**Capstone Project Final:**

**Predicting Heart Disease with Different Machine Learning Techniques**

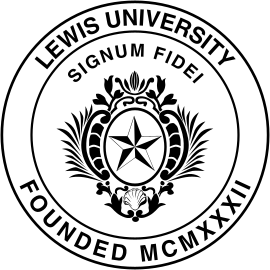
Anabel Aguilar Ramirez

Agata Fietko

Tiffany Cerny-Czykier

Lewis University

Romeoville, IL



**TABLE OF CONTENTS**

**1. INTRODUCTION**……………………………………………………………………………….……3

**2. DATA DESCRIPTION**………………………………………………………………………….……3

**3. METHODOLOGY**……………………………………………………………………………………4

**4. FINDINGS AND ANALYSIS**……………………………………...………….……………………...5

1. Database Management Systems……………………..…………………………………….……5
2. Business Data Warehousing…………………………………………………………………….9
3. Accounting and Business Information Systems………….……………………………………10
4. Decision Science / Business Statistics…………………….…………………………………...12
5. Data Mining Tools…………………………………………………………………………..…13
6. Business Intelligence………….……………………………………………………………….17
7. Business Process Automation………………………………………………………………….22
8. Introduction to Business Analytics…...………………………………………………………..23
9. Business Data Visualization……………..…………………………………………………….28
10. Web Analytics……………...…………………………………………………………………30

**5. CONCLUSION AND RECOMMENDATIONS**…………………...………………..…………….32

**6. LIMITATIONS**…………………………………………………………..………..…………………32

**7. REFERENCES**……………………………………………………………….……………………...32

1. **Introduction**

Heart disease remains one of the leading causes of mortality worldwide, making early detection and prevention strategies critical for improving health outcomes. There are different conditions that can affect the heart, including coronary artery disease, arrhythmias, and congenital heart defects, among others. The complexity of heart disease, combined with the variety of risk factors such as age, gender, blood pressure, cholesterol levels, and lifestyle choices makes it a challenging condition to predict and manage effectively. This project aims to utilize different machine learning techniques to potentially predict the presence of heart disease in patients, thereby facilitating early intervention and increasing the chances of saving lives.

1. **Data Description**

This dataset compiled in 1988 was obtained from the Website Kaggle, it encompasses information from four distinct databases: Cleveland, Hungary, Switzerland, and Long Beach V. The dataset includes a total of 303 observations, each representing a different patient profile, with 14 attributes:

**Age:** Numerical, indicating the age of the patient.

**Sex:** Categorical, with 1 representing male and 0 female.

**CP:** Categorical, encoded as 1 to 4, each number representing a different type of chest pain where 0: Typical Angina, 1: Atypical Angina, 2: Non-anginal Pain, and 3: Asymptomatic

**trtbps:** Numerical, the patient's resting blood pressure in mm Hg at admission to the hospital.

**Chol:** Numerical, the patient's cholesterol measurement in mg/dl.

**fbs:** Categorical, where 1 indicates a fasting blood sugar greater than 120 mg/dl, and 0 otherwise.

**restECG:** Categorical, with values indicating the Resting Electrocardiographic Results as 0: Normal, 1: Abnormality, 2: Hypertrophy based on Estes' criteria.

**thalachh:** Numerical, the maximum heart rate achieved by the patient.

**exng:** Categorical, where 1 indicates the presence of angina induced by exercise, and 0 indicates no angina.

**Oldpeak:** Numerical, the ST depression induced by exercise relative to rest.

**slp:** Categorical, with values representing the slope of the peak exercise ST segment where 0: Upsloping, 1: Flat, and 2: Downsloping.

**caa:** Numerical, indicating the number of major vessels colored by fluoroscopy.

**thall:** Categorical, with values indicating different types of thalassemia as 0: Normal, 1: Fixed Defect, and 2: Reversible Defect.

**Output:** Categorical, indicating the presence (1) or absence (0) of heart disease (target). It is the target variable.

1. **Methodology**

**Data Gathering and Preprocessing:**

* Conducted research about Heart Attack from CDC Website and Mayo Clinic.
* Obtained the Heart Attack Prediction dataset from the Kaggle Website.
* Cleaned and preprocessed the dataset in Excel to remove missing values and inconsistencies and ensure data quality for efficient and accurate analysis.

**Database Management Systems:**

* Created a database with patient information utilizing Lucid Chart and MSSQL.

**Business Data Warehousing:**

* Built a business data warehouse utilizing an Integration Services Project within Visual Studio (SSIS).

**Accounting and Business Information Systems**

* Utilized tools like Microsoft Excel, Python, Microsoft Access, and PowerPoint to collect, display, and process data.

**Decision Science / Business Statistics**

* Gathered quantitative measurements to analyze each of the 14 attributes and summarize their values.

**Data Mining Tools**

* Utilized different tools like Python or Excel to analyze the outliers, correlation, and distribution in the data.

**Business Intelligence**

* Created cost sensitivity analysis with the help of Weka.

**Business Process Automation**

* Created an automation system using Python.
* Implemented BPA tool to improve efficiency of data analysis.

**Introduction to Business Analytics**

* Utilized four forms of analytics: descriptive, diagnostic, predictive, and prescriptive.

**Business Data Visualization:**

* Created visualizations with reports using Tableau to present insights and our findings from the Heart Attack dataset.

**Web Analytics**

* Created infographics to emphasize and show the importance of heart health and preventative measures.
* Created Google alerts of related data.

**4. Findings and Analysis**

**A. Database Management Systems**

Database management system is a tool that allows data retrieval, management, operation, and manipulation. To manipulate our dataset, we created the following subsystem:

**Subsystem Description**

The "Heart Disease Prediction & Management System" is a specific subsystem that is part of the larger digital health infrastructure of a healthcare company. It facilitates the tracking of patient demographics, clinical data, lifestyle factors, and disease outcomes, providing a comprehensive view necessary for heart disease prediction and research.

**List of tables, description of each table, and their relationships.**

patients Table

Data Summary: Stores basic information about the patients involved in the study. It includes columns such as PatientID (integer), Age (integer), Gender (varchar), County (varchar)

Relationship: This is the central table with one-to-many relationships to ClinicalInformation, LifestyleFactors, and DiseaseOutcome tables.

PK, FK: PatientID is the primary key. There are no foreign keys in this table.

clinicalInformation Table

Data Summary: Contains detailed clinical data for each patient visit or assessment. This table includes columns such as ClinicalID (integer), PatientID (integer, FK), Chest Pain (integer), BloodPressure (float), Cholesterol (float), and HeartRate (integer).

Relationship: Has a many-to-one relationship with the Patients table, indicating multiple clinical records can be associated with a single patient.

PK, FK: ClinicalID is the primary key, while PatientID is the foreign key that links to the Patients table.

lifestyleFactors Table

Data Summary: Captures lifestyle-related information that could influence heart health, including SmokingStatus (varchar), PhysicalActivityLevel (integer). Columns include LifestyleID (integer), PatientID (integer, FK), SmokingStatus (varchar), PhysicalActivityLevel (integer),.

Relationship: Exhibits a many-to-one relationship with the Patients table, as patients can have multiple lifestyle factor records over time.

PK, FK: LifestyleID is the primary key, with PatientID serving as the foreign key connecting back to the patients table.

diseaseOutcome Table

Data Summary: Records the outcomes of heart disease assessments for each patient. Includes OutcomeID (integer), PatientID (integer, FK), and HeartDiseaseStatus (varchar) to indicate the presence or absence of heart disease.

Relationship: This table also has a many-to-one relationship with the Patients table since a patient may undergo several assessments over time.

PK, FK: OutcomeID is the primary key, while PatientID is the foreign key linking to the Patients table.

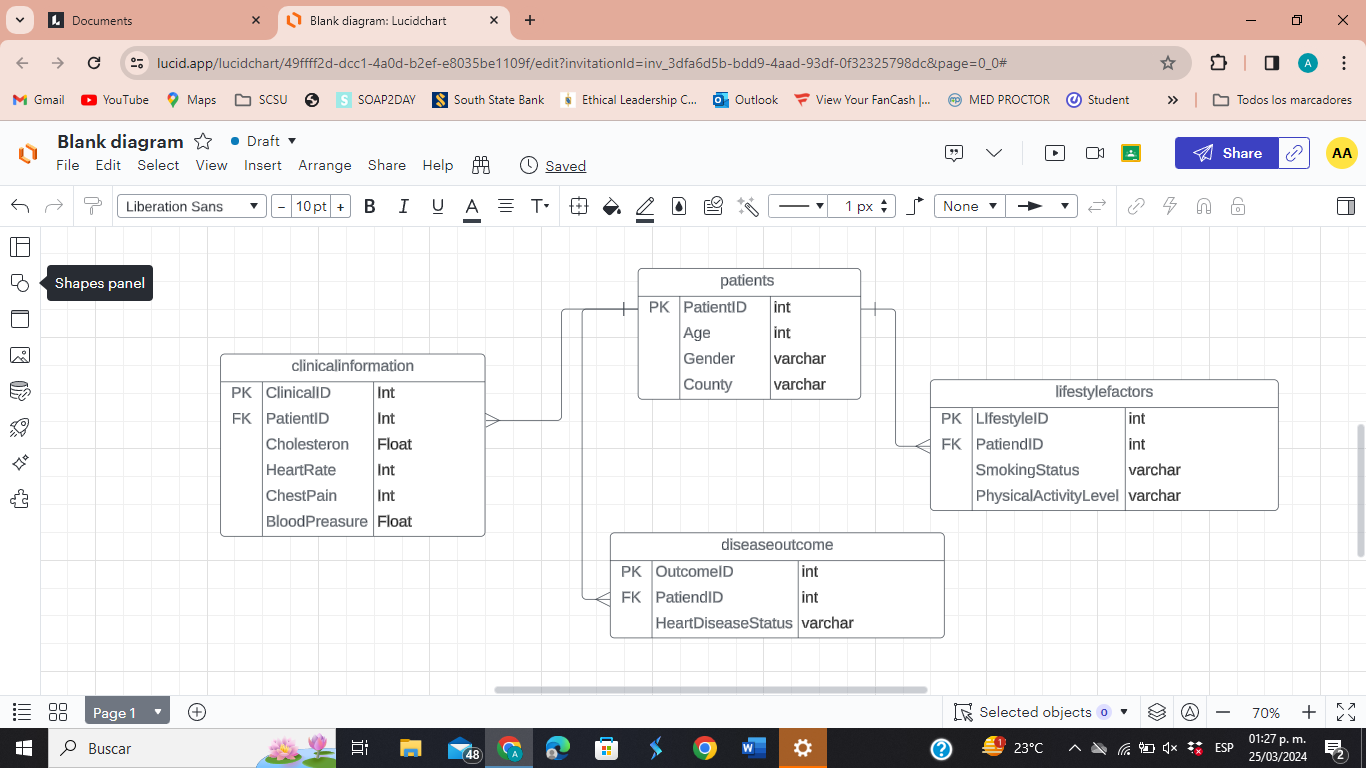
**Descriptions of Relationships between tables**

patients to clinicalInformation: A patient can have multiple clinical records over time, reflecting different visits or assessments.

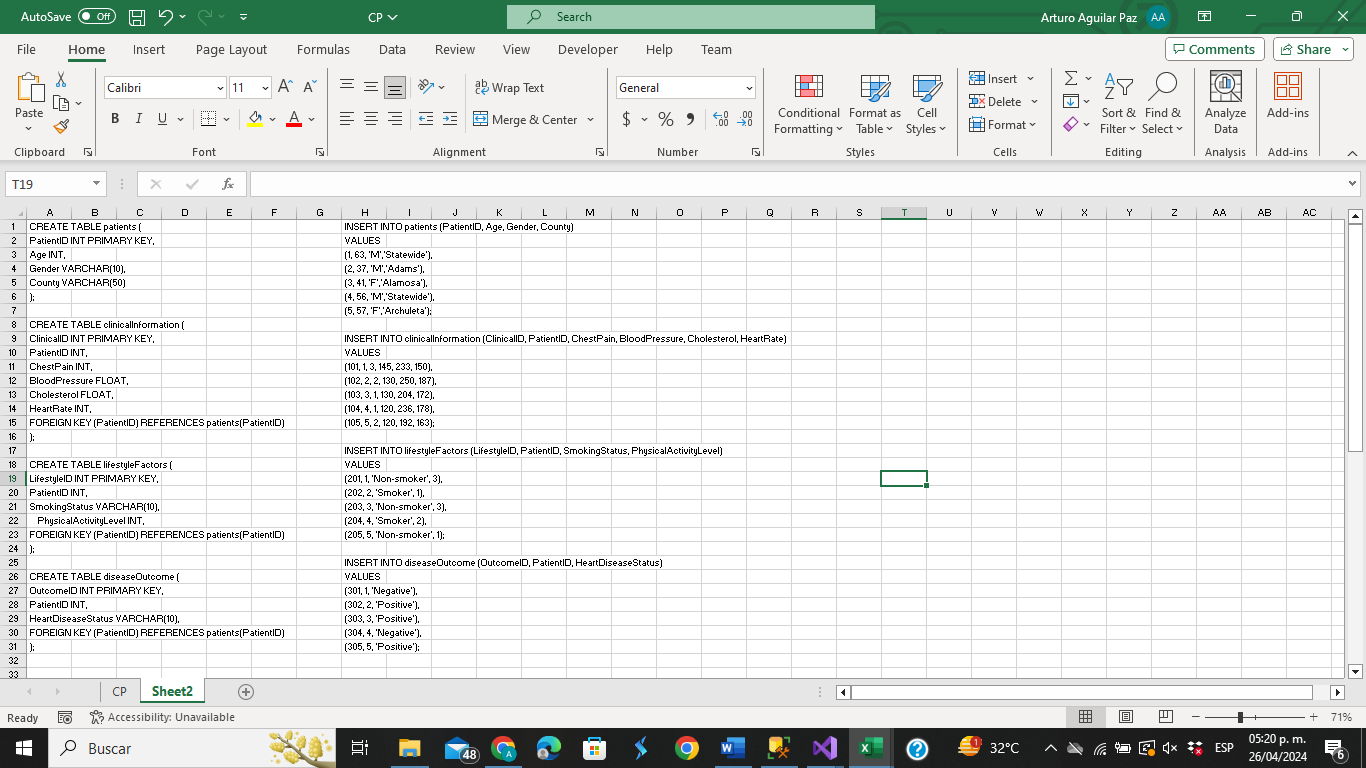
patients to lifestyleFactors: A single patient might change lifestyle factors over time, necessitating new records for each assessment or update.

patients to diseaseOutcome: Patients can have multiple disease outcome records, indicating ongoing assessments of their heart disease status.

**Lucid chart ERD design**



**SQL script**



**SQL Fiddle**



**Database in SQL Server**

****

Healthcare professionals can find patterns, trends, and risk factors related to heart disease by utilizing this system. With the help of machine learning algorithms and predictive analytics, we can predict the probability of heart attacks based on data from the system.

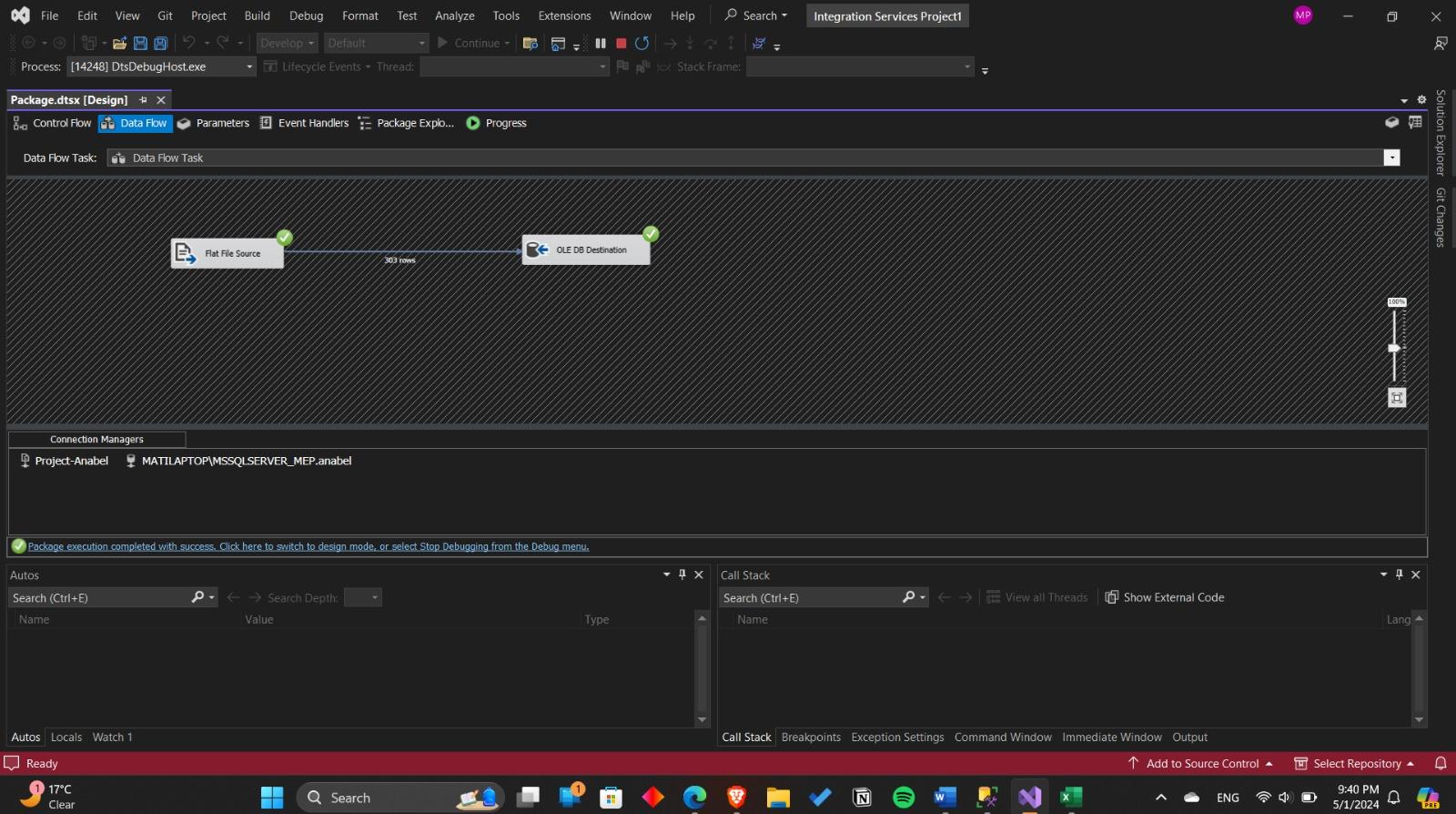
**B. Business Data Warehousing**

The fundamental elements of data warehousing architecture encompass data gathering, storage, and distribution, often referred to as ETL processes.

For our project, to load the data to the Database we extracted the important data from our main dataset to another Excel sheet. Then we cleaned and transformed it into a CSV. file. Later, we set up the Database in the SQL Server to stage the data. After this, we used Visual Studio (SSIS) to upload the data in the SQL Server staging area by using an Integration Services Project.

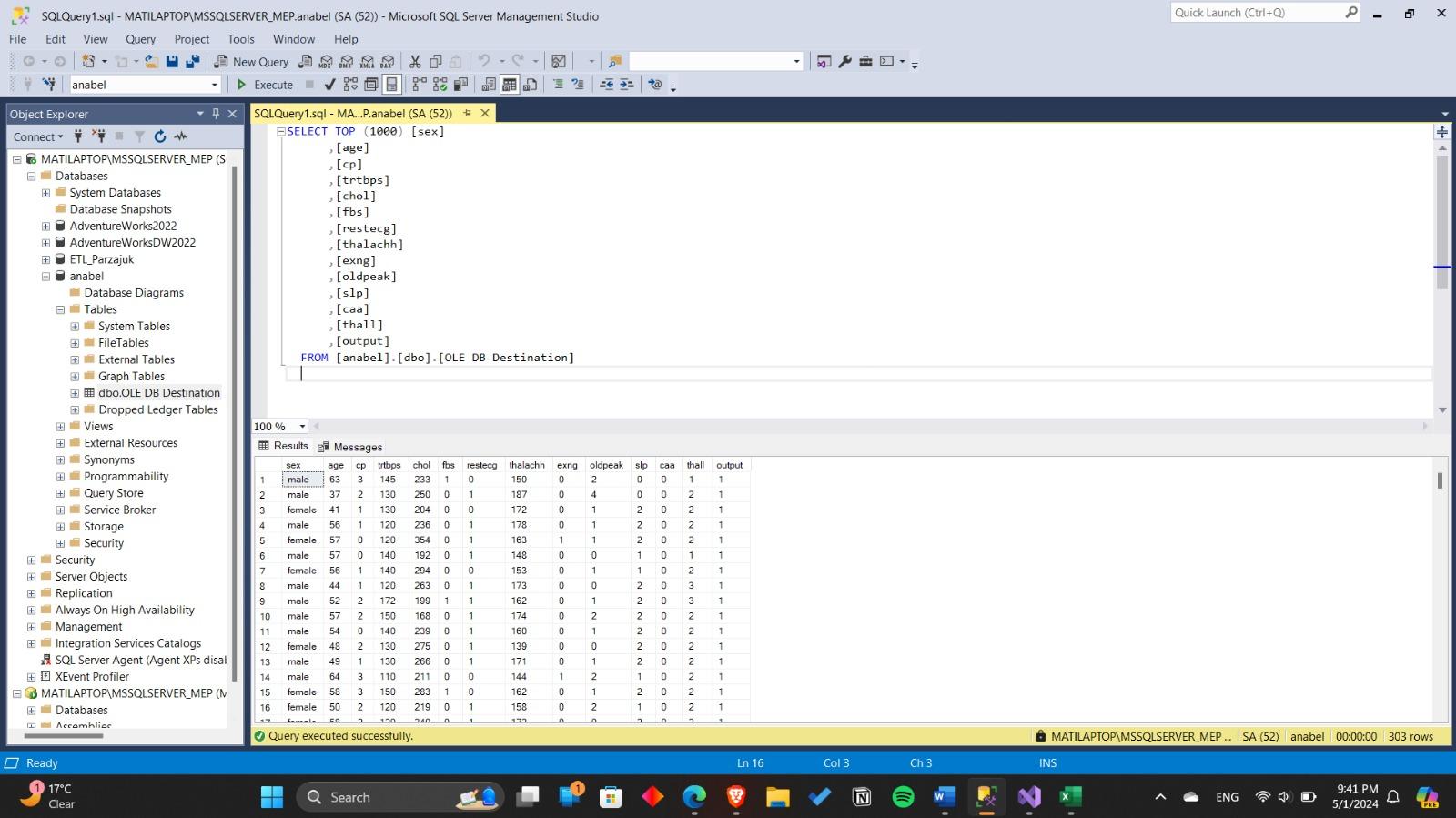
The process began by creating a Data Source (Flat File) with all the required specifications and then we connected it to the OLE DB Destination Editor that we created (Database) and added a table with SQL command, we made sure to confirm mappings and error outputs. After this process, we obtained a new table into our Database.

**Flat File to OLE DB Destinator**

****

The above picture represents the result of successfully creating a connection of our input and destination source as well as effectively making a connection to the SQL server and the new sample file.

**Database after ETL Process**

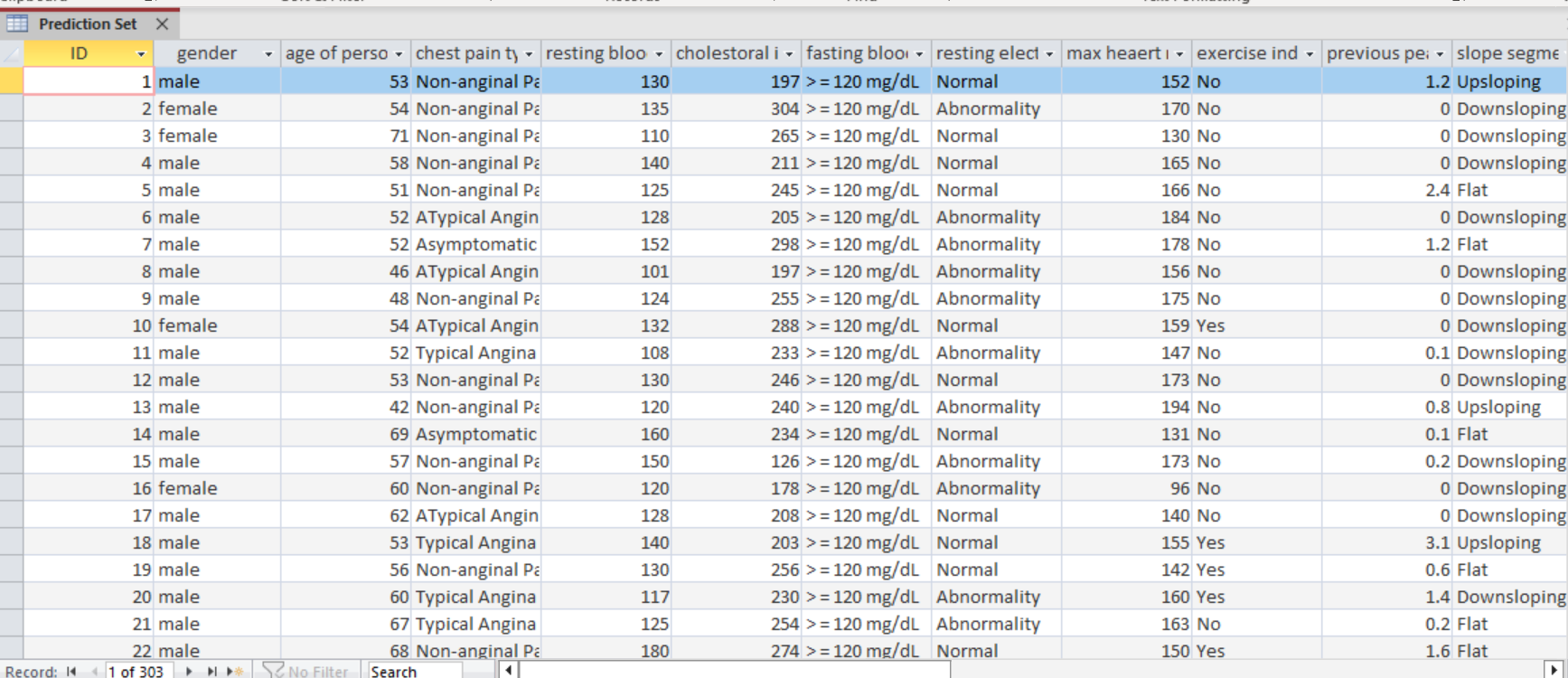


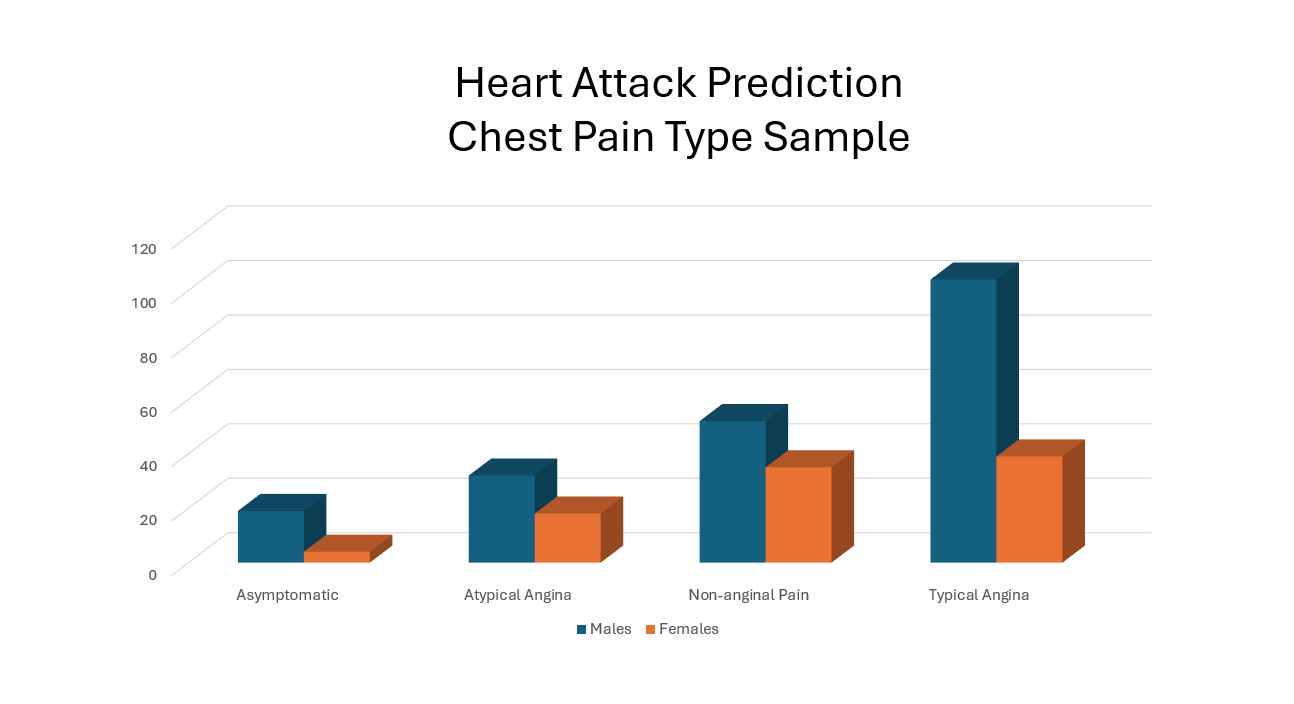
By using ETL (Extract, Transform, Load) processes in data warehousing we can store our data in a structured way to ensure that it is properly prepared, cleaned, and loaded into a database for further analysis and reporting. Once configured, ETL processes can be automated to run at scheduled intervals, reducing manual effort and ensuring data freshness.

**C. Account and Business Information Systems**

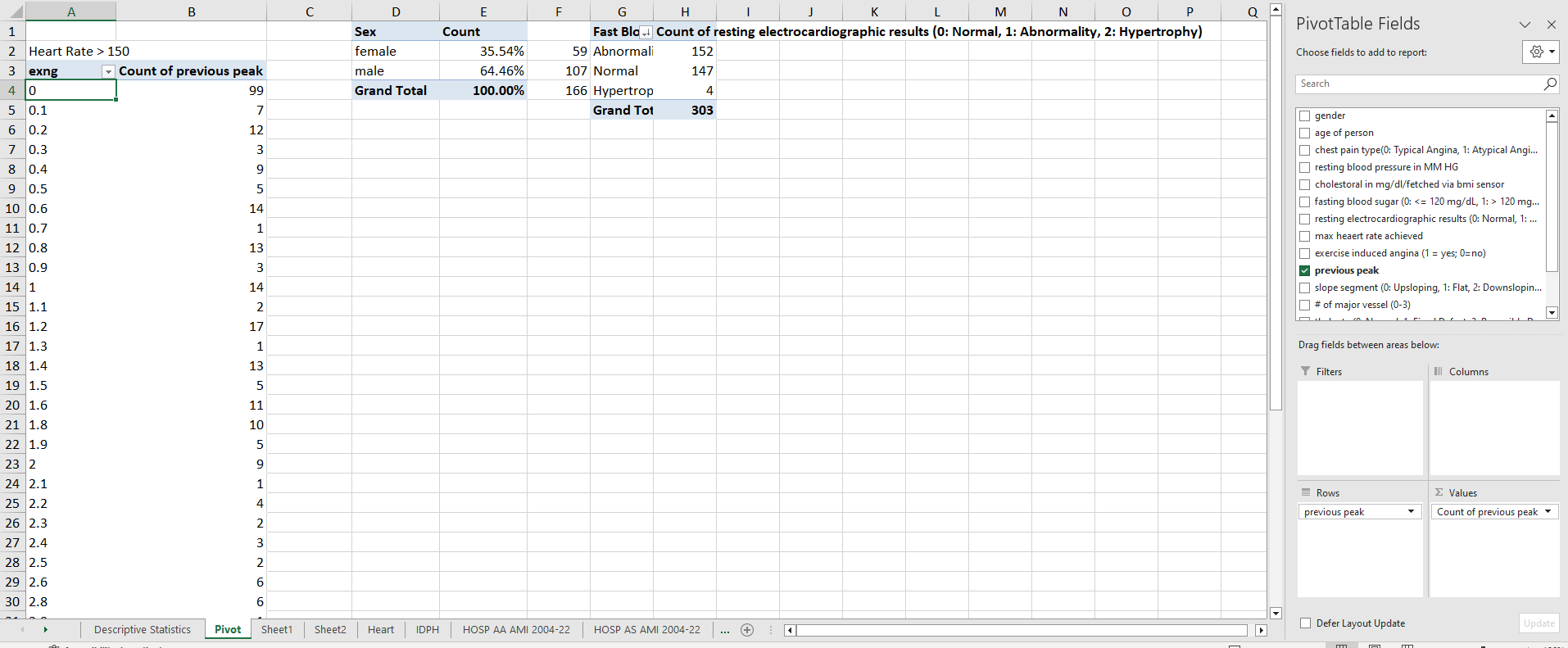
Information systems allow us to collect, store, display, and process information. In our analysis, the data has been stored in Microsoft Excel, Python, SQL, Weka, Microsoft Access, and PowerPoint.

Microsoft Access can be used to store and manage data.

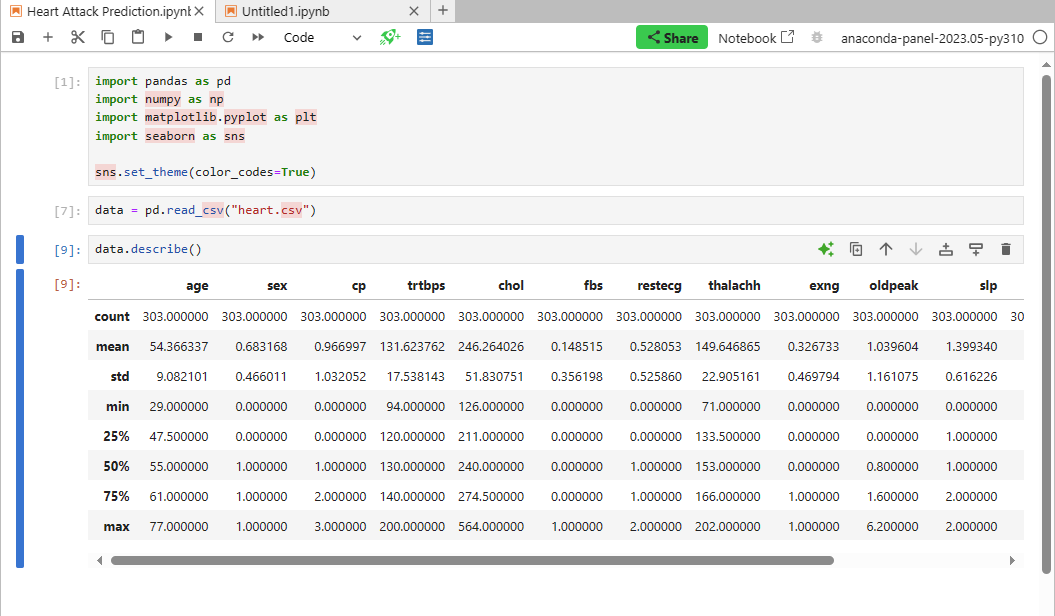


Microsoft PowerPoint to create presentations.

Microsoft Excel for data mining, organizing, and calculating data.



Python/Anaconda for computing, data science, and machine learning.



**D. Decision Science / Business Statistics**

Descriptive statistics were used to gather quantitative measurements to analyze each of the 14 attributes and summarize their values. The measurements include the distribution, central tendency, and variability or dispersion.

**Age** is a numeric value with a range of 48. The largest record is 77 and the smallest is 29 (both are males). The mode is 58 with 19 records of which 13 (68%) were males and 6 (32%) were females. The standard deviation is 9.082, which comes from the large range and low number of records. Skewness is -0.202, which makes sense with the youngest record set at 29.

**Sex** is a categorical attribute and has two values: male and female. With a frequency distribution of 207 males & 96 females. The 303 samples have a relative frequency of .68 males with .32 being females. This can also be shown as percentage with 68% being male and 32% being female.

**Chest pains (CP)** are categorical values with 4 different values: Typical Angina, Atypical Angina, Non-anginal Pain, and Asymptomatic. The values can be counted, ranked, and put into a percentage.

**Resting Blood Pressure(trestbps)** was measured upon admission into the hospital which had a range of 106 with the highest at 200 mmHg and lowest at 94mm Hg. The most common (mode) being 120 with 37 incidents with 73% of those incidents being males. The average resting blood pressure at the time of admission was 131 mm Hg which according to the CDC is “at risk”. The normal range is less than 120 mm Hg. Most readings were skewed to the right with a skewness of 0.714. The data falls within the normal distribution having a kurtosis at 0.930.

**Cholesterol (Chol)** > 200mg puts you at a higher risk for a heart attack. In our dataset, the average cholesterol level was 246.26 with the highest number at 564 (driving skewness of 1.143) which is more than twice the number that puts you at higher risk. A range in level from as low as 126 to 564. Frequency Distribution 100-199 count 50, 200-299 count 209, 300-399 count 40, 400 to 499 count 3 and >500 is 1.

**Fasting Blood Sugar (fbs)** is categorical and set with a 0: <= 120 mg/dL, 1: > 120 mg/dL. Used to gauge diabetes, hyperthyroidism, and pancreatitis. Out of the data sample 258 incidents had a fasting blood sugar under 120 mg/dl which represents 85% of the dataset. Such a low sugar level is an indication of hypoglycemia which is also known as diabetes. This piece of information is critical because during acute hypoglycemia heart rate and systolic blood pressure increase and both can lead to heart attack. Further noted that out of the 85% of hypoglycemia, the male subject is 67% of 258 incidents.

**Resting Electrocardiographic (restECG)** is categorical with three types: Normal, Abnormality (deviation of the typical pattern), and Hypertrophy (thickening of the heart’s left ventricle wall). The test is used to measure activity of your heart during a resting phase to determine rhythm enlargement, and heart attack. The frequency distribution is Abnormality - 152, Normal - 147 & Hypertrophy - 3.

**Max Heart Rate (thalachh)** ranges from largest at 202 to smallest at 71. The average max heart rate in our sample is 149 with a standard deviation of 22.90. The data has a skewness of -0.54 with a mode at 162. 85% of the heart rates > 150 had <=120mg fasting blood sugar level.

**Exercise Induced Angina (exng)** is a chest pain caused by exercise. The values were categorical with 0=No & 1=Yes. 67% of patients experienced no chest pains (0) while 33% experienced pains (1).

**Previous Peak (oldpeak)** represents the ST depression induced by exercise relative to rest in a heart stress test. The largest peak is 6.2 with the smallest at 0. The mean is 1.04 with a standard deviation of 1.16. Skewness at 1.269 with 0 being the most common mode of 99 samples.

**Slope of Peak Exercise ST Segment (SP)** iscategorical using 0: Upsloping, 1: Flat, and 2: Downsloping. The distribution of the data: Downsloping 142 (47%), Flat 140 (46%), and Upsloping 21 (7%). Downsloping is a sign of an unhealthy heart.

**Number of major vessels colored by fluoroscopy(caa)** can range from 0 to 4 in our sample. The frequency distribution 0(102), 1 (65), 2(76), 3(60) & 4(20). The mode is 0 with an average of .79. The higher values suggest better blood flow while lower values put you at risk for heart disease.

**Thal rate (thall)** is a categorical value ranging from 0 to 4. The frequency distribution is 0(2), 1(18), 2(166) & 3 (117). The mode is 2 with an average of 2.31.

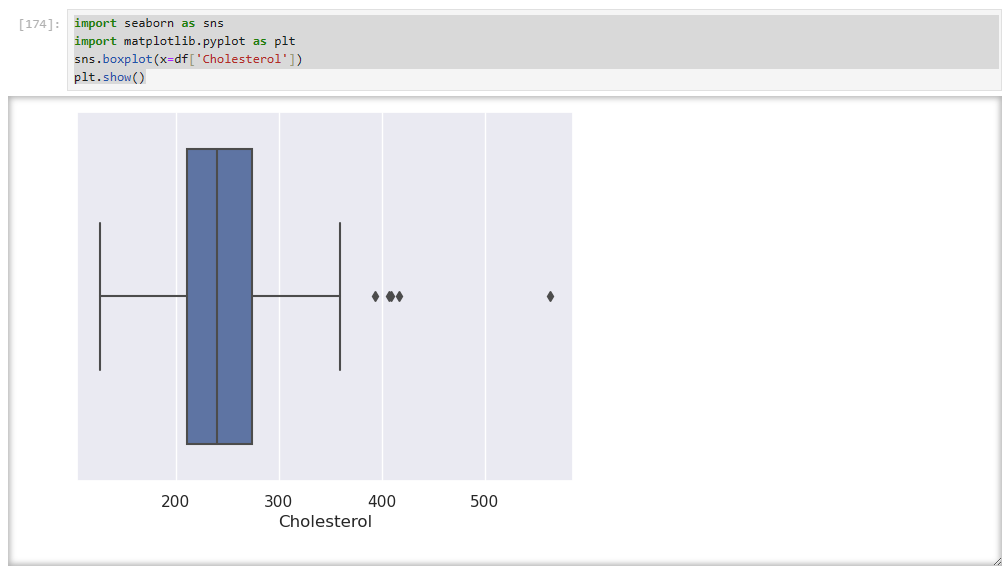
**Target Variable (output)** is categorical at 0 or 1. Distribution of 0 - 138 (46%) and 1-165(54%). The mode is a 1.

**E. Dating Mining Tools**

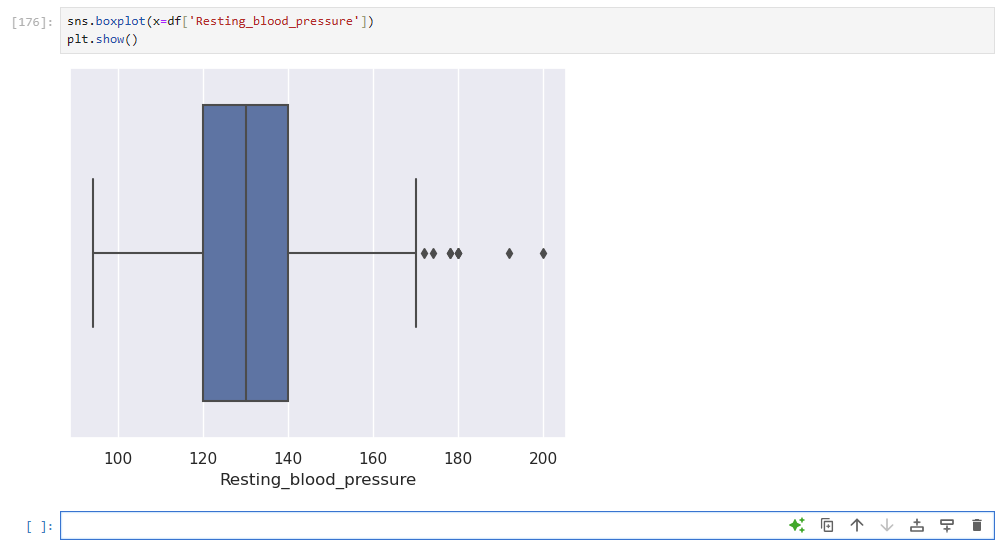
Data mining tools allow you to easily discover valuable insights into large amounts of data. Throughout our time at Lewis, we learned how to use Excel, Python, and Weka to mine data.

Using a Boxplot to determine the quartile distribution, average, and outliers.

Cholesterol has an interquartile 211 to 275 with an average of 246. The sample includes 4 outliers in their data and are 394, 407, 409 & 564



Resting blood pressure has interquartile 120 to 140 with an average of 132. The sample includes 6 outliers in their data and are 172, 174, 178, 180, 192 & 200.



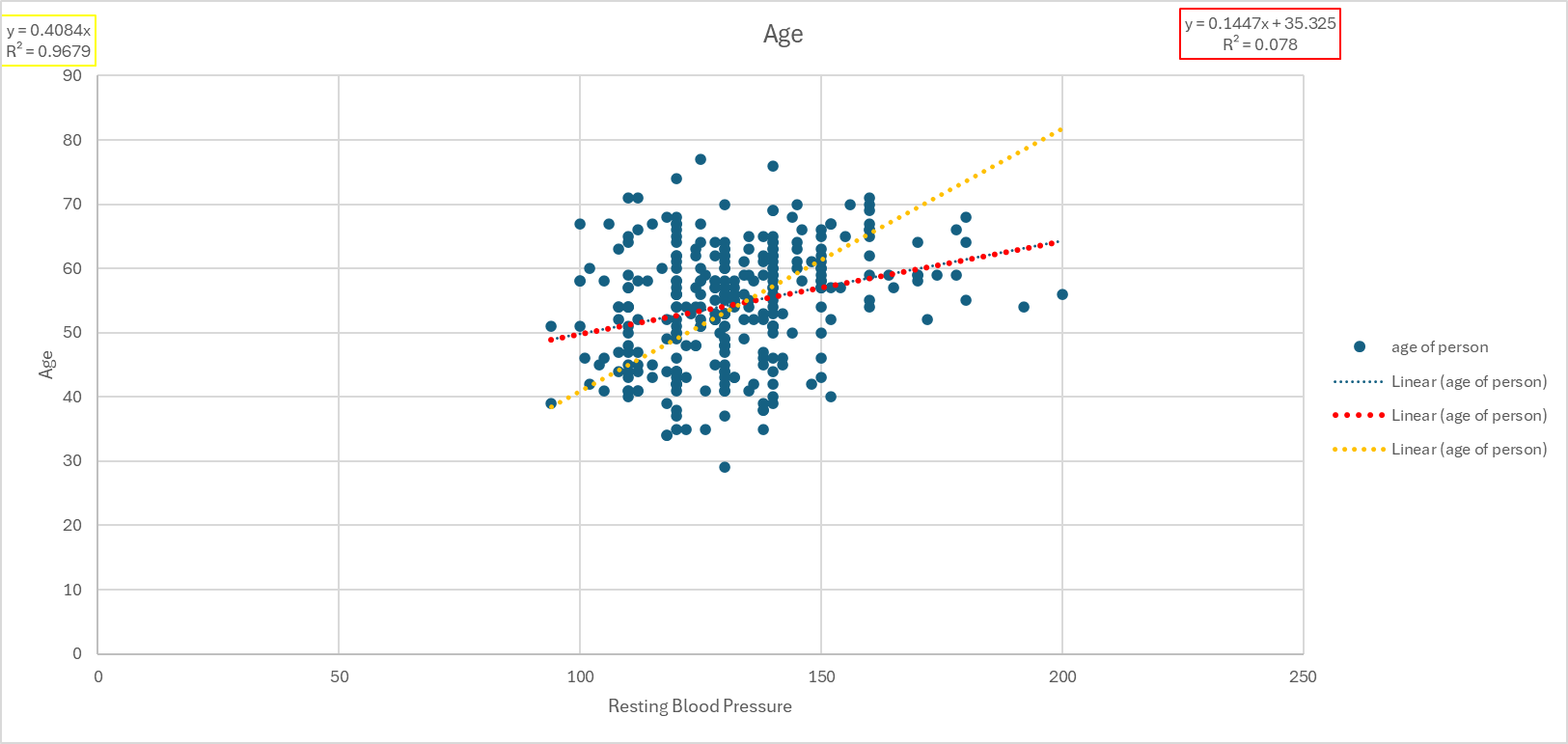
Excel -scatter plot with linear regression and Rsquared value

**Age and Resting Blood Pressure:**

•Positive linear relationship

•R2 shows high relations and variance at 97% when Intercept is set to 0

•Age shows a direct correlation with higher resting blood pressure.

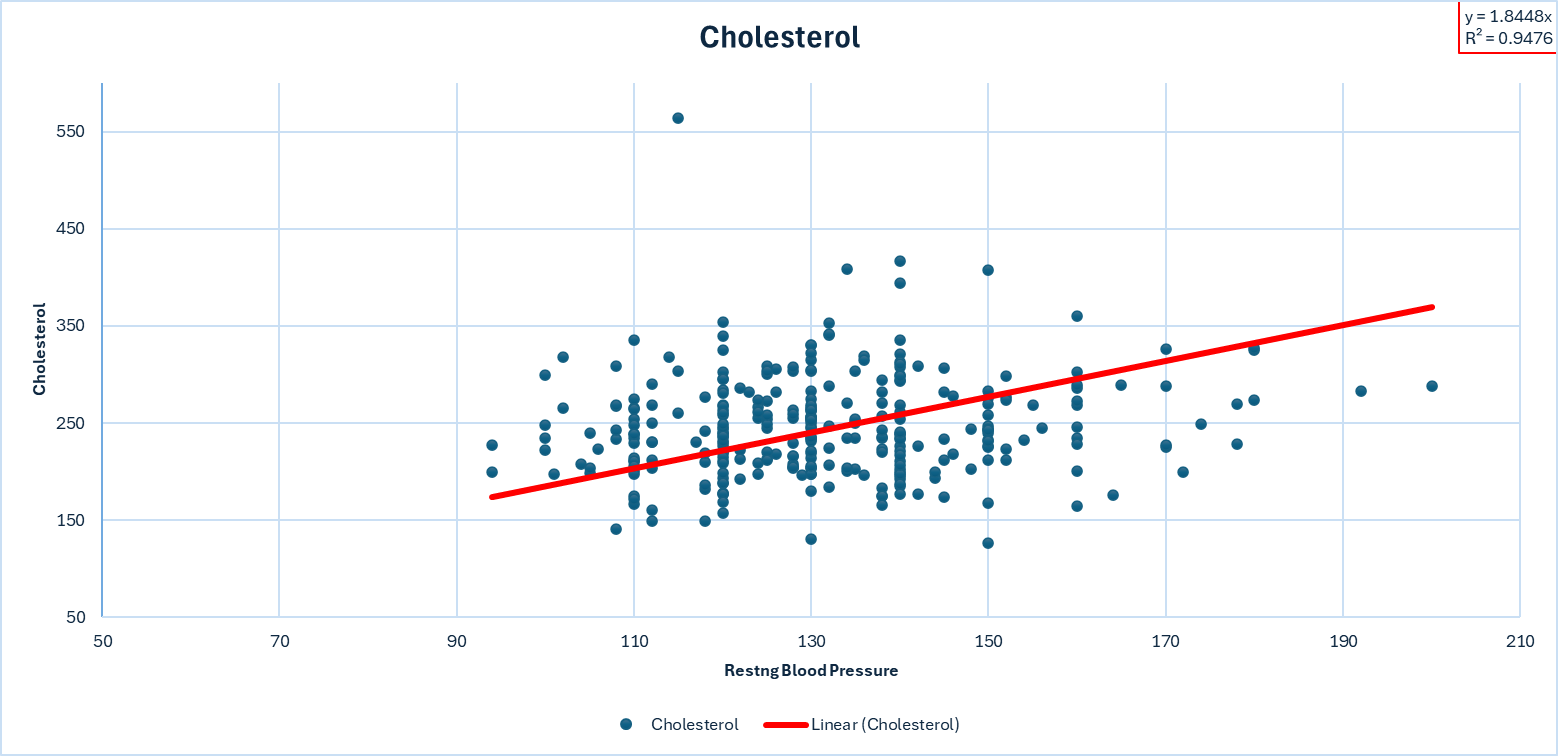


**Cholesterol and Blood Pressure:**

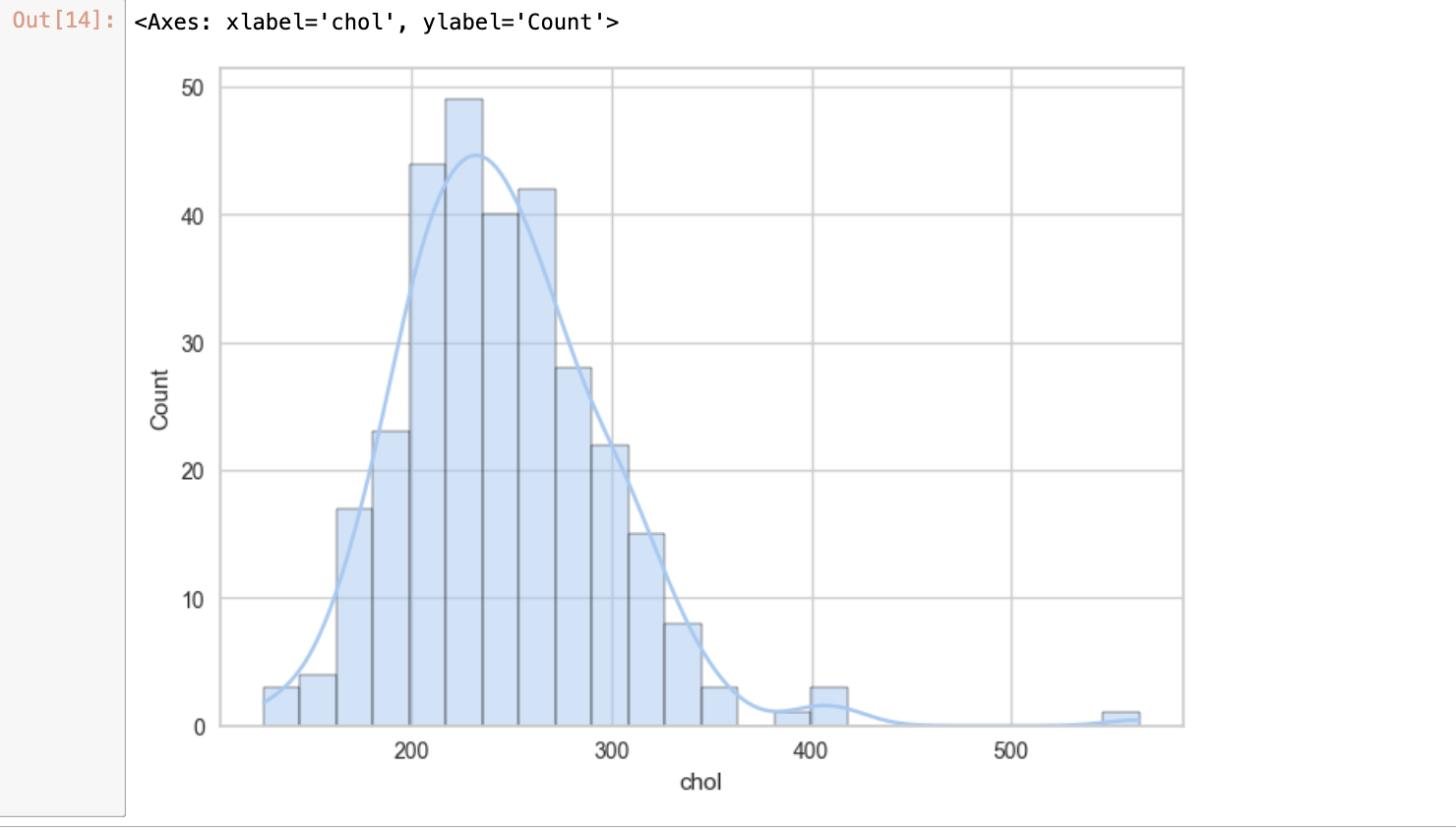
•Positive linear relationship

•R2 shows high relations and variance at 95% when Intercept is set to 0

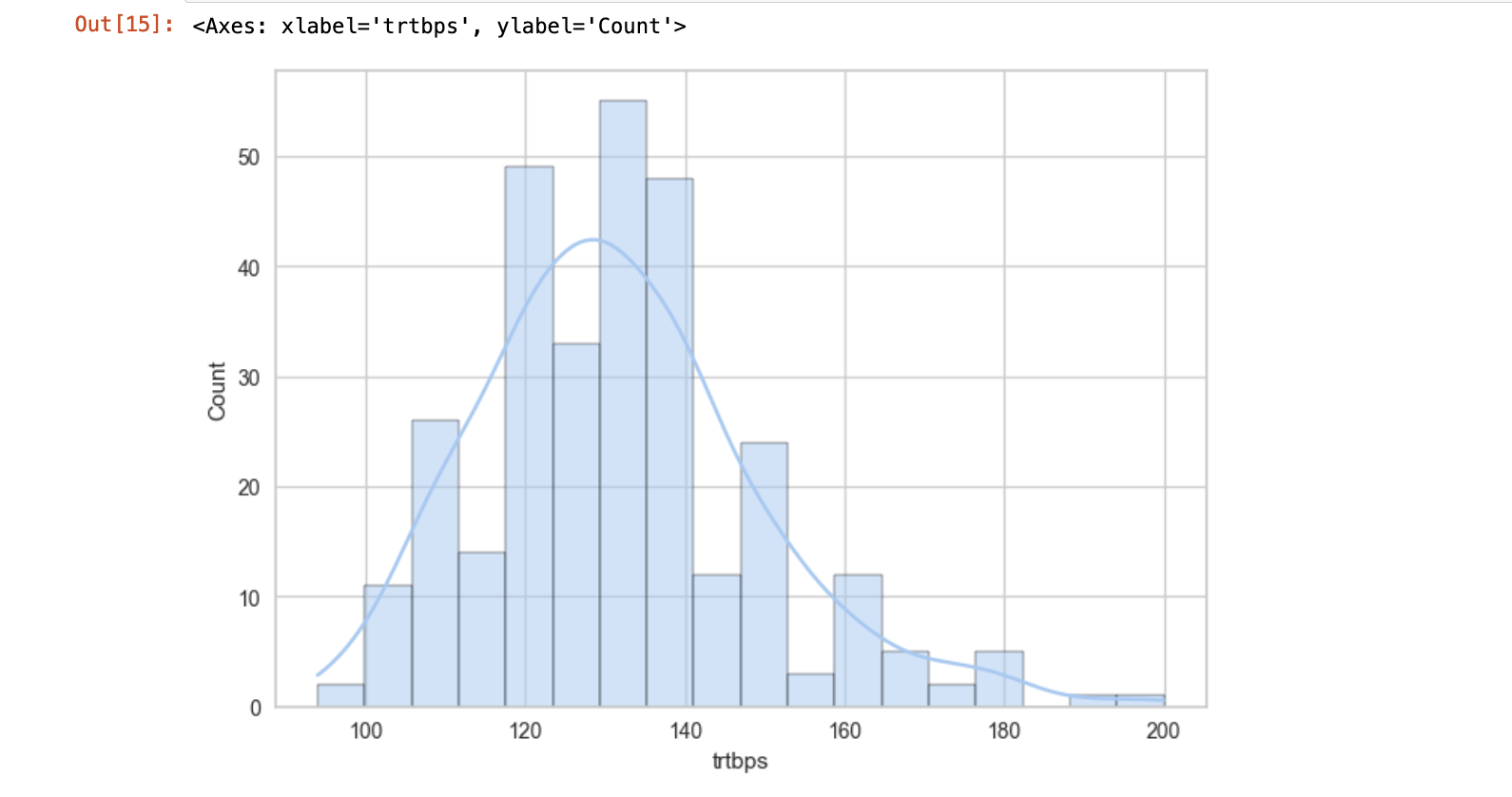
•Age shows a direct correlation with higher resting blood pressure.



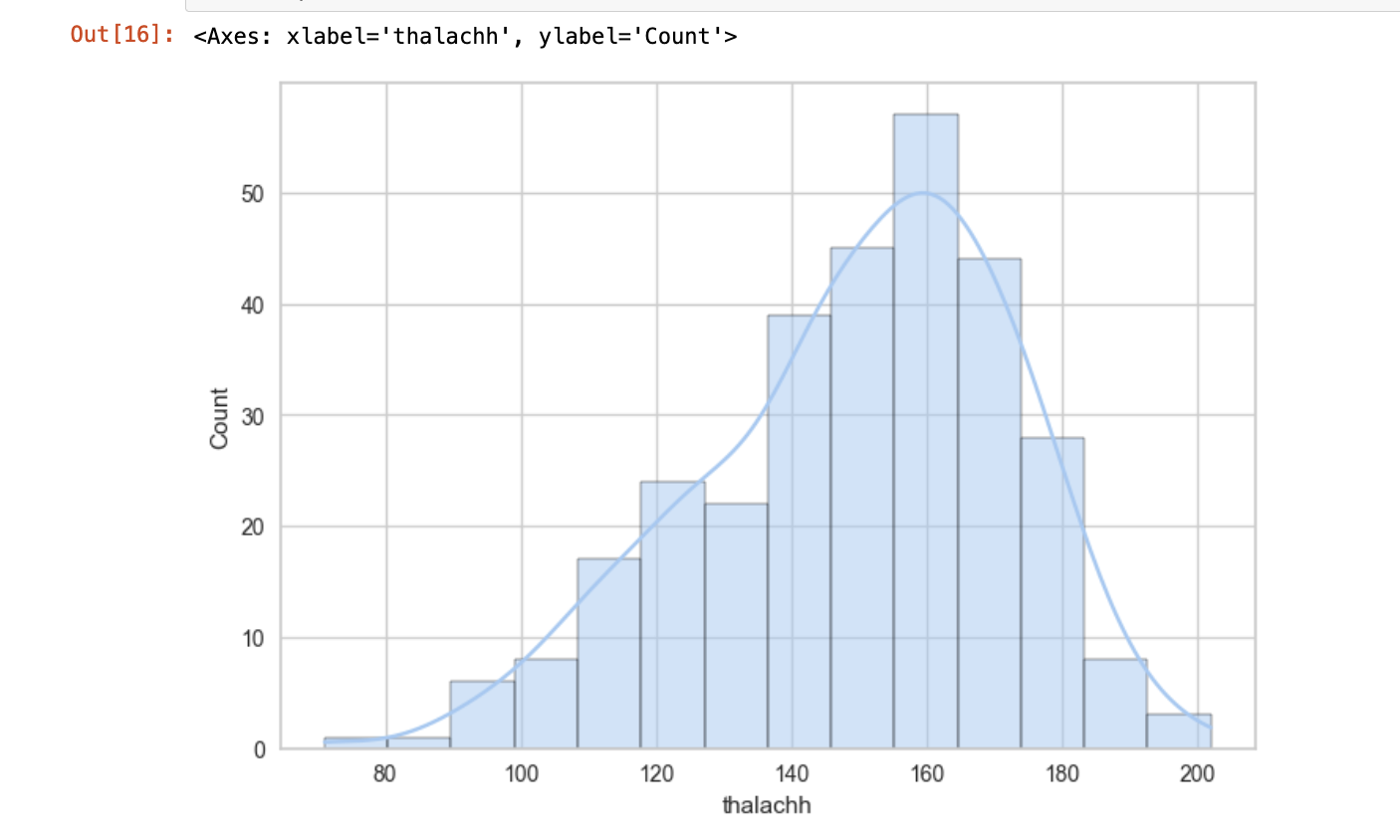
Bell Curve demonstrating the normal distribution of cholesterol.



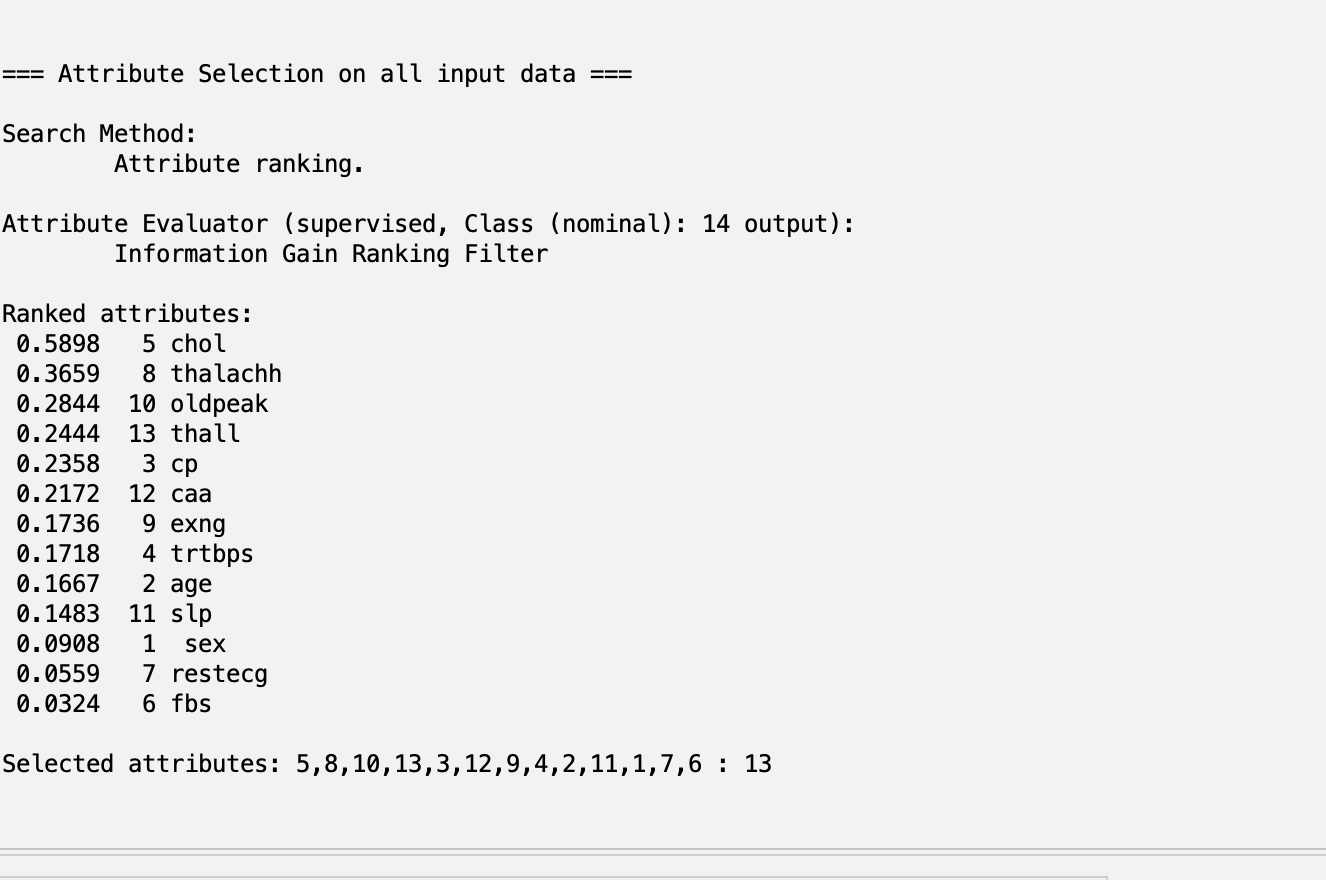
Bell Curve demonstrating the normal distribution of trtbps.



Bell Curve demonstrating the normal distribution of thalachh.



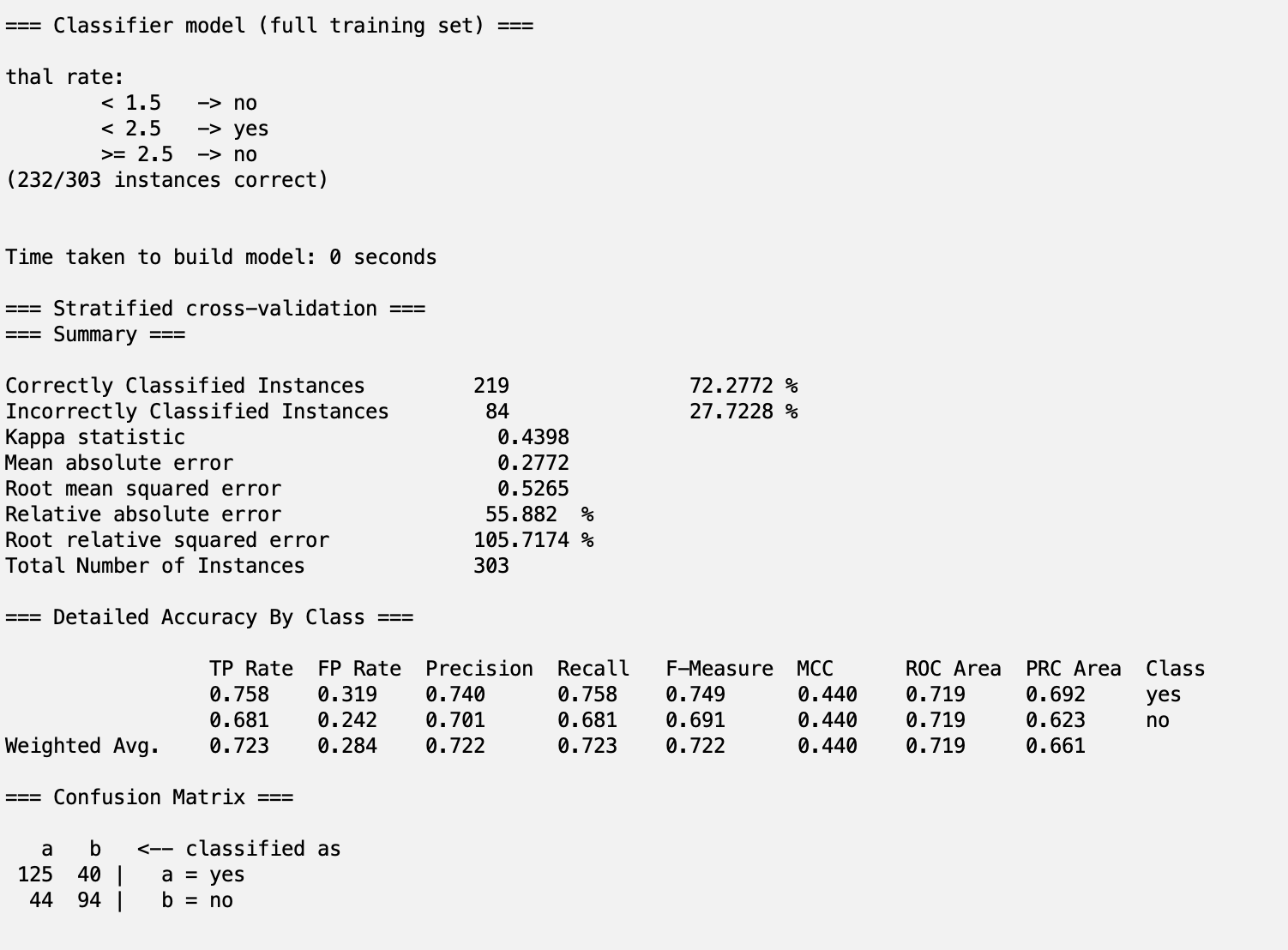
**F. Business Intelligence**

****

1. **OneR:**

Attribute chose thal rate since it had the lowest error,

Accuracy value: 72.3%



2.  **Decision Tree J48:**

a) pruned

- Root node: chest pain type

- Number of leaves:26

- Size of tree: 51

- Accuracy: 78.5%

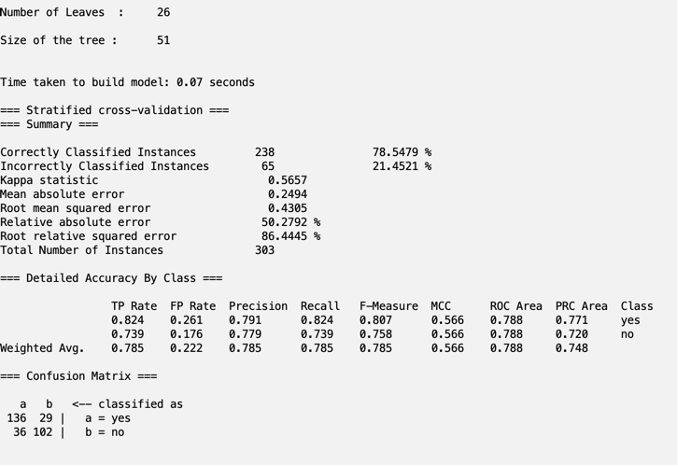


b) unpruned

-Number of leaves: 35

- Size of tree:61

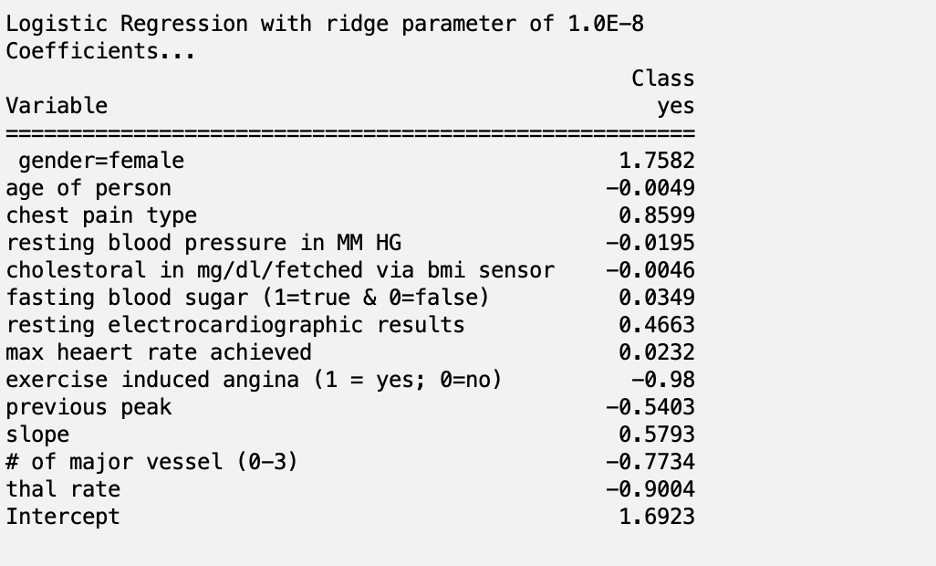
- Accuracy:77.9%



3. **Logistic regression**:

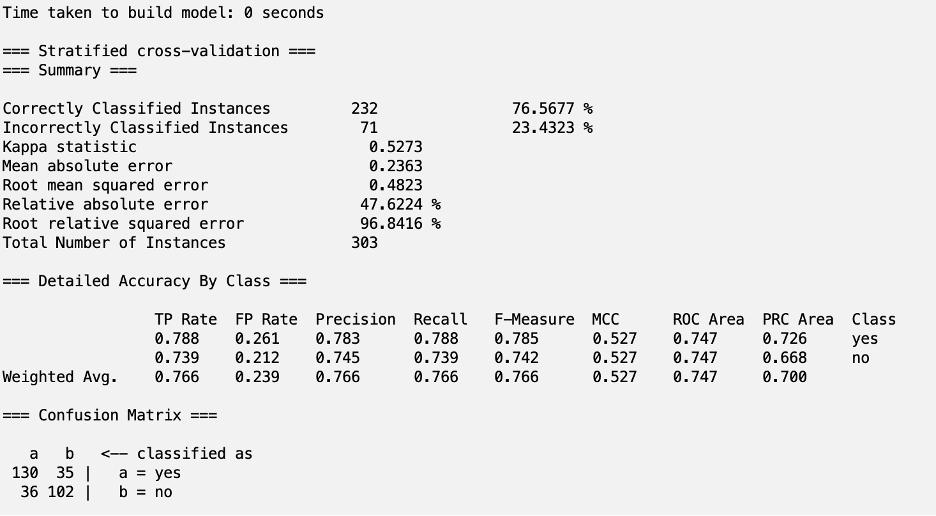
Accuracy: 82.2%

The attributes having positive relationship with the variable which is “yes” are chest pain type, resting electrocardiographic results, ma heart rate achieved, and slope.



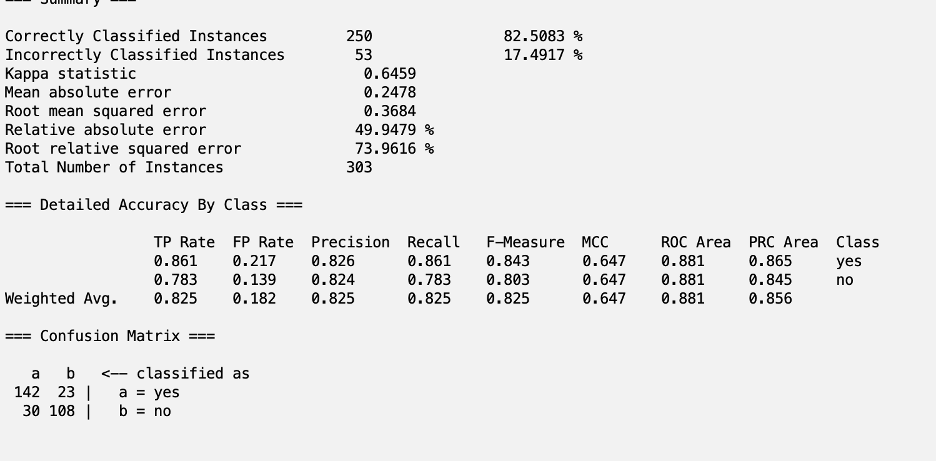
4.  **Instance based learning:**

Accuracy: 76.6%



After changing K value to 9

Accuracy: 82.5%



**A. Cost sensitivity scenario:**

From a hospital perspective, predicting the outcome of potential heart attacks in patients. The hospital wants to minimize the number of heart attacks by being able to predict them or potentially predict them and take steps to prevent them.

**B. Output variables:**

Yes – positive.

No – negative.

**C. Confusion Matrix:**

TP- Predicted positive (correctly predicted heart attack)

TN- Predicted negative (correctly predicted no heart attack)

FP- Predicted positive (predicted heart attack, in reality no heart attack)

FN- Predicted negative (predicted no heart attack, in reality there was heart attack)

**D. Interpretation:**

For this scenario, the highest cost would be FN- predicted no heart attack, in reality that was a wrong prediction. It will cost the reputation of the hospital as well as the doctor for the misdiagnosis, and potentially could cost the life of the patient which is the highest cost possible.

5.  **Based on Instance-based learning with a K value = 9**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted class | |
| A: yes – heart attack | B: no- no heart attack |
| Actual  class | A: positive | TP- 142 | FN-23 |
| B: negative | FP-30 | TN-108 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1. | Classifier 1: Logistic | | | Classifier 2 : J48 | | |
|  | Default | weight =2 | weight =3 | Default | weight =2 | weight =3 |
| Accuracy | 82.2% | 82.8 | 81.5 | 78.6% | 81.8% | 80.9% |
| Precision | 0.810 | 0.787 | 0.766 | 0.791 | 0.789 | 0.764 |
| Recall | 0.879 | 0.939 | 0.952 | 0.824 | 0.909 | 0.939 |
| F-Measure | 0.843 | 0.856 | 0.849 | 0.807 | 0.845 | 0.842 |
| AUC | 0.899 | 0.899 | 0.899 | 0.788 | 0.829 | 0.819 |
| # of FN | 20 | 10 | 8 | 29 | 15 | 10 |
| # of FP | 34 | 42 | 48 | 36 | 40 | 48 |

The results, from the Cost Sensitive Analysis with the Logistic Classifier set with the weight set to 3, are the best. When we raise the weight, the accuracy drops 1%, but the number of FN gets down from 20 to 8. The accuracy being over 80% is still satisfactory, and the Recall value is almost at 1 (which is the ideal value).

**G. Business Process Automation**

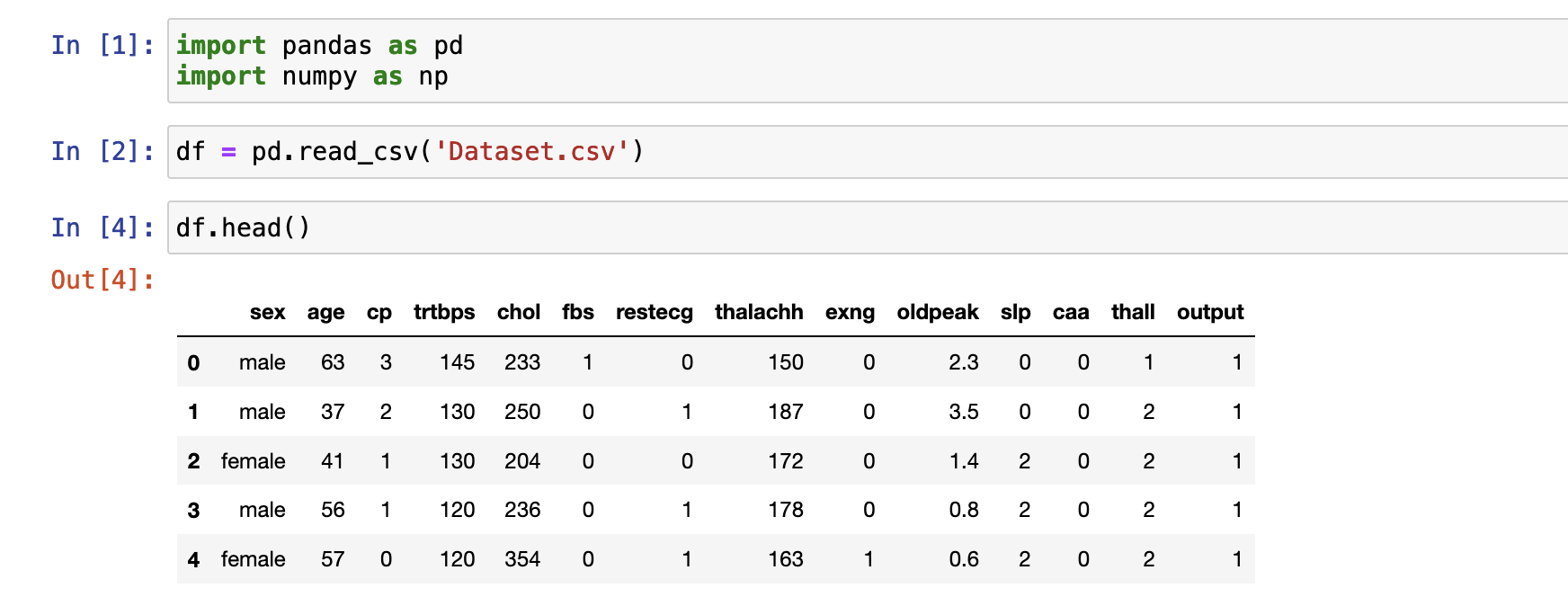
* Identifying the areas in need of improvement or automation:

We can find a few areas for improvement of the process of detecting possible heart attacks. For example, the implementation of utilizing technology for heart health or more frequent measurements if a certain variable seems above average.

Business Process Automation also can improve the time of analysis it takes for the warning signs within patients’ charts to be noticed and acted on. The early signs of the issue with the patient's statistics might not be clear, so putting systems in place to give us a warning to bring our attention to the problem might help prevent heart disease from happening. It is all time sensitive data when it comes to preventing the complications of the bad tests outcomes. The more patients we want to analyze the more useful BPA becomes.

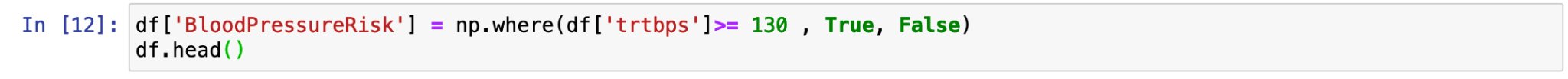
* In this scenario we are looking to implement some BPA into the hospital, or doctor's office to improve heart disease prediction and prevention.

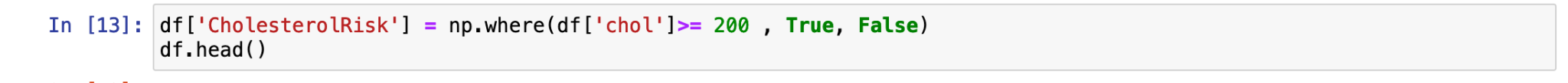
The benefits of this would be not only to the hospital but also to the patients. It could improve productivity as well as efficiency in the process of collecting and analyzing data. To see the outliers within it faster and react almost immediately.



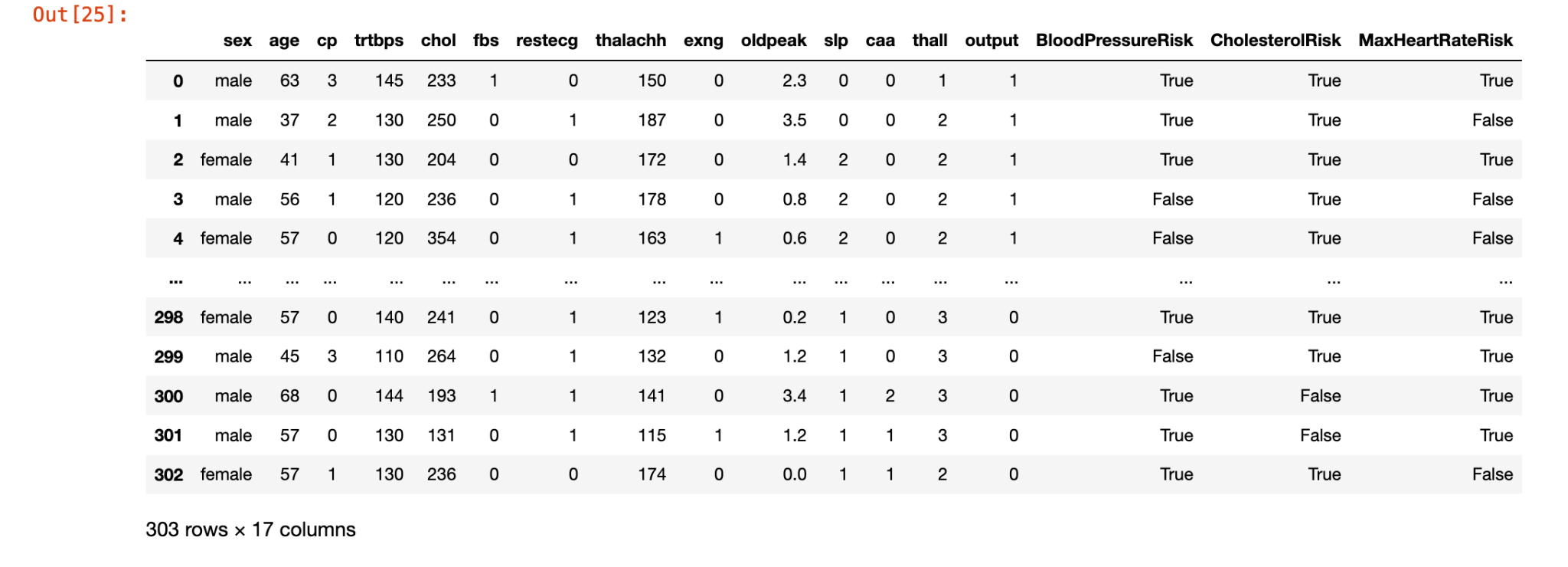
A screenshot of a computer

Description automatically generated



****

****



With the normal value for Blood pressure (trtbps) being under 130, anything above puts people at risk. The condition is set to anything above 130 being True - at risk, anything below False - not at risk. For Cholesterol (chol) the normal values are below 200 so anything above is True - at risk, and anything below False - not at risk. Similarly, for the Maximum heart rate achieved (thalachh) the normal value is below 220-’age of a person, so anything above that will be True - at risk, anything below is normal which will read False. We can see the outcome in the new columns.

With this information, we can quickly see who is at risk from the most important attributes in our data set and act accordingly. When we implement a program like this into a hospital database to detect those risks immediately, we can improve efficiency and minimize human errors.

**H. Introduction to Business Analytics**

**Descriptive Analytics**

The total number of instances is 303 with the sex categorical attribute being split into a majority of 0.68 men and 0.38 women and the age as a numerical value range from 29 years old to 77. Other attributes include Resting Blood pressure with a standard deviation of 17.5, Cholesterol with a standard deviation of 51.8, and max heart rate achieved with a standard deviation of 22.9.

The chest pain attribute has 4 values: Typical Angina, Atypical Angina, Non-anginal Pain, and Asymptomatic with the typical angina being substantially more frequent compared to the other three. When compared with Exercise induced angina, we can see that twice as many instances are a no (204) and only 99 at yes. The fasting blood sugar attribute shows that 258 instances are at <= 120 mg/dL and only 43 at > 120 mg/dL. For the resting electrocardiographic with three values 0- normal, 1- Abnormality, and 2- Hypertrophy, there are only 4 instances of Hypertrophy, 152 Abnormality, and 147 normal results. The resting blood pressure highest frequency value is at 120 with the 37 of the instances.

Visualization of the data gives us an idea of the tendencies within attributes we have (the red being men and blue women). Just by looking at all attributes we can draw some conclusions about what factors might have a higher impact on the probability of suffering from heart attack. There are some straightforward trends within the data like trtbps (resting blood pressure), thalachh (max heart rate), and chol (cholesterol) are positively skewed. The age distribution has a slight positive skewness but overall, it is a multimodal distribution which is not expected as age does not seem to have a big impact on heart attack.

A screenshot of a graph

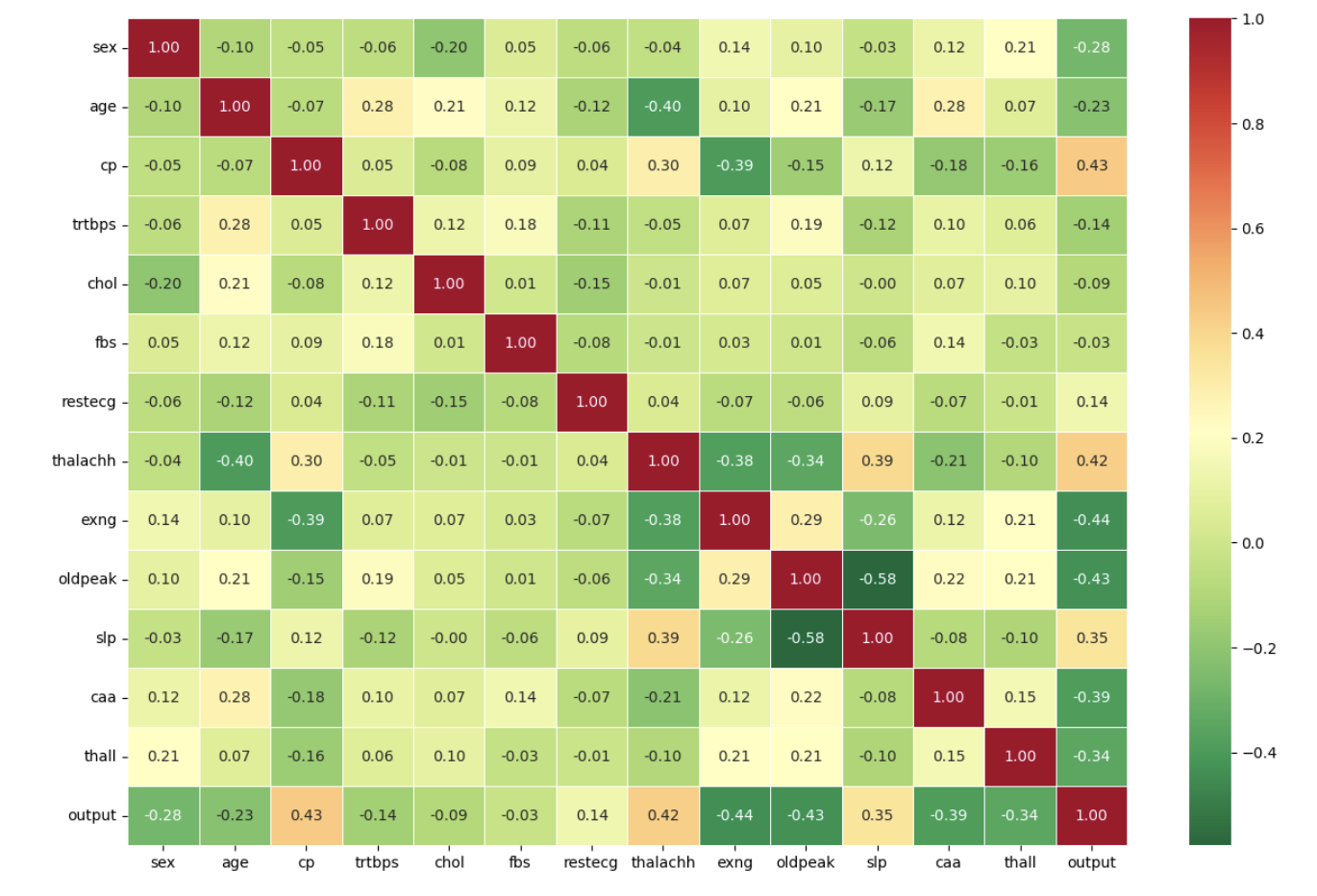
Description automatically generated

**Diagnostic Analytics**

The correlation heat map shows us the different attributes within the data set. With 1 being a perfect positive correlation and -1 perfect negative, we can see from this the positive and negative correlations like the positive correlation between the CP (chest pain) and Thalach(maximum heart rate) at 0.43, and Thalach(maximum heart rate) and slope at 0.39. The CP and output at 0.43, and the thalach and output at 0.42. Age has a negative correlation with thalach -0.39, and sex negative one with output -0.23. Other negative correlations are CP with exang (exercise-induced angina) at -0.39, thalach and exang at -0.38, thalach and oldpeak at -0.34. With the highest negative correlation between the oldpeak and slope at -0.58.

Looking at the data we can see how cholesterol and Maximum rate achieved has the biggest impact on the output. With higher cholesterol, we can see an increase in heart attack occurrence. For the Maximum heart rate achieved positive skewness, with 162 being the highest frequency value.

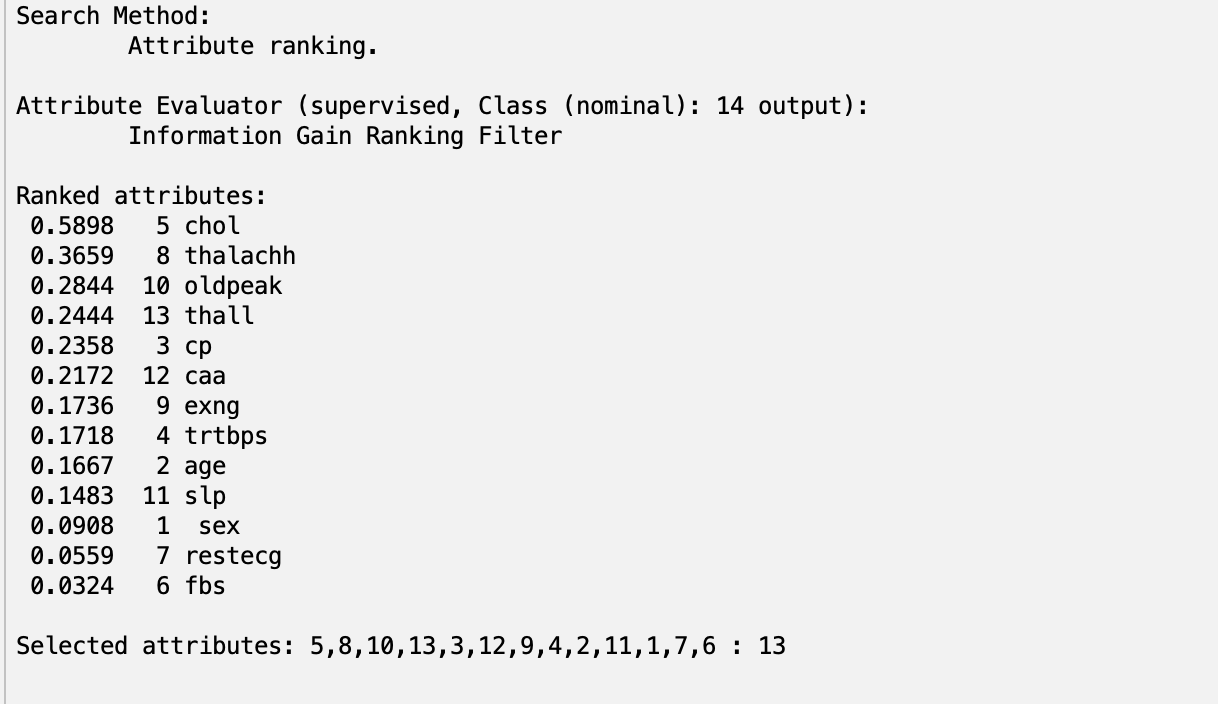




**Predictive Analytics**

After putting the data through the information gain attribute evaluation, we can see how chol- Cholesterol attribute is the most relevant to predicting heart attack at 0.5898, followed by thalachh- the maximum heart rate achieved by the patient at 0.3659.

We can expect from this that any fluctuation in the cholesterol value will have an impact on the output attribute. Similarly with the maximum heart rate, will be an important attribute in predicting the outcome. From this dataset, we can conclude that people with a higher max heart rate have a higher risk of heart attack, and people with lower oldpeak also have a higher chance of heart attack. As well as people with non-anginal chest pain are at higher risk. There is also a significant difference in the number of men with a higher risk of heart attacks compared to women.



**Prescriptive Analytics**

After analyzing the dataset and understanding the impact of various factors on heart disease, prescriptive analytics aims to identify strategies to reduce the risk of heart attack. Our strategies implement lifestyle modifications by individuals, promotion of community health programs, and utilization of technology to provide ongoing health education and monitoring.

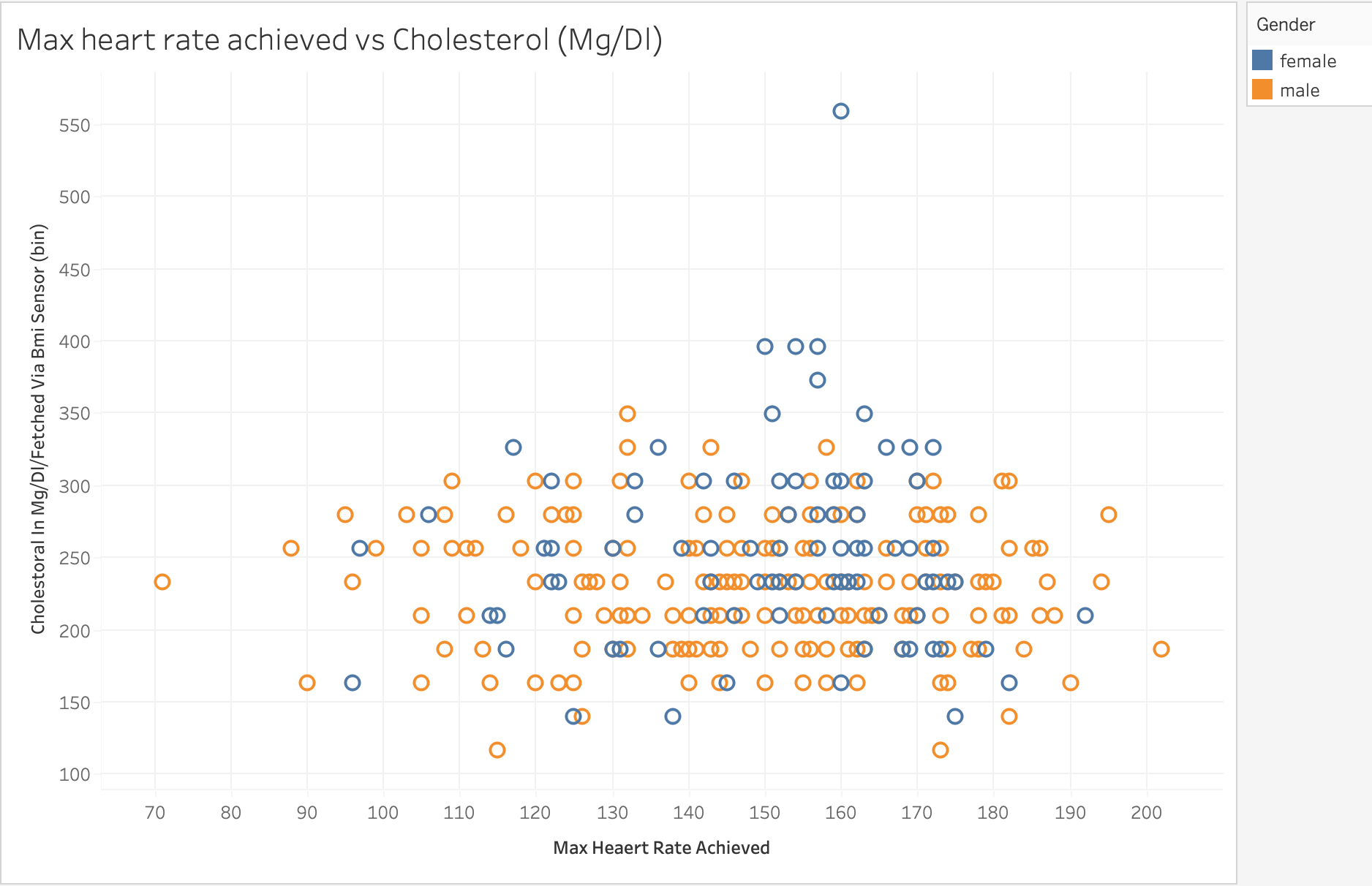
**Strategic Recommendations and Implementation**

* Cholesterol Control:
* Recommend dietary changes emphasizing fruits, vegetables, whole grains, and lean proteins while moderating intake of saturated and trans fats.
* Encourage regular blood testing to monitor cholesterol levels and prompt lifestyle adjustments as necessary.
* Launch public health campaigns to raise awareness about the risks of elevated cholesterol and promote healthy dietary habits.
* Maximizing Heart Rate Through Exercise:
* Develop personalized exercise plans based on individuals’ health status, focusing on cardiovascular activities like running, walking, cycling, and swimming to enhance heart rate efficiency.
* Establish community fitness programs offering diverse physical activities to encourage regular participation across populations.
* Screening and Preventative Measures:
* Implement targeted screening programs for early detection of heart disease among high-risk demographics identified through descriptive analytics.
* Launch educational initiatives to increase public awareness of heart attack risk factors and preventive strategies.
* Utilizing Technology for Heart Health:
* Introduce wearable technology capable of monitoring key heart health indicators such as blood pressure, oxygen saturation, and ECG readings.
* Deploy mobile apps and online platforms offering personalized diet and exercise tracking, mindfulness resources, and health education materials.
* Strategies Evaluation and Continuous Improvement:
* Establish metrics and key performance indicators (KPIs) to assess the effectiveness of implemented strategies, such as reductions in cholesterol levels and increases in physical activity rates.
* Emphasize the importance of ongoing evaluation and adaptation of strategies based on real-world outcomes and feedback to ensure efficiency.

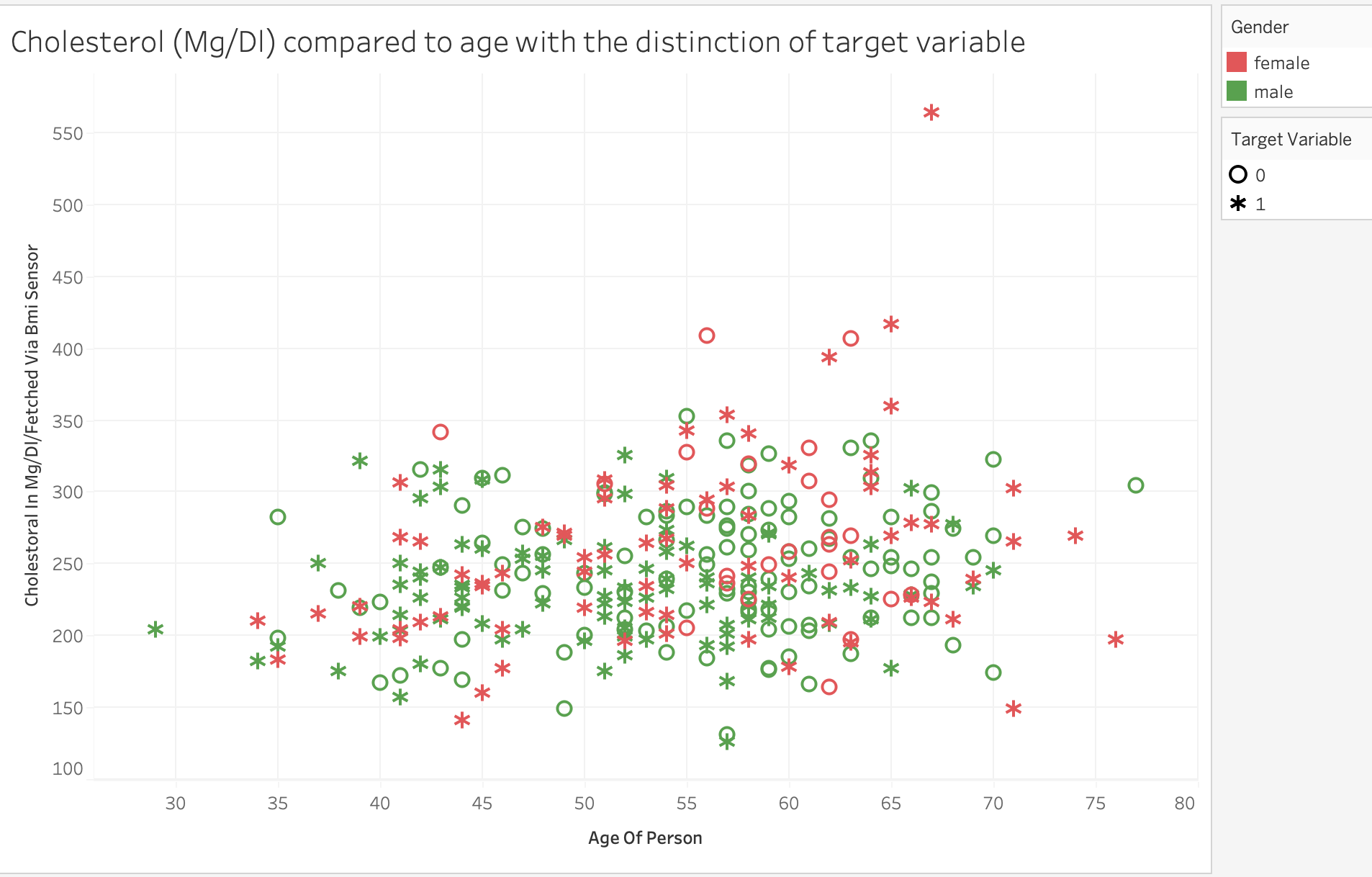
By integrating these recommendations into healthcare systems and community initiatives, we can work towards reducing the incidence and impact of heart attacks, ultimately leading to improved health outcomes and enhanced quality of life.

**I. Business Data Visualization**

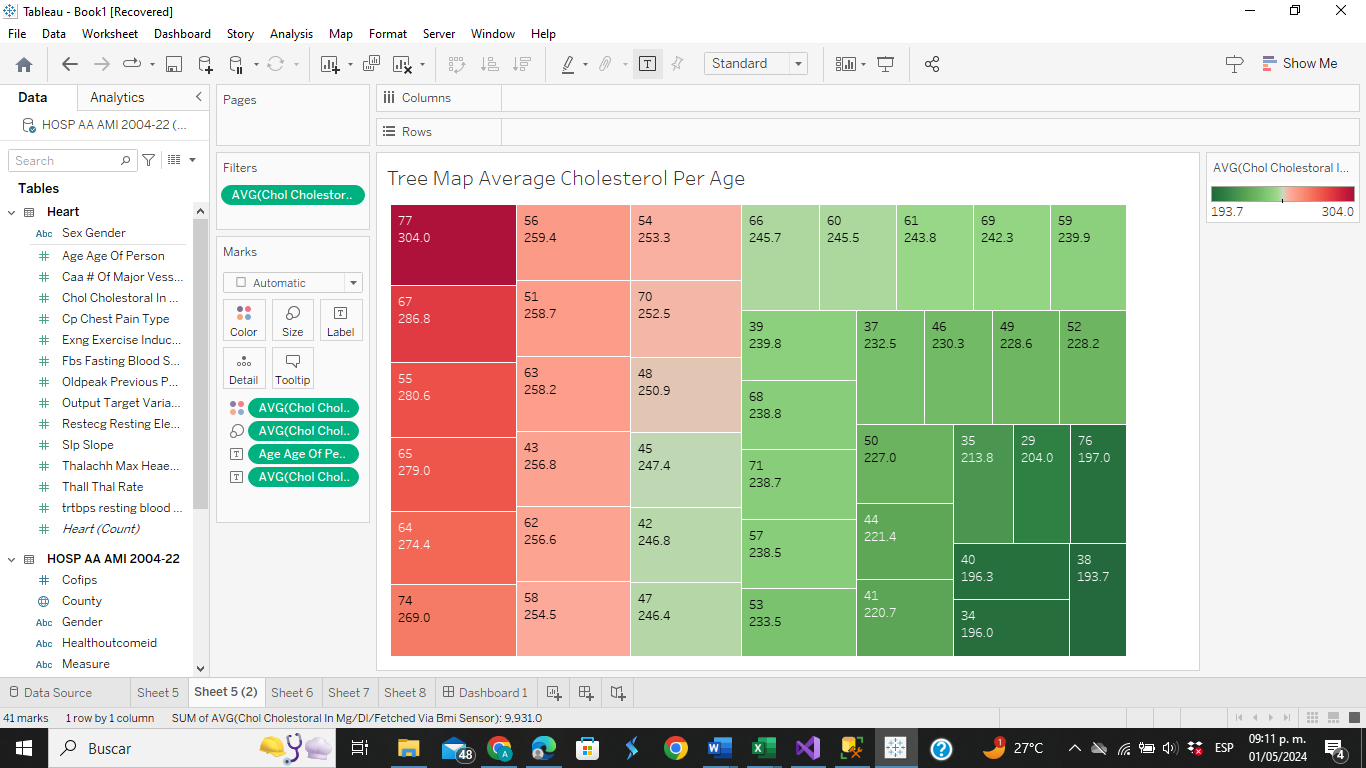
**Tableau:**

****

In the visualization above we compare the X-Maximum heart rate achieved and Y-Cholesterol(Mg/Dl). Organized by gender with color with blue being female and orange for male. It shows us the relationship between the two variables. Both are very important variables to finding out our output value and seeing the distribution of both being aligned however both are skewed into different directions thalachh-Max heart rate to the right and chol-Cholesterol to the left.

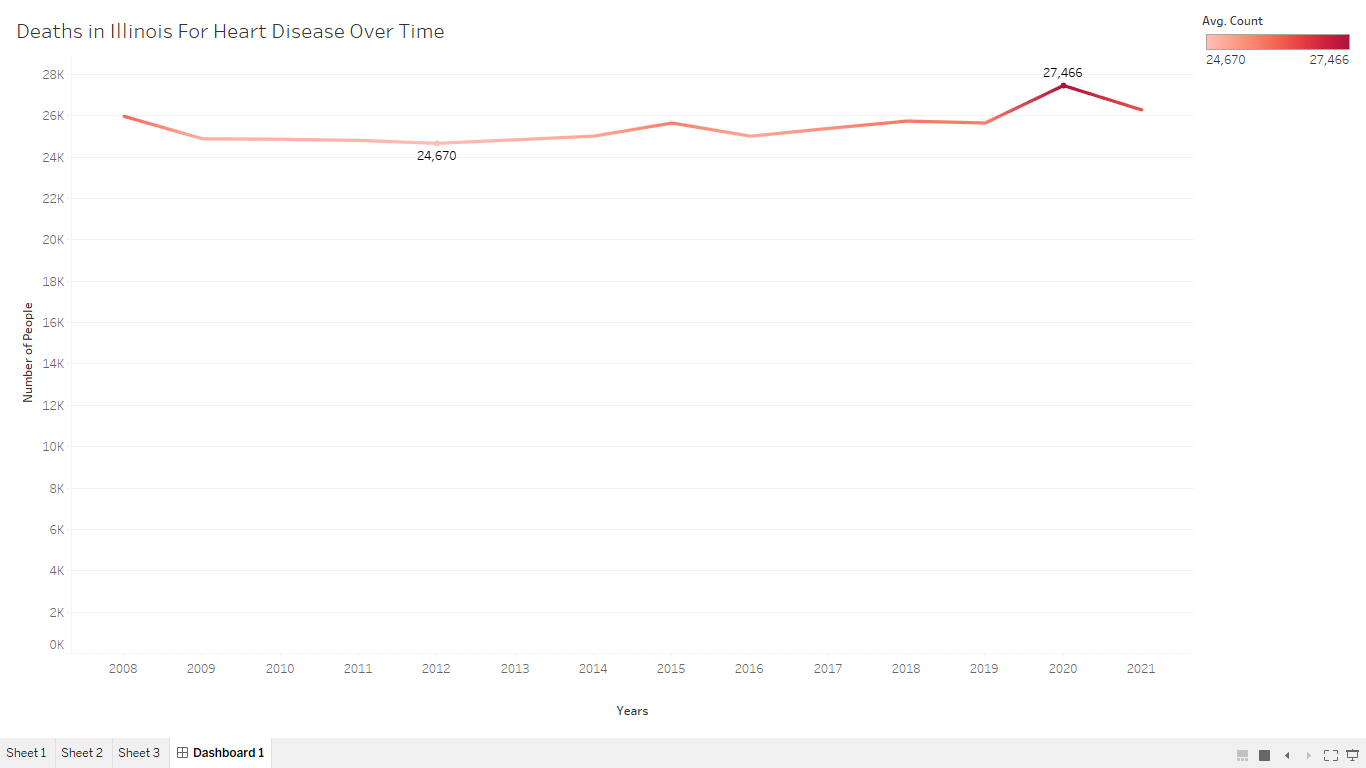
****

In the visualization above we see the cholesterol compared to the age. We can see a few outliers who are very young with low cholesterol and older with high cholesterol both suffering from heart attacks. However, you can also see the oldest in the aged person has not suffered from a heart attack. We can get interesting insights from data visualizations that might not be as easy to notice when looking at a spreadsheet form of data. For example, the straight line of all variables being positive with output regardless of gender.

****

The treemap visualization gives us a helpful overview of the differences in average cholesterol levels among various age groups. People's cholesterol labels tend to rise with age. The dark red color represents the average reading of 149.21, which is noted as highest at the age of 59. This increased cholesterol levels in older adults’ points to a higher risk of heart attack in this population.

On the other hand, lower cholesterol levels are also highlighted by the illustration; these are especially apparent in younger age groups, such as 34 and 41, where the darker green color corresponds to common levels of 118 and 119, respectively. Younger people's lower cholesterol level indicates a lesser chance of having a heart attack.

****

The time series visualization offers a detailed representation of the number of deaths attributed to heart disease in Illinois over a period of time. The data shows that, with 24,670 recorded deaths, 2012 had the lowest death toll. A closer look at the graph reveals that during 2009 and 2010, there was a slight decrease in deaths. Later, there were variations in the number of deaths until 2015. From 2015 to 2019, there was a noticeable increase in the amount of deaths, with 27,466 deaths reported in the next year, which is the highest number on the chart.

Analyzing this data is significant because it highlights the value of heart disease prevention strategies and the need for thorough record-keeping to successfully monitor patients' health and manage their medical histories. Heart disease, particularly heart attacks, is still one of the top causes of death in the US and other countries.

**J. Web Analytics**

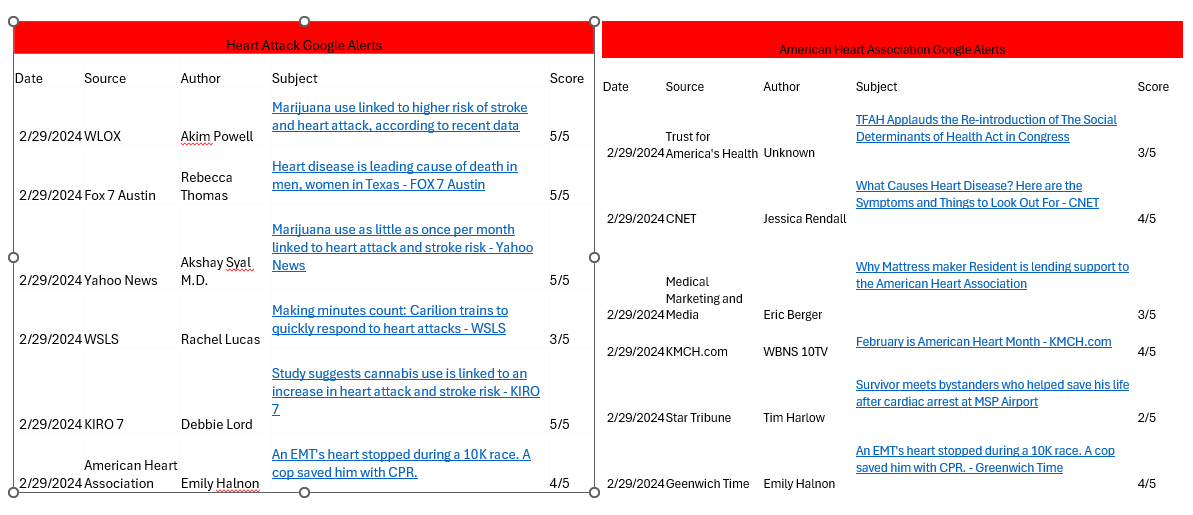
Web Analytics collects, analyzes, and reports data relative to user interactions. The collection can then be used for actionable information such as consumer preferences and driving decision-making.

To improve our results, we must continue to collect relative data and data which is up to date. Additional good data collected from the web will enhance and progress our model which will help make better informed prediction data decisions.

Infographics created using the project's graphs & the Mayo Clinic website help us quickly understand the signs, symptoms, and preventative care.



Google Alerts will allow you to stay notified of information or potential data which can be used to further advance heart attacks predictions.



Relative searches:

[New tool brings big changes to cardiovascular disease predictions | American Heart Association](https://www.heart.org/en/news/2023/11/10/new-tool-brings-big-changes-to-cardiovascular-disease-predictions) incorporates a new calculator for cardiovascular-kidney-metabolic syndrome.

**5. Conclusion and Recommendations**

As death from heart disease is steadily one of the leading causes, with a staggering statistic of one person dying every 33 seconds, it is important to find ways to prevent it from happening. By applying different machine learning techniques this project established that there are certain measures within the most monitored factors causing heart disease that have more impact than others, for example cholesterol, or maximum blood pressure. After analyzing several machine learning techniques, we found that people with increased cholesterol, and higher heart rate were more likely to suffer from a heart attack, which we previously found to be more prominent than others. That suggests that while looking at the bigger picture of patients' charts is important, we can see the outliers and use them in combination with machine learning techniques and potentially lower the number of people dying from heart attacks. Decision Science / Business Statistics helped introduce the data and give us a bigger picture of the statistics. Introduction to Business Analytics we looked at the 4 types of business analysis and found the correlations between the output variable and the other ones. Business Intelligence with cost sensitivity analysis, based on previously found most important variables, gave insight into minimizing errors in predicting heart disease. Data Mining Tools gave us an overview of the whole data set (a larger amount of data) and showed outliers within variables. Business Data Visualization showed us the relationships and distribution within the variables. Database Management Systems using SQL servers we have created a database to store all the patients' information in an organized manner. Database warehousing - by using ETL (Extract, Transform, Load) we can store data in a structured way, and we can prepare it for further analysis. Business Process Automation helped utilize automation processes to improve efficiency and lower human error. Web Analytics was a great tool to advertise prevention techniques and showing the importance of self-care and paying attention to our health. Accounting and Business Information Systems we utilized the collect, store, display, and process information with the use of different programs. Future research into the use of machine learning techniques should focus on finding more variables that are more important than others and utilizing them for preventing heart disease. If the importance of heart health can be promoted and taken seriously with the combination of application of different tools to gain more insight, the statistics of deaths and people suffering from the diseases can be lowered. Analytical teams should be working closely with doctors to find ways to help each other and implement improvements.

**6. Limitations Appendix**

* Data size: 303 samples
* Four different databases
* Slightly outdated from 1988 - as heart disease is a present issue, it has changed over the years with the improvement of technology, the health trends within society, etc.

**7. References**

*Cardiovascular research center - overview*. Mayo Clinic. (2023, May 2). https://www.mayo.edu/research/centers-programs/cardiovascular-research-center/overview

Centers for Disease Control and Prevention. (2024, January 9). *Heart attack symptoms, risk, and recovery*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/heart\_attack.htm

*Department of Public Health*. data.illinois.gov. (n.d.). https://data.illinois.gov/organization/department-of-public-health

Zaganjori, J. (2024, January 19). *Heart attack prediction*. Kaggle. https://www.kaggle.com/datasets/juledz/heart-attack-prediction

*More than half of U.S. adults don’t know heart disease is leading cause of death, despite 100-year reign*. American Heart Association. (n.d.). https://newsroom.heart.org/news/more-than-half-of-u-s-adults-dont-know-heart-disease-is-leading-cause-of-death-despite-100-year-reign