Patient Health Data Analysis

By Amal.V.S

Analysis Requirements

1. Age Grouping - Group patients into age categories (e.g., 18-30, 31-45, 46-60, 61+). Calculate the average blood pressure for each age group.

2. Cholesterol Statistics - Calculate and print the minimum, maximum, and average cholesterol levels.

3. Diabetes Prevalence - Determine the prevalence of diabetes among different age groups.

Requirements

1. Use the `pandas` library for data manipulation.

2. Implement functions for each analysis task.

3. Include error handling for cases where the input file is missing or the data is not in the expected format.

4. Provide a simple command-line interface to specify the input file path. **(Not Required)**

Jupyter Notebook Explanation

- First Cell of code is to import the required libraries.
- Second cell is to give the **input file path of the CSV file**. The path of your CSV file must be given here or else data set cannot be loaded. *(The Existing file path is mine)*
- Execute the Third cell to Load the data set. Any error occurred while loading the data set will be handled.

The code is Divided in to 6 sections: **Basic EDA, Splitting BP, Data Visualization, Task-1, Task-2, Task-3**

1. Basic EDA

This Section is all about Exploring the data set. Running every cell in this Section will give a clear idea about the data set. This cell shows the following:

- Top 10 Columns of the data set
- Shape of the data set
- Data set Info
- Describing the Data set with respect to Numerical Columns and Categorical Columns
- Checking for Null Values – ***Found that No Null values were present in the data set.***
- Checking for Duplicate Values - ***Found that No Duplicates values were present in the data set.***

2. Splitting Blood Pressure Feature in to two - Systolic & Diastolic (Very Important)

This section is a really Important Section as the Blood Pressure data is in the format of Systolic BP/Diastolic BP. This data must be split in to two individually and converted to numerical value to get insights from Blood Pressure Data. **So, Running this Section is Very Important.**

- Running the first cell in this section will Split the data and make it in to numerical values and will create two new columns called Systolic_BP and Diastolic_BP and will be added to the main Data Frame.
- Third and Fourth Cell in this Section will Remove the Initial Blood Pressure Column and Patient Id Column, which are unwanted Columns that give no insights for our Analysis. **So, Running this Two cells is also very Important.**

3. Visualising the Data set

Running the Cells in this Section will give insights on how the data is Distributed and also, we could find any patterns form the visualisations.

- The First cell in this Section will find the Numerical and Categorical Columns in the data set as it is necessary for plotting charts.
- The Second cell will display the Visualisation of Numerical Columns. Since, all the Numerical columns in the data set are Continuous data, Histogram is used to Plot the Numerical Data Distribution.
  Insights: It is found that major portion of age in the data set lies between 30 and 60.
- The Third cell in this Section will Visualise the Categorical Column. Since, there is only one Categorical Column in the data set, the column name is directly given for plotting Count Plot.
  Insights: It is found that "High" Cholesterol value is the highest count and "Low" Cholesterol value is the Lowest Count.
- The fourth cell in this Section will Visualise Cholesterol count with hue as Diabetes to understand the number of people with the respective Cholesterol values who have diabetes or not.
  Insights: It is found that there are no people having diabetes who have cholesterol level as Low or Normal. Also, Only the people with high cholesterol value as High are having Diabetes and also note that all the people in the data set with Cholesterol level as High are having diabetes.
- The Fifth Cell in this section will Visualise sum of Diabetes occurrences by cholesterol level. This Visualisation will strength to our insight that there are no people having diabetes in Low and Normal Cholesterol level. Also, it says only people having high cholesterol level are having Diabetes as Positive.
- The Sixth Cell in this section will Visualise the Count of Diabetes Variable. From the plot, we can understand that the number of False values is more than True values in the Diabetes Column.

4. Task - 1 - Age Grouping

This Section will Group patients into age categories (18-30, 31-45, 46-60, 61+) and the average blood pressure for each age group is Calculated and Printed.

- The first cell in this section is the function for solving the task 1. Just running this cell will create a function for solving task 1. The function returns a pandas series with columns Solystic_BP, Diastolic_BP and

Average Blood Pressure. The values in it are the average values of respective columns for each age group.

- The second cell in this section will call the function with our main data frame as argument and the result returned by function will be stored in a variable and printed.

Insights: The analysis of average blood pressure values reveals a distinct pattern among age categories: individuals aged 18-30 and 31-45 exhibit normal blood pressure levels, while those in the 46-60 and 60+ age groups tend to have higher blood pressure, indicating a potential age-related trend in blood pressure distribution. The conclusion was made by referring this article : here

5. Task - 2 - Cholesterol Statistics

This Section will calculate and print the minimum, maximum, and average cholesterol levels.

- The first cell in this section is the function for solving the task 2. Just running this cell will create a function for solving task 2. The function returns a pandas Data Frame with columns Statistic, Cholesterol_Level, Cholesterol_Category. Statistics Column have values of our requirement. Cholesterol_Level have the result values in numerical. Cholesterol_Category have the result values in Categorical as it were in the initial dataset.
- The second cell in this section will call the function with our main data frame as argument and the result returned by function will be stored in a variable and printed.

Insights: The Minimum Cholesterol value is Low. The Maximum Cholesterol value is High. The Average Cholesterol value is Normal.

6. Task - 3 - Diabetes Prevalence

This section will determine the prevalence of diabetes among different age groups.

- The first cell in this section is the function for solving the task 3. Just running this cell will create a function for solving task 3. The function returns a pandas Data Frame with column Diabetes Percentage(%) were the values in it are Percentage of People having Diabetes as positive with respect each age category.
- The second cell in this section will call the function with our main data frame as argument and the result returned by function will be stored in a variable and printed.

Insights: The Percentage of people having diabetes in Age categories 18-30 & 31-45 are very low that is only 9% and 6% respectively. On the other hand, Percentage of people having diabetes in Age categories 46-60 & 60+ are very High that is about 82% and 100%. The main thing to note is that not even person above 60 age is with no Diabetes, every person above 60 age are having diabetes in the data set.