**National Higher School of Statistics Algeria**

M: SC with R Mr. Hocine Abdelouaheb Date: 23/01/2025

**Notes :**

1. No need to use any generative AI tool work with the documentation of Tidyverse.
2. The test should be R (Script) or Rmd (Notebook)
3. **The deadline Saturday 25 January before 5 pm**

**Data description:**

Csv file contains the data of Cancer Patient.

**Variables:**

* **PatientID**: Unique identifier for each patient.
* **Age**: Age of the patient in years.
* Gender: Gender of the patient (e.g., Male, Female).
* **RaceEthn**: Race or ethnicity of the patient (e.g., Caucasian, Asian, African American, Hispanic, Other).
* **BMI**: Body Mass Index of the patient, representing their weight-to-height ratio.
* **SmokingSt**: Smoking status of the patient (e.g., Smoker, Former Smoker, Non-Smoker).
* **FamilyHist**: Indicates whether there is a family history of cancer (e.g., Yes, No).
* **CancerTyp**: Type of cancer diagnosed (e.g., Breast, Colon, Lung, Skin, Prostate, Leukemia).
* Stage: Stage of cancer at diagnosis (e.g., I, II, III, IV).
* **TumorSize**: Size of the tumor measured in centimeters.
* Treatment: Treatment received by the patient (e.g., Surgery, Radiation, Chemotherapy, Combination Therapy).
* **TreatmentResponse**: Response to the treatment (e.g., Complete Remission, Partial Remission, No Response).
* **SurvivalMo**: Number of months the patient survived post-diagnosis.
* Recurrence: Indicates whether cancer recurred after treatment (e.g., Yes, No).
* **GeneticM**: Presence of specific genetic mutations related to cancer (e.g., BRCA1, KRAS, EGFR, None).
* **HospitalRegion**: Geographic region of the hospital where the patient received treatment (e.g., N

**First :**

Exploratory data Analysis

Use Tidyverse library to answer the following questions?

1. What is the distribution of cancer types?
2. What is the average BMI by cancer stage?
3. What is the survival distribution by hospital region?
4. What is the tumor size distribution?
5. What is the relationship between tumor size and survival months?
6. What is the proportion of smokers in the dataset?
7. What is the average tumor size by cancer type?
8. Is there a difference in average BMI between smokers and non-smokers?
9. What is the distribution of cancer stages?
10. How many patients are in each hospital region?
11. What is the survival rate by cancer type?
12. What is the relationship between age and survival months?
13. What is the most common treatment by cancer type?
14. What is the gender distribution of patients?
15. How does smoking status vary by hospital region?
16. What is the average survival months by genetic mutation type?
17. What is the distribution of patients by age group?
18. What are the most common combinations of cancer type and stage?
19. What is the relationship between BMI and tumor size?
20. How many patients have a family history of cancer?

**Second :**

Function creation and error handling

1. Create a function to classify age groups.
2. Create a function to normalize tumor size.
3. Create a function to calculate BMI categories.
4. Write a function to check if survival months are valid (non-negative).
5. Develop a function to calculate tumor size as a percentage of maximum size.
6. Write a function to generate descriptive statistics for a numeric column.
7. Create a function to filter data by cancer type and stage.
8. Write a function to compute the average survival months for each region.
9. Develop a function to create a summary table for a given variable.
10. Write a function to identify patients with above-average tumor sizes.
11. Create a function to categorize survival months into short, medium, and long terms.
12. Develop a function to check for missing values in each column.
13. Write a function to calculate the median tumor size for smokers and non-smokers.
14. Create a function to count the number of patients by gender and cancer type.
15. Write a function to group data by region and calculate average BMI.
16. Develop a function to identify patients with genetic mutations linked to survival rates.
17. Write a function to summarize treatment response rates by cancer stage.
18. Create a function to check for outliers in numeric columns.
19. Write a function to calculate the proportion of patients with recurrence by region.
20. Develop a function to flag invalid BMI values (e.g., below 10 or above 60).