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 25 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 andstate 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]

EDA and Statistics (INT-351) CA-1

Set-9

Q1. Define Exploratory Data Analysis (EDA) and explain its importance in the data analysis process. What are the primary goals of Exploratory Data Analysis (EDA)? [5]

Q2: Using a Python library of your choice, load a dataset (House price prediction California dataset) and demonstrate three EDA techniques you can use to understand its basic characteristics. [5]

Q3: Distinguish between public and private data. Provide examples of each type. [5]

Q4: Write a Python code snippet to access a public API and retrieve data. Display a summary of the retrieved data. [5]

Q5: Explain the importance of correctly identifying and handling data types during the data analysis process. Why is data cleaning an essential step before performing EDA? Provide specific examples of data quality issues that might require cleaning. Describe the process of handling duplicate records in a dataset. Why is this important for accurate analysis? [5]

Q6: What is data standardization and why is it useful in data analysis? Provide an example of a situation where standardization is beneficial. How does it impact machine learning algorithms? Describe the process of data filtering and provide an example of a filtering criterion that could be applied to a dataset. [5]

EDA and Statistics (INT-351) CA-1

Set-10

Q1: List and briefly explain the main objectives of conducting EDA on a dataset. Discuss the ethical considerations associated with using public and private data for analysis. [5]

Q2: Using a Python library of your choice, load dataset from Kaggle (IRIS dataset) and demonstrate three EDA techniques you can use to understand its basic characteristics. [5]

Q3. Define web scraping and provide an example of a real-world scenario where web scraping could be useful for data collection. [5]

Q4: Write a Python code snippet to access a public API and retrieve data. Display a summary of the retrieved data. [5]

Q5: Given a dataset, write a Python code snippet to determine the data types of each column? [5]

Q6: How does data filtering contribute to focusing the analysis on specific subsets of data? Describe the process of data filtering and provide an example of a filtering criterion that could be applied to a dataset. [5]

EDA and Statistics (INT-351) CA-1

Set-11

Q1: Why is it important to visualize data during the EDA process? Provide at least two reasons. Distinguish between public and private data. Provide examples of each type. [5]

Q2: Using a Python library of your choice, load dataset from Kaggle (Bioassay dataset) and demonstrate three EDA techniques you can use to understand its basic characteristics. [5]

Q3: What are the potential legal and ethical challenges of web scraping? How can these challenges be addressed? [5]

Q4: Define outliers in a dataset and explain their potential impact on statistical analysis. Describe two methods for identifying and handling outliers in a dataset. [5]

Q5: What is data standardization and why is it useful in data analysis? Provide an example of a situation where standardization is beneficial. [5]

Q6: Give examples of invalid values that might be present in a dataset. How can these invalid values be identified and rectified? Discuss the potential consequences of not addressing invalid values in a dataset. How does data filtering contribute to focusing the analysis on specific subsets of data? [5]

EDA and Statistics (INT-351) CA-1

Set-12

Q1: How does EDA help in understanding the distribution and characteristics of a dataset? Explain with help taking a real time scenario for the university management system by taking an example of teacher student relationship. [5]

Q2: Explain the importance of correctly identifying and handling data types during the data analysis process. [5]

Q3: Describe the process of data filtering and provide an example of a filtering criterion that could be applied to a dataset. [5]

Q4: When might it be appropriate to remove rows or columns with missing values instead of imputing them? What are some common reasons for having duplicate rows in a dataset? How can they be identified, provide two common techniques for the missing values? Write a Python function to remove duplicate rows from a Data Frame using pandas. [5]

Q5: Define web scraping and provide an example of a real-world scenario where web scraping could be useful for data collection. [5]

Q6: Give examples of invalid values that might be present in a dataset. How can these invalid values be identified and rectified? Discuss the potential consequences of not addressing invalid values in a dataset. [5]

EDA and Statistics (INT-351) CA-1

Set-13

Q1: What are some common reasons for having duplicate rows in a dataset? How can they be identified? Write a Python function to remove duplicate rows from a Data Frame using pandas. [5]

Q2: Explain the difference between z-score standardization and Min-Max scaling. [5]

Q3: Write a code in python and List out, briefly explain the fundamental data types in Python. How does data filtering contribute to focusing the analysis on specific subsets of data? [5]

Q4: Explain the concept of the Interquartile Range (IQR) and how it's used to identify outliers. Write a Python function that takes a Data Frame column as input and identifies and removes outliers using the IQR method. [5]

Q5: Write a note of Min Max Scoring? [5]

Q6: List and define the main data types typically encountered in datasets. [5]

Student List with Assigned Sets

Registration Number	Name	RollNumber	Alotted Sets
12113501	Shubham Kumar	RK21UTA01	Set-1
12112282	Palli Sai Kiran	RK21UTA02	Set-2
12112093	Khurram Shahin	RK21UTA03	Set-3
12111724	Shahriar Mumin Khan	RK21UTA04	Set-4
12113102	Annamdevula Ravi	RK21UTA05	Set-5
12113229	Gummudu Kishore Kumar	RK21UTA06	Set-6
12109994	Priyanshu Singh	RK21UTA07	Set-7
12110145	Prathipati Venkatesh	RK21UTA08	Set-8
12110626	Marlakunta Kedhareswer Naidu	RK21UTA09	Set-9
12111396	Darsi Venkat Charan	RK21UTA10	Set-10
12100915	Nived Suresan A	RK21UTA11	Set-11
12100863	C S Charithartha Sai	RK21UTA12	Set-12
12109514	Nikhil Singh	RK21UTA13	Set-13
12109665	T Tanusree	RK21UTA14	Set-1
12109211	Karri John Pradeep Reddy	RK21UTA15	Set-2
12108024	Anushka Kashyap	RK21UTA16	Set-3
12108472	Gopidesi Vinod Kumar	RK21UTA17	Set-4
12108725	Dharani K S	RK21UTA18	Set-5
12106386	Pentyala Kumar Govindu	RK21UTA19	Set-6
12106729	Kriti Mishra	RK21UTA20	Set-7
12106692	Garvit Joshi	RK21UTA21	Set-8
12107057	Yaswanth Subrahmanyam Jonnadula	RK21UTA22	Set-9
12107367	Shivansh Ranjan	RK21UTA23	Set-10
12107544	Shaik Latheef	RK21UTA24	Set-11
12107776	Lakshya Sharma	RK21UTA25	Set-12
12107627	Medam Sai Shashank	RK21UTA26	Set-13
12104754	Achanagari Hanu Tejesh	RK21UTA27	Set-1
12104652	Alexander Peter Maliyakkal	RK21UTA28	Set-2
12106234	Vulli B M S Pruthvi	RK21UTA29	Set-3
12105798	Utkrist Ark	RK21UTA30	Set-4
12103929	Velagalapalli Sai Kishore Chandra	RK21UTA31	Set-5
12115897	Kunal Yadav	RK21UTA32	Set-6
12115161	Mahrishi Rathore	RK21UTA33	Set-7
12115398	Rohan Patel	RK21UTA34	Set-8
12116486	Madhan Sai Thupakula	RK21UTA35	Set-9
12019170	K S Namritha	RK21UTA71	Set-10
12102845	Ankur Banerjee	RK21UTB36	Set-11
12102585	Nikhil Pathak	RK21UTB37	Set-12
12102610	S Surjith Subash	RK21UTB38	Set-13
12101918	Indukuri Satya Sudheer Varma	RK21UTB39	Set-1
12101692	Gurram Karthik	RK21UTB40	Set-2
12104702	K Somanath Sai Teja Srinivas	RK21UTB41	Set-3
12104879	Jarugu Mukesh Sai	RK21UTB42	Set-4
12107747	Mahamad Suhail	RK21UTB43	Set-5

12107884	Vaspari Murari	RK21UTB44	Set-6
12107890	Sanjana Umrao	RK21UTB45	Set-7
12107896	Prabhu Varun Puppala	RK21UTB46	Set-8
12107901	Madireddy Bharath Kumar Reddy	RK21UTB47	Set-9
12107624	Kanigelupula Surya Venkata Phanindra	RK21UTB48	Set-10
12107183	Rahul Rajput	RK21UTB49	Set-11
12108436	Saksham Parasher	RK21UTB50	Set-12
12108310	Mohammed Aasif	RK21UTB51	Set-13
12107941	Peyyala Akshay Mathew	RK21UTB52	Set-1
12109517	Adigopula Varun Kumar	RK21UTB53	Set-2
12109549	Pallanti Asrith Vatsal	RK21UTB54	Set-3
12100859	Abhinav Kumar	RK21UTB55	Set-4
12100568	Mandeep Singh Gill	RK21UTB56	Set-5
12100583	Sunkari Vedavyas	RK21UTB57	Set-6
12100403	Poothi Chandrasekhar Reddy	RK21UTB58	Set-7
12110965	Anindita Pandit	RK21UTB59	Set-8
12110943	Shristi Sehwag	RK21UTB60	Set-9
12113036	Siddharth Prahasith Bathula	RK21UTB61	Set-10
12112410	Nikhil Kaundal	RK21UTB62	Set-11
12111711	Kunal Kumar Pandit	RK21UTB63	Set-12
12111702	manish choudhury	RK21UTB64	Set-13
12112264	Bevara Hemanth Kumar	RK21UTB65	Set-1
12113773	Vidhya Bhusan Rath	RK21UTB66	Set-2
12115210	Rohan Stanislaus R	RK21UTB67	Set-3
12115853	Syed Faiq Husain	RK21UTB68	Set-4
12114879	Debasish Chandra Dey	RK21UTB69	Set-5
12114325	Aman Verma	RK21UTB70	Set-6