Roll#:____________________________

Course:	Introduction to Data Science	**Course Code:**	DS 2001
Program:	BS(DS)	**Semester:**	Fall 2022
Duration:	3 Hour	**Total Marks:**	90
Paper Date:	27-12-2023	**Page(s):**	12
Section:	BS (DS) A, B, C	**Section:**	
Exam:	Final	**Roll No:**	

Instructions: Answer in the space provided. You can ask for rough sheets, but they will not be graded or marked. In case of confusion or ambiguity make a reasonable assumption. Questions during exam are not allowed.

Question#1 **[6+10+5=21]**

Given the following dataset

x1	**x2**	**y**	**Predicted y**
1	2	1	
-2	-2	0	
1	-1	0	
3	1	1	

a) Draw the architecture of a Multilayer Perceptron network containing 2 neurons and an output layer with a single neuron. Your network should take input features x1 and x2.

b) Using the architecture from Part a, perform a single forward pass with the given specifications:

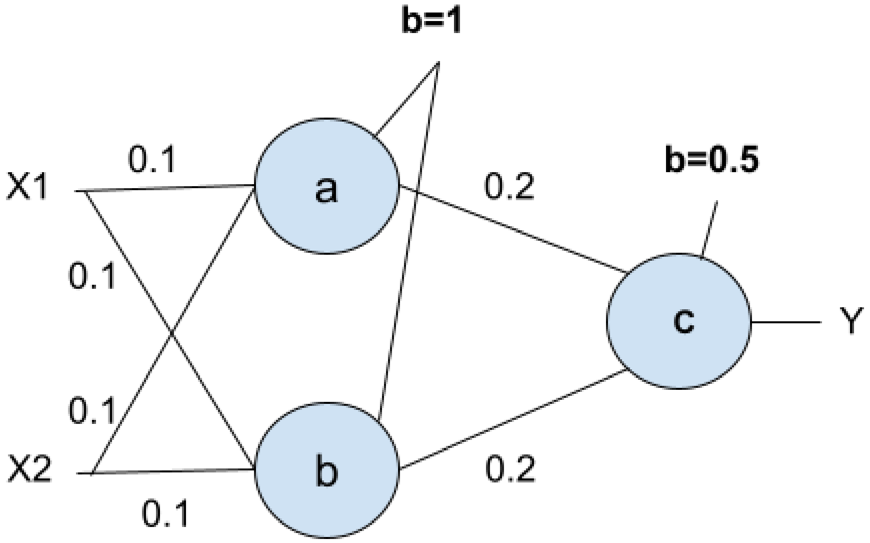
   i) All weights for the input layer are set to 0.1 and Bias is set to 1.

   ii) Output layer weights are set to 0.2 with bias of 0.5.

   iii) Apply the binary step activation function to each neuron

c) Using a threshold of >=0.5, convert the output of the output neuron to 0 and 1. Compare the predicted output with the actual output provided and calculate the accuracy of the model.

Roll#:__________________________



x 1	x2	y	Hidden Layer	Output Layer	Predicted y T>=0.5
1	2	1	a = 1*0.1+2*0.1 + 1 = 1.3 Activation = Sigmoid S(a) = 0.78 b = 1*0.1+2*0.1 +1 = 1.3 S(a) = 0.78	c = 0.78*0.2 + 0.78*0.2 + 0.5 = 0.81	1
-2	-2	0	a = -2*0.1+(-2)*0.1 + 1 = 0.6 S(a) = 0.64 b = -2*0.1+(-2)*0.1 + 1 = 0.6 S(a) = 0.64	c = 0.64*0.2 + 0.65*0.2 + 0.5 = 0.75	1
1	-1	0	a = 1*0.1+(-1)*0.1 + 1 = 1 S(a) = 0.73 b = 1*0.1+(-1)*0.1 +1 = 1 S(a) = 0.73	c = 0.73*0.2 + 0.73*0.2 + 0.5 = 0.79	1
3	1	1	a = 3*0.1+1*0.1 + 1 = 1.4 S(a) = 0.8 b = 3*0.1+1*0.1 +1 = 1.4 S(a) = 0.8	c = 0.8*0.2 + 0.8*0.2 + 0.5 = 0.82	1

Accuracy = 2/4 = 50%

Question#2: **[3x6 = 18]**

A financial institution has implemented a new fraud detection system to identify and prevent fraudulent transactions. The system uses a machine learning model to analyze transaction patterns and flag potentially fraudulent activities. Company is primarily interested in the number of fraudulent transactions that are correctly identified by the machine learning model. The institution wants to assess the performance of the system and has provided a dataset containing labeled examples of transactions (fraudulent or legitimate) for this purpose.

Dataset:

The dataset consists of 501 transactions, with 351 legitimate transactions and 150 fraudulent transactions. Each transaction record includes various features such as transaction amount, transaction type, and time of day. However, the model fails to correctly identify 6 legitimate and 15 fraudulent transactions.

a) Create a confusion matrix for the fraud detection system. Clearly mention the predicted and actual columns.

b) Calculate Accuracy, Precision, Sensitivity, Specificity and the F1 score.

c) Briefly interpret your results.

Predicted	Acual		
		legitimate	fraudulent
	legitimate	345	15
	fraudulent	6	135

True Positive (TP): 150 - 15 = 135
True Negative (TN): 351 - 6 = 345
False Positive (FP): 6
False Negative (FN): 15

Accuracy = 95.8
Precision = 95.74
Recall/Sensitivity = 90
Specificity = 98.29
F1 Score = 0.927

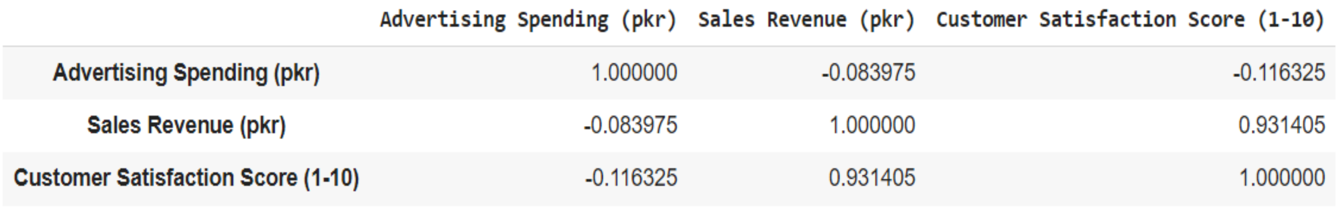
Question#3 **[10]**

A marketing department is analyzing the effectiveness of their advertising campaigns across different channels. They have collected data on the amount spent on advertising (in thousands) and the corresponding sales revenue generated (in thousands) for a set of products. The marketing team is interested in understanding the correlation between the advertising spending and sales revenue to optimize their future campaigns. The dataset contains data for 20 products, with the advertising spending and sales revenue for each product.

Create the heat map/correlation matrix for the above scenario, and comment on the importance of each of the features in ML training.

Product	**Advertising Spending (pkr)**	**Sales Revenue (pkr)**	**Customer Satisfaction Score (1-10)**
1	100	500	7
2	150	470	5
3	200	700	8
4	120	300	4
5	180	800	9
6	250	400	6
7	90	600	7
8	300	250	3
9	170	900	8
10	130	520	6
11	160	630	7
12	220	550	5
13	110	480	4
14	190	730	9
15	140	720	8
16	260	200	2
17	230	950	10
18	120	380	5
19	200	830	9
20	180	700	8

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$



	Advertising Spending (pkr)	Sales Revenue (pkr)	Customer Satisfaction Score (1-10)
Advertising Spending (pkr)	1.000000	-0.083975	-0.116325
Sales Revenue (pkr)	-0.083975	1.000000	0.931405
Customer Satisfaction Score (1-10)	-0.116325	0.931405	1.000000

Advertising Spending (pkr) vs. Sales Revenue (pkr):
Correlation Coefficient: -0.084
Interpretation: There is a weak negative correlation between Advertising Spending and Sales Revenue. This suggests that as Advertising Spending increases, Sales Revenue tends to slightly decrease. However, the correlation is not strong.

Advertising Spending (pkr) vs. Customer Satisfaction Score (1-10):
Correlation Coefficient: -0.116
Interpretation: There is a weak negative correlation between Advertising Spending and Customer Satisfaction Score. This suggests that as Advertising Spending increases, Customer Satisfaction Score tends to slightly decrease. Again, the correlation is not strong.

Sales Revenue (pkr) vs. Customer Satisfaction Score (1-10):
Correlation Coefficient: 0.931
Interpretation: There is a strong positive correlation between Sales Revenue and Customer Satisfaction Score. This suggests that as Sales Revenue increases, Customer Satisfaction Score tends to significantly increase. This strong positive correlation indicates a potential relationship between the two variables.

Question#4: **[10]**

In this dataset:

- "Age" is the age of the customer.
- "Income" is the annual income of the customer.
- "Purchase" is the target variable indicating whether the customer made a purchase (Yes) or not (No).

Calculate entropy and information gain **[For section B and C]**

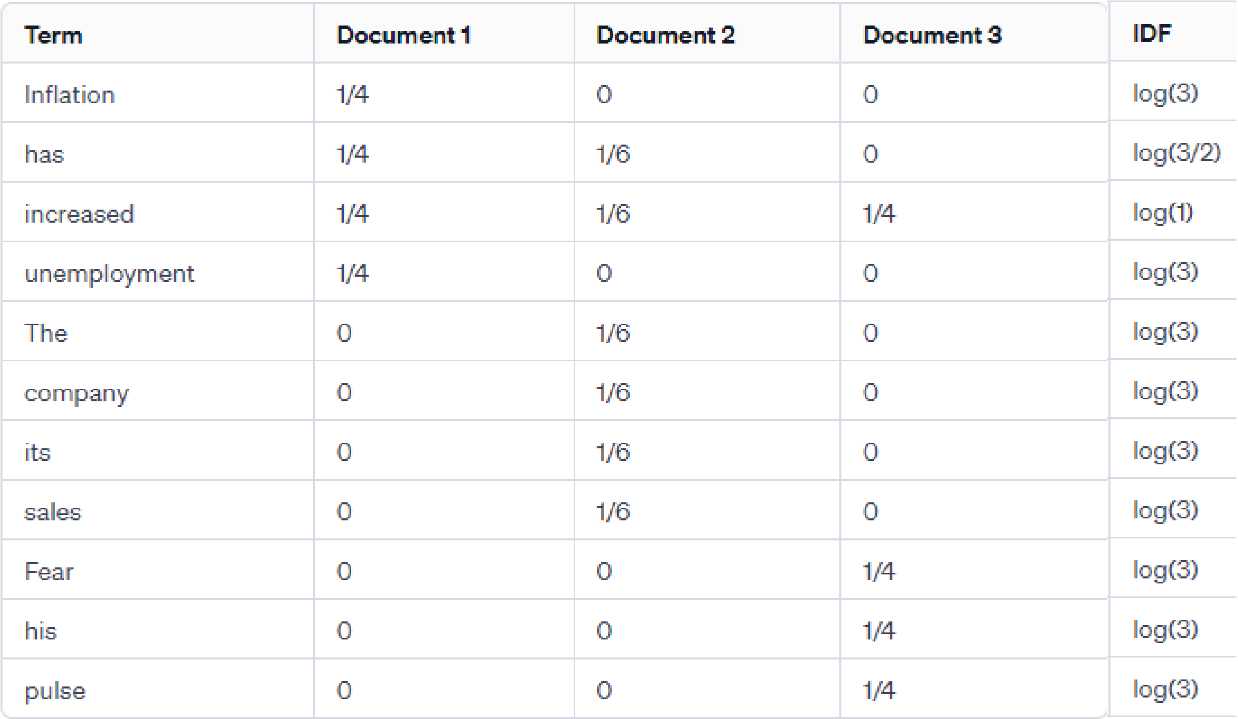
id	age	Income	purchase
1	20 - 25	low	no
2	30 - 35	medium	no
3	20 - 25	high	yes
4	20 - 25	low	no
5	20 - 25	high	yes
6	30 - 35	low	yes
7	20 - 25	high	yes
8	30 - 35	low	no
9	30 - 35	medium	yes
10	20 - 25	medium	no

OR

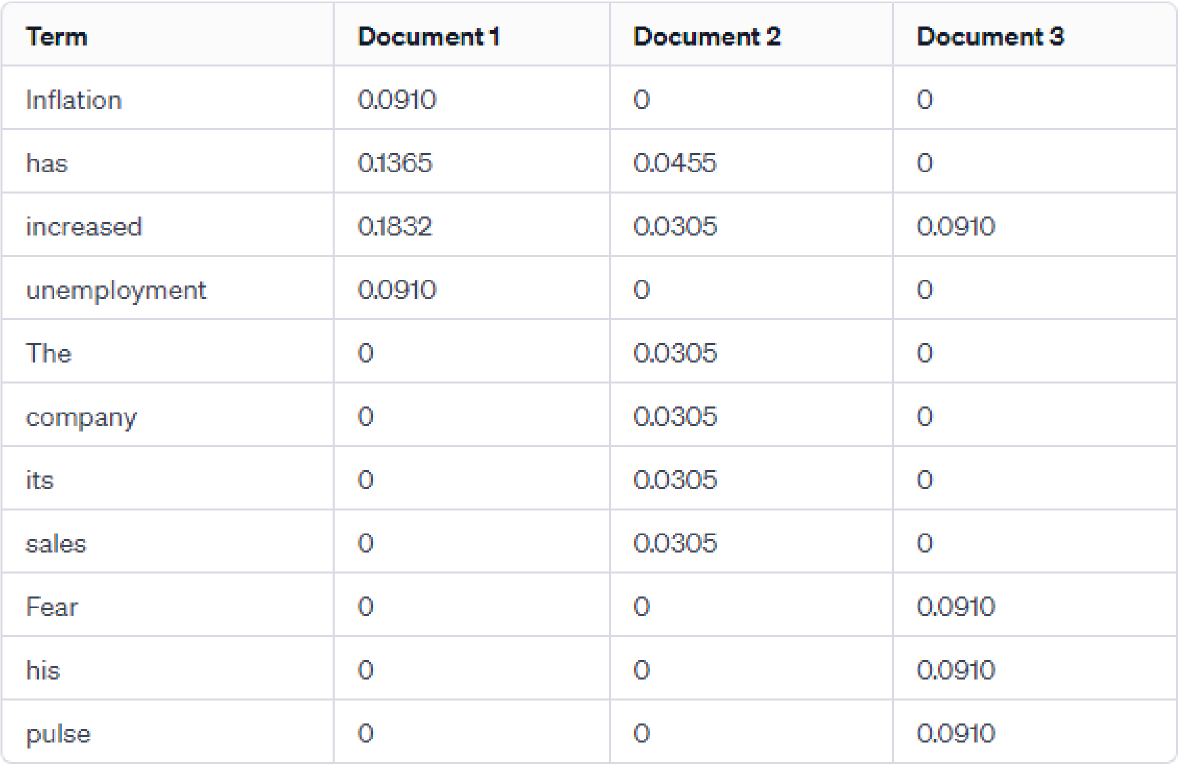
You are given a set of sentences, and your task is to calculate the Term Frequency-Inverse Document Frequency (TF-IDF) for each word in these sentences. **[For section A]**

1. Inflation has increased unemployment
2. The company has increased its sales
3. Fear increased his pulse

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$



Term	Document 1	Document 2	Document 3	IDF
Inflation	1/4	0	0	log(3)
has	1/4	1/6	0	log(3/2)
increased	1/4	1/6	1/4	log(1)
unemployment	1/4	0	0	log(3)
The	0	1/6	0	log(3)
company	0	1/6	0	log(3)
its	0	1/6	0	log(3)
sales	0	1/6	0	log(3)
Fear	0	0	1/4	log(3)
his	0	0	1/4	log(3)
pulse	0	0	1/4	log(3)



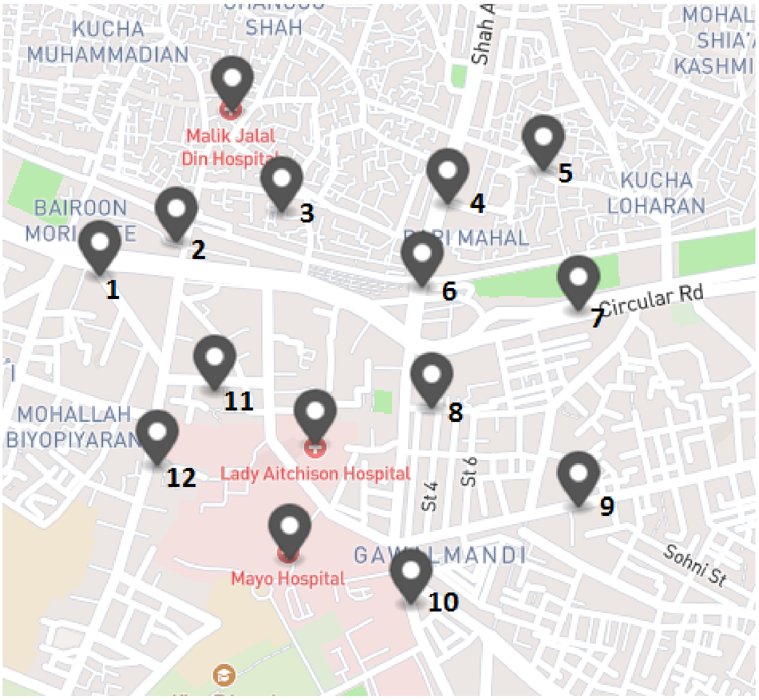
Term	Document 1	Document 2	Document 3
Inflation	0.0910	0	0
has	0.1365	0.0455	0
increased	0.1832	0.0305	0.0910
unemployment	0.0910	0	0
The	0	0.0305	0
company	0	0.0305	0
its	0	0.0305	0
sales	0	0.0305	0
Fear	0	0	0.0910
his	0	0	0.0910
pulse	0	0	0.0910

Roll#:________________________

Roll#:______________________________

Question#5 **[10+5=15]**

In this scenario, you are provided with data on several locations where accidents occurred in the past month, along with the coordinates of three hospitals. Your objective is to analyze accident hotspots and strategically cluster accident locations to determine an optimal allocation of hospitals.

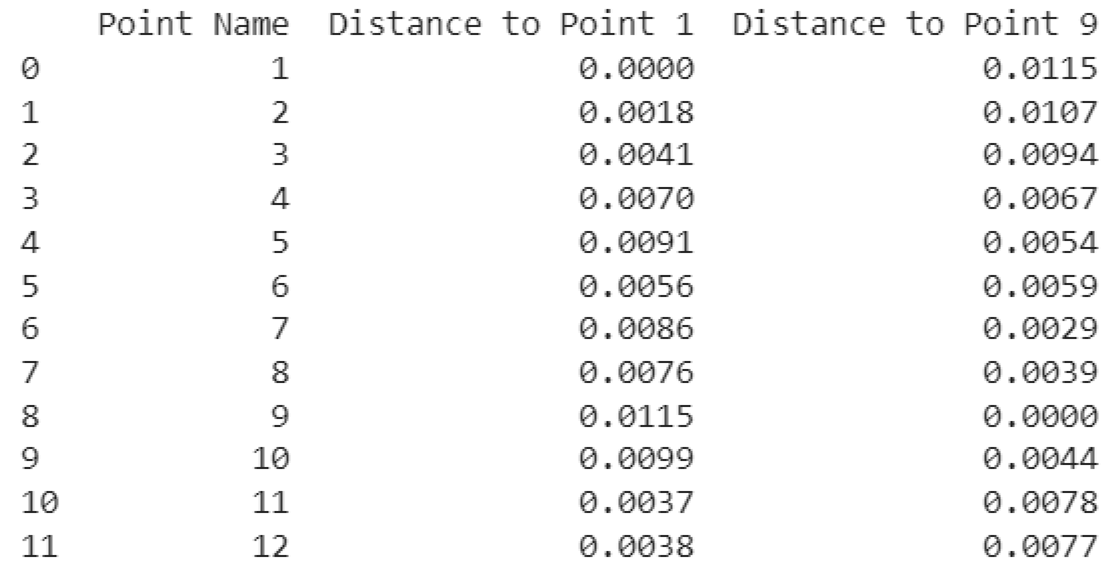


Malik Jalal (74.3140, 31.5788)
Lady Aitchison (74.3155, 31.5739)
Mayo (74.3150,31.5722)

Point Name	Latitude	Longitude
1	31.5764	74.3118
2	31.5769	74.3131
3	31.5774	74.3149
4	31.5775	74.3177
5	31.5779	74.3194
6	31.5763	74.3173
7	31.5759	74.3199
8	31.5745	74.3175
9	31.5730	74.3199
10	31.5716	74.3169
11	31.5747	74.3138
12	31.5736	74.3128

(a) Utilize clustering techniques to identify 2 accident hotspots. Implement a clustering algorithm (e.g., K-Means) to group accident locations based on proximity. Take 1 and 9 as initial centroids. Distance metric is Manhattan distance **(|(x2-x1)+(y2-y1)|)**.

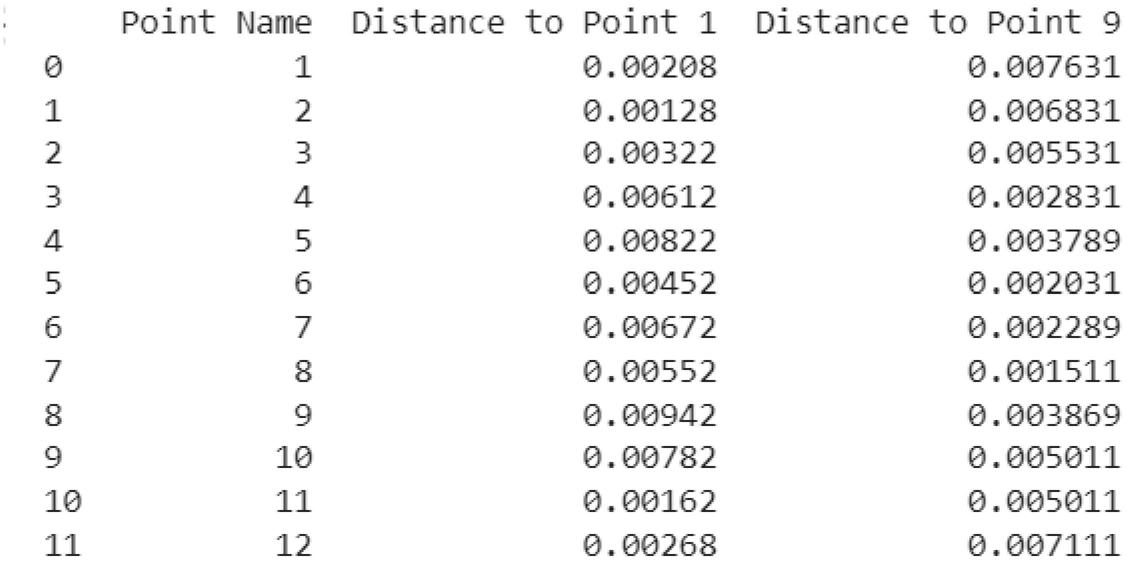
(b) Determine the optimal allocation of hospitals to the identified accident hotspots. Assign each cluster to the nearest hospital. Mark the resultant clusters on the map.



```
    Point Name  Distance to Point 1  Distance to Point 9
0            1               0.0000               0.0115
1            2               0.0018               0.0107
2            3               0.0041               0.0094
3            4               0.0070               0.0067
4            5               0.0091               0.0054
5            6               0.0056               0.0059
6            7               0.0086               0.0029
7            8               0.0076               0.0039
8            9               0.0115               0.0000
9           10               0.0099               0.0044
10          11               0.0037               0.0078
11          12               0.0038               0.0077
```

C1 = {1,2,3,11,12}
C2 = {4,5,6,7,8,9,10}

New centroids = [31.57580,74.31328],[31.57524,74.318271]

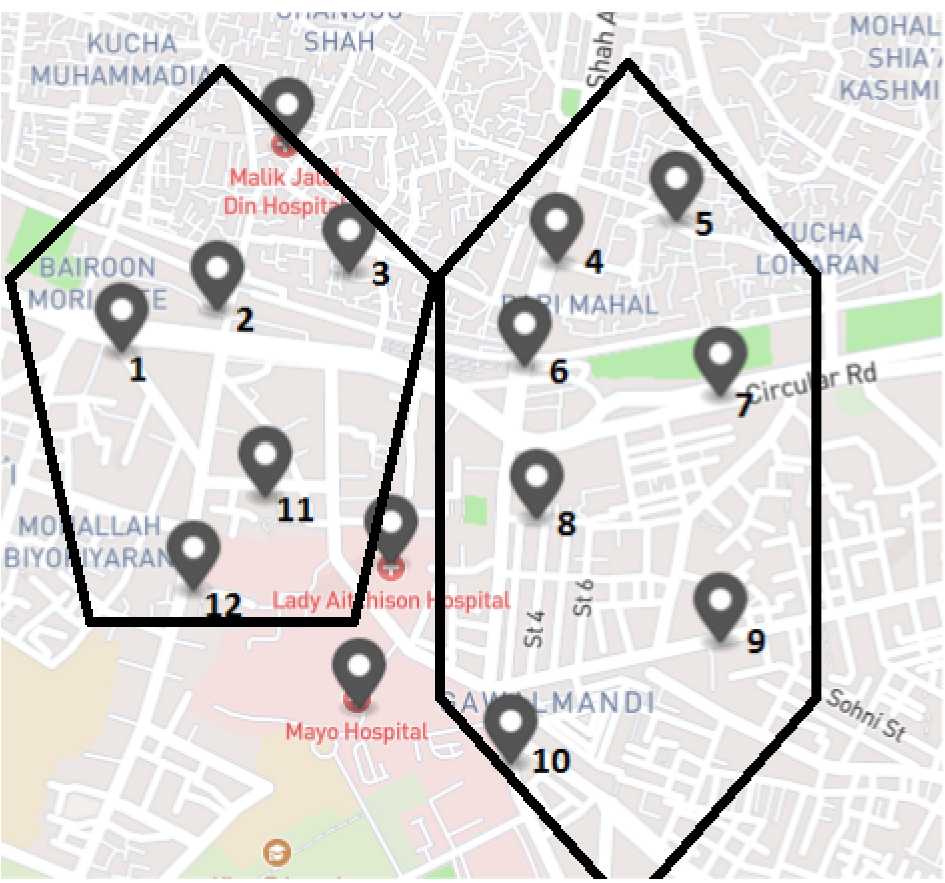


```
    Point Name  Distance to Point 1  Distance to Point 9
0            1              0.00208             0.007631
1            2              0.00128             0.006831
2            3              0.00322             0.005531
3            4              0.00612             0.002831
4            5              0.00822             0.003789
5            6              0.00452             0.002031
6            7              0.00672             0.002289
7            8              0.00552             0.001511
8            9              0.00942             0.003869
9           10              0.00782             0.005011
10          11              0.00162             0.005011
11          12              0.00268             0.007111
```

C1 = {1,2,3,11,12}
C2 = {4,5,6,7,8,9,10}

No change

Roll#:____________________________


KUCHA MUHAMMADIA
SHAH
Malik Jalal Din Hospital
BAIROON MORI GATE
MOHALLAH BIYORIYARAN
Lady Aitchison Hospital
Mayo Hospital
Shah A
MOHALLA SHIA KASHMI
KUCHA LOHARAN
PARI MAHAL
Circular Rd
St 4
St 6
GAWALMANDI
Sohni St
1
2
3
4
5
6
7
8
9
10
11
12

Question#6 **[2+2+2+10 = 16]**

Suppose you are hired to analyze the dynamic relationship between the film industry and socio-economic factors such as crime rate, GDP etc. in the United Kingdom from 2000 to 2022.
You are provided with three CSV files containing data related to IMDb movies and the crime rates. The datasets and their attributes are as follows.

1- imdb_movies.csv:
Attributes: {Title, Genre, Rating, IMDb Rating, Year, Duration, RegionCode}
2- imdb_regions.csv:
Attributes: {RegionCode, Country, Language}
3- crime_rate.csv:
Attributes: {Year, Crime rate, Country}

a) Load the csv files into pandas dataframes and perform basic preprocessing to clean the dataset.

```
# Load CSV files into DataFrames
imdb_movies = pd.read_csv('imdb_movies.csv')
imdb_regions = pd.read_csv('imdb_regions.csv')
crime_rate = pd.read_csv('crime_rate.csv')

#drop duplicates
imdb_movies .drop_duplicates(inplace=True)
imdb_regions .drop_duplicates(inplace=True)
crime_rate .drop_duplicates(inplace=True)

# Handle missing values
imdb_movies .dropna(inplace=True)
imdb_regions .dropna(inplace=True)
crime_rate .dropna(inplace=True)
# Convert data types if needed

crime_rate ['Crime rate'] = crime_rate ['Crime rate'].astype(float)
```

b) Merge the dataframes where necessary and filter the data that belongs to United Kingdom during 2000-2022

```
merged_data = pd.merge(imdb_movies , imdb_regions , on='RegionCode', how='left')

merged_data = merged_data[(merged_data['Country'] == 'United Kingdom') &
(merged_data['Year'].between(2000, 2022))]
```

```
crime_rate = crime_rate [(crime_rate ['Country'] == 'United Kingdom') & (crime_rate
['Year'].between(2000, 2022))]
```

c) Perform EDA and come up with at least two comprehensive graphs/visualizations that can help understand the data.

```
# Time Series Plot for Crime Rate Over the Years (2000-2022)
plt.figure(figsize=(12, 6))
sns.lineplot(x='Year', y='Crime rate', hue='RegionCode', data=crime_rate)
plt.title('Crime Rate Over the Years (2000-2022) in the United Kingdom')
plt.xlabel('Year')
plt.ylabel('Crime Rate')
plt.legend(title='Region Code', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()

# Distribution of Movie Genres in the United Kingdom
plt.figure(figsize=(12, 6))
sns.countplot(x='Genre', data=merged_data, order=merged_data['Genre'].value_counts().index)
plt.title('Distribution of Movie Genres in the United Kingdom (2000-2022)')
plt.xlabel('Movie Genre')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.show()
```

d) Your task is to find any relationship between movie Genre and crime rate over the years. How would you achieve this task? Your answer should contain you strategy accompanied with python code.

```
# Group by Genre and calculate the average crime rate for each genre
genre_crime_avg = merged_data.groupby('Genre')['IMDb Rating'].mean().reset_index()

# Merge the genre data with the crime rate data
merged_genre_crime = pd.merge(genre_crime_avg, crime_rate, on='Genre', how='left')

# Scatter plot to visualize the relationship
plt.figure(figsize=(10, 6))
sns.scatterplot(x='IMDb Rating', y='Crime rate', data=merged_genre_crime)
plt.title('Relationship between Movie Genre (IMDb Rating) and Crime Rate in the UK (2000-2022)')
plt.xlabel('Average IMDb Rating by Genre')
plt.ylabel('Crime Rate')
plt.show()

# Correlation analysis
correlation_matrix = merged_genre_crime[['IMDb Rating', 'Crime rate']].corr()
print("Correlation Matrix:\n", correlation_matrix)
```

Rough Sheet:

Roll#:__________________________

—---------------------Good Luck!----------------------------