1. INTRODUCTION

Staying active is critical for people to live long, happy, and healthy lives. Regular exercise is proven to reduce an individual's risk of type 2 diabetes, heart disease, stroke, and some cancers.[1] Physical activity has also been shown to boost self-esteem, mood, energy, and sleep quality and reduces an individual's risk of anxiety disorders, depression, dementia, and Alzheimer's. [1] Exercise is also thought to manage symptoms of anxiety and depression just as well as medications in certain individuals who are diagnosed with these conditions.[2] There are countless physical activities that individuals can take up to unlock the benefits of an active lifestyle and there is something for everyone.

Even though the benefits and options of exercise are vast, it is often difficult for many people to stay consistently active in large cities where it is easy to get swept up in the hustle and bustle of everyday life. Individuals who are looking to become active may also not know what areas of their city are best suited for the type of physical activity they want to begin. In this analysis, we are seeking to explore the options individuals have to begin an active lifestyle in the city of Boston and take a look at how similarly each neighborhood in Boston likes to stay fit using k-means clustering. We will gain insight on which areas of Boston are best for different kinds of physical activity.

The stakeholders of this project are individuals looking to start living an active lifestyle in the city of Boston. This project could also be of interest to tourists who are looking to stay active during their trip to Boston and are looking for the best neighborhoods to maintain their lifestyle. This project is also useful for businesses looking to open up a specific kind of recreational area in Boston, as they can see where there would be the least amount of competition for them. Finally, this project is useful for already existing recreational venues in Boston, as this analysis may motivate more people to come into their locations.

2. DATA

The following data will be required to complete this project:

- A list of neighborhoods in Boston, Massachusetts.
  - The following Wikipedia article was scraped for this information: List of Neighborhoods in Boston
- Latitude and Longitude coordinates of Boston Neighborhoods.
- Data on venues in each Boston Neighborhood.
  - This data was collected using the Foursquare Places API.

The data required for this project will be collected via web-scraping and by using Foursquare's API. We will use the web-scraping package BeautifulSoup to create a pandas DataFrame of all of the neighborhoods available in Boston. We will then get the Latitude and Longitude values of the neighborhoods using the Nominatim package in python and add those values to our DataFrame. The final data we will be using is location data from Foursquare. We will be querying Foursquare's API for data on what

recreational venues exist within Boston's neighborhoods. This information will be used for determining which recreational venues are most popular in each neighborhood and will allow us to use clustering to see the similarities between neighborhoods in Boston based on their exercise options. Then, we will see how the different neighborhoods in Boston can best stay fit.

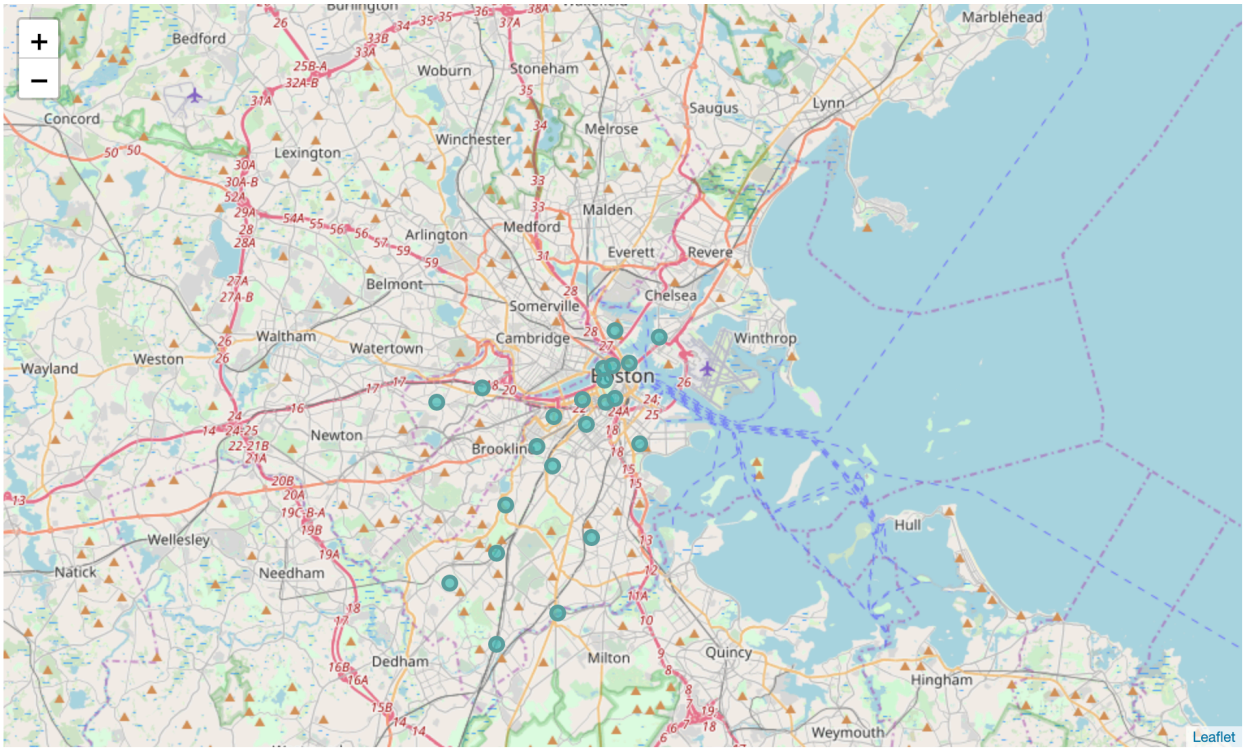
3. METHODS

3.1: Initial Data Scraping and Clean Up

First, we imported all necessary packages and began the web-scraping process by defining our URL to extract the list of Boston Neighborhoods. We user the python Requests library and BeautifulSoup to get our data and appended all the data to a DataFrame. In this case, the neighborhood items that had extraneous information in them were renamed to only be reflective of the neighborhood names as a part of data cleanup.

To begin the process of obtaining the latitude and longitude values for the neighborhoods, we defined a recursive function that would return the geocode of the address passed into it. This was done to ensure that the Geocoder Timeout exception was handled and would not interrupt the geocoding process for our dataset. We used Nominatim to do our geocoding. We appended latitudes to a 'latitude' list and longitudes to a 'longitude' list and then assigned the latitude and longitude lists to be columns in the original DataFrame. For this project, we handled the case where Nominatim returns a 'None' object for a neighborhood. Nominatim usually returns 'None' if the API does not have any data on the address it is queried for.

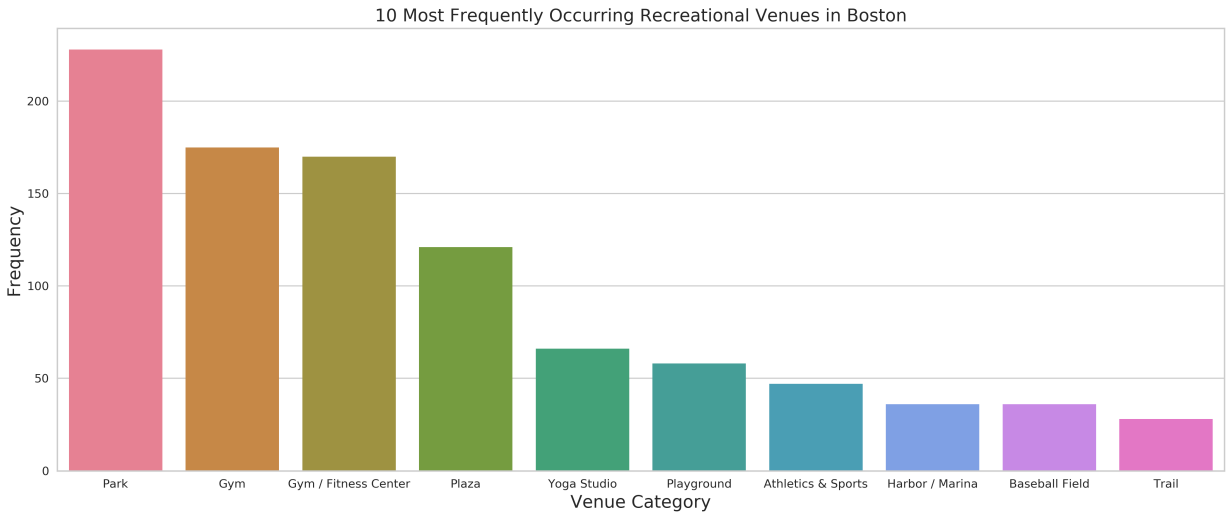
To get a better sense of the are we would be working with, we went ahead and plotted a map of Boston with neighborhoods superimposed on top using Folium. The following map was obtained.



3.2: Foursquare Data

To acquire recreational venue data from Foursquare, we went ahead and defined our Foursquare API credentials to being collecting data from their API. Once the credentials were defined, we went ahead and wrote a function that returns a list of venues that are near to a specific neighborhood. The function takes in the neighborhood name, neighborhood latitude, and neighborhood longitude as arguments and returns a DataFrame of all of the venues that are near to the neighborhood we specified. The URL defined in this function was defined to only return venues within the Foursquare "Outdoors and Recreation" category.

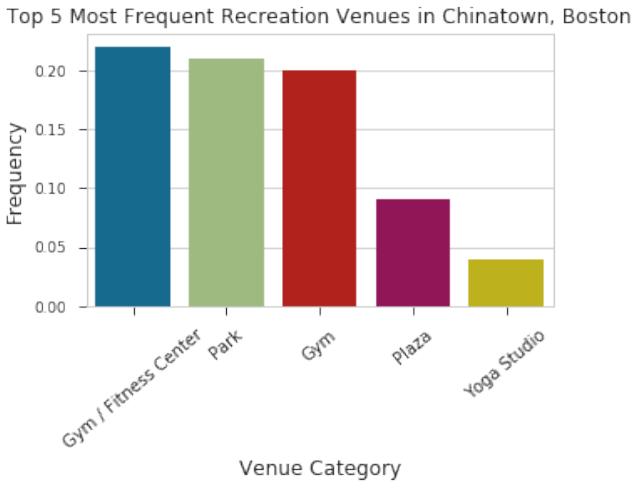
To understand our data, we took a look at the number of venues returned for each neighborhood and the number of unique categories Foursquare returned.
We conducted an analysis on which types of recreational venues are most frequent across all of Boston by creating a visualization using Seaborn. We found that the most popular venues across Boston were in the following order.

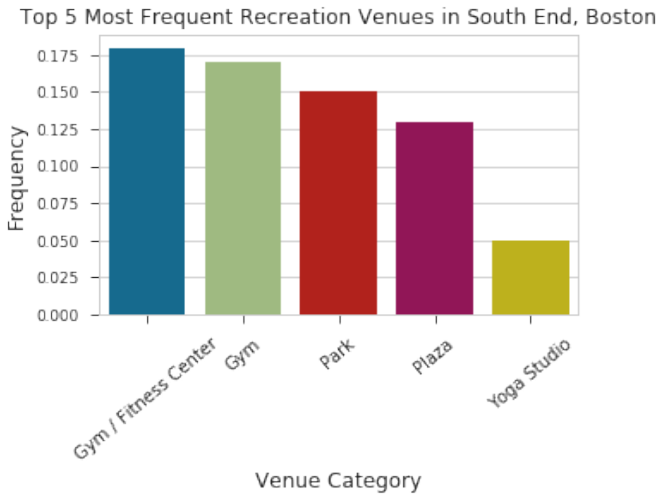


3.3: K-Means Clustering

To begin clustering our neighborhoods and analyzing the types of recreational venues that were most popular in them, we performed one-hot encoding to see what categories of venues are present in which neighborhoods. Then, we prepared our data for our clustering algorithm by grouping our one-hot DataFrame by neighborhood and taking the mean of the frequency of each category. We then assigned this grouping to a new DataFrame. Creating a grouped DataFrame in the manner outlined here allowed us to use a form of weighting to showcase how significant each category is in each neighborhood.

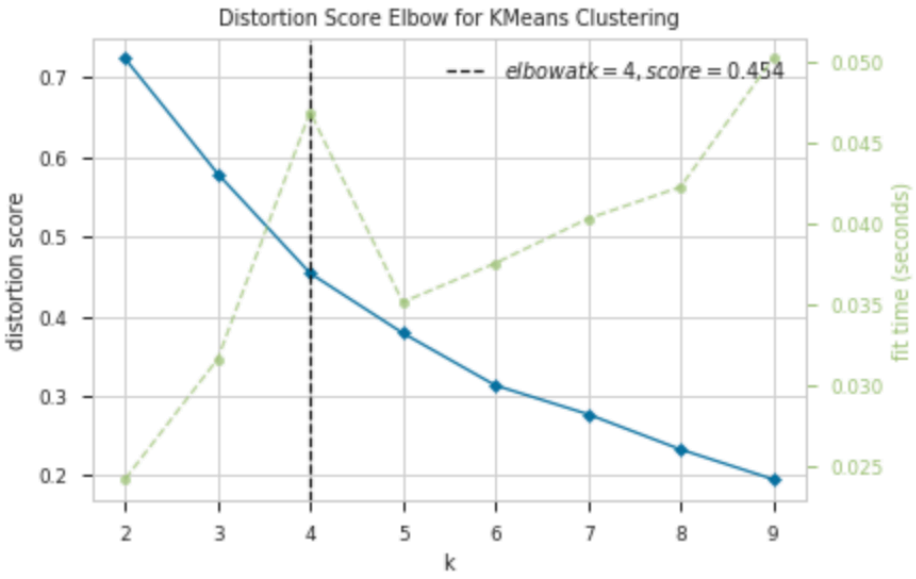
Since we had our grouped DataFrame, we were interested in analyzing what the most frequent recreational venue is in each neighborhood. We created a visualization for each neighborhood to further our understanding. This is helpful for those residing in these areas because they can understand what specific venues their neighborhood has the most of. Frequency can be linked to accessibility for many individuals. The visualizations can be found in the Jupyter notebook for this project, but we have included two of the visualizations here as an example.





We stored the information on the top 10 venues in each neighborhood into a pandas DataFrame. This was because the information would helpful later for visualizing what our clusters should approximately be after running K-Means. It also allowed us to see how K-means clustered our data.

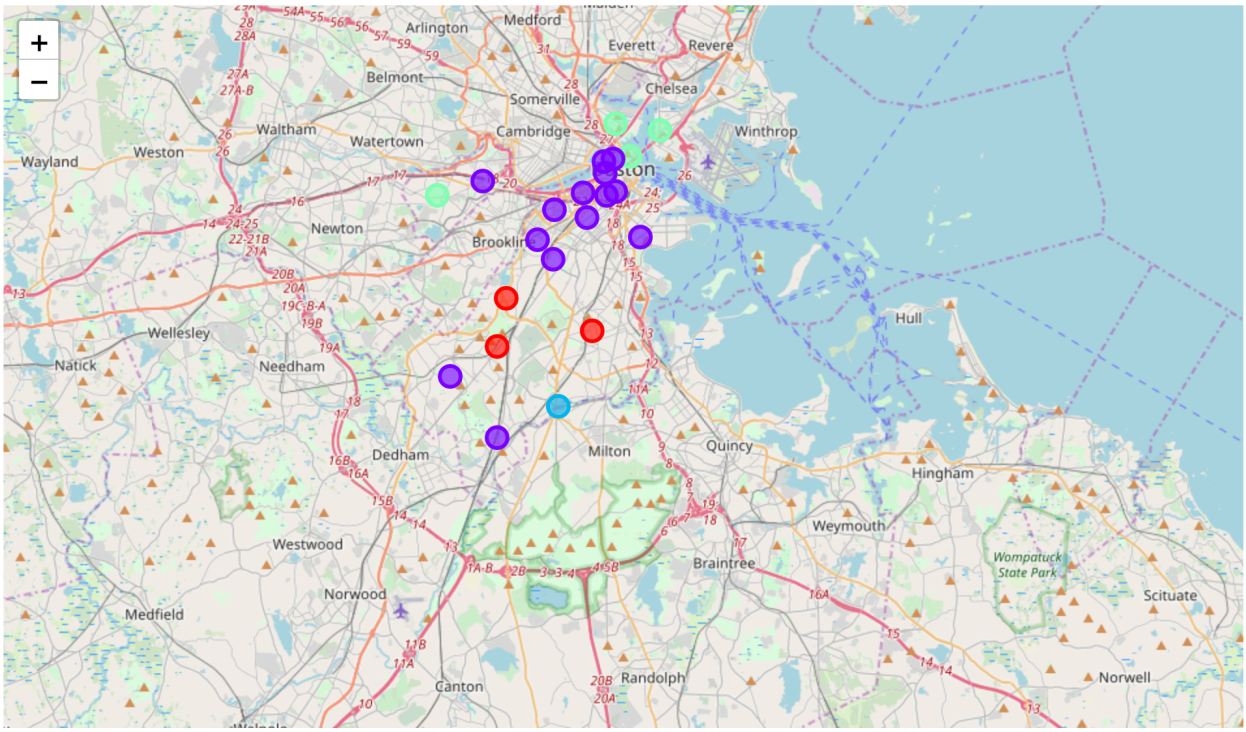
To begin our clustering, we had to determine which k would produce the best clusters for our data. We chose to use the Elbow Method in Yellowbrick to visualize which k would be the best for our data. The elbow method is used as an interpretation and validation of consistency within cluster analysis.



The visualization generated by Yellowbrick indicated that a k value of 4 would produce the most effective clusters for our data. We ran our K-means clustering algorithm on our weighted boston_grouped DataFrame with this k value after removing the Neighborhood name from the DataFrame. After we generated our cluster labels, we added them to our Top 10 Venues DataFrame.

4. RESULTS

After clustering was completed, we created a visualization of Boston with our clusters. The clusters of neighborhoods are differentiated by color.



K-means clustering led us to draw the following conclusions about our data:

- Cluster 1 contains neighborhoods in Boston that have very high numbers of plazas and parks.
- Cluster 2 contains neighborhoods with high numbers of gyms and fitness centers
- Cluster 3 contains the neighborhood Mattapan, where the most frequent recreational venue is a yoga studio.
    - It should be noted that the top 5 venues for this neighborhood all had equal weighting in the DataFrame, which could have contributed to why this neighborhood got its own cluster.
- Cluster 4 has neighborhoods with an extensive number of parks

DISCUSSION

1 , ideal for someone who likes to keep fit in an open public space.
2, which would be ideal for those looking to train in a traditional gym setting.
3, most balanced neighborhood, ideal for those looking to do yoga, esp moms because there's a playground\
4, , an ideal situation for those who like to stay fit outside.

CONCLUSION

SOURCES

1. https://www.nhs.uk/live-well/exercise/exercise-health-benefits/
2. https://adaa.org/living-with-anxiety/managing-anxiety/exercise-stress-and-anxiety