Derivable 1

1:The data analysis presented here is aimed at probing and clarifying the relationships between various medical and personal factors like age, cancer type, stage, treatment, and lifestyle and patient outcomes. The analysis further aims to predict outcomes and pinpoint high-risk cases through the application of supervised machine learning techniques. Doing this, allowing enhanced medical decisions, more efficient treatment plans, and the public of early detection and intervention. The whole process would eventually lead to the generation of insights driven by data which would not only make healthcare quality better but also increase patient survival rates and make more effective, personalized treatment regimens possible.

2:

Python programming is used for data analysis because it supports both procedural and object-oriented programming. It allows performing procedural steps such as data cleaning, analysis, and visualization, while also using object-oriented features through data structures like DataFrames. With powerful libraries such as pandas, NumPy, and matplotlib, Python makes analyzing, processing, and visualizing data fast and efficient.

3:

**Type of Machine Learning Algorithm**

In this data we are using Supervised Learning because the dataset contains labeleddata, where the Outcome ( EX: *Recovered*, *Under Treatment*, *Deceased*) is known for each patient.

This algorithm will help and allow the model to learn from past records and predict the outcome for the new patients based on their data (EX: age, cancer type, treatment type).

**Purpose of the Algorithm**

The purpose of implementing this algorithm is to predict cancer patient outcomes and identify the risk levels based on the medical and personal data such as age, cancer type, stage and treatment type.

By doing that the Algorithm will help to improve the medical decision, improving the treatment planning, and early identification of high-risk patients.

4: **The independent variables** of our chosen dataset are: Age, Gender, Nationality, Emirate, Cancer\_Type, Cancer\_Stage, Treatment\_Type, Treatment\_Start\_Date, Smoking\_Status, Comorbidities, Ethnicity, Weight, Height, Hospital, Primary\_Physician .

These are the factors that might affect or predict the probability of developing cancer or its severity, they are not affected by the disease itself.

**The dependent variables** of our chosen dataset are: Outcome, Death\_Date, Cause\_of\_Death, shaimaa

The outcome represents the final health status of each patient this variable depends on various factors like age, cancer stage and treatment type. The death\_date and cause of death variables are affected by the disease progress and the overall treatment process.

These variables are influenced by the independent variable.

5:

Yes, sampling will be done, the Random sampling is chosen because it gives every record in the dataset an equal chance of being selected, which helps avoid bias and ensures the sample represents the overall population**.** This method is simple, effective, and suitable for large datasets like the cancer dataset, allowing accurate and fair statistical analysis.

**Deliverable 2**

Q1. Justify why you want to perform the descriptive analysis for the chosen dataset.?

- To summarize cancer characteristics It allows us to explore patterns in cancer type, stage, and treatment. This helps determine which cancer types are most common and which stages are frequently diagnosed.

-To understand the population structure. Descriptive analysis provides an overview of patient demographics such as age, gender, nationality, and emirate which helps identify which groups are most affected by cancer in the UAE. Overall, descriptive statistics make the large dataset easier to interpret and support better public health insights.

Q2. A **Python function** was created to calculate descriptive statistics for any field in the dataset.  
 The function includes measures of:

* **Central Tendency:** Mean
* **Dispersion:** Interquartile Range (IQR)
* **Position:** Skewness (Manual Formula)
* **Frequency:** Frequency Table

This reusable function simplifies data analysis and helps describe numerical and categorical variables efficiently.

Q3.

**Q5 - Create a detailed descriptive statistics report about the dependent variable**

**of the chosen dataset.**

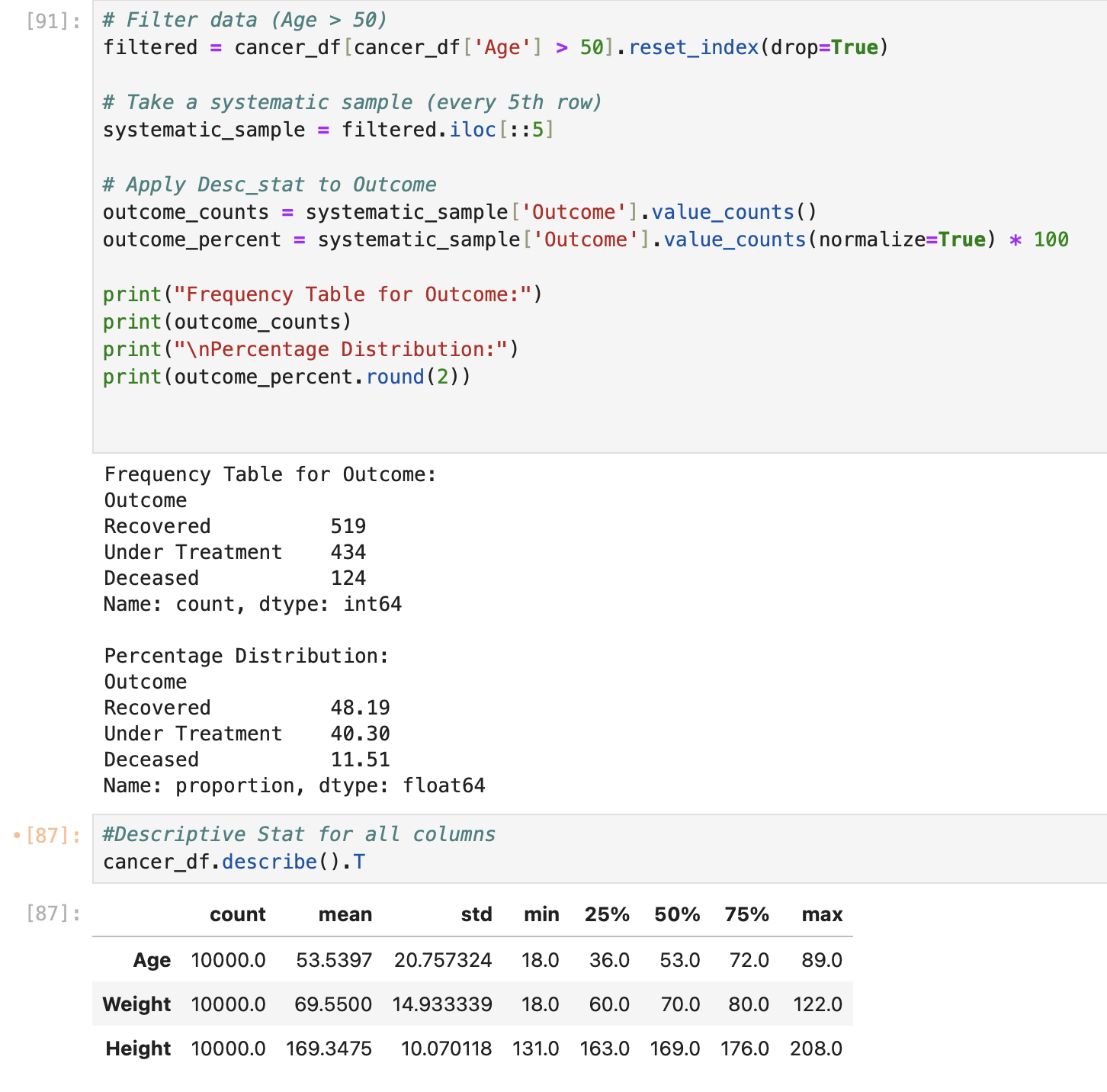
Variable description

The dependent variable in this project is the outcome, which represents the patients health status after cancer treatment . it shows whether the patient has recovered, undertreatment or deceased.

Data and sampling

The data was filtered to include only patients above 50 years old and then the systematic sample , every 5th record was taken to make sure the sample was evenly distributed and representative of the data set.

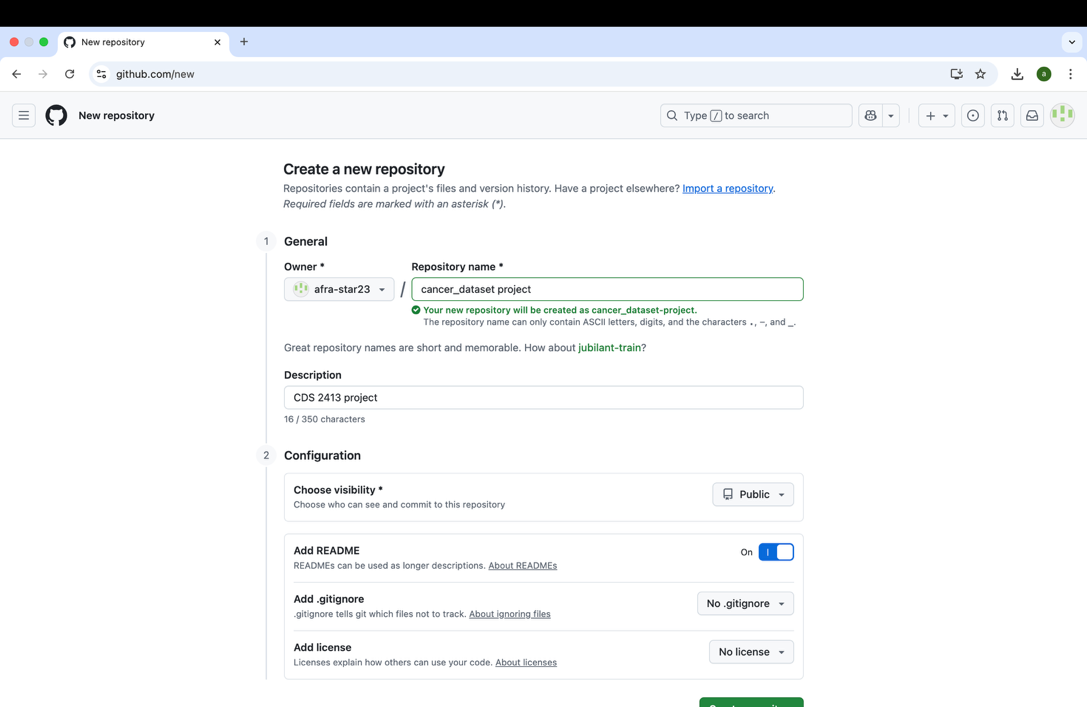
outcome



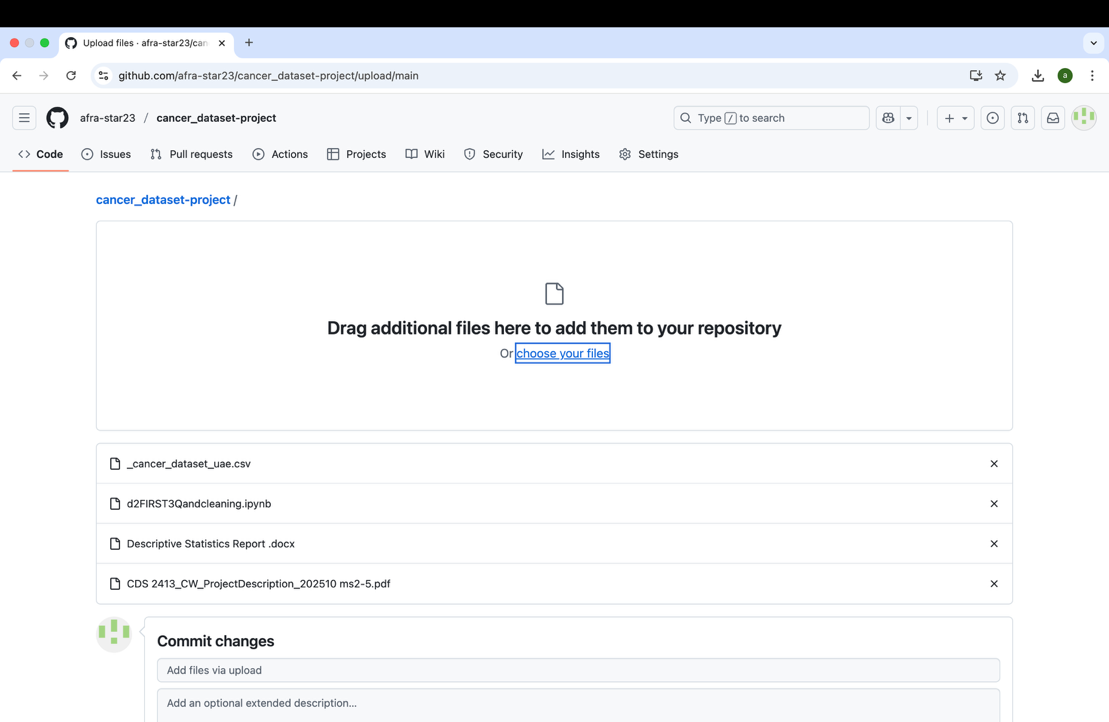
From the output we can see that most patients have recovered by a percentage of 48%,around 40% are still under treatment and only 11% are deceased.This shows that the majority of patients are in good condition, and overall, the treatment outcomes are positive, especially for patients above 50 years old. In conclusion The analysis of the Outcome variable shows that most patients have improved or currently under medical care, while a small percentage have passed away.In general, the results indicate \*effective medical care\* and positive recovery rates among cancer patients in this dataset.

**Deliverable 4**

16) Create a new repo for project in GitHub



17) Upload all the project files created for CLO1,CLO2 and CLO3 to the GitHub repo.



18) Configure Git with GitHub

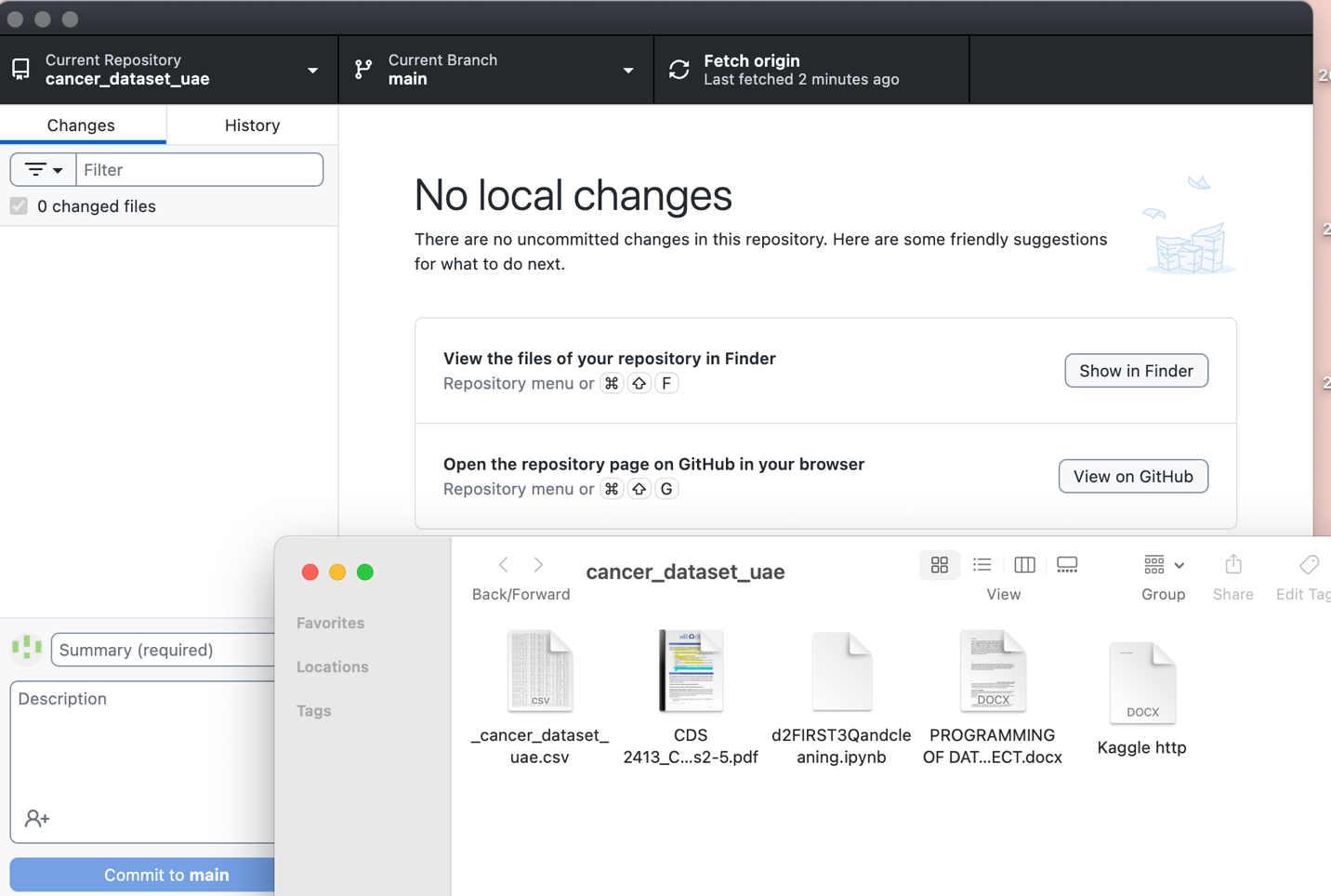
A screenshot of a computer

AI-generated content may be incorrect.

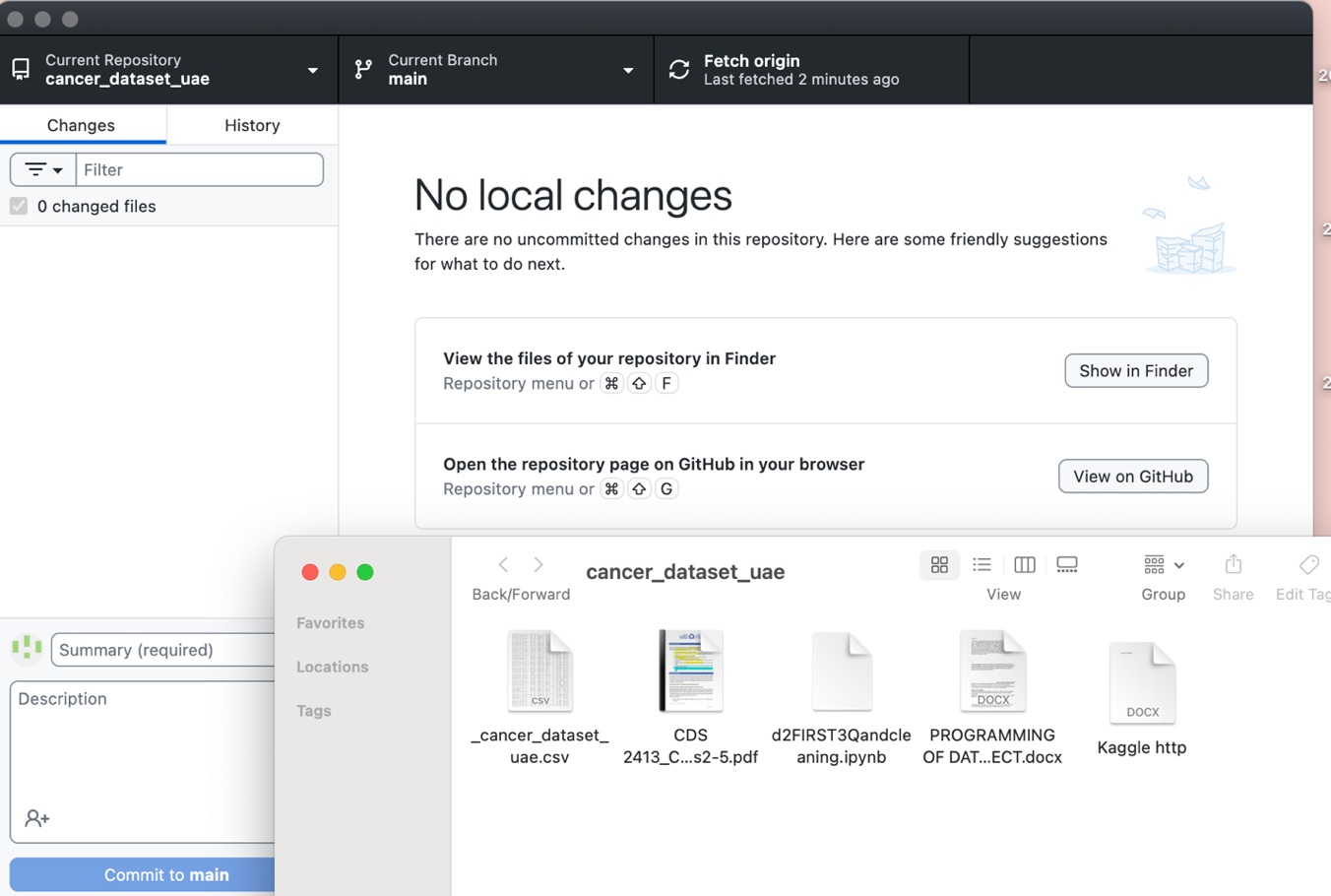
19) Clone Github repo to Git



20) Pull any file from GitHub repo to Git



20) Pull any file from GitHub repo to Git



21) Modify the pulled file and push the modified file to GitHub

A screenshot of a computer

AI-generated content may be incorrect.