

**School of InfoComm Technology**

**Data Exploration & Analysis Assignment**

Diploma in DS

April 2022 Semester

**ASSIGNMENT 1**

(30% of DEA Module)

23rd May 2022 – 10th June 2022

**Submission Deadline:**

**Presentation: 10th Jun 2022 (Friday), 11:59PM**

**Excel files: 10th Jun 2022 (Friday), 11:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **T01 / T02 / T03** |
| **Student Name** | **:** | **Chong Xin Le** |
| **Student Number** | **:** | **S10221973** |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 19th Jun 2022 (Sunday), 11:59PM.

Table of Contents

[Abstract 3](#_Toc105598017)

[Data Cleaning 4](#_Toc105598018)

[Univariate analysis 5](#_Toc105598019)

[Bivariate analysis 6](#_Toc105598020)

[Multivariate analysis 12](#_Toc105598021)

[Summary 14](#_Toc105598022)

### Abstract

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels including, coronary heart disease (heart attacks), cerebrovascular diseases (strokes), heart failure (HF), and other types of pathology. Altogether, cardiovascular diseases cause the death of approximately 17 million people worldwide annually, with fatalities figures on the rise for first time in 50 years the United Kingdom. (Siddique, 2019) Heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases. (National Heart,Lung, Blood Institute, 2022)

In this report, I analyse a dataset of 299 patients that contains 11 features which can be used to predict heart failure. And at the end of this report, my aim was to come out with findings that could provide insights on the variables that could influence heart failure risk.

Starting off, I applied several methods learnt in excel and other resources to predict the survival of each patient having heart failure symptoms and to detect the most important risk factors that may lead to heart failure. By first cleaning the data, I proceeded to make pivot tables and charts. Using these charts, I examined the datasets made in different applications to identify informative variables and their interpretations.

From the various Univariate, Bivariate and Multivariate analysis technique, I was able to come out with charts to predict the patient’s survival and rank the variables according to the one with the highest risk factors. I also created a correlation matrix which allowed me to determine correlations between the variables against our target variable and then prove them using the charts I have created.

Our results showed not only that serum creatinine, ejection fraction, serum sodium and age are the most predictive clinical features of the dataset and are sufficient to predict patients’ survival. While sex, diabetes and smoking were found as the least significant risk factors for mortality among heart failure patients.

This discovery has the potential to impact clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not. Medical doctors may now focus mainly on serum creatinine, ejection fraction, serum sodium and age to predict and understand if a patient will survive after heart failure.

# Data Cleaning

1. Standardizing Data Values

Almost 99% of the values from the ‘age’ feature were whole numbers as seen from the given dataset. After examining the data, I realise that some numbers contain decimals such as ‘60.667’.

Using excels find and replace function I was able convert the values to a whole number. The values for age were all rounded down as we assumed that they passed away on that given age hence it would be more suitable to round down since they have not reached the age if we were use the rounded-up value.

Table

Description automatically generatedTable

Description automatically generated

After changes

1. Replacing errors

There is an underscore in every value of ‘Platelets’ feature which is incorrect. This can be easily corrected in excel by also using the find and replace function to replace ‘\_’ with ‘ ’. We can see how our column now looks like from the pictures below.

Text

Description automatically generated Table

Description automatically generated

After changes

# Univariate analysis

1. Death Event Breakdown

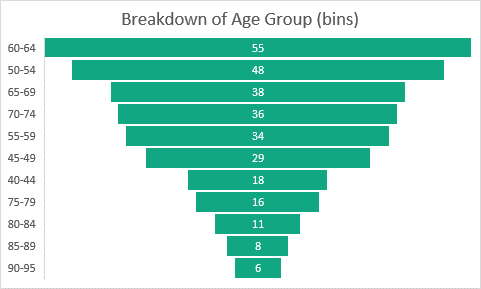
Below is a pie chart made in excel, like a pie chart, a doughnut chart allows us to get a glance of the data in our chart quickly, but it is better used when the only displaying a small number of categories.

Observing this chart, we can see that deaths caused by heart failure take up 32% of our data.

1. Breakdown of Age Groups

Getting an overview of the different ages we have in our dataset which allows us to know what we are working with when doing data exploration. See how most of the people in the dataset falls within the middle age group of 50-74 while the older age group from 75-95 takes up only a small amount.

A funnel chart is very useful here where we want to quickly make sense of trends in data. Here, we only want to look at the ages with largest and smallest population. We are not looking into the ones in the middle which makes a funnel chart the most suitable graph in this case



# Bivariate analysis

1. Death Event by Age

I created a histogram with group bar charts in excel by creating bins of 5 for the age variable with ‘Death Event’ as the Y-axis. Using a histogram offers an insightful look at frequency distribution which in this case we can see the distribution across the different ages as well as make comparisons to our death event.

Here, we can progressively see how the number of deceased starts to overtake the number of alive from the age of 80 onwards compared to younger ages like 40-44 where the fraction of those who died due to heart failure is only 1 out of 17. It is therefore inferred that the older you get, the more likely you are to pass away due to heart failure.

1. Anaemia vs Death Event

This is a clustered column chart made in excel. It counts the number of death event which is split by whether the patient has anaemia or not. See that the death for people who does not have anaemia and has anaemia are very similar with only a difference of 4. However, for a greater number of people who did not die and does not have anaemia the percentage of death is lower compared to that of those who has anaemia and died. This shows that anaemia does affect the chances of having heart failure.

1. Creatinine Phosphokinase vs Death Event

Below is a scatter chart depicting Death Event by Creatinine Phosphokinase. A scatter plot is useful when trying to determine whether there’s a potential relationship between variables. In this case, there seems to be an absence of a relationship as the scatter plot shows that even though at different Creatinine Phosphokinase levels the number of alive and not alive patients are distributed almost equally as well. However, at Creatinine Phosphokinase 582 mcg/L, there is a spike in both the alive and not alive though the spike of not alive is much higher at 30. Though the spike only appears at this level, it is still important as it may be why there is an increase in heart failure.

1. Death Event by Diabetes

Below is a clustered column chart depicting Death Event and Diabetes.

From this chart:

% of people who died but does not have diabetes – 56/118 x 100 = 47.4%

% of people who died and have diabetes – 45/85 x 100 = 47%

This implies that diabetes is likely not a cause of heart failure since the percentage of people who died and has diabetes is almost the same as people who does not have diabetes but still pass away from heat failure. While this also mean diabetes has a very weak relationship with Death Event.

1. Platelets vs Death Event

Generally, the plots in this chart are almost parallel to the x-axis. This means Platelet variable has almost no or a weak relationship with Death Event. However, from the scatter plot the value count of patients who are not alive are slightly higher than those who are alive for the same platelet amount. This could mean having a higher platelet may have some effect on heart failure but might not be significant. Furthermore, if we hover over the points with count 15 and 10, we can see at platelet 263358.03mL the increase in alive and not alive patients increased drastically. It could just represent that more patients happen to fall in this platelet amount, hence the increase in both deceased and alive patients.

1. Death Event by High Blood Pressure

Below are two doughnut charts, one representing patients with no high blood pressure while the other representing patients with high blood pressure.

Having a higher blood pressure, does seem to affect having heart failure as shown by the charts below because 37.14% died from having high blood pressure while only 29.38% died without high blood pressure. This shows not having high blood pressure can decrease your chances of heart failure.

1. % Breakdown of Ejection Fraction against Death Event

I created a stack bar chart in excel plotting Ejection Fraction over the percentage of Death Event. A stacked bar charts, allows shows the relative percentage of multiple data series in stacked bars, where the total of each stacked bar always equal to 100%. This makes it easier for us to compare the dead and alive from every ejection fraction value but also compare within each ejection fraction. From this chart, by looking at ejection fraction of 20, it is the only value that has more than 50% of deceased patients. It is followed by the ejection fraction values of 17,25 and 62 that has 50% of patients that passed away. This could mean between 17-25 the chances of getting heart failure are higher. Although values like 14,15,65 or 80 has a 100% or 0% death rate, as the death event count is low, we cannot conclude anything as the data is too small to get an actual analysis. Thus, these values are ignored.

1. Serum Creatinine vs Death Event

Used an area chart which allows us to compare the variables easily in this case we can see which count of death event is higher by looking at the colour overlaps. We can see the chart that between 0-1.7 the number of deceased patients seem much lower compared to those who were alive. This means that this level of Serum Creatinine is considered a safe level for not getting heart failure. As we get closer to 1.8-3, we can see from the chart that more people are deceased then alive. This is hence the riskier area where being in this level may mean that the chances of heart failure are higher.

1. Serum Sodium vs Death Event

Likewise, to the chart above, we have an area chart portraying Death Event against Serum Sodium. We can easily see that from 131-134 the count of deceased patients is higher than those alive meaning patients with serum sodium levels that are within this range tend to have died from heart failure. While 135-145mEg serum sodium seems to be the range where more patients are alive. This would be the range we want to be in to avoid a higher risk of heart failure.

1. Death Event by Sex

A clustered column chart allows for direct comparison of multiple series in each category, which in this case makes it easier for us to compare the death event with sex. From the bar chart,

Female % death: 34/71 x 100 = 47.9%

Male % death: 62/132 x 100 = 47.0%

The percentage of death for both genders is almost similar, this means there is almost no relationship of sex with death event. Hence, being a female or male does not really mean your chances of dying from heart failure is greater or smaller.

1. Death Event by Smoking

The ratio of people who died from smoking is slightly lesser than the people who died but did not smoke. This does not mean that smoking helps to decrease the chances of heart failure, but it could mean that smoking has no correlation with the risk of heart failure. Even though this chart does not provide meaningful insights to us, research by the World Health Organization have shown that tobacco use is one of the most important risk factors of heart disease. Furthermore, the effect of behavioural risk factors may show up as high blood pressure, obesity, and many others. (World Health Organization, 2021) This could be a reason why smoking has no correlation with the death event as it is closely related to high blood pressure which is in turn closely related to the death event

# Multivariate analysis

1. Correlation heatmap

Using Jupyter Notebook, I was able to code out a correlation heatmaps that can be used to find potential relationships between variables and to understand the strength of these relationships.

Over here, we can see that the variables that have the stronger relationship with our target variable is ejection fraction, serum creatinine, serum sodium and age (-0.27, 0.29, -0.2, 0.25). Based on our Bivariate charts, we can also see that this is true.

It can also be seen that sex, diabetes and smoking has a non-linear relationship from the heatmap. This can also be proven by the charts we have created for these variables against the Death Event.

Thus, this heat map provides us a summarized version of the correlation between each of our variable which we can then prove this correlation with the charts created in this report.

Chart, treemap chart

Description automatically generated

Code used to generate correlation heatmap in Jupyter Notebook :

Text, application

Description automatically generated

1. Breakdown of people who passed away and have an illness

Pie chart below shows the breakdown of the number of deaths caused by Boolean variables. A pie chart is useful here because circle size can be made proportional to the total quantity it represents which makes it visually easier to get a glimpse of information.

We can see from the pie chart that high blood pressure has caused the most deaths due to heart failure while smoking has caused the least and diabetes and anaemia ranges around the middle. Now we can know what the larger causes of heart failure are and what may not have so much effect on having a heart failure.

# Summary

In this report, the fact that our charts have shown that ejection fraction, serum sodium, serum creatinine and age as the most relevant features confirmed the relationship between our variable with the target variable in our correlation heatmap. Likewise, features that had no relationship with mortality by heart failure could also be proven by both the heatmap as well as the charts that we have created.

Like always, a dataset comes with its limitation and in this case, we are only working with 299 patients. It is a small set of data and the findings I have found in this report may only apply to this data. A larger dataset would have permitted us to obtain more reliable results and additional information about the physical features of patients like its height, body mass would have been useful to detect additional risk factors for cardiovascular heart disease.

Nonetheless, with our current given dataset, our results have shown that it might be possible to predict the survival of patients with heart failure solely from ejection fraction (EF at 20%, higher chances of heart failure), serum sodium (levels 131-134 are riskier), serum creatinine (levels 1.8-3 are riskier) and age (the older, the higher the risk). Features like Creatinine Phosphokinase, Anaemia, Platelets could be used to predict heart failure but may not give significant results. However, I have also found that out of all the category variables that were encoded with Boolean values it seems that most of the patients that passed away from heart failure had Anaemia. This was closely followed up by diabetes and high blood pressure. Although, this does not prove much it could still be an underlying cause of mortality by heart failure.

Moreover, features such as diabetes, smoking and sex are seen to have almost no or little relevance to the mortality by heart failure. This is shown by the percentages of people who died having or not having diabetes, smoking, or not smoking and being a female or male are very much similar.  The reason behind why these features may not be significant is because smoking and diabetes are basically causes of heart problem at initial stages. In addition, the data provided were of advanced stages of heart failure. Up to these stages, these factors (diabetes and smoking) may probably be controlled by medications and hence these factors do not have significant effect on deaths due to heart failure. (Tanvir Ahmad, 2017)

Knowing the most important risk factors can be particularly encouraging for the hospital settings: in case many laboratories test results and clinical features were missing from the electronic health record of a patient, doctors could still be able to predict patient survival by just analysing the ejection fraction, serum sodium, serum creatinine and age value.

# References

National Heart,Lung, Blood Institute. (2022). *What Is Heart Failure?* Retrieved from https://www.nhlbi.nih.gov/health/heart-failure

Siddique, H. (2019). *UK heart disease fatalities on the rise for first time in 50 years.* Retrieved from https://www.theguardian.com/society/2019/may/13/heart-circulatory-disease-fatalities-on-rise-in-uk

Tanvir Ahmad, A. M. (2017). *Survival analysis of heart failure patients: A case study.* Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001#sec009

World Health Organization. (2021). *Cardiovascular diseases (CVDs).* Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)