**Individual Assignment 1 (10%)**

Duration: 24 hours

Description of the dataset (titanic.csv)

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Key** |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

**Variable Notes**

**pclass**: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower  
  
**age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
**sibsp**: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
**parch**: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

**Tasks for the assignment:**

1. Python Basics

* Task 1: Write a Python program that:
  1. Asks the user for their name and age.
  2. Prints a message: "Hello [name], you will turn 100 years old in [year]." (Assume the current year is 2025).
  3. Create a function calculate\_age\_in\_100() that takes the current age as input and returns the year when the person will turn 100.
* Task 2: Create a function generate\_summary() that:
  1. Takes a list of numbers as input.
  2. Returns the mean, median, and standard deviation of the numbers.
  3. Handles the case when the list is empty by returning a message: "List is empty."

2. Data Importing and Exploration

* Task 1: Import the Titanic\_Passengers.csv file into a pandas DataFrame.
* Task 2: Display the first 10 rows of the dataset.
* Task 3: Show the data types of all columns using dtypes and identify any potential issues (e.g., a numeric column stored as a string).
* Task 4: Identify the number of missing values in each column and determine a strategy to handle these missing values.
  + For example, for Age, can we use imputation with the median or should we drop rows with missing values?

3. Data Cleaning and Transformation

* Task 1: Handle missing values in the dataset:
  1. Fill missing values in the Age column with the median of the Age column.
  2. Drop rows with missing values in the Embarked column.
* Task 2: Convert the Sex column to a numerical format:
  1. Map male to 0 and female to 1.
* Task 3: Create a new column, FamilySize, by adding the SibSp and Parch columns together.
* Task 4: Perform the following transformations:
  1. Create a new column FarePerPerson, which divides the Fare by FamilySize.
  2. Remove any duplicate rows based on PassengerId.

4. Data Manipulation and Aggregation

* Task 1: Filter the dataset to show only passengers who survived (Survived = 1), but exclude passengers who are under 18 years old.
* Task 2: Group the dataset by Pclass and calculate the following:
  1. Average age of passengers in each class.
  2. Average fare in each class.
  3. Total number of survivors in each class.
* Task 3: Create a pivot table to analyze the relationship between Embarked (port of embarkation) and Survived. Calculate the survival rate for each port.
* Task 4: Filter the dataset to show only the passengers who paid more than the average Fare and survived. What do these passengers have in common (age, class, family size)?

5. Data Visualization

* Task 1: Plot a histogram of the Age distribution with bins that group ages into 10-year intervals.
* Task 2: Create a bar plot comparing the survival rate for each embarkation port (Embarked).
* Task 3: Create a scatter plot comparing Fare vs Age. Use color to indicate whether the passenger survived or not.
* Task 4: Create a box plot comparing the distribution of Fare across different passenger classes (Pclass).

6. Statistical Analysis

* Task 1: Calculate the mean, median, and standard deviation of the Age and Fare columns. Which column has a greater spread of values?
* Task 2: Perform a t-test to compare the average Fare between male and female passengers. Report whether there is a significant difference.
* Task 3: Perform a chi-square test to assess if there is an association between Survived and Pclass. Is the survival rate significantly different across classes?
* Task 4: Using scipy.stats, calculate the correlation between Age and Fare. What does the correlation tell you about the relationship between these two variables?