AoL Data Mining and Visualization

Ghazi Athari

2024-06-22

Introduction

Globally, approximately **17.9 million people die from heart attacks** each year, making it *one of the leading causes of death worldwide*. This staggering number represents 4% of all deaths, equating to roughly 1 in every 25 deaths. The urgency is underscored by the statistic that for individuals who do not receive prompt stenting (percutaneous coronary intervention or PCI), *the death rate increases dramatically by 3.3 deaths per 100 every 10 minutes*. This highlights the critical need for timely medical intervention and the importance of early detection and preventive measures.

Our analysis aims to identify early signs of heart problems, thereby increasing preparation time for patients and reducing the risk of fatal heart attacks. Early detection can lead to timely treatment and lifestyle changes, crucial in preventing severe cardiac events. This effort aligns with the third goal of the **Sustainable Development Goals (SDG)**, which seeks to ensure healthy lives and promote well-being for all ages. By focusing on heart disease, we contribute to achieving this goal, addressing the significant global burden of cardiovascular diseases.

To conduct this analysis, we utilize the **"Heart Attack Analysis & Prediction Dataset"** by Rashik Rahman, available on Kaggle. This comprehensive dataset includes variables related to heart health, such as age, sex, cholesterol levels, blood pressure, and other key indicators of cardiovascular risk. Our methods involve **Exploratory Data Analysis (EDA)** to understand the dataset and uncover initial patterns. with EDA, we can identify *significant predictors of heart attacks*.

Our research also underscores the importance of *public health initiatives focused on heart disease prevention and education*. Increasing awareness of the risk factors and symptoms of heart disease can empower individuals to seek medical advice and adopt healthier lifestyles, contributing to a broader reduction in heart attack incidence and improving overall population health.

Data Description

1. Age: The patient's age.
2. Sex: The patient's gender

- 0: female
- 1: male

3. ChestPainType: Chest Pain Types
- 0: Typical Angina
- 1: Atypical Angina
- 2: Non-Anginal Pain
- 3: Asymptomatic

4. RestBloodPressure: Resting Blood Pressure
5. Cholesterol: Serum Cholesterol Levels
6. FastingBloodSugar: Fasting Blood Sugar
- 0: <= 120 mg/dL
- 1: > 120 mg/dL

7. RestECG: Resting ECG Results
- 0: Normal
- 1: ST-T Wave Abnormality
- 2: Probable or Definite Left Ventricular Hypertrophy

8. MaxHeartExercise: Maximum Heart Rate During Exercise
9. ExerciseAngina: Exercise-Induced Angina
- 0: No
- 1: Yes

10. SegmentDepression: ST-Segment Depression
11. SegmentSlope: Slope of ST Segment
- 0: Downsloping
- 1: Flat
- 2: Upsloping diagnosis

12. MajorVesselsColored: Number of Major Vessels Colored by Fluoroscopy (CAA)
13. ThalasemmiaType: Thalassemia Type
- 0: None (Normal)
- 1: Fixed Defect
- 2: Reversible Defect
- 3: Thalassemia

14. HeartAttack: Risk of Heart Attack
- 0: No
- 1: Yes

Data Preprocessing

We preprocessed the data starting from changing the variables' names to make the data more readable. Then, we changed some variables' datatypes into categorical. Next we see if the dataset have any missing values with a code from ProjectPro.

```
missing <- is.na(data)
count <- colSums(missing)
print(count)
```

```
##                  Age                  Sex       ChestPainType
RestBloodPressure
##                    0                    0                   0
0
##          Cholesterol   FastingBloodSugar              RestECG
MaxHeartExercise
##                    0                    0                   0
0
##      ExerciseAngina   SegmentDepression        SegmentSlope
MajorVesselsColored
##                    0                    0                   0
0
##     ThalassemiaType         HeartAttack
##                    0                    0
```

Our dataset doesn't have any missing values. We can continue to preprocess the dataset by searching for duplicate values with a code from rdocumentation.org

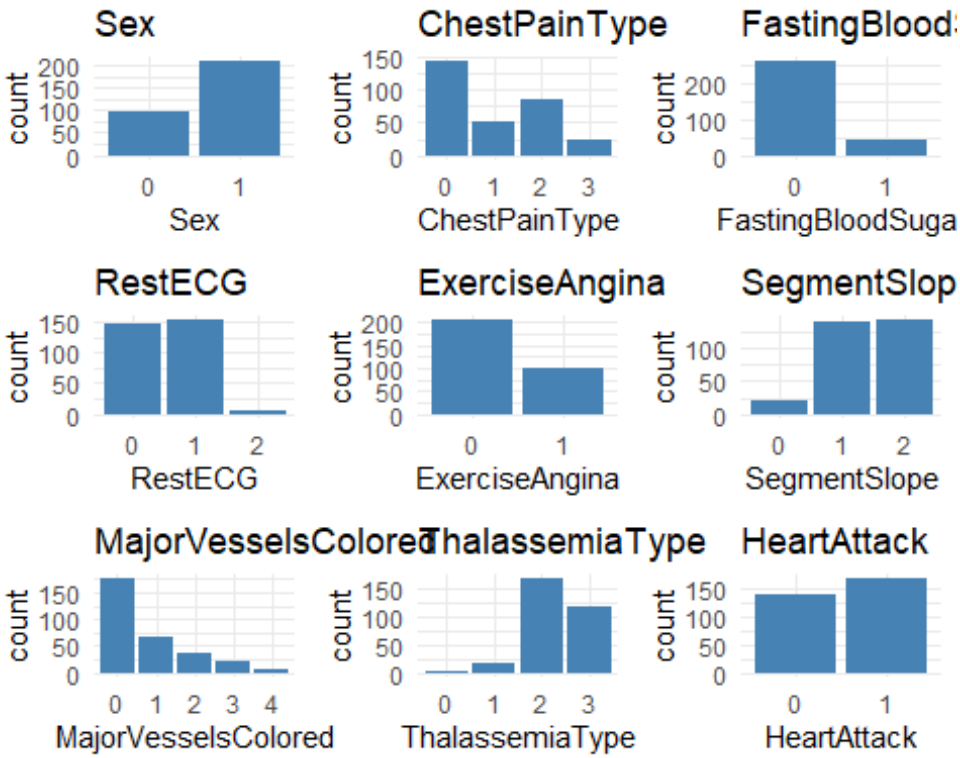
```
duplicate <- duplicated(data)
duplicate_rows <- data[duplicate, ]
print(duplicate_rows)
```

From running that code, we have one duplicate values. We can remove the duplicate with this code.

```
data_fixed <- data %>% distinct()
```

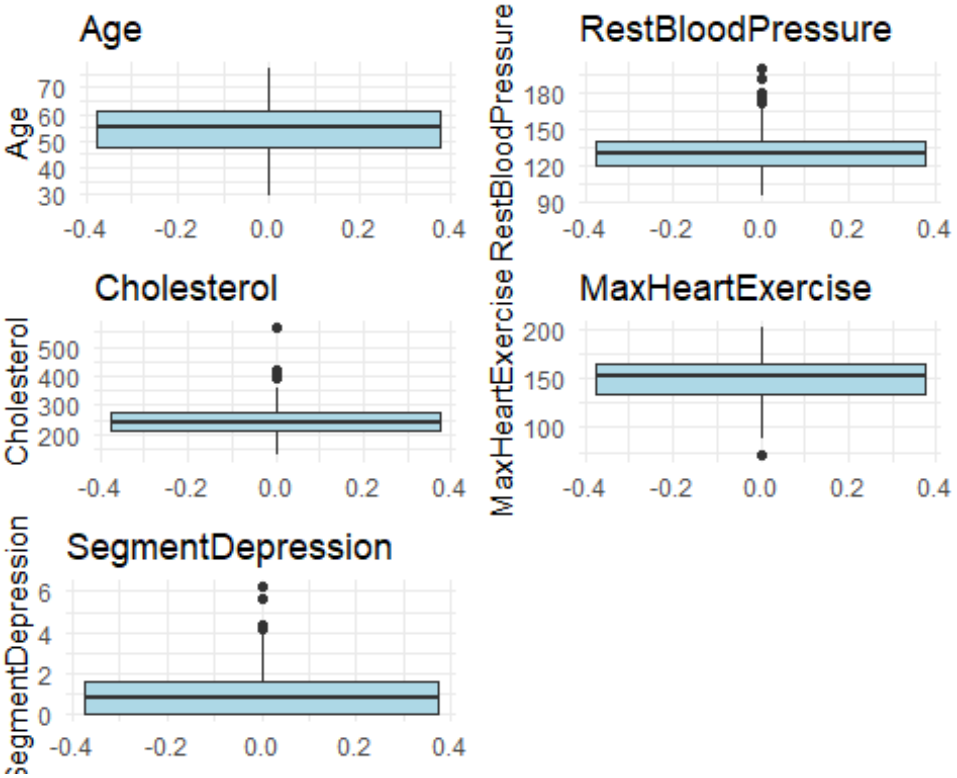
Data Exploration

Now that our data is clean and preprocessed, let's see the categorical variables' barplots look like



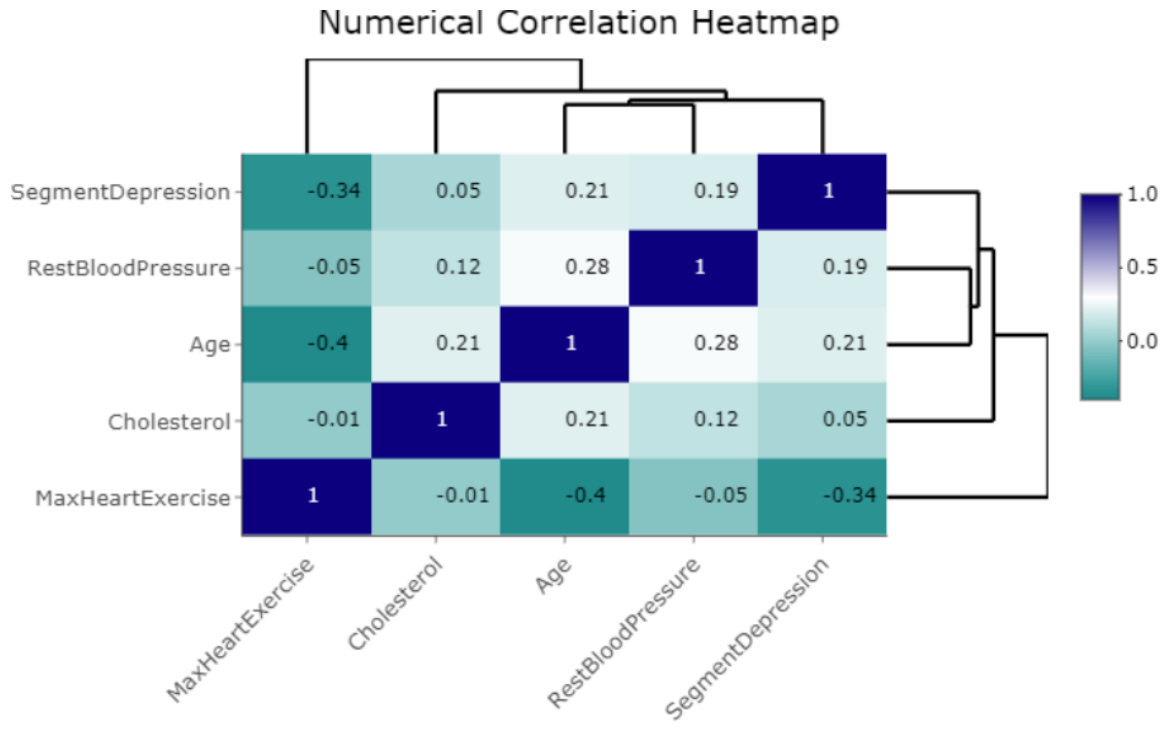
Here, most of the categorical features are imbalanced.

Next, We'll see the numerical ones



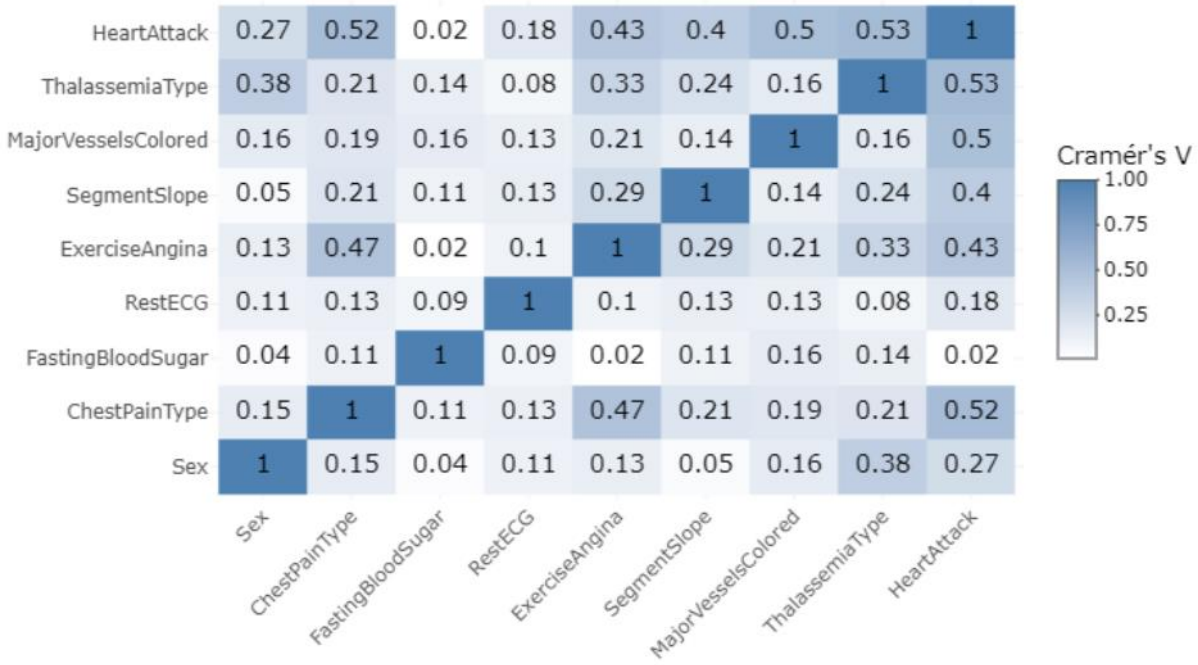
Looks like there are several outliers, but since there's not a lot of outliers and the outliers follow the data trend, we can leave them be.

Now let's see if out numerical data are correlating to eachother



Looks like there's not really any correlation in the numerical data.

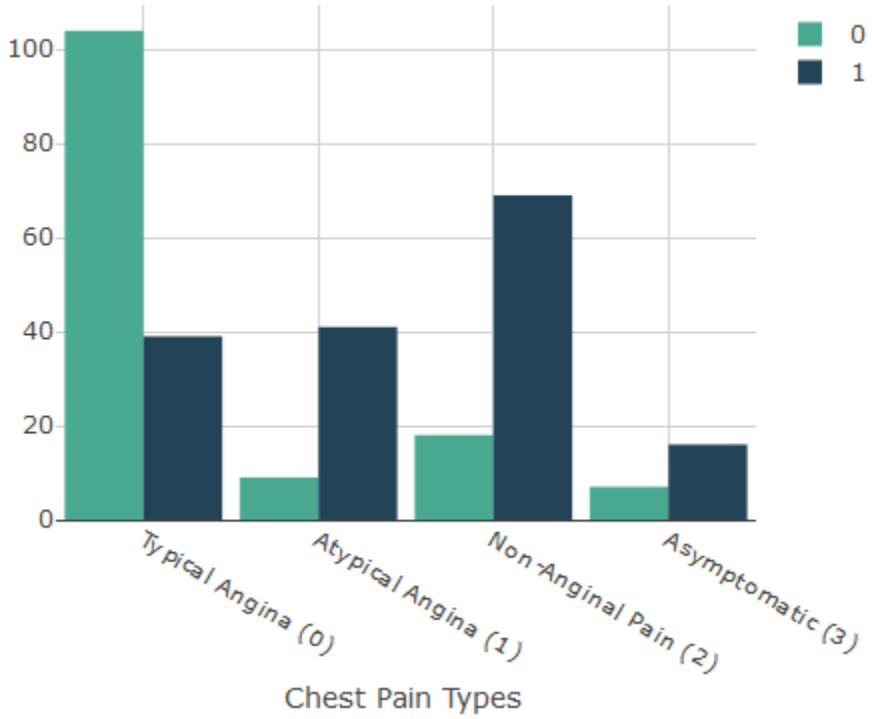
Now let's look at the *Cramér's V* plot for the correlations between categorical data



From the *Cramér's V plot*, it is evident that there is a correlation between the risk of heart attack and the **type of chest pain experienced, the number of major vessels colored by fluoroscopy, and the type of Thalassemia**. Specifically, the risk of heart attack appears to be related to whether a person has typical angina, atypical angina, non-anginal pain, or is asymptomatic. Similarly, the risk of heart attack is also associated with the number of major vessels colored by fluoroscopy, and whether a person has normal Thalassemia, a fixed defect, a reversible defect, or Thalassemia.

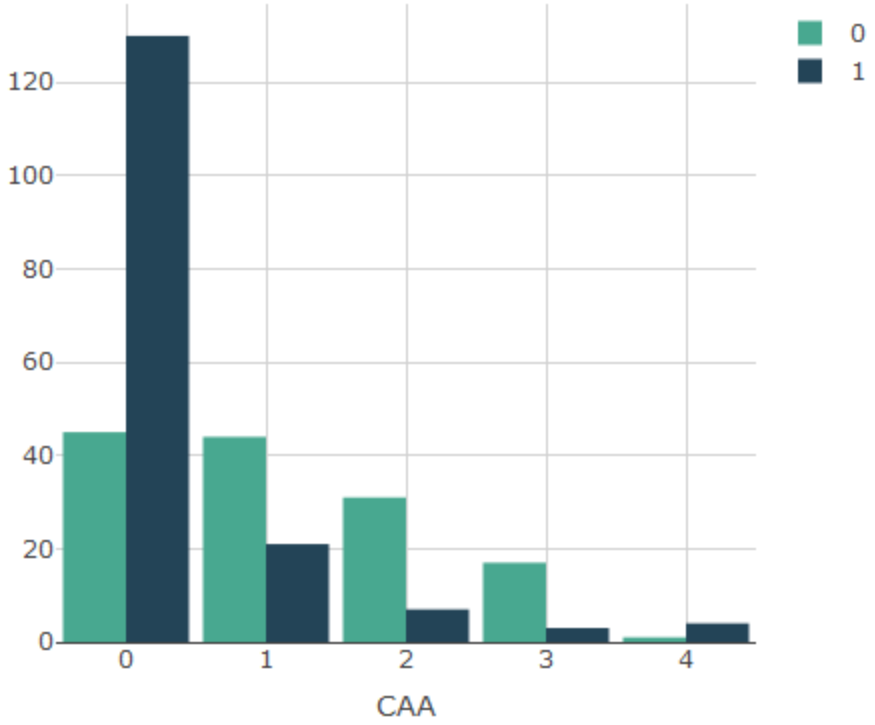
Statistical Analysis

Next, we will see the the relationship between the risk of heart attack and different types of chest pain



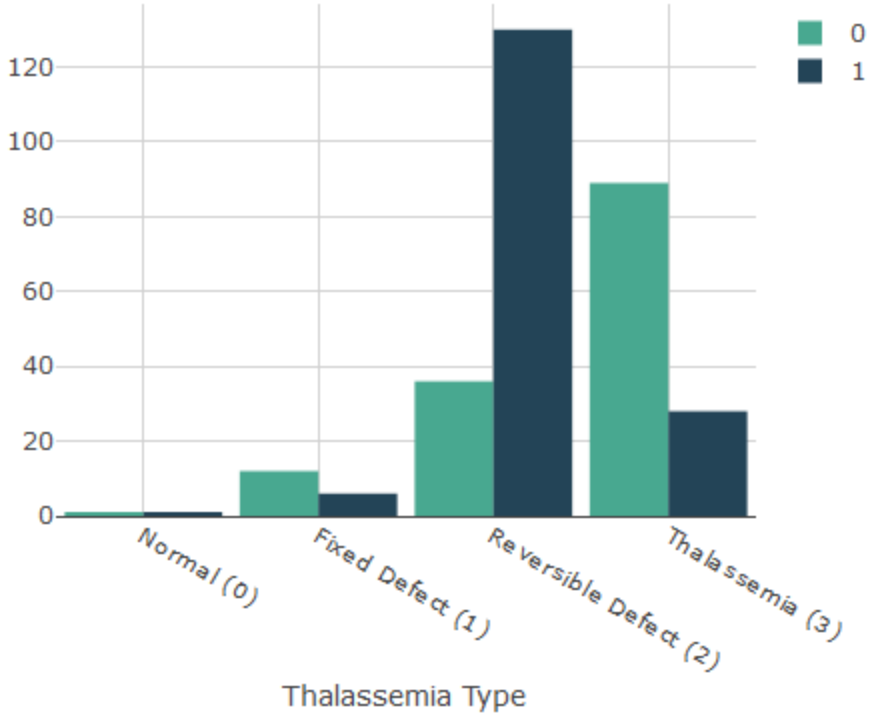
Here, we can see that people with **chest pain type 1 (Atypical Angina)** are more prone to heart attacks.

Next we'll see the connection between the risk of having a heart attack and the number of major vessels colored by fluoroscopy.



In this case, it appears that individuals who have a **fluoroscopy result showing no major vessels colored (a CAA value of 0)** are more susceptible to heart attacks.

And lastly, we'll see the relationship between the risk of heart attack and the type of thalassemia.



From this observation, it seems that individuals with **Thalassemia type 2**, also known as a **Reversible Defect**, show a higher susceptibility to heart attacks.

Discussion

The analysis reveals several noteworthy associations between various factors and the risk of heart attacks. Firstly, a negative correlation is observed between max heart exercise and segment depression, indicating an inverse relationship between these variables. Secondly, individuals experiencing Atypical Angina (chest pain type 1) appear to be more susceptible to heart attacks, suggesting the importance of considering chest pain type as a predictive factor. Additionally, those with a major vessels colored (CAA) value of 0 demonstrate heightened vulnerability to heart attacks, indicating a potential link between the absence or lower count of major vessels colored and increased risk. Lastly, individuals with Thalassemia type 2 (Reversible Defect) exhibit a higher likelihood of heart attacks, underscoring the significance of Thalassemia type in assessing cardiac health. These findings shed light on potential risk factors and underline the importance of comprehensive evaluation and management strategies to mitigate the risk of heart attacks.

Conclusion

In conclusion, our analysis of the **"Heart Attack Analysis & Prediction Dataset"** has led to some significant findings that can contribute to the early detection and prevention of heart attacks. We have identified that individuals with *chest pain type 1 (Atypical Angina)*, those with a *fluoroscopy result showing no major vessels colored (a CAA value of 0)*, and individuals with *Thalassemia type 2 (Reversible Defect)* are more prone to heart attacks.

These findings not only enhance our understanding of the key predictors of heart attacks but also emphasize the importance of early detection and intervention. By identifying these risk factors, we can provide timely medical attention and lifestyle guidance to those most at risk, thereby reducing the global mortality rate from heart attacks.

Moreover, these insights further underline the necessity for **public health initiatives focused on heart disease prevention and education**. By increasing public awareness about these risk factors, we can empower individuals to seek medical advice and adopt healthier lifestyles.

Overall, our analysis aligns with the third goal of the **Sustainable Development Goals (SDG)** - ensuring healthy lives and promoting well-being for all ages. By focusing on heart disease, we are contributing to reducing the global burden of cardiovascular diseases and moving one step closer to achieving this significant goal.

References

- https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2018/february/every-minute-counts-when-it-comes-to-heart-attack-treatment
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6192063/
- https://www.webmd.com/heart-disease/heart-failure/early-diagnosis-heart-failure#:~:text=The%20sooner%20that%20you%20know,and%20prevent%20other%20health%20issues
- https://www.heartfoundation.org.au/your-heart/evidence-and-statistics/key-statistics-heart-attack
- https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2253710/
- https://www.projectpro.io/recipes/find-count-of-missing-values-dataframe
- https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/duplicated
- https://www.kaggle.com/code/abraamsaid/heart-attack-eda-ann