**Module 3 R Practice**

Dia Khosla

College of Professional Studies, Northeastern University  
ALY 6010: Probability Theory and Introductory Statistics

Instructor - Dr. Selcuk Baran

June 13, 2024

**Abstract: Heart Attack Dataset**

This project analyzes a dataset aimed at predicting heart attacks based on 303 individuals and 14 key features. Statistical methods such as one-sample t-tests for mean and hypothesis testing for p-value were utilized to identify significant predictors of heart disease. Key findings underscore age, maximum heart rate achieved, ST depression, and the number of major vessels as critical indicators.

**Introduction**

This report presents an analysis of the Heart Attack dataset, focusing on identifying key predictors and insights related to heart disease risk among 303 individuals. The dataset encompasses 14 crucial variables, including demographic factors such as age and sex, alongside medical indicators like cholesterol levels and exercise-induced angina. By applying statistical methods like hypothesis testing and logistic regression, this analysis aims to uncover significant relationships between these variables and the likelihood of heart disease. The findings contribute to a better understanding of cardiovascular health and offer implications for preventive strategies and clinical decision-making.

This comprehensive dataset provides a robust foundation for analyzing cardiovascular risk factors and developing predictive models to enhance heart disease diagnosis and prevention.

This study analyzed the Heart\_Attack dataset to investigate age and cholesterol levels as risk factors for heart attacks. A one-sample t-test showed a significant difference in average age from 50 years (t = 8.49, df = 301, p < 0.001). A proportion test revealed a significant deviation from 50% for individuals with high cholesterol (> 200 mg/dL) (X-squared = 153.06, df = 1, p < 2.2e-16). These findings highlight age and cholesterol as crucial factors in cardiovascular risk assessment.

**Interpretation of Statistical Analyses**

One-Sample T-Test: To determine if the mean age of patients in the dataset is significantly different from 50 years.

**Hypotheses:**

* Null Hypothesis (\(H\_0\)): The mean age of patients is 50 years.
* Alternative Hypothesis (\(H\_1\)): The mean age of patients is not 50 years.

**Results:**

* t-value: 8.3686
* Degrees of freedom (df): 302
* p-value: 2.195e-15
* Sample Mean: 54.37
* 95% Confidence Interval: [53.34, 55.39]

**Interpretation:**

The p-value is significantly less than 0.05 (p < 0.001), indicating strong evidence against the null hypothesis. We reject the null hypothesis and conclude that the mean age of individuals in the dataset is significantly different from 50 years. The mean age of the patients is approximately 54.37 years. We are 95% confident that the true mean age of the population from which this sample is drawn falls between 53.34 and 55.39 years.

**Key Findings**

The mean age of patients in the dataset is significantly different from 50 years, with a mean age of approximately 54.37 years. This finding indicates that age is a crucial factor in the risk of heart disease.

**Hypothesis Testing for Proportion:** To determine if the proportion of patients with high cholesterol (cholesterol > 200 mg/dL) is significantly different from 0.5.

**Hypotheses:**

* Null Hypothesis (Ho): The proportion of patients with high cholesterol is 0.5.
* Alternative Hypothesis (H1): The proportion of patients with high cholesterol is not 0.5.

**Results:**

* Chi-Squared value: 153.06
* Degrees of freedom (df): 1
* p-value: < 2.2e-16
* Sample Proportion: 0.1424
* 95% Confidence Interval: [0.1060, 0.1881]

**Interpretation:**

* The p-value is significantly less than 0.05 (p < 0.001), indicating strong evidence against the null hypothesis.
* We reject the null hypothesis and conclude that the proportion of patients with high cholesterol is significantly different from 0.5.
* The observed proportion of patients with high cholesterol is approximately 14.24%.
* We are 95% confident that the true proportion of patients with high cholesterol in the population falls between 10.60% and 18.81%.

**Key Findings:**

Cholesterol Levels: Only about 14.24% of patients have cholesterol levels above 200 mg/dL. This proportion is significantly lower than 50%, suggesting that a smaller subset of patients has high cholesterol, which is a known risk factor for heart disease.

**Proportion Test for Low Cholesterol:** To determine if the proportion of patients with low cholesterol (cholesterol < 200 mg/dL) is significantly different from 0.5.

Hypotheses:

* Null Hypothesis (Ho): The proportion of patients with low cholesterol is 0.5.
* Alternative Hypothesis (H1): The proportion of patients with low cholesterol is not 0.5.

**Results:**

* Chi-Squared value: 136.45
* Degrees of freedom (df): 1
* p-value: < 2.2e-16
* Sample Proportion: 0.1623
* 95% Confidence Interval: [0.123, 0.210]

**Interpretation:**

* Given this very low p-value, we reject the null hypothesis. Therefore, we conclude that the true proportion of patients with low cholesterol is significantly different from 0.05 (p < 0.001, indicating strong evidence against the null hypothesis.
* The sample estimate for the proportion of patients with low cholesterol is approximately 16.23%. This is much lower than the hypothesized proportion of 50%. We reject the null hypothesis and conclude that the proportion of patients with low cholesterol is significantly different from 0.5.
* We are 95% confident that the true proportion of patients with low cholesterol in the population falls between 12.3% and 21.0%. The confidence interval does not include 0.5, which further supports the rejection of the null hypothesis.

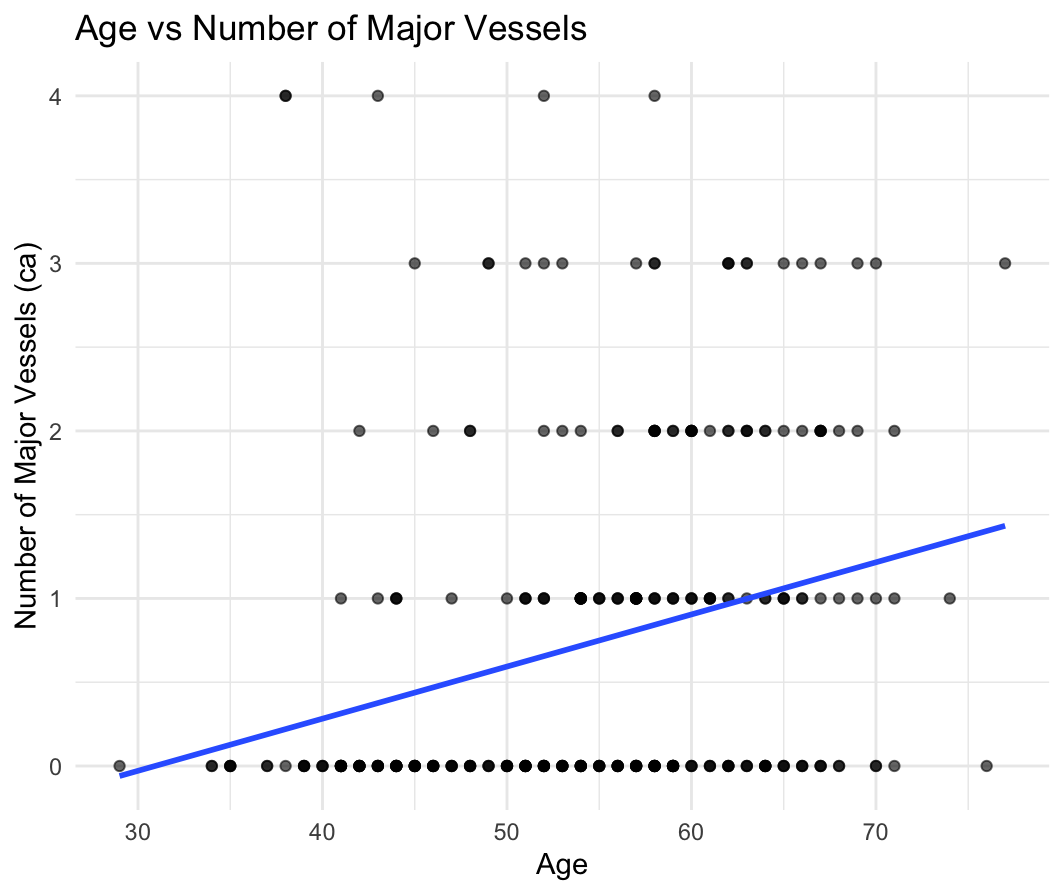
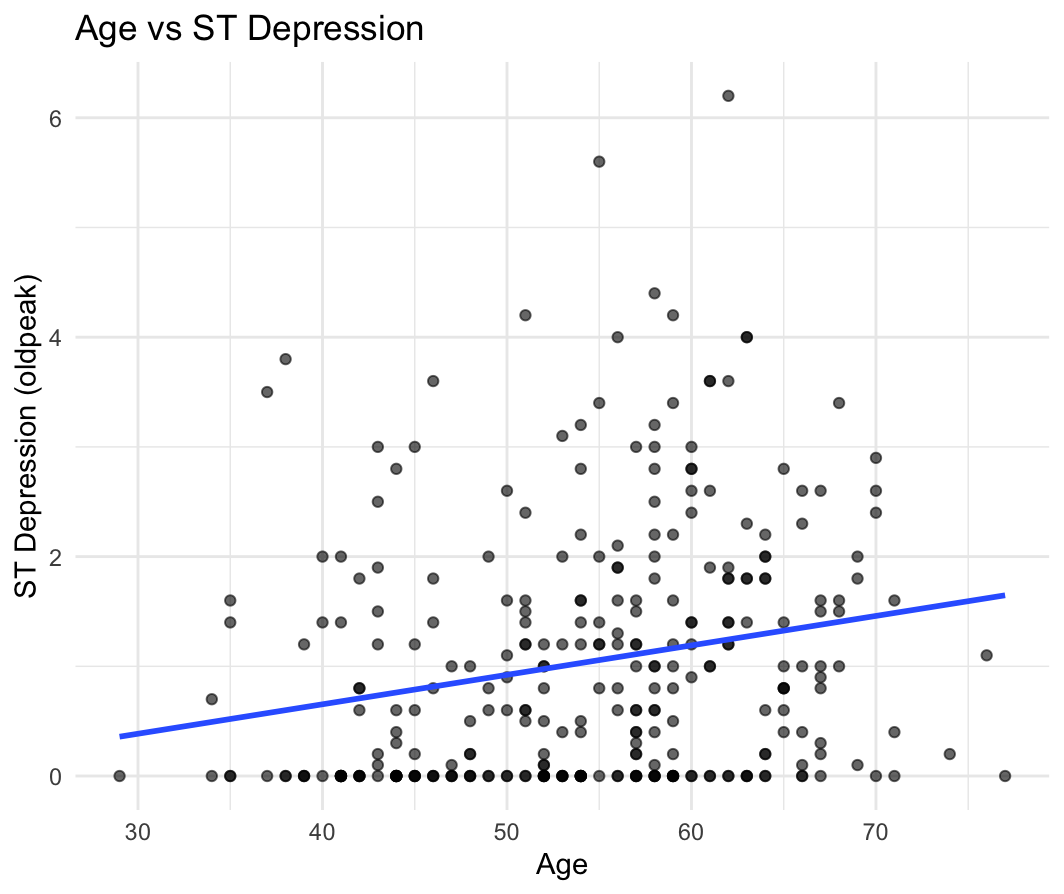
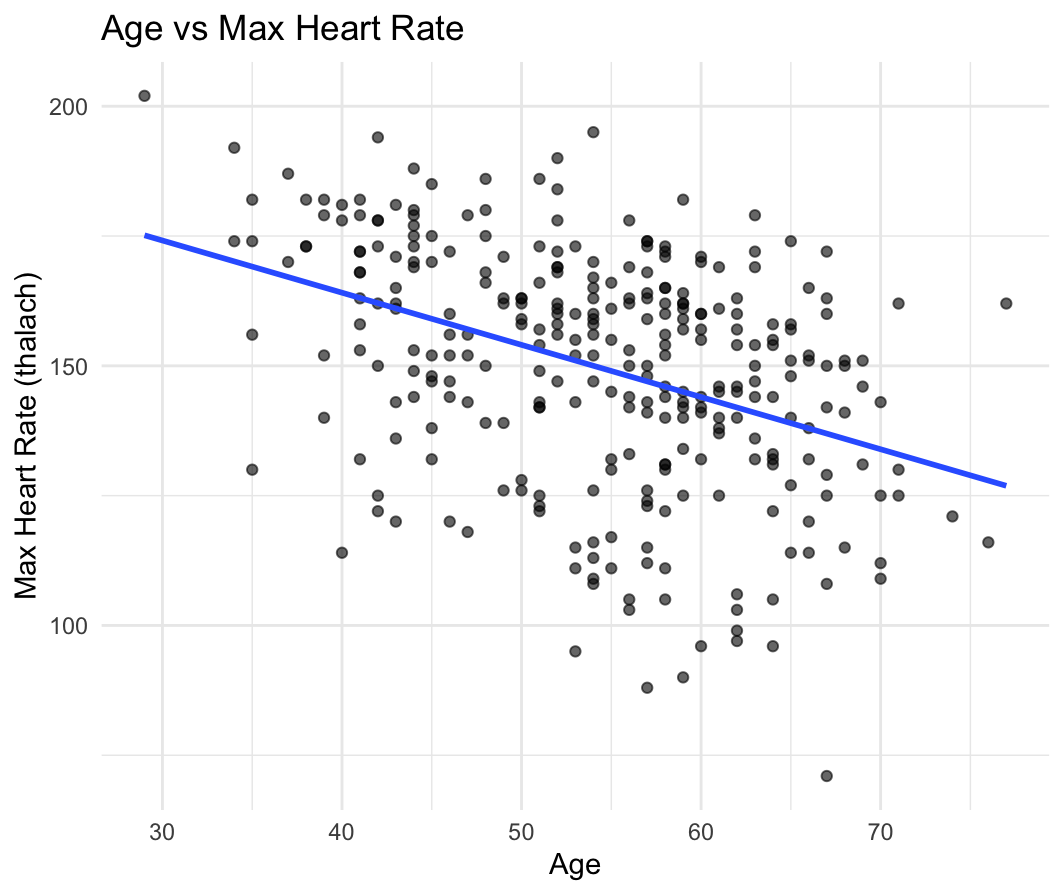
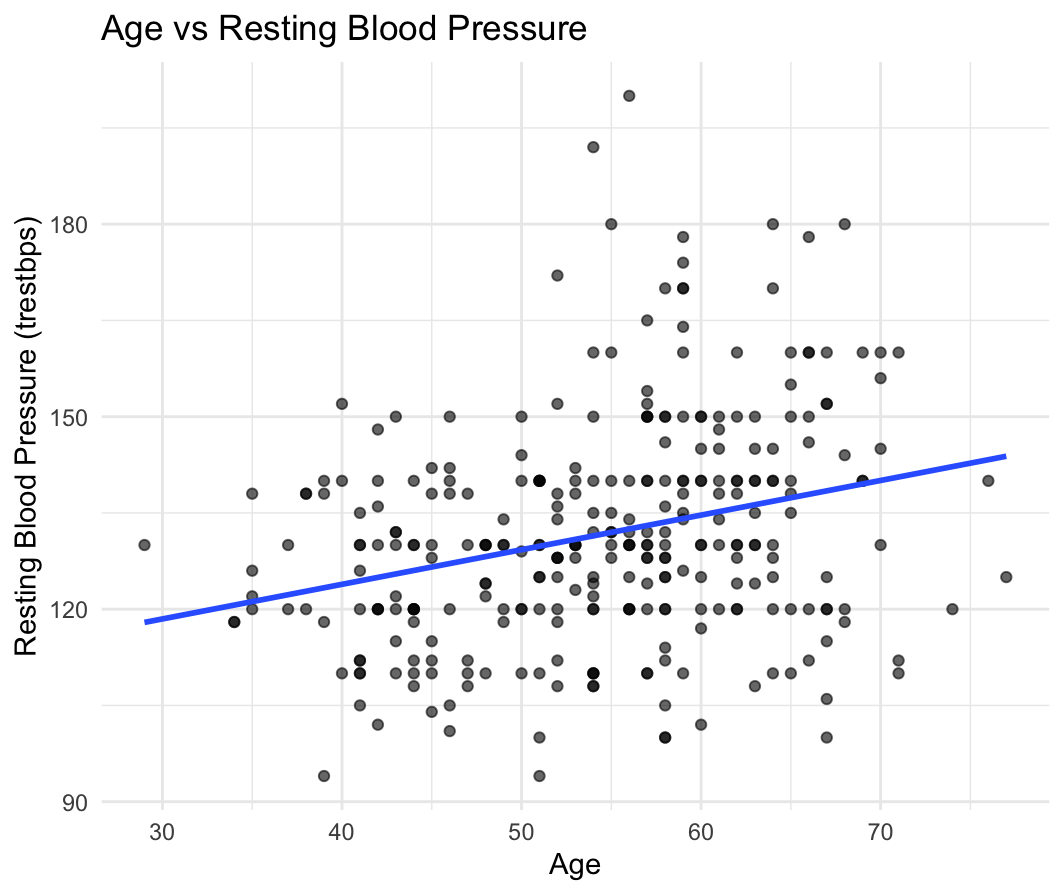
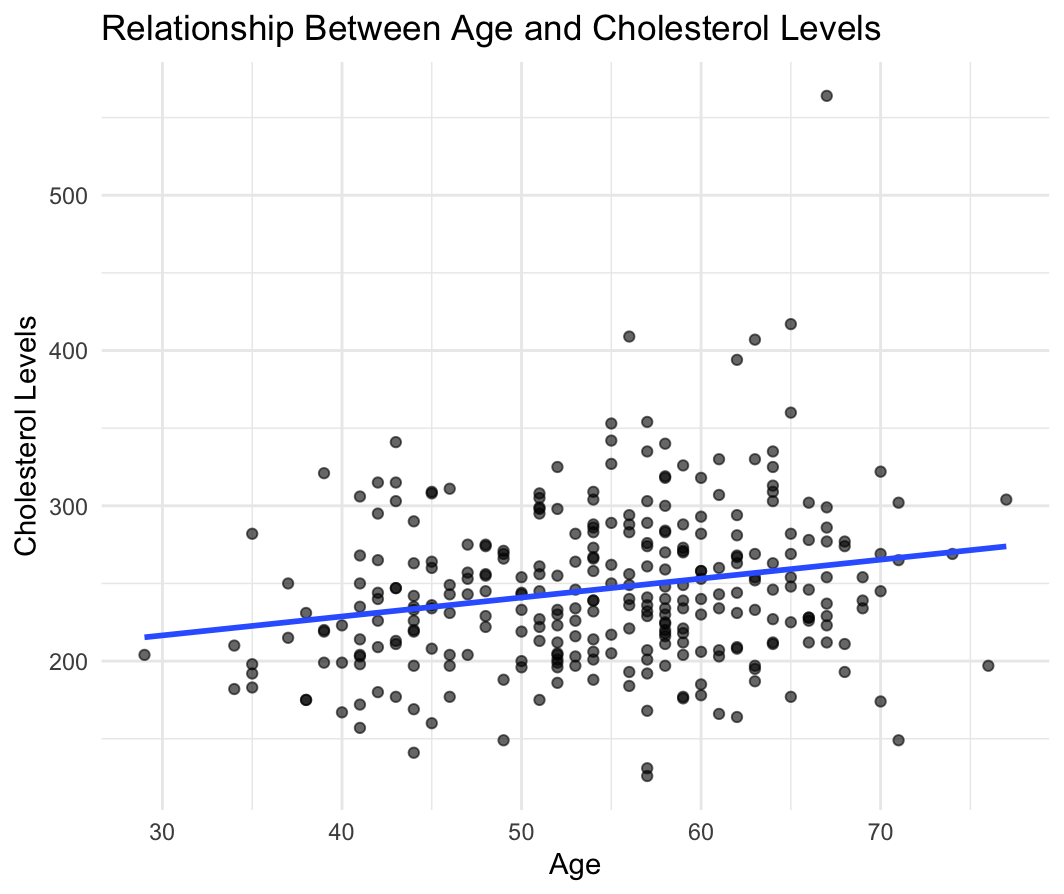
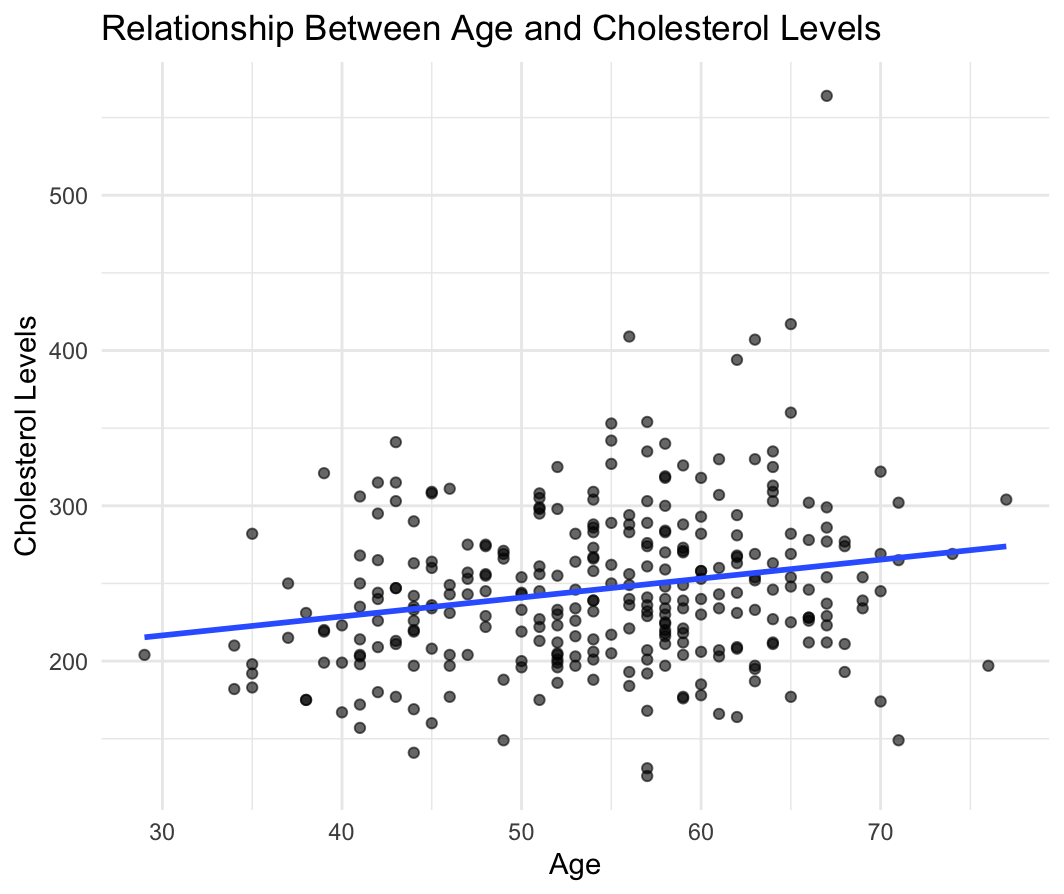
**Key Findings:**

Cholesterol Levels: Approximately 16.23% of patients have cholesterol levels below 200 mg/dL. This proportion is significantly lower than 50%, suggesting that a smaller subset of patients has low cholesterol, which may influence their risk factors for heart disease.

**Analysis of Correlation Matrix for Heart Disease Risk Factors**

The correlation matrix highlights relationships between age, cholesterol levels (chol), resting blood pressure (trestbps), maximum heart rate achieved (thalach), ST depression (oldpeak), and the number of major vessels colored by fluoroscopy (ca).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **chol** | **trestbps** | **thalach** | **oldpeak** | **ca** |
| **age** | 1.0000000 | 0.213678 | 0.279351 | -0.398522 | 0.210013 | 0.276326 |
| **chol** | 0.213678 | 1.000000 | 0.123174 | -0.009940 | 0.053952 | 0.070511 |
| **trestbps** | 0.279351 | 0.123174 | 1.000000 | -0.046698 | 0.193216 | 0.101389 |
| **thalach** | -0.398522 | -0.009940 | -0.046698 | 1.000000 | -0.344187 | -0.213177 |
| **oldpeak** | 0.210013 | 0.053952 | 0.193216 | -0.344187 | 1.000000 | 0.222682 |
| **ca** | 0.276326 | 0.070511 | 0.101389 | -0.213177 | 0.222682 | 1.000000 |

Data Visualization: Here, I have drew scatterplots to observe the relationship between risk factors compared to age which can help predict chances of having Heart Attack in a patient. Each plot shows variations in resting blood pressure, cholesterol levels, maximum heart rate, and ST depression across different ages. These visualizations offer insights into correlations and patterns that could indicate cardiovascular risk. 

### **Key Findings:**

1. **Age**:

* Moderately positively correlated with cholesterol, resting blood pressure, ST depression, and number of major vessels.
* Negatively correlated with maximum heart rate.

1. **Cholesterol (chol)**:

* Weakly correlated with other factors.

1. **Resting Blood Pressure (trestbps)**:

* Positively correlated with age, ST depression, and number of major vessels.

1. **Maximum Heart Rate Achieved (thalach)**:

* Negatively correlated with age, ST depression, and number of major vessels.

1. **ST Depression (oldpeak)**:

* Positively correlated with age and number of major vessels.

1. **Number of Major Vessels Colored by Fluoroscopy (ca)**:

* Positively correlated with age and ST depression

**Interpretation:** Age, maximum heart rate achieved, ST depression, and the number of major vessels colored by fluoroscopy are crucial predictors of heart disease risk. As age increases, other risk factors like cholesterol, resting blood pressure, and ST depression also tend to increase, while maximum heart rate decreases.

**Conclusion:**

The statistical analyses highlight significant predictors of heart disease, including age, cholesterol levels, maximum heart rate, ST depression, and the number of major vessels.

In examining the Heart Attack dataset, we've discovered that age and cholesterol levels play crucial roles in determining heart disease risk. Our analysis, including a one-sample t-test, revealed a significant deviation from the expected average age of 50 years (t = 8.37, p < 0.001), highlighting how age affects susceptibility to heart issues. We also found that both high cholesterol (> 200 mg/dL) and low cholesterol (< 200 mg/dL) levels significantly diverged from expected proportions (p < 0.001), indicating their profound impact on heart health. Through scatterplot visualizations, we saw how variables like resting blood pressure, cholesterol levels, maximum heart rate, and ST depression change with age.

Reference :

* *Heart Disease prediction*. Kaggle. <https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction>
* GeeksforGeeks. (n.d.). How to find Confidence Intervals in R. GeeksforGeeks. <https://www.geeksforgeeks.org/how-to-find-confidence-intervals-in-r/>
* Bluman, A. G. (2016). Elementary statistics: A step by step approach (10th ed.). McGraw-Hill. ISBN 978-1-260-04200-9.
* Kabacoff, R. (2015). R in action (2nd ed.). Manning. ISBN 978-1-617-29138-8.
* Redirecting. (2024). <https://northeastern.instructure.com/courses/174180/modules>
* Lumen. (n.d.). *Concepts in Statistics: Module 8: Inference for One Proportion*. <https://courses.lumenlearning.com/wm-concepts-statistics/chapter/hypothesis-test-for-a-population-proportion-2-of-3/>
* R-bloggers. (2022, December). *Hypothesis Testing in R*. R-bloggers. <https://www.r-bloggers.com/2022/12/hypothesis-testing-in-r/>
* StatQuest. (n.d.). *Hypothesis testing and the null hypothesis* [Video]. YouTube. <https://www.youtube.com/watch?v=VFEDOqpuAeQ>
* GeeksforGeeks. (n.d.). *Hypothesis testing in R Programming*. <https://www.geeksforgeeks.org/hypothesis-testing-in-r-programming/>