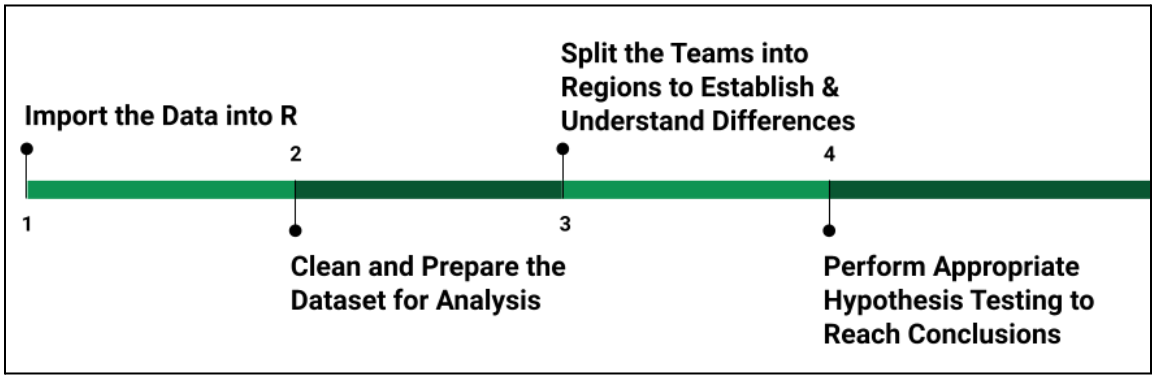
From this table, we can note the slight differences we see between each region. The West region actually has the lowest mean compound sentiment score, indicating least positive sentiments, while the Midwest has the highest mean, indicating most positive sentiments, with the South and Northeast in between.

Lastly, we wanted to get an understanding of how individual teams may be affecting some of the patterns we are seeing in regional differences. The following graph shows the box-and whisker plots of compound sentiment scores for each team:



Throughout the exploratory data analysis, the regionality of the teams was debated. Looking at each difference of medians and distributions in the boxplot above, gives a glimpse of individual teams and how they can differ. 32 levels to compare would increase the error in the hypothesis testing, so it was ultimately decided that regionality must be used for analysis. The above plot is still interesting in possibly predicting the teams that will influence the differences in sentiment the most. For example, the New England Patriots have the lowest distribution, which indicates a more negative fanbase. There is quite a large amount of variation between the individual teams and how the distributions of sentiment scores are represented amongst fanbases.

Analysis Plan



The purpose of our analysis is to ultimately reach our end goal of having a quantifiable measure to our original hypothesis. The process will inform us if there actually is a significant regional difference in the overall sentiments of different NFL fan bases across the U.S. The following steps, which can be seen in the graphic above, will make up our analysis process.

Import Data into R

Our first step to begin the analysis will be to import our data into our analysis tool of choice: the data science-oriented programming language R. We chose to use R for our analysis because of our familiarity with the language as well as its ability to easily create visualizations that will effectively convey our findings. Using our data which is in a CSV file format, we will import the data into a dataframe in R and from there we can use various libraries found in R to create visualizations that coincide with our analysis.

Clean Data

Once we have imported our data into a dataframe in R, our first step before we are able to organize the data and create visualizations will be to clean the data. The principal ways we intend to do this is by removing the text column, checking for outliers and removing NA values. The text column contains the actual tweet corresponding to each row. Since this column will not be needed in our analysis and since the text takes up the most space in the data, we will remove it so

that manipulation of the data will be easier. We will also check for values that could interfere with our analysis by looking for major outliers that could represent errors in the sentiment analysis. Finally, we will remove any NA values since they will not be useful for our analysis.

Divide Data into Regions

After data cleaning, we will separate the data into four main geographic regions of the United States: Northeast, South, West, and Midwest. This will form the basis with which we can begin to compare the regions. The way we will decide which team will fall into each region is based on where they play. One item to note in this step is that there are more teams in the South of the United States compared to the other regions, which may interfere with our data analysis and may need to be compensated for in the future.

Perform Hypothesis Testing

With our data now imported into R, cleaned and separated into geographic regions, we can begin to perform hypothesis testing to evaluate whether our initial hypothesis is true. Our initial choice is to use a statistical t-test to find if there is a statistically significant difference between each geographic region compared to the Northeast. Once we start to analyze the data and begin t-testing, we will decide what type of t-test will work best to give the best results to convey our data. We may also decide to use another type of statistical test such as a pairwise test to either reinforce the results from our t-test or show a different perspective. Different t-tests will allow for analysis for distributions while attempting to eliminate error. These tests can be completed using R and previous code examples.

References

[1] "Geographic Levels," United States Census Bureau. Oct. 8, 2021. [Online], Available: https://census.gov. [Accessed Sept. 14, 2022].

[2] "Choosing the Correct Statistical Test In SAS, STATA, SPSS and R," UCLA Advanced Research Computing. [Online], Available: https://stats.oarc.ucla.edu. [Accessed Sept 14, 2022].