CS 506 Project Draft Analysis
Libby James, Harrison Richmond, Alex Lin

NAACP Media Research project with BU Spark! - Draft

Summary:

The broader goal of the NAACP is "to ensure the political, educational, social, and economic equality and rights of all persons and to eliminate racial hatred and racial discrimination." In this project, we are working to analyze The Boston Herald and their coverage of Black communities in order to assess any differences in what stories are told about Black communities. Hopefully, the results will lead us to possible ways in which The Boston Herald can do its part to encourage elimination of racial hatred and descrimination.

Data Collected:

Our data collection was done through Scrapy utilizing the Wayback Machine. We pulled articles from the local news section of The Boston Herald, aiming to get coverage of the past five years of articles.

So far, we have been able to collect 2,295 applicable articles from The Boston Herald website. This number is far lower than we would have liked. The Boston Herald is an interesting site to navigate, as it is constantly changing. The pulls we have done so far have ranged over three different URLs for starting the pathway of local news coverage, and four different possible xpath responses. While all of these pathways were used, there are still chunks of time within the past five years that must have used a completely separate URL or pathway than any of these, as not every time period has coverage. This was not anticipated by us as many other websites, such as The Boston Globe, have only one URL and pathway to follow to cover the entirety of the last five years.

Of the time periods that fell into these pathways, we were able to pull 8,126 articles in total. However, many of these articles were not able to easily be classified into a neighborhood. Instead, only 2,295 articles were able to be classified. We remained resilient and decided to push on with this data while we try to figure out possible ways to fetch more.

Once we had this data in place, we used python to get word counts of different words that we believed could provide insight into the ways The Boston Herald was reporting upon different Boston communities. The words with negative association included things such as "violence",

"crime", and "dangerous", and those with positive association included things such as "heroes", "charity", and "education".

On a deeper level than just positive and negative associations, these words hold associations as to what the paper is reporting on as well. For example, while the word "victim" does hold a positive association, as that person is seen in a positive light, the words is still usually used when reporting on crime and danger- stories that have to do with current events. On the contrary, words such as "celebration" and "charity" are usually community based stories, discussing a neighborhood hero or a business achieving a goal.

Possible Answers:

Below is a list of questions we provided in our proposal. Below each question is the possible information that our group has at this point that we may be able to provide.

(1) What is the volume of media discussing Black people and communities?
    (a) For The Boston Herald, there is a large quantity of articles discussing Black communities. Out of the four communities that have the highest number of articles, three of them are the top three communities with the highest concentration of Black Community members.

Article Count
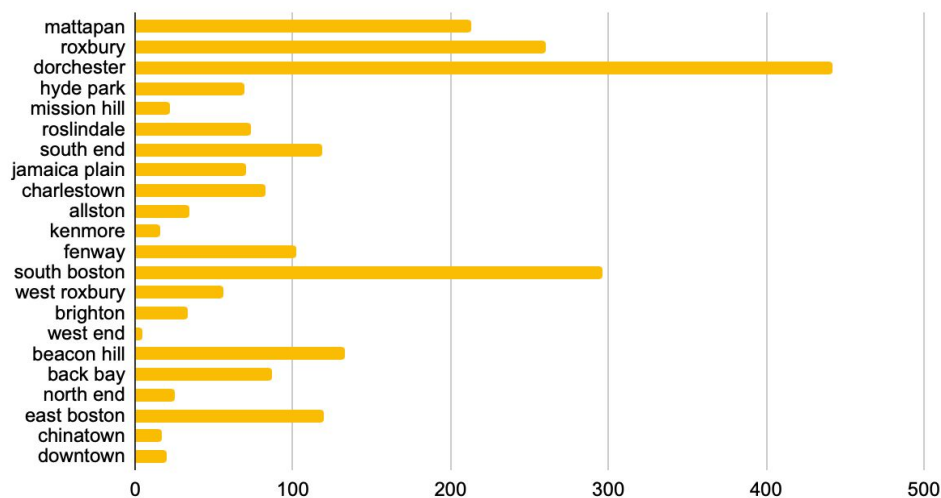
Above is a chart showing how many articles were assigned to each neighborhood

Neighborhoods are listed from top to bottom in order of largest percentage of Black community members to lowest percentage of Black community members.

(2) What topics are being covered in relation to Black people and communities?
  (a) The topics most commonly covered in relation to Black people and communities from the topics we studied are all related to violence: murder, weapons, guns, etc. There were far more articles discussing these topics than there were discussing topics of theater, sports, fund-raising, or the like. The most common articles not relating to violence discussed school and education; this was a common topic for communities that are not predominantly Black as well.
(3) What tone is being used when describing Black people and Black communities?
  (a) As the topics are most commonly related to violence, there is mostly negative connotations when relating to Black communities.

Below are some questions we are unable to answer at this point in time. Below each question here, there is an explanation as to why we do not currently have a response.
(1) Which outlets are covering Black people and communities?
  (a) We are unable to provide too much information on this topic as we were only able to look into this one resource, however we could possibly provide further information on this by discussing with the adjacent group.
(2) How are outlets distributing their coverage of Black people and communities?
  (a) This is a talking point we merely did not have the time to get around to. We explored simply through The Boston Herald website, and we did not delve into their social media presence, physical presence, or otherwise to see how they are getting this information to the community.
(3) What community members are getting covered?
  (a) We were unable to pinpoint specific names in the articles with the time constraints and roadblocks we have encountered.
(4) When are Black communities being covered?
  (a) We did not have time to look at spikes and drops throughout the years, especially as we are missing small chunks of data as The Boston Herald website has changed so often.

Additional Data Processing:

With our remaining time this semester, we would love to take a deeper dive into all of this data. For one thing, we currently only have the raw numbers, counting different occurrences of the different words. If we could use tf-idf to show how these counts correspond to frequency percentages, it may make it easier to see why these raw numbers are meaningful.

If we had much more time to go about this, we would love to also explore how these word frequencies actually correspond to happenings in these communities. For example, are there actually more charity drives in communities that are not primarily Black, and are there actually more murders in communities that are primarily Black, or is the data just skewed in terms of what is being reported to the audience?

Additionally, if we decide there are more data points of words we would like to get counts of, we could do further searches for those as well.

Visualizations:

Most Prominent Discrepancies in the Frequency of Word Counts:

Area of Boston:	Included Zipcodes	Black Population by Zipcode	Article Count:	Murder	Shot	Gun	Killed	Weapon
all studied	all studied	25%	2295	559	533	374	411	128
mattapan	02126, 02124	86%, 64%	213	32	76	125	78	13
roxbury	02119, 02121	58%, 69%	260	58	115	68	54	20
dorchester	02122, 02125, 02121, 02124	31%, 33%, 69%, 64%	442	187	138	84	104	34
hyde park	02136	45%	69	36	30	3	13	8
mission hill	02120	25%	22	26	9	3	7	8
roslindale	02131	25%	73	12	60	8	37	5
south end	02118	16%	119	54	31	6	20	8
jamaica plain	02130	13%	70	16	16	12	7	2
charlestown	02129	9%	83	21	7	5	12	3
allston	02163, 02134	9%, 6%	34	2	0	1	1	0
kenmore	02115, 02215	8%, 4%	16	0	0	0	1	0
fenway	01125, 02215	8%, 4%	102	6	6	20	6	0
south boston	02210, 02127	5%, 7%	296	77	19	5	25	14
west roxbury	02132	6%	56	9	3	10	5	5
brighton	02135	5%	33	16	5	7	9	2
west end	02114	5%	5	0	0	0	0	0
beacon hill	02108	5%	133	1	8	13	4	2
back bay	02116, 02199	6%, 0%	87	0	5	3	16	0
north end	02110, 02019, 02113	4%, 4%, 0.7%	25	0	0	0	5	0
east boston	02128	3%	120	3	0	1	7	4
chinatown	02111	2%	17	3	5	0	0	0
downtown	02203	0%	20	0	0	0	0	0

Above is a chart that is an option to include in our poster. While we do have a much larger excel sheet showing all of the words and data that we have pulled, these are the top 5 words in which we saw the biggest discrepancies as to how often they appeared in articles relating to Black communities as opposed to non-predominantly Black communities.
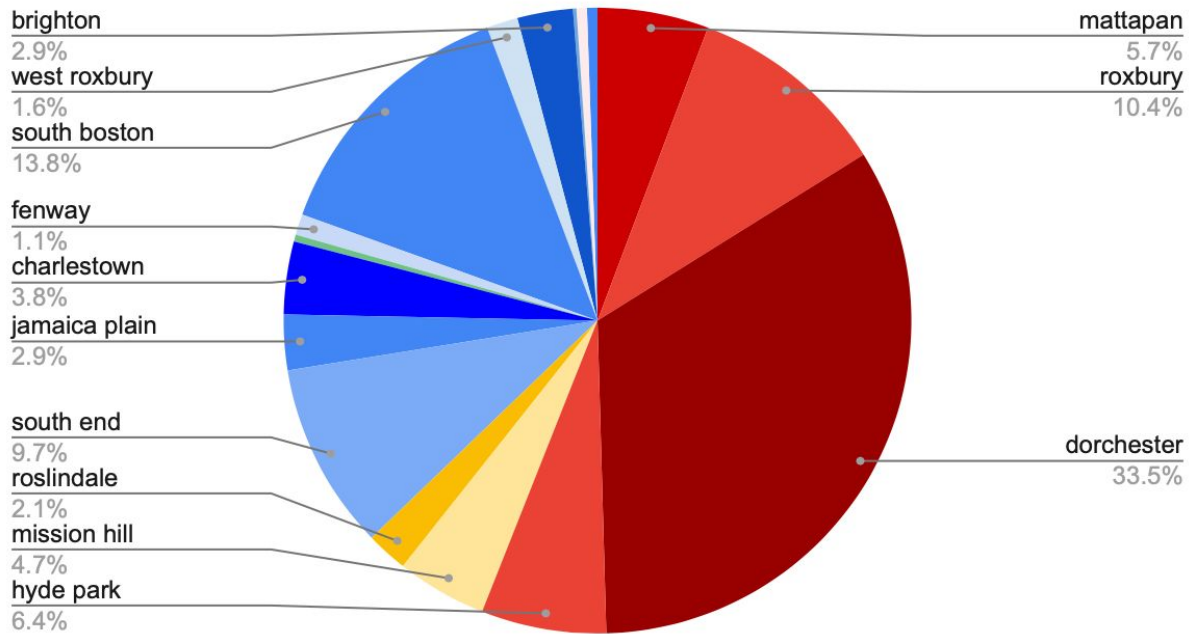
We believe it could also be beneficial to instead represent the data more visually, here in the form of bar graphs and pie charts.

For the pie charts, we have chosen to represent communities with three colors: blue for the communities that have a lower percentage of members of the Black community than Boston as a whole, yellow for the communities that have an equal percentage of members of the Black community than Boston as a whole, and red for the communities that have a greater percentage of members of the Black community than Boston as a whole. As 25% of the population of Boston is composed of members of the Black community, this means blue represent neighborhoods with less than 25% of their community members being a part of the Black community, yellow represents 25%, and red represents more than 25%..
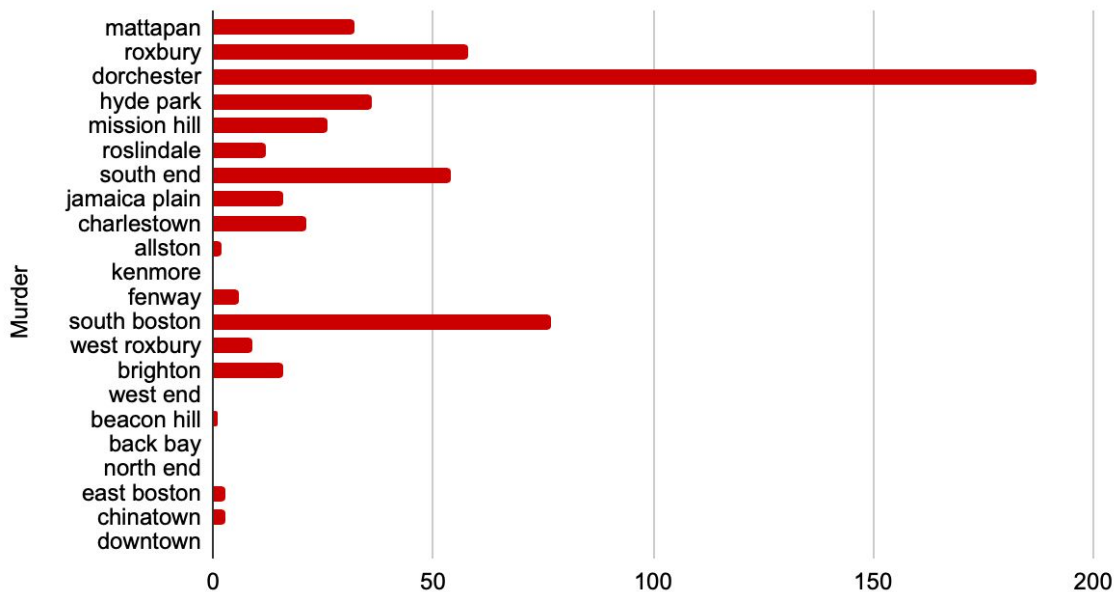
For the bar graphs, we put the communities with the largest percentage of Black community members at the top, flowing all the way down to the communities with the lowest percentage of Black community members at the bottom.
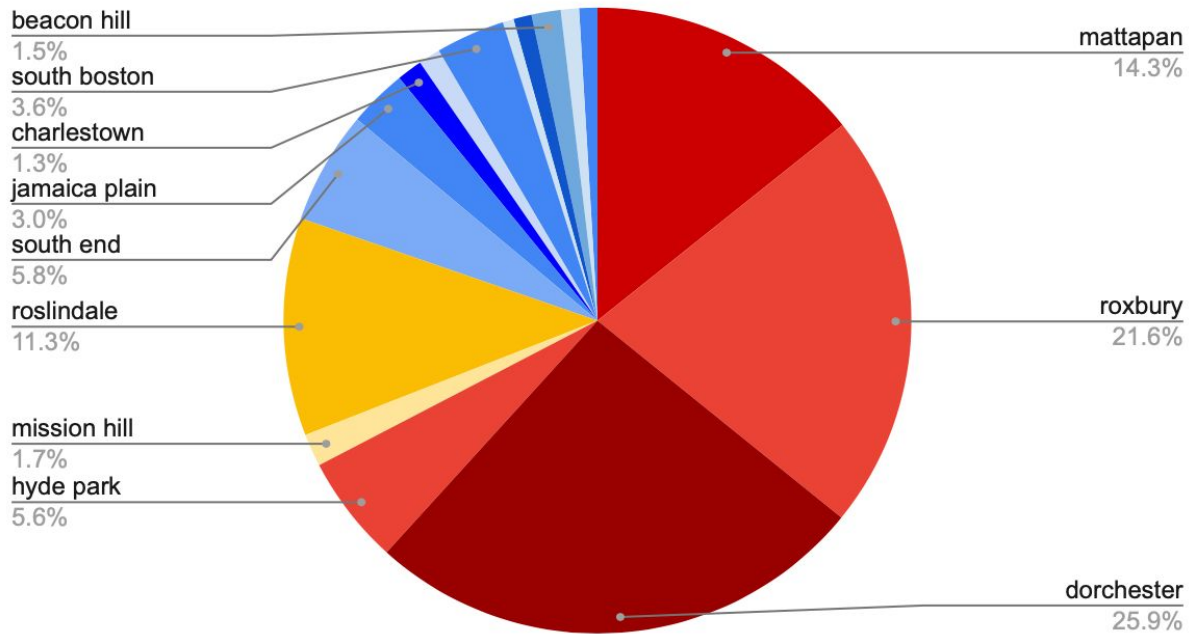
Charts on the Word Count for "Murder":

Murder

Region	Percentage
brighton	2.9%
west roxbury	1.6%
south boston	13.8%
fenway	1.1%
charlestown	3.8%
jamaica plain	2.9%
south end	9.7%
roslindale	2.1%
mission hill	4.7%
hyde park	6.4%
mattapan	5.7%
roxbury	10.4%
dorchester	33.5%

Murder

Charts on the Word Count for "Shot":

Shot

Location	Percentage
beacon hill	1.5%
south boston	3.6%
charlestown	1.3%
jamaica plain	3.0%
south end	5.8%
roslindale	11.3%
mission hill	1.7%
hyde park	5.6%
mattapan	14.3%
roxbury	21.6%
dorchester	25.9%

Shot
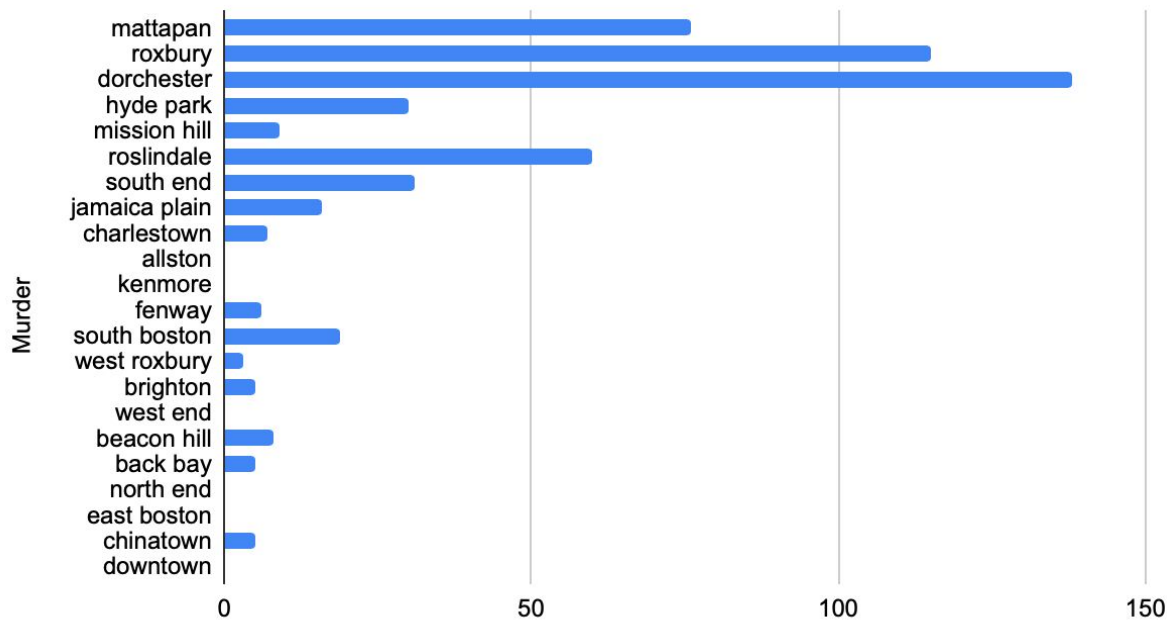
Charts on the Word Count for "Gun":

Gun

Region	Percentage
beacon hill	3.5%
brighton	1.9%
west roxbury	2.7%
fenway	5.3%
jamaica plain	3.2%
south end	1.6%
roslindale	2.1%
dorchester	22.5%
mattapan	33.4%
roxbury	18.2%
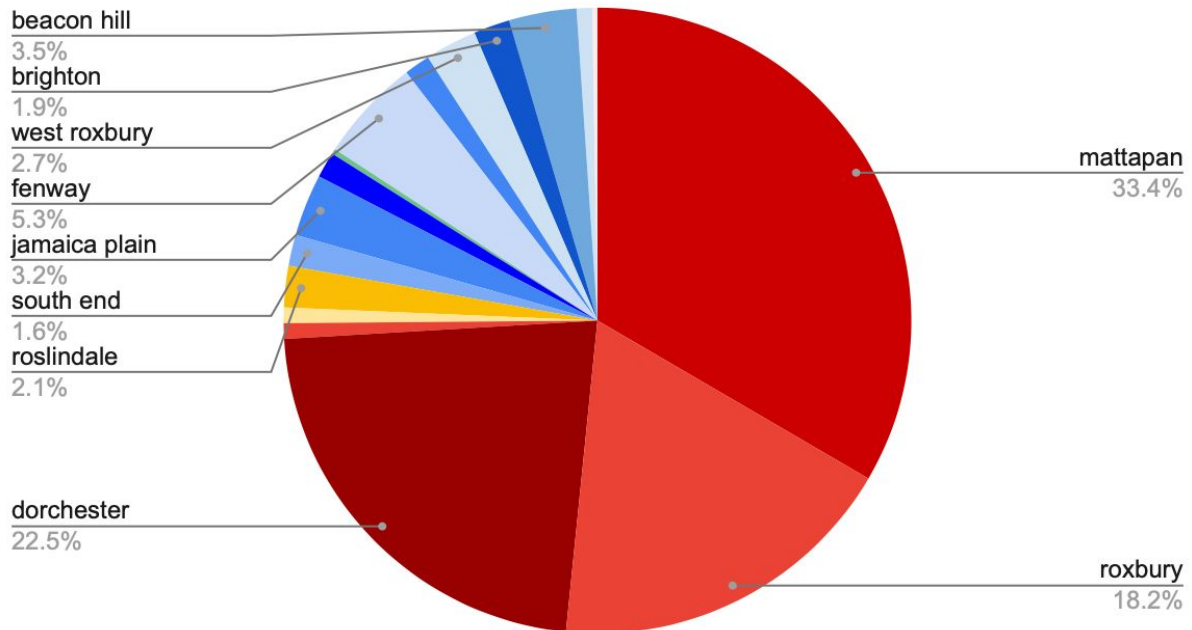
Gun
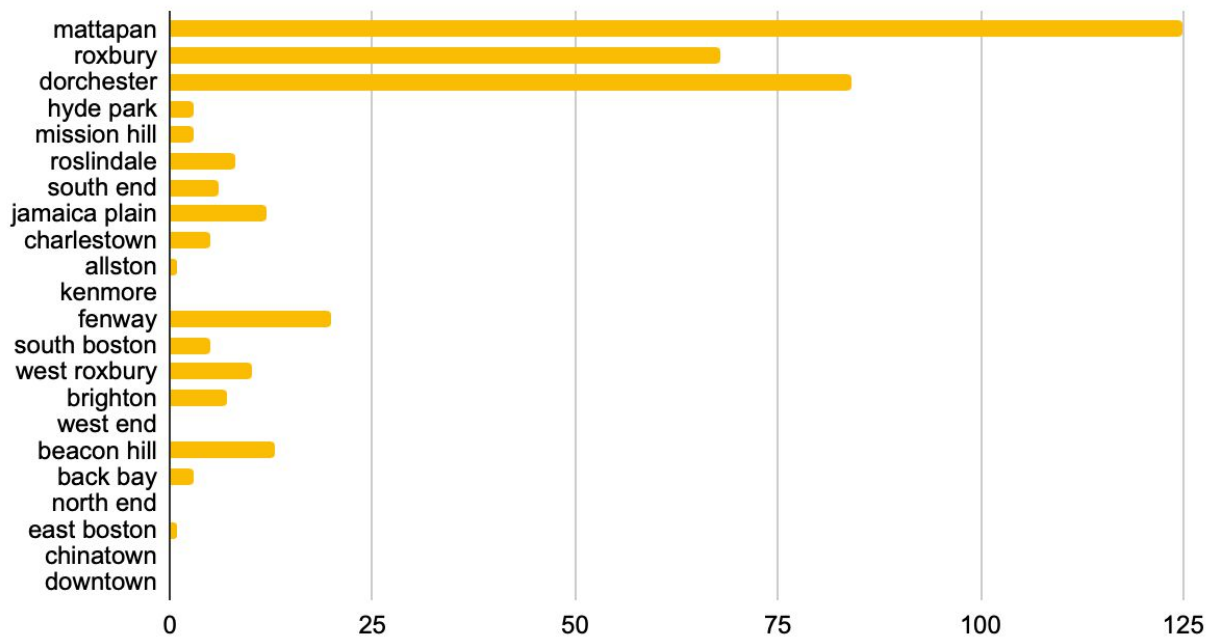
Charts on the Word Count for "Killed"

Killed

east boston
1.7%

back bay
3.9%

brighton
2.2%

south boston
6.1%

fenway
1.5%

charlestown
2.9%

jamaica plain
1.7%

south end
4.9%

roslindale
9.0%

mission hill
1.7%

hyde park
3.2%

mattapan
19.0%

roxbury
13.1%

dorchester
25.3%

Killed

mattapan
roxbury
dorchester
hyde park
mission hill
roslindale
south end
jamaica plain
charlestown
allston
kenmore
fenway
south boston
west roxbury
brighton
west end
beacon hill
back bay
north end
east boston
chinatown
downtown

0   25   50   75   100   125

Charts on the Word Count for "Weapon":

Weapon

east boston	3.1%
west roxbury	3.9%
south boston	10.9%
charlestown	2.3%
south end	6.3%
roslindale	3.9%
mission hill	6.3%
hyde park	6.3%
mattapan	10.2%
roxbury	15.6%
dorchester	26.6%
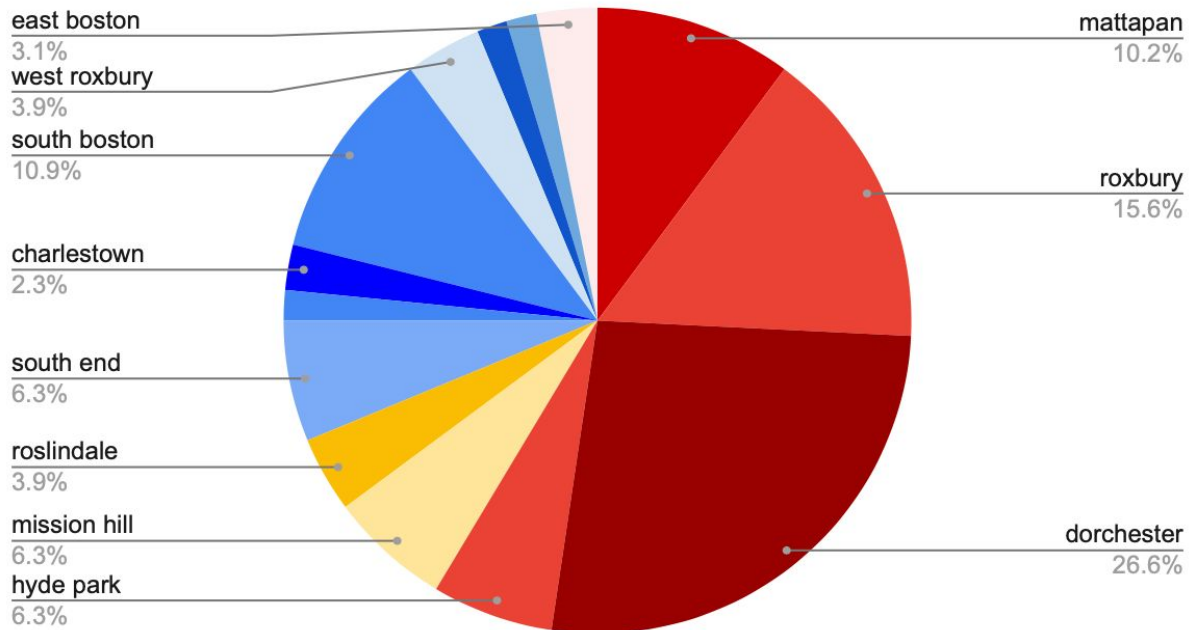
Weapon
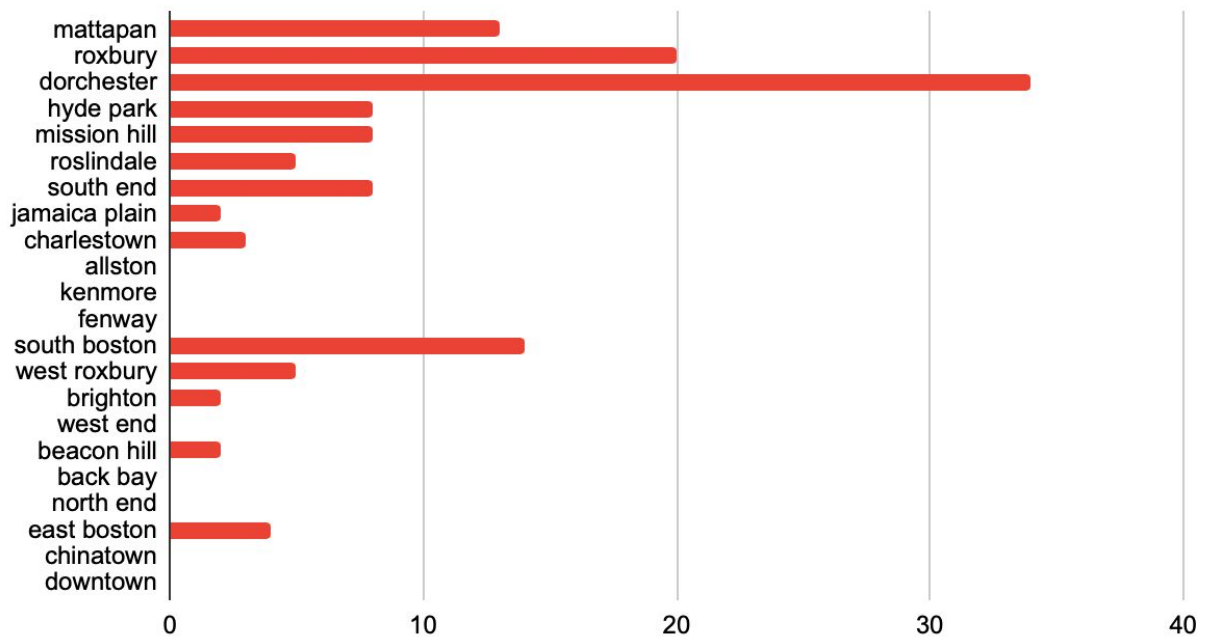
Out of all of these visual tools, we believe the pie charts may be most helpful in getting the point across that the topics of murder and gun violence are covered far more often in Black communities. This is because it shows the percentages, not the count. Notice that there is a specific focus on murder and gun violence as these were the areas of the data where there was the largest difference between predominately Black communities and communities that are not predominately Black.

Proposal Changes:

Our original proposal focused a lot on how the media changed over time, however we have shifted the focus to be more upon how the Boston Herald reports on Black communities differently than communities that are not predominantly Black. Additionally, we had wanted to use more resources than just the Boston Herald, however with how difficult that site has been to navigate, and as that was the main source we were asked to pull from, we have not yet been able to expand our research outside of that domain.

Additional Information:

Additional information about our steps so far, including the code we used to analyze the data, the json files representing the data collections we have pulled, and the spreadsheets we have created as a visualization of some of our data, can be found at https://github.com/etjames/NAACP-Media-Research. Please do not hesitate to reach out to the group leader, Libby James, at etjames@bu.edu, with any questions.