CAPSTONE

ACME Startup Relocation

ABSTRACT

This Capstone project is a demonstration of key data science concepts applied to the scenario of a company relocation using factors determined by an employee survey.

Manja, Frank

IBM Data Science Professional

Introduction/Business Problem

ACME Start is a Machine Learning (ML) startup in New York City that recently received Series A Funding in the amount of $20M USD. ACME currently has 10 employees and is looking to grow to 20 employees by years end. The startup has outgrown its current location and needs to relocate to accommodate current and future employees.

Data

The Chief Executive Officer (CEO) is concerned that the relocation may result in staff attrition so the CEO asks the Human Resources (HR) Director to perform a survey of employees. The survey asks employees to rank factors that the company will incorporate in the selection of the new location. These factors include proximity to mass restaurants, parks, and gyms. The HR Director will use these results of the survey and Foursquare location data to prepare a report of suitable locations for the relocation.

Methodology

The Data Science team follows the standard Data Science Methodology illustrated in Figure 1.

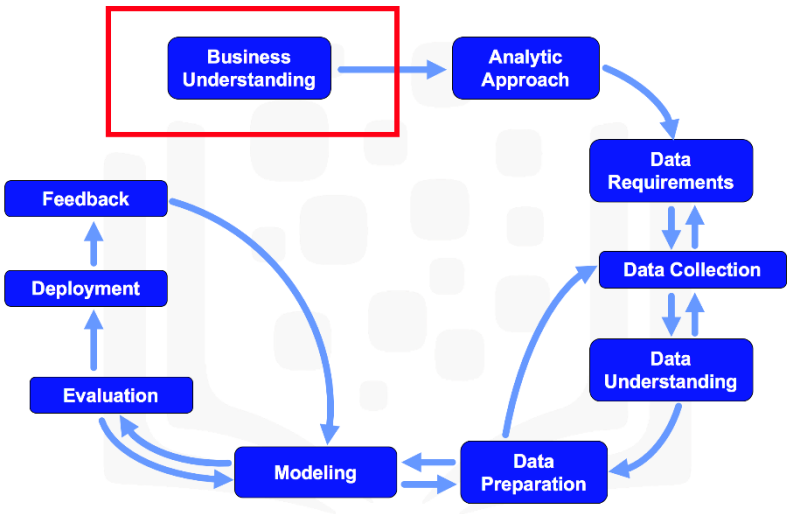


Figure 1. Data Science Methodology

Business Understanding

The HR Director is the lead for this analysis. She understands that office relocations can be stressful for employees and at time bring about unwanted attrition so she will prioritize employee engagement.

Analytics Approach

The HR Director has asked the Data Science team to conduct a survey to of employees to gauge employee sentiment and use the information learned in that survey to guide the data analysis.

Data Requirements

The Data Science team will center their analysis on three data sources:

- Employee Survey
- New York City Maps
- Foursquare Location Data

Data Collection

The Data Science team has collected the raw survey results from employees using SurveyMonkey. The team downloaded mapping files from the NY city website. The team signed up for a Foursquare account to access data using the Foursquare API.

Data Understanding

The Data Science team will use standard data exploration techniques on each data source to learn the structure and dimensions of all data.

Data Preparation

The Data Science team will transform data, making sure there are no missing data values, and incorrect data types. The team will impute missing data from available data. Also, the team will map Foursquare Venue Categories to the ranked features that employees called for in the survey. All data transformation steps will be documented in the Notebook.

Modeling

The Data Science team will furnish maps of potential sites and segment the sites by their features. The team decided to use clustering and segmentation techniques as they perform well on geographical data where unsupervised learning is needed. The team will use K-means with K=5.

Evaluation

The Data Science team will evaluate sites for suitability by determining what venues are within 200 meters of potential locations.

Deployment

The Data Science team will present findings to the HR Director and CEO.

Feedback

The Data Science team will solicit feedback from the HR Director and CEO, both of whom may ask for additional factors to be included in the analysis.

Data Analysis

Survey Data

The raw survey contained ten records. One record for each employee in the startup.

```
4]: raw_survey_df
```

4]:

	SurveyID	JobFamily	Resturants	Gyms	Parks
0	1	Developer	3.0	1	2.0
1	2	Management	5.0	NaN	NaN
2	3	Management	2.0	5	4.0
3	4	Developer	NaN	1	5.0
4	5	Developer	1.0	1	NaN
5	6	Marketing/Sales	3.0	NaN	1.0
6	7	Marketing/Sales	NaN	5	1.0
7	8	Developer	2.0	3	5.0
8	9	Data Scientist	2.0	3	1.0
9	10	Data Scientist	5.0		3.0

As common in most surveys, the survey takers did not answer every question. The Data Science team worked to identify missing data and incorrect datatypes.

Fix data issue with empty data not identified with NaN

```
5]: raw_survey_df.replace(" ", np.nan, inplace = True)
```

Check the datatypes

```
6]: raw_survey_df.dtypes
```

```
6]: SurveyID        int64
    JobFamily      object
    Resturants    float64
    Gyms           object
    Parks         float64
    dtype: object
```

Change the data type of Gyms to float64

```
7]: raw_survey_df['Gyms'] = raw_survey_df['Gyms'].astype("float")
```

To deal with the missing data the Data Science team replaced missing data with the mean of values of available data.

Deal with missing data

```
[9]: avgResturants = raw_survey_df['Resturants'].astype("float").mean(axis=0)
     raw_survey_df["Resturants"].replace(np.nan, avgResturants, inplace=True)
     print('Set missing Resturants to Average Resturants:',avgResturants)

     avgGyms = raw_survey_df['Gyms'].astype("float").mean(axis=0)
     raw_survey_df["Gyms"].replace(np.nan, avgGyms, inplace=True)
     print('Set missing Gyms to Average Gyms:',avgGyms)

     avgParks = raw_survey_df['Parks'].astype("float").mean(axis=0)
     raw_survey_df["Parks"].replace(np.nan, avgParks, inplace=True)
     print('Set missing Parks to Average Parks:',avgParks)
```

```
Set missing Resturants to Average Resturants: 2.875
Set missing Gyms to Average Gyms: 2.7142857142857144
Set missing Parks to Average Parks: 2.75
```

The team performed exploratory data analysis on the survey.

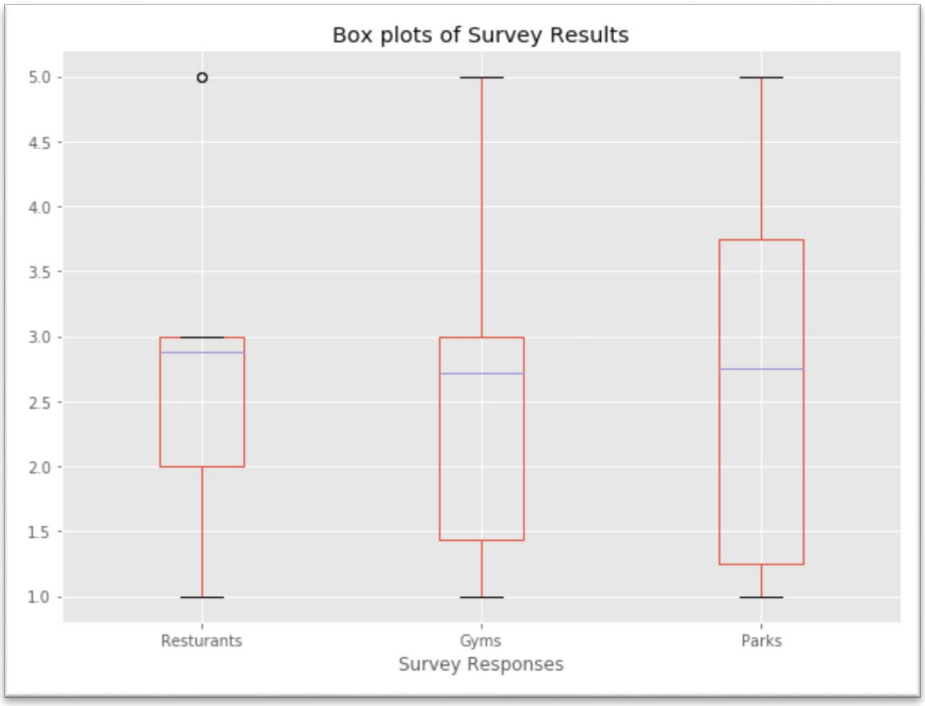
Perform 'Exploratory Data Analysis'

```
11]: survey_df.describe()
```

11]:

	SurveyID	Resturants	Gyms	Parks
count	10.00000	10.000000	10.000000	10.000000
mean	5.50000	2.875000	2.714286	2.750000
std	3.02765	1.285604	1.469262	1.545603
min	1.00000	1.000000	1.000000	1.000000
25%	3.25000	2.000000	1.428571	1.250000
50%	5.50000	2.875000	2.714286	2.750000
75%	7.75000	3.000000	3.000000	3.750000
max	10.00000	5.000000	5.000000	5.000000

The team visualize the survey data.



The team calculated the means of survey factors to determine the ranking of location factors that staff want to inform the relocation decision.

Calculate the means to determine the ranking of factors that staff want to inform the decision

```
12]: factors_df=survey_df[['Resturants','Gyms','Parks']].mean(axis=0).reset_index()

13]: factors_df
```

13]:

	index	0
0	Resturants	2.875000
1	Gyms	2.714286
2	Parks	2.750000

```
14]: factors_df.columns=['Factor','Average']
```

Rank the factors

```
15]: factors_df['Rank']=factors_df.index + 1
```

The ranked factors are in order: Restaurants, Gyms, and Parks.

```
16]: factors_df
```

16]:

	Factor	Average	Rank
0	Resturants	2.875000	1
1	Gyms	2.714286	2
2	Parks	2.750000	3

New York Neighborhood Dataset

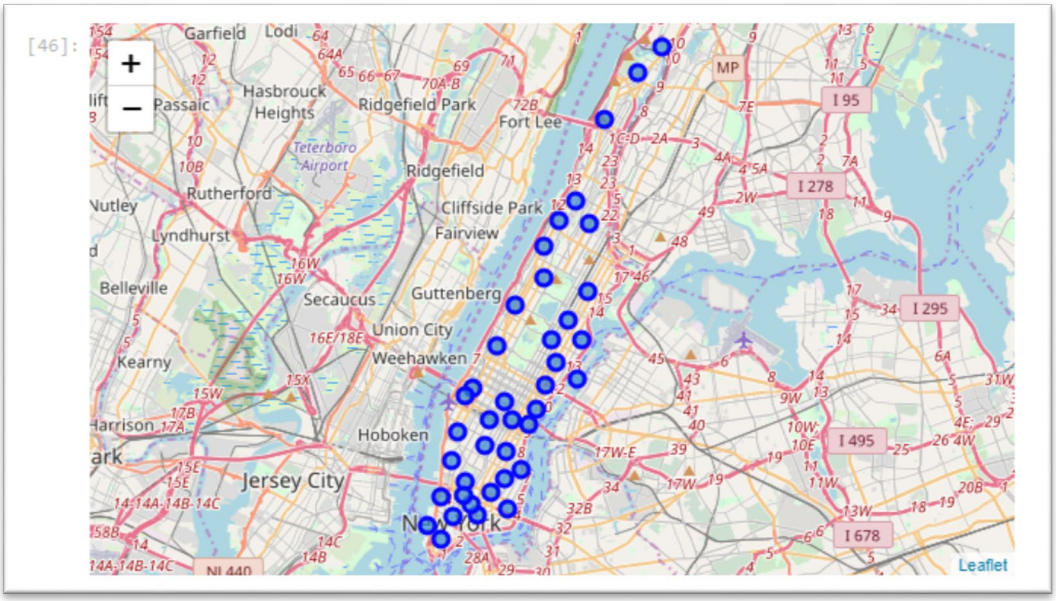
The New York Neighborhood dataset has a total of 5 boroughs and 306 neighborhoods. A link to the dataset is at https://geo.nyu.edu/catalog/nyu_2451_34572. Sample data from the dataset is shown below.

```
[20]: neighborhoods_data[0]
```

[20]:


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


The CEO decided to center the relocation search to neighborhoods in Manhattan.



The Data Science team used the Foursquare API to get venue information for each of the Manhattan neighborhoods.

```
[59]: print(manhattan_venues.shape)
      manhattan_venues.head()

      (800, 7)
```

[59]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

However, the venue categories provided by Foursquare varied significantly from the location factors the employees were surveyed on. The Data Science team wrote a function to map Foursquare venue categories to location factors and in the process consolidated all place to eat into the Restaurants factor, all places of leisure to the Parks factor, and all places for physical activity to the Gyms factor. The function is provided below.

```
# Function to map venue categories to employee factors
def Map_Venue_Category(Venue_Category):

    #print (Venue_Category)
    if re.search('Restaurant', Venue_Category, re.IGNORECASE):

        return 'Restaurant'

    if re.search('Joint', Venue_Category, re.IGNORECASE):
```

```
        return 'Restaurant'
    if re.search('Tea Room', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Pizza Place', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Donut Shop', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Coffee Shop', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Bagel Shop', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Bakery', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Bar', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Bistro', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Beer', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Burrito', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Breakfast', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Tea Shop', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
```

```
    if re.search('Candy', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Café', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Cheese', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Chocolate', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Bridge', Venue_Category, re.IGNORECASE):
        return 'MassTransit'
    if re.search('Bus Line', Venue_Category, re.IGNORECASE):
        return 'MassTransit'
    if re.search('Community Center', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Cupcake', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Deli', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Dessert', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Diner', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Dog Run', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
```

```
    if re.search('Farmers Market', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Food', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Fountain', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Yogurt', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Gourmet', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Grocery', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Gym', Venue_Category, re.IGNORECASE):
        return 'Gym'
    if re.search('Harbor', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Histor', Venue_Category, re.IGNORECASE): #for historic and history
        return 'Park'
    if re.search('ice cream', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Liquor', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Memorial', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Pie', Venue_Category, re.IGNORECASE):
```

```
        return 'Restaurant'
    if re.search('Poke Place', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Pool', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Snack', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Steak', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Taco', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Tennis Stadium', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Trail', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Field', Venue_Category, re.IGNORECASE):
        return 'Park'
    if re.search('Spa', Venue_Category, re.IGNORECASE):
        return 'Gym'
    if re.search('Golf Course', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Playground', Venue_Category, re.IGNORECASE):
        return 'Restaurant'
    if re.search('Plaza', Venue_Category, re.IGNORECASE):
```

```
        return 'Park'

    if re.search('Basketball Court', Venue_Category, re.IGNORECASE):

        return 'Park'

    if re.search('GastroPub', Venue_Category, re.IGNORECASE):

        return 'Restaurant'

    if re.search('Momument', Venue_Category, re.IGNORECASE):

        return 'Park'

    if re.search('Sandwich', Venue_Category, re.IGNORECASE):

        return 'Restaurant'

    if re.search('Gastro', Venue_Category, re.IGNORECASE):

        return 'Restaurant'

    if re.search('Tennis', Venue_Category, re.IGNORECASE):

        return 'Park'

    if re.search('Waterfront', Venue_Category, re.IGNORECASE):

        return 'Park'

    else:
        # if clean up needed return the same name
        return Venue_Category
```

After the Data Science Team mapped the venue categories to employee categories the resulting data was aligned with the objectives of the location analysis.

Neighborhood	Employee_Venue_Category	
Battery Park City	Boat or Ferry	1
	Cooking School	1
	Gym	1
	Park	6
	Performing Arts Venue	1
	Restaurant	8
	Shopping Mall	1
	Smoke Shop	1
Carnegie Hill	Bookstore	1
	Dance Studio	1
	Gym	5

	Park	2
	Restaurant	9
	Shoe Store	1
	Wine Shop	1
Central Harlem	Cycle Studio	1
	Gym	1
	Jazz Club	1
	Library	1
	Music Venue	1
	Restaurant	15
Chelsea	Hotel	1
	Nightclub	2
	Restaurant	14
	Speakeasy	1
	Theater	2
Chinatown	Bike Shop	1
	Garden Center	1
	Gym	1
	Hotel	1
	Museum	1
	Noodle House	2
	Restaurant	13
Civic Center	Antique Shop	1
	Dance Studio	1
	General Entertainment	1
	Gym	4
	Monument / Landmark	1
	Park	1
	Restaurant	10
	Yoga Studio	1
Clinton	Building	1
	Comedy Club	1
	Gym	3
	Hotel	1
	Indie Theater	1
	Lounge	1
	Movie Theater	1
	Restaurant	7
	Sporting Goods Shop	1
	Theater	3
East Harlem	Clothing Store	1
	Dance Studio	1
	Gym	1
	Park	1
	Pet Store	1
	Pharmacy	1
	Restaurant	14
East Village	Restaurant	19
	Wine Shop	1

Financial District	Doctor's Office	1
	Gym	4
	Jewelry Store	2
	Monument / Landmark	1
	Park	1
	Restaurant	10
	Salad Place	1
Flatiron	Bookstore	1
	Cycle Studio	2
	Furniture / Home Store	2
	Gym	4
	Miscellaneous Shop	1
	Restaurant	7
	Salad Place	1
	Sports Club	1
	Wine Shop	1
Gramercy	Bike Rental / Bike Share	1
	Comedy Club	1
	Gym	1
	Park	1
	Restaurant	14
	Thrift / Vintage Store	1
	Yoga Studio	1
Greenwich Village	Clothing Store	1
	Cosmetics Shop	1
	Jazz Club	1
	Optical Shop	1
	Park	1
	Restaurant	14
	Yoga Studio	1
Hamilton Heights	Park	1
	Pub	1
	Restaurant	15
	Smoke Shop	1
	Yoga Studio	2
Hudson Yards	Art Gallery	1
	Department Store	1
	Furniture / Home Store	1
	Gym	1
	Hotel	1
	Music School	1
	Park	1
	Pet Store	1
	Public Art	1
	Residential Building (Apartment / Condo)	1
	Restaurant	8
	Supermarket	1
	Theater	1
Inwood	Park	2

	Pet Store	1
	Restaurant	14
	Veterinarian	1
	Wine Shop	1
	Yoga Studio	1
Lenox Hill	College Academic Building	1
	Gift Shop	1
	Gym	2
	Park	1
	Restaurant	11
	Salad Place	1
	Smoke Shop	1
	Wine Shop	1
	Women's Store	1
Lincoln Square	Concert Hall	3
	Gift Shop	1
	Gym	1
	Indie Movie Theater	3
	Indie Theater	1
	Library	1
	Opera House	2
	Park	2
	Performing Arts Venue	2
	School	1
	Theater	3
Little Italy	Animal Shelter	1
	Clothing Store	1
	Newsstand	1
	Park	2
	Restaurant	13
	Salad Place	1
	Women's Store	1
Lower East Side	Art Gallery	2
	Clothing Store	1
	Park	1
	Performing Arts Venue	1
	Restaurant	14
	Yoga Studio	1
Manhattan Valley	Arts & Crafts Store	1
	Bike Shop	1
	Cosmetics Shop	1
	Furniture / Home Store	1
	Hostel	1
	Park	1
	Restaurant	12
	Wine Shop	1
	Yoga Studio	1
Manhattanville	Gym	1
	Lounge	1

	Other Nightlife	1
	Park	2
	Restaurant	14
	Supermarket	1
Marble Hill	Bank	1
	Department Store	1
	Discount Store	1
	Gym	1
	Miscellaneous Shop	1
	Park	1
	Pharmacy	1
	Restaurant	10
	Supplement Shop	1
	Video Game Store	1
	Yoga Studio	1
Midtown	Clothing Store	1
	Gym	2
	Hotel	2
	Miscellaneous Shop	1
	Park	3
	Restaurant	8
	Salad Place	1
	Smoke Shop	1
	Sporting Goods Shop	1
Midtown South	Boutique	1
	Building	1
	Clothing Store	1
	Cosmetics Shop	2
	Hotel	1
	Lingerie Store	2
	Restaurant	12
Morningside Heights	Bookstore	3
	Outdoor Sculpture	1
	Park	2
	Pub	1
	Restaurant	12
	Salad Place	1
Murray Hill	Hotel	2
	Jazz Club	1
	Lounge	1
	Museum	1
	Restaurant	14
	Speakeasy	1
Noho	Boutique	1
	Gym	1
	Hotel	1
	Park	1
	Restaurant	13
	Rock Club	2

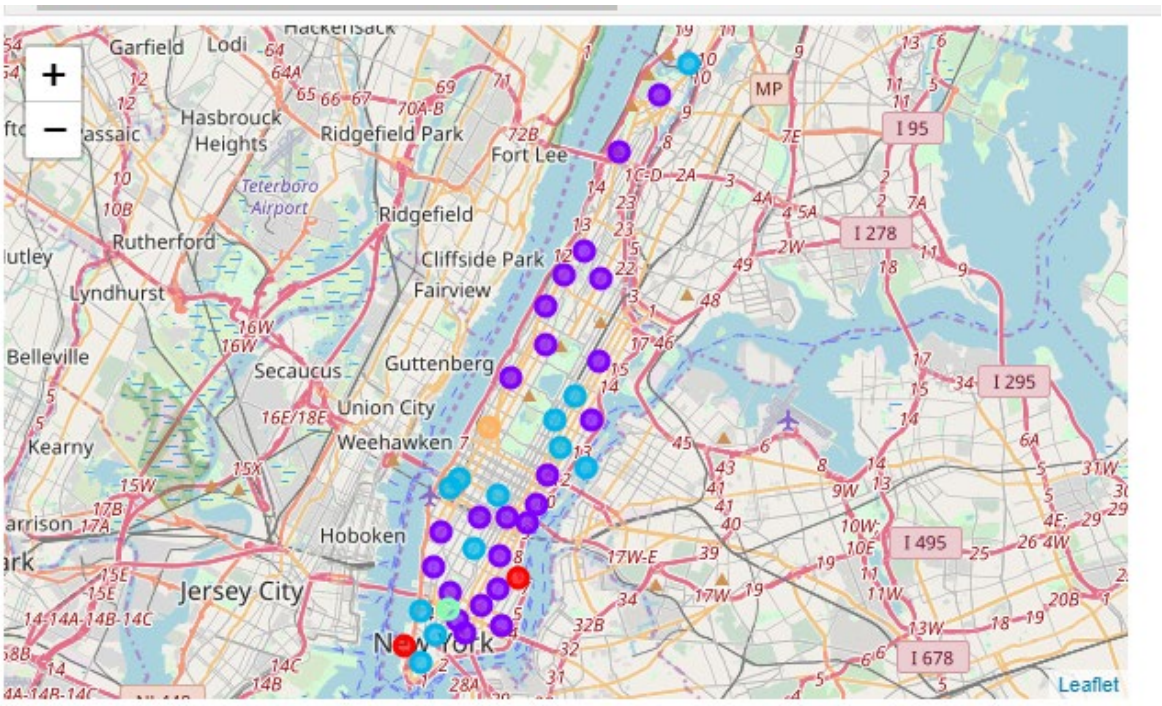
	Wine Shop	1
Roosevelt Island	Cosmetics Shop	1
	Gym	2
	MassTransit	1
	Outdoors & Recreation	1
	Park	3
	Residential Building (Apartment / Condo)	1
	Restaurant	9
	Scenic Lookout	1
	School	1
Soho	Art Museum	1
	Arts & Crafts Store	1
	Boutique	1
	Clothing Store	2
	Cycle Studio	1
	Dance Studio	1
	Men's Store	2
	Miscellaneous Shop	1
	Optical Shop	1
	Restaurant	5
	Women's Store	3
	Yoga Studio	1
Stuyvesant Town	Boat or Ferry	2
	Gas Station	1
	Heliport	1
	Park	6
	Pet Service	1
	Restaurant	9
Sutton Place	Adult Boutique	1
	Gym	4
	Restaurant	13
	Spiritual Center	1
	Yoga Studio	1
Tribeca	Cycle Studio	1
	Gym	1
	Hotel	1
	Indie Theater	1
	Park	2
	Restaurant	10
	Salad Place	1
	Wine Shop	2
	Yoga Studio	1
Tudor City	Convenience Store	1
	Martial Arts Dojo	1
	MassTransit	1
	Park	4
	Restaurant	12
	Yoga Studio	1
Turtle Bay	Duty-free Shop	1

	Gift Shop	1
	Lounge	1
	Martial Arts Dojo	1
	Museum	1
	Park	1
	Residential Building (Apartment / Condo)	1
	Restaurant	12
	Tourist Information Center	1
Upper East Side	Art Gallery	1
	Bookstore	1
	Boutique	1
	Gym	2
	Hotel	2
	Jazz Club	1
	Optical Shop	1
	Restaurant	11
Upper West Side	Bookstore	1
	Cosmetics Shop	1
	Gift Shop	1
	Movie Theater	1
	Pub	1
	Restaurant	15
Washington Heights	Market	1
	Park	3
	Restaurant	14
	Wine Shop	2
West Village	Accessories Store	1
	Boutique	1
	Cosmetics Shop	2
	Park	2
	Restaurant	13
	Speakeasy	1
Yorkville	Gym	1
	Hobby Shop	1
	Monument / Landmark	1
	Park	1
	Restaurant	13
	Video Store	1
	Wine Shop	2

Results

The Data Science team used K-means with K=5 to cluster the Manhattan neighborhoods using the employee factors.

[76]:



Then the Data Science team found the one location in all of Manhattan that had all the venues in the same rank order as the employees requested in the survey. That location was Carnegie Hill.

Find the location that has all of the venues in the rank order chosen by the employees

```
83]: manhattan_merged.loc[(manhattan_merged['1st Most Common Venue'] == 'Restaurant') &
                          (manhattan_merged['2nd Most Common Venue'] == 'Gym') &
                          (manhattan_merged['3rd Most Common Venue'] == 'Park')]
```

83]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
30	Manhattan	Carnegie Hill	40.782683	-73.953256	2	Restaurant	Gym	Park	Bookstore	Wine Shop	Shoe Store	Dance Studio

From Wikipedia we learned the following about Carnegie Hill:

> Carnegie Hill is a neighborhood within the Upper East Side, in the borough of Manhattan in New York City. Its boundaries are 86th Street on the south, Fifth Avenue (Central Park) on the west, with a northern boundary at 98th Street that continues just past Park Avenue and turns south to 96th Street and proceeds east up to, but not including, Third Avenue.

Discussion

The key element of this analysis was clearly the ranking of factors by employees. This was by design as the HR Director wanted to ensure employees felt they were heard and respected in this office relocations. As a startup it is critical to maintain key staff members. The analysis that the Data Science team undertook found the one location in Manhattan, NY that met all criteria.

Potentially, the analysis could be tweaked by adding new survey questions for employees or weighting the response of developers and data scientist over others in the company. Also, the search radius around each location could be increased from 200 meters (1 city block) to 500 meters (2.5 city blocks) to get more information on nearby venues. However, that could lead to noise and potentially overlapping locations.

Conclusion

The analysis met its desired outcome by identifying the single location in Manhattan that clear matched all the ranked desires of the staff.