Research proposal:

Analysis of Heart Disease Dataset

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. Heart disease cost the United States about $239.9 billion each year from 2018 to 2019. This includes the cost of healthcare services, medicines, and lost productivity due to death (from *CDC, Centers for Disease Control and Prevention, https://www.cdc.gov/heartdisease/facts.htm*).

By analyzing factors that may be associated with heart disease, and how they interact with each other, we can possibly help people/doctors, healthcare providers, government, etc. prevent and/or decrease the number of positive diagnoses of heart disease.

The dataset that I will be working with presents 14 columns, and was downloaded from Kaggle. These are described by different factors associated with heart disease as following:

1) **rest_bp:** resting blood pressure (mm Hg) - numeric
2) **chest_pain**: chest pain type - categorical
    a) 0: Typical angina
    b) 1: Atypical angina
    c) 2: Non-anginal pain
    d) 3: Asymptomatic
3) **Thal:** Thallium stress testing - categorical
    a) 0: Normal
    b) 1: Fixed defect
    c) 2: Reversible defect
4) **age:** age of the patient (years) - numeric
5) **fasting_bs:** fasting blood sugar > 120 mg/dl - categorical
    a) 0: fasting blood sugar <= 120 mg/dl
    b) 1: fasting blood sugar > 120 mg/dl
6) **max_hr:** maximum heart rate achieved (beats per minute) - numeric
7) **exercise_angina:** exercise-induced angina - categorical
    a) 0: no exercise-induced angina
    b) 1: exercise-induced angina
8) **gender:** gender of patient - categorical
    a) 0: female
    b) 1: male
9) **st_slope:** ST segment slope - categorical
    a) 0: upsloping
    b) 1: flat
    c) 2: downsloping
10) **cholesterol:** serum cholesterol level (mg/dl) - numeric
11) **st_depression:** ST depression induced by exercise relative to rest - numeric
12) **rest_ecg:** resting electrocardiographic results - categorical
    a) 0: normal
    b) 1: abnormality in ST-T wave
    c) 2: probable or definite left ventricular hypertrophy

13) **num_vessels:** number of major vessels colored by fluoroscopy - categorical
    a) 0: none
    b) 1: one
    c) 2: two
    d) 3: three

14) **diagnosis:** diagnosis of heart disease - categorical
    a) 0: no heart disease (negative diagnosis)
    b) 1: presence of heart disease (positive diagnosis)

For this research I will be focusing on three of the factors that are associated with heart disease: *max_hr* (numerical)*, age* (numerical)*,* and *cholesterol* (numerical).

During this analysis I will be answering the following questions:

A) Is there any correlation between these factors?
B) From these four factors, which one is the one that presents higher correlation with heart disease diagnosis?
C) Do patients with heart disease have higher cholesterol than patients without heart disease?
D) Do patients with heart disease are older than patients without heart disease?
E) Do patients with heart disease have higher max. heart rate than patients without heart disease?

My hypotheses are:

1) There is a strong correlation between max_hr and heart disease diagnosis.
2) There is a correlation between age and heart disease diagnosis.
3) H0: There is no significant difference between positive and negative heart disease diagnosis for these factors.
    Ha: There is a significant difference between positive and negative heart disease diagnosis for these factors.

These hypotheses will be tested with the following methods:

- ***Correlation function***. A correlation coefficient of zero indicates that no linear relationship exists between two variables, and a correlation coefficient of −1 or +1 indicates a perfect linear relationship. The correlation coefficient can fall within (-1,1) interval.
- **Statistical tests**: since the number of samples in this dataset is relatively small, and half of the variables are categorical. If the data sets follow a normal distribution, I could perform parametric tests. If they don’t follow a normal distribution, I will have to run non-parametric tests.