DATA TELLS WHERE TO BUILD NEW HOTELS IN BERLIN, GERMANY

IBM Data Science Course

Capstone Project

INTRODUCTION

- The need for accommodation in Berlin is always high since this city is the major center of politics, culture, media and science.
- A new entrepreneur wants to ***set up a hotel chain*** in Berlin, so he wants to find the best and optimized places to build hotels.
- Tourists' sharing about their experience (likes, ratings) on social media (Foursquare) is an important data source to utilize

- Problem: *How to find the best possible places in Berlin to build up a new hotel chain in Berlin using the data retrieved from Foursquare?*

DATA SET DESCRIPTION

- The data required to solve the problem is the **Foursquare** data for interesting places in Berlin, Germany.
- Using Foursquare API to get the venues data within the radius as 10,000m from Berlin city center.
- We pass **query=tourist** into the URL to filter these values which will helpful for other tourists.

```
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}&offset={}
&query=tourist'.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, radius, LIMIT, offset)
```

	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng	name	neighborhood	postalCode	state
0	Behrenstr. 55-57	Opera House	DE	Berlin	Deutschland	NaN	188	[Behrenstr. 55-57, 10117 Berlin, Deutschland]	4adcda8af964a520bd4921e3	[{'label': 'display', 'lat': 52.51596828902978...	52.515968	13.386701	Komische Oper	Unter den Linden	10117	Berlin
1	Gendarmenmarkt	Concert Hall	DE	Berlin	Deutschland	Charlottenstr.	429	[Gendarmenmarkt (Charlottenstr.), 10117 Berlin...	4adcda7cf964a520584721e3	[{'label': 'display', 'lat': 52.513652, 'lng':...	52.513652	13.391911	Konzerthaus Berlin	NaN	10117	Berlin
2	Markgrafenstr.	Plaza	DE	Berlin	Deutschland	Mohrenstr.	466	[Markgrafenstr. (Mohrenstr.), 10117 Berlin, De...	4adcda7df964a5206a4721e3	[{'label': 'display', 'lat': 52.51357005399756...	52.513570	13.392720	Gendarmenmarkt	NaN	10117	Berlin
3	Bebelplatz	Plaza	DE	Berlin	Deutschland	Unter den Linden	342	[Bebelplatz (Unter den Linden), 10117 Berlin, ...	4adcda7df964a520ab4721e3	[{'label': 'display', 'lat': 52.51653012045096...	52.516530	13.393847	Bebelplatz	NaN	10117	Berlin
4	Am Festungsgraben 1	Theater	DE	Berlin	Deutschland	NaN	481	[Am Festungsgraben 1, 10117 Berlin, Deutschland]	4adcda89f964a52095492 1e3	[{'label': 'display', 'lat': 52.51894077519612...	52.518941	13.395245	Maxim Gorki Theater	NaN	10117	Berlin

DATA SET DESCRIPTION (CONT.)

- To retrieve venue likes and ratings, we pass venue ID for each location to URL and call the Foursquare API. The return data set combines of ratings and likes for each location.

```
def get_venue_likes_and_rating(venue_id):
    url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(venue_id,CLIENT_ID,
                                                                                             CLIENT_SECRET, VERSION)

    results = requests.get(url).json()
    try:
        likes = results['response']['venue']['likes']['count']
    except:
        likes = np.nan
        print('Venue {} has no like info'.format(venue_id))

    try:
        rating = results['response']['venue']['rating']
    except:
        rating = np.nan
        print('Venue {} has no rating info'.format(venue_id))

    return (likes,rating)
```

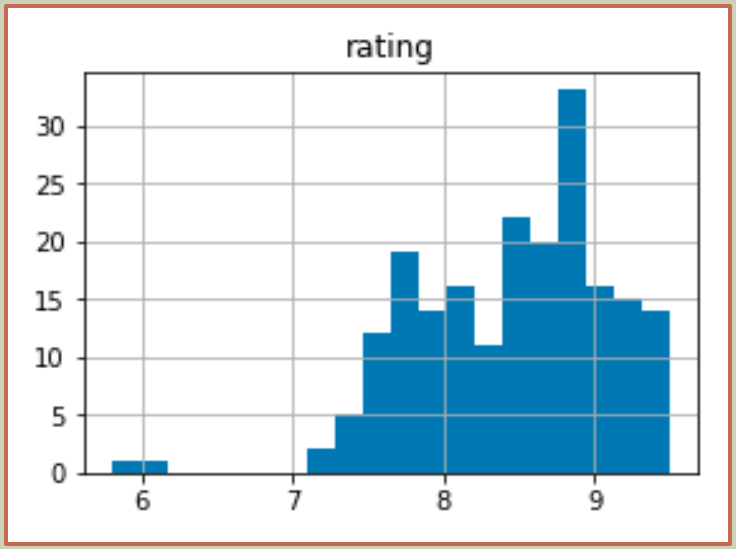
	address	categories	cc	city	country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng	name	neighborhood	postalCode	state	no_of_likes	rating
0	Behrenstr. 55-57	Opera House	DE	Berlin	Deutschland	NaN	188	[Behrenstr. 55-57, 10117 Berlin, Deutschland]	4adcda8af964a520bd4921e3	[{'label': 'display', 'lat': 52.51596828902978...	52.515968	13.386701	Komische Oper	Unter den Linden	10117	Berlin	182.0	8.7
1	Gendarmenmarkt	Concert Hall	DE	Berlin	Deutschland	Charlottenstr.	429	[Gendarmenmarkt (Charlottenstr.), 10117 Berlin...	4adcda7cf964a520584721e3	[{'label': 'display', 'lat': 52.513652, 'lng':...	52.513652	13.391911	Konzerthaus Berlin	NaN	10117	Berlin	327.0	9.2
2	Markgrafenstr.	Plaza	DE	Berlin	Deutschland	Mohrenstr.	466	[Markgrafenstr. (Mohrenstr.), 10117 Berlin, De...	4adcda7df964a5206a4721e3	[{'label': 'display', 'lat': 52.51357005399756...	52.513570	13.392720	Gendarmenmarkt	NaN	10117	Berlin	1293.0	9.4
3	Bebelplatz	Plaza	DE	Berlin	Deutschland	Unter den Linden	342	[Bebelplatz (Unter den Linden), 10117 Berlin, ...	4adcda7df964a520ab4721e3	[{'label': 'display', 'lat': 52.51653012045096...	52.516530	13.393847	Bebelplatz	NaN	10117	Berlin	176.0	8.8
4	Am Festungsgraben 1	Theater	DE	Berlin	Deutschland	NaN	481	[Am Festungsgraben 1, 10117 Berlin, Deutschland]	4adcda89f964a520954921e3	[{'label': 'display', 'lat': 52.51894077519612...	52.518941	13.395245	Maxim Gorki Theater	NaN	10117	Berlin	134.0	9.1

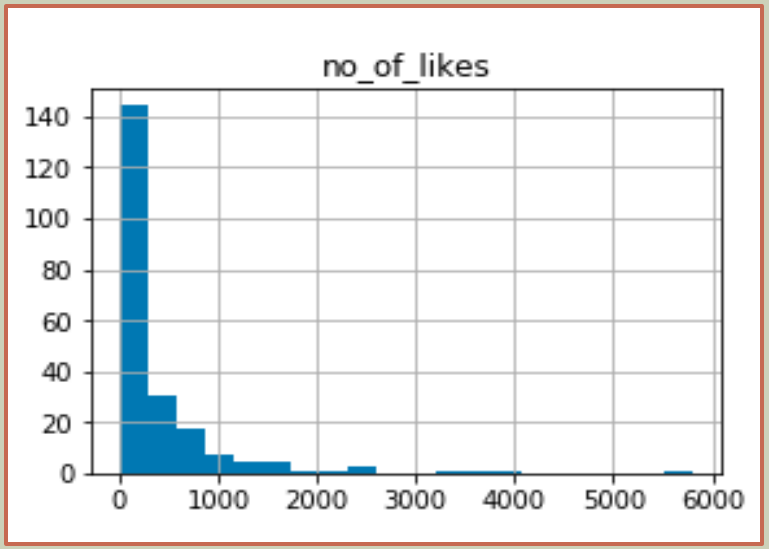
METHODOLOGY

- Clustering is a most common exploratory data analysis technique which often used to get an intuition about the data structure.
- ***K-means clustering*** is the unsupervised machine learning algorithm that we select for this project. k-means clustering is among the most popular and simplest unsupervised machine learning algorithms.
- To implement k-means clustering, we use ***sklearn*** library from python

RESULT

- Data exploratory analysis





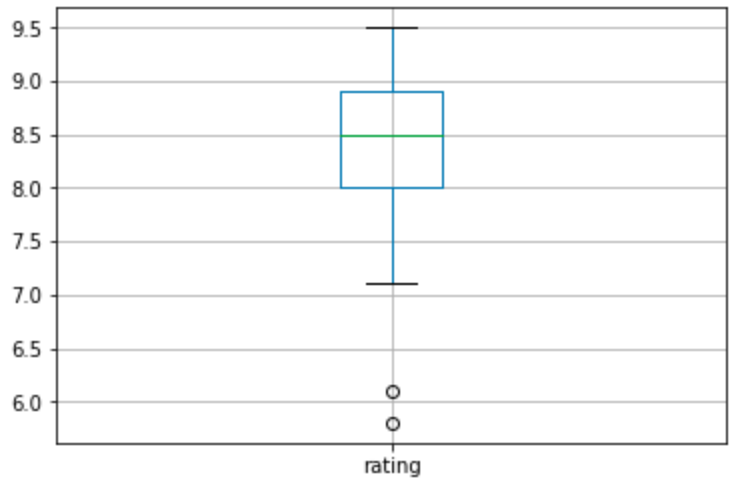
- The majority of places receiving the rating from travelers above 7.0.
- Most of travelers give the rating around 9.0 for places in Berlin

- The majority of places will receive around 200-500 likes from people who give feedbacks on Foursquare.

RESULT (CONT.)

- Exploratory analysis – Box plot

```
rates_df = explore_df[['rating']]
boxplot = rates_df.boxplot()
```

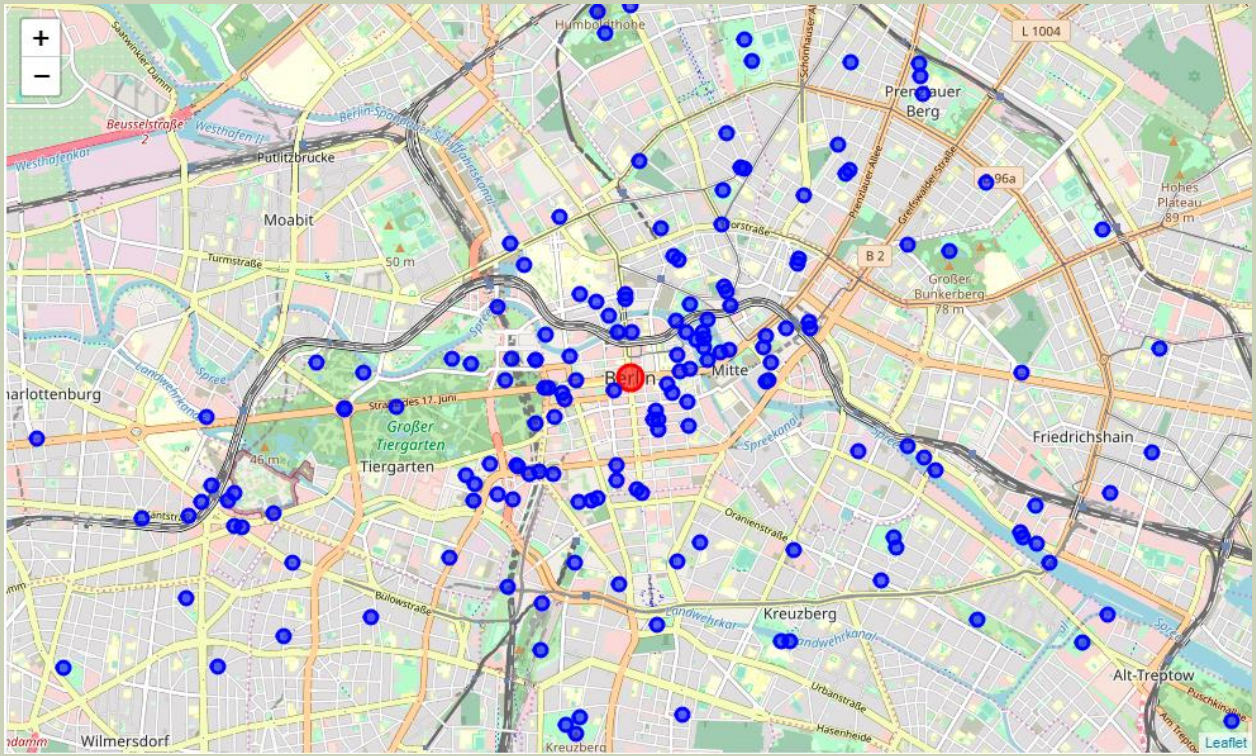


- Having outliers for rating
- Median is 8.5, first quartile is slightly above 7.0
- We decide to drop places which have rating less than 7.5

```
# Get names of indexes for which column rating has value less than 7.5
indexNames = explore_df[explore_df['rating'] < 7.5 ].index

# Delete these row indexes from dataFrame
explore_df.drop(indexNames , inplace=True)
```

RESULT (CONT.)

- Raw data visualization

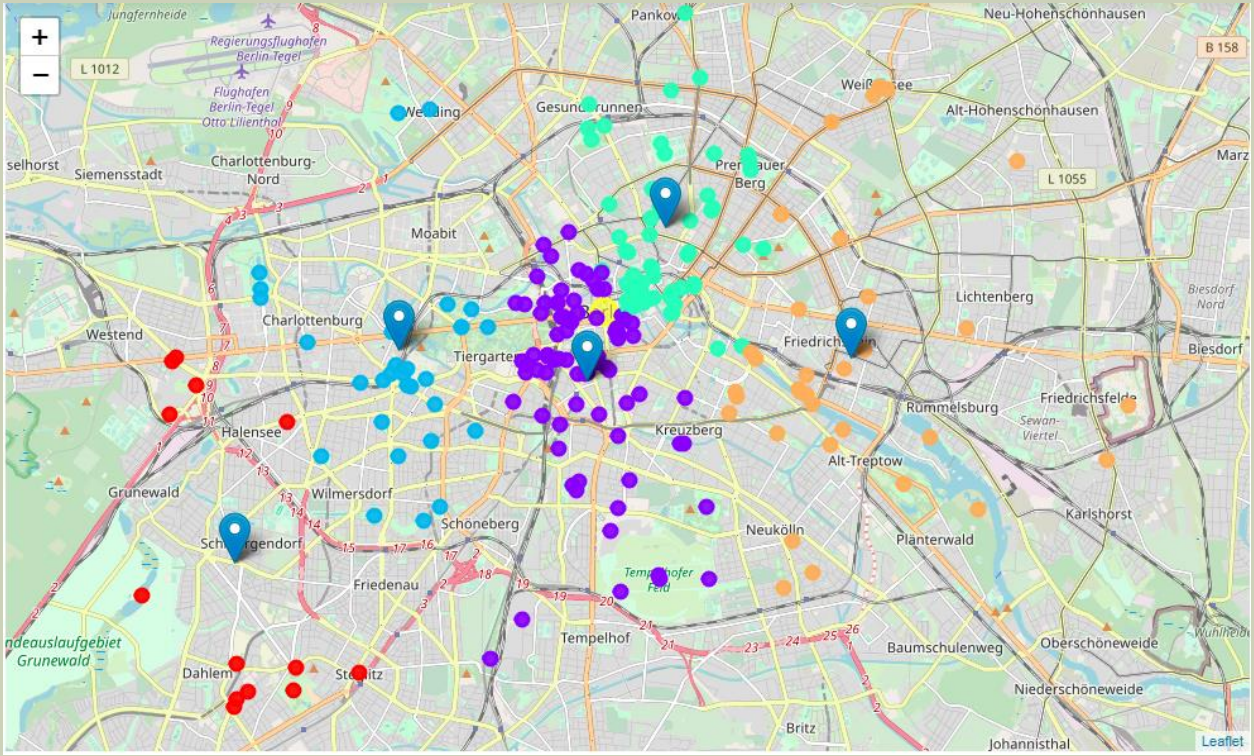


Observation:

- High concentration near city center
- No clearly visible cluster

RESULT (CONT.)

- K-means clustering with k = 5. The centroids of clusters are obtained using ***cluster_centers_*** attribute



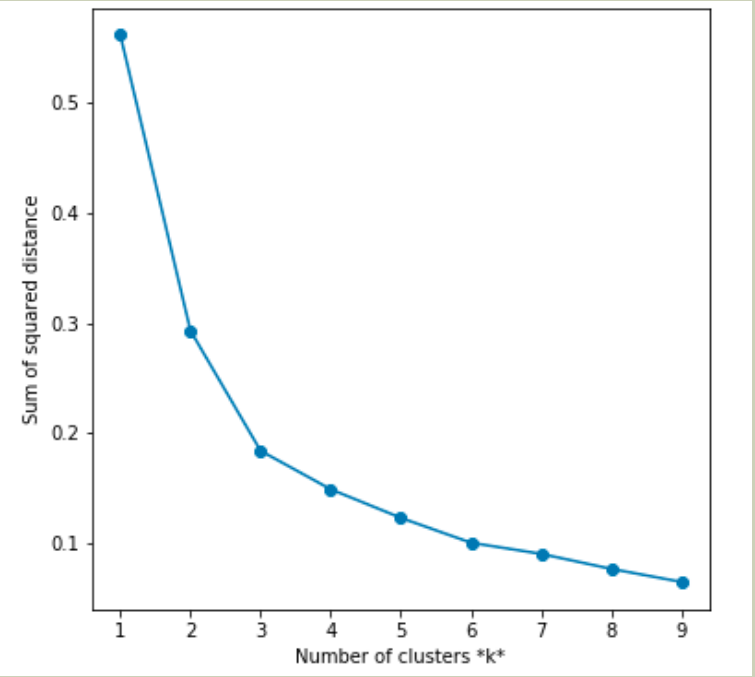
Observation:

- High concentration near city center
- The density of clusters are not equal, especially for the cluster in the West of city, quite a few places belong to this cluster (the red cluster)

DISCUSSION

- Model evaluation using Elbow method

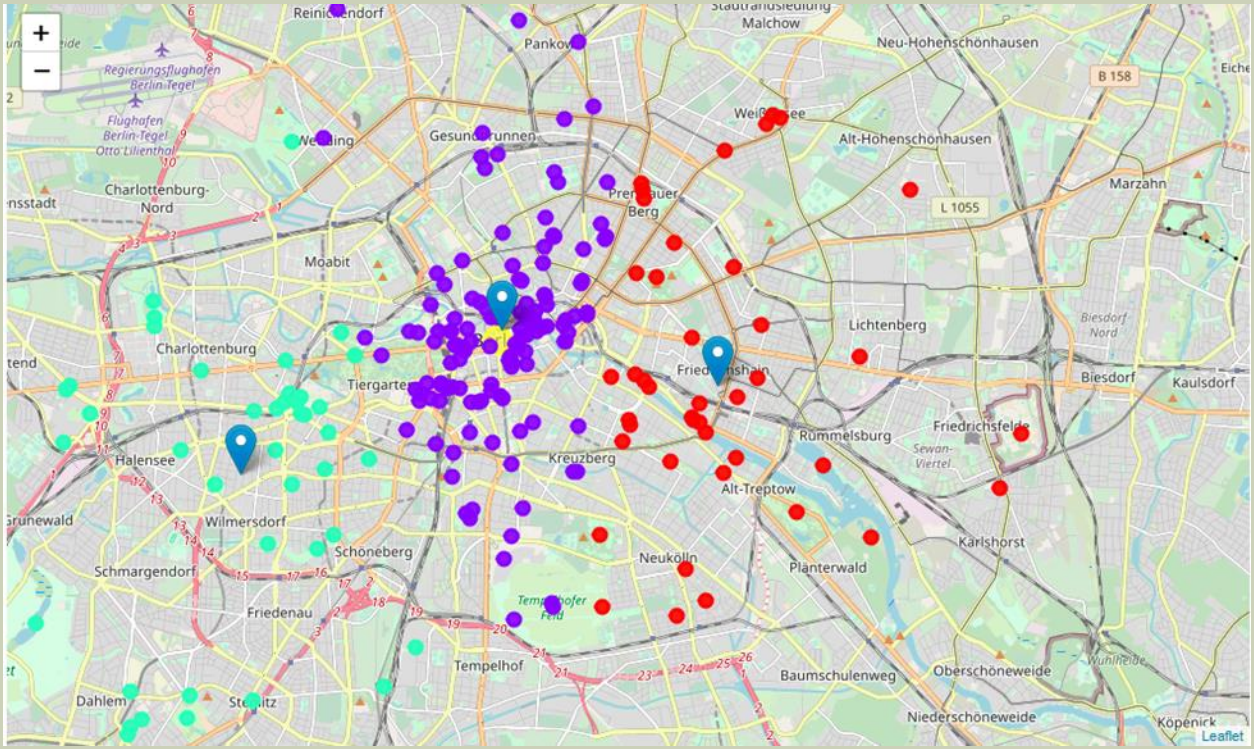
```
'''Run the Kmeans algorithm and
 get the index of data points clusters'''
import matplotlib.pyplot as plt
sse = []
list_k = list(range(1, 10))

for k in list_k:
    km = KMeans(n_clusters=k)
    km.fit(explore_df_clustering)
    sse.append(km.inertia_)

# Plot sse against k
plt.figure(figsize=(6, 6))
plt.plot(list_k, sse, '-o')
plt.xlabel(r'Number of clusters *k*')
plt.ylabel('Sum of squared distance');
plt.show()
```



- k =3 is possibly the best k to select for this project

DISCUSSION (CONT.)

- Re-run the k-means clustering with k = 3



<u>Observation:</u>

- More significant number of data points in each cluster
- Less clusters required, but the centroids (or the place for building hotels) can cover as effectively as 5 clusters.

CONCLUSION

- Successfully use k-means clustering algorithm to tackle the business problem
- Data retrieved from Foursquare about interesting places in Berlin are properly pre-processed, analyzed and applied the algorithm.
- K-means clustering model is evaluated by Elbow method to find the most optimal k value
- Re-build the new model with k value yield from Elbow graph
- Business decision made based on the outcome from the machine learning model