Karthiga Thangavelu

V00925048

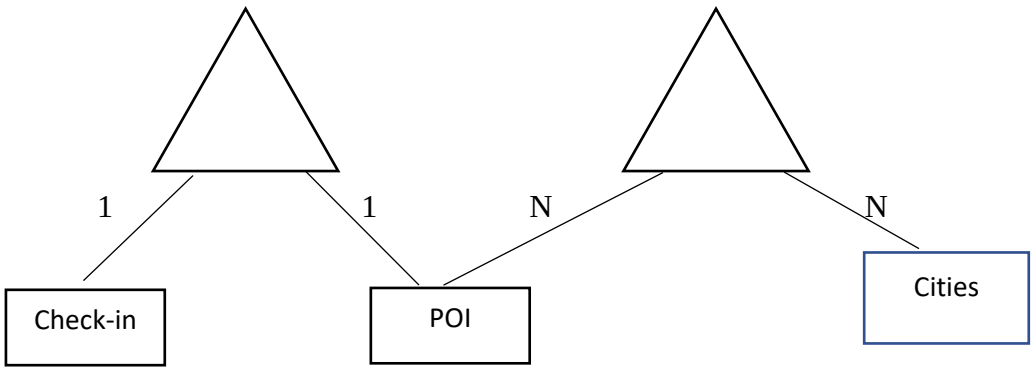
CSC – 501

Assignment -1

Contents:

1. Dataset:

The dataset used for this assignment was global-scale check-in data collected from FourSquare. The dataset which includes 18 months data which is from April 2012 to September 2013. The datasets are check-in dataset which has details such as user Id, venue Id by the user and UTC time with offset. The second dataset contains point of interest which has details such as venue Id and which country it belongs to. The third dataset is cities which contains cities and countries it belongs to. Considering this as three different tables and its relationship can be given as, the table check-in has Venue Id as a foreign key which is from the table points of interests. The point of interest table and cities table both as country code. It is possible to take the venues that belongs to countries and cities that belongs to countries. The cities table can be joined with point of interest table through country code. The table cities and check-ins cannot be joined. The relational schema for the given dataset can be represented as,



The relation between the table check-in and POI is many to many. The relation between table POIs and Cities is 1 to 1. Since one venue Id belongs to one city.

For this assignment, I have used Python, panda's library, Linux command, and Tableau for visualization.

Analyzing the dataset given, I have come up with questions that helps to find insight of Hospitality Industry in west cost of US. Taking 3 popular cities, I have answered the questions from analyzing the given data. The questions are:

1. Geographic Distribution of Hospitality Industry in Cities (How many Hotel visits in each category (near city, sub-urban, and outskirts) of the city?)
2. Hotel Stay in a Week (Which day of the week has most Hotel visit during summer and winter?)
3. Hotel Popularity (Which is the popular Hotel and which category it belongs to?)

2. Geographic Distribution of Hospitality Industry in Cities:

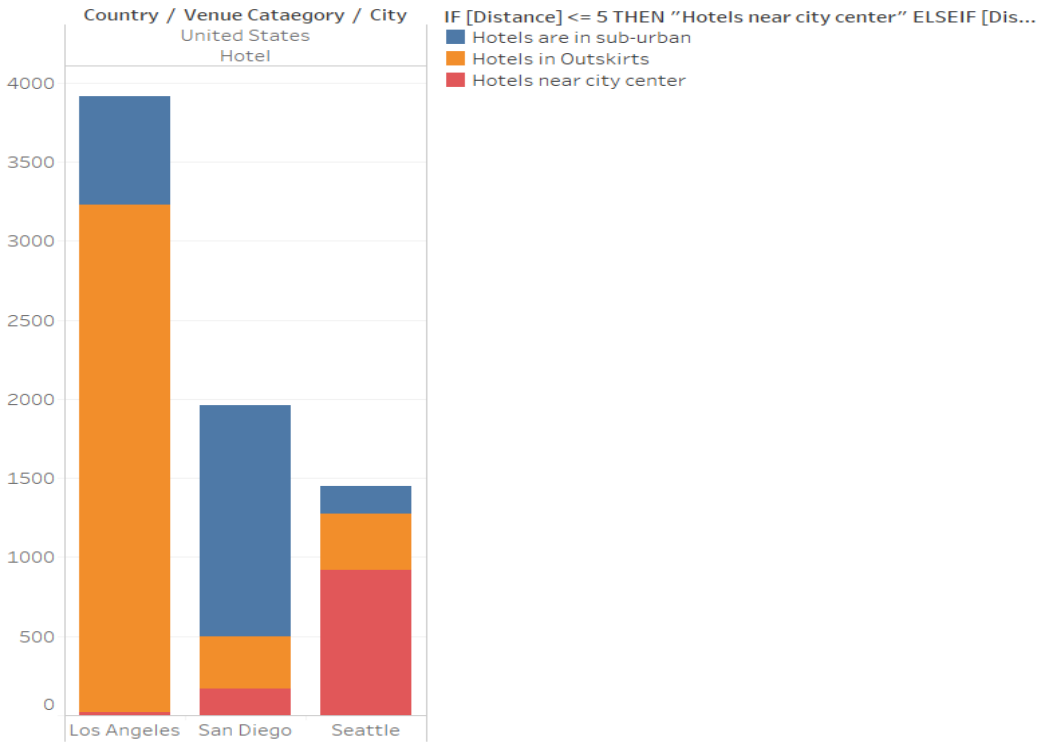
In Geographic Distribution, I have analyzed business in hospitality industry in west coast of US taking popular cities (Los Angeles, San Diego, and Seattle) and taking one venue category hotel. From this visualization, it is observed that most visited hotels were in Los Angeles comparing to other two cities

2.1.Expected:

In popular cities, it is expected that most visited hotels are in cities.

2.2.Actual insight:

The actual insight of these three countries are: In Los Angeles, the most visited hotels are in outskirts of cities. From this we can infer that the hotels in city are costly and there might be more tourist places in outskirts of city. Whereas, in Seattle most people visited hotels within city where the cost of hotels is cheaper as there is no tax. In San Diego, people mostly visited hotels which are in sub-urbans. As we see in this though these cities are all in west coast the check-ins of people vary from city to city depending on the cost of living of the city.



2.3.Modelling:

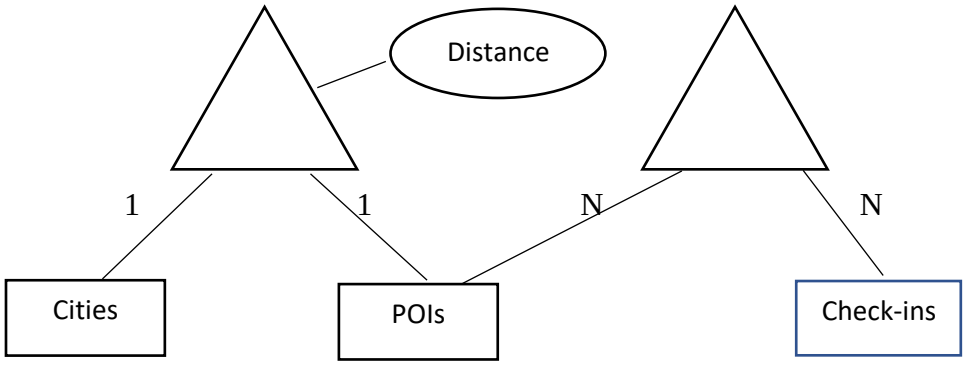
Spatial Modelling:

Here I have taken all three table Checkins, POIs, and Cities. By taking the longitude and latitude of the cities and venue calculated the distance between the city center and venue. Setting 20 Km radius as the criteria, mapped which venue belongs to which city. The venue Id which are 20 KM that belongs the city in that column which reduced most of the data which helps narrowing down to take only related data. Further to analyze data, I have split the data into within city, sub-urbans, and outskirts. Within city limit data are less than 5 km, sub-urban limit is between 5km and 10 km, and above 10km are outskirts.

I have used radial distance for spatial modelling to consider city limit as this would give equal distance in all side. But considering minimum bounding box for setting the city limit which will not be equal on all side and which leaves the points that belongs to the city. Grouping of points will not give same accuracy as radius which is taking all the points.

Relational Modelling:

Joined the POIs and Cities dataset calculating distance. Joined the data acquired from POIs and Cities with check-ins dataset.



Code for joining Cities and POIs:

```
for j,l,n,q in zip(country_UC_city['Latitude'], country_UC_city['Longitude'],co untry_UC_city['Name'],country_UC_city['Country_Name']):
        distance_city = distance.distance((i,k),(j,l))
        if distance_city <= 20:
            writer.writerow({'Venue_ID': p,'Distance': distance_city, 'Venune_Category_name': m, 'City_name': n, 'Country_name': q})
```

Command used for joining Cities and checkins:

```
join -1 1 -2 2 <(sort -k 1 places_US.txt) <(sort -k 2 dataset_TIST2015_Checkins.txt) -t $'\t' > result_hotel_checkins_US.txt &
```

Used join and sort command in Linux for joining two datasets. Stored the dataset in new table which extracted all the data requires for doing spatial analysis. The data is sorted and based on the key value give it maps with the other table to find the match and join. The unmatched data are not considered.

The result_hotel_checkins_US table which contains all the data from checkins table and places_US (join of POIs and Cities) table which are matched.

3. Hotel Stay in a Week:

From the given three cities (Los Angeles, San Diego, and Seattle) above, analyzing which day of the week was busy and number of visits by people during summer and winter.

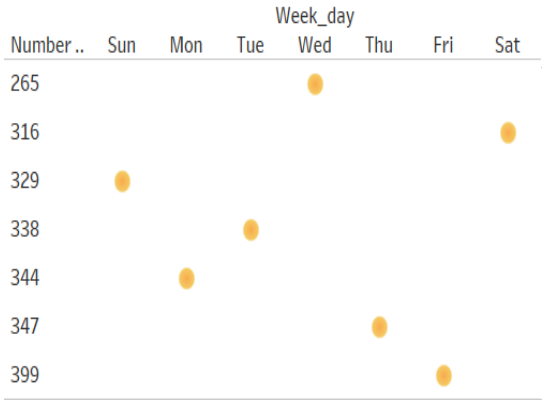
3.1.Expected:

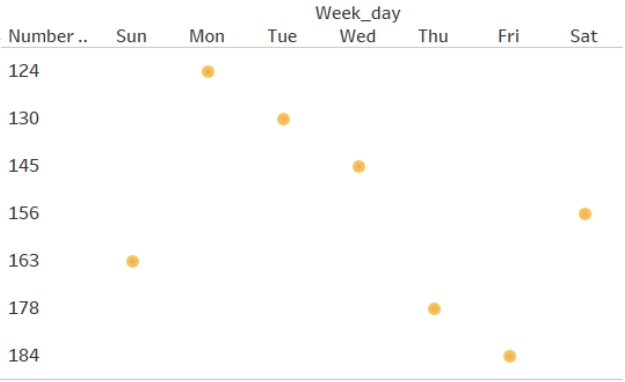
It is expected that there are more visits during summer than winter and more visit during weekends than weekday.

3.2.Actual Insight:

Observing the visualization, we can tell the visits are more during summer in Los Angeles. Also, there are more visit during weekends which are Friday and Saturday. In San Diego, which is like Los Angeles the visit are more during summer and weekend. Whereas, in Seattle number of visits during summer are same but from observation we can see there are busy during weekdays.

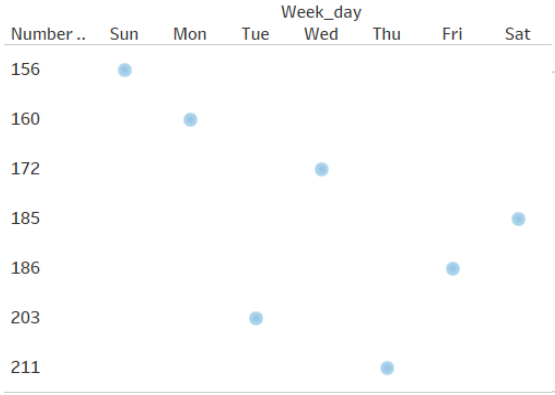
Los Angeles:

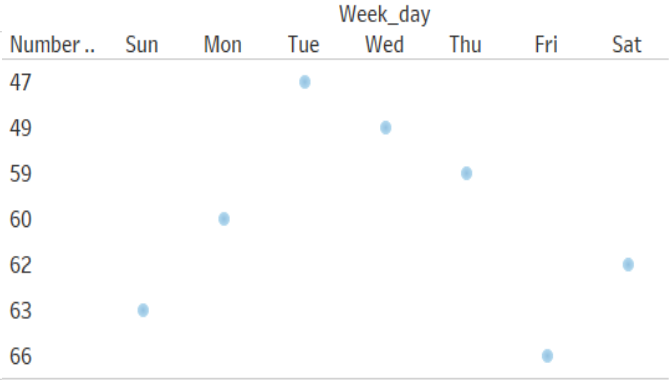




Summer Winter

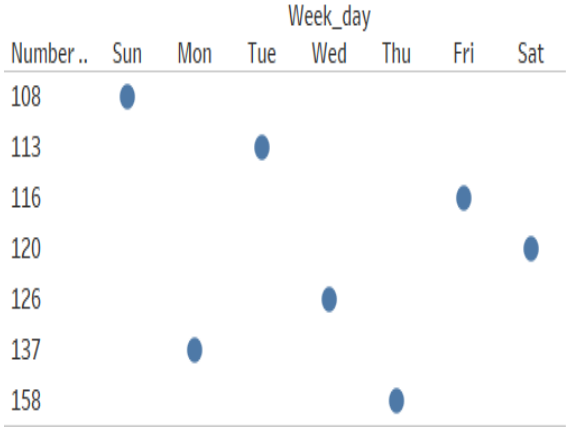
San Diego:

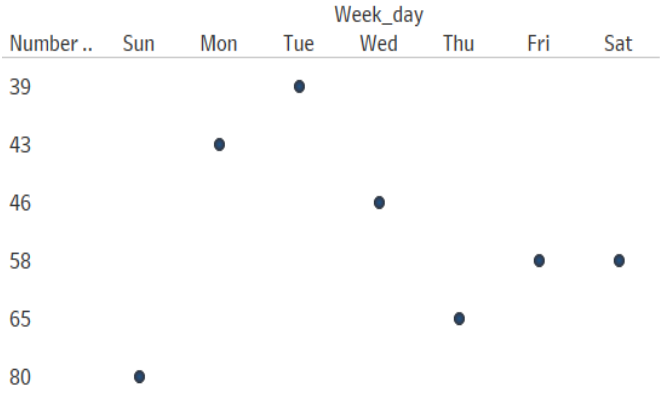




Summer Winter

Seattle:





Summer Winter

3.3.Modelling:

Temporal Modelling:

The above analysis is a spatio-temporal modelling. In which the user check-ins the place during desired week of the time. This is event based spatio-temporal modelling where the check-in happens during in a place with respect to time like weekday or weekend.

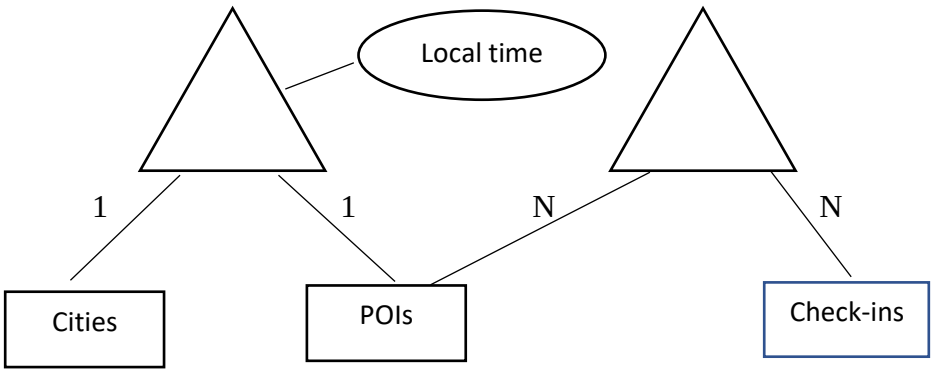
For this dataset, Joining the cities and POIs table and joined it with check-ins table. The time given in the check-ins table is UTC time and the offset also given. Converting UTC time to local time, the total number of check-in weekly is counted. Observing the table, it seems that same city has two offset times.

4abed6b4f964a520339020e3	13.031854282139234 km	Hotel	Los Angeles	United States	136877	Tue May 22 14:29:01 +0000 2012	-420
4abed6b4f964a520339020e3	13.031854282139234 km	Hotel	Los Angeles	United States	53525	Tue Nov 20 20:29:57 +0000 2012	-480

From this table we could see for November the offset is -480 and for May the offset is -420. Which is observed as daylight saving.

The local time converted data is further filter into winter and summer. Here, I have considered months for winter are Nov, Dec, Jan, and Feb and months for summer are Apr, May, Jun, and July.

Relational modelling:



Similar to the previous relational modelling, the cities and POIS table are joined. The new column local time is added to the table and it is joined with check-in.

Code for converting to local time:

```
date_time = datetime.datetime.strptime(i,"%a %b %d %H:%M:%S %z %Y").timestamp()
date_time1 = datetime.datetime.utcfromtimestamp(date_time+(j*60.0))
```

Code for taking only weekly check-in count:

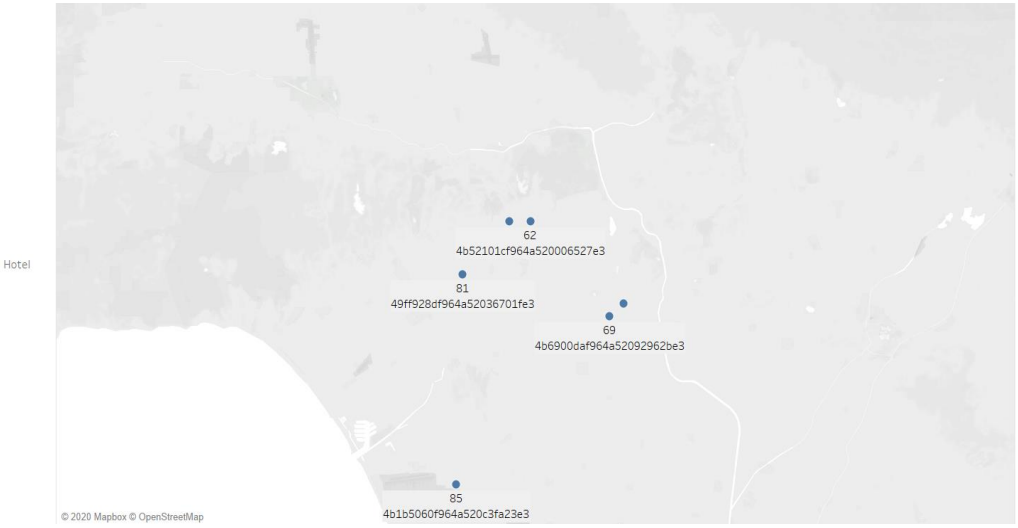
```
date_time = datetime.datetime.strptime(i,'%Y-%m-%d %H:%M:%S')
day = date_time.weekday()
week_dict[day] = week_dict[day] + 1
```

4. **Hotel Popularity:**

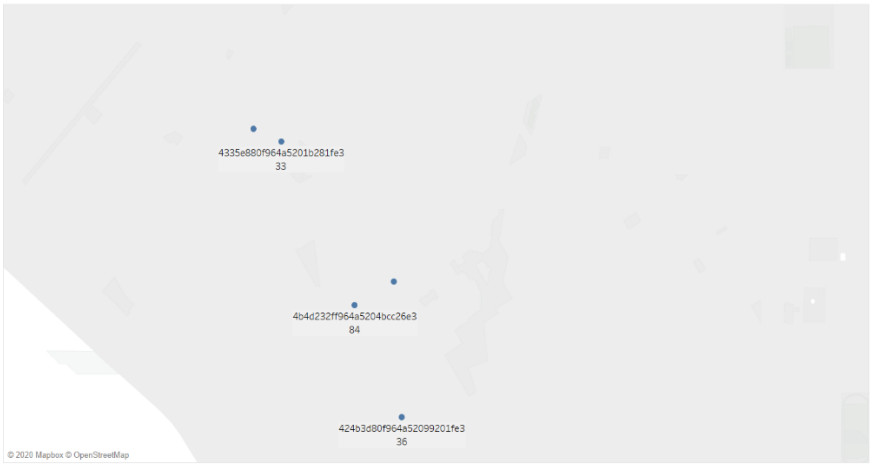
From the above cities, here, I have taken two cities which is Los Angeles and Seattle, also, taken summer to analyze the data. Here, I have shown the most visited venue in two cities during summer.

In Los Angeles, '**438cf623f964a520302b1fe3**' Venue Id is the most visited hotel, the total visits by the people is 103 and which is in outskirt of the city. In Seattle, the most visited hotel is '**4b4be2e5f964a52080aa26e3'** which is located outskirts of the city. Most of the hotel are in visits are in city, though the most visited hotel is in outskirt.

Los Angeles



Seattle:



Venue with most check-ins

Here, I have used the tables that are extracted from above questions.

Code for number of check-ins count:

```
for venueId in to['Venue_name']:
    if venueId in venue_count:
        venue_count[venueId] = venue_count[venueId] + 1
    else:
        venue_count[venueId] = 1
sort_venue_category_count = sorted(venue_count.items(), key=lambda x:x[1],reverse=True)
tmp_sorted=sort_venue_category_count[0]
all_venue_count = {}
for venue_category in tmp_sorted:
    all_venue_count[venue_category] = {}
    venue_count = all_venue_count[venue_category]
    for (venueId, ven_cat) in zip(to['Venue_ID'],to['Venue_name']):
        if venue_category == ven_cat:
            if venueId in venue_count:
                venue_count[venueId] = venue_count[venueId] + 1
            else:
                venue_count[venueId] = 1
final_sort = []
for sort_venue_cat in all_venue_count.keys():
    sort_venue_Id_count = sorted(all_venue_count[sort_venue_cat].items(), key=lambda y:y[1],reverse=True)
    sort_top3 = sort_venue_Id_count[0:6]
    final_sort.append(sort_top3)
```

I have filtered according to cities and month and taken only two cities using filter in dataframe.

From the above questions it is observed that Los Angeles has the most visited Hotels which are in outskirts of city. Weekends are busier comparing to weekdays and wintertime the check-in are less. The most visited hotel is also located in outskirts of city. In case of Seattle, though the most check-ins are in the city, the most visited hotel is in outskirt of the city. Also, there are lot of check-ins in weekdays than weekends