Somaiya Vidyavihar University
K. J. Somaiya College of Engineering
Department of Electronics and Telecommunication Engineering
Semester: July 2024 – Oct 2024

Max. Marks: **Class: TY B.Tech_DE**
20 Semester: V **Branch: EXTC**
Subject: Principles of Data Science (Departmental Elective-116U03E515)

IA2_Project

Roll No	**Name of Student**	**IA-2**
16010321101	Kanse Vinay Sitaram Pushpalata	**3-Python**

Guidelines:

1. Your respective assignment is mentioned in the table above. Refer to the serial number of the assignment and the tool you will use.

For eg: 1-Excel means Sr No 1 Dataset and tool used is Excel.

2. **In case of Datasets 14 & 15, there are multiple files present in the link, students should accordingly refer to their allotted dataset.**

For eg: 15-Excel-Apple means Sr No 15 Dataset, tool used is Excel and file to be used of the dataset is Apple.

3. For students receiving Excel Assignment, each question should be answered as a new sheet in a workbook. The **entire workbook** must then be added in a folder along with the downloaded dataset which should be submitted on a drive link which will be shared. Along with this, a **report** should be prepared answering all the questions. Attach all necessary screenshots and a conclusion drawn for every question.

4. For students using Python/R, must share the **entire source code** along with the downloaded dataset in a folder. This will be submitted on a drive link which will be shared. Along with this, a **report** should be prepared answering all the questions. Attach all necessary screenshots and a conclusion drawn for every question.

5. Reports should be submitted in **PDF format** only.

6. **Exploratory Data Analysis** (Cleaning & Modifications if required) are to be conducted for all the datasets before answering the questions.

3) Dataset: Titanic

Link: https://docs.google.com/spreadsheets/d/1SF7_RQi8nxf6ppd8cp2ZYh-BCwGVqp4QKf0ze0eUyT8/edit?ref=hackernoon.com#gid=116838508
This popular open-source dataset offers information on the passengers onboard the Titanic ship when it sank on April 15, 1912. It can be used by data analytics beginners interested in data cleaning and preprocessing, descriptive statistics, data visualization and predictive modeling.
Some of the variables included in the dataset:

- PassengerId - A unique identifier for each passenger.
- Survived - This shows whether the passenger survived or not (0 = No, 1 = Yes).
- Pclass - A passenger's class (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name - A passenger's name.
- Sex - A passenger's gender.
- Age - A passenger's age.
- SibSp - The number of siblings/spouses aboard.
- Parch - The number of parents/children aboard.
- Ticket - The ticket number.
- Fare - The fare paid for the ticket.
- Cabin - The cabin number.
- Embarked - The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Questions (For Python/R):

1. What is the survival rate of the passengers? Calculate the percentage of passengers who survived compared to the total number of passengers.
2. What is the proportion of male and female passengers? Calculate the percentage of male and female passengers in the dataset.
3. Which class of passengers had the highest survival rate? Calculate the survival rate for each passenger class (1st, 2nd, and 3rd) and determine which class had the highest survival rate.
4. How many passengers had siblings or spouses aboard? Calculate the number of passengers who had siblings or spouses aboard (SibSp > 0).
5. What is the average fare paid by passengers in each class? Calculate the mean fare for each passenger class (1st, 2nd, and 3rd).
6. How many passengers embarked from each port? Calculate the count of passengers who embarked from each port (Cherbourg, Queenstown, Southampton).
7. What is the age distribution of the passengers? Create a histogram or kernel density plot to visualize the distribution of passenger ages.
8. What is the fare distribution of the passengers for each class? Create a box plot or violin plot to visualize the fare distribution for each passenger class (1st, 2nd, and 3rd).
9. What is the average survival rate of male and female passengers? Calculate the average survival rate for both male and female passengers.
10. How many unique ticket numbers are there? Use the Pandas library to find the count of unique ticket numbers in the dataset.