Exploratory Data Analysis of Heart-Attack Prediction dataset

What is EDA:

EDA is the process of identifying patterns, observing trends, and formulating hypothesis.

Dataset : https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Target column:

output – whether Heart-Attack is possible or not ~ 0- NO , 1= YES

Predictors/Feature Columns:

- age - Age of the patient
- sex - Sex of the patient
- cp - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
- trtbps - Resting blood pressure (in mm Hg)
- chol - Cholestoral in mg/dl fetched via BMI sensor
- fbs - (fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
- restecg - Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
- thalachh - Maximum heart rate achieved
- oldpeak - Previous peak
- slp - Slope
- caa - Number of major vessels
- thall - Thalium Stress Test result ~ (0,3)
- exng - Exercise induced angina ~ 1 = Yes, 0 = No

To Do in EDA:

1. Clean the data, Transform categorical variables to int handle missing values and outliers
2. Find how features are correlated with target
3. Identify which features that are significant in predicting the target variable.(Feature Engineering)
4. Use P-value test to test the hypothesis

What is Done Here: Attaching Python notebook.

1. **Clean the data, Transform categorical variables to int handle missing values and outliers**
    a. Explored the distribution of the population in each column using histogram/box plot to find the outliers
    b. Transformed categorical column for the sake of analysis.
    c. Used log transformation to eliminate skewness in some columns
    d. Checked if there is null values in the dataset. No missing values

2. **Find how features are correlated with target**
    a. Visualised correlation matrix using heatmaps (triangle heatmap, exclusively with the output column)
    b. Effect of different features on other features and target variable

3. **Identify which features that are significant in predicting the target variable. (Feature Engineering)**
    a. we compare the correlation between features and remove one of two features that have a correlation higher than 0.4.

4. **P-Value Test**
    1. We assume to null hypothesis to be “The selected combination of dependent variables do not have any effect on the independent variable”.
    2. Then we build a small regression model and calculate the p values.
    3. If the p values is higher than the threshold, we discard that combination of features

    1. **Hypothesis testing of Serum cholesterol levels and target**
        a. I will perform a hypothesis test to check the difference between the population mean of serum cholesterol levels of the person who have heart disease and the person who doesn't have heart disease at alpha=0.05.
        Null Hypothesis :- H0: u0 = u1
        Alternative Hypothesis :- H0: u0 != u1

    the p-value is 0.186 which is greater than 0.05 so i retain null hypothesis that there is no significant difference between the population mean of serum cholesterol levels of the person who has heart disease and the person who doesn't have heart disease.

    2. **Hypothesis testing to find average age when a person got Heart disease**
        a. Two Groups are there Group A(0) consist of the people who have heart disease and Group B(1) consist of people who don't have heart disease. The sample mean and the standard deviation is known in this case. So, at alpha=0.05 (95 % confident) test whether the average age of the person having a heart disease is more than the person who doesn't have heart disease.

        Null Hypothesis -: H0 : uA - uB <= 0
        Alternative Hypothesis: - HA : uA - uB > 0

I achieved a p-value less than 0.05 so we reject null hypothesis and our alternative hypothesis is true that is the average age of the person having a heart disease is more than the person who doesn't have heart disease.

3. It's believed by a doctor that among the patients who have heart disease 30% of them have asymptomatic chest pain, 25% have an atypical angina chest pain, 40% have a non-anginal pain and 5% have typical angina. From the given sample of 136 patients who have heart **disease test whether the belief of the doctor is true at alpha=0.05**.
    a. **Null Hypothesis** -: Probability distribution of chest pain is P(asymptomatic)=0.30, P(atypical angina)=0.25, P(non-anginal pain)=0.40, P(typical angina) = 0.05
    b. **Alternative Hypothesis**: - Probability distribution of chest pain is not defined in null hypothesis

The p-value is less than 0.05 so we reject null hypothesis(original claim) it means that the doctor's belief is False.

Result of analysis:

1. middle aged (45 to 60 years) persons have higher chance of heart attack
2. trtbps (Resting Blood Pressure) has week or slightly negative relation with heart attack
3. thalach (Maximum Heart Rate Achieved) has positive relation with heart attack
4. oldpeak (Previous peak) has negative relation with heart attack
5. Women are more likely to have heart problem than men (based on ratio)
6. People with High Cholestoral (more than 200) have very higher chance of heart attack
7. People with Maximum Heart Rate Achieved > 150 has higher chance of heart attack
8. Age has negative relation with thalachh (Maximum Heart Rate Achieved)
9. Chest Pain has higher chance of heart attack
10. typical angina has lower chance of heart attack than other chest pains
11. People with lower major vessels (caa) have much higher chance of heart attack
12. Age has positive relation with n major vessels (caa) (older people are more likely to have vessels)
13. People with Thall == 2 have much higher chance of heart attack
14. Women are likely to have higher levels of cholesterol compared with men
15. Cholestoral has positive relation with Age
16. Slope has positive relation with heart attack and People with Slope == 2 have much higher chance of heart attack
17. trtbps (Resting Blood Pressure) has positive relation with Age
18. oldpeak (ST depression induced by exercise relative to rest) has highly negative relation with heart attack