Course Name- EDA and Statistics

Course Code- INT 351

Continuous Assessment-I

Important Guidelines:

1. All questions in this Academic Task are compulsory.

2. It is mandatory to attempt all questions of the assignment in your own handwriting on A4 size sheets/pages with a blue color ink pen. Any other mode of attempt (typed or printed codes or table) except handwritten/drawn will not be accepted/considered as valid submission(s) under any circumstances.

3. Every attempted sheet/page should carry clear details of student such as Name, Registration number, Roll number, Question number and Page number. The page numbers should be written clearly on the bottom of every attempted sheet in a prescribed format as: for page 1; Page 1 of 4, for page 2; Page 2 of 4, for page 3; Page 3 of 4 and for page 4; Page 4 of 4, in case your assignment/document is of 4 pages.

4. After attempting the answer(s) single pdf format document (can be done with many free online available converters).

5. This PDF file should be uploaded onto the UMS interface on or before the last date of the submission.

6. Refrain from indulging into plagiarism as copy cases will be marked zero.

7. This Document contains multiple sets of papers. The allocation sheet is also attached in the CA file. All the students are advised to attempt the Set allocated to him/her.

8. If any student found indulge in malpractices like plagiarism from internet or classmates, attempting wrong set of question paper or any other, will be awarded with zero (0) marks in CA.

EDA and Statistics (INT-351) CA-1

Set-1

1. Select a csv/excel file of your own and answer the following questions.

i) What are the libraries used to read the file?

ii) What is the shape of your data-frame?

iii) Visualize the first and last 10 rows of the dataset. [1+1+3]

2. Find the null values in your dataset, and calculate the percentage of it, plot it in a bar chart. [1+2+2]

3. i) Drop the columns with more than 20% of null values and return the shape of the new data-frame.
   ii) Fill the remaining null values with mean/median/mode. Mention reason behind using any of the mentioned method [2.5+2.5]

4. Give a brief description about the column contained in the dataset. And give an intuition behind the statistical summary of the numerical columns. [2+3]

5. Let X be a random variable with PDF given by

$$f_X(x) = \begin{cases} cx^2 & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the constant c.

[5]

6. Perform univariate analysis of three numerical columns and state analysis of their distributions.

[5]

EDA and Statistics (INT-351) CA-1

Set-2

1. Select a csv/excel file of your own and answer the following questions.

i) What are the libraries used to read the file?

ii) What is the shape of your data-frame?

iii) Visualize the first and last 10 rows of the dataset. [1+1+3]

2. What do you mean by discrete and continuous random variables. Explain with examples. [2+1+2+1]
3. What is an outlier? Use boxplot on the numerical columns and state whether outlier exists in the feature or not. [2+3]
4. If outlier exists how will you treat the rows containing the outlier, do the needful and state reason behind it. [5]
5. Perform univariate analysis of three categorical columns of your choice, write the unique values, count of values and number of unique values in those columns. [5]
6. Let X be a discrete random variable with the following PMF [5]

$$P_X(x) = \begin{cases} 0.1 & \text{for } x = 0.2 \\ 0.2 & \text{for } x = 0.4 \\ 0.2 & \text{for } x = 0.5 \\ 0.3 & \text{for } x = 0.8 \\ 0.2 & \text{for } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

a. Find R_X, the range of the random variable X.
b. Find $P(X \leq 0.5)$.
c. Find $P(0.25 < X < 0.75)$.
d. Find $P(X = 0.2 | X < 0.6)$.

EDA and Statistics (INT-351) CA-1

Set-3

1. Import any dataset into the Jupyter Notebook of the following files
    a. Comma separated value
    b. Text file
    c. Excel file

    Using drive, import the new dataset of any source into the Jupyter Notebook. [5]

2. Perform the following functions using the given web-page and retrieve the information http://www.imdb.com/search/title?sort=num_votes,desc&start=1&title_type=feature&year=1950,2012
    a. Duration of the movie
    b. Name of the movie
    c. Releasing year. [5]
3. Insert Iris dataset into the Jupyter NoteBook and perform the Analysis
    a. Quantity Analysis of the dataset
    b. Finding the relationship among the values in the dataset
    c. Analyse the Gaussian Distribution of the data [5]
4. Using "bank.csv"
    a. Display the quantity of empty values
    b. Delete a particular column which has the highest number of empty values
    c. Replace the other missing values using inferential statistics
    d. Check if the data is redundant (if so remove it) [5]
5. With the help of the Boston House Pricing dataset, perform the following
    a. Using Univariate outlier graph, detect the outlier for the column 'DIS' and display the position of the outliers.
    b. Perform the analysis for the Multivariate Outlier graph and visualize it.
    c. By using the technique of the Z – Score, detect the outliers with the threshold value=3 [5]

6. With the User defined dataset, find out the outliers and perform the analysis

a. Using Quartile method find out the outlier of dataset
b. Find out the position of the detected outlier.
c. Eliminate the detected outliers.
d. Visualize the eliminated outliers. [5]

EDA and Statistics (INT-351) CA-1

Set-4

1. Import the below mentioned files into the Jupyter NoteBook from any source
    d. Comma separated value
    e. Text file
    f. Excel file

    Using drive, import the new dataset of any source into the Jupyter Notebook. [5]

2. Perform the following functions using the given web-page and retrieve the information. [5]

    https://www.flipkart.com/search?p%5B%5D=facets.brand%255B%255D%3DSamsung&sid=tyy%2F4io&sort=recency_desc&wid=1.productCard.PMU_V2_1

3. Insert Titanic dataset into the Jupyter NoteBook and perform the Analysis
    d. Quantity Analysis of the dataset
    e. Finding the relationship among the values in the dataset
    f. Analyse the Gaussian Distribution of the data. [5]
4. Using "bank.csv"
    e. Display the quantity of empty values
    f. Delete a particular column which has the highest number of empty values
    g. Replace the other missing values using inferential statistics
    h. Check if the data is redundant (if so remove it) [5]
5. Display the outlier of inbuilt dataset from the packages and perform the following analysis
    c. Using Univariate outlier graph, detect the outlier for the column 'DIS' and display the position of the outliers.
    d. Perform the analysis for the Multivariate Outlier graph and visualize it.
    e. By using the technique of the Z – Score, detect the outliers with the threshold value=3 [5]

6. With the User defined dataset, find out the outliers and perform the analysis
    e. Using Quartile method find out the outlier of dataset
    f. Find out the position of the detected outlier.
    g. Eliminate the detected outliers.
    h. Visualize the eliminated outliers. [5]

EDA and Statistics (INT-351) CA-1

Set-5

1. Create the following data files and import into Jupyter Notebooks
    a. Comma Separated Values
    b. Excel files
    c. Text file. [5]

    Import any data set from your drive into the python environment

2. Try to retrieve details from the below mentioned web-page and store it into a csv file http://www.imdb.com/search/title?sort=num_votes,desc&start=1&title_type=feature&year=1950,2012

    a. Name of the movie
    b. Releasing Year
    c. Duration of Movie. [5]

3. Insert Boston Dataset into the notebook
    a. Perform Quantity Analysis
    b. Find Correlation among the attributes
    c. Figure out Normality of the dataset [5]

4. Using “bank.csv”
    a. Display the quantity of empty values
    b. Delete a particular column which has the highest number of empty values
    c. Replace the other missing values using inferential statistics
    d. Check if the data is redundant (if so remove it) [5]

5. Using “Boston.csv”
    a. Display a boxplot for DIS attribute and the position of outliers
    b. Select any two attributes of your choice and detect the outliers

    Use Z-Score method to detect the outlier when threshold is greater than 3. [5]

6. Take any dataset of your choice which contains outliers
    a. Detect outliers using IQR method
    b. Display the position of those outliers
    c. Eliminate those outliers from the dataset
    d. Verify if the outliers got eliminated using visualization [5]

EDA and Statistics (INT-351) CA-1

Set-6

1. Create the following data files and import into Jupyter Notebooks
    d. Comma Separated Values
    e. Excel files
    f. Text file

    Import any data set from your drive into the python environment. [5]
2. Try to retrieve any details of your choice from the below mentioned web-page and store it into a csv file [5]

    https://www.flipkart.com/search?p%5B%5D=facets.brand%255B%255D%3DSamsung&sid=tyy%2F4io&sort=recency_desc&wid=1.productCard.PMU_V2_1

3. Insert Titanic Dataset into the notebook
    d. Perform Quantity Analysis
    e. Find Correlation among the attributes
    f. Figure out Normality of the dataset [5]
4. Using "bank.csv"
    e. Display the quantity of empty values
    f. Delete a particular column which has the highest number of empty values
    g. Replace the other missing values using inferential statistics
    h. Check if the data is redundant (if so remove it) [5]
5. Using "Boston.csv"
    c. Display a boxplot for DIS attribute and the position of outliers
    d. Select any two attributes of your choice and detect the outliers
    e. Use Z-Score method to detect the outlier when threshold is greater than 3 [5]
6. Take any dataset of your choice which contains outliers
    e. Detect outliers using IQR method
    f. Display the position of those outliers
    g. Eliminate those outliers from the dataset
    h. Verify if the outliers got eliminated using visualization [5]

EDA and Statistics (INT-351) CA-1

Set-7

1. Replace missing value in the bank dataset using mean, median, mode and remove the duplicate. [5]

2. Analyse quantity of column and its distribution using functions in pandas library. [5]

3.Import necessary libraries and scrap the data from the link given below into jupyter notebook and convert it into DataFrame in pandas with the columns (name, year and runtime)

Link

(http://www.imdb.com/search/title?sort=num_votes,desc&start=1&title_type=feature&year=1950,2012) [5]

4. Perform Heatmap using the inbuilt Boston dataset and describe the variable correlation and describe its variate type. [5]

5. Perform Data visualisation to find the outlier in univariate and bivariate.

[5]

6. Perform Heatmap using the inbuilt Boston dataset and describe the variable correlation and describe its variate type. [5]

EDA and Statistics (INT-351) CA-1

Set-8

1. Find the null values from the data set and remove the null value if it is numeric and if it is character replace it with NAN. [5]

2. Perform Heatmap using the inbuilt Boston dataset and describe the variable correlation and describe its variate type. [5]

3. Import dataset from csv, excel, text and from the google drive link given below LINK (https://drive.google.com/uc?export=download&id=1lqNKpOTdf5va7sQOsJKLShIWBvighaAI) [5]

4. Analyse the data quality issue and give a solution using missing value imputation. [5]

5. Perform distplot in target variable using the inbuilt Boston dataset and answer whether is randomly distributed or not. [5]

6. Replace missing value in the bank dataset using mean, median, mode and remove the duplicate. [5]

Student List with Assigned Sets

SerialNo	RegisterationNumber	Name	RollNumber	Set Allocation
1	12015947	Aaditya Mishra	RK20MPA01	Set – 1
2	12015838	Dev Rajput	RK20MPA02	Set – 2
3	12016262	Suthar Ghanshyam	RK20MPA03	Set – 3
4	12019402	Abhishek Kumar	RK20MPA04	Set – 4
5	12019311	Bairi Tulasi Ram	RK20MPA05	Set – 5
6	12019047	Katapally Yashas Vinay Reddy	RK20MPA06	Set – 6
7	12019170	K S Namritha	RK20MPA07	Set – 7
8	12020850	Hardik Bhatt	RK20MPA08	Set – 8
9	12020911	Khapate Srikanth Gangadhar	RK20MPA09	Set – 1
10	12015821	Bhabesh Kumar Jena	RK20MPA10	Set – 2
11	12015051	Devu Selva Raj Reddy	RK20MPA11	Set – 3
12	12013915	Abhishek Kumar	RK20MPA12	Set – 4
13	12014130	Rachamalla Shiva Shankar Goud	RK20MPA13	Set – 5
14	12013929	Venkata Naga Kalyan Nandavaram	RK20MPA14	Set – 6
15	12014433	Ishan Vivek	RK20MPA15	Set – 7
16	12009352	M V S K Lalith Kumar	RK20MPA16	Set – 8
17	12009254	Gembali Rakesh	RK20MPA17	Set – 1
18	12008336	Immidichetty Reddy Swethak	RK20MPA18	Set – 2
19	12007036	Vedullapalli Shanmukh Sri Sai	RK20MPA19	Set – 3
20	12006692	Atla Tarun Kumar	RK20MPA20	Set – 4
21	12007894	Sattu Revanth	RK20MPA21	Set – 5
22	12014609	Devangam Nikhil Sekhar	RK20MPA22	Set – 6
23	12015282	Pulkit Patodia	RK20MPA23	Set – 7
24	12015877	Shubham Chandak	RK20MPA24	Set – 8
25	12016138	Challa Rohan Reddy	RK20MPA25	Set – 1
26	12020686	Sanjay Kumar Swami	RK20MPA26	Set – 2
27	12012614	Veeri Dheeraj	RK20MPA27	Set – 3
28	12013075	Muthyallh Mmihiraansh	RK20MPA28	Set – 4
29	12011557	Katappagari Naveen Reddy	RK20MPA29	Set – 5
30	12019386	Prathapa Shivateja	RK20MPA30	Set – 6
31	12019635	Dudekula Sindhu Padmaja	RK20MPA31	Set – 7
32	12016080	Sabjit Singh	RK20MPA32	Set – 8
33	12018805	C Chandini	RK20MPA33	Set – 1
34	12019163	Ankit Pandey	RK20MPA34	Set – 2
35	12019967	Vineet Srinivas Dasari	RK20MPA35	Set – 3
36	12004188	Aman Kumar Sharma	RK20MPB36	Set – 4
37	12009326	Md Mazidul Islam	RK20MPB37	Set – 5
38	12013468	Mohit Singh	RK20MPB38	Set – 6
39	12013580	Pasumarthi Jaya Sanjay	RK20MPB39	Set – 7
40	12015083	Jaismeen	RK20MPB40	Set – 8
41	12014496	Rohit Kumar Yadav	RK20MPB41	Set – 1
42	12020315	Shivam Singh	RK20MPB42	Set – 2
43	12020413	Mayank Pant	RK20MPB43	Set – 3
44	12020309	Khureshi Imran	RK20MPB44	Set – 4

45	12019024	Gaddam Nimitha Reddy	RK20MPB45	Set – 5
46	12019279	D Sri Ram Vamsi Kalavagunta	RK20MPB46	Set – 6
47	12019359	Chinthalagattu MadhuKumar	RK20MPB47	Set – 7
48	12019674	Amrit Kumar Samantaray	RK20MPB48	Set – 8
49	12019711	Aryan Jain	RK20MPB49	Set – 1
50	12019978	Param	RK20MPB50	Set – 2
51	12018291	Rishika Pandeya	RK20MPB51	Set – 3
52	12000061	Abubakar Wazih Tushar	RK20MPB52	Set – 4
53	12020439	Ritik Kumar	RK20MPB53	Set – 5
54	12019246	Kartik Yadav	RK20MPB54	Set – 6
55	12019266	Ronak Jain	RK20MPB55	Set – 7
56	12019283	Raqeeb Shaikh	RK20MPB56	Set – 8
57	12019715	Akshay Kumar	RK20MPB57	Set – 1
58	12012095	Sriraj Bhogi	RK20MPB58	Set – 2
59	12012994	Andriyala Manidathu	RK20MPB59	Set – 3
60	12015095	Dappili Bharat Kumar Reddy	RK20MPB60	Set – 4
61	12002346	Duddu Yaswanth	RK20MPB61	Set – 5
62	12007218	Aryaman Singh Tomar	RK20MPB62	Set – 6
63	12008460	Ranak Debnath	RK20MPB63	Set – 7
64	12014538	Mohit	RK20MPB64	Set – 8
65	12019295	Vendikattu Naveen	RK20MPB65	Set – 1
66	12019441	Badavath Srikanth	RK20MPB66	Set – 2
67	12019541	Vaibhav Mohan Mishra	RK20MPB67	Set – 3
68	12019982	Rajiv Ranjan	RK20MPB68	Set – 4
69	12020440	Aditya Raj	RK20MPB69	Set – 5
70	12020500	Sanjeev Thakur	RK20MPB70	Set – 6
71	12010696	Md Ananur Islam	RK20MPB71	Set - 7
72	12014891	Piyush Kumar Singh	RK20MPB72	Set - 8