|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | Select the correct flow of data wrangling   1. Data Collection → Data Cleaning → Data Transformation → Data Analysis 2. Data Collection → Data Modeling → Data Visualization → Data Storage 3. Data Cleaning → Data Collection → Data Storage → Data Reporting 4. Data Analysis → Data Visualization → Data Cleaning → Data Collection | 1 | 1 | 3 | 5 | 2.1.3 |
| 2 | What does the following code snippet do?  df.fillna(0, inplace=True)  a) Drops all rows with NaN values b) Replaces all NaN values with 0 c) Removes duplicate rows d) Standardizes the dataset | 1 | 1 | 3 | 5 | 2.1.3 |
| 3 | Which function is used to combine two datasets by matching rows on a specific column?  a) concat() b) pivot() c) groupby()  d) merge() | 1 | 1 | 3 | 5 | 2.1.3 |
| 4 | What is the purpose of the pd.concat() function?  a) To combine datasets horizontally or vertically  b) To find missing data  c) To standardize columns  d) To create histograms | 1 | 2 | 3 | 5 | 2.1.3 |
| 5 | You have a column with mixed data types (e.g., numbers and strings). How can you clean it for numerical analysis?  a) Convert all values to strings b) Remove non-numeric values or replace them with NaN c) Sort the values in ascending order d) Create a pivot table | 1 | 2 | 3 | 5 | 2.1.3 |
| 6 | What is the primary purpose of the annotate() function in Matplotlib?  a) To add legends to the plot b) To add text annotations to specific points on the plot c) To configure plot styles d) To create 3D surface plots | 1 | 1 | 4 | 5 | 2.1.3 |
| 7 | Which of the following plots is best suited for visualizing relationships between multiple variables in a dataset?  a) Histogram b) Line plot c) Pair plot d) Box plot | 1 | 1 | 4 | 5 | 2.1.3 |
| 8 | Which of the following code snippets correctly sets the style of a Matplotlib  plot to "seaborn-darkgrid"?  a) plt.style.use('seaborn-darkgrid') b) sns.set\_style('darkgrid') c) plt.set\_style('seaborn-darkgrid') d) sns.use('seaborn-darkgrid') | 1 | 2 | 4 | 5 | 2.1.3 |
| 9 | You want to visualize the distribution of a variable along with its density. Which Seaborn function should you use?  a) sns.scatterplot()  b) sns.histplot()  c) sns.lineplot()  d) sns.boxplot() | 1 | 1 | 5 | 5 | 2.1.3 |
| 10 | What does the following Seaborn function achieve?  sns.pairplot(data, hue='species')  a) Plots a scatter plot matrix for all numerical variables b) Creates histograms for all numerical variables c) Creates a box plot for all numerical variables d) Plots a scatter plot matrix grouped by species | 1 | 2 | 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 | Discuss the general programming tips to deal with large data sets. | 5 | 2 | 3 | 5 | 2.1.2 |
| 12 | When merging two DataFrames in pandas that have columns with the same name, how can you ensure the column names are distinguishable? | 5 | 3 | 3 | 5 | 2.1.2 |
| 13 | Given the dataset data ={'Ages': [3, 18, 22, 10, 25, 29, 34, 14, 40, 45, 50, 55, 60, 12, 65, 70, 75, 80, 85]}, categorize the continuous Ages values into the groups of children, young, middle, and elder. Define appropriate age ranges for each category and implement the conversion. | 5 | 3 | 3 | 5 | 2.2.3 |
| 14 | Compare a box plot and a histogram, highlighting their use cases and strengths. | 5 | 3 | 4 | 5 | 2.2.3 |
| 15 | How can you control the line properties (e.g., color, style, and width) of a chart in Matplotlib. Write the python code and explain. | 5 | 3 | 5 | 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 | You are working as a data analyst for a retail company. The company has provided you with a dataset containing sales information from multiple regions. However, the dataset has several issues:   * **Missing Data**: Some rows in the Sales column are blank, while some entries in the Region column are marked as "Unknown." * **Irregular Formats**: The dates in the Date column are in various formats like 01-01-2023, 2023/01/01, and 1st Jan 2023. * **Duplicate Records**: There are rows where the same Product and Region are listed multiple times for the same Date. * **Irrelevant Data**: There are columns in the dataset, such as Transaction ID, which are irrelevant for the analysis. * **Outliers**: There are unusually high Sales values, such as $1,000,000, which seem unrealistic. * **Categorical Inconsistencies**: The Product column has entries like "Appl" instead of "Apple" and "Bananaa" instead of "Banana." * **Merging**: The dataset also needs to be combined with a separate dataset containing information about the profit margins for each product and region.   **Question:**  You have been tasked with preparing the dataset for analysis. Outline the data wrangling steps you would perform to clean, transform, and structure the data. Discuss how you would handle missing values, standardize formats, remove duplicates, filter irrelevant data, detect and handle outliers, resolve inconsistencies, and merge datasets. Ensure your process results in a clean and analysis-ready dataset. | 10 | 3 | 3 | 5 | 2.2.3 |
| **(OR)** | | | | | | |
| 16 b | data = {  'Date': ['2023-01-01', '2023-01-01', '2023-01-02', '2023-01-02'],  'Product': ['Apple', 'Banana', 'Apple', 'Banana'],  'Region': ['North', 'North', 'South', 'South'],  'Sales': [100, 150, 200, 50]  }  df = pd.DataFrame(data)  Write the output of the following commands using the above code snippet and discuss the output   * pivot\_df = df.pivot(index='Date', columns='Product', values =   'Sales')   * stacked\_df = df.stack() * stacket\_pivot= pivot\_df.stack() | 10 | 3 | 3 | 5 | 3.3.1 |
|  | | | | | | |
| 17 a | Discuss the functionalities of the Seaborn library for advanced visualizations. Provide examples of pair plots, box plots, and histograms. | 10 | 2 | 4 | 5 | 2.2.3 |
| **(OR)** | | | | | | |
| 17 b | Illustrate how to create and interpret 3D surface plots. Discuss scenarios where such visualizations are beneficial. | 10 | 2 | 5 | 5 | 3.3.1 |

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

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