

# 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** | **:** | **T02** |
| **Student Name** | **:** | Lim Wee Liang Kelven |
| **Student Number** | **:** | S10221788K |

Contents

[School of InfoComm Technology 1](file:///C:\Users\Kelve\Downloads\Year%202.1\DEA\ASG1\Q3%20Heart%20Failure%20Prediction\Q3_report_Lim%20Wee%20Liang%20Kelven_S10221788K.docx#_Toc105680887)

[Overview of Report 4](#_Toc105680888)

[Problem Understanding 4](#_Toc105680889)

[Data Exploration 5](#_Toc105680890)

[Univariate Analysis 5](#_Toc105680891)

[Bivariate Analysis 6](#_Toc105680892)

[Others 11](#_Toc105680893)

[Summary 12](#_Toc105680894)

[Appendices 14](#_Toc105680895)

[Univariate 1 14](#_Toc105680896)

[Univariate 2 14](#_Toc105680897)

[Univariate 3 14](#_Toc105680898)

[Univariate 4 15](#_Toc105680899)

[Univariate 5 15](#_Toc105680900)

[Univariate 6 15](#_Toc105680901)

[Univariate 7 16](#_Toc105680902)

[Bivariate 1 16](#_Toc105680903)

[Bivariate 2 16](#_Toc105680904)

[Bivariate 3 17](#_Toc105680905)

[Bivariate 4 17](#_Toc105680906)

[Bivariate 5 17](#_Toc105680907)

[Bivariate 6 18](#_Toc105680908)

[Bivariate 7 18](#_Toc105680909)

[Bivariate 8 18](#_Toc105680910)

[Bivariate 9 19](#_Toc105680911)

[Bivariate 10 19](#_Toc105680912)

[Bivariate 11 19](#_Toc105680913)

[Bivariate 12 20](#_Toc105680914)

[Bivariate 13 20](#_Toc105680915)

[Bivariate 14 20](#_Toc105680916)

[Bivariate 15 21](#_Toc105680917)

[Bivariate 16 21](#_Toc105680918)

[Bivariate 17 21](#_Toc105680919)

[Bivariate 18 22](#_Toc105680920)

[Bivariate 19 22](#_Toc105680921)

[Bivariate 20 22](#_Toc105680922)

[Bibliography 23](#_Toc105680923)

# Overview of Report

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. In Singapore, nineteen people die from cardiovascular diseases such as heart diseases and stroke every day. CVD accounted for 31.7% of all deaths in Singapore in 2020. This means that almost 1 out of 3 deaths in Singapore is due to CVD. Among all CVD deaths, heart failure is a common cause.

This report will cover the findings from the exploration of factors of CVD deaths and any relationships discovered.

These factors include age, gender, presence of anaemia, presence of diabetes, presence of high blood pressure (HBP), presence of smoking, ejection fraction, and levels of creatinine phosphokinase and platelets in the blood.

This report will explain the choices of visualisations and the critical findings from the visualisations from a univariate, bivariate, and multivariate perspectives.

This report will include any references from third-party websites to support its findings.

# Problem Understanding

Before making any visualisations, we must explore the data from three perspectives: univariate, bivariate, and multivariate. In univariate exploration, we can explore a factor’s distribution in a histogram or a box plot. In bivariate and multivariate exploration, by comparing one or more factors together, we can identify relationships and correlations between factors.

(Visualisations with minimal correlation and their respective explanations will not be included to no exceed the word count)

# Data Exploration

## Univariate Analysis

*(With reference to Appendices)*

Univariate 1 shows the proportion of females and males. A pie chart was aptly shows how big something is compared to the whole. From here, we can see that 105 people, or about 35%, are females, and 194 people, or about 65%, are males.

Univariate 2 shows the distribution of age. The histogram was chosen because it shows the frequency of people in a certain age group and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people are between 60 and 65 years old and is a bit right skewed. From the box plot, we can see that the minimum age is 40, the lower quartile age is 51, the median age is 60, the mean age is 60.83, the upper quartile age is 70, and the maximum age is 95.

Univariate 3 shows the distribution of creatine phosphokinase. The histogram was chosen because it shows the frequency of data in a certain range and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people have around 100 mcg/L to 199 mcg/L of creatine phosphokinase and is right skewed. From the box plot, we can see that the minimum is 23 mcg/L, the lower quartile is 115 mcg/L, the median is 250 mcg/L, the mean is 581.84 mcg/L, the upper quartile is 582 mcg/L, and the maximum is 1211 mcg/L. In the histogram, since there are values all the way to 8000 mcg/L, the mean value of the box plot is close to the upper quartile value. In the box plot, any values beyond 1211 mcg/L were removed as they were outliers.

Univariate 4 shows the distribution of ejection fracture. The histogram was chosen because it shows the frequency of data in a certain range and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people have an ejection fracture of 35% to 39%. From the box plot, we can see that the minimum is 14%, the lower quartile 30%, the median is 38%, the mean is 38.08%, the upper quartile is 45%, and the maximum is 65%.

Univariate 5 shows the distribution of platelet count. The histogram was chosen because it shows the number of people in a certain range and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people have a platelet count of 253,000 to 265,000 kiloplatelet/mL. From the box plot, we can see that the minimum is 75,000 kiloplatelet/mL, the lower quartile is 212,000 kiloplatelet/mL, the median is 262,000 kiloplatelet/mL, the mean is 263,358 kiloplatelet/mL, the upper quartile is 304,000 kiloplatelet/mL, and the maximum is 427,000 kiloplatelet/mL.

Univariate 6 shows the distribution of serum creatine. The histogram was chosen because it shows the number of people in a certain range and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people have 1 mg/dL to 1.5 mg/dL and is right skewed. From the box plot, we can see that the minimum is 0.5 mg/dL, the lower quartile is 0.9 mg/dL, the median is 1.1 mg/dL, the mean is 1.39 mg/dL, the upper quartile is 1.4 mg/dL, and the maximum is 2.1 mg/dL. In the histogram, since there are values all the way to 9.5 mg/dL, the mean value of the box plot is close to the upper quartile value.

Univariate 7 shows the distribution of serum sodium. The histogram was chosen because it shows the number of people in a certain range and the box plot was chosen because it shows the minimum, lower quartile, median, mean, upper quartile, and maximum values of the distribution. From the histogram, we can see that most people have 136 mEq/L to 137 mEq/L and is left skewed. From the box plot, we can see that the minimum is 125 mEq/L, the lower quartile is 134 mEq/L, the median is 137 mEq/L, the mean is 136.63 mEq/L, the upper quartile is 140 mEq/L, and the maximum is 148 mEq/L.

## Bivariate Analysis

*(With reference to Appendices)*

Bivariate 1 shows the number of deaths in each age group. A line chart was chosen because it aptly shows the increase and decrease of values without occupying much space. From the graph, we can see that most deaths occur between 60 and 70 years old. The graph does not have a monotonic relationship because the line fluctuates as age increases. From here, we could deduce that age does not impact your chances of dying by CVD’s. However, the correlation number beside the graph shows that there is a correlation between age and death; as age increases, you are more likely to die by CVD.

Bivariate 2 shows the average creatine phosphokinase level of non-death and death events. A bar chart was chosen because it aptly compares the value between two factors. The left blue bar represents a non-death event, and the right orange bar represents a death event. The number above the bars is the average creatine phosphokinase level for both bars. From here, we can see that the orange bar has a higher number than the blue bar. The correlation number shows that there is a correlation between creatine phosphokinase and death. This shows that having higher levels of creatine phosphokinase could lead to CVD deaths.

Bivariate 3 shows the average ejection fracture percentage of non-death and death cases. A bar chart was chosen because it aptly compares the value between two factors. The left blue bar represents a non-death event, and the right orange bar represents a death event. The number above the bars is the average ejection fracture percentage for both bars. From here, we can see that the orange bar has a lower number than the blue bar. The correlation number shows that there is a negative correlation between creatine phosphokinase and death. This shows that having a lower ejection fracture percentage could lead to CVD deaths.

Bivariate 4 shows the average platelet count of non-death and death cases. A bar chart was chosen because it aptly compares the value between two factors. The left blue bar represents a non-death event, and the right orange bar represents a death event. The number above the bars is the average platelet count for both bars. From here, we can see that the orange bar has a lower number than the blue bar. The correlation number shows that there is a negative correlation between platelet count and death. This shows that having a lower platelet count could lead to CVD deaths.

Bivariate 5 shows the average serum creatine level of non-death and death events. A bar chart was chosen because it aptly compares the value between two factors. The left blue bar represents a non-death event, and the right orange bar represents a death event. The number above the bars is the average serum creatine level for both bars. From here, we can see that the orange bar has a higher number than the blue bar. The correlation number shows that there is a correlation between serum creatine and death. This shows that having higher levels of serum creatine could lead to CVD deaths.

Bivariate 6 shows the average serum sodium level of non-death and death cases. A bar chart was chosen because it aptly compares the value between two factors. The left blue bar represents a non-death event, and the right orange bar represents a death event. The number above the bars is the average serum sodium level for both bars. From here, we can see that the orange bar has a lower number than the blue bar. The correlation number shows that there is a negative correlation between platelet count and death. This shows that having a lower serum sodium level could lead to CVD deaths.

Bivariate 7 shows the proportion of smokers by gender. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents female, and the right represents male. The orange portion represents smokers, and the blue portion represents non-smokers. And to the right of the graph, the number shows the correlation between the two factors: gender and smoking. From here, we can see that the orange portion for males is much larger than the females. There is strong correlation between the two factors. This shows that there are more male smokers than female smokers.

Bivariate 8 shows the number of smokers by age. A tree map was chosen as it is colourful and eye catching. The squares represent the number of smokers within an age group; the bigger the square, the more smokers there are. From here, we can see that most smokers are between 60 and 65 years old. There is a slight correlation between the two factors. This shows that age has a slight impact on smoking.

Bivariate 9 shows the proportion of anaemics who died. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents non-anaemics, and the right represents anaemics. The orange portion shows a death event, and the blue portion shows a non-death event. And to the right of the graph, the number shows the correlation between the two factors: death and anaemia. From here, we can see that the death rate for anaemics is larger than that of the non-anaemics. There is also a slight correlation between the two factors. This shows that anaemia slightly affects death probability.

Bivariate 10 shows the proportion of anaemics by gender. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents female, and the right represents male. The orange portion represents anaemics, and the blue portion represents non-anaemics. And to the right of the graph, the number shows the correlation between the two factors: gender and anaemia. From here, we can see that the orange portion for males is smaller than the females. There is a negative correlation between the two factors. This shows that there are fewer anaemic males than females.

Bivariate 11 shows the number of anaemics by age. A tree map was chosen as it is colourful and eye catching. The squares represent the number of anaemics within an age group; the bigger the square, the more anaemics there are. From here, we can see that most anaemics are between 60 and 65 years old. There is a slight correlation between the two factors. This shows that age has a slight impact on anaemia.

Bivariate 12 shows the proportion of diabetics by gender. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents female, and the right represents male. The orange portion shows diabetics, and the blue portion shows non-diabetics. And to the right of the graph, the number shows the correlation between the two factors: gender and diabetes. From here, we can see that the orange portion for males is larger than the females. There is a negative correlation between the two factors. This shows that there are more diabetic males than females.

Bivariate 13 shows the number of diabetics by age. A tree map was chosen as it is colourful and eye catching. The squares represent the number of diabetics within an age group; the bigger the square, the more diabetics there are. From here, we can see that most diabetics are between 60 and 65 years old. There is a negative correlation between the two factors. This shows that the frequency of diabetes reduces with age.

Bivariate 14 shows the proportion of people with HBP who died. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents people without HBP, and the right represents people with HBP. The orange portion shows a death event, and the blue portion shows a non-death event. And to the right of the graph, the number shows the correlation between the two factors: death and HBP. From here, we can see that the death rate for people with HBP is more than that of people without HBP. There is a correlation between the two factors. This shows that HBP influences death probability.

Bivariate 15 shows the proportion of people with HBP by gender. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents female, and the right represents male. The orange portion shows people with HBP, and the blue portion shows people without HBP. And to the right of the graph, the number shows the correlation between the two factors: gender and HBP. From here, we can see that the orange portion for males is smaller than the females. There is a negative correlation between the two factors. This shows that there are fewer males with HBP than females.

Bivariate 16 shows the number of people with HBP by age. A tree map was chosen as it is colourful and eye catching. The squares represent the number of people with HBP within an age group; the bigger the square, the more people with HBP there are. From here, we can see that most people with HBP are between 60 and 65 years old. There is a correlation between the two factors. This shows that the frequency of people with HBP increases with age.

Bivariate 17 shows the relationship between anaemia and HBP. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents non-anaemics, and the right represents anaemics. The orange portion shows people with HBP, and the blue portion shows people without HBP. And to the right of the graph, the number shows the correlation between the two factors: anaemia and HBP. From here, we can see that the orange portion for anaemics is slightly larger than the non-anaemics. There is a slight correlation between the two factors. This shows that anaemia has a slight influence on HBP.

Bivariate 18 shows the relationship between anaemia and smoking. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents non-anaemics, and the right represents smoking. The orange portion shows smokers, and the blue portion shows non-smokers. And to the right of the graph, the number shows the correlation between the two factors: anaemia and smoking. From here, we can see that the orange portion for anaemics is smaller than the non-anaemics. There is a negative correlation between the two factors. This shows that less anaemics smoke than non-anaemics.

Bivariate 19 shows the relationship between smoking and diabetes. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents non-diabetics, and the right represents diabetics. The orange portion shows smokers, and the blue portion shows non-smokers. And to the right of the graph, the number shows the correlation between the two factors: smoking and diabetes. From here, we can see that the orange portion for diabetics is smaller than the non-diabetics. There is a negative correlation between the two factors. This shows that diabetics are less likely to smoke.

Bivariate 20 shows the relationship between smoking and HBP. A 100% stacked bar chart was chosen as it can show proportionality for two or more variables. The left bar represents non- smokers, and the right represents smokers. The orange portion shows HBP, and the blue portion shows non-HBP. And to the right of the graph, the number shows the correlation between the two factors: smoking and diabetes. From here, we can see that the orange portion for smokers is smaller than the non- smokers. There is a negative correlation between the two factors. This shows that smokers are less likely to have HBP.

# Others

This dataset is not 100% accurate as it does not show ages below 40 years old. Hence, some of the graphs may be inaccurate, skewed, or biased towards adults. When viewing the conclusions of this analysis, view it with scepticism and a pinch of salt. These conclusions were drawn from the given dataset, and some may disagree with factual medical reports and hence might not have references cited with them.

# Summary

Factors that increase your likelihood of CVD deaths are:

* Old age
* Having anaemia
  + (Wiley Online Library, 2021)
* Having HBP
  + (National Library of Medicine, 2015)
* Having a high creatine phosphokinase level
* Having a low ejection fracture percentage
  + (National Library of Medicine, 2005)
* Having a high serum creatinine level
  + (S G Wannamethee, 1997)
* Having a low serum sodium level
  + (Xin He, 2018)
* Having a low platelet count

Factors that increase your likelihood of getting diabetes are:

* Gender
  + Males are more likely to get diabetes (EA Gale, 2001)
* Age
  + Younger people are more likely to get diabetes
* Smoking
  + (Centers for Disease Control and Prevention, 2022)

Factors that increase your likelihood of smoking are:

* Gender
  + Males are more likely to smoke (National Library of Medicine, 2022)
* Age
  + Older people are more likely to smoke
* Not having anaemia

Factors that increase your likelihood of getting anaemia are:

* Gender
  + Females are more likely to have anaemia.
* Age
  + Older people are more likely to get anaemia

Factors that increase your likelihood of getting HBP are:

* Gender
  + Females are more likely to have HBP.
* Age
  + Older people are more likely to get HBP
* Not smoking
* Having anaemia

# Appendices

## Univariate 1

Chart, pie chart

Description automatically generated

## Univariate 2

Chart, histogram

Description automatically generated

## Univariate 3

Table

Description automatically generated with medium confidence

## Univariate 4

Chart

Description automatically generated

## Univariate 5

A picture containing chart

Description automatically generated

## Univariate 6

Graphical user interface, application

Description automatically generated

## Univariate 7

Chart, histogram

Description automatically generated

## Bivariate 1

Chart, line chart

Description automatically generated

## Bivariate 2

Chart, bar chart

Description automatically generated

## Bivariate 3

Chart, bar chart

Description automatically generated

## Bivariate 4

Chart, bar chart, waterfall chart

Description automatically generated

## Bivariate 5

Chart, bar chart

Description automatically generated

## Bivariate 6

Chart, bar chart

Description automatically generated

## Bivariate 7

Chart, bar chart, waterfall chart

Description automatically generated

## Bivariate 8

Chart, treemap chart

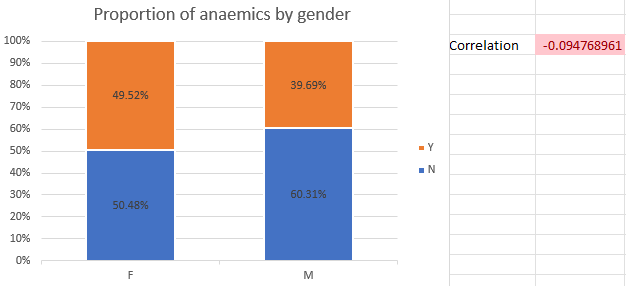
Description automatically generated

## Bivariate 9

Chart, waterfall chart

Description automatically generated

## Bivariate 10

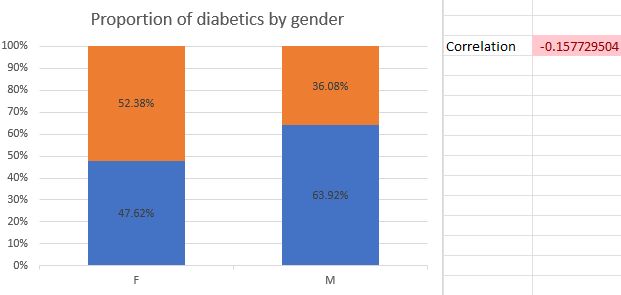


## Bivariate 11

Chart, treemap chart

Description automatically generated

## Bivariate 12



## Bivariate 13

Chart, treemap chart

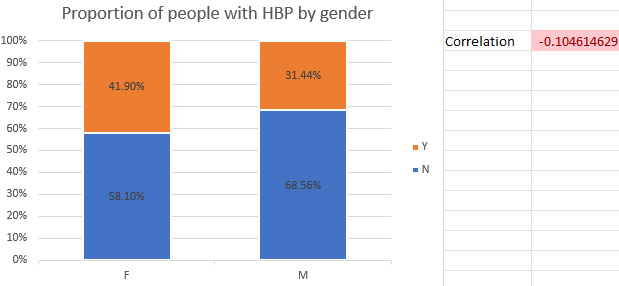
Description automatically generated

## Bivariate 14

Chart, waterfall chart

Description automatically generated

## Bivariate 15



## Bivariate 16

Chart, treemap chart

Description automatically generated

## Bivariate 17

Chart, waterfall chart

Description automatically generated

## Bivariate 18

Chart, waterfall chart

Description automatically generated

## Bivariate 19

Chart, waterfall chart

Description automatically generated

## Bivariate 20

Chart, waterfall chart

Description automatically generated

# Bibliography

Centers for Disease Control and Prevention. (5 May, 2022). *Smoking and Diabetes*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html#:~:text=We%20now%20know%20that%20smoking%20is%20one%20cause%20of%20type%202%20diabetes.&text=In%20fact%2C%20people%20who%20smoke,people%20who%20don't%20smoke.&text=People%20with%20diabetes%2

EA Gale, K. G. (Jan, 2001). *Diabetes and gender*. Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/11206408/#:~:text=Men%20seem%20more%20susceptible%20than,II%20diabetes%20to%20their%20offspring.

National Library of Medicine. (13 December, 2005). *Influence of ejection fraction on cardiovascular outcomes in a broad spectrum of heart failure patients*. Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/16330684/

National Library of Medicine. (November, 2015). *High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059018/

National Library of Medicine. (May, 2022). *Are there gender differences in tobacco smoking?* Retrieved from National Library of Medicine: https://nida.nih.gov/publications/research-reports/tobacco-nicotine-e-cigarettes/are-there-gender-differences-in-tobacco-smoking#:~:text=Introduction-,Tobacco%2C%20Nicotine%2C%20and%20E%2DCigarettes%20Research%20Report,gender%20differences%20in%20tobacco%

S G Wannamethee, A. G. (March, 1997). *Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke*. Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/9056611/

Wiley Online Library. (5 October, 2021). *Association of iron deficiency with incident cardiovascular diseases and mortality in the general population*. Retrieved from Wiley Online Library: https://onlinelibrary.wiley.com/doi/10.1002/ehf2.13589

Xin He, C. L. (August, 2018). *Risk of Cardiovascular Mortality Associated With Serum Sodium and Chloride in the General Population*. Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/29935992/#:~:text=In%20the%20multivariate%20model%2C%20low,CI%2C%200.97%2D1.12%20per%20standard