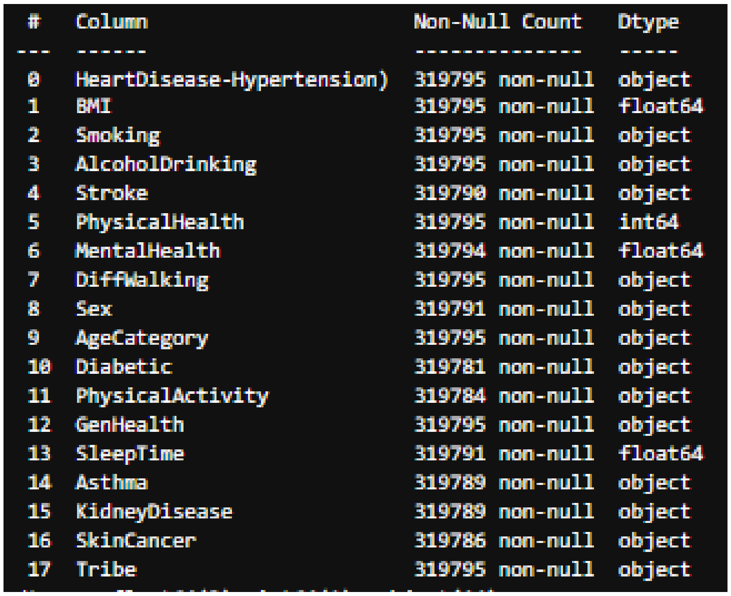
Data Preprocessing Methodology for Hypertensive Heart Disease Dataset

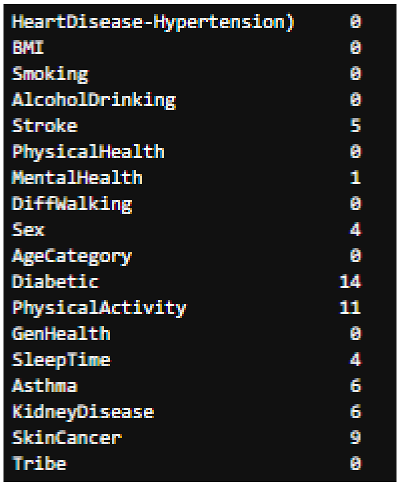
Introduction

The dataset presented with us has over 310,000 rows of records of different persons surveyed on different possible factors of hypertension. Containing 18 different features. On checking, it was realized that some features in the datasets contains missing values and needed to be preprocessed in order to aid our analysis and avoid bias and error. The tables below show the features in the dataset and the features with the missing values.

The following highlighted steps are the steps taken to clean the dataset according to the Stake holder after consultation.



```
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   HeartDisease-Hypertension)  319795 non-null  object
 1   BMI                         319795 non-null  float64
 2   Smoking                     319795 non-null  object
 3   AlcoholDrinking             319795 non-null  object
 4   Stroke                      319790 non-null  object
 5   PhysicalHealth              319795 non-null  int64
 6   MentalHealth                319794 non-null  float64
 7   DiffWalking                 319795 non-null  object
 8   Sex                         319791 non-null  object
 9   AgeCategory                 319795 non-null  object
 10  Diabetic                    319781 non-null  object
 11  PhysicalActivity            319784 non-null  object
 12  GenHealth                   319795 non-null  object
 13  SleepTime                   319791 non-null  float64
 14  Asthma                      319789 non-null  object
 15  KidneyDisease               319789 non-null  object
 16  SkinCancer                  319786 non-null  object
 17  Tribe                       319795 non-null  object
```

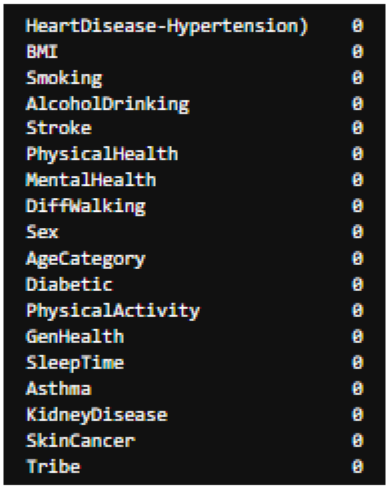


```
HeartDisease-Hypertension)     0
BMI                            0
Smoking                        0
AlcoholDrinking                0
Stroke                         5
PhysicalHealth                 0
MentalHealth                   1
DiffWalking                    0
Sex                            4
AgeCategory                    0
Diabetic                      14
PhysicalActivity              11
GenHealth                      0
SleepTime                      4
Asthma                         6
KidneyDisease                  6
SkinCancer                     9
Tribe                          0
```

Data Cleaning Process

Step 1. Filling the features with missing values with the instructed values.

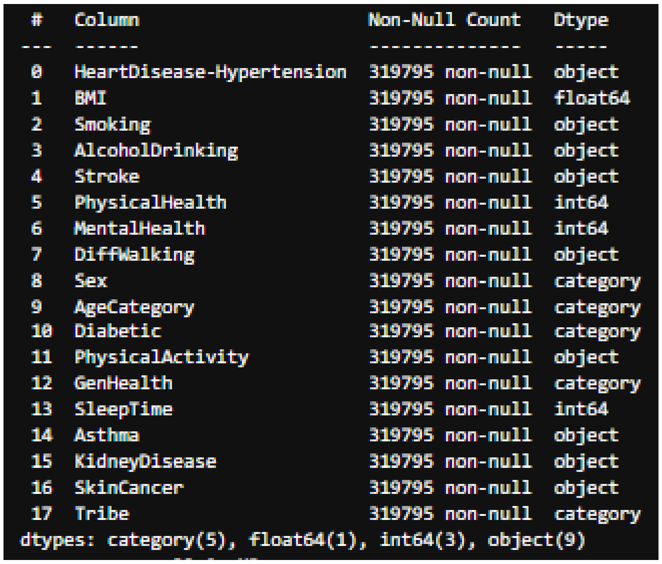
A new dataset of same numbers of rows and features was given on request to impute the missing values of the first dataset we were provided with. The integrity of this new dataset was confirmed and was used to impute the missing values after which the original dataset was confirmed to the imputed already. After which the condition was confirmed as below.



```
HeartDisease-Hypertension)    0
BMI                           0
Smoking                       0
AlcoholDrinking               0
Stroke                        0
PhysicalHealth                0
MentalHealth                  0
DiffWalking                   0
Sex                           0
AgeCategory                   0
Diabetic                      0
PhysicalActivity              0
GenHealth                     0
SleepTime                     0
Asthma                        0
KidneyDisease                 0
SkinCancer                    0
Tribe                         0
```

After this was done we moved to the next step.

Step 2. Data Validation: Here we ensured that all features are in the right data types. This is done by using the method astype. And all features were put in the right types such as int, objects. Below is the outcome of this step.



```
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   HeartDisease-Hypertension  319795 non-null  object
 1   BMI                        319795 non-null  float64
 2   Smoking                    319795 non-null  object
 3   AlcoholDrinking            319795 non-null  object
 4   Stroke                     319795 non-null  object
 5   PhysicalHealth             319795 non-null  int64
 6   MentalHealth               319795 non-null  int64
 7   DiffWalking                319795 non-null  object
 8   Sex                        319795 non-null  category
 9   AgeCategory                319795 non-null  category
 10  Diabetic                   319795 non-null  category
 11  PhysicalActivity           319795 non-null  object
 12  GenHealth                  319795 non-null  category
 13  SleepTime                  319795 non-null  int64
 14  Asthma                     319795 non-null  object
 15  KidneyDisease              319795 non-null  object
 16  SkinCancer                 319795 non-null  object
 17  Tribe                      319795 non-null  category
dtypes: category(5), float64(1), int64(3), object(9)
```

Step 3: The next step that was taken to prepare our data for analysis was Feature Engineering. Here, based on research on the features we have in the dataset, we were able to derive a feature that is useful to our analysis and granted more insights.

Based on the Medical Standards, the following are the category

Underweight --- Less than 18.5

Healthy weight --- 18.5 - 24.9

Overweight --- 25.0 - 29.9 -- Not Obese

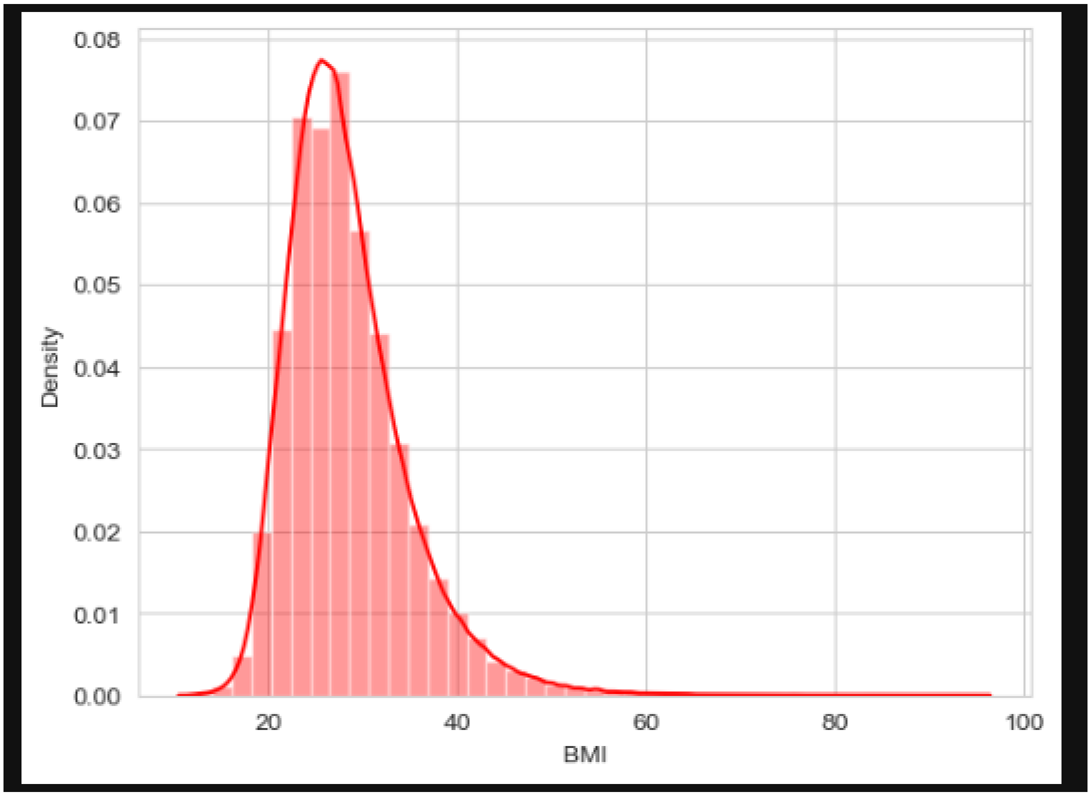
Low Risk --30.0 - 34.9 -- Class 1 Obesity

Moderate Risk --- 35.0 - 39.9 Class 2 Obesity

High Risk --- equal or greater than 40.0 Class 3 Obesity

The only additional feature we created is **BMI Status.** Which helps us to categorize each person into a clearer category that tells more about their BMI status and implications.

Step 4: The last step that was taken was to check for outliers in our features in order to determine the right measure of centrality for aggregation.



The present of outliers was confirmed by plotting a density plot which shows extreme values but these values were not dropped based on the rationale that the present of outliers is normal as there is not standard range for the values in the features.

The was all of our Data Preprocessing steps.