**A picture containing logo

Description automatically generated**

**Project: Using Exploratory Data Analysis and Visualization to predict heart diseases.**

**Technical Report**

**Course: Data visualization**

**Instructor: Dr. Osama Abdulhay**

**Group Members:**

**Reema Ma’en - 20190150**

**Osama Rakan - 20190341**

**Amr Amr - 20190633**

**Saif Abu Al Raisi - 20190465**

**Abdulrahman Shabaneh - 20190546**

**Rama Al Alawneh – 20190153**

**Detailed contribution of each member:**

We worked in parallel but mainly each one contributed the following:

Reema maen: analytical

Rama alawneh: technical

Osama rakan: code

Amr amr; code

Abulrahman shabaneh: technical

Saif abu al raisi: powerpoint

# **Abstract**

As a requirement for our data visualization project it has been requested of us to write a technical report, thus we made this report which is to be read by a technician. After a long process of exploring, cleaning, and modeling the data, it was time to come up with this reliable report that would provide an overview for the data statistically and visually, such as data discrepancies, issues, and irregularities After some time of brain-storming ideas we came up with solutions for each and every problem we faced while exploring the data; which were dropping null, duplicate rows which in fact did make up a small percentage of the data, as well as adding new columns such as Max\_Heart\_Rate\_Difference which displayed the difference between the standard maximum heart rate for a patient and the patients actual maximum heart rate.

We came up with this analysis as a result of using some python libraries such as seaborn , matplotlib, and plotly to plot the data. We used scatter plots, heat maps, boxplots, and made a new data frame which we made for comparison reasons, it’s named DF1, DF1 didn’t have nulls and duplicate values.

# **Introduction**

The data set in which this report is based upon represents the information of people who were admitted to the hospital, it displays medical information regarding the patient the final task is to use the features available to come up with the outcome of interest which is whether the patient has CHD (coronary heart disease) or not. This paper discusses methodology used to reach the desired aim. The methodology used is centered around visualizations and feature selection.

# **Methodology**

* basic data exploration was made using ‘describe()’, ‘info()’ which are python built in functions.
* Column ‘outcome’ values were manipulated
* Nulls and duplicates were discovered in the dataset the appropriate measure was taken
* “thal” and “ca” columns were also found out to be of type object while they are supposed to be of type integer.
* An unnecessary column “ID” that adds no new information to the process was noted and dealt with accordingly.
* Discrepancies and inconsistencies where found in “chol”,” thalack”,” age” and” oldpeak” columns in which 5 values were suspected to be wrong entries by using boxplot to detect them.
* The distribution of the features was viewed using histogram and normality was checked by qqplot.
* Heat map was used to detect the correlations between “outcome” column and the other features.
* We plotted each feature with the other
* By that we decided to plot the most relevant features with the outcome
* Based on the results in the previous step, new features were added that contributed more to the desired objective.
* Summary statistics were made again after performing data cleaning.

# **Results**

* The result of the summary statistics before cleaning the data shown in Appendix A
* Null values where dropped, as well as the duplicated rows since they don’t make up a large percentage of the data, the null values were 2% of the entire dataset presented by “?” while duplicates were 2.2%, thus they were dropped.
* Column ‘outcome’ values were changed to zeros or ones only since the outcome of interest is concerned with only whether the patient has or doesn’t have CHD
* Convert “ca”,” thalack”, “thal” into columns of type integer.
* “ID” column was dropped since it adds no useful information in the process, Afterall the main concern is not about the “who’s”.
* Data inconsistencies, after viewing the boxplots for each numerical feature, we noticed that the so-called outliers where more or less human errors, thus we contacted an expert and verified our hypothesis which then resulted in the changes shown in Table 2 Adjustments.
* Through using the qqplot, it was clear that columns.” thalack”,“age”, “Restbps”, and “chol” had normal distribution Another notable feature about “age” column is that most of the ages are around 54 meaning that most of the test subjects were elders. As for “oldpeak”, it was a right skewed distribution as shown in **Appendix C**.
* A heatmap was used by function ‘corr()’ to find out the correlations between all the features and the target “outcome”, “chol” had the highest correlation of 0.66 ,”ca’ with 0.52, “thal” with 0.51,”oldpeak ” 0.50 as shown in Appendix D.
* Bivariate relations are shown in Figure 8
* A new column was added “Normal\_Max\_Heart\_Rate” which is calculated by subtracting the patients age from 220.
* Another new column was added “Max\_Heart\_Rate\_Difference” which represents the difference between “Normal\_Max\_Heart\_Rate” and “thalack” columns. It displays the difference between the standard maximum heart rate for a patient and the patients actual maximum heart rate, if the number was negative then it could be safely assumed that the patients actual heart rate is higher than the standard maximum heart, otherwise it’s a positive number.
* To check how useful it will be we checked its values through the correlation matrix and it gave us 0.36
* The result of the summary statistics after the preprocessing is shown in Table 3

# **References**

[1] CDC (2019). *Target Heart Rate and Estimated Maximum Heart Rate | Physical Activity | CDC*. [online] www.cdc.gov. Available at: <https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm>.

# **Appendix**

## **Graphical user interface, text Description automatically generatedAppendix A**

Table summary statistics

## **Appendix B**

Chart, bar chart, box and whisker chart

Description automatically generatedGraphical user interface, application, Teams

Description automatically generatedGraphical user interface, application, Teams

Description automatically generated with medium confidence

Figure 1 age, chol, oldpeak boxplots

Table 2 Adjustments

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Noise** | **After adjusting the Noise** |
| chol | 3600 | 360 |
| thalack | 1,71 and 1,42 | 171 and 142 |
| age | 5.7 | 57 |
| oldpeak | 34 | 3.4 |

## Chart, line chart Description automatically generated**Shape Description automatically generatedAppendix C**

Figure chol distribution, histogram (left) qqplot (right)

Chart, line chart

Description automatically generated**Chart, histogram

Description automatically generated**

Figure oldpeak distribution, histogram (left) qqplot (right)

**Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated**

Figure age distribution, histogram (left) qqplot (right)

**Chart, line chart, scatter chart

Description automatically generated**Chart, histogram

Description automatically generated

Figure restbps distribution, histogram (left) qqplot (right)

**Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated**

Figure thalack distribution, histogram (left) qqplot (right)

## **Appendix D**

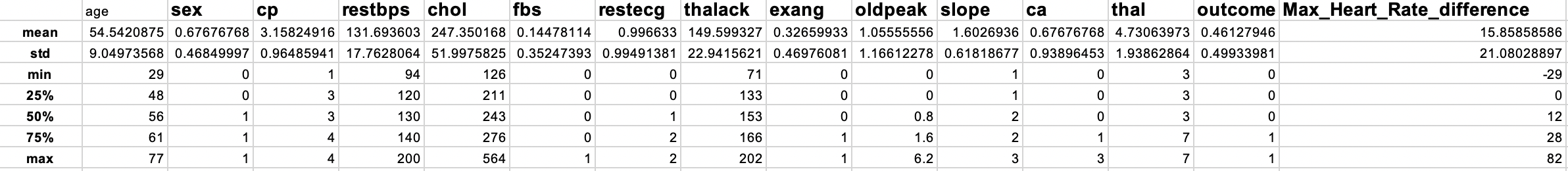
Chart

Description automatically generated

Figure 7 Correlation matrix

## **Appendix E**

Table



‌

A picture containing graphical user interface

Description automatically generated

Figure