National University of Computer and Emerging Sciences

Islamabad Campus

Introduction to Data Science (DS2001)

Sessional-II Exam

Total Time (Hrs): 1

Total Marks: 40

Total Questions: 3

Course Instructor(s):
Ms. Saira Qamar
Section(s): DS-A, B, C

Date: Nov 4, 2024

_____ _____ _____

Roll No **Course Section** **Student Signature**

Do not write below this line.

Attempt all the questions.

[CLO-1 Comprehend the fundamental constructs of programming language for data analysis and representation.]

Q1: Write the outputs of the following python codes in the respective output column. There are no syntax mistakes in the code. Assume that all libraries and modules are already imported. **[14 marks]**

	Code	**Output**
i.	data = pd.DataFrame({ 'Price': [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]}) filtered_data = data[(data['Price'] > data['Price'].quantile(0.25)) & (data['Price'] < data['Price'].quantile(0.75))] print(filtered_data)	
ii.	data = [12, 15, 13, 10, 8, 16, 18, 11, 17, 19, 3] plt.boxplot(data) plt.show() Which values are considered outliers based on the boxplot?	
iii.	data = pd.DataFrame({ 'City': ['New York', 'Los Angeles', 'New York', 'Chicago', 'Chicago', 'Los Angeles'], 'Category': ['A', 'B', 'A', 'A', 'B', 'A'], 'Sales': [100, 200, 150, 80, 60, 90]}) grouped_data = data.groupby(['City', 'Category']).sum() print(grouped_data)	

iv.	text = "Data science is fascinating!" tokens = word_tokenize(text) print(tokens)	
v.	data = pd.DataFrame({ 'Age': [25, 32, 47, 51, 38], 'Income': [40000, 50000, 75000, 82000, 61000]}) print(data.describe())	
vi.	df = pd.DataFrame({ 'A': [1, 2, 3, None, 5], 'B': [5, None, 3, 2, 1]}) print(df.fillna(df.mean()))	
vii.	data = pd.DataFrame({ 'A': [1, 2, np.nan, 4], 'B': [5, np.nan, np.nan, 8]}) missing_values = data.isnull().sum() print(missing_values)	

[CLO-3 Solve and analyze programming and data analysis problems using standard libraries and/or toolboxes of the programming language.]

Q2: Write the short answers of the following questions: [10 marks]

i. Explain skewness and kurtosis in a dataset. Given that high skewness or kurtosis might indicate outliers, how would you interpret these values in the context of EDA, and what strategies could you apply to handle highly skewed data? [2.5 marks]

ii. Write a Python function that takes a DataFrame and returns a list of columns with missing values, along with the percentage of missing data for each column. Suggest a strategy for handling missing data based on these percentages. [2.5 marks]

iii. Write Python code to load an image, convert it to grayscale, and then apply a Gaussian blur to reduce noise. Explain how Gaussian blurring helps with further image processing tasks and its effect on different image details. [2.5 marks]

iv. Write Python code that applies word frequency analysis to a text corpus and identifies the top 5 most frequent words, excluding common stop words. Explain how word frequency can give insights into the text and situations where it might be misleading. [2.5 marks]

[CLO-3 Solve and analyze programming and data analysis problems using standard libraries and/or toolboxes of the programming language.]

Q3: You work for an online book retailer, and your team has collected customer reviews along with ratings for each book. The goal is to analyze general trends in customer satisfaction by genre and identify common words or phrases that indicate positive or negative sentiments. You need to conduct an exploratory data analysis (EDA) and apply NLP preprocessing techniques, including tokenization, stemming, and lemmatization. The dataset includes the following columns: **[16 Marks]**

- BookID: Unique identifier for each book.
- Genre: Genre of the book (e.g., "Mystery," "Romance," "Fantasy").
- Rating: Customer rating for the book on a scale of 1 to 5.
- ReviewText: Customer review for the book (e.g., "An amazing journey, couldn't put it down

BookID	Genre	Rating	ReviewText
101	Mystery	5	"An amazing journey, couldn't put it down!"
102	Romance	2	"Boring and predictable, didn't enjoy it."
103	Fantasy	4	"Great world-building, but a bit too slow."
104	Mystery	3	"Interesting, but the ending was disappointing."
105	Fantasy	1	"Terrible characters and poor story development."

Using this data, write code to complete the following tasks:

---

1. Exploratory Data Analysis (EDA): [9 marks]
    - Genre vs. Average Rating: Write a python code to calculate the average rating for each genre, then plot this information as a bar chart. Based on your findings, print which genres have the highest and lowest average ratings.
    - Rating Distribution by Genre: Write a python code to create boxplots of Rating for each genre to visualize the spread and detect any outliers.
    - Positive and Negative Reviews Analysis: Write a python function to define reviews with ratings 4 or 5 as "Positive" and reviews with ratings 1 or 2 as "Negative." Count the number of positive and negative reviews for each genre and plot the counts. Identify any genres with notably higher or lower counts of positive or negative reviews.

2. NLP Text Preprocessing: [7 marks]
    - Tokenization: Write a python code to tokenize each review in the ReviewText column. For the first five reviews, display the tokens alongside the original text.
    - Apply stemming and lemmatization to the tokens in the ReviewText column. Display the original tokens, stemmed tokens, and lemmatized tokens for the first five reviews.
    - Write code to calculate the percentage of words that changed after stemming and lemmatization. Print this percentage and discuss what this indicates about the complexity of the reviews.