Problem Statement: Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner.

Description of Data Attributes:

Variable	**Description**
Age	Age in years
Sex	1 = male; 0 = female
cp|	Chest pain type
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Target	1 or 0

Then, we performed the following steps:

1. **Preliminary Analysis**
   a. Null values- None, no missing values in the dataset
   b. Duplicate samples- One duplicate sample , this was dropped from the data set
   c. Attributes classified under each variable type based on their unique variables.
      - Numerical Variables
        - thalach - The person's maximum heart rate achieved
        - chol- The person's cholesterol measurement in mg/dl
        - oldpeak - ST depression induced by exercise relative to rest
        - ca- number of major vessels (0-3) colored by flourosopy
        - Age
        - trestbps (The person's resting blood pressure)

      - Categorical variables
        - Sex
        - Cp
        - fbs
        - restecg
        - exang
        - Slope
        - Thal

   d. Distribution of Disease and Data/ Relationship and other Factors
      - Create a heatmap plot (correlation matrix) to have an overview of the mean and median difference, and generated histplots and boxplots through matplotlib library to show the spread of the data
        - The following variables showed outliers.
          - Positively skewed – outliers belonging in the upper Quartile range
            - trestbps
            - chol
            - fbs
            - oldpeak
            - ca
          - Negatively skewed- outliers belonging in the lower Quartile range
            - Thalach
            - Thal
      - Based on the p-values of the variables from the initial dataset, the following are insignificant in relation to the target variable ‘ target’
        - 'age', 'trestbps', 'chol', 'fbs', 'restecg', 'slope'
      - Based on the correlation matrix, there shows no significant linear relation between all the variables.
      - Occurrence of heart attacks (Cardio Vascular Disease) based on Age Groups:
        - Highest number of samples were taken from age group ***51-61--Fifties_Early Sixties***, followed by ***38-50-- Early_ Late Forties***
        - The highest number of hearts across all age groups were found from the age group:***51-61--Fifties_Early Sixties*** but from the total

number of samples of this age group, the patients detected with No- heart attack is slightly greater than the number of patients with heart attacks
    - The least number of occurrence of heart attacks were found from age group: ***29-37-- Early_Late Thirties***
    - From the total number of samples from the age group: ***38-50-- Early_ Late Forties***, there is a significant gap between patients with No- heart attack and with heart attacks wherein patients with heart attacks is around 50% higher than no heart attacks
- Composition of patients with respect to Sex Category
    - There were more male patients studied than female
    - From the total number of female samples, there is a significant gap between patients with No Heart Attacks and with, around 75% had heart attacks
    - Although there is an unbalanced sample size between Male and Female groups, patients detected with heart attacks and without are normally distributed across both genders
- Detect heart atacks based on anomalies in the resting trestbps blood pressure () of a patient
    - Normal levels of trestbps is from 130-140 for adults.
    - Outliers were found outside Q3, or upper limit outliers, There are 13 sample (patients) who have extreme trestbps ranging from 170-200 mm/Hg
    - Only 3 out 13 samples with extreme values for trestbps were diagnosed with cvd or heat attack, whilst 9 out of 13 patients with extreme values were diagnosed with no heart attack
    - No Heart Attacks from the 9 patients with extreme values could be the result of errors in measurements or factors
    - Based on the p-value score of trestbps, this is insignificant with respect to existence or non of heart attacks(dependent variable), depending on the performance of the model, the outliers may also be insignificant
- Relationship between the following ***Numerical Variables and Target:***
    - Cholesterol levels- Although in general cholesterol levels should be under 200 mg/dl for adults, there is no correlation between chol and hear attack(target). This is supported by positive skewness of the distribution and the p-value
    - Oldpeak- is positively skewed as well, data shows that a high number of patients with “0” ST depression induced by exercise had heart attacks
    - Trestbps- positively skewed with few outliers, even with normal bps of 130-140, the data shows that cvd occurs in cases of bps lower than the normal level. P-value is high than alpha, supporting that trestbps is not significant with target
    - Thalac- patients with maximum heart rate of 140, are likely to have heart attacks based on the data. Thalac is also a significant variable with respect to heart attacks.

    - Ca- Though significant in terms of p-value, showed a positively skewed relationship with the target variable, heart attack The relationship shows that the lower the number of vessels, the higher the chances of heart attack.

  - Relationship between the following categorical variables and target variable:
    - Cp- significant variable. While chest pain may be caused by other factors, the occurrence of chest pains from the patients, led to more cases of heart attacks.
    - Fbs- non-significant variable, it does not show that patients with >120 mg/ml level of fbs had heart attacks.
    - Restecg- non-significant variable, most patients with left ventricular hypertrophy, the condition which is high risk for heart problems were not diagnosed with heart attacks, whilst with normal levels did show high number of heart attacks
    - Exang-significant variable, however most of those with no exercised induced angina has heart attacks as opposed to those who have "exang' had no heart attacks.
    - Slope- non significant variable, however, patients with downsloping(signs of unhealthy heart), were mostly diagnosed with heart attack
    - Thal (thalium stress result)- significant variable, People with thal value equal to 2 (fixed defect) are more likely to have heart disease. This maybe the reason why thal is negatively skewed, However, higher risk of heart for fixed defects maybe the case if it is large or multiple, and involves the left anterior descending artery. This is an indicator of a heart attack as this indicate damaged heart muscle or a previous heart attack incidences.

2. **<u>Build a baseline model</u>** to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection:
    - Prior to creating the models, a second DF was created to have dummies to address outliers to check if the model will perform better
    - In order to avoid going by assumptions, scaling was performed
    - Logistic and Random Forest Models for both data sets generated resulted to;
        - Accuracy score of 1 for all models for all data frames (original, with dummies and with dropped insignificant variables)
        - Prediction for False Negatives and False Positives are all at 0 for all models done for all dataframes
        - Precision and Recall Scores for both Yes and No are 1.0 (100% accuracy)
        - No class imbalance evident– Support values for both 0 and 1 are not too far from each other
0 = 18
1=13
        - The baseline LR and RF models for all dataframes are perfectly predicting heart attacks