Lecturer's Name: Dr James Connolly / Vini Vijayan
Assessment Title: Heart failure prediction data analysis
Work to be submitted to: Dr James / Vini
Date for submission of work: 08-05-2023
Place and time for submitting work: Blackboard

To be completed by the Student
Student's Name: Ashwitha Lakshmy Sunitha
Class: Msc Big Data Analytics
Subject/Module: Data Science
Word Count (where applicable):
I confirm that the work submitted has been produced solely through my own efforts. **Student's signature: Ashwitha Date:** 08-05-2023

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero.

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment shall normally be carried forward from the original examination to the repeat examination.

Declaration:

I declare that this work is entirely my own and does not contain the words or ideas of someone else, whether published or not, without specific acknowledgement by relevant referencing. I have read and understood the LYIT Plagiarism Policy on the "Student & Academic Policies" section of the LYIT Website and understand plagiarism to include:

Direct copying of text, images and other materials (electronic or otherwise) from a book, article, fellow student's essay, handout, web page or other source without proper acknowledgement.

Claiming individual ideas derived from a book, article etc. as one's own and incorporating them into one's work without acknowledging the source of these ideas.

Overly depending on the work of one or more other sources without proper acknowledgement of the source, by constructing an essay, project etc., extracting large sections of text from another source and merely linking these together with a few of one's own sentences.

I understand that it is my responsibility to familiarise myself with and to follow the Institute's Assessment Regulations. I acknowledge that Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations and that penalties will be applied if I breach this policy.

Signed: Ashwitha	**Date: 08-05-2023**

HEART FAILURE PREDICTION

DATA ANALYSIS

GitHub Link: https://github.com/Ashwitha1997/CA1.git

ABSTRACT

This data analysis investigates the correlation between a number of variables and the occurrence of fatal events in people with heart failure. The influence of diabetes, age, ejection fraction, platelet counts, and time since diagnosis on the chance of having a fatal event is one of the study's five research topics. In order to evaluate the hypotheses and reach conclusions during the data analysis, statistical techniques like chi-squared tests are used. The study's findings can aid in a better understanding of the risk factors for heart failure and aid medical practitioners in creating efficient preventative and treatment plans.

Two competing hypotheses are put up for hypothesis testing: the null hypothesis (H0) and the alternative hypothesis (H1). While H1 makes the premise that there is a link or difference between the variables, H0 makes the opposite assumption. The likelihood of receiving the observed data or more extreme results if H0 is true is calculated using statistical methods to evaluate the hypotheses. A modest p-value gives support for H1 by showing that H0 should be rejected since the observed data are unlikely to have happened if H0 is true. The results' ultimate goal is to make inferences about the data based on the statistical analysis.

RESEARCH QUESTIONS

Q1: Is there a difference between those with and without diabetes in their incidence of death events?

Q2: Is there a relationship between age and the likelihood of experiencing a death event?

Q3: Is there a relationship between ejection fraction and the likelihood of experiencing a death event?

Q4: Are there any differences in platelet levels between individuals who experience a death event and those who do not?

Q5: Does the length of time since diagnosis of heart failure impact the likelihood of experiencing a death event?

DATA PREPARATION

Examined the data for mistakes, anomalies, and missing numbers. If a variable is not normally distributed, it should be changed using a log transformation to make it so, then the most pertinent variables should be chosen for analysis.

HYPOTHESIS TESTING Q1

H_0: Patients with diabetes is not related to risk of heart failure

H_1: Patients with diabetes is related to risk of heart failure

Here, using statistical method to evaluate strength of evidence against a null hypothesis. The null hypothesis (H0) in this case is that patients with diabetes are not related to the risk of heart failure, while the alternative hypothesis (H1) is that they are related. We are using variables ‘diabetes’ and ‘DEATH_EVENT’ for the hypothesis.

Name	Description	Type	Comments
Diabetes	If the patient has diabetes	Categorical Independent variable	Variable is categorical and need not be converted
Death event	If the patient died during the follow up period	Categorical Dependent variable	Variable is categorical and need not be converted

STATISTICAL TESTING Q1

We may apply a statistical test to see whether there is a significant correlation between the two variables "diabetes" and "death_event" in order to test the hypothesis that individuals with diabetes are connected to the risk of heart failure.

Below figure shows the correlation between the variables of the dataset

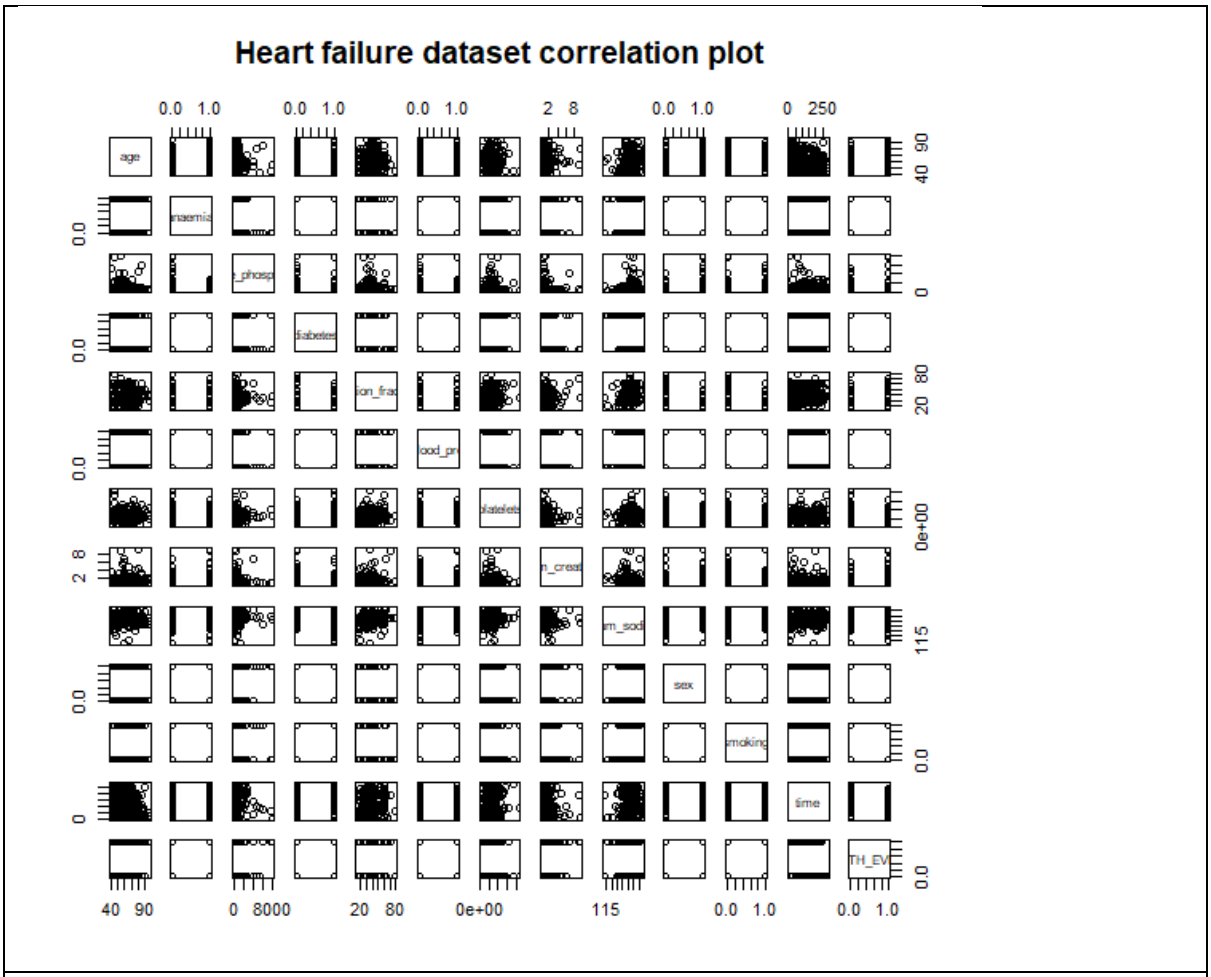

Figure: Shows the correlation between the variables of the dataset

Below figure shows the correlation between the variables of the dataset

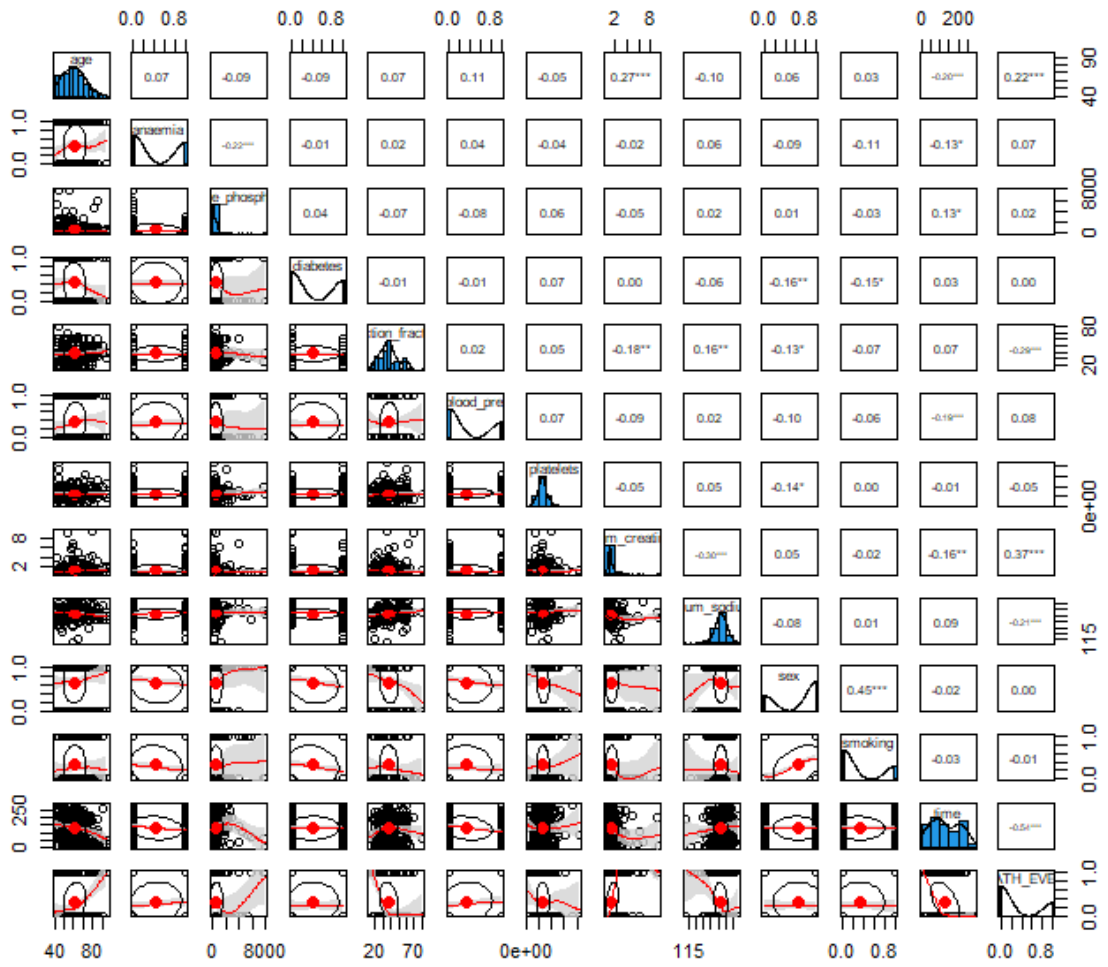

Figure: Shows the correlation between the variables of the dataset

Below is Q-Q plot, where in a Q-Q plot, the quantiles of the data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the points on the plot will fall approximately along a straight line. In the case of the "diabetes" and "death_event" variables, a Q-Q plot could be used to assess whether the distribution of "diabetes" or "death_event" is approximately normal.

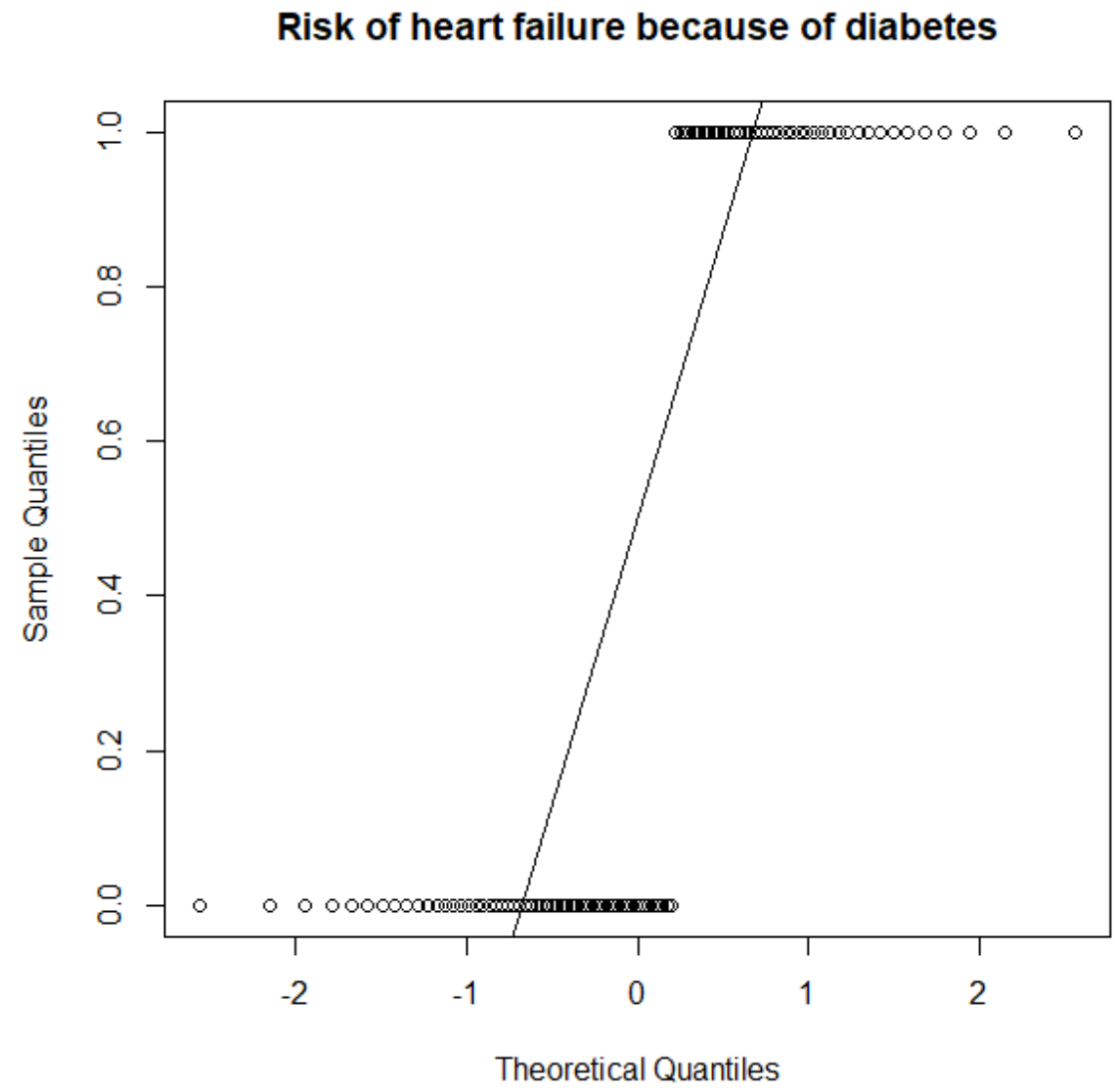

Figure: shows Q-Q plot plotted against ‘diabetes’ and ‘death-event’

Below figure shows the distribution of heart failure data which correlates patients with diabetes have the risk of heart failure of not

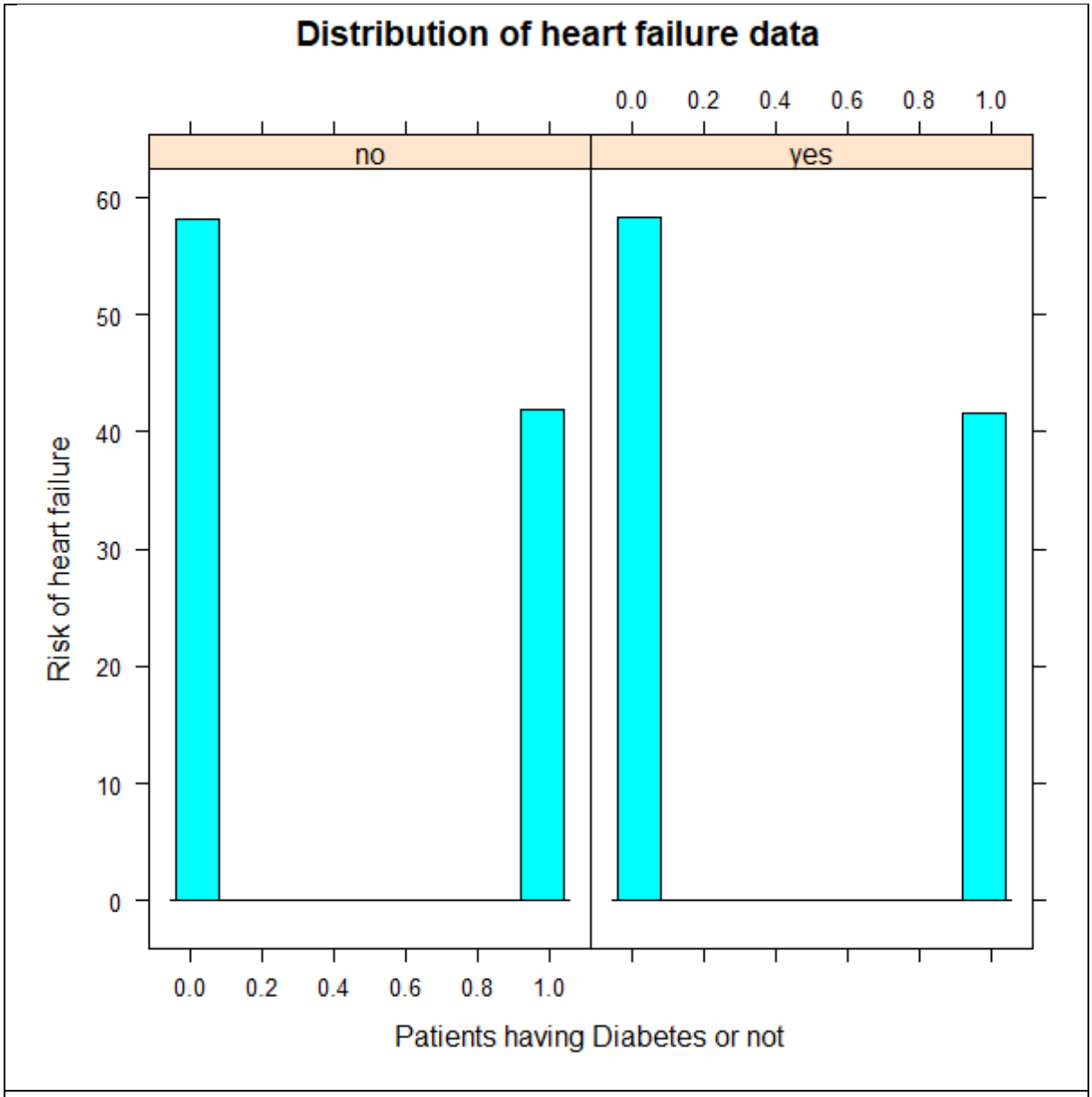

Figure: Histogram of heart failure data for diabetic patients having risk of heart failure or not

We may apply a statistical test to see whether there is a significant correlation between the two variables "diabetes" and "death_event" in order to test the hypothesis that individuals with diabetes are connected to the risk of heart failure. A contingency table displaying the observed frequencies of 'death_event' for patients with and without diabetes would be first created in order to do the chi-square test. Then, assuming there is no correlation between the two variables, we would compute the predicted frequencies for each column in the table. Finally, we would compute the test statistic and accompanying p-value using the chi-square test formula.

RESULTS

p-value obtained from Chi-squared test : p-value = 1

Wilcox test p-value = 0.9739

CONCLUSION

The alternative hypothesis (H1) in this example was that there is a link between "diabetes" and "death_event," contrary to the null hypothesis (H0), which stated that there is no association. The chi-square test's p-value was 1, indicating that we did not successfully reject the null hypothesis and that there is insufficient data to support a link between the two variables. In other words, we cannot conclude that patients with diabetes are related to the risk of heart failure based on the data and the statistical test that we performed

HYPOTHESIS TESTING Q2

H_0: There is no relationship between age and the likelihood of experiencing a death event.

H_1: There is relationship between age and the likelihood of experiencing a death event.

The variable "age" and the variable "death_event" have no association, according to the null hypothesis (H0) for the study topic. This suggests that a death event can happen regardless of the patient's age. In other words, age has no bearing on the likelihood of witnessing a death experience. On the other hand, the competing hypothesis (H1) asserts that there is a connection between the variables "age" and "death_event". This indicates that there may be a correlation or reliance between the two factors, and that age may affect the likelihood of encountering a death event. In other words, the chance of observing an increased risk event may be influenced by a person's age.

Name	Description	Type	Comments
Age	Patient age	Continuous Independent variable	Variable is continuous and need not be converted
Death event	If the patient died during the follow up period	Categorical Dependent variable	Variable is categorical and need not be converted

STATISTICAL TESTING Q2

We may apply a statistical test to see whether there is a significant correlation between the two variables "age" and "death_event" in order to test the hypothesis that the individual’s age is connected to the risk of heart failure.

Below figure shows the correlation between the Age and Death_Event variables of the dataset

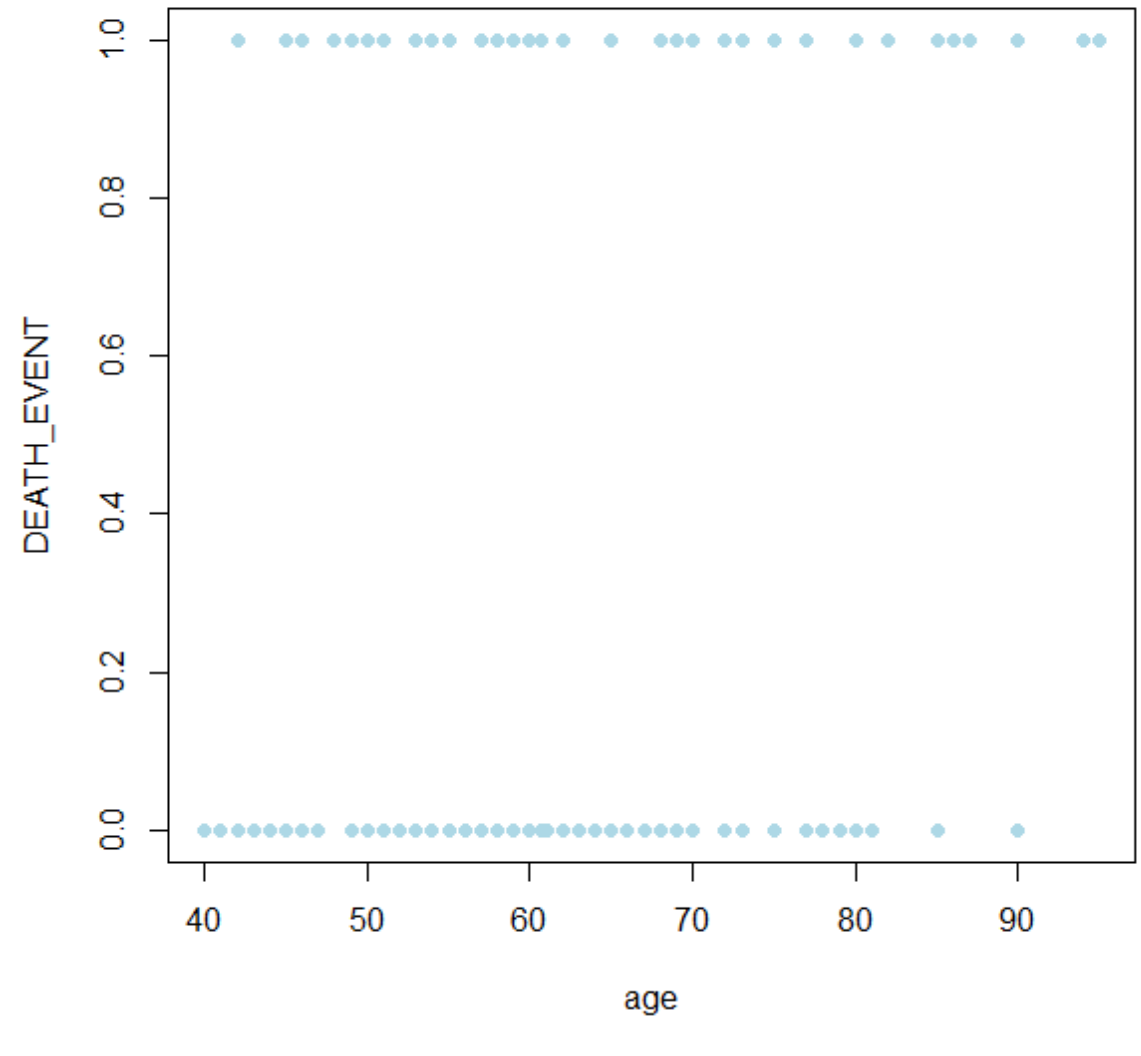

Figure: shows the correlation between the variables ‘age’ and ‘death_event’ of the dataset

Below is Q-Q plot, where in a Q-Q plot, the quantiles of the data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the points on the plot will fall approximately along a straight line. In the case of the "age" and "death_event" variables, a Q-Q plot could be used to assess whether the distribution of "age" or "death_event" is approximately normal.

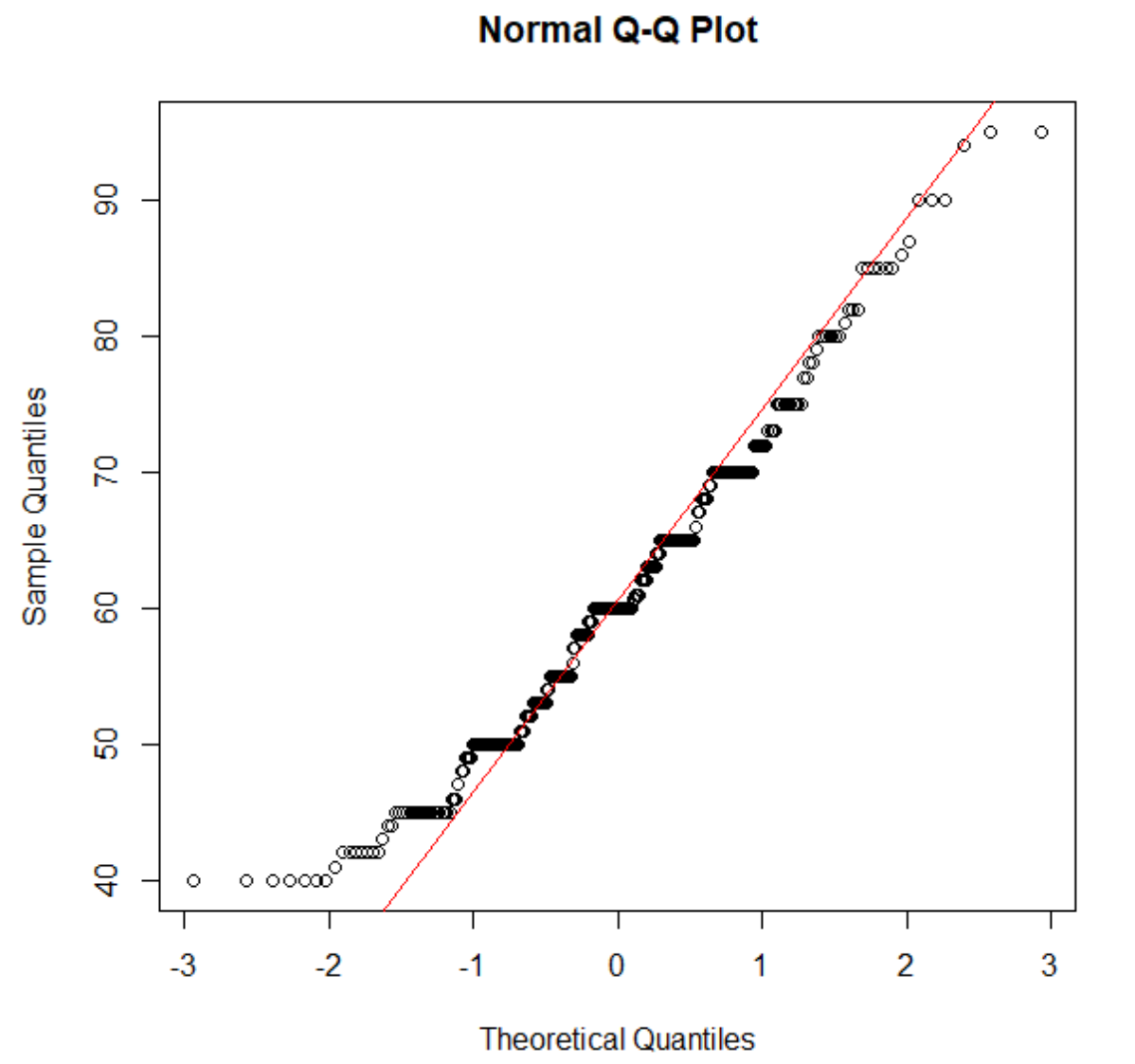

Figure: Q-Q plot against ‘age’ and ‘death_event’ variables

Below figure shows the distribution of heart failure data which correlates patients age to have the risk of heart failure of not

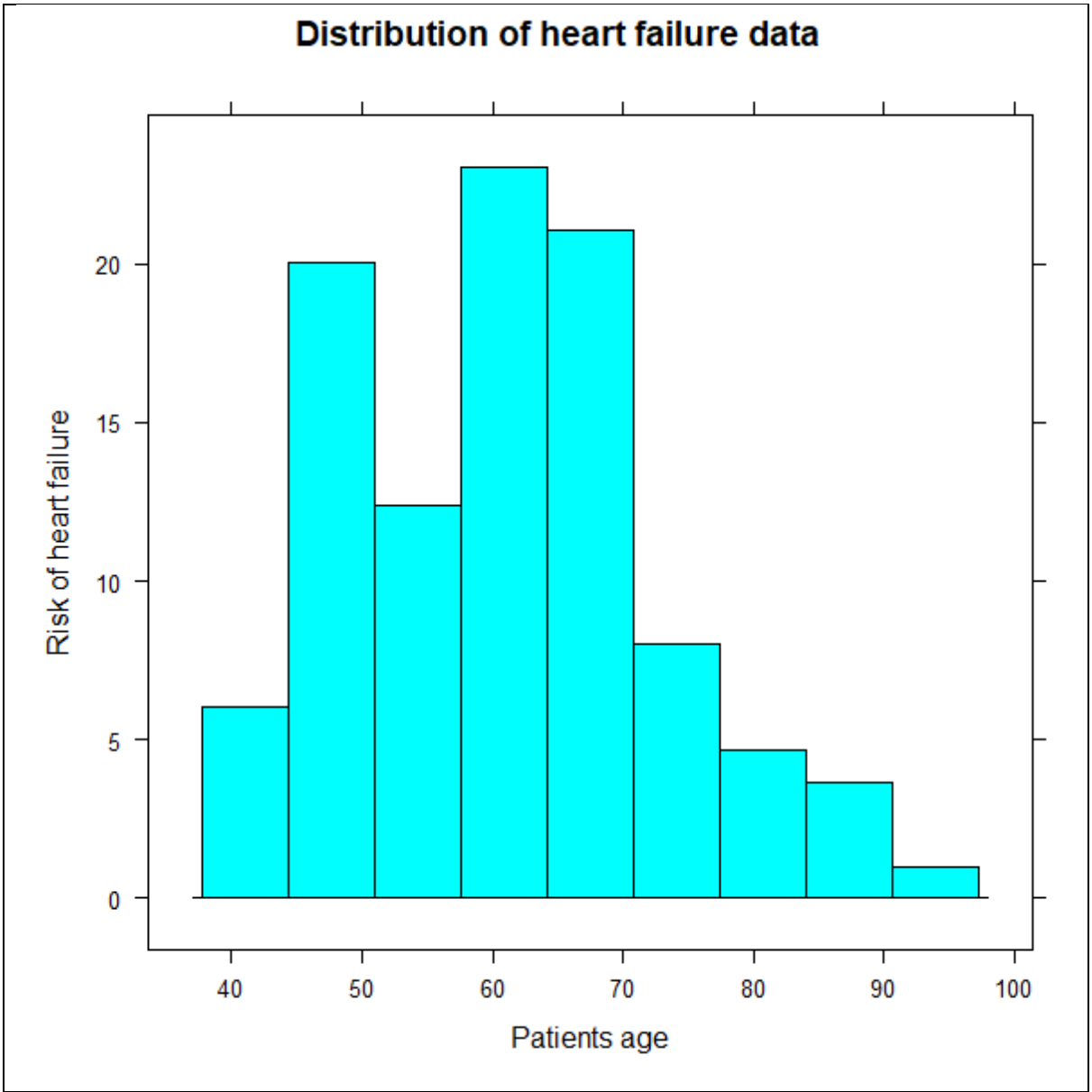

Figure: Histogram of heart failure data to check if patients age related to having risk of heart failure

RESULTS

p-value for Chi-squared test is p-value = 0.01523

CONCLUSION

A p-value of 0.01523 suggests that there is some evidence against the null hypothesis. Specifically, it means that if the null hypothesis were true (i.e., if there were truly no relationship between age and the likelihood of experiencing a death event). Therefore, we can reject the null hypothesis at a significance level of 0.05 (i.e., a commonly used threshold for statistical significance). This suggests

that there is a statistically significant relationship between age and the likelihood of experiencing a death event.

HYPOTHESIS TESTING Q3

H_0: There is no relationship between ejection fraction and the likelihood of experiencing a death event.

H_1: There is a relationship between ejection fraction and the likelihood of experiencing a death event.

Here we are trying to understand whether there is a relationship between ejection fraction and the likelihood of experiencing a death event. The null hypothesis (H0) posits that there is no relationship between ejection fraction and the likelihood of experiencing a death event, while the alternative hypothesis (H1) posits that there is a relationship between the two variables.

Name	Description	Type	Comments
Ejection fraction	Percentage of blood leaving the heart each time it contracts	Continuous Independent variable	Variable is categorical and need not be converted
Death event	If the patient died during the follow up period	Categorical Dependent variable	Variable is categorical and need not be converted

STATISTICAL TESTING Q3

We may apply a statistical test to see whether there is a significant correlation between the two variables "ejection_fraction" and "death_event" in order to test the hypothesis that individual's ejection fraction are connected to the risk of heart failure.

Below figure shows the correlation between the variables of the dataset

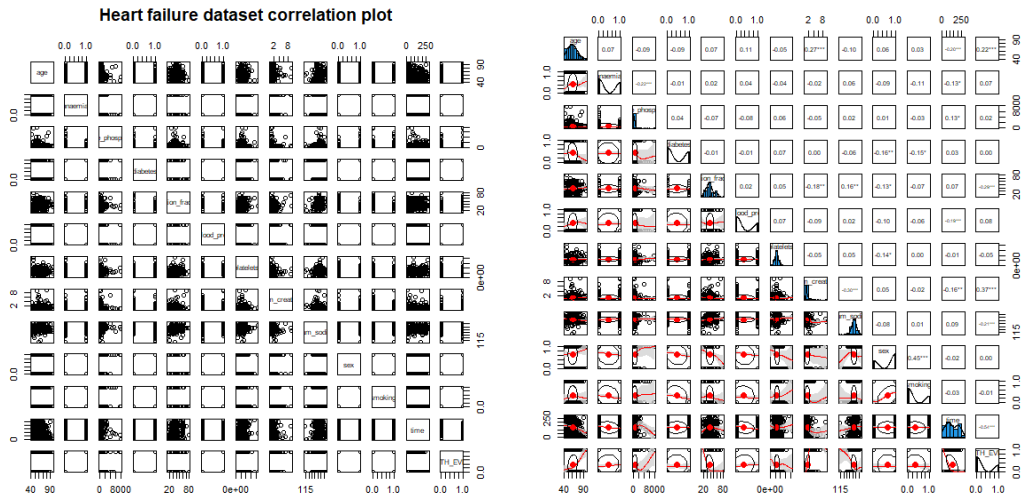

Figure: shows the correlation between the variables of the dataset

Below figure shows the correlation between the ejection_fraction and death_event variables of the dataset

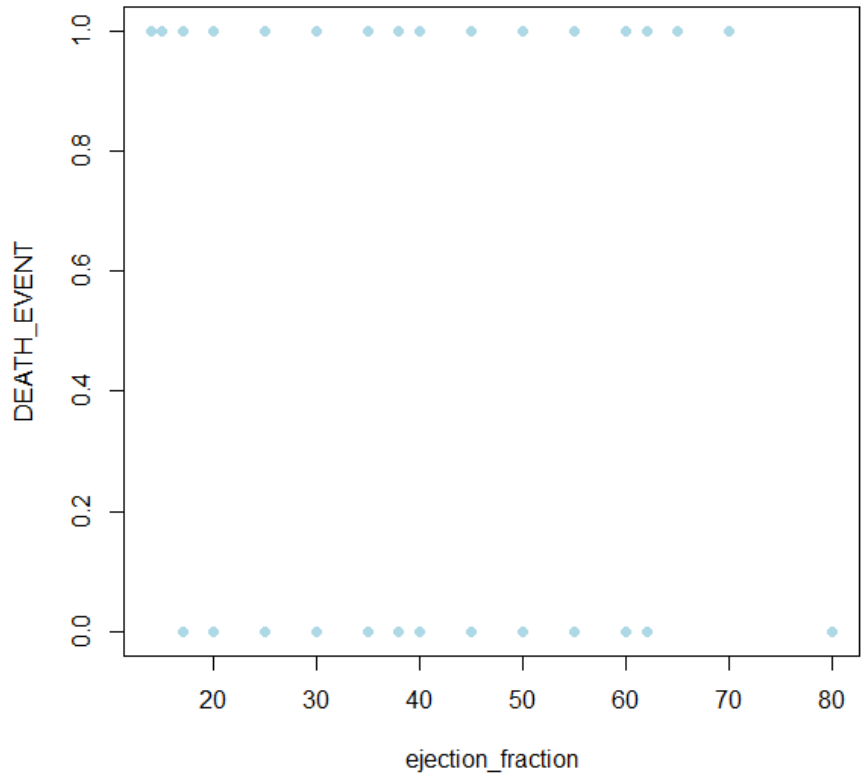

Figure: Below figure shows the correlation between the ejection_fraction and death_event variables of the dataset

Below figure shows the distribution of heart failure data which correlates patients ejection fraction to have the risk of heart failure of not

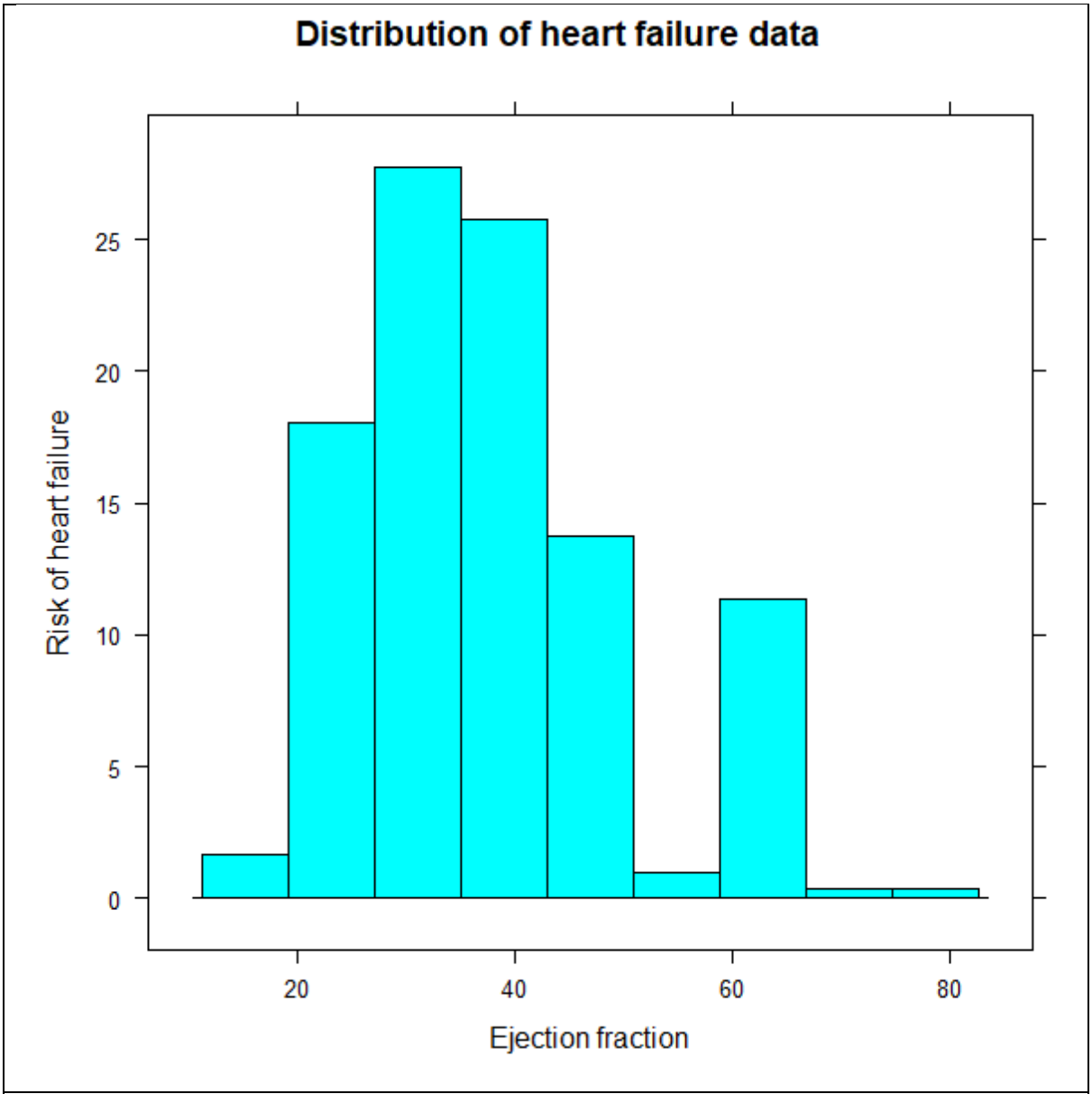

Figure: Histogram of heart failure data for patients having ejection fraction having risk of heart failure or not

Below is Q-Q plot, where in a Q-Q plot, the quantiles of the data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the points on the plot will fall approximately along a straight line. In the case of the "ejection_fraction" and "death_event" variables, a Q-Q plot could be used to assess whether the distribution of "ejection_fraction" or "death_event" is approximately normal.

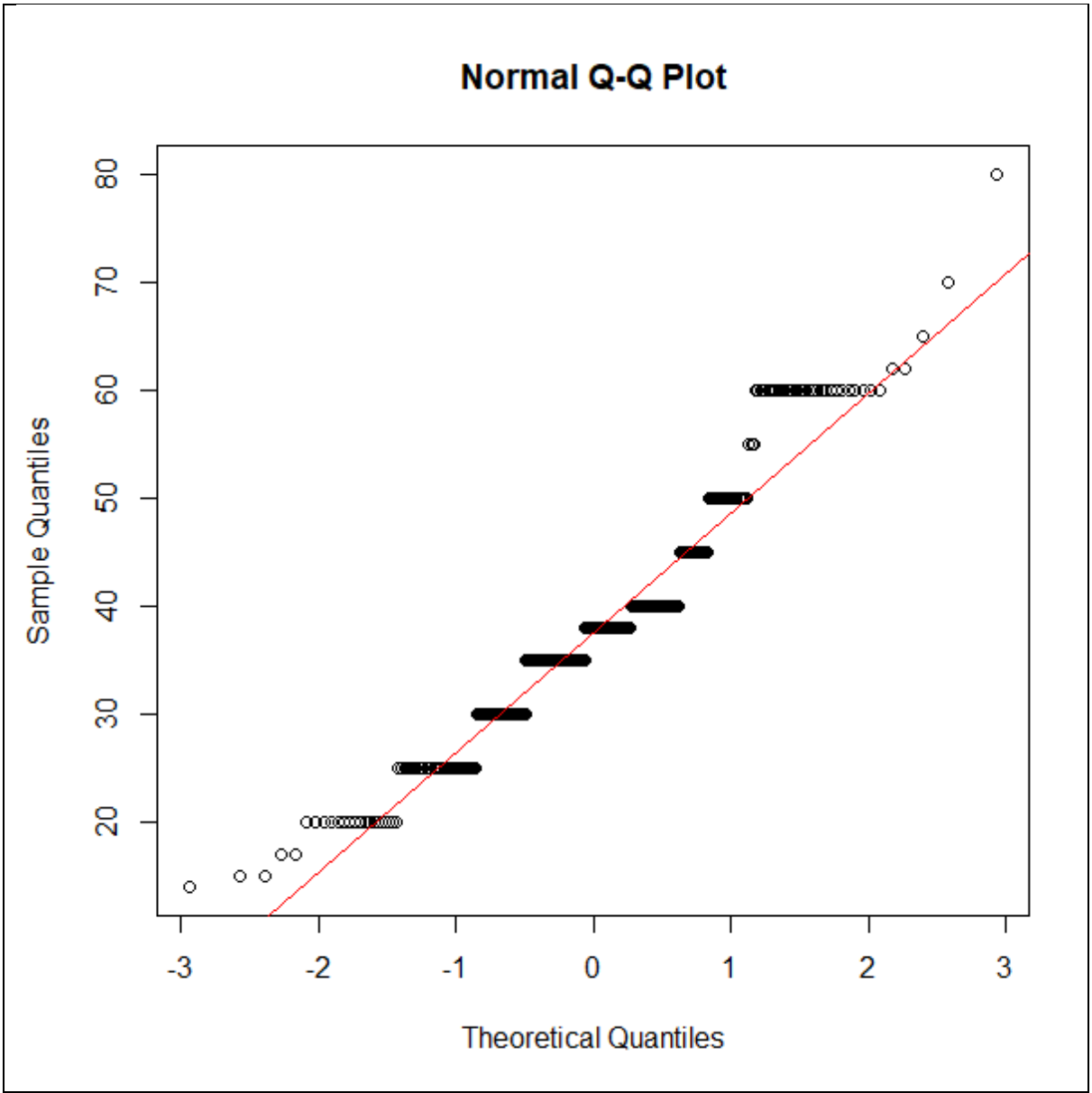

Figure: shows Q-Q plot plotted against 'ejection_fraction' and 'death-event'

RESULTS

p-value for the chi-squared test is p-value = 6.459e-08

CONCLUSION

A p-value of 6.459e-08 (very small value) obtained from the chi-squared test indicates strong evidence against the null hypothesis (H0) and suggests that there is a significant relationship between ejection fraction and the likelihood of experiencing a death event. Therefore, we reject the null hypothesis and accept the alternative hypothesis (H1).

In other words, the evidence suggests that ejection fraction is associated with the likelihood of experiencing a death event.

HYPOTHESIS TESTING Q4

H_0: There is no relationship between platelets and the likelihood of experiencing a death event.

H_1: There is a relationship between platelets and the likelihood of experiencing a death event.

Here, using statistical method to evaluate strength of evidence against a null hypothesis. The null hypothesis (H0) in this case is that patient's platelets count are not related to the risk of heart failure, while the alternative hypothesis (H1) is that they are related. We are using variables 'platelets' and 'DEATH_EVENT' for the hypothesis.

Name	Description	Type	Comments
platelets	Number of platelets in the blood.	Continuous Independent variable	Variable is continuous and will remain the same
Death event	If the patient died during the follow up period	Categorical Dependent variable	Variable is categorical and need not be converted

STATISTICAL TESTING Q4

We may apply a statistical test to see whether there is a significant correlation between the two variables "platelets" and "death_event" in order to test the hypothesis that the individual's platelet count is connected to the risk of heart failure.

Below figure shows the correlation between the platelets and Death_Event variables of the dataset

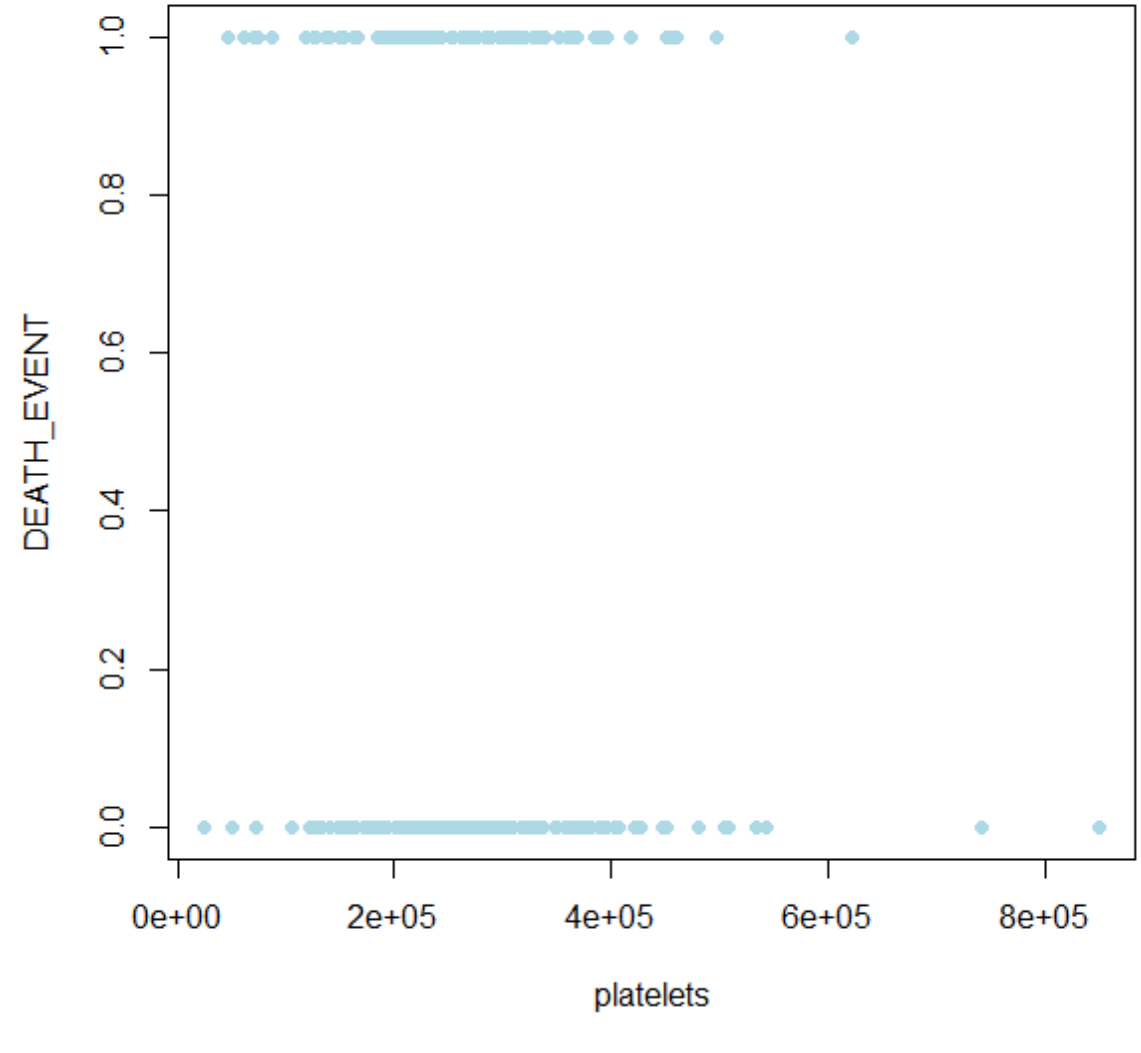

Figure: shows the correlation between the variables ‘platelets’ and ‘death_event’ of the dataset

Below is Q-Q plot, where in a Q-Q plot, the quantiles of the data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the points on the plot will fall approximately along a straight line. In the case of the "platelets" and "death_event" variables, a Q-Q plot could be used to assess whether the distribution of "platelets" or "death_event" is approximately normal.

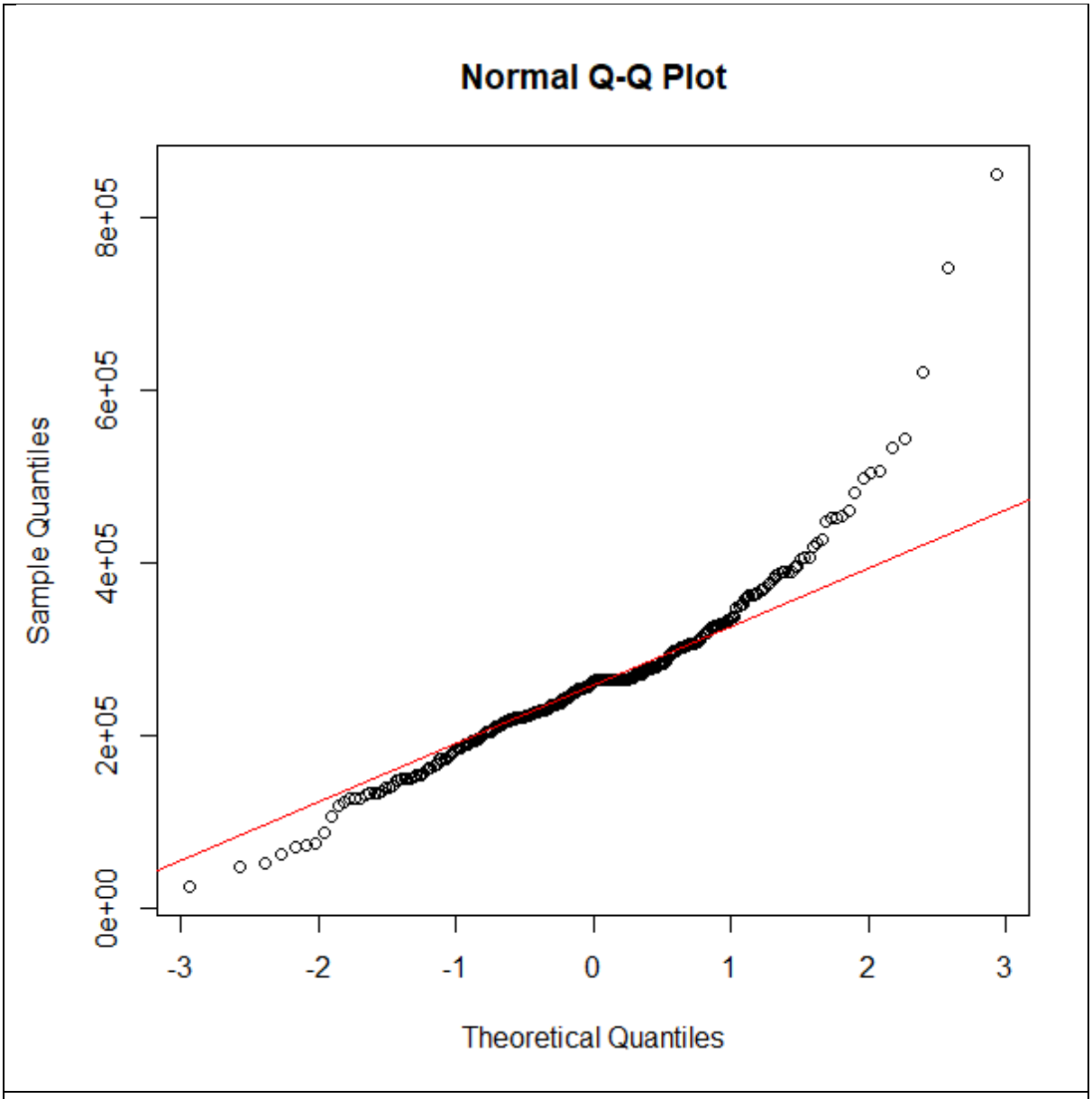

Figure: Q-Q plot against ‘platelets’ and ‘death_event’ variables

Below figure shows the distribution of heart failure data which correlates patient's platelets count to have the risk of heart failure of not

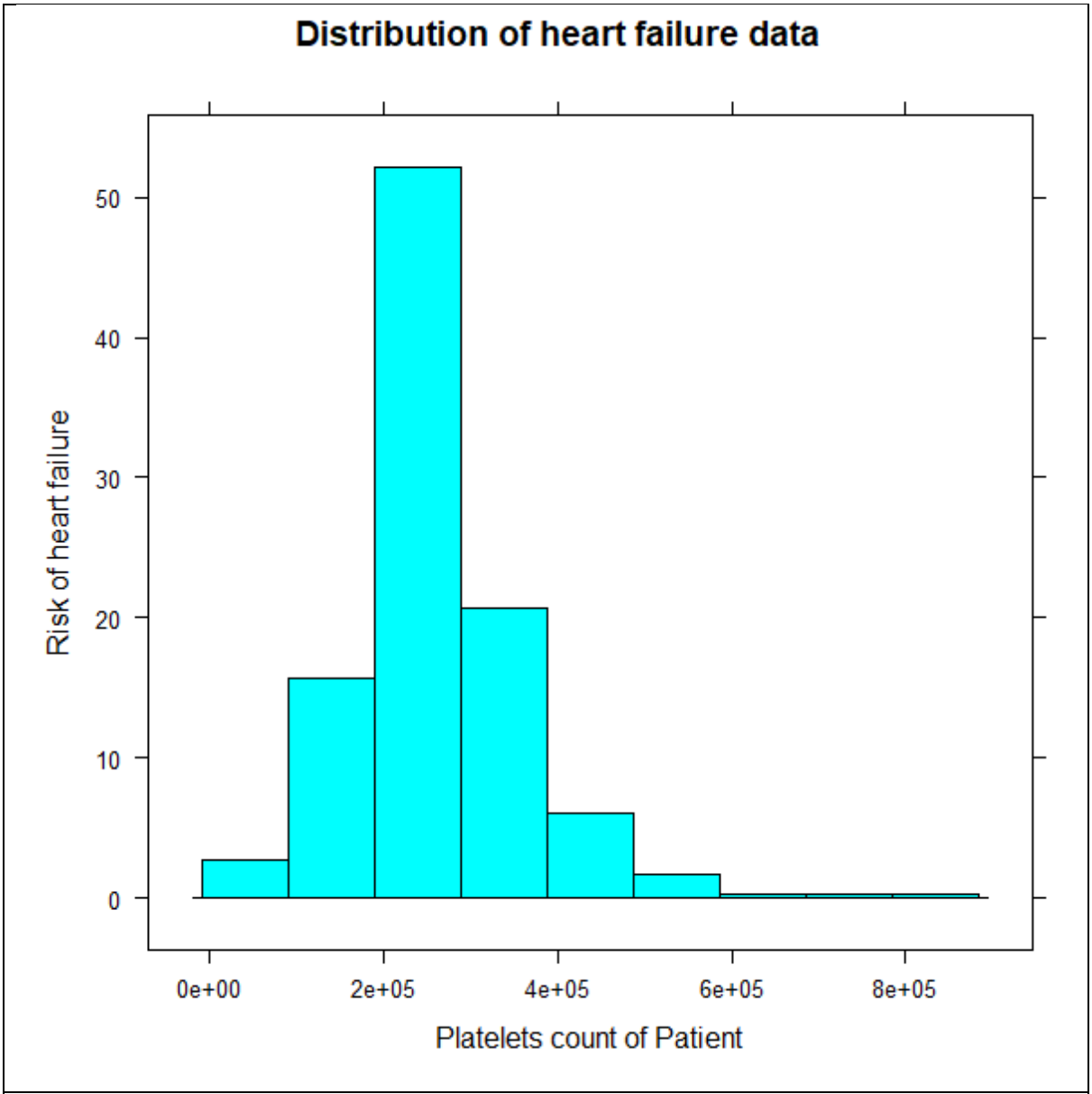

Figure: Histogram of heart failure data to check if patient's platelets count related to having risk of heart failure

RESULTS

Wilcox p-value = 0.4256

CONCLUSION

A p-value of 0.4256 suggests that there is no significant difference between the two groups being compared.

Based on this result, we can conclude that there is no evidence to suggest that the variable being tested has a significant impact on the outcome being analysed.

HYPOTHESIS TEST Q5

H_0: The length of time since diagnosis of heart failure has no impact on the likelihood of experiencing a death event.

H_1: The length of time since diagnosis of heart failure does impact the likelihood of experiencing a death event.

The null hypothesis (H0) is that the length of time since diagnosis of heart failure has no impact on the likelihood of experiencing a death event, which means there is no difference in mean age between patients who experienced a death event and those who did not. The alternative hypothesis (H1) is that the length of time since diagnosis of heart failure does impact the likelihood of experiencing a death event, which means there is a significant difference in mean age between the two groups.

Name	Description	Type	Comments
Time	Follow-up period	Continuous Independent variable	Variable is continuous and will remain the same
Death event	If the patient died during the follow up period	Categorical Dependent variable	Variable is categorical and need not be converted

STATISTICAL TEST Q5

We may apply a statistical test to see whether there is a significant correlation between the two variables "time" and "death_event" in order to test the hypothesis that the individual's follow up time is connected to the risk of heart failure.

Below figure shows the correlation between the time and death_event variables of the dataset

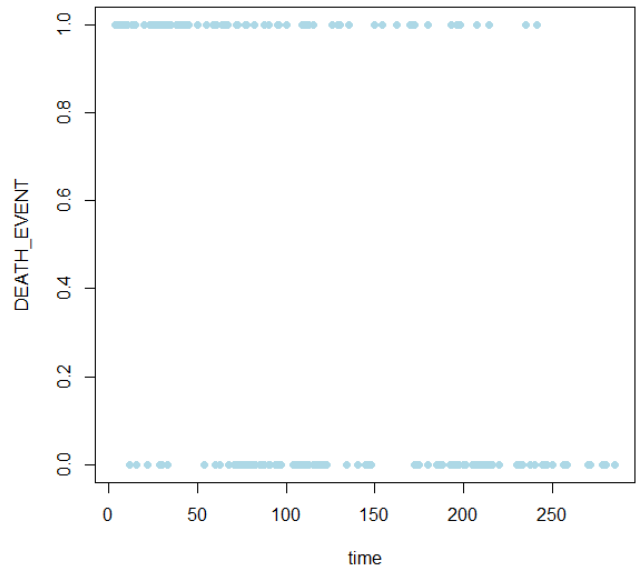

Figure: shows the correlation between the variables 'time' and 'death_event' of the dataset

Below is Q-Q plot, where in a Q-Q plot, the quantiles of the data are plotted against the quantiles of a theoretical normal distribution. If the data is normally distributed, the points on the plot will fall approximately along a straight line. In the case of the "time" and "death_event" variables, a Q-Q plot could be used to assess whether the distribution of "time" or "death_event" is approximately normal.

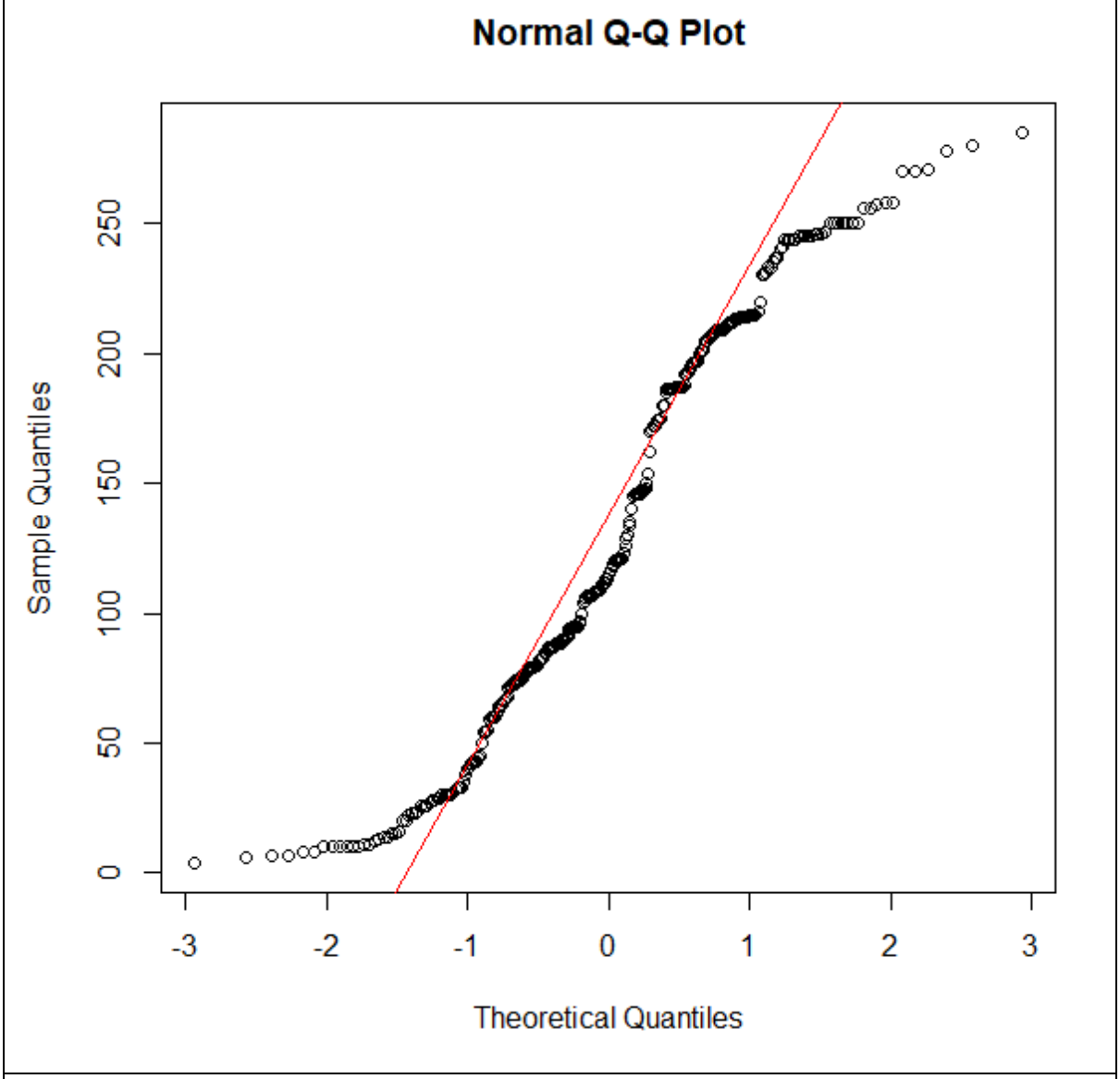

Figure: Q-Q plot against ‘time’ and ‘death_event’ variables

Below figure shows the distribution of heart failure data which correlates patients follow up time to have the risk of heart failure of not

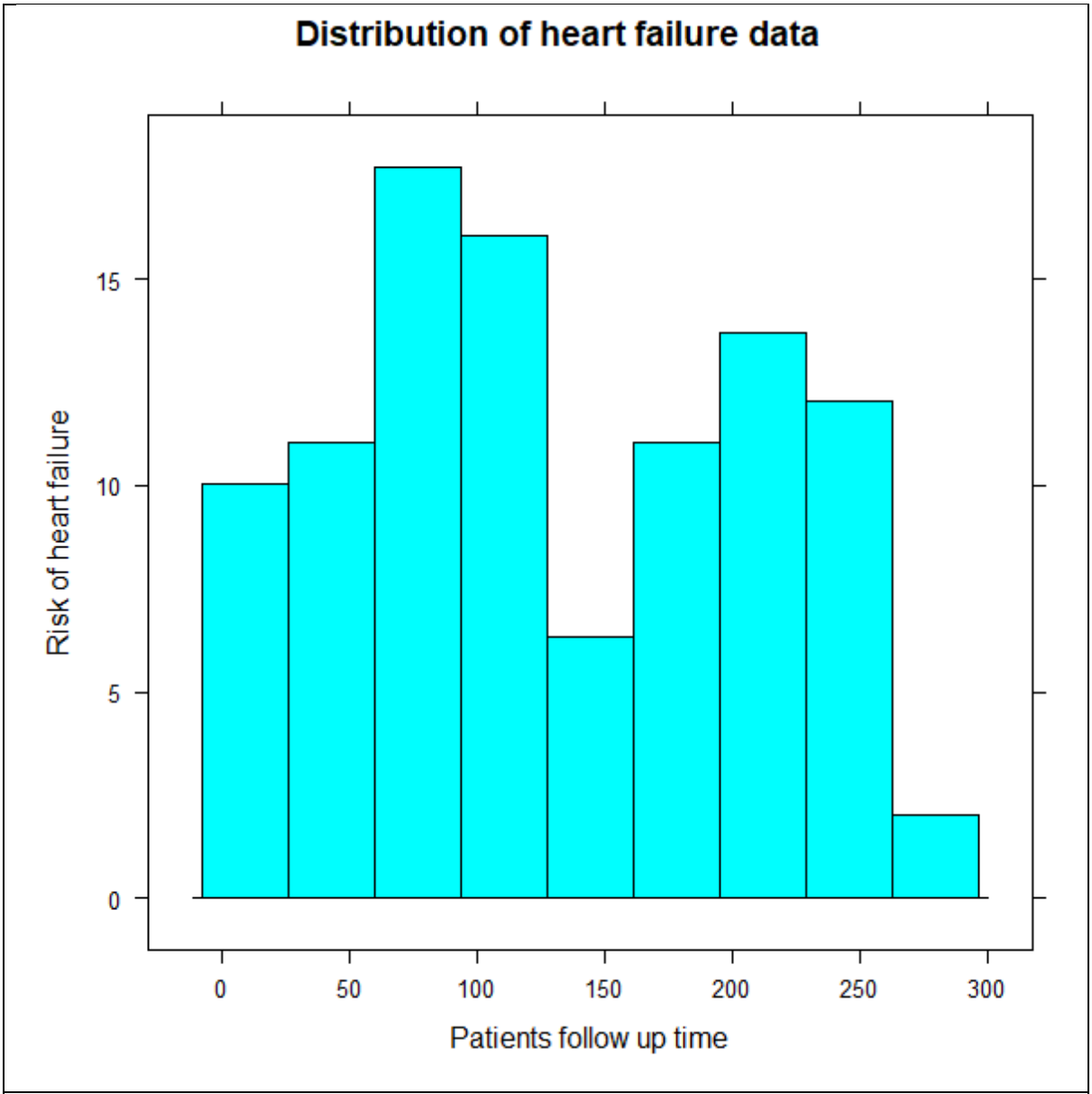

Figure: Histogram of heart failure data to check if patients follow up time related to having risk of heart failure

RESULTS

Chi-squared p-value = 6.59e-07

CONCLUSION

A p-value of 6.59e-07 suggests strong evidence against the null hypothesis (H0) and in favour of the alternative hypothesis (H1). We can conclude that there is a significant association between the length of time since diagnosis of heart failure and the likelihood of experiencing a death event.