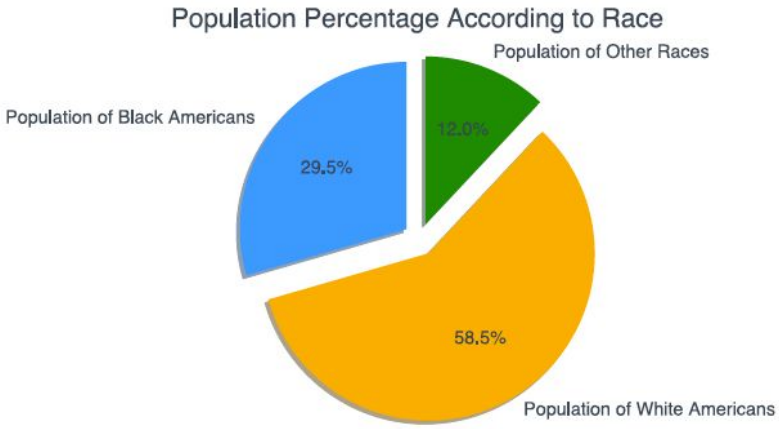
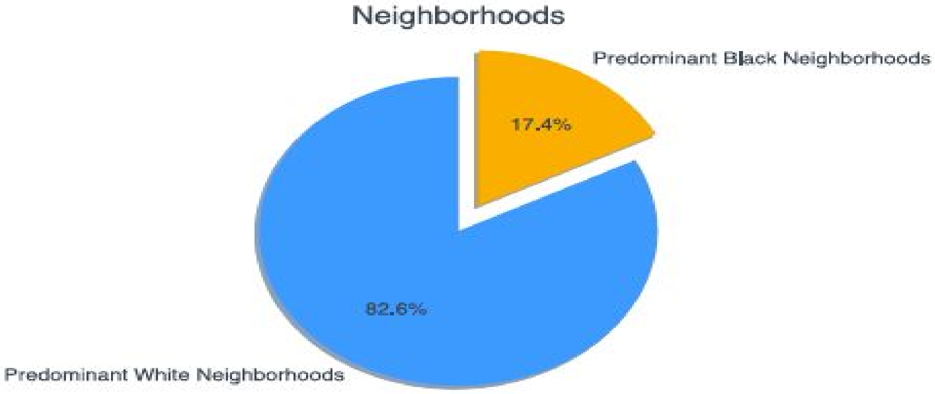
Exploratory Data Analysis:

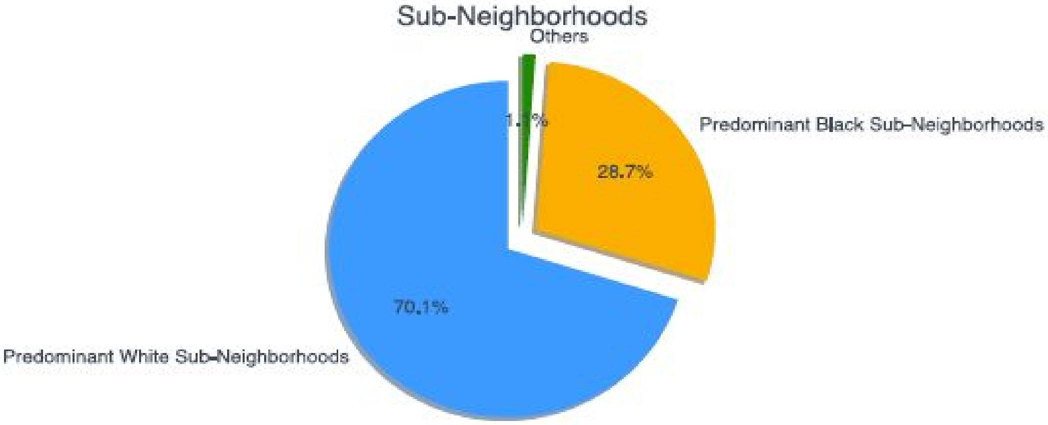
Revised Census Data:

The revised census data provided us with a view of racial segregation across the Boston area. Initially we identified the population of Black Americans, White Americans and Other races. It was determined that White Americans constitute 58.5%, while Black Americans 29.5% of the total population. The remaining 12% constitute of American-Indians and Asians. This in turn gave us an understanding of how we can interpret the data we retrieve later in our exploration.



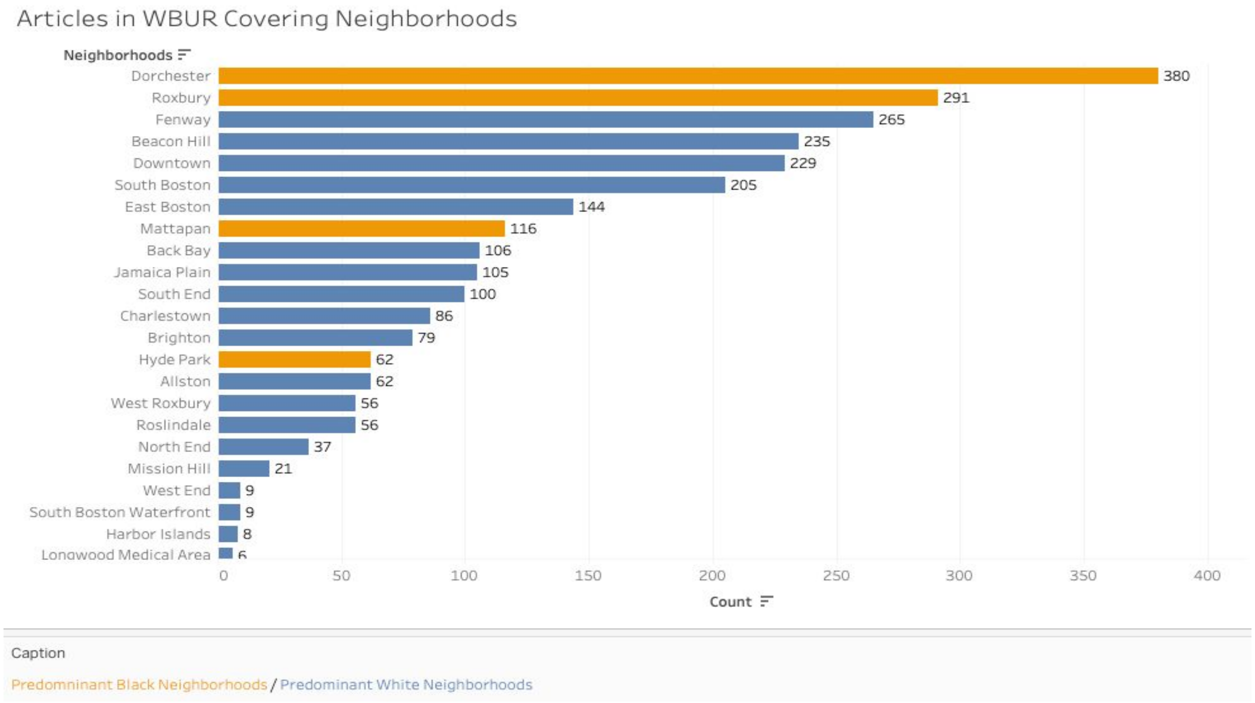
Next, to further understand the segregation. Analysis was done on both neighborhoods and sub-neighborhoods. Out of the total 23 neighborhoods in the area, 19 were found to be predominant White neighborhoods, while the remaining 4 predominant Black neighborhoods. Similarly, out of the 87 sub-neighborhoods, 61 were found to be predominant White sub-neighborhoods, 25 predominant Black sub-neighborhoods, and 1 predominant Asian sub-neighborhood.



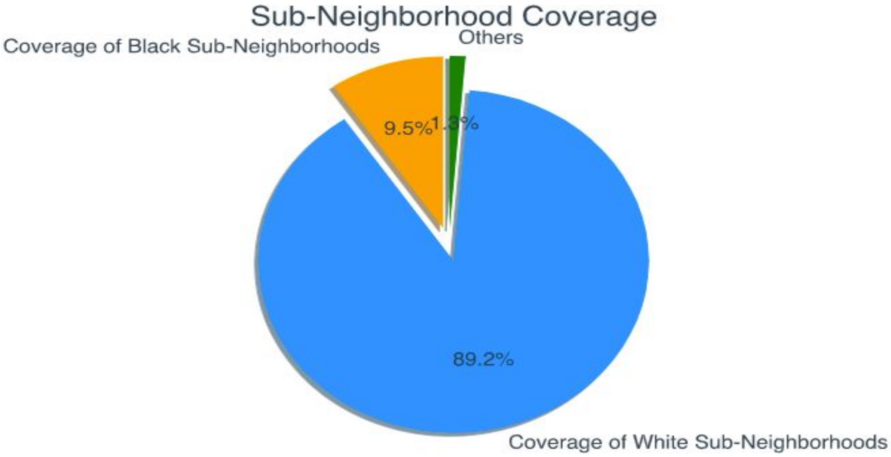


WBUR and Census Data: What is the correlation?

After aggregating the data for all the five years for the WBUR, a preliminary analysis was done to identify the relationship between the census data and the articles. A total of 12,656 articles were identified between the years 2014 – 2018. Between these years, a total of 1818 articles were about white neighborhoods, and they covered all 19 white neighborhoods. The most prominent white neighborhoods covered were Beacon Hill, Downtown, East Boston, Fenway, Jamaica Plain, South Boston and South End. As for the Black predominant neighborhoods, 849 articles were about Black neighborhoods, and the coverage was spread throughout the 4 predominant black neighborhoods (Fig 2). Therefore, predominantly Black neighborhoods were subject to 31.8% of news articles that had a geographic mention. Thus, it was observed, the coverage of predominantly Black Neighborhoods compared to predominantly White Neighborhoods was coherent with percentage population (29.5%)

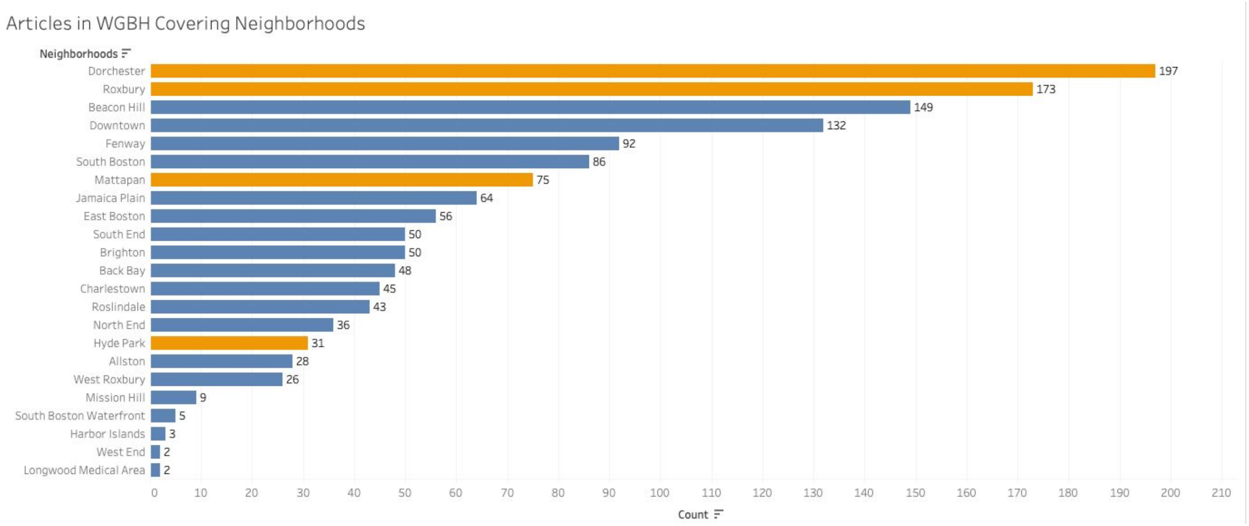


Even though this was the case, we believed that exploring the coverage of sub-neighborhoods will give us an even better understanding. Between the years, there were 8267 articles covering white sub neighborhoods. It was found that out of the 61 predominant white sub neighborhoods 33 sub-neighborhoods were covered. The most prominent white sub-neighborhoods covered were Allston, Boston, Brighton, Charlestown, Columbus, and Roslindale. Similar analysis was done for the black sub-neighborhoods, there were 878 articles covering the sub-neighborhoods, and out of the total 25 predominant Black sub-neighborhoods only 9 were covered. The most prominent black sub-neighborhoods covered were found to be Dorchester, Hyde Park, Mattapan, and Roxbury. Predominantly Black sub-neighborhoods were subject to only 9.5% of news articles that had a geographic mention, as seen in the figure below.

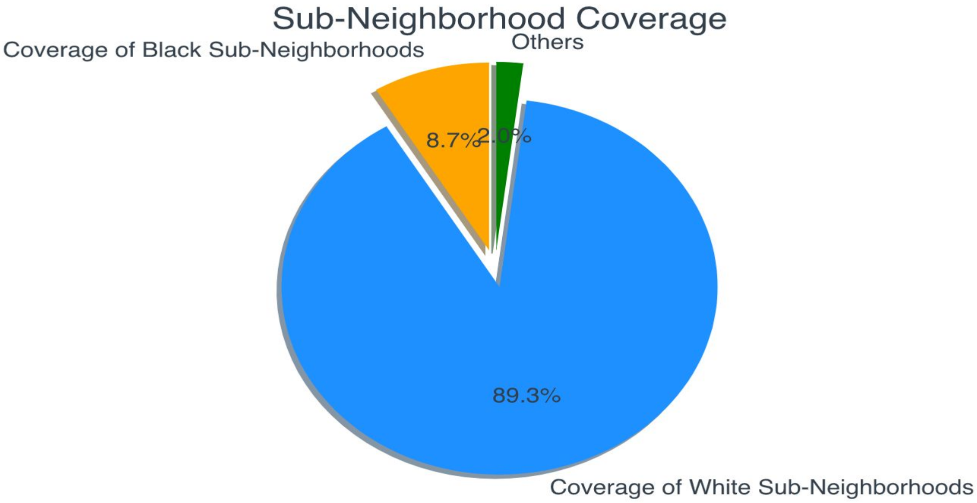


WGBH and Census Data: What is the correlation?

Similar to what was done using the WGBH scraped data set, an analysis was done to identify the relationship between the census data and the articles. A total of 8,324 articles were identified between the years 2014 – 2018. Between these years, a total of 926 articles were about white neighborhoods, and this covered all 19 white neighborhoods. The most prominent white neighborhoods covered were Beacon Hill, Downtown, East Boston, Fenway, Jamaica Plain, and South Boston. As for the Black predominant neighborhoods, 476 articles were about Black neighborhoods, and the coverage was spread throughout the 4 predominant black neighborhoods (Fig 2). Therefore, predominantly Black neighborhoods were subject to 33.6% of news articles that had a geographic mention. The results were very similar to the ones obtained from the analysis done on the WGBH.



The same process was done on the sub-neighborhoods. Between the years, there were 5064 articles covering white sub neighborhoods. It was found that out of the 61 predominant white sub neighborhoods 23 sub-neighborhoods were covered. The most prominent white sub-neighborhoods covered were Allston, Boston, Brighton and Charlestown. Similar analysis was done for the black sub-neighborhoods, there were 492 articles covering the sub-neighborhoods, and out of the total 25 predominant Black sub-neighborhoods only 8 were covered. Predominantly Black sub-neighborhoods were subject to only 8.7% of news articles that had a geographic mention, as seen in the figure below. We concluded even though the amount of articles found on the WBUR is larger than that found on the WGBH during these years, the results were very similar on every aspect. This in turn gave us a solid understanding of our data set so we can further analyze it.

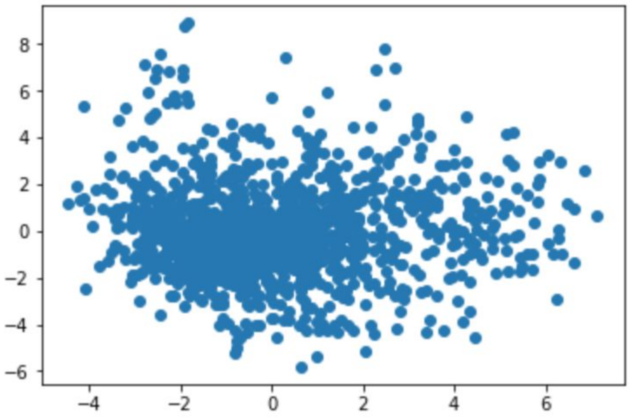


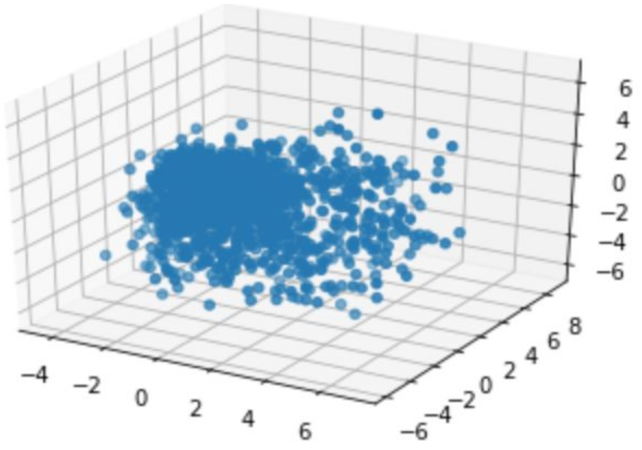
WGBH Web-Scraped Data: Is there a "better" way to represent the data?

At the beginning of the semester our PM's showed us what they had previously done in order to capture clusters of news articles for downstream processing; they used LDA (Latent Dirichlet Allocation) clustering to capture the topics of news articles. The purpose of clustering these articles was to study what topics were being associated with whichever race was being mentioned. The problem here is that the LDA model uses the BOW (bag-of-words) approach to cluster text data which neglects word ordering; this can result in mediocre results at best.

One of the things we sought to improve was the feature representation of the textual data (i.e. news articles) and capture more meaning behind them in order to unlock more meaningful results. We focused on representing the data using document level embeddings via an algorithm called Doc2Vec. This algorithm is an extension of the famous Word2Voc model but instead of learning word embeddings it takes in entire documents and learns document embeddings that represent the underlying meaning of each document.

However, it turned out that these document embeddings were not as helpful in categorizing the articles as we hoped. Below is a 2D and 3D visualization of the document embeddings obtained from one of the web-scraped data sets we collected (WGBH 2017):

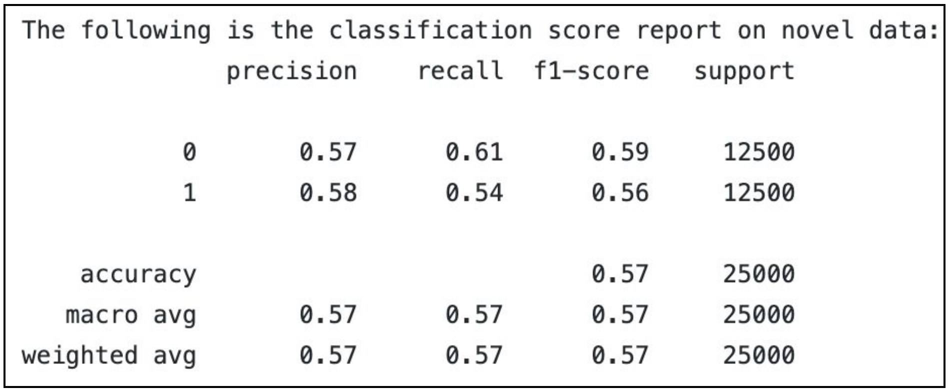




It is evident that the data is forming a cluster around the bottom left of the 2D graph even though there is a large number of points scattered away. This is most probably due to the underlying nature of news articles these days and how they can be convoluted with many topics (e.g. an article about how NBA players kneel during the national anthem can be about both sports and politics). This proved that there is more to be done in regards to better topic classification of news articles.

WGBH Web-Scraped Data: Can we at least capture sentiment?

We then focused back on the traditional Word2Vec embeddings to represent the news articles in order to perform sentiment analysis. By transforming word documents into embeddings and then feeding them into a recurrent neural network we were able to successfully perform sentiment analysis on the data with an average accuracy score of around 57% for unseen data:

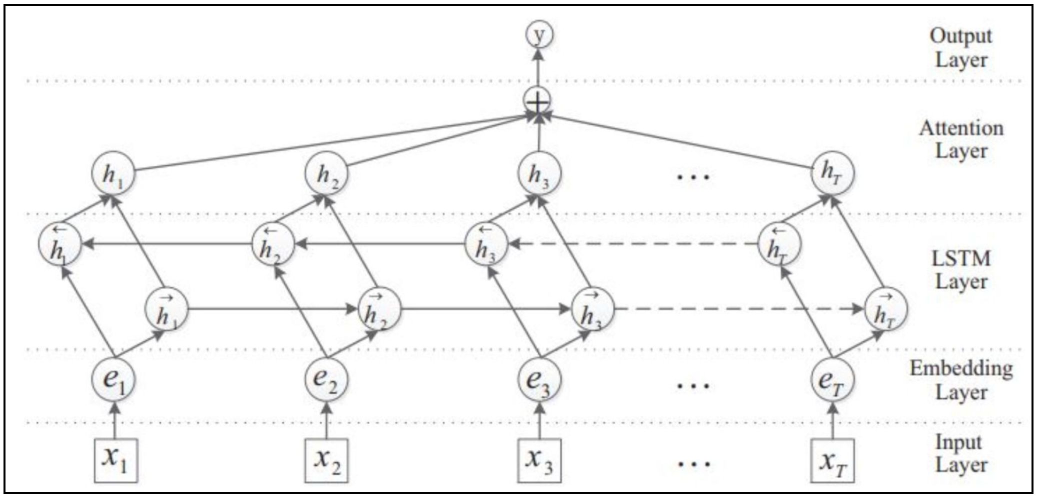


```
The following is the classification score report on novel data:
              precision    recall  f1-score   support

           0       0.57      0.61      0.59     12500
           1       0.58      0.54      0.56     12500

    accuracy                           0.57     25000
   macro avg       0.57      0.57      0.57     25000
weighted avg       0.57      0.57      0.57     25000
```

The specific architecture we used was a bidirectional LSTM (Long Short Term Memory) neural network with the option of specifying an attention layer; the attention layer is nothing more than one-layer neural network appended to the LSTM's outputs to help the neural network better focus on the right words when performing sentiment analysis. The following is a high-level depiction of the neural network architecture:


y
Output Layer
Attention Layer
h_1 h_2 h_3 ... h_T
LSTM Layer
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_T$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_T$
Embedding Layer
e_1 e_2 e_3 ... e_T
Input Layer
x_1 x_2 x_3 ... x_T

WBUR Web-Scraped Data: Is topic modeling better than explicit race mention for sentiment analysis?

Previously, explicit race mention was used to create two datasets: one where the articles are about black people and another where the articles are about white people. Then sentiment analysis can be used on each dataset to determine whether or not the media is negatively reporting on black people. However, we decided that instead of using explicit race mention to construct these two datasets, we would use topic modeling. We developed a Doc2Vec that takes as input an array of words and returns a similarity score along with the indices of articles that embody the input array of words. We used words like “black”, “african american”, and “haitian” to create a topic that encompasses articles discussing black people. We did the same for white people as well. The similarity score is a number between 0 and 1 so the closer the score is to 1, the better the article embodies the topics. We decided on a threshold by manually experimenting with values in the range of 0.1 to 0.9 and then randomly sampling ten articles to check to see if they were about the given topic. We found that 250 was the optimal number of articles. We had four WBUR datasets, so we ended up with 1250 articles about black people and 1250 articles about white people.

Since the WBUR data was not labeled, we tried unsupervised methods for sentiment analysis. Specifically, we used VADER and TextBlob polarity. VADER uses a list of lexical features, like words, that are already labeled as positive or negative in order to determine the sentiment of the entire text. TextBlob uses polarity scores, which take into account the semantic relations and frequency of each word in a sentence. VADER returns three scores, the positive, negative, and neutral score. Each range from 0 to 1 where values close to 1 mean the articles exhibit a particular sentiment. So a negative score of 0.87 would be highly negative while a positive score of 0.87 would be highly positive. For TextBlob, articles with scores greater than 0 are considered positive while scores less than 0 are negative. The results can be seen in the tables below.

VADER

Score Type	Black Articles	White Articles
Positive	0.0567	0.0591
Negative	0.0835	0.0822
Neutral	0.8598	0.8586

TextBlob

Score Type	Black Articles	White Articles
Positive	0.1168	0.1164
Negative	-0.0663	-0.0622

The results from VADER indicate that the sentiment for white and black articles is roughly the same. Also, the coverage for both black and white articles is very neutral. The results from TextBlob also show that the sentiment for black and white articles is very similar. However, the TextBlob results show a slightly positive sentiment for both black and white articles. The negative score for both black and white articles is very close to zero so they are not highly negative. These results match the results that were produced last year when running these algorithms on articles with explicit race mention.

WBUR Web-Scraped Data: What topics are being covered in predominantly white versus predominantly black neighborhoods?

In order to look at what topics were being covered in different neighborhoods of Boston, we used a simple model to look for mentions of both the neighborhood in question, and various keywords for several different topics. We looked at crime, culture, education, politics, race, and sports as topics, and attributed 30 unique keywords to search for in each article to be considered a "mention". We used one of the original LDA models to influence some of the keywords that we attributed to these topics. The histogram below shows the distribution of mentions across these topics. Note that some articles can have multiple topics attributed to them in the model that we had created.

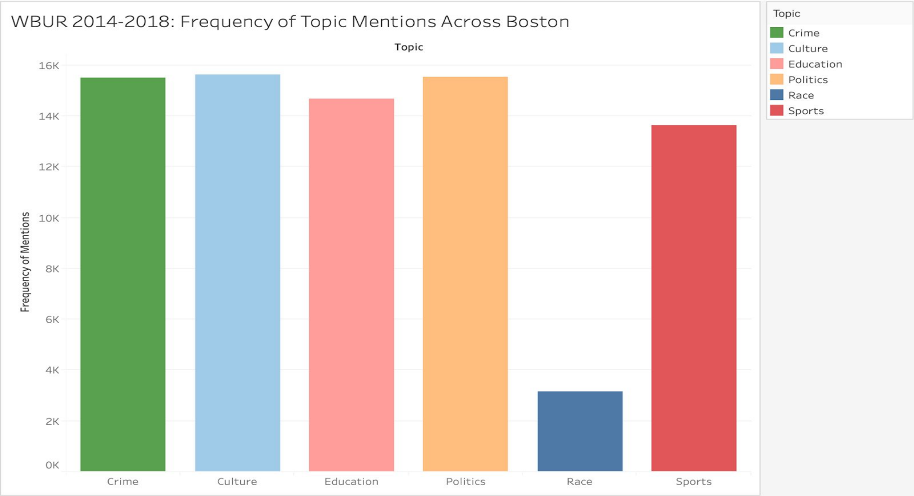
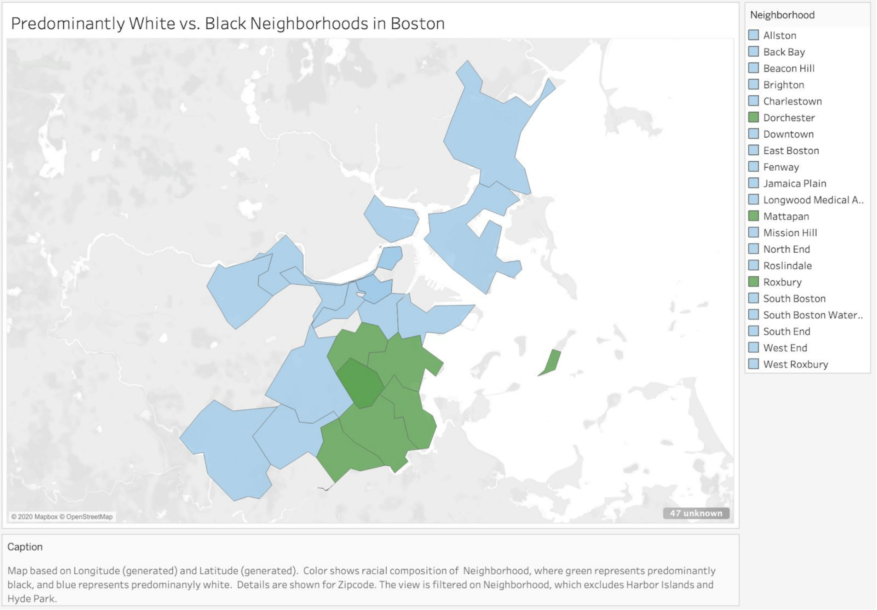


Figure 7: Frequency of Topic Mentions in WBUR 2014-2018

We then used the census data and looked at black proportion and white proportion of neighborhoods to determine which neighborhoods were “predominantly black” versus “predominantly white”, based on the proportion being greater than .50. The following map depicts predominantly black and white neighborhoods in Boston, with blue representing predominantly white, while green represents predominantly black neighborhoods.



Caption

Map based on Longitude (generated) and Latitude (generated). Color shows racial composition of Neighborhood, where green represents predominantly black, and blue represents predominanyly white. Details are shown for Zipcode. The view is filtered on Neighborhood, which excludes Harbor Islands and Hyde Park.

Figure 8: Predominantly White vs. Black Neighborhoods in Boston by Population Proportion

The top 3 predominantly black neighborhoods are Dorchester, Mattapan, and Roxbury, while 3 of the largest, most predominantly white neighborhoods are Charleston, the North End, and South Boston. We filtered out these neighborhoods in particular to look at the topic mention breakdown more specifically to see if there was, in fact a difference in the topics being covered and the frequency among them. The next figure shows the breakdown for Charleston, the North End, and South Boston, followed by the breakdown for Dorchester, Mattapan, and Roxbury.

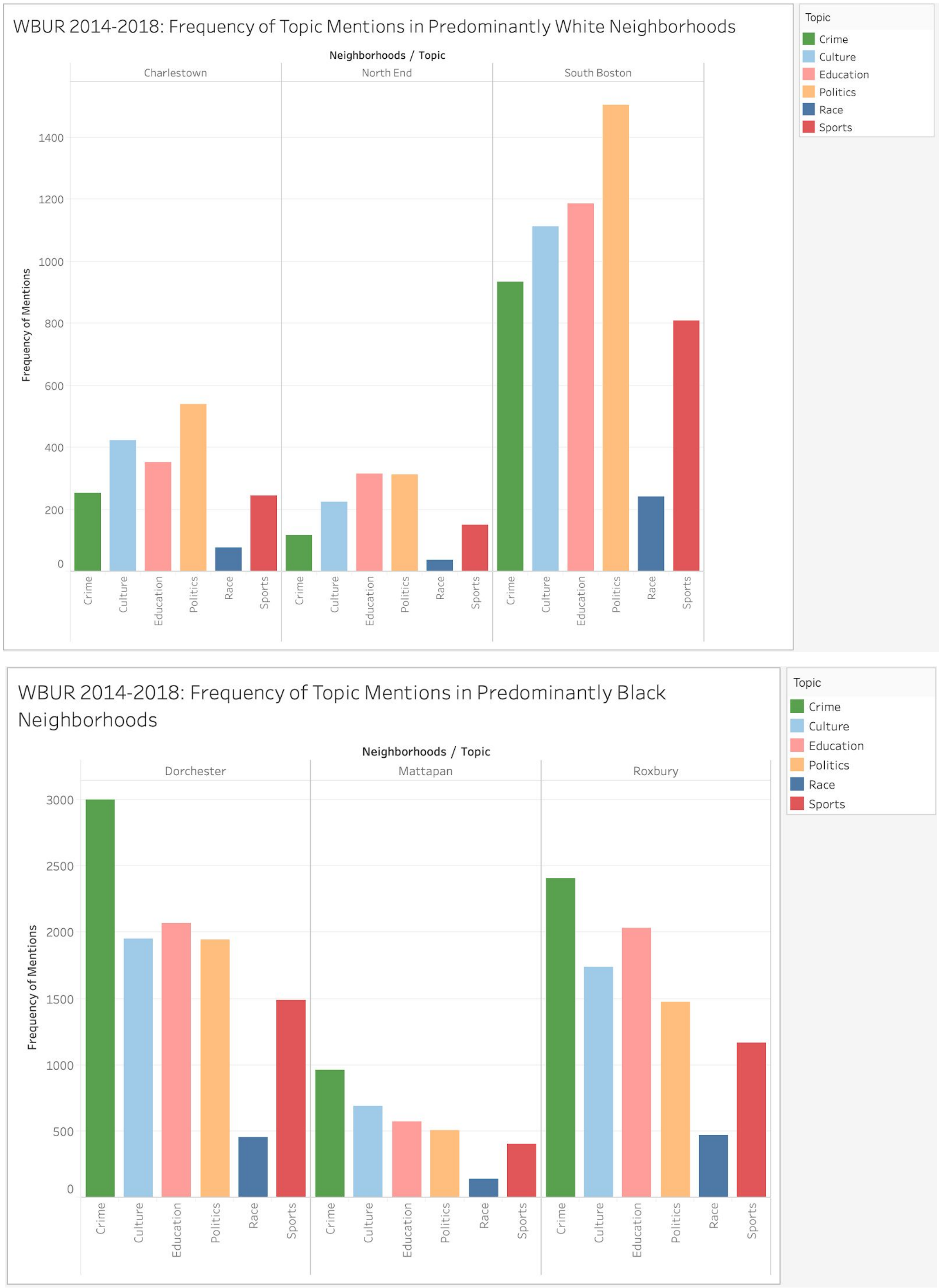


Figure 9: Topic Mentions in Predominantly White vs. Black Neighborhoods

These histograms showed us that crime was being covered significantly more in black neighborhoods relative to other topics when compared to white neighborhoods. A next step would be to look into the actual crime rate in these neighborhoods, as there is actual data for that, relative to the frequency of articles and mentions of crime in these areas. This could be used to indicate if there is actually a bias present in WBUR's reporting.

Given these topics and the racial distribution across Boston, we layered the frequency of mentions of these topics over a map of Boston showing the proportion of blacks living in each area. The following figure gives more insight into what topics are covered in each neighborhood of Boston for further analysis. The size of each point represents the frequency of mentions of that respective topic. Some of the major trends that this figure highlights is that sports are covered heavily in the Fenway area, politics are covered heavily in the financial district, and crime is covered heavily in the Dorchester/Roxbury areas.

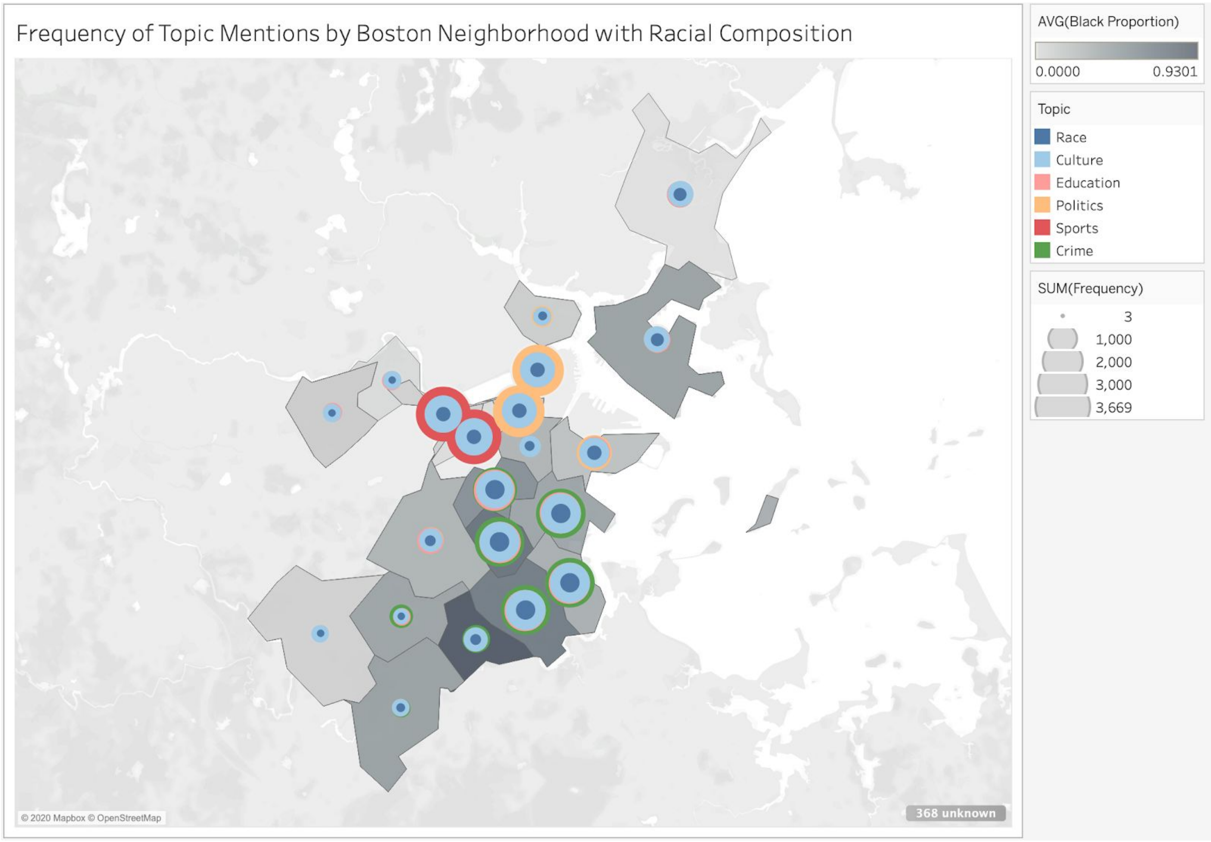


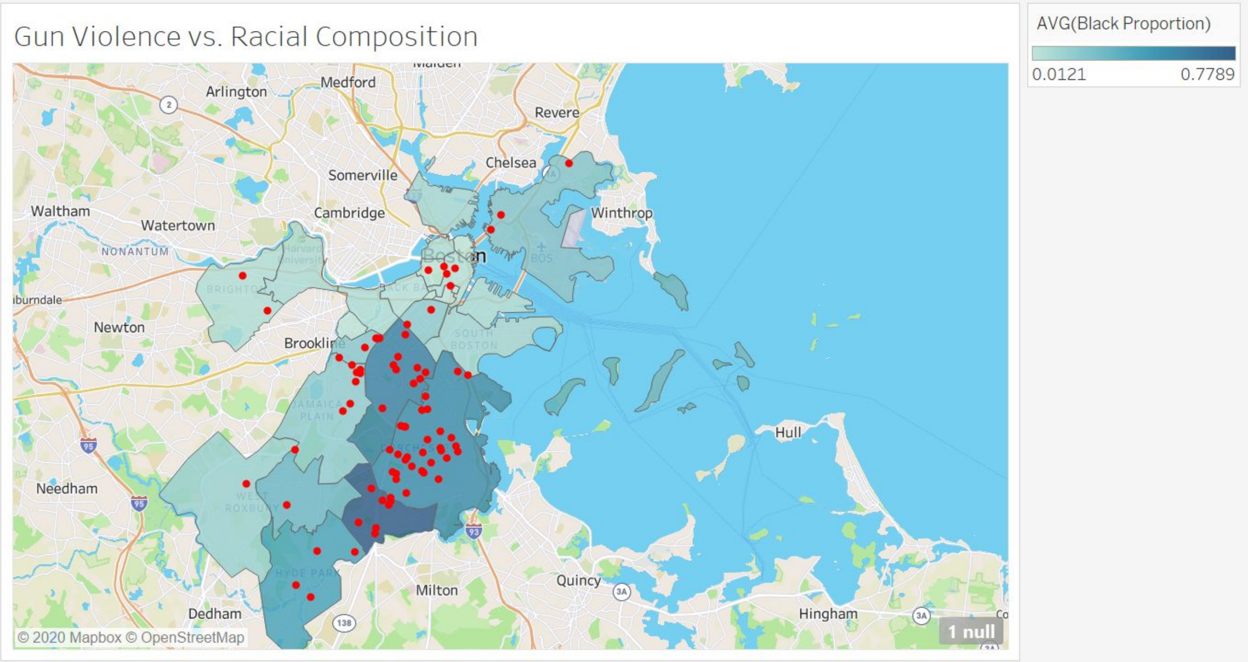
Figure 10: Frequency of Topic Mentions by Boston Neighborhood with Racial Composition

WBUR Web-Scraped Data: Is coverage of gun violence (including homicides) consistent with higher incidence or are there variations across different neighborhoods?

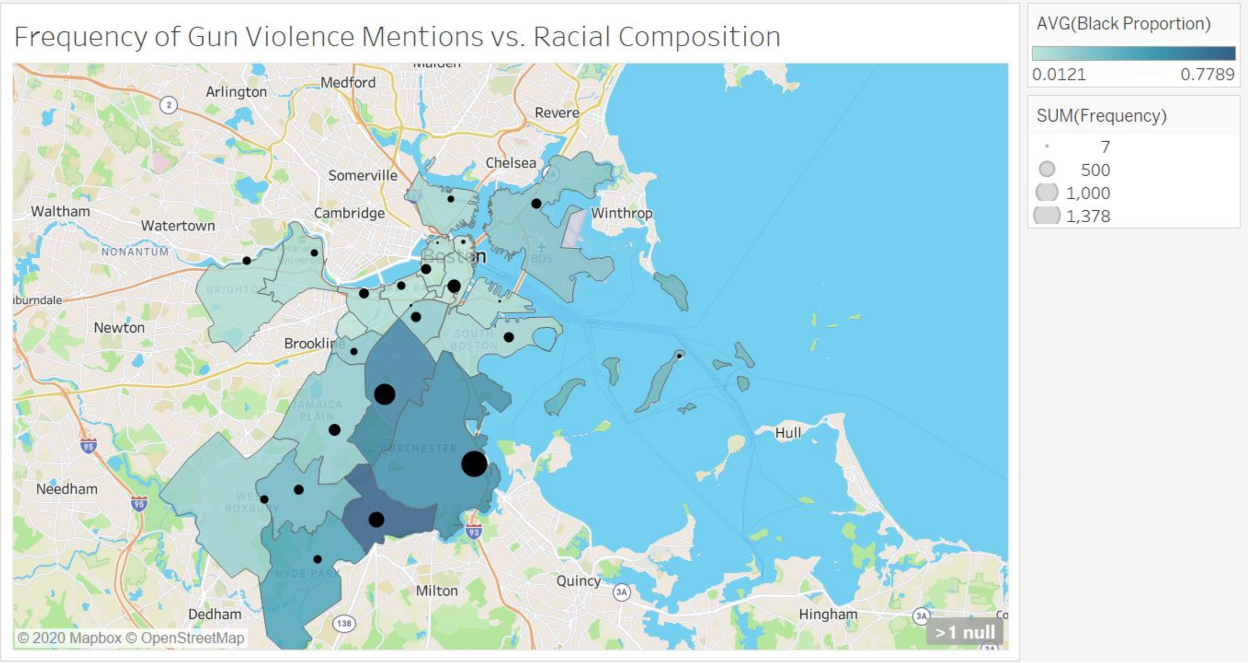
Originally, we were given a list of homicide victims from Boston, with each individual's name, age, race, etc. However, this information on its own lacked an element to connect it to news coverage by neighborhood. Thus, using the open database known as the *Gun Violence Archive* (https://www.gunviolencearchive.org/), we were able to create a query to extract a CSV file containing all incidents of gun violence in Boston from 2014 to 2018. Notably, each incident also had an address included (amongst other new information) so now we had a geographical component to work with. Yet this address on its own would still not be able to yield a specific point on a data visualization, so we created an instance of a geocoder in python using the geopy package. This would give us the ability to go through each of the roughly 2000 incidents and pinpoint it to an exact latitude and longitude (and altitude, although that is certainly not relevant for our purposes), which could then be put onto Tableau. One thing to note is that the geocoder requires us to timeout between each entry (that is, a RateLimiter built into geopy, where we must specify how long we delay between each iteration) so that we do not get locked out of access to the server, so going through every single entry requires quite a bit of time. Thus, for the figures shown below, only a portion of the gun violence incidents were geocoded and put into Tableau. However, all of them will be geocoded and visualized for this project.

With the average proportion of the Black residents in each of Boston's neighborhoods previously found by the aforementioned Census data, we would be able to see how incidents of gun violence varied across the neighborhoods. Furthermore, and more importantly, we could then combine this with the output from code that goes through each article, and look for mentions of specific neighborhoods and a new set of key words relating to gun violence (similar to the method used in the previous section). The significance here is that we could now see if greater incidence of gun violence in a particular neighborhood is correlated to greater coverage of it in news articles or if there is some discrepancy, and if this discrepancy can be attributed to the racial compositions of the neighborhoods. The presence of such a discrepancy could be indicative of a bias present (or lack thereof) in the WBUR's reporting as well.

In more technical terms, this visualization begins with the map of Boston first layered with the proportion of Black residents in each neighborhood, then with dots to indicate each incident of gun violence and where it occurred, or a larger mark above each neighborhood to specify how frequently gun violence was mentioned in the WBUR articles (where the size of the dot denotes the frequency of such).



Incidents of Gun Violence (onl y100) Compared to Racial Composition of Boston Neighborhoods

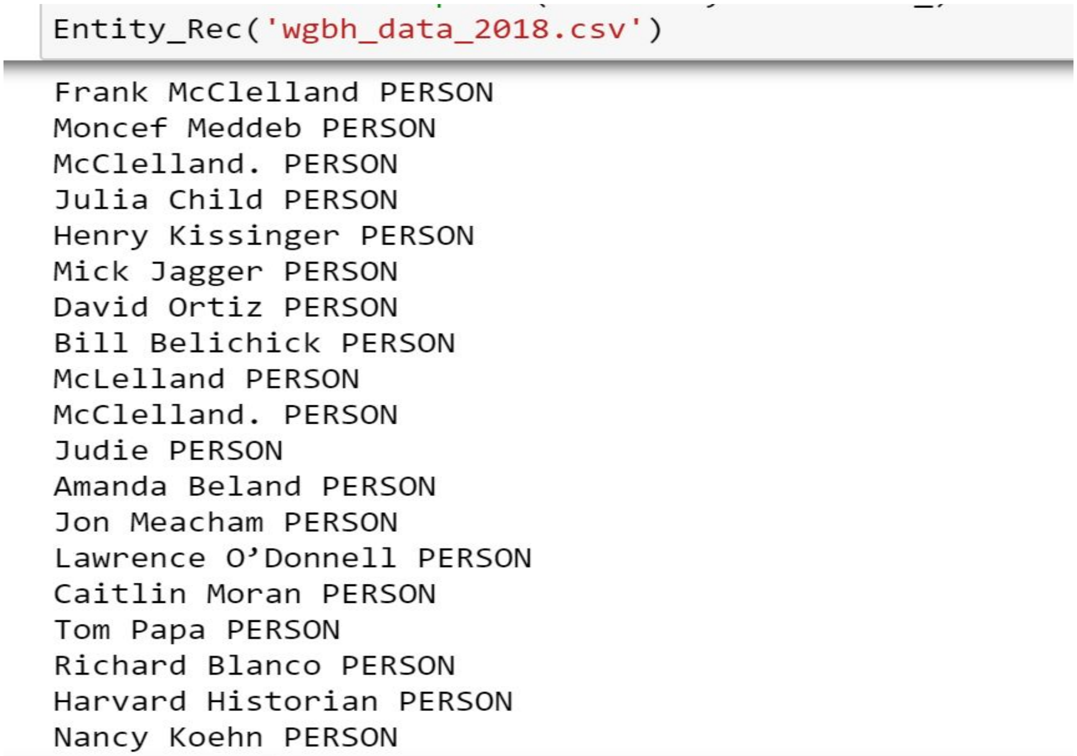


Frequency of Mentions of Gun Violence in WBUR Articles Compared to Racial Composition of Boston Neighborhoods

The above figures represent a sample output using 100 of the approximately 2000 gun violence incidents that have been geocoded. In the top image, each red dot is an incident of gun violence in Boston that occurred between 2014 and 2018. The darkness of the blue hue of each block represents the proportion of Black residents in each neighborhood. Regarding the lower image, each black dot indicates the frequency of gun violence (through a set of keywords related to the topic) being mentioned for articles in the WBUR (2014 - 2018) for each neighborhood. The larger the dot, the more frequently gun violence was mentioned in articles about neighborhoods at that point. At least for this subset of the gun violence incidents, there seems to be no discrepancy between the density of these incidents and frequency of being talked about by the WBUR. However, moving forward, applying this same logic and visualization model to the rest of the gun violence incidents (and hopefully, other news outlets) can provide a basis for evaluating bias in reporting gun violence in areas depending on their racial composition.

Entity Recognition:

We implemented an entity recognition method as the client asked. Entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. Our clients want us to find all entities that in the given Boston dataset articles with the label person's name and company. We used NLP and python to achieve the goal.



```
Entity_Rec('wgbh_data_2018.csv')
```

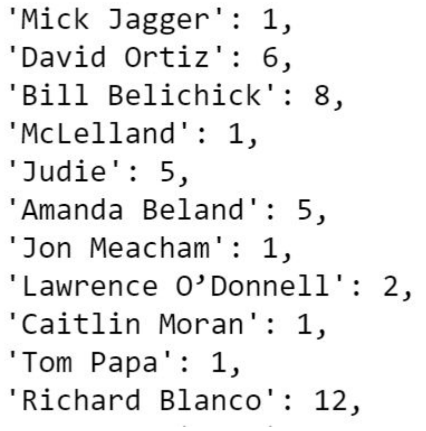
```
Frank McClelland PERSON
Moncef Meddeb PERSON
McClelland. PERSON
Julia Child PERSON
Henry Kissinger PERSON
Mick Jagger PERSON
David Ortiz PERSON
Bill Belichick PERSON
McLelland PERSON
McClelland. PERSON
Judie PERSON
Amanda Beland PERSON
Jon Meacham PERSON
Lawrence O'Donnell PERSON
Caitlin Moran PERSON
Tom Papa PERSON
Richard Blanco PERSON
Harvard Historian PERSON
Nancy Koehn PERSON
```

By using python package Spacy. We are able to identify the entities in the given dataset, showing as the figure below.

This algorithm will successfully identify entities that exist in a given dataset with an accuracy almost nearly 100%. But sometimes it will give an entity not fully correct like include some . or [ ] in the result it returns.

Static Function

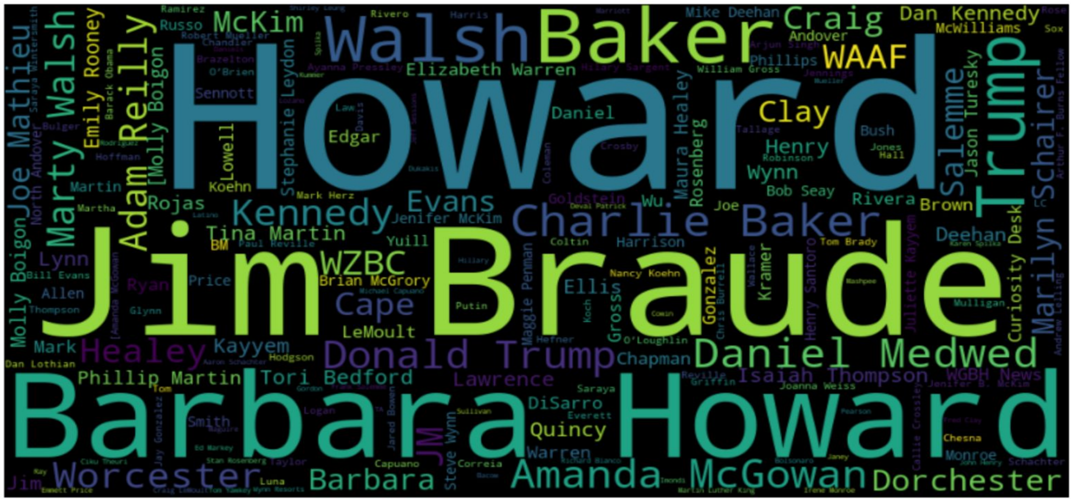
Besides that, we also implemented a static function and a word cloud function to make the data more informative and clear to people with little computer science background. The static function returns a result that tells people how many times an entity appears in the dataset. Also have the ability to rank.



```
'Mick Jagger': 1,
'David Ortiz': 6,
'Bill Belichick': 8,
'McLelland': 1,
'Judie': 5,
'Amanda Beland': 5,
'Jon Meacham': 1,
'Lawrence O’Donnell': 2,
'Caitlin Moran': 1,
'Tom Papa': 1,
'Richard Blanco': 12,
```

Word Cloud Function

The word cloud function is just to make data look more intuitive. The more a word appears in the database, the more frequent the name, the bigger the size is shown in word cloud.


Howard
Jim Braude
Barbara Howard
Walsh
Baker
Craig
Dan Kennedy
McKim
WAAF
Trump
Charlie Baker
Kennedy
Evans
Marty Walsh
Adam Reilly
Joe Mathieu
Emily Rooney
Clay
Henry
Wynn
Salemme
Jason Turesky
Schairer
Marilyn
Maura Healey
Rosenberg
Elizabeth Warren
Edgar
Daniel
Tina Martin
WZBC
Cape
Donald Trump
Daniel Medwed
Healey
Kayyem
Ellis
Harrison
Rivera
Brown
Deehan
Rojas
Lynn
Allen
Price
Brian McGrory
LeMoult
Phillip Martin
Tori Bedford
Isaiah Thompson
Lawrence
Chapman
DiSarro
Quincy
Warren
Smith
Worcester
Barbara
Amanda
McGowan
Dorchester
Molly Boigon
Stephanie Leydon
Lowell
Koehn
Mark
Curiosity Desk
Gonzalez
Kramer