# مبادرة رواد مصر الرقمية

DEPI Graduation Project

HealthCare Predictive analysis

**Team Members:**

* Amina Zeyad Atta Ibrahim
* Amr Ashour Mohamed Ali
* Elham Mokhtar Ahmed Mohammed
* Enas Samir Abbas Mohamed
* Esraa Mohamed Sayed Mohamed

**Training Company:** CLS

**Group Number:** CLS\_\_CAI2\_AIS4\_G4

**Instructor:** Mahmoud El-Sayed

# **HealthCare Predictive analysis**

Contents

[1. Project Overview 3](#_heading=h.amleyvsthfjk)

[2. Milestone1 3](#_heading=h.joleahk7gptk)

[2.1 Data Collection: 3](#_heading=h.u0a8ml366yij)

[2.2 Data Exploration: 3](#_heading=h.r61pnbffc4ti)

[2.2.1 Check Data Statistics; First 5 records, Last 5 records, NULLs , Duplicates in Dataset and check if there is any categorical column needs encoding: 4](#_heading=h.5h6g7lv2frcw)

[2.2.2 Data Distribution and Handling: 6](#_heading=h.vsgnek635l53)

[2.2.2.1 Features Distribution and Box plots after data handling 14](#_heading=h.yhjylaqkzbh4)

# List of Figures

Figure 1 Original Dataset sample 3

Figure 2 First 5 records in the dataset 4

Figure 3 Last 5 records in the dataset 4

Figure 4 Checks on Data(columns type,NULLS) 5

Figure 5 Checks on Duplicates and Nulls 5

Figure 6 Data Description Part1 6

Figure 7 Data Description part2 6

Figure 8 Sample of Data after age conversion 6

Figure 9 Data Description after age conversion. Changes are in age 7

Figure 10 Features Distribution part1 7

Figure 11 Features Distribution part2 7

Figure 12 Features Distribution part3 8

Figure 13 Number of records in which gender=3 8

Figure 14 Features Box plots part1 9

Figure 15 Features Box Plots part2 9

Figure 16 Possible Lowest Systolic (ap\_hi) 10

Figure 17 Possible Highest Systolic(ap\_hi) 10

Figure 18 Thresholds for ap\_lo 11

Figure 19 Mean of ap\_hi of subset (ap\_hi < 40mmHg & ap\_hi > 150) 11

Figure 20 Distribution of ap\_lo filtered on (<40 & >150 & cardio=1) 12

Figure 21 Q-Q Plot of ap\_lo after removing outliers 12

Figure 22 Sample of Dataset after cleaning 13

Figure 23 Data Description after cleaning 13

Figure 24 Features Distribution Part1 14

Figure 25 Features Distribution Part2 14

# **Project Overview**

The Healthcare Predictive Analytics specially **Cardiovascular Disease (CVD)** project focuses on developing a classification model to classify the patient if he/she suffers from cardiovascular disease or not by providing data-driven insights. The model will be designed to help healthcare professionals with tasks such as patient risk detection, making informed decisions based on predictive analytics. The project will utilize a classification model.

# **Milestone1**

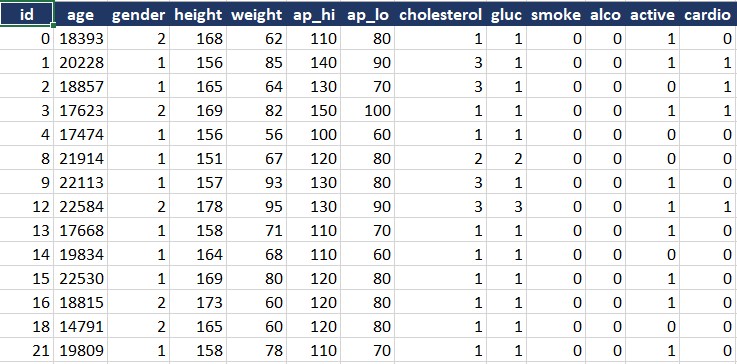
## **Data Collection:**

Our Dataset was collected from Kaggle Link:<https://www.kaggle.com/datasets/scientificstephen/medical-examination-dataset-analysis>

Link of Dataset on our drive: <https://drive.google.com/file/d/1uD5d16AkU_fdwTs3Fq6xqlGCN09A7o_7/view?usp=drive_link>

## **Data Exploration:**

Healthcare cardiovascular disease dataset consists of 70000 records of people and 13 factors taken into consideration in the dataset. The below figure shows a sample of the dataset



*Figure 1 Original Dataset sample*

Description of each column (Key Features):

* id: this is just a number used to identify the patient.
* age: contains the age of each patient in days.
* gender: The column identify the sex of each patient (1: Female,2: Male).
* height: contains the height of each patient in meters (m).
* weight: contains the weight of each patient in kilograms (kg).
* ap\_hi: Systolic of the patient in mmHg.
* ap\_lo: Diastolic of the patient in mmHg.
* cholesterol: categorize the cholesterol level of each patient (1:Low, 2:Medium, 3:High).
* gluc: categorize the glucose level of each patient (1:Low, 2:Medium, 3:High).
* Smoke: Categorize the patient if he/she is a smoker or not.(1: Smoker, 0: Not a smoker)
* alco: Categorize the patient if he/she drinks alcohol or not. (1: drinker, 0: not a drinker)
* active: Categorize if the patient practices any sport or not (1: practice any activity, 0: not practice any activity).
* cardio: this is the target column in which classifies the patient if he/ she suffers from cardiovascular disease or not.

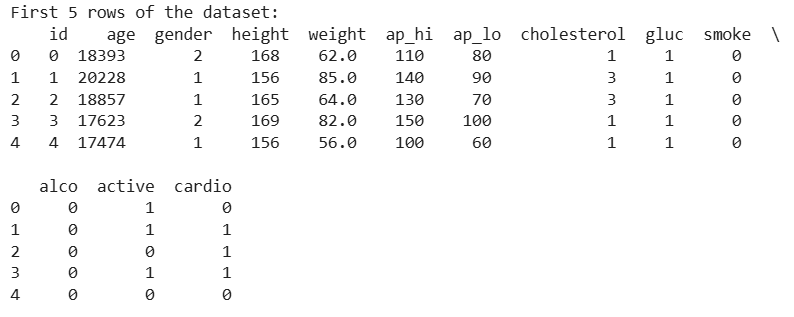
**Data Summary:**

1- Size: (13\*70000) Thousands of individual records.

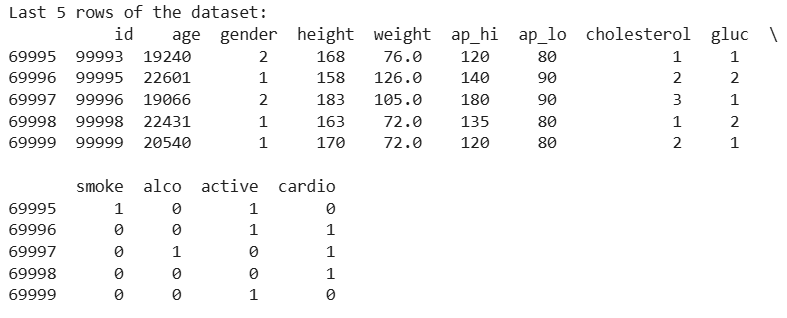
2- Type: Mixed numeric and categorical data.

3- Challenges: Includes outliers and categorical data requiring cleaning and preprocessing.

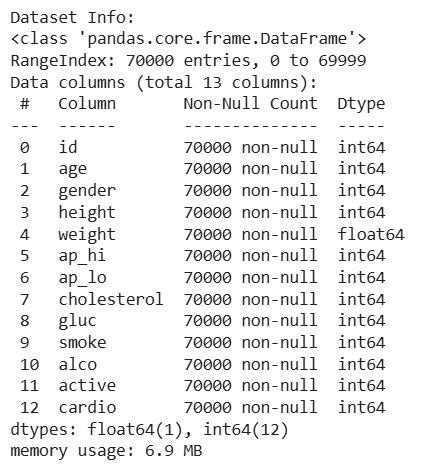
### **Check Data Statistics; First 5 records, Last 5 records, NULLs , Duplicates in Dataset and check if there is any categorical column needs encoding:**



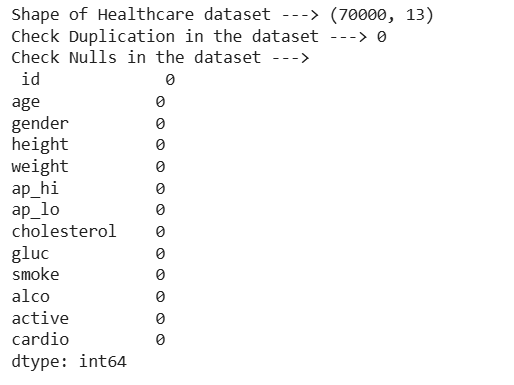
*Figure 2 First 5 records in the dataset*

**

*Figure 3 Last 5 records in the dataset*

**

*Figure 4 Checks on Data(columns type,NULLS)*

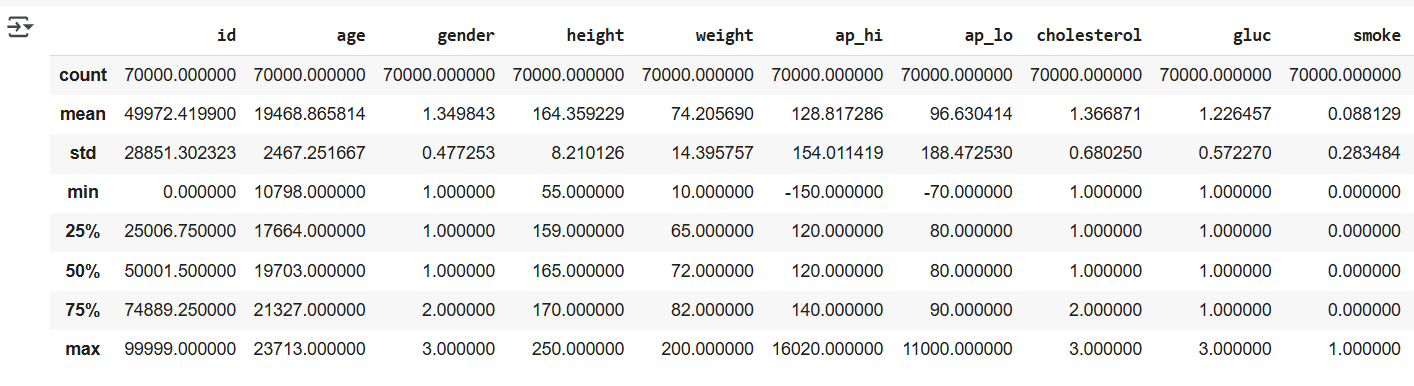
**

*Figure 5 Checks on Duplicates and Nulls*

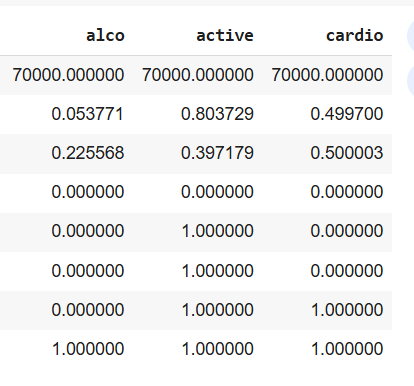
From the above figure it is found that:

* No Duplicates.
* No NULLs.
* No need for encoding.

## **Data Distribution and Handling:**



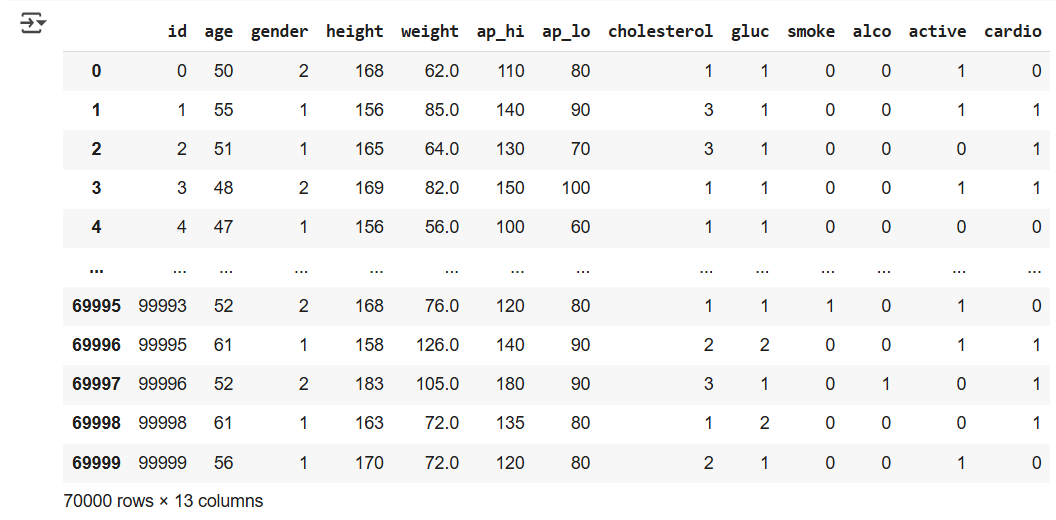
*Figure 6 Data Description Part1*



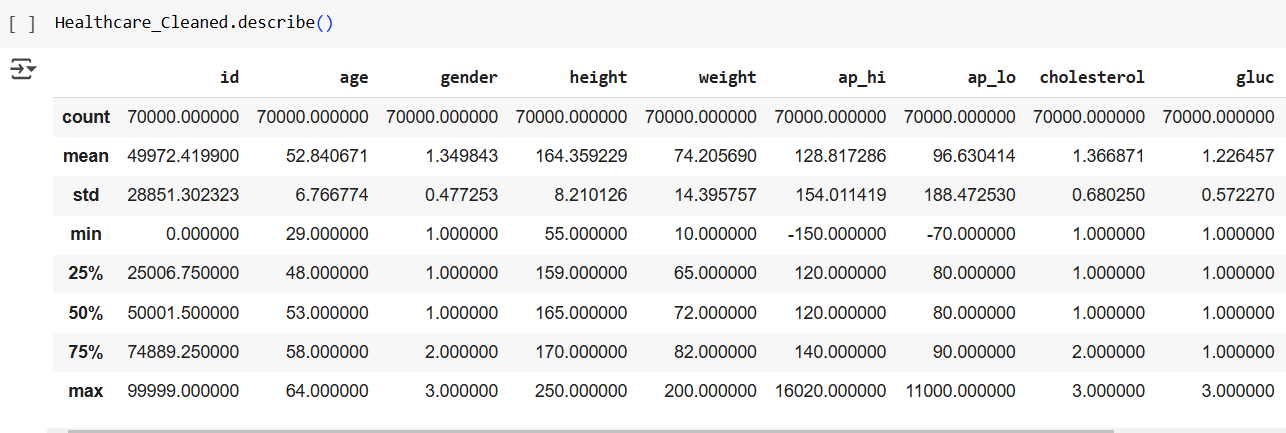
*Figure 7 Data Description part2*

From the above figures, it is found that the age has very large numbers as it is calculated in days.

So we have converted days into years then we have shown a sample of the data after conversion as follow



*Figure 8 Sample of Data after age conversion*



*Figure 9 Data Description after age conversion. Changes are in age*

Then we have drown distribution of each key feature as follow:

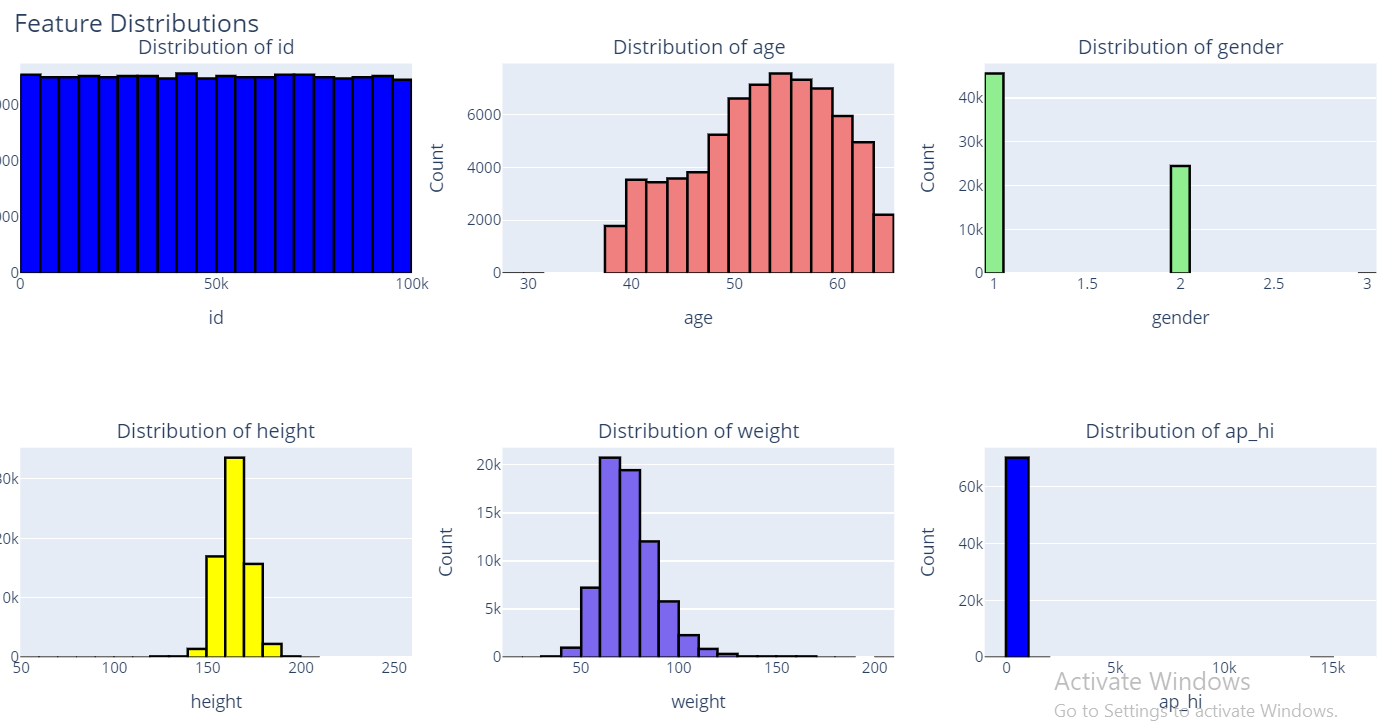


Figure 10 Features Distribution part1

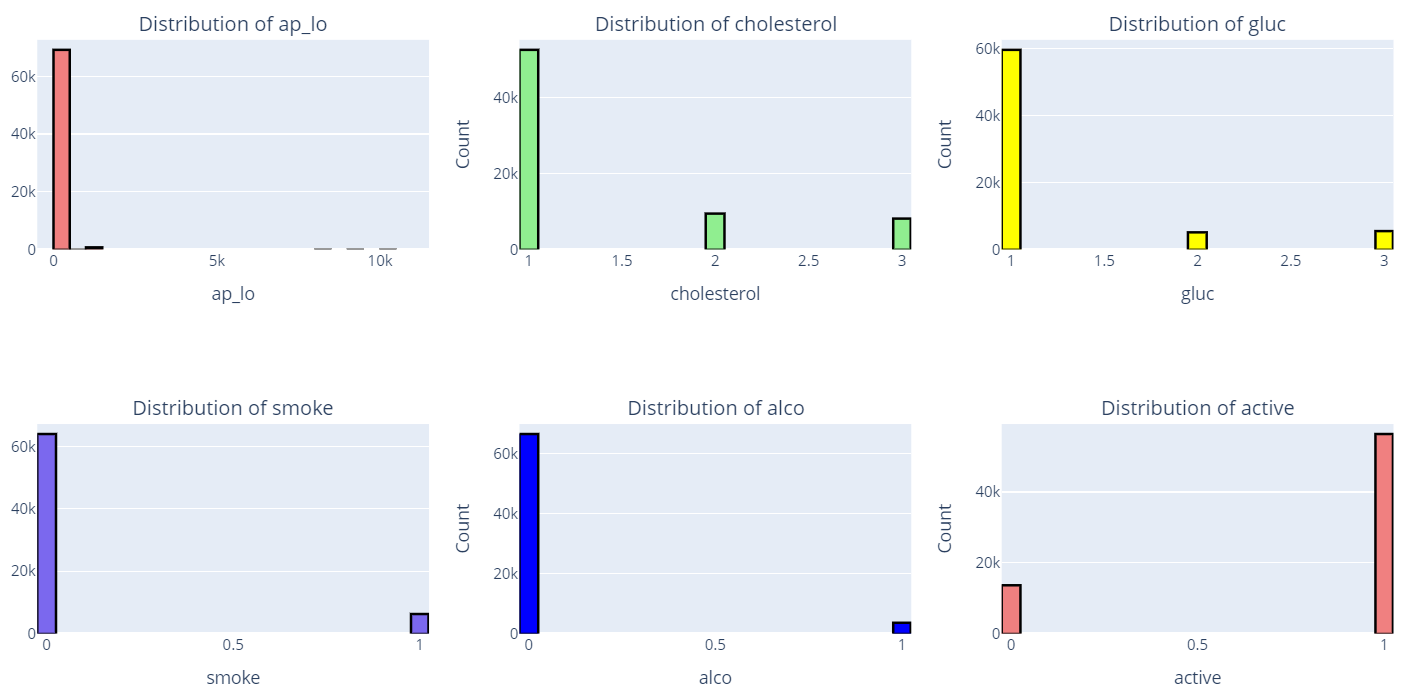


Figure 11 Features Distribution Part2

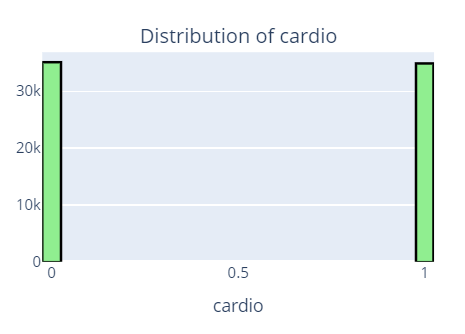


Figure 12 Features Distribution Part3

From the above figure we found that:

* Age distribution is left skewed data so it needs normalization to convert the left skewed to Gaussian distribution.
* Gender contains 3 categories which is not logic as 0 for Males, 1 for Females and 2 for what?! so it depends on the number of samples of this category. the number of samples is 11 records as shown below, it is recommended to eliminate it.

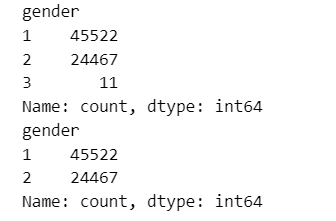
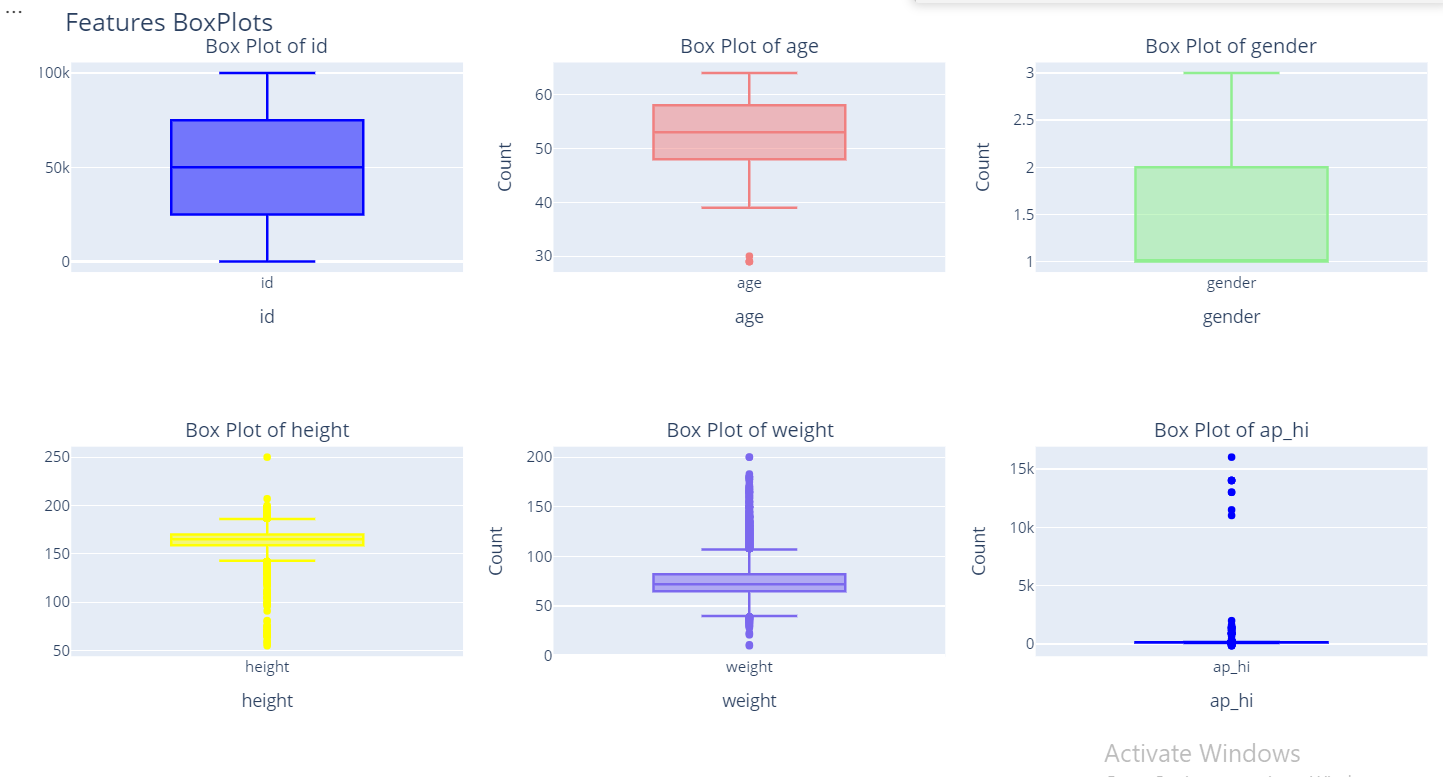


Figure 13 Number of records in which gender=3

* Height and weight are almost Gaussian distribution.
* ap\_hi and ap\_lo seem to have outliers as the maximum values from data description are 16020 and 11000 respectively which are not logic values and number of samples at these values are not large so we can handle them by elimination as follow.
* In features the most dominant samples are the normal samples of people but cardiovascular column is balanced which means that there are outliers in the dataset.



*Figure 14 Features Box plots part1*

**

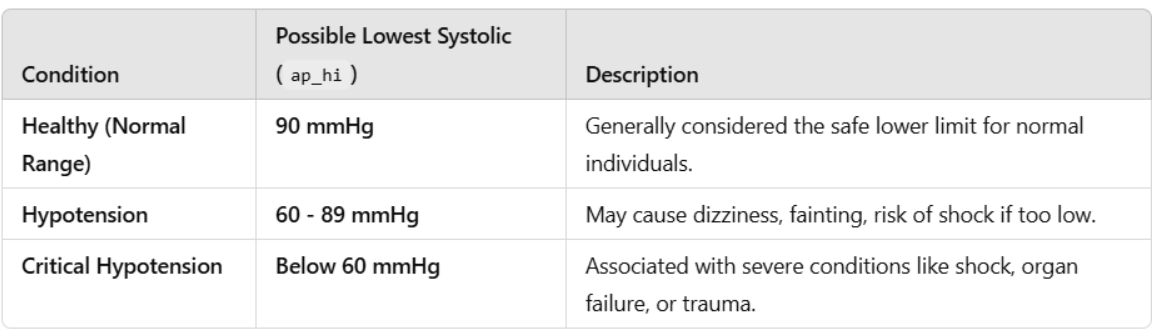
*Figure 15 Features Box Plots part2*

From the above figures we found that:

* The minimum and maximum values of age are 29 and 64 respectively which are normal values so we decided to keep it as it is.
* In height column minimum and maximum values are 55 and 250 cm respectively. But some of these values are not logic and not correlated with age. Thresholds we have chosen are 100 cm for the lower threshold and 200 cm for the upper threshold.

Based on these thresholds outlier samples count was 31 sample so we decided to eliminate them.

* In weight column minimum and maximum values are 10 and 200 kg respectively. Value 10 does not match with the minimum value of age 29 so our thresholds for weight were 45 kg as a lower threshold and 190 kg as an upper threshold. We found that 304 samples were out of this range (outlier) so we decided to eliminate them.
* For ap\_hi and ap\_lo thresholds chosen are dependent on medical domain as follow:



*Figure 16 Possible Lowest Systolic (ap\_hi)*

From the above figure we found that the possible lowest systolic value is 60 mmHg below that the human will not be alive. But if there is a noise in the device used to measure the pressure this may affect the measurement to we have chosen the lower threshold to be 50 mmHg.

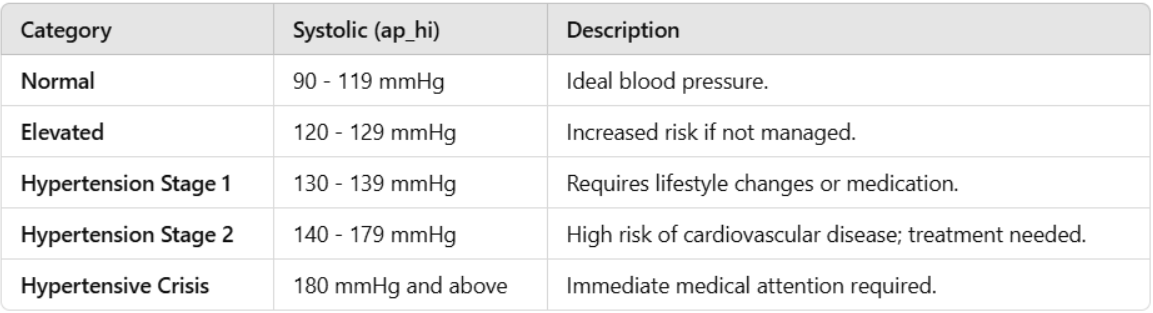


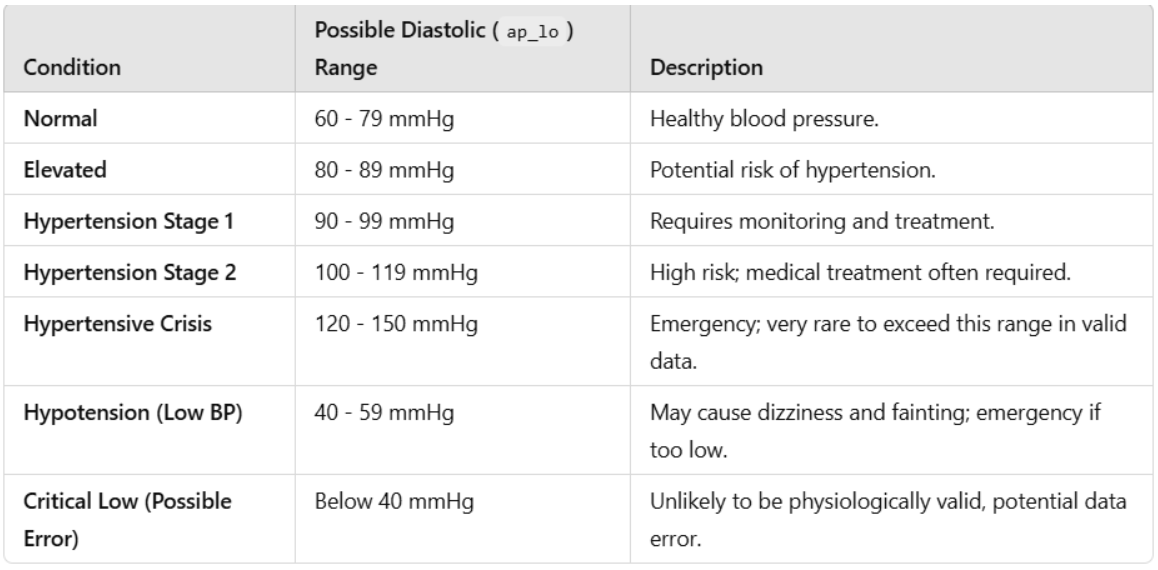
Figure 17 Possible Highest Systolic(ap\_hi)

From the above figure we found that the maximum value for ap\_hi is 180 mmHg but during search we found that the maximum value ap\_hi can be taken and the human is a live is around 250 mmHg.

So thresholds chosen for ap\_hi 🡪 from 50 mmHg to 250 mmHg.

Based on the above thresholds it is found that 224 samples are considered as outliers so we decided to eliminate these samples.

* For ap\_lo from search we found that:



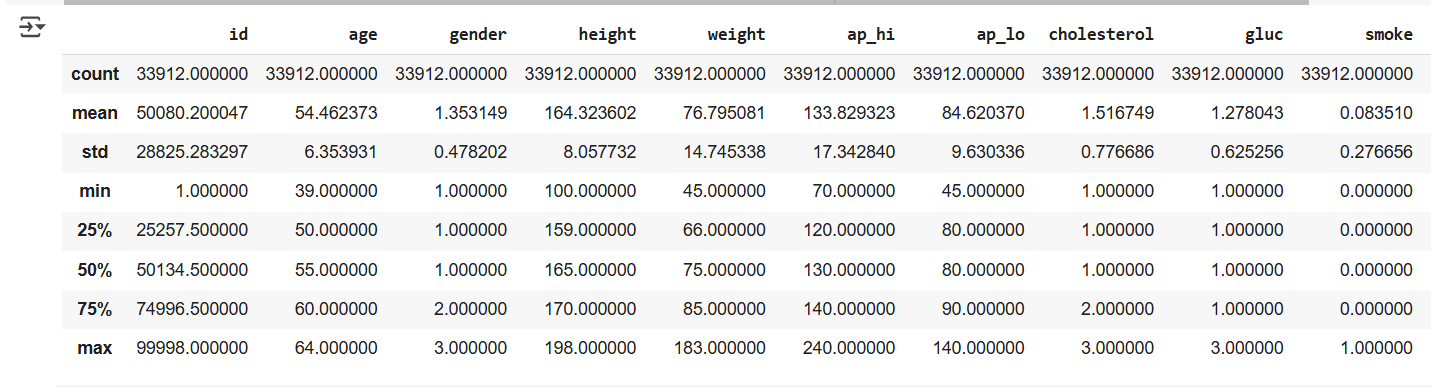
*Figure 18 Thresholds for ap\_lo*

From The above figure we found that the minimum value for ap\_lo is 40 mmHg and the upper threshold 150 mmHg.

Based on the previous thresholds we found that 1013 samples as outlier we decided not to eliminate these samples as they represent 1.4%.

The logic used to impute the outliers is as follow:

* By logic, people whose ap\_lo is lower than 40 mmHg and higher than 150 mmHg definitely suffer from cardiovascular disease which means cardio flag=1. This will give us subset of data. Number of samples in this subset is 837 records. we decide to calculate the mean of ap\_lo of these people and impute the outlier values with this mean.



*Figure 19 Mean of ap\_hi of subset (ap\_hi < 40mmHg & ap\_hi > 150)*

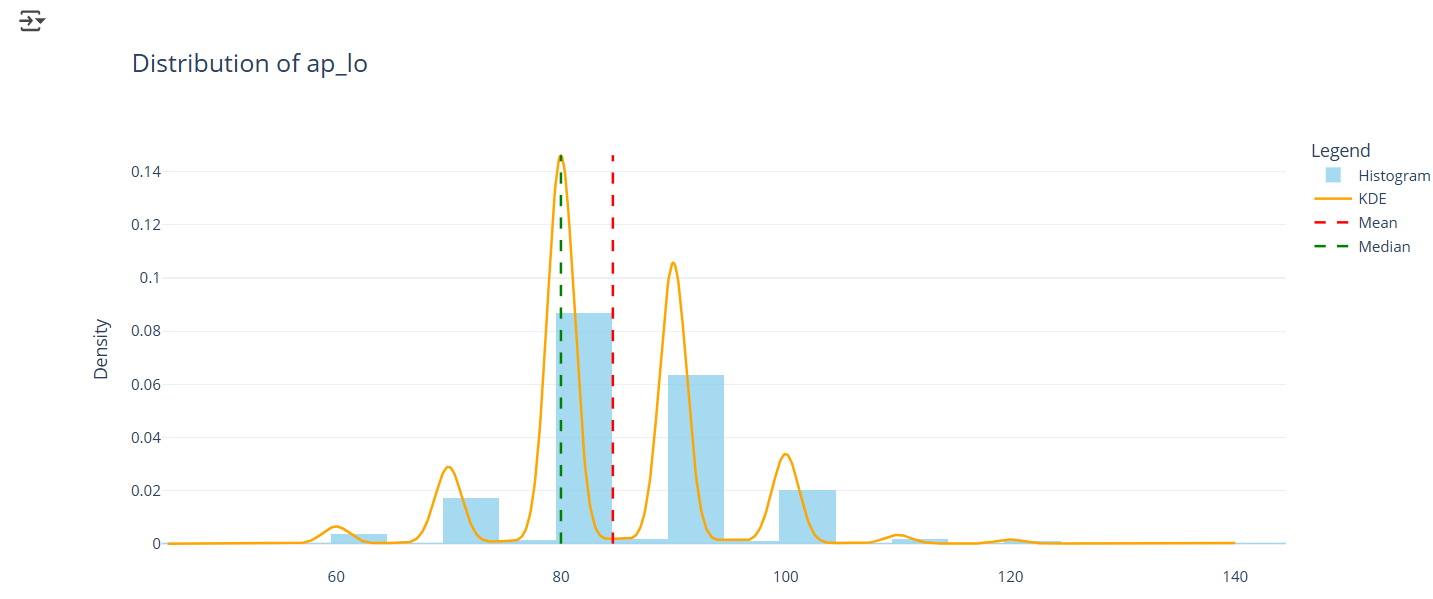
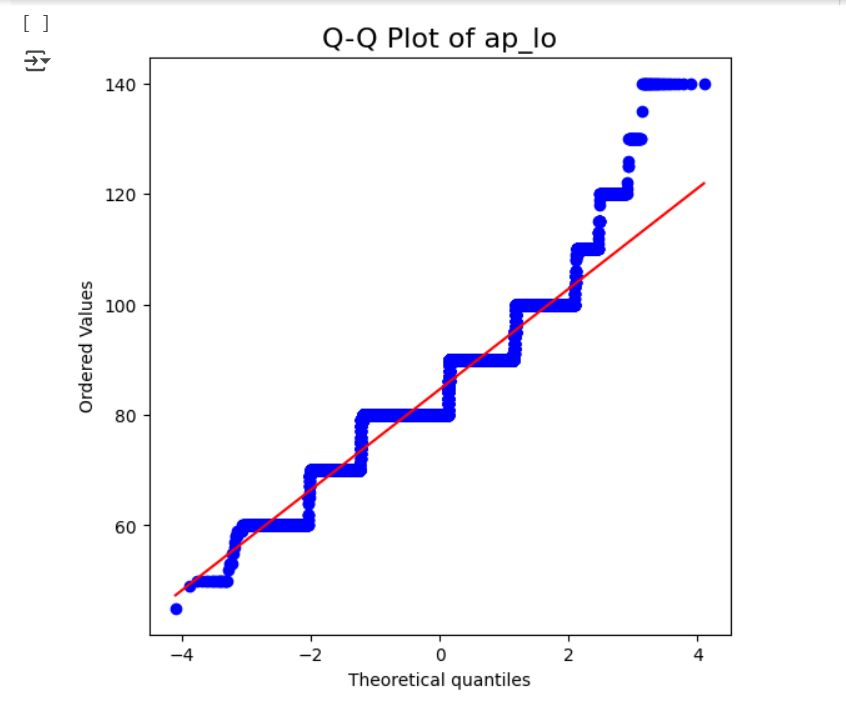


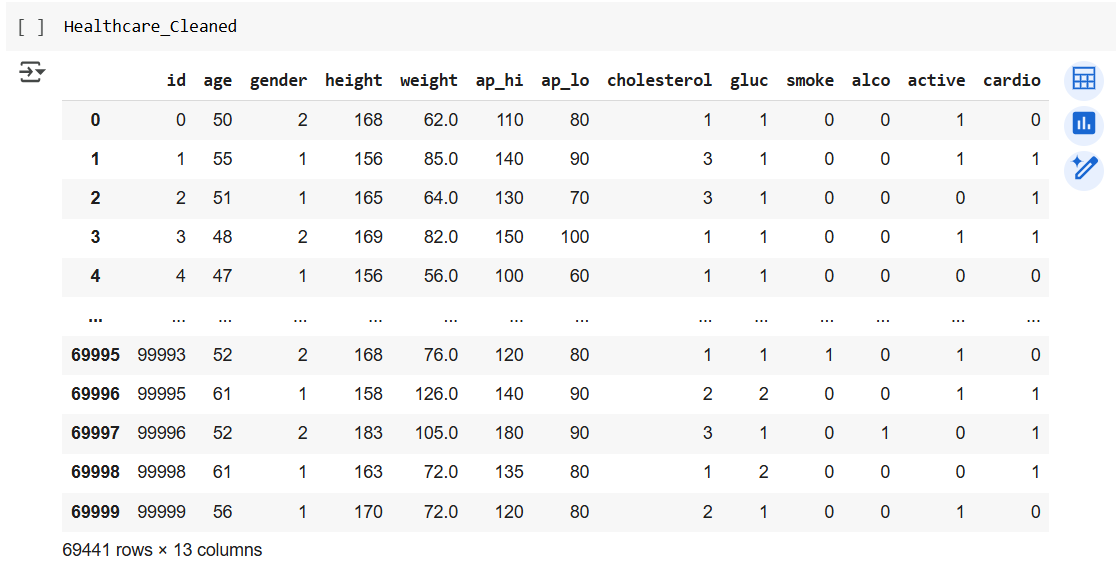
Figure 20 Distribution of ap\_lo filtered on (<40 & >150 & cardio=1)

From the above figure we found that mean=84 and median=80 so we substituted by mean.

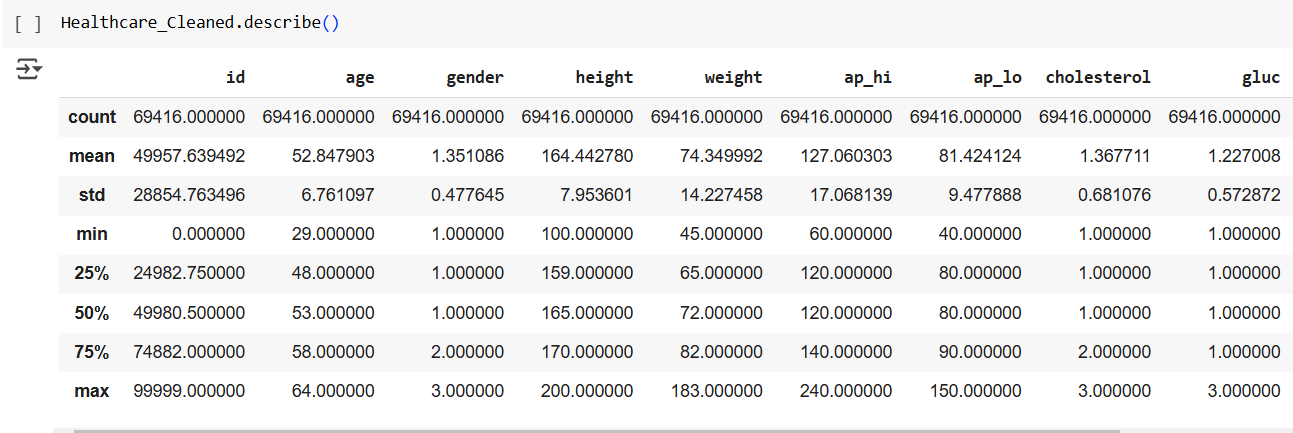


*Figure 21 Q-Q Plot of ap\_lo after removing outliers*

This is a Q-Q plot of ap\_lo after substitution which means that the data distribution almost became normal distribution.



*Figure 22 Sample of Dataset after cleaning*

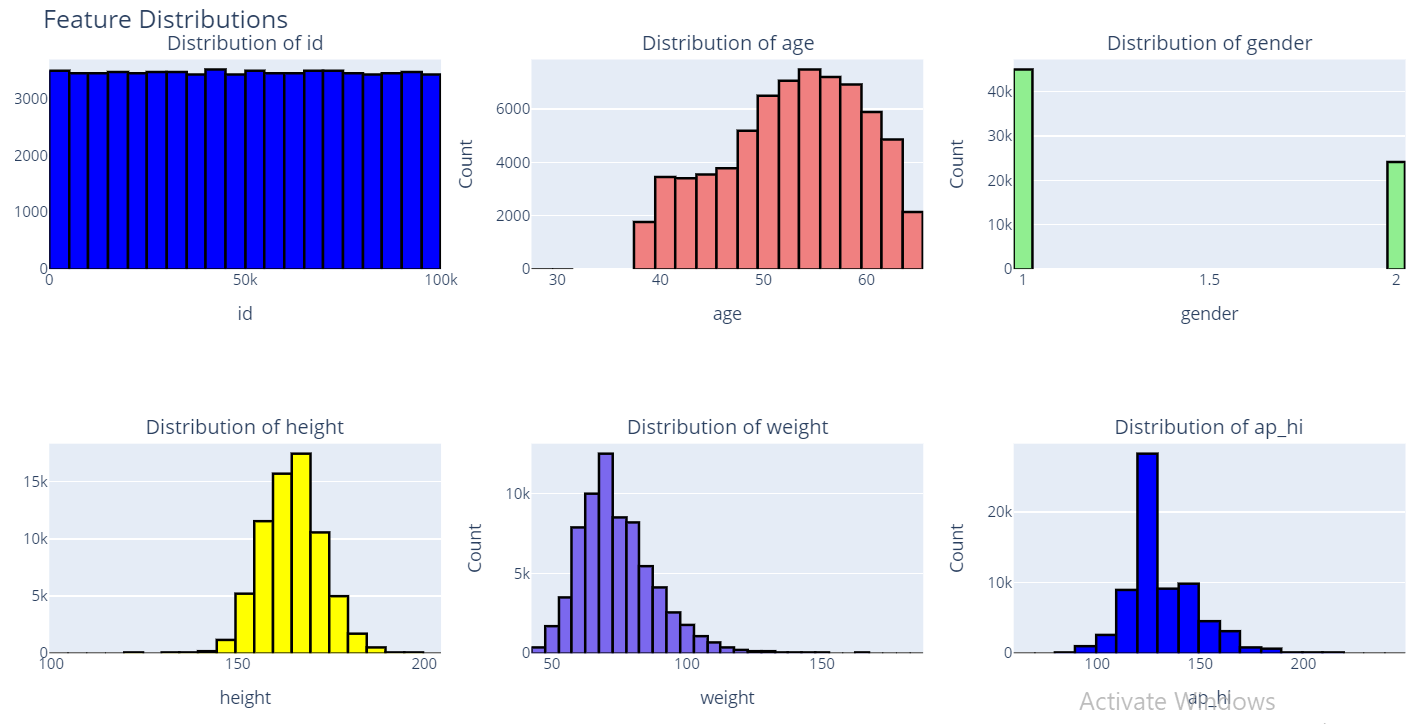


*Figure 23 Data Description after cleaning*

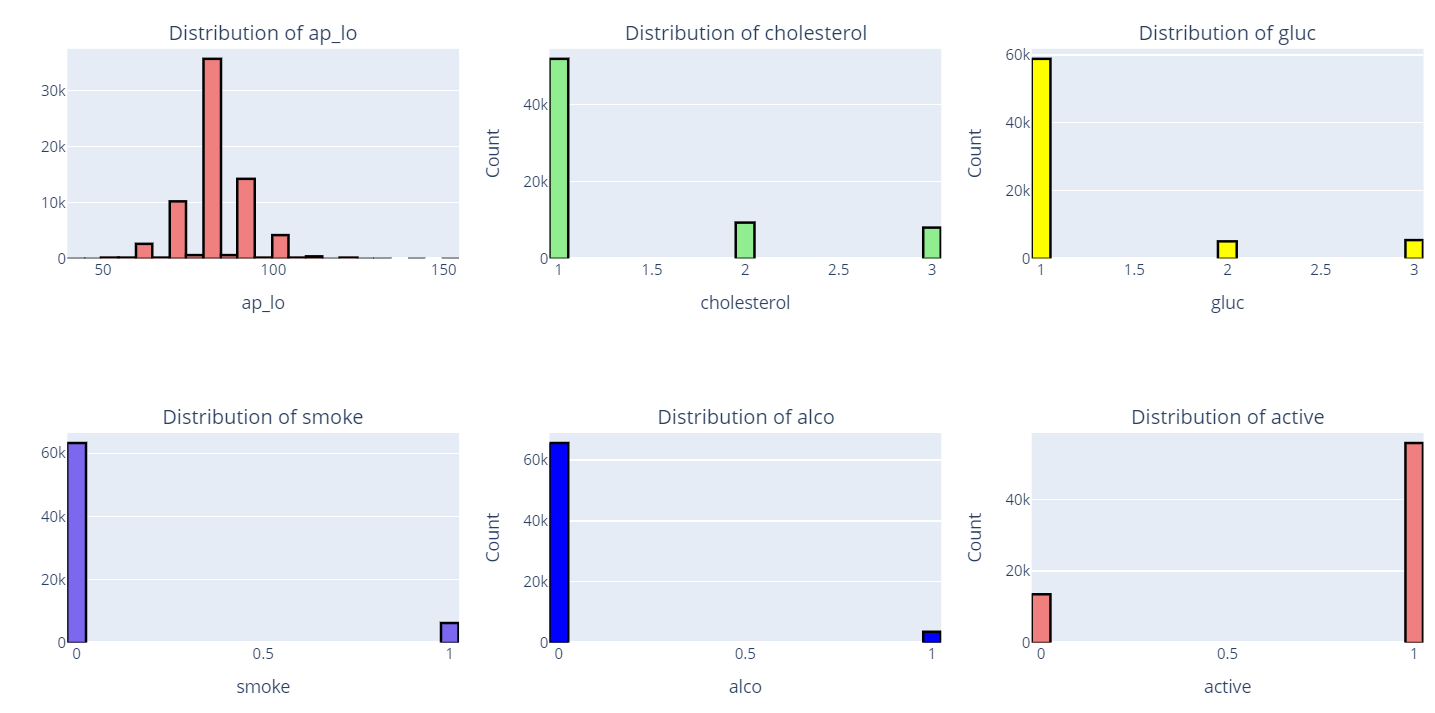
*There is a bug found during data investigation:*

*It is found that there are some records in which ap\_lo > ap\_hi which does not make sense.*

### **2.2.2.1 Features Distribution and Box plots after data handling**

**

*Figure 24 Features Distribution Part1*

**

*Figure 25 Features Distribution Part2*

*From the above figures it is found that:*

* *Gender feature now contains only two categories.*
* *Distribution of ap\_hi & ap\_lo are enhanced*