Find My Home

A Predictive Model for Finding a Neighborhood that Matches a User's Lifestyle

Introduction

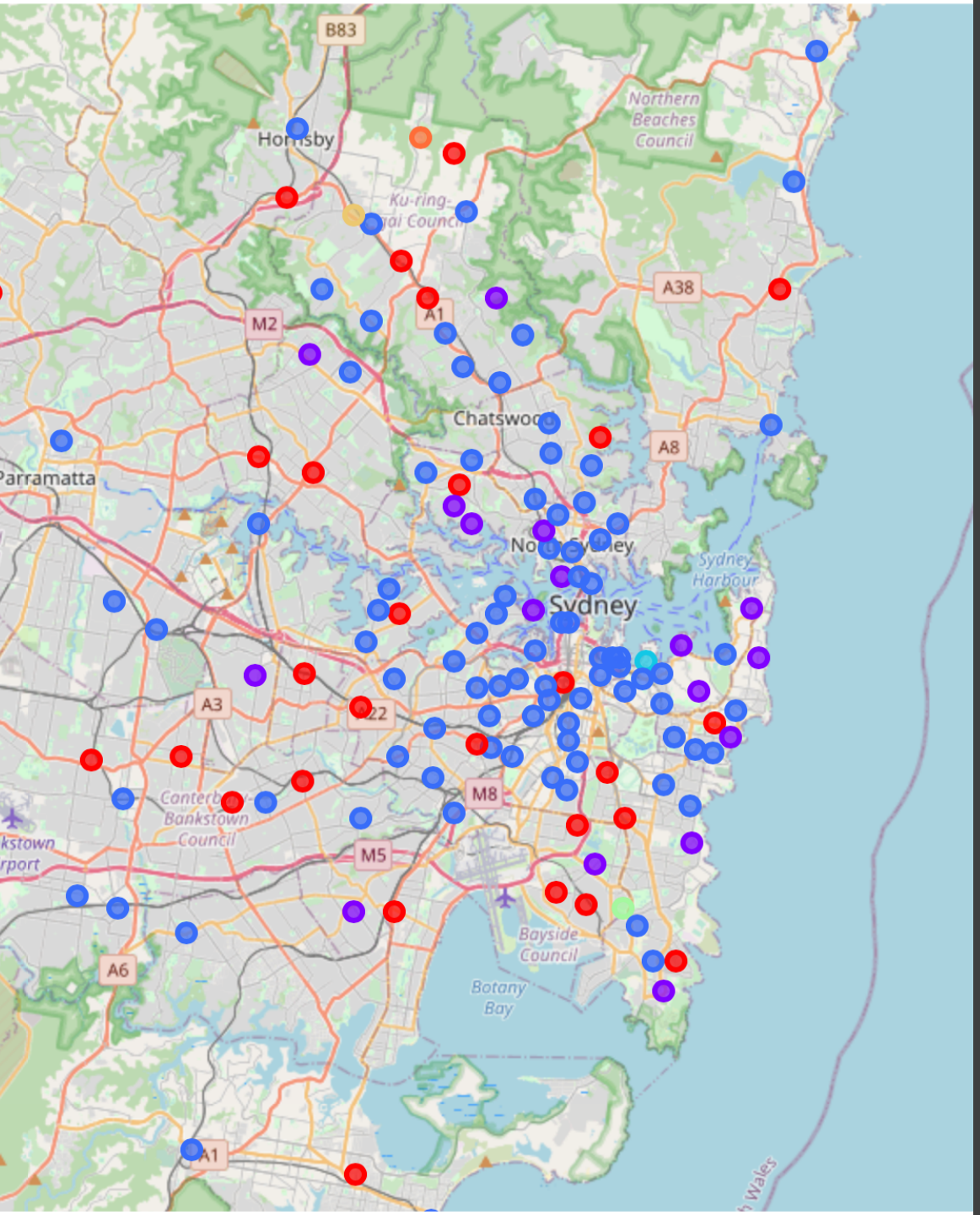
- The city of Sydney, Australia is made up of a wide variety of neighborhoods and it is difficult when arriving for the first time to know where to move that will fit your lifestyle
- This project’s goal is to be able to find the neighborhoods that best match the user.
- To accomplish this, I developed profiles for each neighborhood and matched them to a user's lifestyle preferences (described by analyzing their Foursquare profile).

Data

- Utilizing an HTML scraping of a list of Sydney neighborhoods, a dataframe is created that contained name, postal code, and geo-coordinates for each neighborhood.
- From here Foursquare data was used to create a cluster analysis based on each neighborhood's top 10 most common venue types.
- I used a user profile in Foursquare to develop a top 10 most visited venue type list.
- Data will be processed using Python and associated libraries (scikit-learn, pandas, numpy, BeautifulSoup4, folium, and matplotlib).

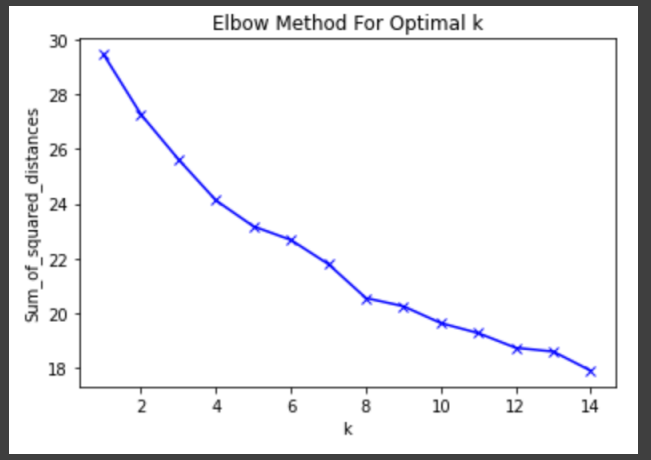
Methodology

- For data scraping, BeautifulSoup4 was employed. Data cleaning was performed to acquire a table of neighborhoods in Sydney with associated geographic coordinates and postal codes.
- Using the Foursquare API, this table was used to extract the nearby venues by category for each neighborhood in Sydney. With one-hot encoding, these categories were turned into dummy variables. The nearby venues were then grouped together so that every neighborhood row would have mean-weighted counts for each venue category. This data was further refined by only taking the highest 10 values for each neighborhood (e.g. the top 10 venue categories for that neighborhood).
- With this processed data, a KMeans cluster analysis could be performed. These clusters create unsupervised relationships among neighborhoods.
- Because KMeans clustering is not meant for the classification of data. It can't be used to apply our user data. Therefore, after the development of cluster labels. We can apply K-nearest neighbor methods to classify a user to a particular cluster.
- To get the user data, we can use the Foursquare API to pull a user account's visited venues undergo the same transformation as neighborhoods to achieve a list of top 10 venue categories for the user. With this information, the K-Nearest neighbor model can be developed with the neighborhood data and fitted to the user data.
- Based on the predicted cluster label for this user, the user will then have a list of neighborhoods that are best fitted to their lifestyle.

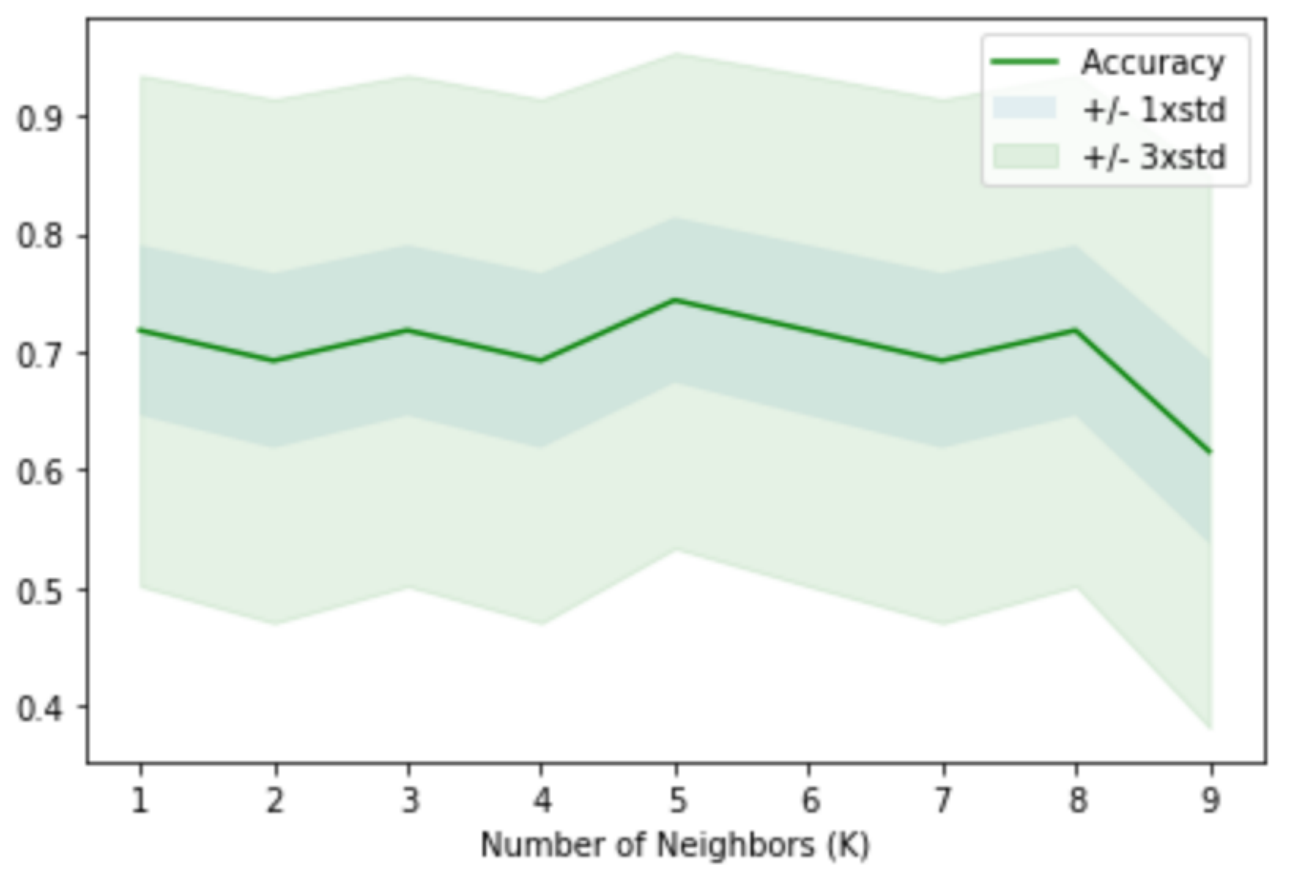


Results – Cluster Analysis

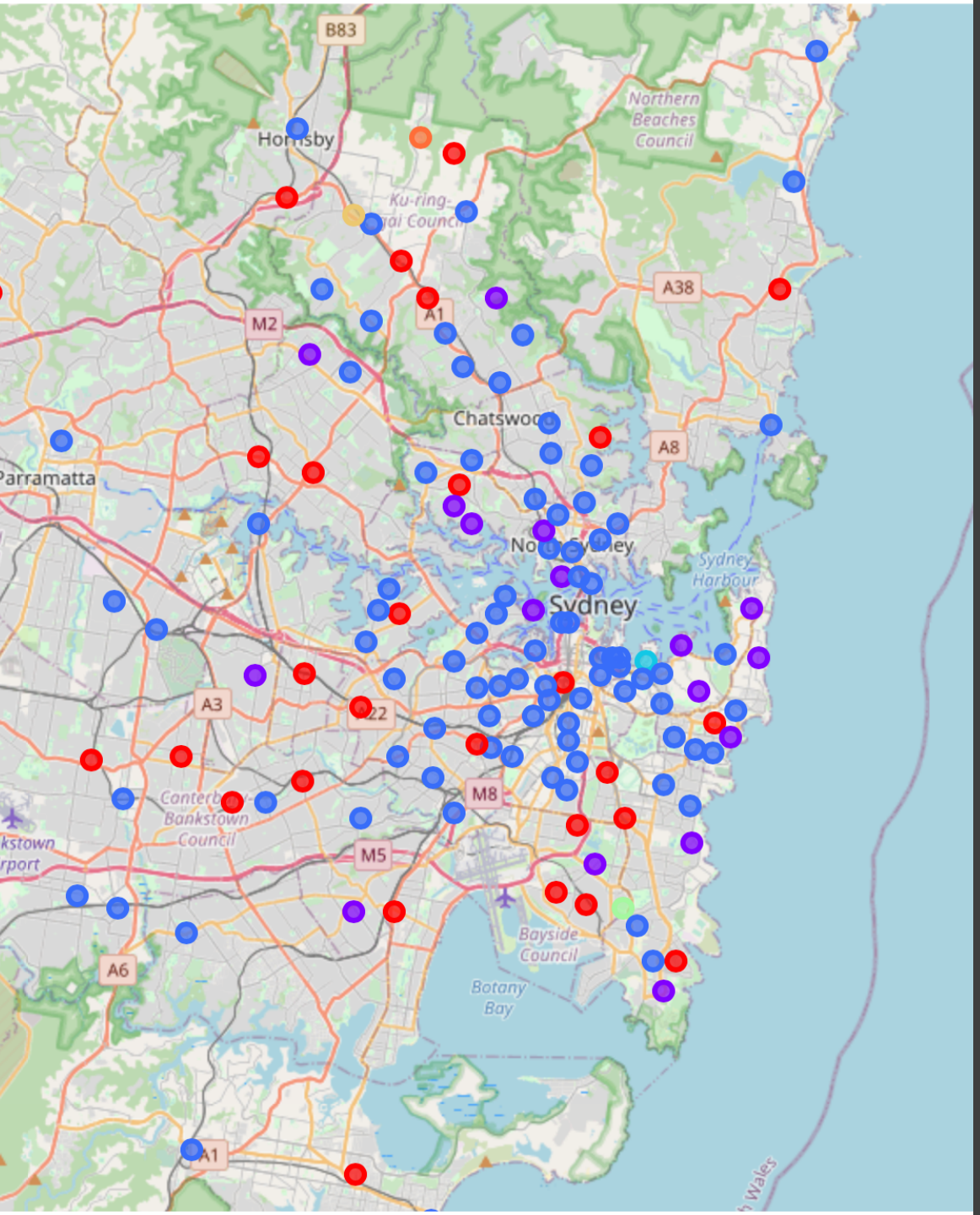
- Cluster using venue category by neighborhood
- The elbow method was employed to decide on the best number of clusters. The graph that was produced had a very minor elbow, but it was decided that 8 was the best choice.
- With 8, there isn't as much differentiation though.



Results



- The second part of the project involved using the cluster labels as a target variable for a prediction model
- The training of the KNN model showed an accuracy of 0.74 using k=5.
- It was able to place the user profile in the 2nd cluster.



Conclusions

- The user was matched to the 2nd cluster (Blue dots on map)
- Poor neighborhood differentiation
- Test user was a profile with poor usage

- Future directions would be to overlay more features onto the neighborhoods (avg. price of house, location based on user's work, etc)