**University of Hertfordshire**

**7COM1079 - Team Research and Development Project**

Final Report

Done By

**Group - 136**

08/01/2021

School of Physics, Engineering and Computer Science

**Table of Contents**

[1. Introduction 3](#_Toc60782375)

[1.a. Purpose 3](#_Toc60782376)

[1.b. Dataset 4](#_Toc60782377)

[1.c. Research Question 5](#_Toc60782378)

[1.d. Hypotheses 5](#_Toc60782379)

[2. Visualization 7](#_Toc60782380)

[2.a. Scatter Plot of Cholestoral vs. Age 7](#_Toc60782381)

[2.b. Frequency Plot of Ages vs Frequency 8](#_Toc60782382)

[3. Analysis 9](#_Toc60782383)

3.a. Histogram Plot on Age……………………………………………………………………………………………………9

3.b. Histogram Plot on Cholestoral………………………………………………………………………………………..10

3.c. Box plot on Age and Cholestoral……………………………………………………………………………………11

3.d. Scatter plot on Age and Cholestoral………………………………………………………………………………11

[4. Conclusion 15](#_Toc60782384)

[Source 16](#_Toc60782385)

**1.Introduction**

**1.a Purpose**

Heart diseases are one of the concerning things in today’s world, previously these kinds of problems were only seen mostly in people aged 40+ but nowadays due to other factors, these kinds of diseases can also be seen in people aged below 40. Among those other factors, cholestoral level in the body can be considered as one of the main factors for that. When there is too much cholesterol in your blood, it builds up in the walls of your arteries, causing a process called atherosclerosis, a form of heartdisease. The arteries become narrowed and blood flow to the heart muscle is slowed down or blocked. The very purpose of our research is about the analysis of heart diseases based on the dataset obtained from the Kaggle website.

There are several ways to test or analyze data like correlation test, students t-test, Chi-squared test, Welch’s Test, Kruskal- Wallis H test, etc., depending on the kind of data, research question and hypotheses.

In our analysis, the data is analyzed by making use of a concept called “correlation” which means that variation of a dependent variable with respect to the variation of an independent variable.

It’s always essential to assess whether there exist a proper dependent and independent variable before analyzing data.

In this analysis, “age” is an independent variable and “cholestoral” is a dependent variable.

**1.b Dataset**

The dataset taken for this project is about the patients suffering or have suffered from heart diseases, it comprises all the attributes of the patients which I would like to explain each of them in detail

age: - Age of patients is years.

sex: - It indicates the gender of patients (“1” denotes male and “0” denotes female).

cp: - This attribute denotes the type of chest pain suffered by patients.

trestbps: - It denotes resting blood sugar of patients (in mm Hg on admission to the hospital).

chol: - It denotes the serum cholestoral level in the body of the patients.

fbs: - It denotes the fasting blood sugar of the patients, “1” indicates true and “0” indicates false.

restecg: - It denotes electrocardiographic results in resting position.

thalach: - It denotes the maximum heart rate achieved by the patients.

exang: - It denotes whether the exercise induced in patients is due to angina (Angina is a type of chest pain caused by reduced blood flow to the heart) or not, “1” indicates yes, “0” indicates no.

oldpeak: - It indicates the amount of depression induced in patients due to exercise relative to rest.

This dataset is free from missing(or) null values.

This project’s research question and hypotheses are about the correlation between age and cholestoral level in the body.

**1.c Research Question**

**RQ**: - Is there any correlation between cholestoral level in the body and age?

It assesses whether there exists a significant or substantial correlation between Cholestoral and Age. This question drives towards conclusion too.

**1.d Hypotheses**

As this project is about assessing the correlation between cholestoral and age from the dataset taken from Kaggle website.

Hence, it’s hypotheses are mentioned as below.

**Null Hypothesis (Ho)**: - There is NO correlation between age and cholestoral level in the body.

**Alternative Hypothesis (H)**: - Yes, there is a correlation between age and cholestoral level in the body.

**2. Visualization**

**Chart, scatter chart

Description automatically generated**

**2.a Scatter plot of Cholestoral level vs Age**

The above graph is obtained by plotting the “Age” variable on

X-axis and “Cholestoral” variable on Y-axis. The above scatterplot is obtained using R studio. This plot shows that there exists a relationship between Age and Cholesterol. The trend line shows that there is a significant increase in the cholesterol level with the increase in age.  It depicts that people or patients of age group between 56 to 70 have acquired more cholesterol than other groups which, these groups are most likely susceptible to heart diseases. The trend line passing through the points in space denotes the trend of the data been plotted, which provides us the significant information.

**2.b Frequency plot of Ages vs Frequency**

Chart, histogram

Description automatically generated The above graph is a frequency plot which is formed by plotting Age variable on X-axis and Cholesterol variable on Y-axis. It is apparent that Age is distributed normally. This plot also appeals that whether the data is distributed normally and whether this corresponding attribute is parametric or non-parametric so that to assess which correlation test must be used which will be discussed in a further section. i.e., Analysis.

**3. Analysis**

Before performing the test on hypotheses, we need to perform several plots like histograms and boxplots to check the normalcy of the data. i.e., to check whether the data is distributed normally or not. This is to assess which correlation must be used as if the data is normal, we use Pearson method or else Spearman or Kendall’s method.

**3.a Histogram plot on Age**

Chart, histogram

Description automatically generated

The histogram(figure) above denotes the frequency of the age variable. The purpose of plotting histogram on Age variable is to find out whether the data of the variable Age is distributed normally or not. The graph apparently depicts that the variable Age is distributed normally as it contains two lower peaks and one highest peak which resembles a bell shape curve.

**3.b Histogram plot on Cholesterol**

Chart, histogram

Description automatically generated

The above (figure) plotted is obtained by plotting cholesterol and it’s frequency. Even this plot also shows that the attribute Cholesterol is distributed normally as it resembles a bell curve.

So, by using histogram we got to know that variables Age and Cholesterol are normally distributed and are parametric.

**3.c Box plot on Age and Cholesterol**

Chart, box and whisker chart

Description automatically generated

The above box plot(figure) is plotted on Age and Cholesterol, in that figure 1 denotes Age variable and 2 denotes Cholesterol variable. In the boxplot, it’s apparent that mean of the variables of Age and Cholesterol are exactly in the middle from their minimum and maximum ranges respectively. By plotting this boxplot, now we are evident that data is distributed normally.

**3.d Scatterplot on Age vs Cholesterol**

Chart, scatter chart

Description automatically generated

The scatterplot above is plotted and the trend line is also plotted too for finding the trend but the main purpose is just to know whether there exists a certain relationship between age and cholesterol. This plot apparently shows that there does exist a certain relationship between them.

After performing the plots like histogram and boxplot, now we can perform the test of hypotheses. As the data is distributed normally over the space, it is evident that Pearson’s correlation test can be used.

The formula for calculating the corresponding correlation is:

* If r = 1, the correlation is said to be perfect positive where all points fall on a line with a positive slope.
* If r = -1, the correlation is said to be perfect negative correlation where all points fall on a line with a negative slope.
* If r = 0, it indicates that there exist no correlation.

This correlation test is the appropriate method because the data is distributed normally. It also utilizes appropriate features while performing a correlation test.

The correlation coefficient is calculated between the two variables of interest using the ***corr()*** function in R. The variables used for testing hypotheses in this project are Age and Cholesterol.

***Graphical user interface, text, application

Description automatically generated***

As shown in the above figure, corr() function is used to test the correlation between the two variables, the obtained value appeals that there exists a significant correlation is it is a positive correlation. Also in plot (figure (plot 3.2)), it can be seen that the trend line has a positive slope where a significant number of points lie on it, as the line follows the trend from low to high, it shows that there exists a correlation and it is positive.

To perform the final test of correlation which supports evidence for one of the hypotheses, we use ***corr.test()*** function to evaluate the p-value with which we will get a better idea to either accept or reject the null hypothesis(Ho).

Graphical user interface, text, application

Description automatically generated

As shown in the above figure, ***corr.test()***  function is used to evaluate the p – value by using Pearson’s method of correlation test. As the obtained p – value is 0.0001786. Which is

P – value = 0.0001786 << 0.05. It suggests that we can reject the null hypothesis with the 95% level of confidence.

**4. Conclusion**

The purpose of this project is all about analyzing the heart diseases dataset obtained from Kaggle website to study the age groups which are prone to heart diseases and to prevent that.

Based on the result obtained after performing the Pearson correlation test, it shows that there exists a significant relationship or correlation between the Cholesterol level in the body and Age of the patients. Therefore, this test provides evidence to reject the Null Hypothesis at 95% confidence level and provides substantial evidence to support the alternative hypothesis.

**Source**

https://www.kaggle.com/ronitf/heart-disease-uci