Roll Number: 221-7503

43

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-3A
Exam:	Mid I	Roll No:	221 7503

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.

33.5

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.

1 Employee_ID	2 Department	3 Age	4 Gender	5 Experience (Years)	6 Rating (1-5)	7 Joining Date	8 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

4 "assuming NULL as '' missing value, not string 'NULL'."

Answer the following questions: 3.5

a) What is the type of each feature?

1- object | ordinal
2. object | ordinal ✓
3- int | continuous
4. Object | Binary
5- int | continuous
6- int | Discrete ✗
7- object | nominal
8- int | Continuous

0025

Roll Number: 221-7503

b) Identify at least three quality issues with this data. 4/

→ Duplicate Entries (key ⇒ E006 twice)
→ Null & noisy data
→ Outliers (311 age, -8 Experience)
→ Informatted data (data object in invalid format)

c) Is there a correlation between years of experience and salary? If so, what is the nature of this correlation? 4

Yes, years of Experiences and salary are directly correlated if we assume that the data is cleaned - More the years of experience more the salary. Also added that the rating too affects the salary in the data.

d) Can you figure out imbalance distribution in any of the features? 0/

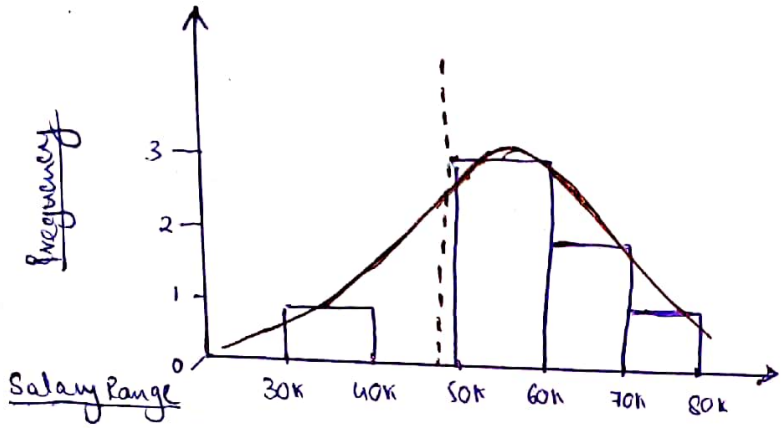
⇒ Salaries are imbalanely distributed without any proper method, seems like most of the data lies left of graph. X

⇒ ~~Age Status mass employment~~ Experience has too imbalance distribution X

Roll Number: 221-7503

e) Create a histogram of salaries. Identify the type of distribution. 4

df.hist()



30-40 => 1
41-50 => 0
51-60 => 3
61-70 => 2
71-80 => 1

Data is negatively skewed

f) Write a python command to display data types and non-null values for each feature. 4

df.info()

g) Write python code to group the data by "Gender" and calculate the average age for each gender. 4

a = df.groupby("Gender")["Age"].mean()

=> a[0]
=> a[1]

Roll Number: 22l-7503

h) Write python code to Determine the number of male and female employees in the dataset. 02

```
df.groupby("Gender").sum()
```

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

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

=> a

j) Write a python code to calculate the mean, median, and standard deviation of the "Salary" column. 4

```
a = df["Salary"].mean()
b = df["Salary"].median()
c = df["Salary"].std()
```

=> a
=> b
=> c

0025

Roll Number: 221-7503

Assuming the current DataSet for Example

4.5

Question#2: 2x5 = 10 Marks

a) What are the key challenges in data cleaning, and how do you address them? 4.5

Real-life data is dirty, in order to the data we have problems like multiple/duplicate entries, null value problem, outliers, invalid formatting.

=> we can remove duplicate entries from a dataframe use drop_duplicate() member function.

=> In order to resolve null-value problem, if the data is unnecessary we can ignore it, if it is important we can replace with centroid, lastly if we have much data we can drop the entry.

=> For outliers we can do same, but for invalid formatting one has to look entries individually.

b) Why is it important to identify outliers in a dataset, and what methods can be used for outlier detection? 5

=) Outliers can disturb the distribution of data. it can create biasedness in our model of machine learning and eventually a false prediction model.

=> To deal with outliers we have two methods:

(i) z-score $< > 3$

(ii) lower & upper bounds using IQR

=) To deal with data, we can do this

(i) If not important like employee ID, we can ignore it

(ii) If we need to maintain data distribution, we can fill it with median

(iii) If it is categorical replace with mode

(iv) If we have large dataset we can drop it