|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Register Number** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



**SRM Institute of Science and Technology**

Set -

**College of Engineering and Technology**

**School of Computing**

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamil Nadu

# Academic Year: 2024-25 (EVEN)

Test: FT4 Date: 29-04-2025

Course Code & Title: 21CSS303T-Data Science Duration: Two periods

Year& Sem: III Year /VI Sem Max.Marks:50

Course Articulation Matrix:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Course  Outcome | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 |
| CO3 | - | - | - | - | 1 | - | - | - | - | - | - | - |
| CO4 | - | - | - | - | 1 | - | - | - | - | - | - | - |
| CO5 | - | - | - | - | 1 | - | - | - | - | - | - | - |

**Note:** CO3 – To identify data manipulation and cleaning techniques using pandas

CO4 – To constructs the Graphs and plots to represent the data using python packages

CO5 – To apply the principles of the data science techniques to predict and forecast the outcome of real- world problem

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Part – A** (10 x 1 = 10 Marks)  *Instructions:*  1) Answer **ALL** questions.  2) The duration for answering Part A is **15 minutes** (this sheet will be collected after 15 minutes).  3**) Encircle the correct answer**. | | | | | | |
| S.No | Question | Marks | BL | CO | PO | PI Code |
| 1 | Which of the following is **NOT** a commonly used tool for **data wrangling?**  a) **Pandas** b) **NumPy** c) **Matplotlib** d) **OpenRefine** | 1 | 1 | 3 | 5 | 2.1.3 |
| 2 | Techniques used to handle the missing values?  a) permutation and imputation  b) insertion and deletion  c) imputation and deletion  d) insertion and deletion | 1 | 1 | 3 | 5 | 2.1.3 |
| 3 | A cricket analyst is looking at a player's scores from 10 matches: 45, 50, 60, 55, 48, 52, 49, 51, 47, 200. When should the score of 200 be considered an outlier and possibly removed from analysis?  a) When it is much higher than all the other scores b) When it is the highest score ever in cricket c) When it helped win the match d) When it is lower than all the other scores | 1 | 2 | 3 | 5 | 2.1.3 |
| 4 | What is the primary condition that requires reshaping or pivoting a dataset?  a) When the data is already in a well-organized format with one row for each observation. b) When the data is in a wide format with many columns and needs to be transformed into a fewer columns. c) When the data contains only numerical values and no categorical columns. d) When the dataset is too small to analyse effectively. | 1 | 2 | 3 | 5 | 2.1.3 |
| 5 | Which is not a Data transformation technique?  a) Attribute Construction  b) Smoothing  c) Data augmentation  d) data Discretization | 1 | 1 | 3 | 5 | 2.1.3 |
| 6 | Which of the Following statement true about Seaborn Library in python?  a) Provide High level interface with less complex syntax and default themes  b) Provides High level interface with high level complex syntax and customizable themes.  c) Provides interactive visualization library with complex syntax  d) Provide Interactive and web-ready visualization with no themes. | 1 | 1 | 4 | 5 | 2.1.3 |
| 7 | A teacher collected data on the number of hours students studied for a math test and their corresponding test scores. She plotted this data on a scatter plot, where the x-axis represents **hours studied** and the y-axis represents **test scores.** The scatter plot showed a cluster of points that generally increased from left to right.  a) There is a **negative correlation** between hours studied and test scores. b) The scatter plot shows **no relationship** between hours studied and test scores. c) There is a **positive correlation** between hours studied and test scores. d) Students who studied **less** always scored **higher** than those who studied more. | 1 | 2 | 4 | 5 | 2.1.3 |
| 8 | Which situation appropriate to use 3D plot?  a) It allows the analyst to ignore one of the variables and focus only on two. b) A 3D plot helps display the relationship between all three variables simultaneously. c) 3D plots are only used for representing time series data. d) It makes the data look more attractive, even if it doesn’t add any analytical value. | 1 | 2 | 4 | 5 | 2.1.3 |
| 9 | A school wants to compare the math test scores of students from three different classes (Class A, Class B, and Class C). The data science teacher uses Matplotlib to create a box plot for each class.What is the main reason for using a box plot in this situation?  a) To show the relationship between two continuous variables. b) To compare the spread, central tendency, and outliers of scores across the three classes. c) To identify the exact score of each student in each class. d) To visualize the trend of test scores over time. | 1 | 2 | 5 | 5 | 2.1.3 |
| 10 | Which of the following plot types is NOT directly available in the Matplotlib library?  a) Line Plot b) Violin Plot c) Heatmap d) Bar Plot | 1 | 1 | 5 | 5 | 2.1.3 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Register Number** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



**SRM Institute of Science and Technology**

Set -

**College of Engineering and Technology**

**School of Computing**

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamil Nadu

# Academic Year: 2024-25 (EVEN SEM)

Test: FT4 Date:29-04-2025

Course Code & Title: 21CSS303T-Data Science Duration: Two periods

Year& Sem: III Year /VI Sem Max.Marks:50

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Part – B** (4 x 5 = 20 Marks)  Instructions: Answer **ANY FOUR** Questions | | | | | | |
| Q.  No | Question | Marks | BL | CO | PO | PI Code |
| 11 | List out different approaches used to combine different datasets with example. Approaches to Combine Different Datasets:  1. **Concatenation (Vertical/Horizontal)**    * Example: pd.concat([df1, df2], axis=0) (vertical) or pd.concat([df1, df2], axis=1) (horizontal) 2. **Merging (SQL-style joins)**    * Example: pd.merge(df1, df2, on='common\_column', how='inner') 3. **Joining**    * Example: df1.join(df2, on='common\_column', how='left') 4. **Appending**    * Example: df1.append(df2, ignore\_index=True) 5. **Union**    * Example: Combining rows from two datasets with the same columns: pd.concat([df1, df2], axis=0, ignore\_index=True) 6. **Cross Join**    * Example: Using a Cartesian product to combine datasets: df1.merge(df2, how='cross') 7. **Concatenation by Index**    * Example: df1.append(df2, ignore\_index=False) | 5 | 2 | 3 | 5 | 2.1.2 |
| 12 | What are the conditions used to choose the data binning techniques with example? Conditions to Choose Data Binning Techniques:  1. **Nature of Data**    * Uniform Data: Equal-width Binning    * Skewed Data: Equal-frequency Binning 2. **Number of Bins**    * Fixed Number of Bins: Equal-width or Equal-frequency Binning    * Adaptive Binning: Custom Binning or Clustering-based Binning 3. **Distribution of Data**    * Normal Distribution: Equal-width Binning    * Non-Normal Distribution: Equal-frequency Binning 4. **Handling Outliers**    * Outlier-prone Data: Adaptive Binning or Clustering-based Binning 5. **Interpretability of Bins**    * Interpretable Bins: Custom Binning based on Domain Knowledge   **Example: Equal-width**: import numpy as np  data = np.random.normal(0, 1, 1000)  bins = np.linspace(min(data), max(data), 6) # 5 equal-width bins | 5 | 3 | 3 | 5 | 2.1.2 |
| 13 | What are the methods used to categorize the Noise and Outliers in the dataset? Methods to Categorize Noise and Outliers in a Dataset:  1. **Statistical Methods:**    * Z-Score (Standard Deviation Method)    * IQR (Interquartile Range) Method    * Modified Z-Score 2. **Visual Methods:**    * Box Plot    * Scatter Plot    * Histogram 3. **Machine Learning Methods:**    * DBSCAN (Density-Based Spatial Clustering of Applications with Noise)    * Isolation Forest    * One-Class SVM 4. **Domain Knowledge:**    * Expert-defined thresholds or rules for outlier detection 5. **Proximity-Based Methods:**    * k-Nearest Neighbors (k-NN)    * Local Outlier Factor (LOF)   These methods help identify and categorize noise and outliers based on statistical properties, clustering, or domain-specific rules. | 5 | 2 | 3 | 5 | 2.2.3 |
| 14 | Write the python code to plot 3D and Scatter plot using Matplotlib?  **3D –Plot code**:  import matplotlib.pyplot as plt  from mpl\_toolkits.mplot3d import Axes3D  import numpy as np  # Create data  x = np.linspace(-5, 5, 100)  y = np.linspace(-5, 5, 100)  x, y = np.meshgrid(x, y)  z = np.sin(np.sqrt(x\*\*2 + y\*\*2))  # Create 3D plot  fig = plt.figure(figsize=(8, 6))  ax = fig.add\_subplot(111, projection='3d')  # Plotting the surface  ax.plot\_surface(x, y, z, cmap='viridis')  # Labels and title  ax.set\_xlabel('X Axis')  ax.set\_ylabel('Y Axis')  ax.set\_zlabel('Z Axis')  ax.set\_title('3D Surface Plot')  # Show plot  plt.show()  **Scattar Plot:**  # Create data for scatter plot  x = [1, 2, 3, 4, 5]  y = [5, 4, 3, 2, 1]  # Create scatter plot  plt.scatter(x, y, color='blue', label='Data points')  # Adding labels and title  plt.xlabel('X Axis')  plt.ylabel('Y Axis')  plt.title('2D Scatter Plot')  plt.legend()  # Show plot  plt.show() | 5 | 3 | 4 | 5 | 2.2.3 |
| 15 | What are the different annotations used to plot the subplots in Matplotlib? Give example. ****Annotations in Matplotlib for Subplots**** In **Matplotlib**, annotations are used to add text, arrows, and other labels to a plot. Below are the different types of annotations commonly used when plotting subplots: plt.text()// Adds text at a specific (x, y) position on the plot. Example: plt.text(2, 3, 'This is a point', fontsize=12, color='red')plt.annotate()  * Adds an **annotation** with optional arrows, highlighting specific data points.   Example: plt.annotate('Peak Point', xy=(3, 5), xytext=(4, 6), arrowprops=dict(facecolor='blue', arrowstyle='->')) ax.text()  * Similar to plt.text(), but used within a specific **Axes** object for **subplots**.   Example: fig, ax = plt.subplots()  ax.text(2, 3, 'Subplot Text', fontsize=12, color='green') ax.annotate()  * Used for **subplots** to annotate with arrows or text at a given data point.   Example: fig, ax = plt.subplots()  ax.plot([1, 2, 3], [4, 5, 6])  ax.annotate('Max Point', xy=(3, 6), xytext=(2, 5), arrowprops=dict(facecolor='red', arrowstyle='->')) Explanation includes:  * **plt.text() and ax.text()** are used to add simple text annotations at specific locations. * **plt.annotate() and ax.annotate()** are used for adding more complex annotations with arrows and labels pointing to specific data points. | 5 | 3 | 4 | 5 | 2.2.3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Part – C (2 x 10 = 20 Marks)**  Instructions: Answer ALL questions. | | | | | | |
| Q.  No | Question | Marks | BL | CO | PO | PI  Code |
| 16 a | Discuss about various Data transformation techniques in detail with example.  **Answer key:** Discussion about transformation techniques with example. ****List of Data Transformation Techniques Used in Data Science:****  * Normalization * Standardization * Log Transformation * Power Transformation * Binning (Discretization) * Encoding Categorical Variables * One-Hot Encoding * Label Encoding * Feature Scaling * Quantile Transformation * PCA (Principal Component Analysis) * Polynomial Transformation * Handling Skewed Data * Text Vectorization (TF-IDF, Count Vectorizer) * Date-Time Feature Extraction | 10 | 2 | 3 | 5 | 2.2.3 |
| **(OR)** | | | | | | |
| 16 b | Consider you are a data analyst for a smart city initiative that monitors Electric Vehicle (EV) charging station usage across different locations. Your goal is to clean, transform, and analyze the data to optimize charging station efficiency, reduce waiting times, and improve user experience. The dataset contains EV charging session logs collected from multiple charging stations and includes the following attributes: Session ID, User ID, Station ID, Location, Charging Start Time, Charging End Time, Charging Duration, Energy Consumed (kWh),Cost ($),Payment Method etc.,  Apply various type of Data wrangling techniques to clean and pre-process the dataset for further analysis with example.  **Answer Key:** ****Data Wrangling Techniques for EV Charging Station Dataset:****  1. **Handling Missing Values**    * Example: Fill missing Payment Method with mode or mark as "Unknown". 2. **Data Type Conversion**    * Example: Convert Charging Start Time and Charging End Time to datetime. 3. **Feature Engineering**    * Example: Calculate Charging Duration using start and end times if missing. 4. **Removing Duplicates**    * Example: Drop duplicate Session ID entries. 5. **Normalization/Standardization**    * Example: Scale Energy Consumed and Cost for machine learning models. 6. **Filtering Invalid Data**    * Example: Remove records with negative Charging Duration or Energy Consumed. 7. **String Cleaning**    * Example: Strip whitespace from Location and Payment Method. 8. **Data Aggregation**    * Example: Group by Station ID to calculate total energy used or peak hours. 9. **Date-Time Feature Extraction**    * Example: Extract hour/day/week from Charging Start Time to find usage patterns. 10. **Outlier Detection and Treatment**  * Example: Identify outlier sessions with extremely high Cost or duration using IQR. | 10 | 3 | 5 | 5 | 3.3.1 |
|  | | | | | | |
| 17 a | Discuss about Matplotlib configuration using different plot styles with example python code and graphs.  **Answer Key:**  Answer may include various element configuration to plot the graph under the Matplotlib with sample code. Matplotlib Configuration Using Different Plot Styles (with Code & Graphs)  1. **Matplotlib supports multiple built-in plot styles** to change the appearance of graphs easily. 2. **Common Styles**:    * 'default': Standard Matplotlib look    * 'ggplot': Inspired by R's ggplot2    * 'seaborn': Attractive statistical plots | 10 | 2 | 4 | 5 | 2.2.3 |
| **(OR)** | | | | | | |
| 17 b | Consider a healthcare data analyst at a research institute studying the connection between dietary habits and common lifestyle-related diseases. A survey was conducted across different age groups, and the collected data includes:   * Participant\_ID * Age\_Group (e.g., Teen, Adult, Senior) * Diet\_Type (e.g., Vegetarian, Non-Vegetarian, Vegan, Junk Food) * Common\_Disease (e.g., Obesity, Diabetes, Hypertension, Heart Disease, None) * Exercise\_Hours\_per\_Week   Write the sample python code for visualization and plot the suitable graphs using Matplotlib for the following:  i) Show how many people follow each diet type? ii) Visualize which diet types are more frequently associated with specific diseases? iii) Identify which age groups are more prone to specific diseases? iv) Plot the distribution of exercise hours for people with and without diseases?  **Answer Key:**   1. **Bar Plot for Diet Types**: Displays the count of people following each diet type. 2. **Stacked Bar Plot for Diet and Disease**: Visualizes the relationship between diet types and common diseases using a stacked bar chart. 3. **Stacked Bar Plot for Age Group and Disease**: Shows which age groups are more prone to specific diseases. 4. **Box Plot for Exercise Hours and Disease**: Compares the distribution of exercise hours between people with and without diseases.   **Example plots:**        These visualizations give a clear understanding of the dataset in terms of diet, age group, disease prevalence, and exercise habits. | 10 | 3 | 5 | 5 | 3.3.1 |

**Course Outcome (CO) and Bloom’s level (BL) Coverage in Questions:**

|  |  |
| --- | --- |
|  |  |