Project ID: P25
Project Title: Tourist Trajectories Bangkok
Group ID: 5
Team Members with ID numbers: Ashutosh Wagh(2022H1030052H), Kushal Chakraborty(2022H1030089H), S Shashank(2022H1030067H), Srinidhi P Katte(2022H1030075H)

1. **Background of the Dataset:**

   This dataset is about tourist trajectories in Bangkok during the Songkran Festival period, which took place from April 11th to April 17th, 2019. It contains information related to the movements and activities of tourists in the city during this time frame.

   The dataset was collected using GPS data from taxi rides. It is stored as a CSV file where each entry represents a tourist trajectory consisting of semantic information including trip information, place information, and trajectory type.

   Fields descriptions:
   - Vehicle ID: Unique vehicle ID
   - lat / lon: GPS location up to 5 decimal places
   - date_time: GPS timestamp (GMT+7)
   - category: place category at taxi origin/destination points
   - subcategory: place subcategory at taxi origin/destination points distance:
   - distance between taxi origin point and taxi destination point (km.)
   - time_used: traveling time
   - tourist_type: tourist trajectory based on tourist expense
   - trajectory_type: tourist trajectory based on tourist activity

2. **Literature Review**

1. **Paper Title and Authors:** GIS Based Route Network Analysis for Tourist Places: A Case Study of Greater Imphal, T. Prameshwori, Wangshimenla. J, L. Surjit, L. Ramananda

   Summary: This study uses ArcGIS network analysis to boost tourism development by establishing efficient travel routes, providing travel directions, locating adjacent facilities, and determining service regions by journey time directions, locating adjacent facilities, and determining service regions by journey time and distance. GIS technology aids tourist site visitation and decision-making.

2. **Paper Title and Authors:** Sentiment-Aware and Personalized Tour Recommendation, Prarthana Padia, Kwan Hui Lim, Jeffrey Cha, Aaron Harwood

   Summary: Visit frequency or photo counts are used to determine tourist interests in existing methods. A novel sentiment-aware personalized tour planner is presented in this research. This method analyzes tourists' written comments about areas they've visited to determine their interests.

3. **Paper Title and Authors:** Cruise passengers' behavior at the destination: Investigation using GPS technology, Stefano De Cantis, Mauro Ferrante, Alon Kahani, Noam Shoval

   Summary: This paper focuses on segmenting cruise passengers based on their behavior at the destination port, using a combination of traditional surveys and GPS technology. The goal is to create indicators to analyze how passengers move around and spend their time at the destination. Data from the Port of Palermo is used for this analysis. The results of the study identify seven distinct activity patterns among cruise passengers at the destination.

4. **Paper Title and Authors:** Identification of Optimum Path for Tourist Places Using GIS-Based Network Analysis: A Case Study of New Delhi, N. Gill, B.D. Bharath

   Summary: This paper aims to explore GIS-based network analysis to optimize routes for tourists. The aim is to determine the optimal routes from the starting point to various tourist destinations, considering the time spent at each location. Route analysis is based on impedance, which can be either time or length of the road network.

3. Experimental Dataset

a. Exploratory Data Analysis:

1) Exploratory Data Analysis

a) Dataset Characteristics

Observation:

Number of rows in the dataset : 33029

1.1) Data Preparation

a) Check for missing values

Observation

Number of missing values or NAN or NULL values : No missing values

1.2) Analysis of Feature 'Distance' from the perspective of 'Trajectory_Type

a) Initial Data Observations

Observations of Part 1

1) For 'trajectory_type' FoodAndDrinkTrip mean distance between source and destination is 5914.376177

2) For 'trajectory_type' LeisureTrip mean distance between source and destination is 3902.376744

3) For 'trajectory_type' NightlifeTrip mean distance between source and destination is 5308.597369

4) For 'trajectory_type' ReligiousTrip mean distance between source and destination is 4474.611502

5) For 'trajectory_type' ShoppingTrip mean distance between source and destination is 5364.915328

6) For 'trajectory_type' SpaTrip mean distance between source and destination is 5281.356976

Observations of Part 2

1) For 'trajectory_type' FoodAndDrinkTrip minimum distance between source and destination is 501

2) For 'trajectory_type' LeisureTrip minimum distance between source and destination is 643

3) For 'trajectory_type' NightlifeTrip minimum distance between source and destination is 509

4) For 'trajectory_type' ReligiousTrip minimum distance between source and destination is 525

5) For 'trajectory_type' ShoppingTrip minimum distance between source and destination is 503

6) For 'trajectory_type' SpaTrip minimum distance between source and destination is 502

Observations of Part 3

1) For 'trajectory_type' FoodAndDrinkTrip maximum distance between source and destination is 397390

2) For 'trajectory_type' LeisureTrip maximum distance between source and destination is 23940

3) For 'trajectory_type' NightlifeTrip maximum distance between source and destination is 610606

4) For 'trajectory_type' ReligiousTrip maximum distance between source and destination is 78652

5) For 'trajectory_type' ShoppingTrip maximum distance between source and destination is 997684

6) For 'trajectory_type' SpaTrip maximum distance between source and destination is 792990

Inferences drawn from the Observations

During their tour or trip in Bangkok, tourists have travelled long distances for FoodAndDrinkTrip, NightLifeTrips and ShoppingTrip. Even For SpaTrip, Significant distances were covered.

1.3) Analysis of Feature 'Time_Used' from the perspective of 'Trajectory_Type'

a) Initial Data Observations

Observation of Part 1 Analysis

1) For 'trajectory_type' FoodAndDrinkTrip mean time required to reach destination from source is 26.218770

2) For 'trajectory_type' LeisureTrip mean time required to reach destination from source is 29.271209

3) For 'trajectory_type' NightlifeTrip mean time required to reach destination from source is 21.695671

4) For 'trajectory_type' ReligiousTrip mean time required to reach destination from source is 21.954906

5) For 'trajectory_type' ShoppingTrip mean time required to reach destination from source is 22.408091

6) For 'trajectory_type' SpaTrip mean time required to reach destination from source is 24.069492

Observation of Part 2 Analysis

1) For 'trajectory_type' FoodAndDrinkTrip minimum time required to reach destination from source is 1.75

2) For 'trajectory_type' LeisureTrip minimum time required to reach destination from source is 3

3) For 'trajectory_type' NightlifeTrip minimum time required to reach destination from source is 1.05

4) For 'trajectory_type' ReligiousTrip minimum time required to reach destination from source is 2.57

5) For 'trajectory_type' ShoppingTrip minimum time required to reach destination from source is 1.98

6) For 'trajectory_type' SpaTrip minimum time required to reach destination from source is 1.85

Observation of Part 3 Analysis

1) For 'trajectory_type' FoodAndDrinkTrip maximum time required to reach destination from source is 1114.95

2) For 'trajectory_type' LeisureTrip maximum time required to reach destination from source is 941.65

3) For 'trajectory_type' NightlifeTrip maximum time required to reach destination from source is 1219.85

4) For 'trajectory_type' ReligiousTrip maximum time required to reach destination from source is 663.45

5) For 'trajectory_type' ShoppingTrip maximum time required to reach destination from source is 1009.05

6) For 'trajectory_type' SpaTrip maximum time required to reach destination from source is 1439.90

Inference drawn from the Analyses

During their tour or trip in Bangkok, tourists spent a significant amount of time for NightLifeTrip, FoodandDrinkTrip, SpaTrip and Shopping Trip.

Analysis of Figure 1

Observation

It is a highly left skewed graph with a fairly long thin tail. It can be observed that most of the trip durations approximately lie between 10 to 30 minutes

Inference

The trip durations are somewhat of medium duration. The durations are neither too short nor too long.

1.4) Analysis of Distance and Time_Used feature together

a) Initial Data Observations

Observation of Part 1 Analysis

1) There are 33029 values for 'distance' and 'time_used' features
2) The mean value for feature 'distance' is 5457.901632
3) The mean value for feature 'time_used' is 23.855
4) The minimum value for feature 'distance' is 501
5) The minimum value for feature 'time_used' is 1.05
6) The maximum value for feature 'distance' is 997684
7) The maximum value for feature 'time_used' is 1439.9
8) 25% of value in 'distance' feature is <=1819 ie it is the 25th percentile
9) 50% of value in 'distance' feature is <=3423 ie it is the 50th percentile or median
10) 75% of value in 'distance' feature is <=6513
11) 25% of value in 'time_used' feature is <=8 ie it is the 25th percentile
12) 50% of value in 'time_used' feature is <=13.2 ie it is the 50th percentile or median
13) 75% of value in 'time_used' feature is <=24

Analysis of Figure 2

Observation

This graph indicates that small distances between source and destinations is taking more time to travel.

Inferences

From this we can infer that the path leading to tourist destinations may have high traffic and congestion.

1.5) Analysis of Source and Destination based Features

a) Initial Data Observations

Observation of Part 1 Analysis

1) For 'category_o' feature there is only 1 value i.e 'Accomodation'. So this feature is not very useful.

2) For 'subcategory_o' feature there are 12 unique values i.e all the 33029 data points will have one of the 12 values as its 'subcategory_o'.

3) For 'category_d' feature there are 6 unique values i.e all the 33029 data points will have one of the 6 values as its 'category_d'.

4) For 'subcategory_d' feature there are 146 unique values i.e all the 33029 data points will have one of the 12 values as its 'subcategory_d'.

5) For Feature 'subcategory_o' value 'Hotel' has the highest frequency i.e. 24668

6) For Feature 'subcategory_d' value 'Spa' has the highest frequency i.e. 9432

7) For Feature 'category_d' value 'Restaurant' has the highest frequency i.e. 9982

Analysis of Figures 3(a) and 3(b)

Observation

1) The Sources are not very far from the Destinations and vice-versa.

2) The Sources and Destinations are clustered around in almost the same region.

1.6) Analysis of Source Based Features

Analysis of Figure 4 and Figure 5

Observation

1) From the above barplot and pie chart we can see that for the feature 'subcategory_o' Hotel has the highest occurence followed by Hostel then Residential Building and others.

Inference

From this we can infer that most of the tourists of Bangkok prefer to reside in Hotels and Hostels.

Analysis of Figures: Figure 6 (a) and Figure 6(b)

Observation

1) From the above two plots we can observe that the sources of origin of the trip is mostly clustered between 13.7 and 13.8 degrees latitude, 100.5 and 100.7 degree longitude.

2) The places of origin of the trip are mostly clustered in a certain radius almost to the central region of Bangkok.

Inference

From this we can infer that the tourists can search for their place of residence in central Bangkok. They need not search for the residence in several different locations of Bangkok. They can search for it only within central Bangkok around Phra Pradaeng and Bang Kruai region. We can infer this as central Bangkok has highly clustered region of places of origin.

Analysis of Figures: Figure 7 (a) and 7(b)

Observation

Hotels form large and dense clusters between 13.75 and 13.65 degrees latitudes and 100.4 and 100.7 degrees longitude.

Analysis of Figures: Figure 8 (a) and 8 (b)

Observation

Hostels are also present in the same region as Hotels. They also form clusters but the clusters are not as dense as hotels.

Analysis of Figures: Figure 9 (a) and 9 (b)

Observation

Residential Building (Apartment / Condo) are also present in the same region as Hotels. They also form clusters but the clusters are not as dense as hotels and are also wide spread.

Analysis of Figures: Figure 10 (a) and 10 (b)

Observation

Number of Resorts are much less as compared to Hotels and Hostels. The clusters are Very less dense.

Analysis of Figures: Figure 11 (a) and 11 (b)

Observation

Number of Inns are much less as compared to Hotels, Hostels and Resorts. The clusters are very less dense which indicates that the inns are not very popular in this region.

Analysis of Figures: Figure 12 (a) and 12 (b)

Observation

Number of Inns are much less as compared to Hotels, Hostels and Resorts. The clusters are very less dense which indicates that the inns are not very popular in this region.

Analysis of Figures: Figure 13 (a) and 13 (b)

Observation

Number of Bread and BreakFast based lodges are much less as compared to Hotels, Hostels and Resorts. The clusters are very less dense which indicates that the inns are not very popular in this region.

Analysis of Figures: Figure 14 (a) and 14 (b)

Observation

Number of Apartment based lodges are much less as compared to Hotels, Hostels and Resorts. The clusters are very less dense which indicates that the inns are not very popular in this region.

Analysis of Figures: Figure 15 (a) and 15 (b)

Observation

Number of Mansion based lodges are much less as compared to Hotels, Hostels and Resorts. The clusters are very less dense which indicates that the inns are not very popular in this region.

Analysis of Figures: Figure 16 (a) and 16 (b)

Observation

Number of Homestay based lodges are present in the least amount in the area.

Analysis of Lodgings booked on different dates

Observation

1) The above output shows the place of tourist residence which has the highest bookings on each date.

2) Hotel has the highest bookings on each date.

Inference

Tourists prefer to stay at hotels during the Bangkok trip.

1.7) Analysis of Destination Based Features

Analysis of Figures: Figure 17 and 18

Restaurants, Spa and Night Clubs are the favorite tourist destinations in Bangkok.

Analysis of Destination Bookings on each Date

Observation and Inference

1) The table in the section shows the destination tourist spots which had the maximum bookings on each date.

2) Night/Club, Bar, Spa, Restaurant are the favorite destination tourist spots.

Analysis of Figures: Figure 19

Observation

From the bar plot and pie chart we can observe that around 30.2 % tourists visit restaurants, 29.7 % tourists visits NightClub/Bar and 28.6 % tourists visit Spa.

Inference

From this we can infer that Restaurants, Night Clubs, Bar and Spa are the famous tourist spots of Bangkok.

Analysis of Figures: Figure 20 (a) and 20 (b)

Observation

1) Most of the destinations are clustered between 100.5 and 100.7 degree longitude and between 13.65 and 13.85 degrees latitude with occasional line based clustering extending outside of the radius of the cluster.

2) The destinations lie between Pak Kret and Samut Praken regions.

Analysis of Figures: Figure 21 (a) and 21 (b)

Observation

All the Spa are almost forming a single cluster near central Bangkok. Although there are very small number of Spa which lies outside the densely populated clusters like an outlier.

Analysis of Figures: Figure 22 (a) and 22 (b)

Observation

All the Night Clubs and Bar are almost forming a single cluster near central Bangkok. Although there are very small number of them which lies outside the densely populated clusters like an outlier. The Spa, Night Clubs and Bars are almost present in the same region in Bangkok. This shows that the region is a LandMark and a great tourist Spot for Bangkok.

Analysis of Figures: Figure 23 (a) and 23 (b)

Observation

1) The Shopping Malls are also present in the same region as Spa, Night Clubs and Bars.

2) The cluster of Shopping Malls are less dense than Spa and Night Clubs indicating the number of shopping malls are less when compared with Spa, Night Clubs, Bar.

3) The cluster of Shopping Malls are also more spread out over a greater region when compared with Spa, Night Clubs and Bars.

Analysis of Figures: Figure 24 (a) and 24 (b)

Observation

1) The Restaurants are forming dense clusters around central Bangkok.

2) Restaurants are also present in the same region as Spa, Night Clubs , Bars and Shopping Malls.

3) The density of the cluster is more than Shopping Malls but equal to or less than Spa and Night Clubs.

Analysis of Figures: Figure 25 (a) and 25 (b)

Observation

1) The majority of the Religious Spots are present in central Bangkok but there are some that spreads out to the outskirts as well.

2) The density of the cluster of Religious spots are much less as compared to other spots.

3) This indicates that the Religious places are much less popular as tourist spots when compared with that of Spa, Night Clubs and Bars.

Analysis of Figures: Figure 26 (a) and 26 (b)

Observation

1) The cluster of Leisure based tourist spots are least dense showing that it is not much popular as a tourist spot.

2) The clusters are not only present in central Bangkok but are also spread out in different regions of Bangkok.

1.8) Analysis of Tourist Types Visiting Bangkok

Analysis of Figures: Figure 27

Observation

From the above graph we can infer that high spending tourists are more than economical spending tourists.

Inferences

From this we can infer that mostly aristocratic people with high spending capabilities visit Bangkok. This also leads to another inference that Bangkok may be an expensive place for a tourist.

Analysis of Figures: Figure 28

Observation

1) From the above graph we can infer that high spending tourists and Economical spending tourists mostly go for FoodAndDrinkTrip, NightLifeTrip and Spa Trip.

2) FoodAndDrinkTrip, NightLifeTrip and Spa Trips are mostly done by High Spending Tourists.

Inferences

Mostly aristocratic people go for Different Trips. They mostly prefer FoodAndDrinkTrip, NightLifeTrip and Spa Trips

1.9) Analysis of Daily Trip at various dates

Analysis of Figures: Figure 29

Observation

1) April 12 had the most number of trips to the Tourist spots.

2) From April 12 onwards the trip counts decrease but still it is very high.

Inference

From this we can infer that from April 12 onwards it could be either a vacation season or festive season in Bangkok.

b. Statistics and Measures:

i) Summary statistics of numerical columns

Feature	Min	Max	Mean	25th Percentile	50th Percentile	75th Percentile
Distance	501	997684	5457.901632	<= 1819	<= 3423	<= 6513
Time	1.05	1439.9	23.855	<= 8	<= 13.2	<= 24

ii) Analysis of Feature 'distance' from the perspective of 'trajectory_type'

Trajectory Type	Min	Max	Mean
Food and Drinks Trip	501	397390	5914.376177
Leisure Trip	643	23940	3902.376744
Nightlife Trip	509	610606	5308.597369
Religious Trip	525	78652	4474.611502
Shopping Trip	503	997684	5364.915328
Spa Trip	502	792990	5281.356976

Observation

During their tour or trip in Bangkok, tourists have to travel long distances for FoodAndDrinkTrip, NightLifeTrips and ShoppingTrip. Even For SpaTrip, Significant distances were also needed to be covered.

iii) Analysis of Feature 'Time_Used' from the perspective of 'Trajectory_Type'

Trajectory Type	Min	Max	Mean
Food and Drinks Trip	1.75	1114.95	26.218770
Leisure Trip	3	941.65	29.271209
Nightlife Trip	1.05	1219.85	21.695671
Religious Trip	2.57	663.45	21.954906
Shopping Trip	1.98	1009.05	22.408091

Spa Trip	1.85	1439.90	24.069492

Observation

During their tour or trip in Bangkok, tourists spent a significant amount of time on NightLifeTrip, FoodandDrinkTrip, SpaTrip and Shopping Trip.

iv) Analysis of Statistics and Measures of the Graph Created

For this project we have considered a multidimensional network which is explained in section(c) suitable network types and shown in **Figure 31**. We have considered to build a Random Graph, since the dataset contains graph of the following format i.e

source 1 connected to destination 1

source 2 connected to destination 2

Hence the graph is of disjoint bipartite format consisting of disconnected components containing two nodes and one edge. So we are using randomization to connect source 1 to destination 1, source 1 to destination2 and destination 1 to destination 2. An edge is connected between them only if the random number generated between 0 and 1 is >= 0.8.

a) Analysis of Lower Layer Graph

In **Figure 32**, we have shown the lower layer of the proposed graph.
This is a bipartite graph from Source Category Nodes to Dest Category Nodes.
In **figure 33**, a plot of degree distribution with respect to rank is shown.
In **figure 34**, a plot of degree histogram is shown.
The exact values of all the measures are mentioned in the notebook.

Measures of the graph	**Observation/Inference**
Average Clustering Coefficient	It is 0 since there are no closed triads or triangles which are present in the above bipartite graph
Degree Centrality	1. The Sources have less degree of centrality than the Destinations. 2. Spa , Restaurant, Shopping Mall, Night Clubs and Bar have high degree centrality. 3. Destinations have higher prestige and gregariousness than Sources.

	4. Destinations are more important tourist spots than Sources.
Average shortest path length	This indicates that the various places are located very close to each other.
Page Rank	1. Night Clubs, Spa, Shopping Mall, Restaurants are the nodes with the most in degree and hence higher page rank. 2. Night Clubs, Spa, Shopping Mall, Restaurants are the most important spots or places in Bangkok which are visited by tourists.

In **figure 35** of **section 2.3**, we have shown a bipartite graph for only the date 2019-04-11.
In **figure 36 and 37 of section 2.3**, we have shown the degree distribution for **figure 35**.

Measures of the graph	**Observation/Inference**
Average Clustering Coefficient	It is 0 since there are no closed triads or triangles which are present in the above bipartite graph
Degree Centrality	1. The Sources have less degree of centrality than the Destinations. 2. Spa , Restaurant, Shopping Mall, Night Clubs and Bar have high degree centrality. 3. Destinations have higher prestige and gregariousness than Sources and are considered to be visited by various tourists staying at different sources or places of origin. 4. Destinations are more important tourist spots than Sources.
Average shortest path length	This indicates that the various places are located very close to each other.
Page Rank	1. Night Clubs, Spa, Shopping Mall, Restaurants are the nodes with the most in degree and hence higher page rank. 2. Night Clubs, Spa, Shopping Mall, Restaurants are the most important spots or places in Bangkok which are visited by tourists.
Betweenness Centrality	1) Betweenness Centrality of Spa, Restaurant, Shopping Mall, Night Clubs and Bar are

	higher than other nodes. 2) High Betweenness Centrality of Spa , Restaurant, Shopping Mall, Night Clubs and Bar indicates that they connect certain pairs of Source, Destination nodes better than others i.e they are important or essential bridges. From this we can infer that tourists who are visiting one tourist spot, also consider visiting other tourist spots only after visiting Spa, Restaurant, Shopping Mall, Night Clubs and Bar.
Closeness Centrality	1) Spa, Restaurant, Shopping Mall, Night Clubs and Bar have high closeness centrality. 2) High Closeness centrality of these places indicate that tourists visiting other tourist spots often visit Spa , Restaurant, Shopping Mall, Night Clubs.
Degree Distribution	The degree distribution is well spread between various nodes and approximately 4 nodes have degree 6

b) Analysis of Upper Layer Graph

From the **figure 39 and 40 of section 2.4),** in the trip trajectories network graph we find that it is a disjoint bipartite graph.Euclidean distance is calculated based on the GPS latitude and longitude data present in the dataset. Using this feature we will create weighted edges between nodes. Using this feature we will also add edges between destinations in order to create connected components in the graph as much as possible.

Since the graph is disjoint bipartite and enough features are not present in the dataset to create connected graph in order to calculate network measures, so we are generating a random graph using the following approach:

1. Take two nodes which don't have any edge between them.
2. Generate a random number between 0 and 1.
3. If the number is greater than 0.8 i.e if the probability is greater than 80 % then we add an edge between two sources.
4. If the number is greater than 0.5 i.e if the probability is greater than 50 % then we add an edge between two destinations.

We do not have the computational resources to compute the statistics and measures of such a huge graph
As we are getting Memory Error due to a large number of edges, we try to build a graph using only 600 source nodes and 600 destination nodes.

We are trying to connect or find relationships between different destinations.

The reason for doing this is that it is a natural tendency of the tourists to go from one tourist destination to other tourist destinations during the trip after commencing from the place of residence.

The graph is shown in **Figure 41**. A sample of the graph is shown in **Figure 42**.
In figure **43**, a plot of **degree distribution** with respect to rank is shown.
In figure **44**, a plot of **degree histogram** vs nodes is shown.

Measures of the graph	**Observation/Inference**
Betweenness Centrality	1) From this we can infer that a 'Restaurant' is an important bridge which connects different components of sources and destination nodes. 2) After commencing their journey from a particular source, tourists have the option to follow two different routes. They can either choose to visit specific destinations first, then proceed to the restaurant, and finally continue to other destinations. Alternatively, they may opt for a direct route, beginning their journey by visiting the restaurant immediately after departing from the source and then exploring other destinations.
Closeness Centrality	The Karaoke Bar with the highest closeness centrality indicates that it has the lowest average path length towards other nodes and is an important tourist spot.
Degree Centrality	The observation indicates that the prestige and gregariousness of the Thai Restaurant is high. This is a very important Tourist spot as most of the tourists visit this place during their Bangkok trip.
Average Clustering Coefficient	1. The score indicates that the neighbors of nodes are not very well connected 2. This can happen due to the fact that sources are not connected to all destinations, destinations are not connected to all other destinations and sources are not connected.
PageRank	From the dataset we can infer that a Thai

	Restaurant is the most important tourist place in Bangkok as most of the tourists with different sources of origin visit this place.
Number of Triangles or Triads	The node with the highest number of traiangles with itself as center is 'Karaoke Bar' with latitude of 13.7213 and longitude of 100.77644.
Diameter	The diameter is found to be 2
Average Shortest Path Length	It is found to be around 1.77 which indicates that the nodes are present close to one another
SimRank	Similarity of nodes is decided based on behavior of the tourists with respect to commencing the tri p starting from a source and going to a destinati on. It calculates how similar are the destination a nd source nodes with respect to other source no des and destination nodes which is based on the trip pattern of the tourists.
Transitivity	The transitivity of the graph is less than 0.5 so it i ndicates that the graph is not at all close to a co mplete graph
Degree Distribution	Most of the nodes have degree between 100 an d 150 and also between 400 and 500

c. Suitable Network Types:

We have proposed a Multidimensional graph for the given dataset as shown in **Fig.31** of ipython notebook. The graph is described as follows:

1. The upper layer forms a heterogeneous graph $G(V, E, f1, f2, VC, EC)$.

$$VC = \{Src\ Node, Dest\ Node\};\ EC = \{Euclidean\ Distance\};\ f1 : V -> VC;\ f2 : E -> EC$$

Since $|VC| > 1$ and $|EC| = 1$ so graph G is a node heterogeneous graph.

a) The graph G has Source Nodes (denoted by green) and Destination Nodes (denoted by orange).

b) There are directed edges from Src Nodes to Dest Nodes and from one Dest Nodes to another Dest Nodes.

c) The weights of these directed edges indicate Euclidean Distance from origin to destination.

2. The lower layer forms a directed bipartite graph $G1(V, E, f3, f4, VC1, EC1)$

$VC1 = \{Src\ Category, Dest\ Category\};\ EC1 = \{Trip\ Frequency\};\ f3 : V \rightarrow\ VC1;\ f4 : E \rightarrow\ EC1$

Since $|VC1| > 1$ and $|EC1| = 1$ so graph G is a node heterogeneous bipartite graph.

a) Graph G1 has Source Category Nodes (denoted by brown) and Destination Category Nodes (denoted by orange).

b) The edges are directed only from Source Category Nodes to Dest Category Nodes.

c) There are no directed edges from one Source Category Node to another Source Category Node and also from one Dest Category Node to another Dest Category Node. d) The weights of the edges indicate the trip frequency from Source Category Node to Dest Category Node.

3. The inter layer directed edges from nodes in the upper layer to nodes in the lower layer indicate 'belongs to' relationship i.e which nodes in the upper layer belong to which category nodes in the lower layer.

a) Description of Lower Layer Graph

In **Figure 32 and 35**, we have shown the lower layer of the proposed graph.
This is a bipartite graph from Source Category Nodes to Destination Category Nodes.

b) Description of Upper Layer Graph

From the **figure 39 and 40 of section 2.4),** in the trip trajectories network graph we find that it is a disjoint bipartite graph. Euclidean distance is calculated based on the GPS latitude and longitude data present in the dataset. Using this feature, we will create weighted edges between nodes. Using this feature, we will also add edges between destinations in order to create connected components in the graph as much as possible.

Since the dataset contains graph of the following format i.e

source 1 connected to destination 1

source 2 connected to destination 2

the graph is disjoint bipartite and enough features are not present in the dataset to create connected graph in order to calculate network measures, so we are generating a randomized graph using the following approach:

1. Take two nodes which don't have any edge between them.
2. Generate a random number between 0 and 1.

3. If the number is greater than 0.8 i.e if the probability is greater than 80 % then we add an edge between two sources.
4. If the number is greater than 0.5 i.e if the probability is greater than 50 % then we add an edge between two destinations.

The graph is shown in **Figure 41**. A sample of the graph is shown in **Figure 42**.

Kindly Note: As we are creating a Randomized Graph, so each time we run the code we get a different graph. So we have provided a Graph_created.pickle file which contains the graph with which we have performed the experiment.

4. Problem Statements:

1. Assuming that every passenger travels using the same vehicle ID (i.e., the same vehicle), if a tourist wants to travel to all the destinations mentioned in their itinerary, what will be the optimal route to cover all those places?

2. Providing personalized recommendations for tourists based on their original and destination categories.

3. Observing tourist patterns to identify the most popular places visited in a given time period.

5. Solution Approach:

1. This problem modeled into an optimization problem where the aim will be to find the shortest route to travel all the locations and return back to the starting point and it can be solved using heuristic algorithms like A*. Further constraints like any specific preferences the tourist has, such as visiting certain destinations at specific times could be considered.

2. By analyzing the tourist's profile and taking into account their category preferences, we can curate a personalized list of the top N destinations that would be most appealing to the traveler. This tailored approach ensures that the recommendations align perfectly with the individual's interests and desires, enhancing their overall travel experience.

3. Make use of community detection methods in order to organize locations into communities or clusters. It's possible that these settlements are representative of larger regions with comparable tourism patterns. In addition, temporal analysis can be used to monitor shifts in popularity over time and to spot patterns that are unique to particular seasons.

6. Takeaways:

1. From the exploratory data analysis, we gather that tourists in Bangkok tend to travel long distances for Food and Drinks, Nightlife, and Shopping trips. They also spend a significant amount of time on Nightlife, Food and Drinks, Spa, and Shopping trips.
2. The most popular tourist destinations in Bangkok are found to be Restaurants, Night Clubs, Bars, Spa, and Shopping Malls. These destinations were clustered mainly in central Bangkok, making it a convenient area for tourists to find accommodation.
3. High-spending tourists are more common in Bangkok, indicating that it may attract tourists with higher budgets.

7. References:

1. Prameshwori, T., et al. "GIS Based Route Network Analysis for Tourist Places: A Case Study Of Greater Imphal." *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN* (2021): 2395-1990.
2. Padia, Prarthana, et al. "Sentiment-aware and personalized tour recommendation." *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.
3. De Cantis, Stefano, et al. "Cruise passengers' behavior at the destination: Investigation using GPS technology." *Tourism Management* 52 (2016): 133-150.
4. Gill, N., and B. D. Bharath. "Identification of optimum path for tourist places using GIS based network analysis: a case study of New Delhi." *International Journal of Advancement in Remote Sensing, GIS and Geography* 1.2 (2013): 34-38.