Roll Number _____ 39.5

Course:	Introduction to Data Science	CourseCode:	DS 2001
Program:	BS(DS)	Semester:	Fall 2023
Duration:	1 Hour	Total Marks:	50
Paper Date:	02-10-2023	Page(s):	6
Section:	BS (DS) A, B, C	Section:	BDS 2A
Exam:	Mid I	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: **10x4 = 40 Marks**

The dataset represents a sample of employee performance evaluation data, containing various attributes related to individual employees within an organization. It includes information such as employee IDs, department affiliations, ages, genders, years of experience, performance ratings, joining dates, and salaries. Each row corresponds to a unique employee, and the dataset provides insights into factors affecting employee performance and compensation.

Employee_ID	Department	Age	Gender	Experience (Years)	Rating (1-5)	Joining Date	Salary
E001	Sales	35	Male	8	4	2020-06-15	60000
E002	HR	28	Female	4	3	2021-01-20	55000
E003	Engineering	42	Male	15	5	2019-03-10	75000
E004	Marketing	311	NULL	0	4	2020-11-05	32000
E005	Sales	29	Male	7	NULL	2021-09-18	58000
E006	Engineering	36	Male	10	4	2020-04-25	70000
E006	Sales	36	Male	-8	4	20-04-2020	70000

Answer the following questions: 3.5

a) **What is the type of each feature?**

Employee-ID = string
Department = string
Age = integar
Experience = integar
Rating = integar
Joining Date = String
Salary = integar

0020

Roll Number: _____

4/

b) Identify at least three quality issues with this data.

i) Null values: there are two null/missing values in "Gender" column, row 4 and "Rating" column, row 5 and negative value in col 5 row 6

ii) Index error: In last two rows, employee_ID has same value "E006"

iii) ~~Inconsis~~ Date format: in joining date column last row "20-04-2020" has wrong format or rest of the dates are in wrong format

c) Is there a correlation between years of experience and salary? If so, what is the nature of this correlation?

4/

Correlation is that they have direct relationship between them.

As experience increases from 0 to 4, salary increases highly. But from 4 years onwards, salary increases at lesser rate comparatively

d) Can you figure out imbalance distribution in any of the features?

0/

Age: There is "311" value which is an outlier and impossible outcome

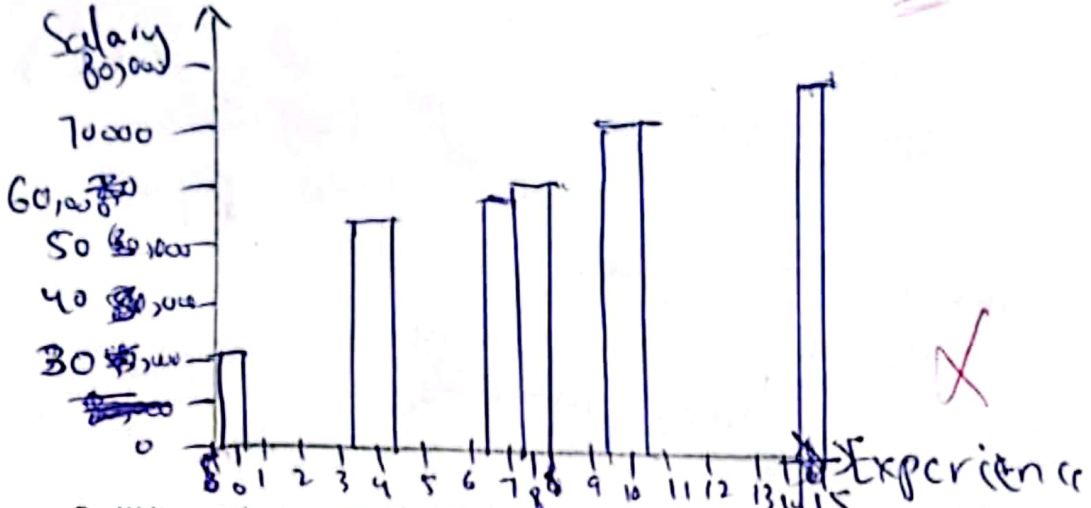
Experience: "-8" in last row, experience is always positive.

Salary: there is gap between $ values

0020

Roll Number: _____

e) Create a histogram of salaries. Identify the type of distribution.



f) Write a python command to display data types and non-null values for each feature.

```
df.info()
```

g) Write python code to group the data by "Gender" and calculate the average age for each gender.

```
ge_gender = df.groupby("Gender")["Age"].m
```

Roll Number: _____

h) Write python code to Determine the number of male and female employees in the dataset. 4

~~de no = df.groupby("Gender").size().unstack()~~

~~[0]~~

~~female_no = df.groupby("Gender").nuniqque~~

~~male_no = df.groupby("Gender")[df[~~

```
male_no = 0
female_no = 0
for i in range(df.shape[0][1])
    if (df["Gender"].iloc[i] == "Male"):
        male_no += 1
    elif(df["Gender"].iloc[i] == "Female"):
        female_no += 1
```

i) Write a python code to Group the data by "Department" and calculate the average salary for each department. 4

```
avg salary = df.groupby("Department")["Salary"].
                mean()
```

j) Write a python code to calculate the mean, median, and standard deviation of the "Salary" column. 3

```
Smean = df["Salary"].mean()
Smedian = df["Salary"].median()
Listsmode = df["Salary"].mode()
Smode = Listsmode[0]
```


0020

3/ 2x5 = 10 Marks

a) What are the key challenges in data cleaning, and how do you address them?

The key challenges in data cleansing are :-

→ Filling missing values.
→ Dropping desired data which is not useful.
→ Replacing the missing values with mean, or median etc.
→ Detecting outliers.
→ Removing or replacing outliers.

3/

b) Why is it important to identify outliers in a dataset, and what methods can be used for outlier detection?

→ It is important to identify outliers in a dataset. because data having outliers in it will not give accurate results. Data will produce unusual results.

→ Outlier detection method is following:-

$Q_1 = 0.25 \times N$ ∴ where n is sample number.

$Q_3 = 0.75 \times N$

→ upper bound = $Q_3 + 1.5\ IQR$. ∴ $IQR = Q_3 - Q_1$

→ lower bound = $Q_1 - 1.5\ IQR$.

If entries of dataset, are greater than upper bound & lower than lower bound, then the dataset will have outliers in it.