Sandrine FOMO

**REPORT 1(Exploratory data analysis)**

**TITLE**

**Exploring Cardiovascular Health: A Comprehensive Analysis of Heart Attack Risk**

**A-Introduction:**

This report aims to conduct a thorough analysis of cardiovascular health by exploring various factors that may contribute to the risk of heart attacks. The investigation is motivated by the critical importance of understanding and mitigating heart attack risks, given the significant impact cardiovascular diseases have on global health.

**A-1 Background:**

Cardiovascular diseases, including heart attacks, are leading causes of morbidity and mortality worldwide. Understanding the factors associated with heart attack risk is crucial for preventive healthcare strategies and improving overall heart health. The dataset under examination was downloaded from Kaggle comprises diverse variables related to age, gender, heart rate, blood pressure, glucose levels, and cardiac biomarkers.

(https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset).

The motivation for examining this dataset lies in its potential to uncover patterns and associations between these variables and the likelihood of experiencing a heart attack. Such insights can inform targeted interventions, personalized healthcare plans, and public health initiatives to reduce the incidence of cardiovascular events. By conducting a comprehensive analysis of this dataset, we aim to enhance preventative measures and promote heart health at both individual and population levels.

**B- Data**

**B-1 Data Description.**



***Add comments:***

eight independent variables (or predictors) and one dependent variable (or response variable). Age, gender (0 for Female, 1 for Male), heart rate (impulse), systolic BP (pressurehight), diastolic BP (pressurelow), blood sugar(glucose), CK-MB (kcm), and Test-Troponin (troponin) are representing the predictors variables, while the response variable pertains to the presence of heart attack (class), which is divided into two categories (negative and positive); negative refers to the absence of a heart attack, while positive refers to the presence of a heart attack.

CK-MB (Creatine Kinase-MB): CK-MB is an enzyme found in the heart muscles. It is often measured in blood tests to assess whether there has been damage to the heart muscle, such as during a heart attack.

Troponin: is a group of proteins found in cardiac and skeletal muscle fibers. Elevated troponin levels in the blood indicate damage to the heart muscle.

**B-2 Summary Statistics**



***add comments:***

The output from the describe function provides summary statistics for each variable in the dataset.

**Age (Variable 1):**

Mean age is approximately 56.19 years, with a standard deviation (sd) of 13.65.

The age distribution is slightly negatively skewed (-0.24), suggesting that there are more individuals with ages above the mean.

The kurtosis is close to 0 (-0.03), indicating a relatively normal distribution.

**Gender (Variable 2):**

Gender is encoded as 0 for Female and 1 for Male.

The mean for this binary variable is 0.66, indicating a slightly higher proportion of males (as 1 represents males).

The skewness and kurtosis suggest a moderate departure from a normal distribution, with a negative skewness and a negative kurtosis.

**Impluse (Variable 3):**

The mean heart rate (impulse) is approximately 78.34 beats per minute.

The distribution has a relatively high skewness (18.21), indicating a rightward (positive) skew.

The kurtosis is extremely high (361.02), suggesting heavy tails and potential outliers.

**Pressure High (Variable 4):**

The mean systolic blood pressure is approximately 127.17 mmHg.

The skewness is close to 0.68, indicating a moderate rightward skew.

The distribution has a positive kurtosis (0.94), suggesting a slightly heavier tail than a normal distribution.

**Pressure Low (Variable 5):**

The mean diastolic blood pressure is approximately 72.27 mmHg.

The skewness is close to 0.49, indicating a slight rightward skew.

The kurtosis is positive (0.85), suggesting a distribution with a slightly heavier tail than a normal distribution.

**Glucose (Variable 6):**

The mean blood sugar level is approximately 146.63 mg/dL.

The distribution has a positive skewness (1.93), indicating a rightward skew.

The kurtosis is 4.09, suggesting heavy tails and potential outliers.

**CK-MB (Variable 7):**

The mean CK-MB level is approximately 15.27.

The distribution is highly positively skewed (4.97), indicating a rightward skew.

The kurtosis is 25.35, suggesting heavy tails and potential outliers.

**Troponin (Variable 8):**

The mean Troponin level is approximately 0.36.

The distribution is highly positively skewed (5.79), indicating a rightward skew.

The kurtosis is 39.80, suggesting heavy tails and potential outliers.

**Class (Variable 9):**

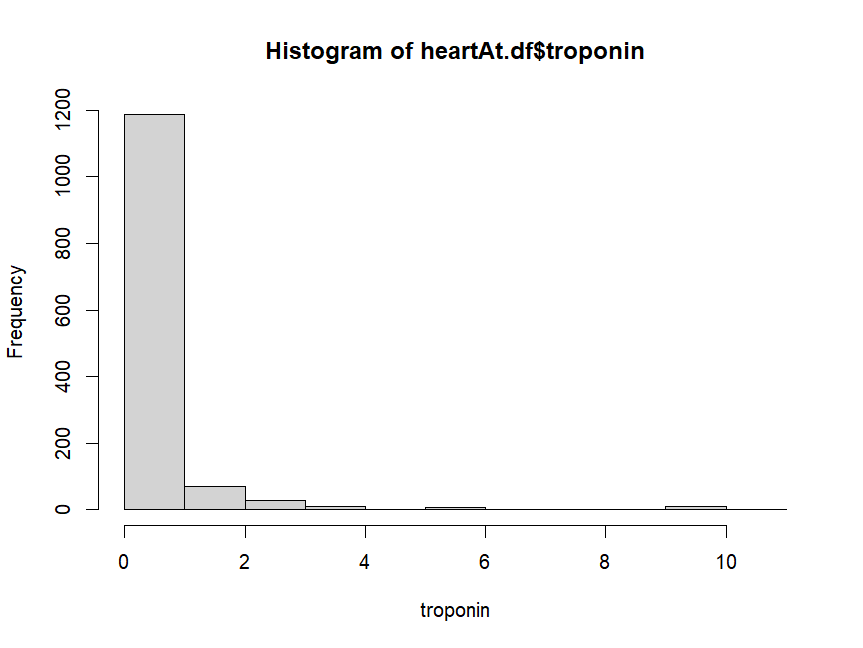
The 'class' variable, representing the presence (positive) or absence (negative) of a heart attack, has a mean of approximately 1.61.

The skewness is negative (-0.47), indicating a slight leftward skew.

The kurtosis is negative (-1.78), suggesting a distribution with lighter tails than a normal distribution.

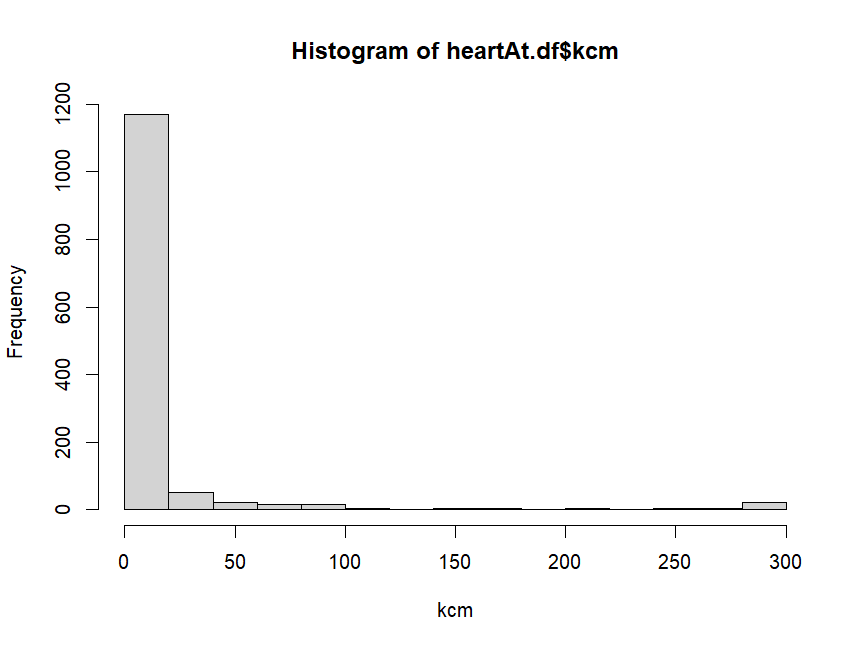
**C- Data Visualization**

**C-1 Histograms and Density Functions**

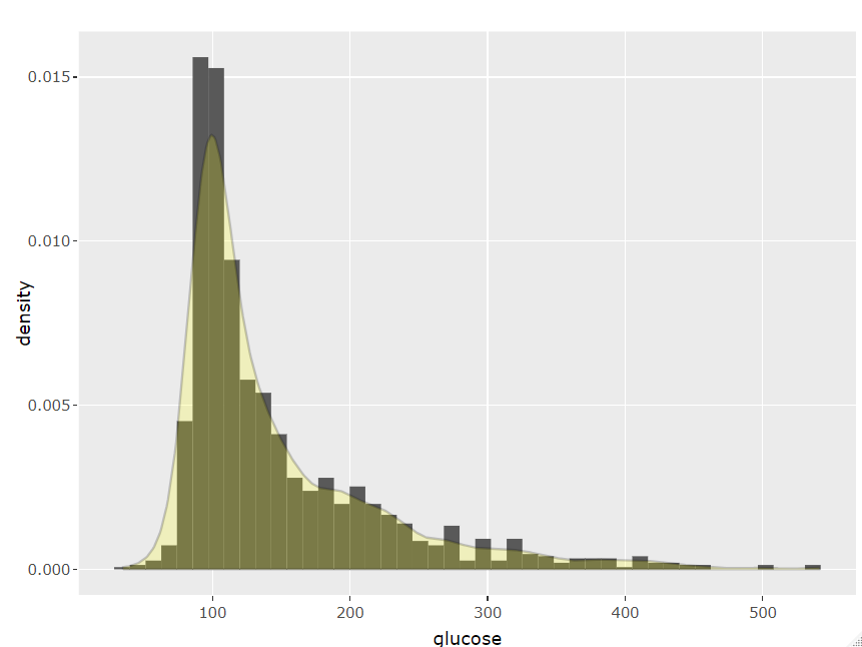
****

*Add comments:*

The distribution of troponin is skewed to the right.

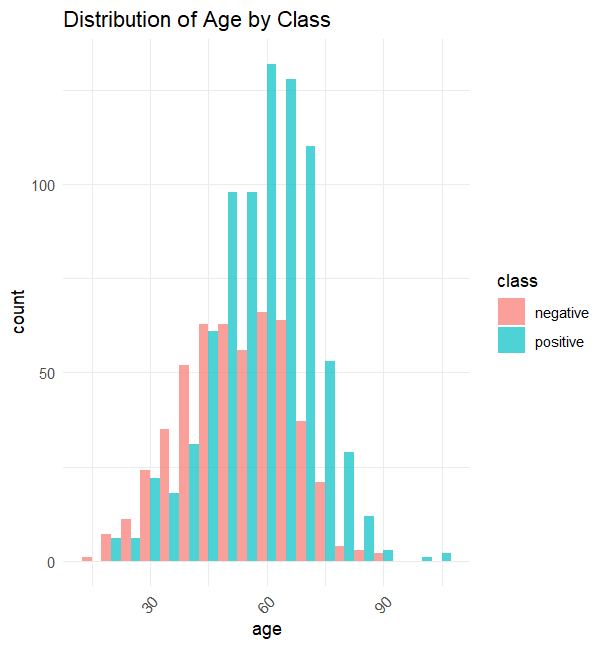
 *Add comments:*

The distribution of kcm is skewed to the right.



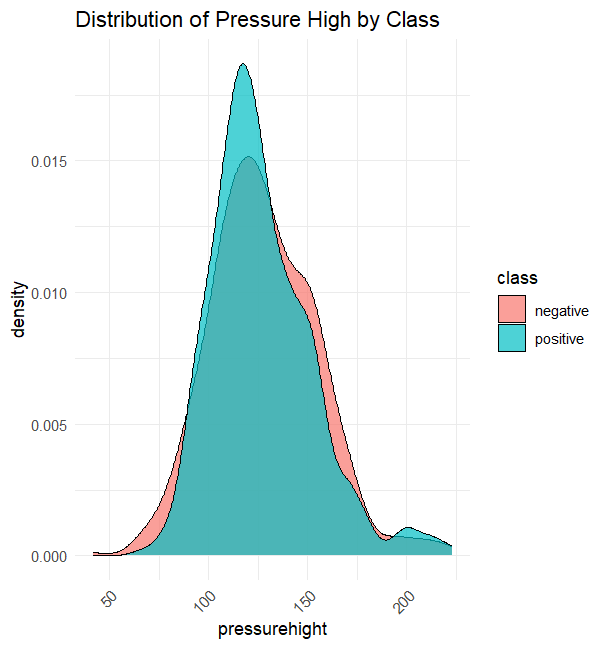
*Add comments:*

The distribution of kcm is skewed to the right.



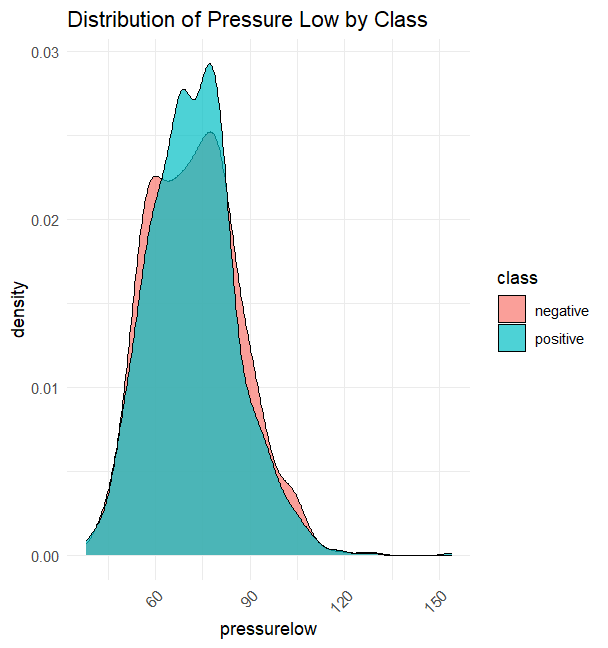
*add comments:*

The distribution shows that people aged 60-65 are at a higher risk of having a heart attack compared to other age groups. The data for both groups show some right skewness

**

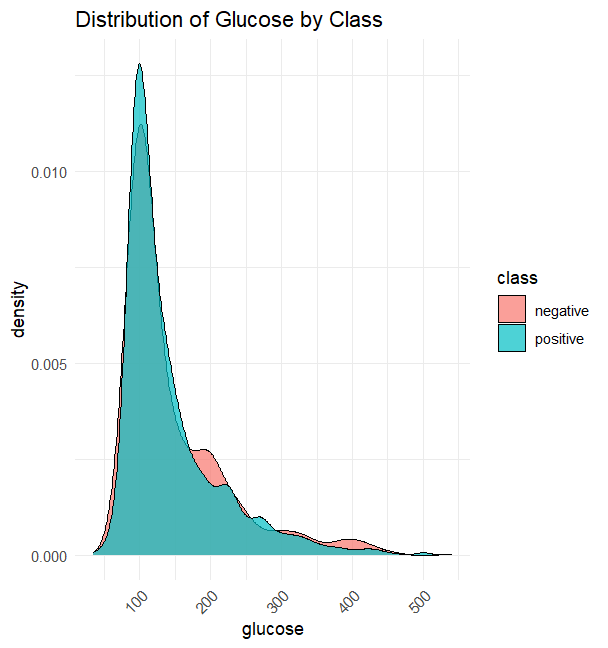
*add comments:*

The plot shows that people with a heart attack tend to have higher systolic blood pressure (pressurehight) compared to those in the negative group. Both plots are skewed to the right.



*add comments:*

The plot shows that people with a heart attack tend to have higher diastolic blood pressure (pressurelow) compared to those in the negative group. Both plots are slightly skewed to the right.

**

*add comments:*

The plot shows that people with a heart attack tend to have high blood sugar(glucose), compared to those in the negative group. Both plots are slightly skewed to the right.

**C-2 Scatter plots**

**

*add comments:*

There seems to be no correlation between diastolic blood pressure (pressurelow) and blood sugar(glucose) in both groups.



*add comments:*

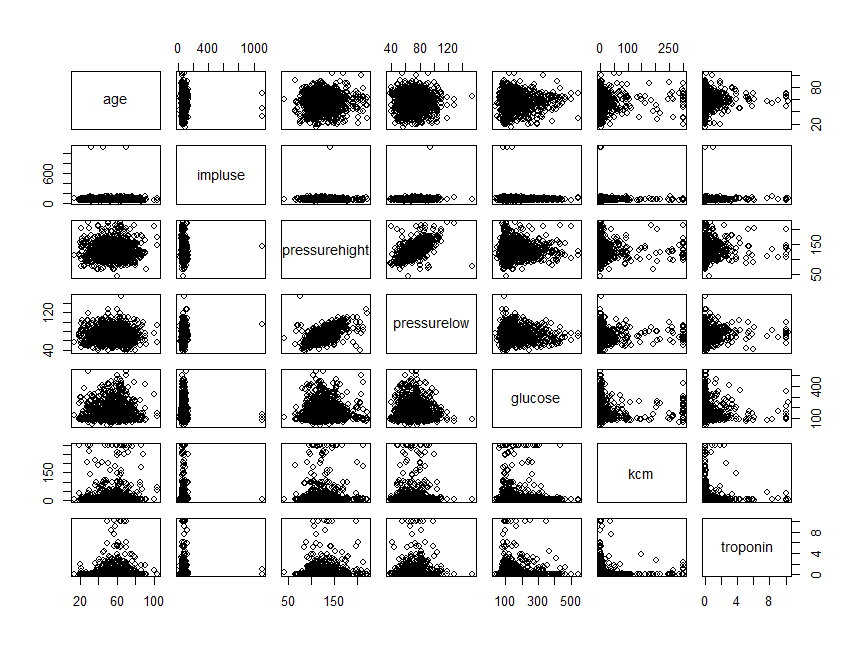
There seems to be no correlation between systolic blood pressure (pressurehight) and blood sugar(glucose) in both groups.



*add comments:*

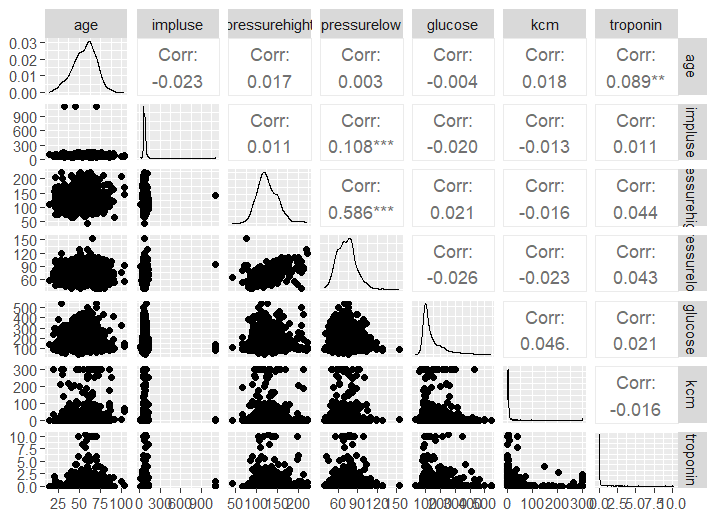
There seems to be a positive correlation between diastolic blood pressure (pressurelow) and systolic blood pressure (pressurehight) in both groups.

***C-3 Matrix plot***

******

*Add comments:*

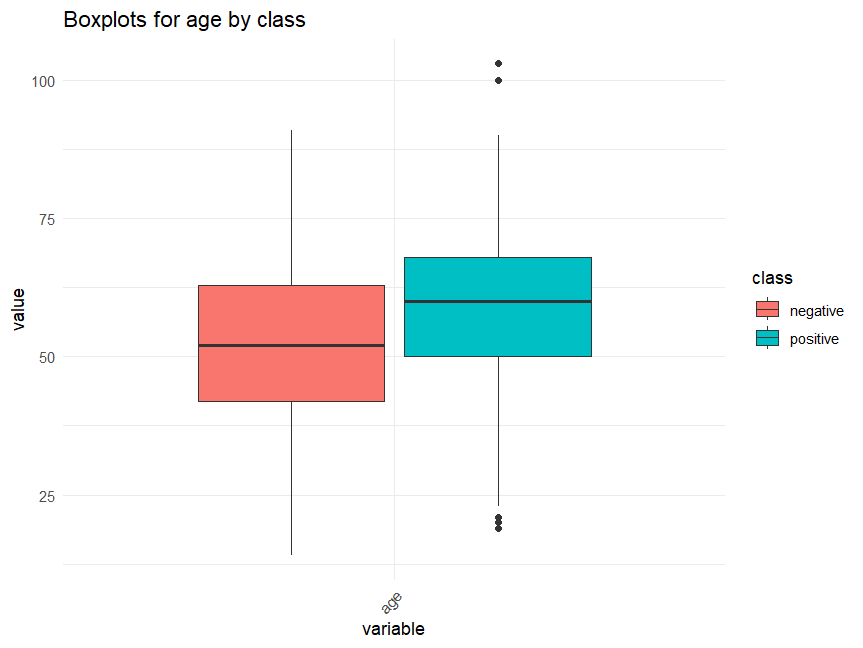
The Matrix plot shows somewhat low to no correlation between variables except for pressurelow and pressurehight which displays a moderately positive correlation.

**

*Add comments:*

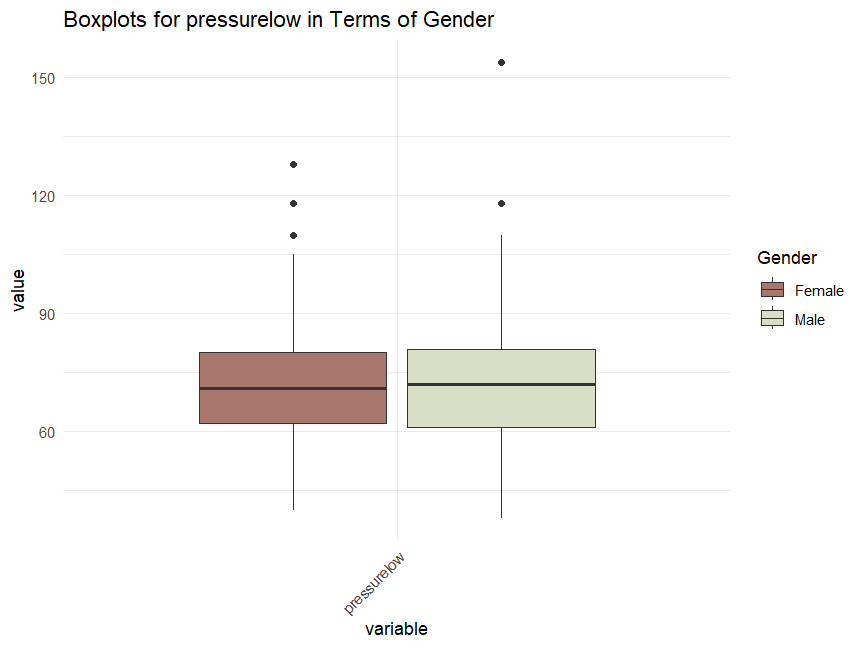
There is a positive correlation (0.586) between diastolic blood pressure (pressurelow) and systolic blood pressure (pressurehight).

**C-4 Boxplots**

**

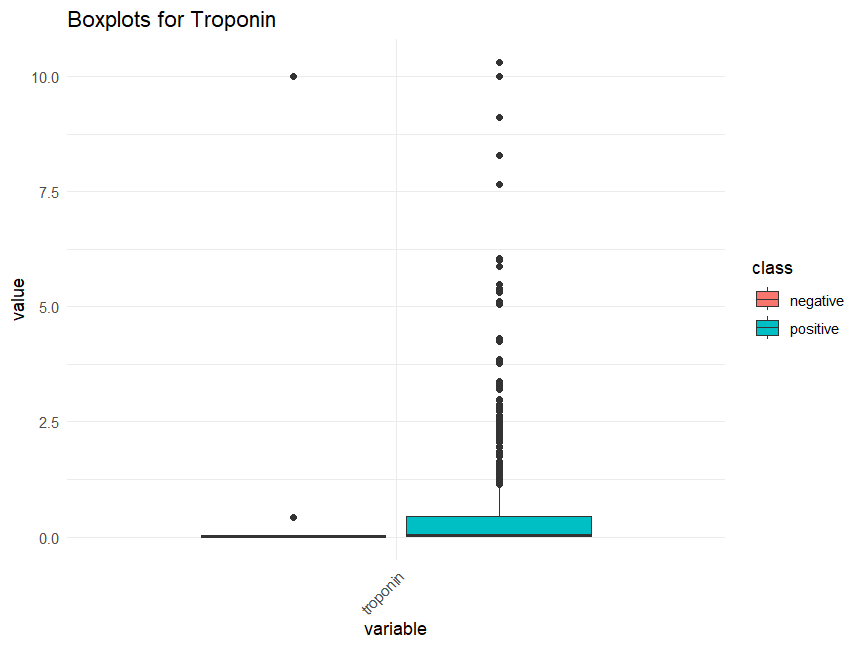
*Add comments:*

The plot indicates a higher median age for individuals with a heart attack compared to those without. The group of people experiencing a heart attack also exhibits outliers.



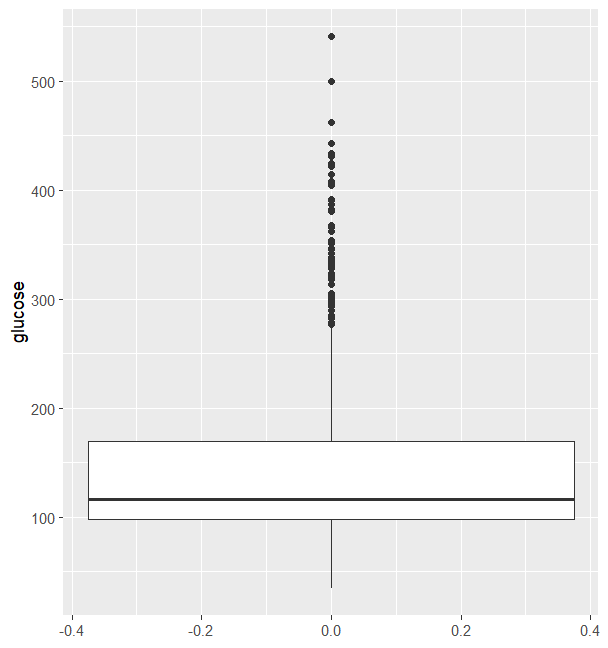
*Add comments:*

The plot indicates a slightly higher median diastolic blood pressure (pressurelow) for individuals with a heart attack compared to those without. Both group exhibit outliers.



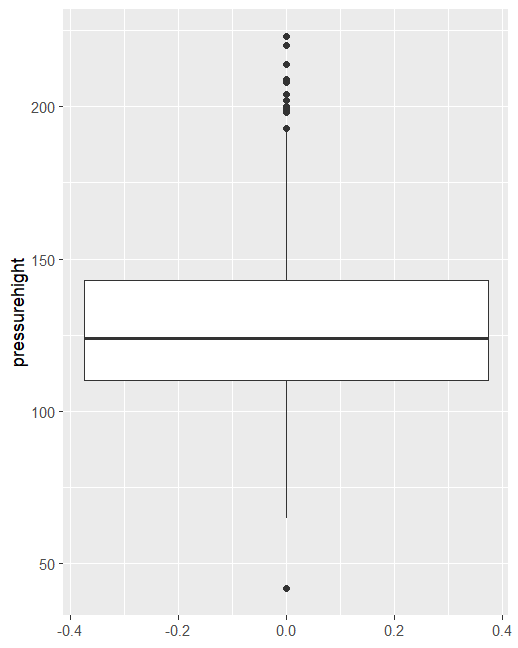
*Add comments:*

Although the median troponin level is 0 for both groups, individuals without a heart attack almost have troponin values of zero. Meanwhile, at least 75% of those testing positive for troponin have values ranging from 0 to approximately 0.35. The group of people testing positive also exhibits a notable number of outliers compare to those without.

**

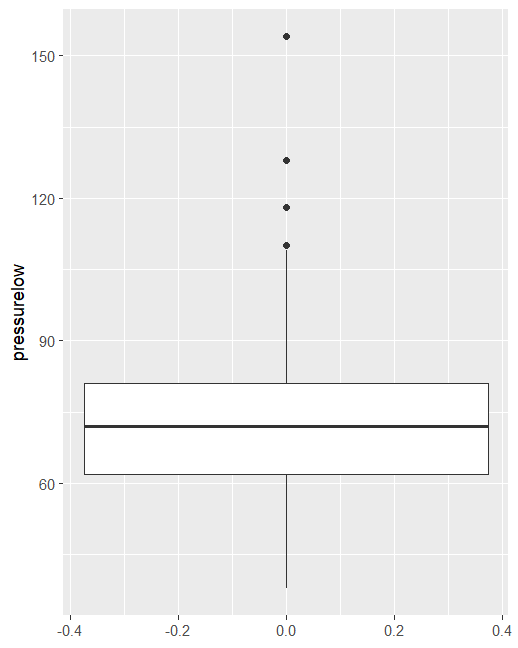
*Add comments:*

The plot represents the blood sugar levels (glucose) of all participants. Twenty-five percent have their blood sugar equal to 100 mg/dL. The median value is approximately 120 mg/dL, and the third quartile is around 170 mg/dL. The plot also reveals the presence of outliers.

**

*Add comments:*

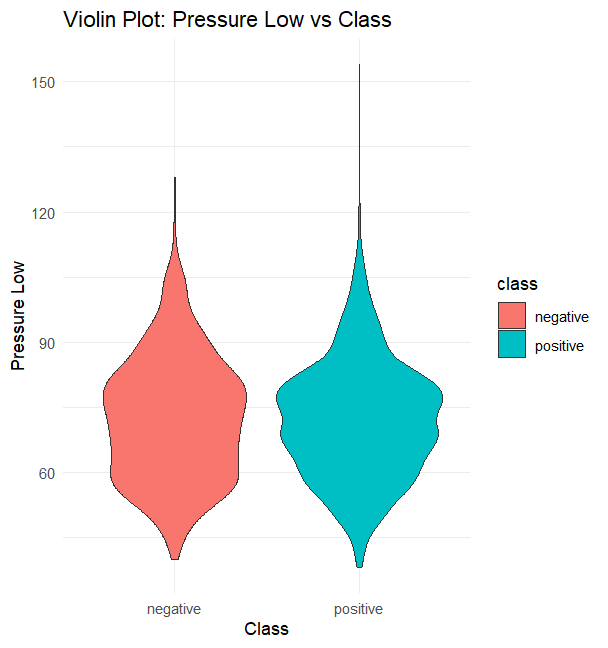
The plot represents the systolic blood pressure (pressurehight) of all participants. Twenty-five percent have their systolic blood pressure equal to 115 mmHg. The median value is approximately 125 mmHg, and the third quartile is around 140 mmHg. The plot also reveals the presence of some outliers.

**

*Add comments:*

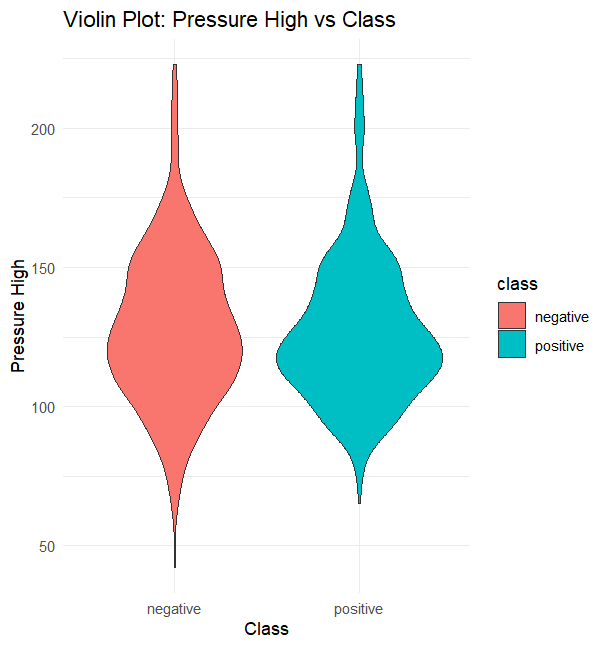
The plot represents the diastolic blood pressure (pressurelow) of all participants. Twenty-five percent have their diastolic blood pressure equal to 62 mmHg. The median value is approximately 72 mmHg, and the third quartile is around 80 mmHg. The plot also reveals the presence of a few outliers.

**C-5 Violin plots**

**

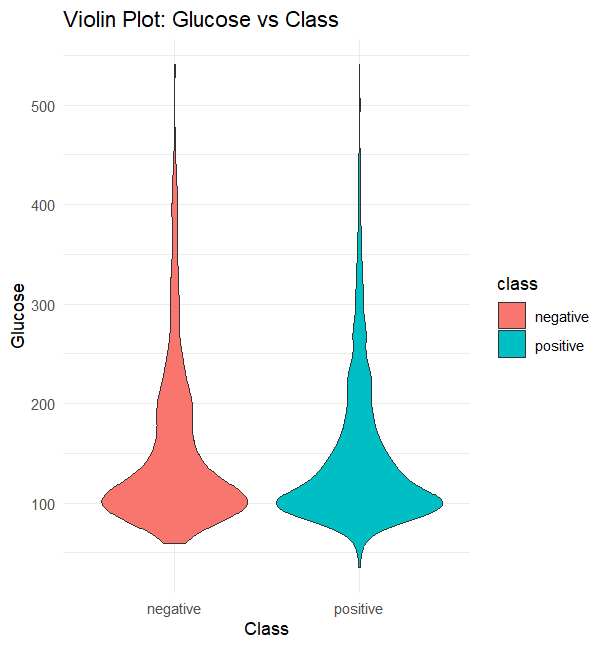
*Add comments:*

The group of people with a heart attack has a wider distribution around 75 mmHg diastolic blood pressure, representing higher density than the group without. Additionally, the group of people with a heart attack is taller than the group without, indicating greater skewness.

**

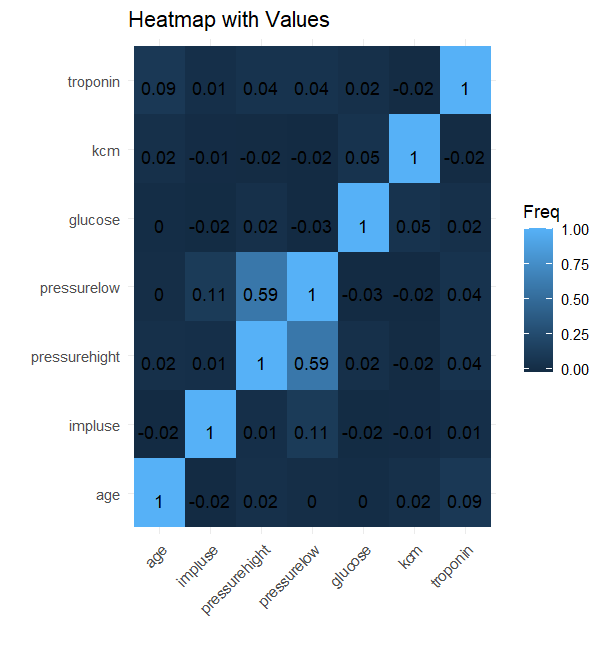
*Add comments:*

The group of people with a heart attack has a wider distribution around 120mmHg, representing higher density than the group without.

**

*Add comments*

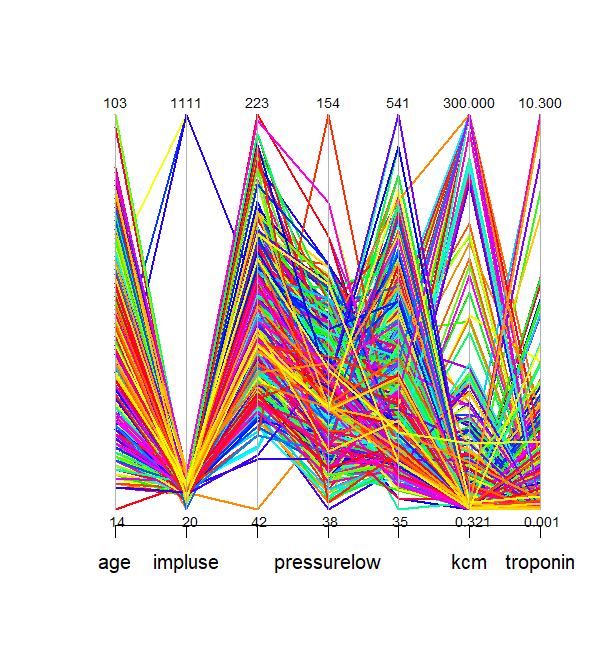
The group of people with a heart attack has a wider distribution around 100 mg/dL, representing higher density than the group without. Both groups are skewed to the right.

**C-6 Heatmap****

*Add comments:*

The highest correlated variables are pressurelow and pressurehight, followed by pressurelow and impluse.

**C-7 Parallel coordinates plot**

**

*Add comments:*

Parallel lines suggest that impulse, presurelow, kcm and troponin are not strongly correlated.

**D- Summary and Conclusion**

This comprehensive analysis aimed to uncover patterns and associations related to cardiovascular health, specifically examining factors that may contribute to the risk of heart attacks. The dataset, sourced from Kaggle, encompasses diverse variables such as age, gender, heart rate, blood pressure, glucose levels, and cardiac biomarkers. Cardiovascular diseases, including heart attacks, are significant contributors to global morbidity and mortality. Therefore, understanding the factors associated with heart attack risk is crucial for developing effective preventive healthcare strategies.

The findings reveal several key insights. Individuals in the age group of 60-65 face a higher risk of heart attacks compared to other age groups. Blood pressure emerges as a notable factor, with individuals experiencing heart attacks demonstrating higher systolic and diastolic blood pressure levels. A positive correlation (0.586) between diastolic and systolic blood pressure further underscores their interconnected influence on heart health.

Blood sugar levels also play a role, with individuals who have experienced heart attacks exhibiting higher glucose levels compared to those in the negative group. Cardiac biomarkers, specifically CK-MB and Troponin, display highly positively skewed distributions, suggesting the presence of potential outliers. Notably, individuals without heart attacks often present Troponin values of zero.

Gender differences indicate a slight male predominance in the dataset. The distribution of heart rate (impulse) is rightward skewed, suggesting potential outliers. These findings collectively contribute to a nuanced understanding of the multifaceted factors influencing heart attack risk.

While this analysis has identified meaningful associations, it is crucial to acknowledge that correlation does not imply causation. The dataset's limitations, including the potential impact of outliers, may influence the generalizability of the findings. Further exploration of outliers could offer additional insights into potential data anomalies or specific conditions affecting the variables under consideration.