Scoring Hotels According to the Aspect of the Nightlife Concept

ONUR SAHIL CERIT

Problem Definition

Enrich and restructure the hotel search architecture with adding a new feature of nightlife aspect of the hotels.

There are 100 hotels from 4 different cities to rank by their nightlife concept.

Rank/Score them by using given features of cities and hotels for users to have the best nightlife experience possible.

Problem Approach

First, analyze the features of given hotels and the list of point of interests:

- Most suitable features that could contribute mostly to the nightlife score are, **distance_to_center(float-meters)**, **longitude/latitude**, **club_club_hotel(boolean)**, **party_people(boolean)**

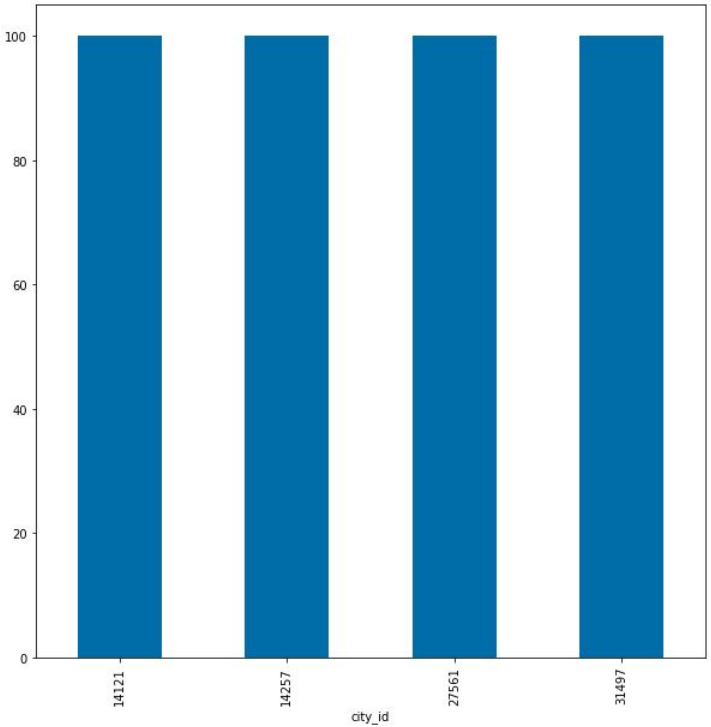
"**club_club_hotel**" and "**party_people**" features have *97.7%* and *81.7%* unknown information(nan value) which is quite high percentage. Thus, I ignored these features.

- Extract the list of unique point of interest types to eliminate the ones that are going to effect the nightlife score of the hotels. I picked "**Bar / Pub**" **,** "**Disco / Nightclub**" **,** and "**Casino**"

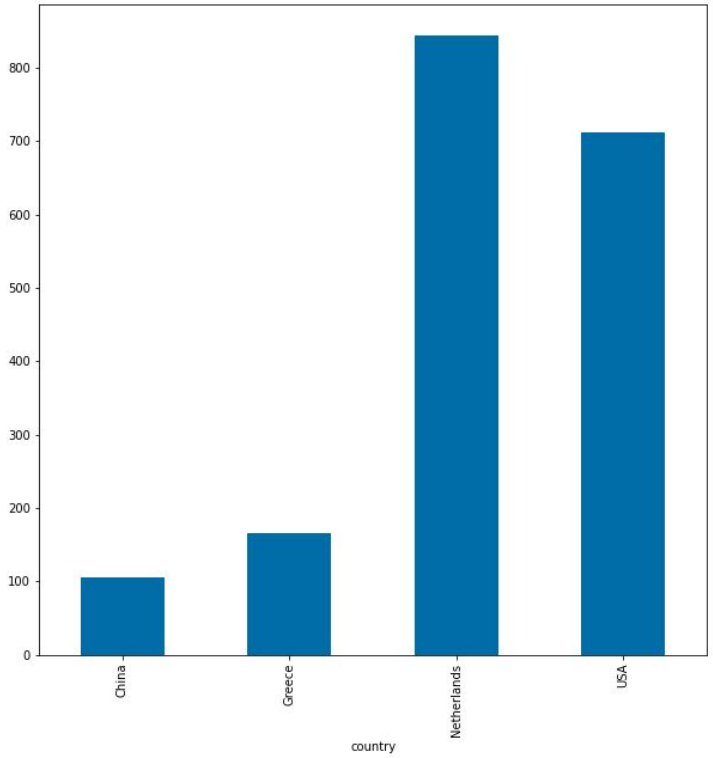
Note: There are other possible types of pois that could have impact on changing the score of the hotels, such as "**Restaurant**", "**Food / Drink**", etc. But, since there is an ambuguity of the description for these types, so that I did not use the location of them.

Problem Approach

Distribution of number of hotels per city



Distribution of number of selected pois per country



Problem Approach

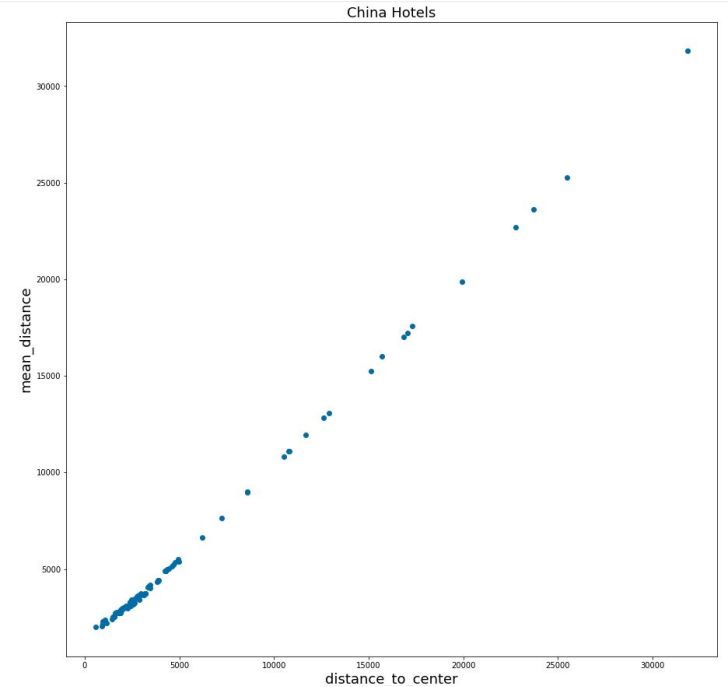
Second, split the hotels and pois datasets into cities:

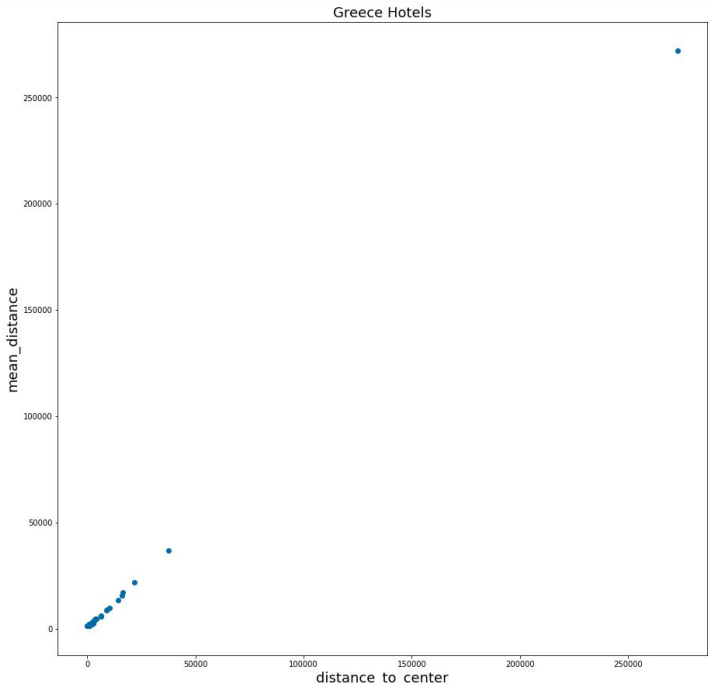
- Split each 100 hotel into their own cities. The reason to do this is that not to make a calculation mistake by mixing the hotels in different countries when determining the distance from each hotel to selected poi locations.

- Also, split the poi locations according to which city they belong to. This supports the distance calculation for each hotel.

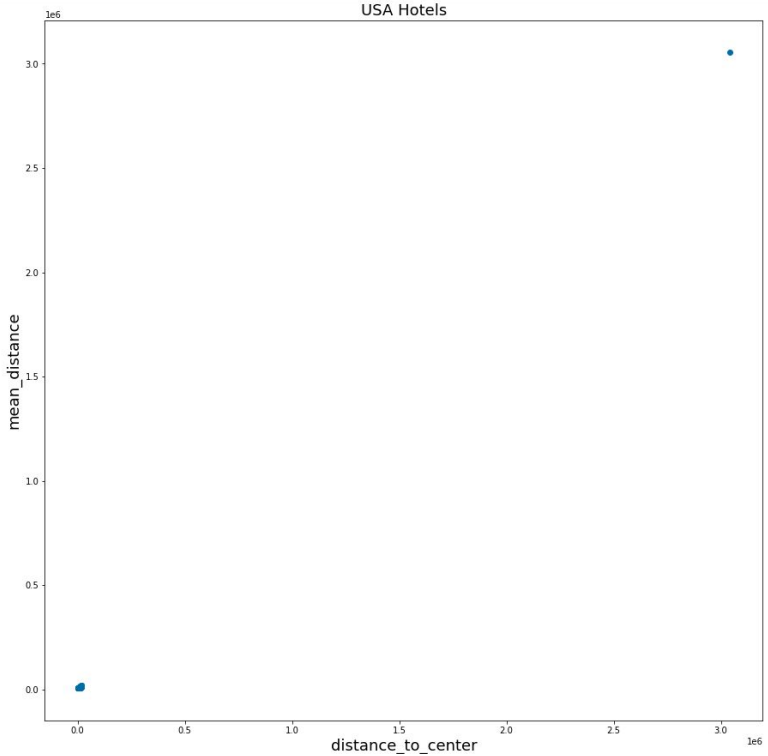
- Calculate the distance from the hotel in city A to all nightlife locations given for that city. Take the mean of distances to obtain average distance of that hotel to nightlife locations.

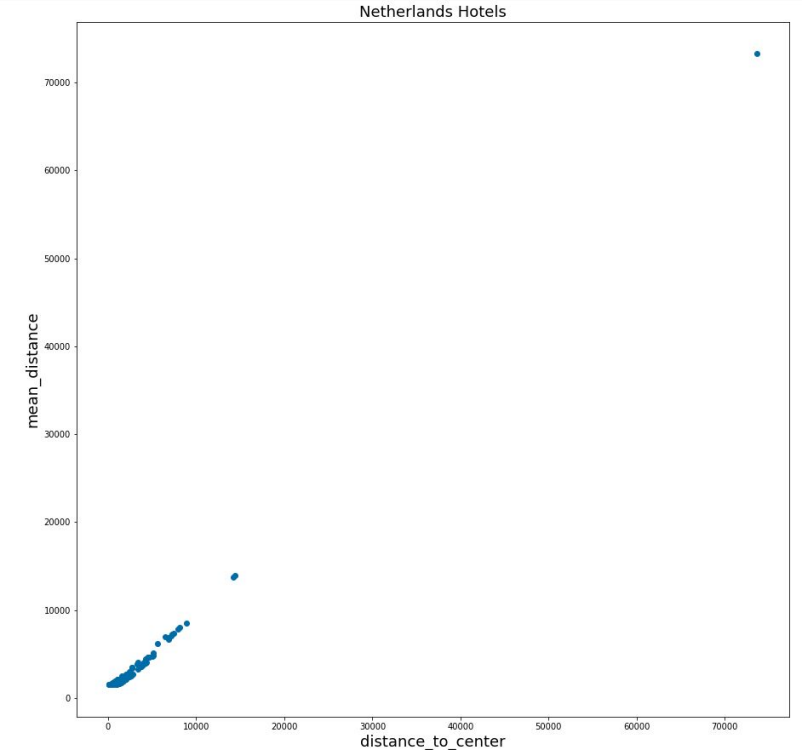
Problem Approach





Problem Approach





Problem Approach

Third, merge, winsorize, and cluster the hotels for scoring:

- After being done with the calculations for each city, I merged them to get one dataset of hotels with distances. From this point, each hotel is going to be clustered and scored regardless of their city, but by the "**mean_distance**" and "**distance_to_center**" center.

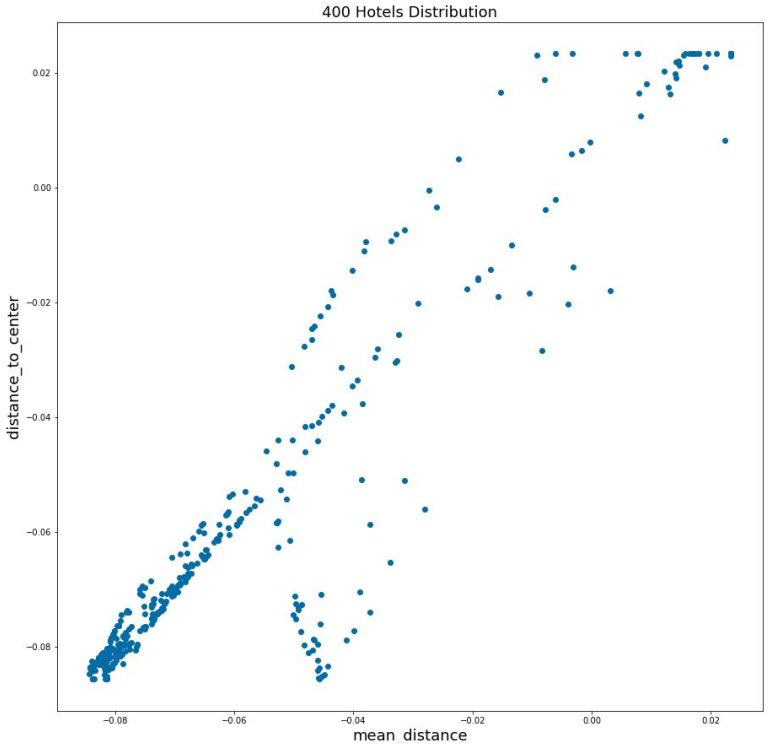
Note: The reason of picking the distance_to_center feature is because of its correlation with the mean_distance feature.

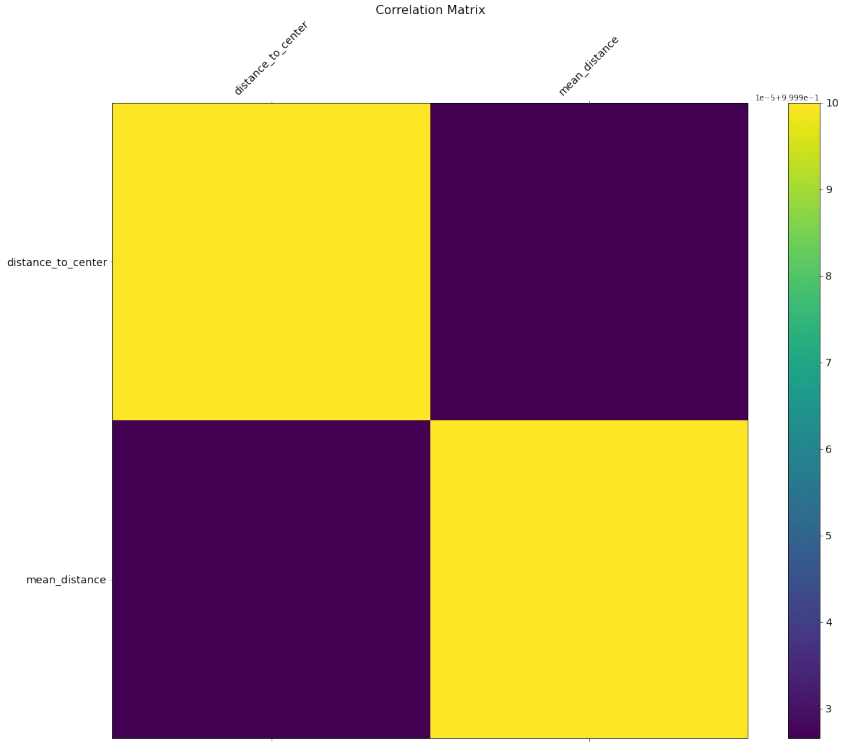
- As it was shown in previous plots, there are outliers in the data. Thus, to handle outliers I used winsorizing method. This limits the extreme data points and reduce the effect of outliers to the final result. I winsorized the 1%th quantile and the 5%th quantile.

- Finally to assign scores to hotels, I needed to group the hotels by the two distance feature and rank them. For this, I used k-means clustering algorithm. I clustered them into 100 groups, so that each group corresponds to a score. After clustering hotels into 100 groups, I scored them by the magnitude(distance) to their centroids.

Problem Approach

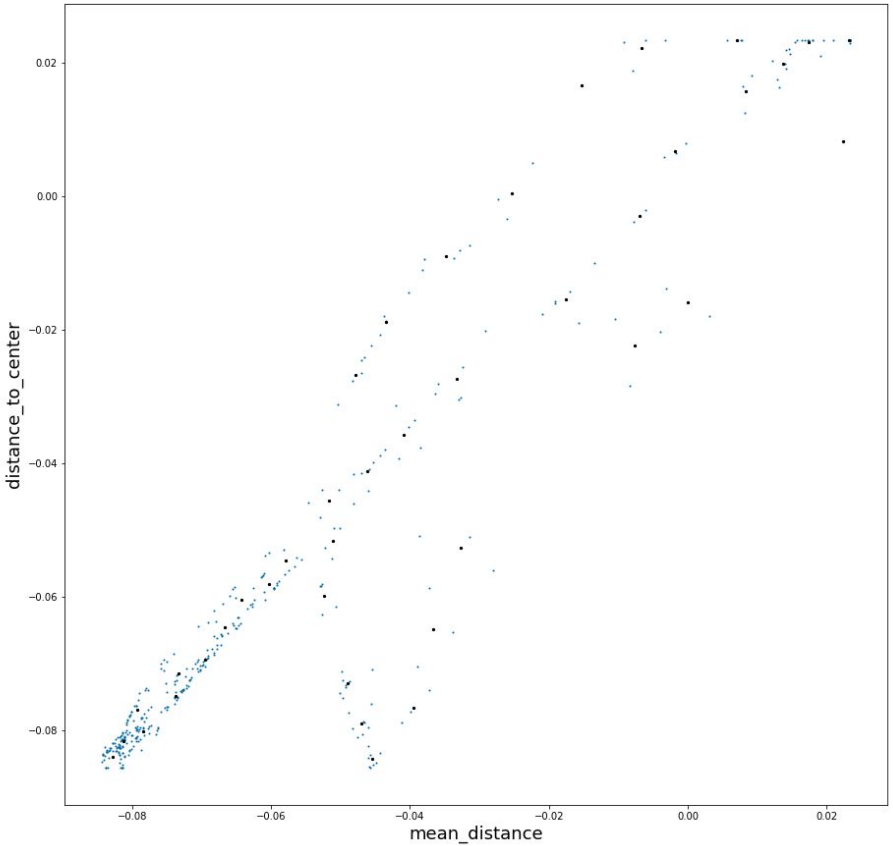
Hotels' mean_distance/distance_to_center scatter plot after winsorizing(handling outliers)





Problem Approach

Hotels Clustered



Validation of Scores

Silhouette Score and Calinski-Harabazs Score:

- I used silhouette score and calinski-harabazs score to validate my clustering results with different number of clusters.

#clusters = 40:

- Silhouette_score: 0.44191264992755674
- Calinski_score: 4660.001221331823

#clusters = 100:

- Silhouette_score: 0.47214150099452956
- Calinski_score: 10535.495935993727

=> As a result 100 gives better clustering result, as well as being a proper number of clusters since I rank the hotels by scoring them from 1 to 100.

Implementation & Testing

There are number of possible ways to implement the nightlife score feature.

1. Implement it for the cities where there are more nightlife locations, so that the calculations for nightlife score would be more concrete as the accuracy increases by the data given.

2. Implement it with a supporting criteria such as nightlife ranking for higher starred hotels or the entertainment hotels.

Implementation & Testing

There are number of possible ways to test this feature
and improve for better user experience.

1. Ratings from actual users that had been to those specific hotels.

2. Field research for the bars, clubs, pubs, casinos or possible venues that would effect the nightlife score in those cities and compare the results of the current scoring algorithm.

Maintenance & Development

There are features that could be included/obtained or
added for the future advancements of this feature

1. Survey from the users, local people, venue owners for the information about the nightlife locations inside the city.

2. To do a field research in these areas:
   2.a. Size of the places
   2.b. Quality of the places
   2.c. User rating of the places
   2.d. Menu prices of the places etc.

3. Apply different effective scoring algorithms possible, such as bidding algorithms for given features, comparing results with logistic regression model(fisher scoring), etc.