INTRODUCTION:

This report is aimed at a study on 'Coronary Artery Disease' for which, steps were taken to bring betterment in the diagnosis. This was done by the hospitals based in Budapest, Zurich, Basel and the US and was performed in the early 2000s. In the given sample, we can observe a total of 211 cases of different patients. In the given sample, there are 6 health-related features and 1 identification feature that have been recorded for each patient. They are - ID of the Patient, Age of the Patient (in years), Sex (0=Male, 1=Female), Chest Pain of 4 different types (0=Typical Angina, 1=Atypical Angina, 2=Non-Anginal, 3=Asymptomatic), Resting Blood Pressure (in mmHg), Cholesterol (in mg/dl), and the Maximum Heart Rate (bpm). The research questions require us to assess two important findings. Firstly, if there is a significant difference in the average age of people with two types of chest pains - 1) Atypical Angina Chest Pain & 2) Asymptomatic Chest Pain. And secondly, we have decide whether Age is a better predictor of Resting Blood Pressure, Cholesterol Level or Maximum Heart Rate with the help of a linear model. These questions are important to assess as Coronary artery disease is a disease in the heart's major blood vessels which limits the blood flow to the arteries due to their narrowing and is of a huge concern amongst doctors, patients & medical researchers. It is of utter importance and in everyone's best interest to know which factors affect this disease and which factors don't.

METHODS:

a) The first research question is based on 24 samples for Atypical Angina Pain & 24 samples for Asymptomatic Chest Pain. It can be answered with the help of a 'Two Sample T-Test'. The columns of interest here are **'Age' & 'ChestPain'**. The reason for choosing this test is because the two samples (Atypical Angina & Asymptomatic) are drawn from two different independent samples. The other reason for choosing this test is because the boxplots of the two samples are approximately similar as well as the histograms are nearly normally distributed. (Even though the histogram for both samples appear slightly skewed, we can consider them due to the less number of samples). Therefore, we assume the hypotheses as follows-

Null Hypothesis (H_0) : There is no significant difference between the two sample average ages for Atypical Angina Chest Pain & Asymptomatic Chest Pain.

Alternative Hypothesis (H_1): There is significant difference between the two sample average ages for Atypical Angina Chest Pain & Asymptomatic Chest Pain.

b) For the second research question, we use a linear model as asked in the question to examine the relationship. The columns of interest here are **'Age', 'RestingBP', 'Cholesterol' & 'MaxHeartRate'**. The two hypotheses (Null & Alternative) are made to examine this linear relationship. But to perform this linear regression, we need to ensure that the three conditions are fulfilled before we proceed. They are as follows-

1) **Identifying if there is a linear relation between the population attributes with the help of a scatter plot. (There should be a linear relation in order to proceed.)**
2) **Checking the normal distribution of residuals by the means of a histogram where the residuals should be normally distributed. (Residuals should be normally distributed to proceed.)**
3) **The residual vs fitted values plot should have a uniform standard deviation on both sides of the horizontal line (zero value line). (There should be a uniform distribution on both sides of the horizontal line to proceed.)**

After the above assumptions are satisfied, we compare the r^2 & r values to verify age is better predictor variable for which dependent variable. Higher the r^2 & closer the value of r to 1, better the fit of the model.

RESULTS OF ANALYSIS:

a) We assume standard deviation to be similar. We have found t-value, degrees of freedom & p-value. No outliers have been found here. As p>0.05, we do not reject the null hypothesis.

H	Write down the null and alternative hypotheses	Stated in Methods (a)
A	Are the assumptions met?	Yes (Stated in Methods (a))
T	What is the value of the test statistic & degrees of freedom?	t=**1.17**, df=(24+24)-2=**46**
P	What is the p-value?	p=**0.247**
D	Do you reject or not reject the null hypothesis?	Do Not Reject (Because p>0.05)
C	Write your conclusion in words:	Stated in Conclusion (a)

b) As we proceed with Age vs Resting Blood Pressure & Cholesterol Levels, we discover that the p-values for them are 0.00 & 0.017 respectively, which are both less than 0. Thus, we reject the null hypothesis for both the attributes. For MaxHeartRate, all assumptions aren't fulfilled.

X & Y Variables	**Age vs Resting Blood Pressure**	**Age vs Cholesterol Levels**	**Age vs Maximum Heart Rate**
1) Is there a linear relationship between the two samples?	Roughly Linear	Roughly Linear	Roughly Linear
2) Is the histogram of residuals approximately normal?	Normally Distributed	Normally Distributed	Right Skewed
3) Are the residuals versus fitted values scattered evenly on either side of the horizontal line?	Constant Spread	Constant Spread	Uneven Spread (more values on the upper end of the horizontal line)
Should we proceed for hypothesis testing?	Yes	Yes	No (Not all assumptions have been met)

Predictor	**r^2 (Goodness-of-Fit)**	**r (Correlation)**
Resting Blood Pressure	0.0613	0.248
Cholesterol	0.0268	0.164

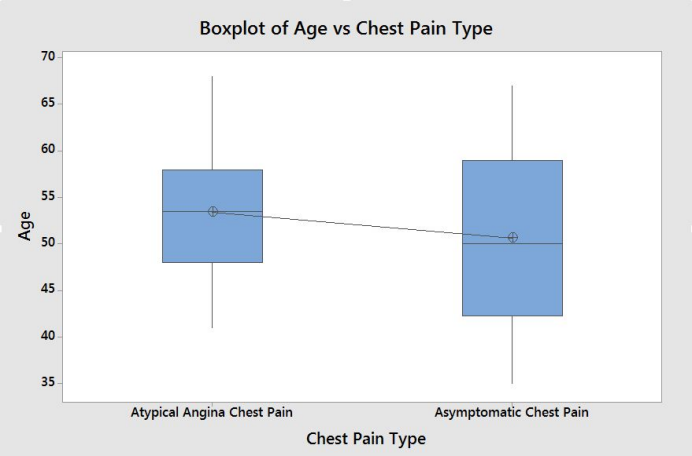
CONCLUSION:

a) Since, P-value = 0.247 which is not less than 0.05 (95% Confidence Interval= (-2.00, 7.58)), so we cannot reject the null hypothesis.
Thus, we can say that evidence suggests that **there is no significant difference in the average age of those admitted with atypical angina chest pain and those admitted with asymptomatic chest pain.**

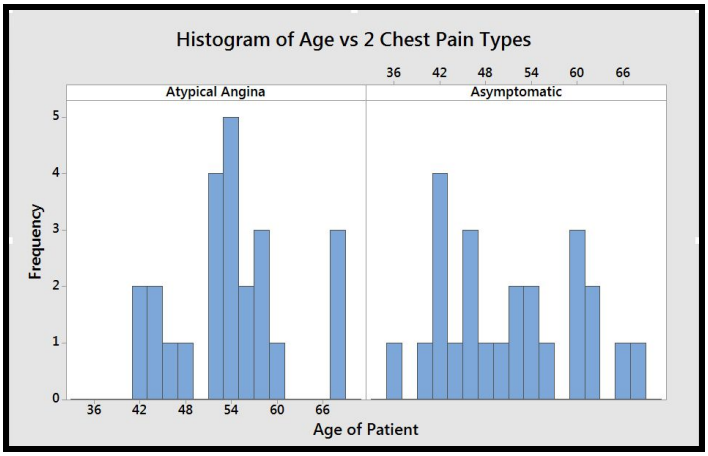
b) Since, P-Value is less than 0.05 for both Resting Blood Pressure & Cholesterol Levels, the linear null hypothesis is rejected for both. Thus, we can say that evidence from the linear model suggests that **Age is better predictor of 'Resting Blood Pressure'** because of higher r^2 value and its 'r' being closer to 1 compared to that of 'Cholesterol's r'. Maximum Heart Rate is out of the question as not all assumptions are satisfied for it. The Linear Regression Equation is **Resting Blood Pressure = 108.28 + 0.396 Age**

APPENDIX

a) 1) Figure 1: Boxplot of Age vs Chest Pain Type



2) Figure 2: Histogram of Age



3) Figure 3: Numerical Summary for Two Sample T-Test

Two-Sample T-Test and CI: Age, Chest Pain Type

Method

μ_1: mean of Age when Chest Pain Type = Atypical Angina Chest Pain
μ_2: mean of Age when Chest Pain Type = Asymptomatic Chest Pain
Difference: $\mu_1 - \mu_2$

Equal variances are assumed for this analysis.

Descriptive Statistics: Age

Chest Pain Type	N	Mean	StDev	SE Mean
Atypical Angina Chest Pain	24	53.42	7.46	1.5
Asymptomatic Chest Pain	24	50.63	8.97	1.8

Estimation for Difference

Difference	Pooled StDev	95% CI for Difference
2.79	8.25	(-2.00, 7.58)

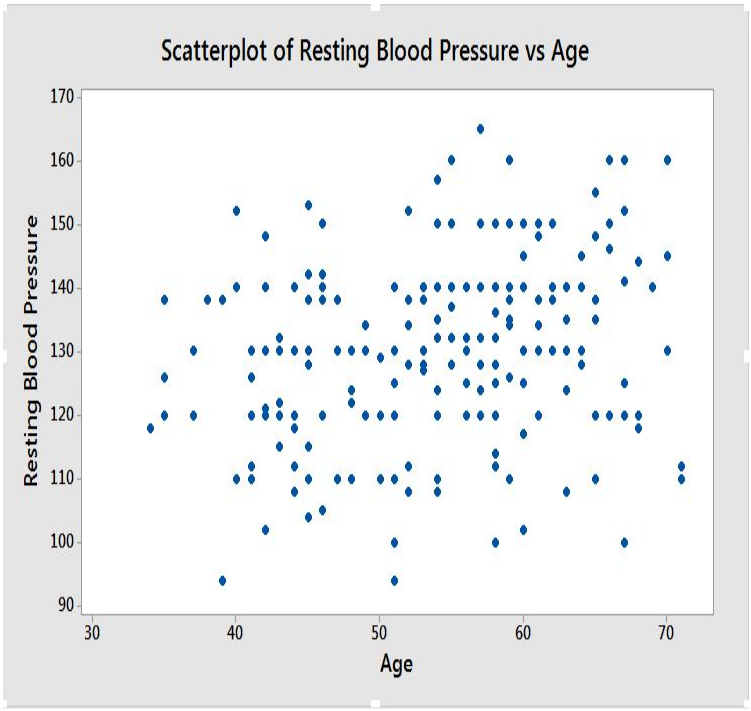
Test

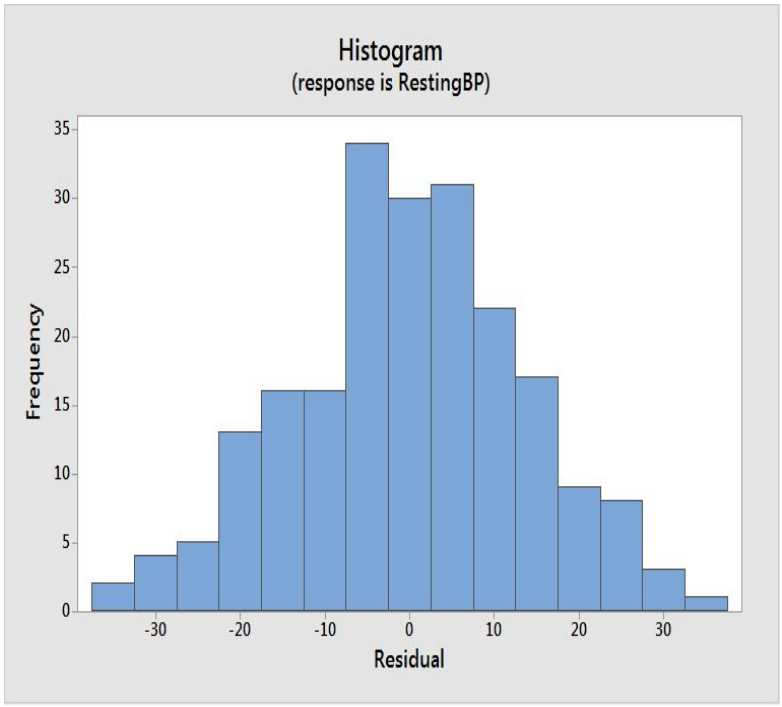
Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

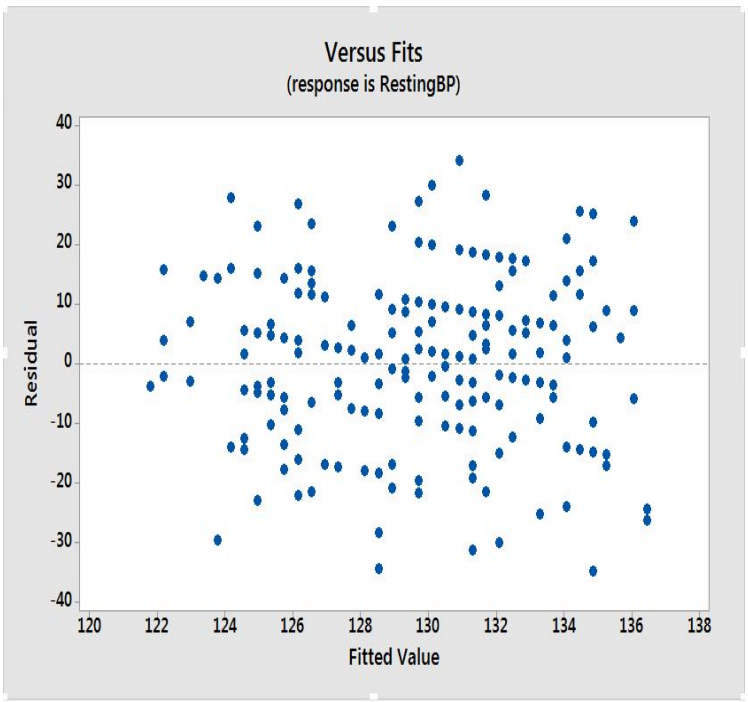
T-Value	DF	P-Value
1.17	46	0.247

b)

Figure 4: Scatterplot, Histogram, Residual vs Fit & Numerical Summary for Age vs Resting Blood Pressure







Regression Analysis: RestingBP versus Age

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13.7408	6.13%	5.68%	4.23%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	108.28	5.84	18.54	0.000	
Age	0.396	0.107	3.69	0.000	1.00

Regression Equation

RestingBP = 108.28 + 0.396 Age