

## 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.

## 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.

## METHODS

First, import all necessary packages. Now that all of the packages have been installed, begin the web-scraping process by defining a URL and extracting the list of Boston Neighborhoods. Append all data to DataFrame and do clean-up if necessary. In this case, the neighborhood items that had extraneous information in them were renamed to only be reflective of the neighborhood names.

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.

Get the latitude and longitude values for each neighborhood in the DataFrame using the function defined above and append latitudes to a 'latitude' list and longitudes to a 'longitude' list.

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.

Assign the latitude and longitude lists to be columns in the original DataFrame. We are now going to create a visualization of Boston to see the area we will be segmenting. To do this, first get the coordinates of Boston using Nominatim. Then, create the visualization of Boston with the neighborhoods superimposed on top using Folium.

Now that we have our preliminary data cleaned up and have a sense of what we are looking to do, we will go ahead and define our Foursquare API credentials and will begin collecting data from them. Once the credentials are defined, we will go ahead and write 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 and returns a DataFrame of all of the venues that are near to the neighborhood we specified.

The URL defined in this function has been defined to only return venues within the Foursquare "Outdoors and Recreation" category.

To understand our data, we will take a look at the number of venues returned for each neighborhood and the number of unique categories Foursquare returned.

Now that we have our venue data, we can analyze which types of recreational venues are most populous across Boston. We can do this by creating a visualization.

First, create a DataFrame of the top 10 most frequent recreational venue types in Boston. Now, we will use Seaborn to generate our visualization to better understand the most frequently occurring recreational venues in all of Boston.

To begin clustering our neighborhoods and analyzing the types of recreational venues that are most popular in them, perform one-hot encoding to see what categories of venues are present in which neighborhoods.

Then, we prepare our data for future clustering by grouping our one-hot DataFrame by neighborhood and taking the mean of the frequency of each category. Assign this grouping to a new DataFrame.

Creating a grouped DataFrame in the manner outlined here allows us to use a form of weighting to showcase how significant each category is in each neighborhood.

Now that we have our grouped DataFrame, we are interested in knowing what the most frequent recreational venue is in each neighborhood. We will create 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.

We can store the information on the top 10 venues in each neighborhood into a pandas DataFrame.

This will be helpful later for visualizing what our clusters should approximately be after running K-Means.

Now we begin our clustering. First we have to determine which k will 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 visualization generated by Yellowbrick indicated that a k value of 4 would produce the most effective clusters for our data.

We will be running 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 generate our cluster labels, we will add them to our Top 10 Venues DataFrame. Again, this will make it easier for us to understand how k-means clustered our neighborhoods in the future.

## RESULTS

Now that our clustering is complete, we will create a visualization of Boston with our clusters. The clusters of neighborhoods are differentiated by color

Cluster 1 contains neighborhoods in Boston that have very high numbers of plazas and parks, ideal for someone who likes to keep fit in an open public space.

Cluster 2 contains neighborhoods with high numbers of gyms and fitness centers which would be ideal for those looking to train in a traditional gym setting.

Cluster 3 contains the neighborhood Mattapan, where the most frequent recreational venue is a yoga studio. It should also be noted that the weighting for the top 5 venues for this neighborhood all had equal weight, which could have contributed to why this neighborhood got its own cluster.

Finally, Cluster 4 has neighborhoods with an extensive number of parks, an ideal situation for those who like to stay fit outside.

## DISCUSSION

## 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>