Business Problem

It is hard to pick hotels in cities like New York, as there are many options but there is not an easy way to rate the proximity of a hotel to places of interest. In this project, I categorized hotels in Manhattan, New York according to their proximity to “Art and Entertainment” venues, “Food” venues and “Night Life” venues.

Data

As the data source, I used Foursquare. Foursquare provides comprehensive data about places through its API. In this project, I started by getting a list of hotels in Manhattan, NewYork and proceeded by finding places of interest within a quick walking distance (250 metre), by using Foursquare API.

Methodology

1. First, I built a dataframe which stored list of hotels in Manhattan, New York, along with the number of “Art and Entertainment” venues, “Food” venues and “Night Life” venues within 250 metres:

```
manhattan_hotels=pd.merge(manhattan_hotels, venues_list_df, on='id')
manhattan_hotels
```

30]:

	id	name	lat	lng	num_of_art_ent_venues	num_of_food_venues	num_of_nightlife_venues
0	49d18dfdf964a5208f5b1fe3	The Plaza Hotel	40.764519	-73.974488	33	49	16
1	5093c236830214706abb75db	citizenM Hotel New York Times Square	40.761691	-73.984953	70	34	35
2	4ae6f117f964a520a6a721e3	The NoMad Hotel	40.744981	-73.988819	33	42	26
3	4bec60a5f909ef3b2808a9c6	Kimpton Hotel Eventi	40.747224	-73.989960	20	34	33
4	4a0e0f85f964a520bf751fe3	Ace Hotel New York	40.745858	-73.988121	31	52	45
5	4a67bbd8f964a520f9c91fe3	The Peninsula New York	40.761658	-73.975384	41	36	24
6	438d6b12f964a520322b1fe3	W New York - Times Square	40.759296	-73.985573	100	27	34
7	4ae0fa2ff964a5203d8421e3	Mandarin Oriental	40.768987	-73.983017	21	40	16
8	4a9f2ec3f964a520d73c20e3	Ink48 , A Kimpton Hotel	40.764505	-73.995987	2	25	11
9	4b58b0a5f964a520d66528e3	Distrikt Hotel	40.756707	-73.992873	51	30	26
10	4ac8d0d3f964a520b3bc20e3	The Carlyle	40.774413	-73.963301	23	18	5

Please note that Foursquare API returns 100 recommended hotels around the geographical location which is searched.

2. then, normalized the data:

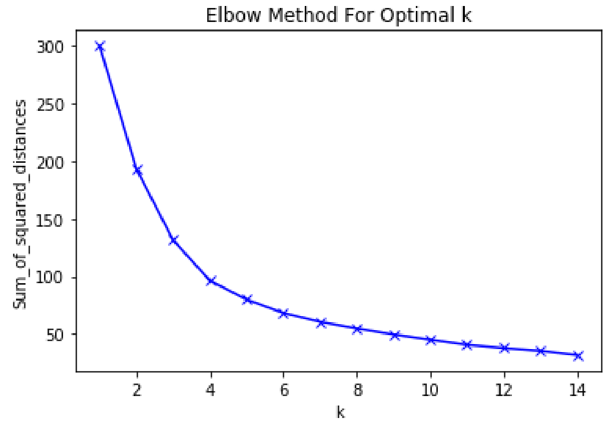
```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

X = manhattan_hotels.values[:,4:]
#X = np.nan_to_num(X)
cluster_dataset = StandardScaler().fit_transform(X)
cluster_dataset
#X
```

3. then, used elbow method to find the optimal K value:

```python
Sum_of_squared_distances = []
K = range(1,15)
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(cluster_dataset)
    Sum_of_squared_distances.append(km.inertia_)
    print(k)
    print(km.inertia_)

import matplotlib.pyplot as plt
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



Elbow method suggests that optimal value for K may be 4 or 5.
By manually evaluating both values according to resulted clusters, I decided to pick K=4.

4. The mean values, that is the average number of “Art and Entertainment” venues, “Food” venues and “Night Life” venues within 250 metres, for the 4 clusters turned out to be as follows:

```python
manhattan_hotels_grouped = manhattan_hotels.groupby('Labels').mean().reset_index()
manhattan_hotels_grouped
```

	Labels	lat	lng	num_of_art_ent_venues	num_of_food_venues	num_of_nightlife_venues
0	0	40.752463	-73.984131	23.000000	36.731707	20.951220
1	1	40.756813	-73.989170	76.944444	31.944444	26.777778
2	2	40.750741	-73.985681	27.722222	49.666667	34.722222
3	3	40.757904	-73.980874	18.217391	20.478261	8.434783

Results

The 4 hotel clusters formed can be named/defined as below:

- Cluster 0: **Decent** Cluster: Hotels in this cluster have decent number of venues form each group (Art & Entertainment, Food, Night Life) within walking distance.
- Cluster 1: **Art & Entertainment** Cluster: Hotels in this cluster have very high number of “Art & Entertainment” venues and also a reasonable number of food venues within walking distance.
- Cluster 2: **Foodie** Cluster: Hotels in this cluster have high number of “Food” venues and a good number of “Night Life” venues within walking distance.
- Cluster 3: **So-So** Cluster: Hotels in this cluster have less venues in each group group (Art & Entertainment, Food, Night Life) compared to other hotel clusters.

It is not surprising that hotels in cluster 1 are geographically located around Times Square.

Discussion

In real life, I usually encounter the problem stated in this project, it is hard to find an overall rating of a hotel according to the venues nearby.

In this analysis I simply focused on the number of venues within short distance of hotels. This analysis may further be improved by including ratings and pricings of both the hotels and venues, however for such data Foursquare premium membership might be required.

Conclusion

Hotels around Times Square offers more in terms of options nearby. Most likely this is reflected to prices.