Giving recommendations about moving company office from New York to Berlin

Boiarchenkov Ramil

April 08, 2020

1. Introduction

Company office is located in one of the neighbourhood in New York. But because of business reasons stakeholders decided to move the office to Berlin. People are used to the infrastructure that they had in New York neighbourhood and they want to feel the same level of comfort in Berlin.

The target audience - company stakeholders and employees.

The problem - selecting the most similar neighbourhood to New York neighbourhood in Berlin.

The main reason - having the same infrastructure and the same level of comfort.

We will need to leverage the Foursquare location data for all neighbourhoods in both cities to make the right decision.

2. Data

We will take he Foursquare location data for all neighbourhoods in both cities. We will gather data on all vanues, preprocces it, so we have the mean amount of all vanues and cluster neighbourhoods in both cities.
The final dataset will have the data on all vanues that are located in neighbourhoods. This will allow us to make a proper clustering.

3. Methology

3.1 Data cleaning and gathering

First, we repeat the same process for New York Queens as we did for Manhattan before. As a result we get the nest data frame:

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant
0	Arverne	0.0	0.000000	0.0000	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
1	Astoria	0.0	0.000000	0.0000	0.010000	0.000000	0.0	0.0	0.0	0.0	...	0.01	0.000000	0.0	0.0
2	Astoria Heights	0.0	0.000000	0.0000	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
3	Auburndale	0.0	0.000000	0.0000	0.055556	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
4	Bay Terrace	0.0	0.027027	0.0000	0.054054	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.027027	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
76	Sunnyside Gardens	0.0	0.000000	0.0000	0.030000	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.010000	0.0	0.0
77	Utopia	0.0	0.000000	0.0625	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
78	Whitestone	0.0	0.000000	0.0000	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
79	Woodhaven	0.0	0.000000	0.0000	0.000000	0.038462	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0
80	Woodside	0.0	0.000000	0.0000	0.036585	0.012195	0.0	0.0	0.0	0.0	...	0.00	0.000000	0.0	0.0

81 rows × 269 columns

Next we will repeat the process for Berlin. First, we gather data about Berlin neighbourhoods from Wikipedia page ('https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin'):

	Neighborhood	Borough
0	(0101) Mitte	Mitte
1	(0102) Moabit	Mitte
2	(0103) Hansaviertel	Mitte
3	(0104) Tiergarten	Mitte
4	(0105) Wedding	Mitte
...	...	...
91	(1207) Waidmannslust	Reinickendorf
92	(1208) Lübars	Reinickendorf
93	(1209) Wittenau	Reinickendorf
94	(1210) Märkisches Viertel	Reinickendorf
95	(1211) Borsigwalde	Reinickendorf

96 rows × 2 columns

Then we get coordinates as we did before and add information about venues:

	Neighborhood	Zoo Exhibit	ATM	African Restaurant	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Asian Restaurant	Austrian Restaurant	...	Vietnamese Restaurant	Vineyard	Volleyball Court	Warehouse Store
0	(0101) Mitte	0.0	0.0	0.0	0.000000	0.0	0.04	0.020000	0.000000	0.000000	...	0.020000	0.0	0.0	0.0000
1	(0102) Moabit	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.015152	...	0.015152	0.0	0.0	0.0000
2	(0103) Hansaviertel	0.0	0.0	0.0	0.000000	0.0	0.00	0.074074	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000
3	(0104) Tiergarten	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000
4	(0105) Wedding	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
88	(1207) Waidmannslust	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0769
89	(1208) Lübars	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000
90	(1209) Wittenau	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000
91	(1210) Märkisches Viertel	0.0	0.0	0.0	0.083333	0.0	0.00	0.000000	0.083333	0.000000	...	0.000000	0.0	0.0	0.0000
92	(1211) Borsigwalde	0.0	0.0	0.0	0.000000	0.0	0.00	0.000000	0.000000	0.000000	...	0.000000	0.0	0.0	0.0000

93 rows × 231 columns

And finally, before clustering we gather data frames:

In [194]:
```
final_df = queens_grouped.append(berlin_grouped)
```

```
C:\Anaconda3\lib\site-packages\pandas\core\frame.py:7123: FutureWarning: Sorting because non-concatenation axis is not aligned.
A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  sort=sort,
```

In [195]:
```
final_df.fillna(value=0, inplace=True)
```

Place Neighborhood column first

In [196]:
```
col = list(final_df.columns)
n = col.index('Neighborhood')
newcol = [col[n]] + col[:n] + col[n + 1:]
final_df = final_df[newcol]
```

In [197]:
```
final_df.head()
```

Out[197]:

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	...	Water Park	Waterfront	Weight Loss Center	Windmill
0	Arverne	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
1	Astoria	0.0	0.000000	0.0	0.0	0.010000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0

3.2 Clustering

We use machine learning algorithm of KMeans to cluster gathered neighbourhoods. After applying the method we get the next result, data frame with cluster labels:

	Cluster Labels	Neighborhood	ATM	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Water Park	Waterfront	Weight Loss Center	Windmill
0	1	Arverne	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
1	1	Astoria	0.0	0.000000	0.0	0.0	0.010000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
2	1	Astoria Heights	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
3	1	Auburndale	0.0	0.000000	0.0	0.0	0.055556	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
4	1	Bay Terrace	0.0	0.027027	0.0	0.0	0.054054	0.0	0.0	0.0	...	0.0	0.0	0.027027	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
88	3	(1207) Waidmannslust	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
89	12	(1208) Lübars	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
90	3	(1209) Wittenau	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
91	1	(1210) Märkisches Viertel	0.0	0.000000	0.0	0.0	0.083333	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0
92	0	(1211) Borsigwalde	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0

Now we can recommend some pattern of behaviour for any given Neighborhood. But if company is located in 'Cambria Heights', it can move to any Berlin Neighborhood cause there is no any similar:

```
In [201]: # First, check the cluster
          cluster_num = final_df[final_df['Neighborhood'] == 'Cambria Heights']['Cluster Labels'].values[0]
          cluster_num
Out[201]: 17

In [202]: final_df[final_df['Cluster Labels'] == cluster_num]
Out[202]:
```

	Cluster Labels	Neighborhood	ATM	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Water Park	Waterfront	Weight Loss Center	Windmill
16	17	Cambria Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
43	17	Laurelton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
73	17	St. Albans	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

3 rows × 350 columns

If company is located in South Ozone Park it can move to (0605) Dahlem Berlin Neighborhood. And it's one to one match.

4. Results

As a result of this project we achieved the next:
1) We gathered the data on Berlin and New York Neighborhoods from Foursquare;
2) We clustered Neighborhoods according to characteristics;
3) We can say if there is a similar Neighborhood in Berlin to a given Neighborhood in New York.

5. Discussion

We noticed next interesting things:
1) If New York Neighborhood is from the first cluster then it is very easy to find a matching Neighborhood in Berlin;
2) Some Berlin Neighborhoods are really different and therefor they produce the third cluster;
3) Even though cities are really different we can find some similar Neighborhoods.

We recommend:
1) If company is located at 'Cambria Heights' to move to any Berlin Neighborhood cause there is no any similar;
2) If company is located at 'South Ozone Park' to move to (0605) Dahlem Berlin Neighborhood. And it's one to one match;

We also can recommend some pattern of behavior for any given Neighborhood

6. Conclusion

In this project we clustered Neighborhoods of two different cities in order to recommend some Berlin Neighborhood for a company to move to.
We gathered the full peacture of the situation and now can recommend some pattern of behaviour for any given Neighborhood in New York or Berlin.

Further developement can be next:

1) Gather more information about Neighborhood to make clustering more accurate;
2) Gat new complex features from gathered ones for the same reason.