

Faculty of Information Science and Technology (FIST)

TDA3121 Data Analytics Fundamentals

Trimester 3 2022/2023

Group Assignment (30%)

Due date: **29 August 2023**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Student Name | Student ID | Academic Program | Tutorial Session |
| 1 | LEE EN | 1191100578 | AI | 1B |
| 2 | FOO HAW LIANG | 1191101497 | ST | 1B |
| 3 | SIAH KAH CHUAN | 1191100577 | ST | 1B |
| 4 | GRAYSON GOH JIN YI | 1191101340 | ST | 1B |

**MARKING RUBRICS (30%)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 5 | 4 | 3 | 2 | 1 | 0 |
| Relevant dataset |  |  |  |  |  |  |
| Data Manipulation and Exploration |  |  |  |  |  |  |
| Relevant Visualization Techniques |  |  |  |  |  |  |
| Customization and Interpretation |  |  |  |  |  |  |
| Code Quality |  |  |  |  |  |  |
| Documentation |  |  |  |  |  |  |
| TOTAL: |  |  | **/30** | |  |  |

Contents

[Introduction 3](#_Toc143525125)

[Data Dictionary 4](#_Toc143525126)

[Data Cleaning 5](#_Toc143525127)

[Check Duplicated 5](#_Toc143525128)

[Stacked Bar Chart 8](#_Toc143525129)

[Stacked Bar Chart for relationship understanding 8](#_Toc143525130)

[Relationship between PClass(Ticket Class) and Survival 8](#_Toc143525131)

[Boxplot 9](#_Toc143525132)

[Interpreting result 9](#_Toc143525133)

[Why boxplot was chosen 9](#_Toc143525134)

[Conclusion 10](#_Toc143525135)

[DECLARATION: 11](#_Toc143525136)

[**Group Member’s Declaration** 12](#_Toc143525137)

[**Group Member’s Declaration** 13](#_Toc143525138)

[**Group Member’s Declaration** 14](#_Toc143525139)

[**Group Member’s Declaration** 15](#_Toc143525140)

# Introduction

In this assignment, we have chosen the titanic dataset. The Titanic dataset is a classic public dataset, which contains 418 unique records about the Titanic's passengers who were victims of the most infamous shipwrecks in history on April 15, 1912. The data consists of demographic and traveling information for each passenger, such as name, sex, age, ticket class, fare, survival status, etc. The goal of the dataset is to understand the relationship between the survival of these passengers and their features.

The Titanic dataset is widely used as an introductory dataset for predictive analytics and machine learning. It is suitable for exploring various aspects of data analysis and visualization, such as data cleaning, feature engineering, model selection, evaluation, and interpretation. The dataset is also interesting from a historical and social perspective, as it reveals the inequalities and biases that existed among the passengers of different classes, genders, and nationalities.

In this report, we use two visualization techniques to explore the Titanic dataset and create graphical representations that effectively communicate the information. The two techniques we have chosen are bar chart and boxplot. A stacked bar chart is a type of bar chart that shows the proportion of different categories within each group. A boxplot is a type of plot that shows the distribution of a numerical variable by using a box to represent the median, quartiles, and outliers.

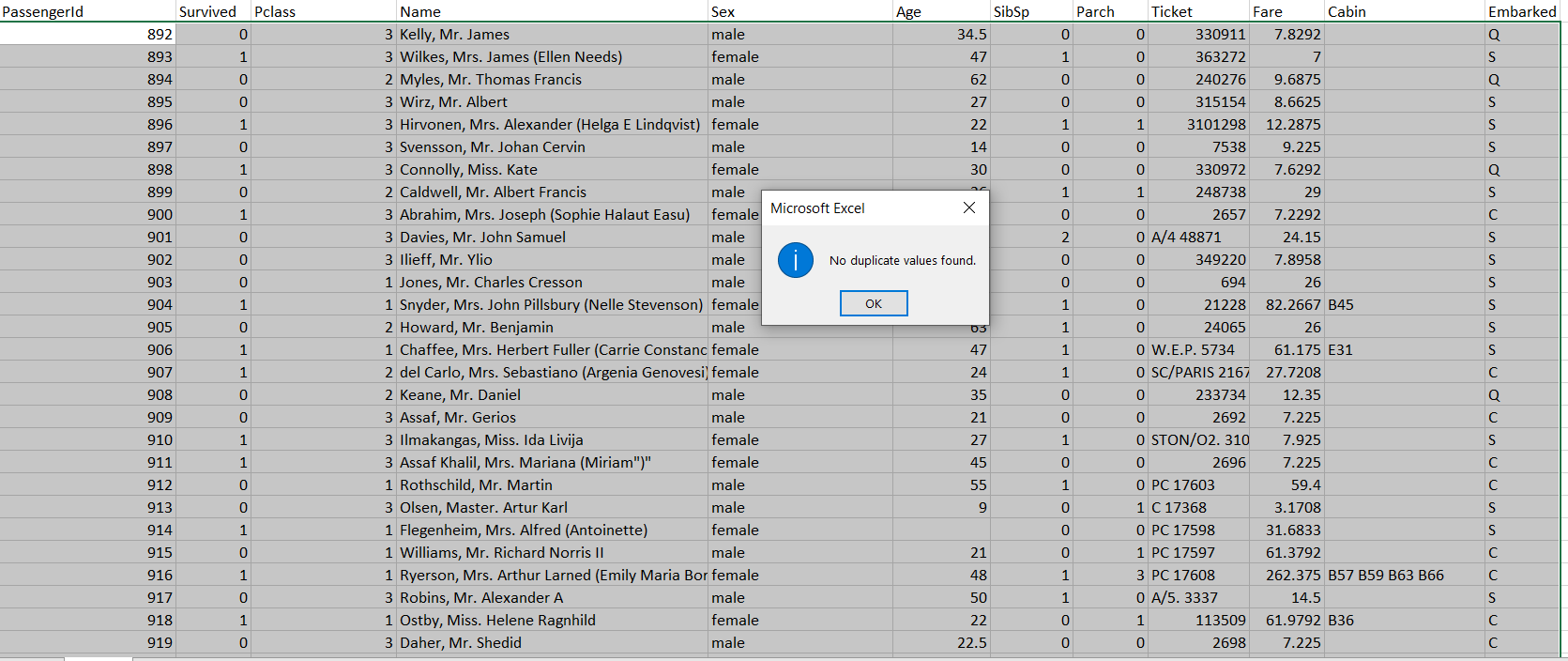
# Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| PassengerId | Integer | A unique identifier assigned to each passenger on the Titanic. |
| Survived | Integer | This column indicates whether the passenger survived the Titanic disaster. It was record as 1 if the passenger survived and 0 if the passenger did not survive. |
| Pclass | Integer | This column represents the class of the ticket purchased by the passenger. It represents the socio-economic status which is 1st class (upper), 2nd class (middle), or 3rd class (lower). |
| Name | String | This column shows the full name of the passenger, including the title. |
| Sex | String | This column presents the gender of the passenger included male or female. |
| Age | Float | This column shows the age of the passenger at the time of boarding the Titanic. |
| SibSp | Integer | This column shows the number of siblings or spouses the passenger had aboard the Titanic. |
| Parch | Integer | This column shows the number of parents or children the passenger had aboard the Titanic. |
| Ticket | String | This column included the ticket number associated with the passenger's ticket. |
| Fare | Float | This column represents the fare paid by the passenger for the ticket. |
| Cabin | String | This column records the cabin number where the passenger stayed. This information might be missing for some passengers. |
| Embarked | String | This column shows the port of embarkation for the passenger to board the Titanic: C = Cherbourg, Q = Queenstown, S = Southampton. |

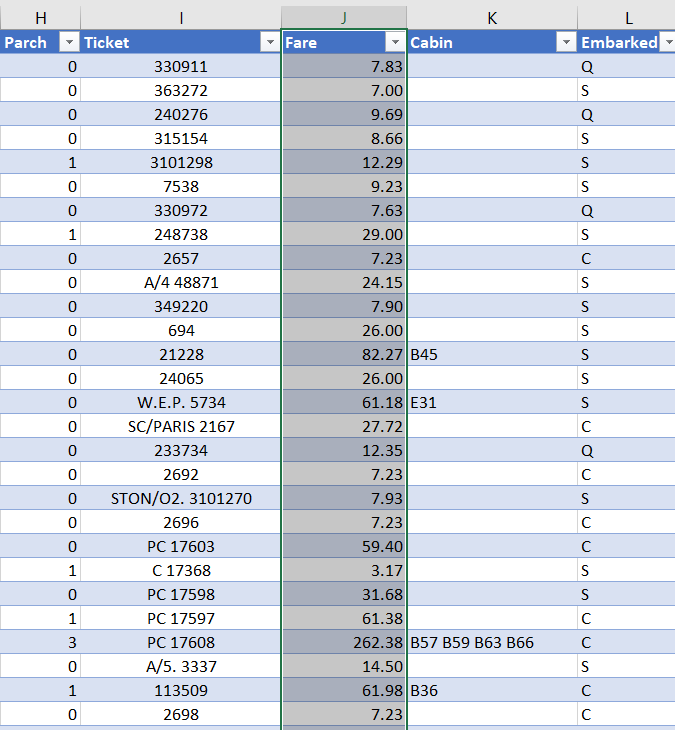
# Data Cleaning

## Check Duplicated

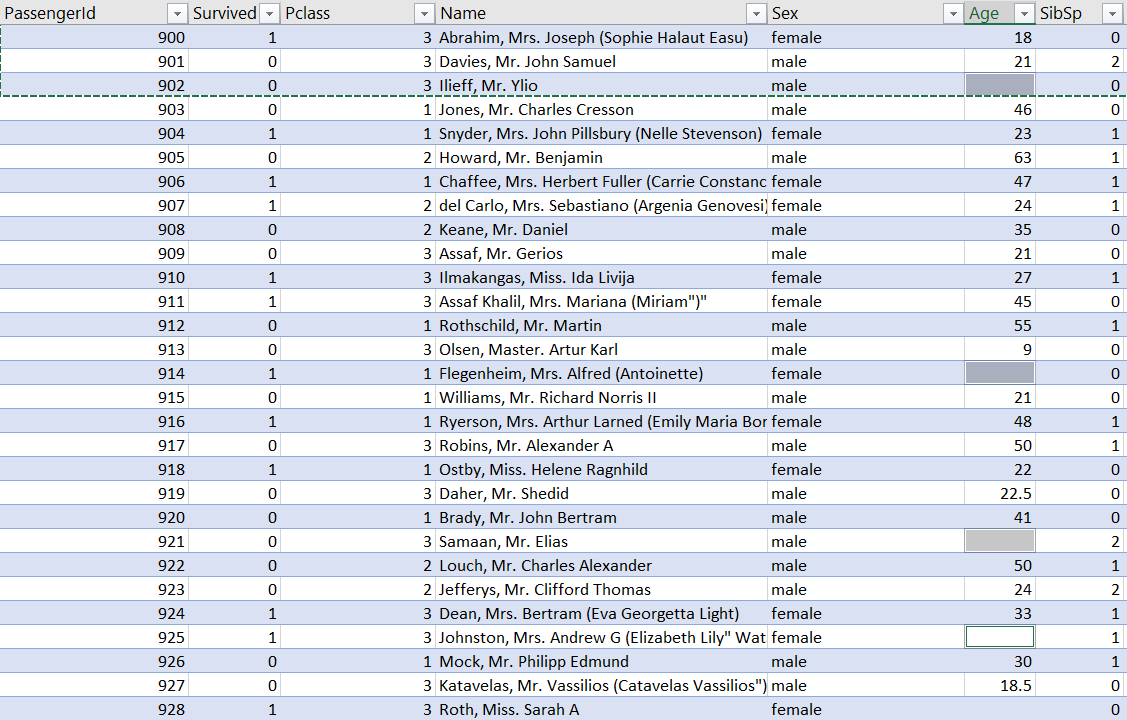
Firstly, check is that any duplicated data in the dataset, if any, remove it.



Next, we make the Fare column to a fixed decimal point.



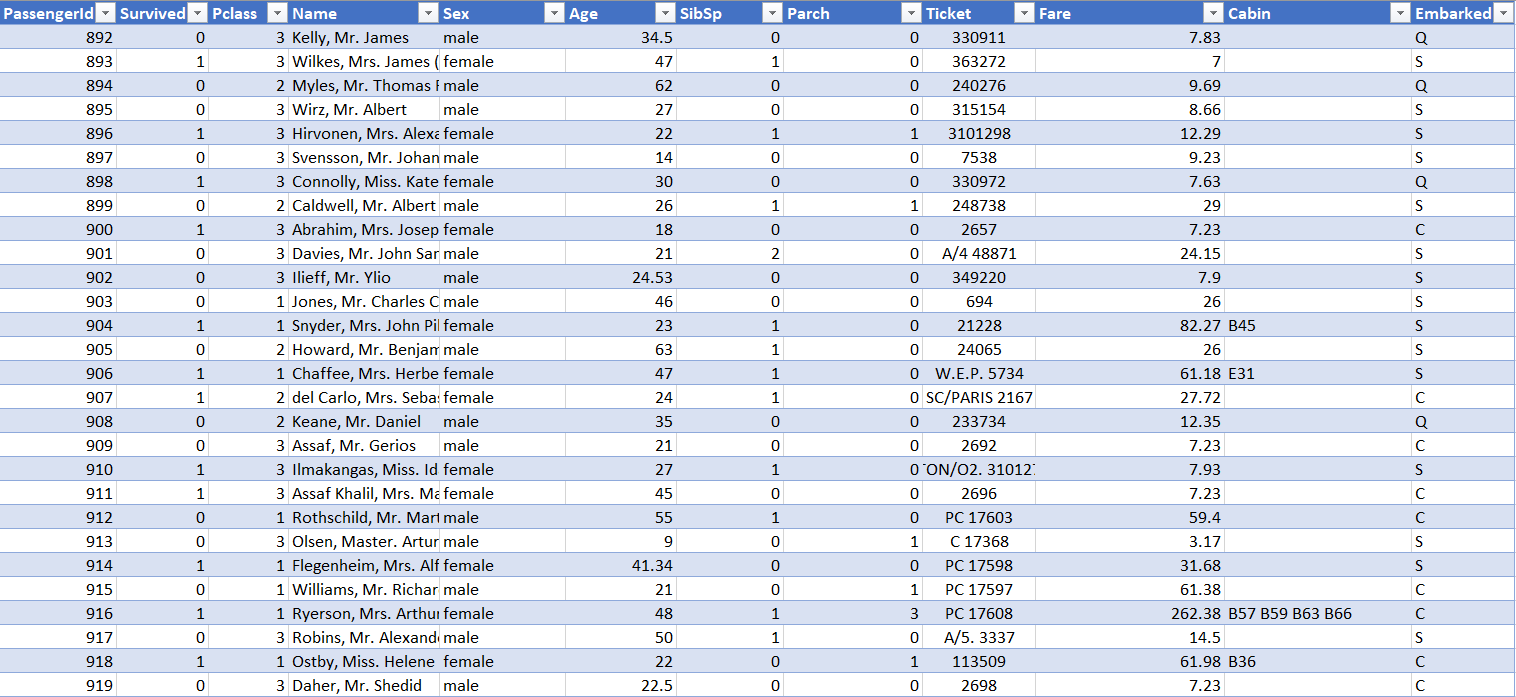
Next, as there have some missing values in the dataset in the Age and Cabin column, we will need to fill up for the missing values especially for the Age column as the age is one of the important factors for us to determine the survival during our analysis. Our approach uses to filling the missing age value is using the mean age of similar people. We divided the people into 6 category which is based on the Pclass and Sex of the passenger as shown in the below table. After getting the mean age of 6 category of passenger, we can use these mean age value to fill in the missing ages.



**Table 1.1: Reference table for filling missing age values.**

|  |  |
| --- | --- |
| Passenger Category | Mean Age |
| Pclass1 + male | 40.52 |
| Pclass2 + male | 30.94 |
| Pclass3 + male | 24.53 |
| Pclass1 + female | 41.34 |
| Pclass2 + female | 24.38 |
| Pclass3 + female | 23.1 |

For the Cabin column, as more than 70% of values are missing. It is hard for us for finding accurate way to fill in the missing value, also in this column we cannot calculate the value’s mean. So, we will simply use the mode of the Known Cabin based on each Pclass to estimate their Cabin type. Table 1.2 show the Pclass and their mode cabin type.



**Table 1.2: Reference table for filling the missing Cabin values**

|  |  |  |
| --- | --- | --- |
| Pclass | Mode cabin type | Quantity |
| 1 | C | 35 |
| 2 | F | 5 |
| 3 | FE | 2 |

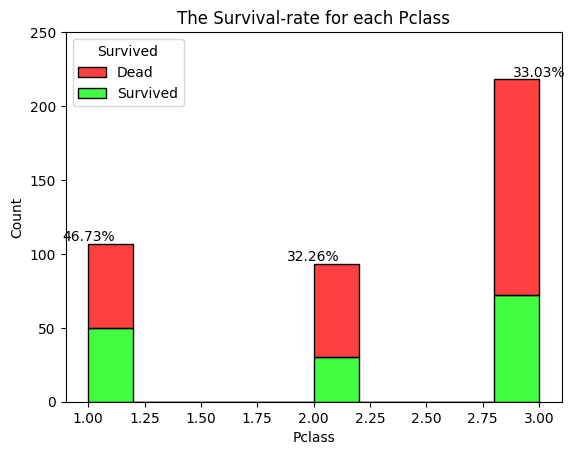
# Stacked Bar Chart

## Stacked Bar Chart for relationship understanding.

In titanic dataset, like most of the data analysis project, the main thing in here is to understand what and how the variables contribute to their target variable, which is Survival. By carry out Exploratory data analysis (EDA), we gain prior knowledge about the relationship between variables, before moving further into later analysis.

The 1st chart selected is bar chart. Bar chart is a simple visualization to show the count of observations. However, simple bar chart is a univariate visualization technique. As said above, to understand the relationship between variable and target variable, we used stacked bar chart here. Stacked bar chart allows more than 1 variable to be included inside it.

The bar chart selected will be used to understand the relationship between *Pclass* and the *Survival* variables. *Pclass* and *Survival* are both categorical variables, thus bar chart is suitable in our case, as it shows the average survival rate (*Survival*) among the *PClass*.



## Relationship between PClass(Ticket Class) and Survival

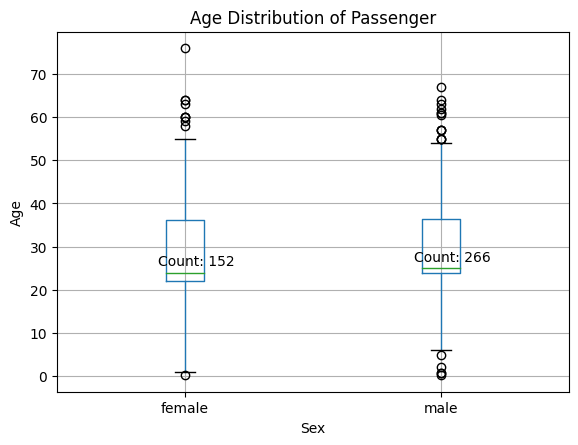
The histogram above shows the number of passengers for each Ticket Class. We can see from here that most of the passenger with 1st ticket class survived, as compared to 2nd and 3rd class. As the ticket class increase (from 1st to 2nd& 3rd), the survival rate decrease (refer the proportion of green colour in each *Pclass* bar, or the percentage above each bar). However, *Pclass* 2 have lower survival rate as compared to 3, but the difference is minor. Hence, the conclusion here we can make is:

Passengers that purchased Ticker Class 1st have the highest survival rate.

# Boxplot

## Interpreting result

The median *age* value in female passenger class is lower than male passenger class. This mean that the female passenger tends to have higher *age* compared male passenger. Apart from that female passenger have larger box size compared to male passenger. This mean that the female passenger *age* is more dispersed. Furthermore, female passenger class have larger range of *age*, which shows that the *age* data is more scattered. Meanwhile, for the skewness of data, both female and male passenger class are appearing to be positive skew data. We also observed that male passenger class have more outliers compared to female class. This tells us that this class have more unique age characteristics compared to the rest of class. In overall, we can conclude that there are lesser female passenger taking the ship with relatively less unique *age* characteristic, but having a wider range of age data.



## Why boxplot was chosen

This is because that this technique is able to summarize the distribution of groups of data better compared to other types of visualization techniques. This includes outliers’ detection, comparison among different groups of data, visualization of data skewness and its interquartile range.

# Conclusion

The exploratory data analysis (EDA) of the Titanic dataset revealed that ticket class and gender were two of the most important factors affecting survival rates. Passengers with first class tickets had the highest likelihood of survival, followed by passengers with second class tickets, and then passengers with third class tickets. This is likely because passengers with first class tickets were given priority for the lifeboats. Additionally, female passengers tended to be younger and had a wider age range than male passengers, which may have also contributed to their higher survival rate. The use of stacked bar charts and boxplots were effective visualization techniques for summarizing and comparing data distributions, detecting outliers, and illustrating trends in the data. Overall, this EDA provides valuable insights into the dynamics of the Titanic disaster and the factors that played a role in passengers' survival outcomes.

# DECLARATION:

We agree that all members deserve equal marks for this project. We confirm that we have contributed equally to produce an original report in our best effort.

Write down your ID (clearly) and sign (by all members):

A black line on a white background

Description automatically generated

Student ID: 1191100578 \_\_\_\_Signature:



Student ID: 1191101497 Signature:

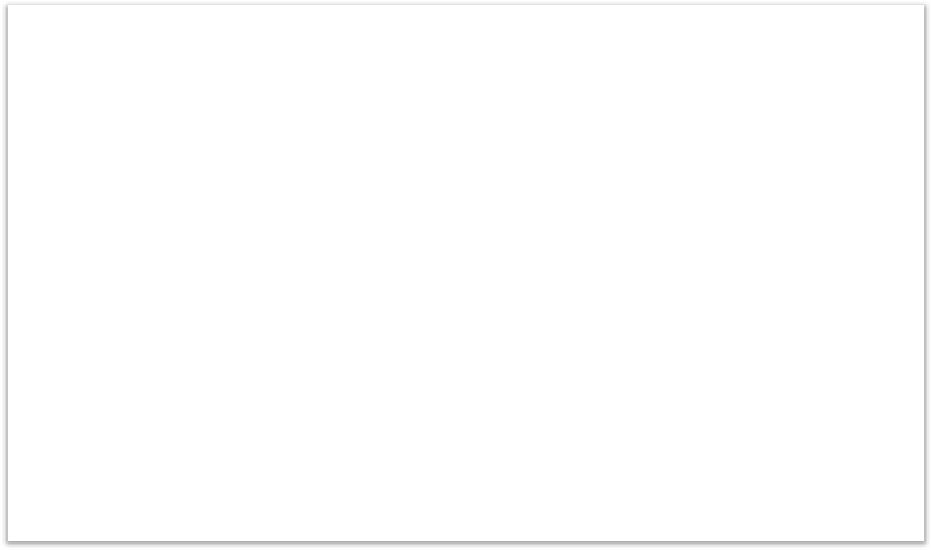


Student ID: 1191100577 Signature:

Student ID: 1191101340 Signature:

A black line on a white background

Description automatically generated



**Declaration by Group Leader**

I hereby declare that all group members’ names are correctly included in the above section. I hold a copy of this assignment which I can produce if the original is lost or damaged. I certify that not part of this assignment has been copied from any other student’s work or from any other source except where due acknowledgement is made in the assignment/project/etc.

Group Leader’s Signature: Group Leader’s Name: Lee En Group Leader’s ID: 1191100578 Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Lee En

Student ID: 1191100578

For the purpose of completing this assignment, I have performed the following tasks:

The bar chart using Python

The interpretation of graph

Reason to choose visualization

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.

A black line on a white background

Description automatically generated

Group member’s signature:

Group member’s name: Lee En

Group member’s ID: 1191100578

Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Foo Haw Liang

Student ID: 1191101497

For the purpose of completing this assignment, I have performed the following tasks:

Introduction

Conclusion

Report Compiling

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.



Group member’s signature:

Group member’s name: Foo Haw Liang

Group member’s ID: 1191101497

Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Siah Kah Chuan

Student ID: 1191100577

For the purpose of completing this assignment, I have performed the following tasks:

Generate boxplot based on age and sex by using python.

Interpret the result.

Give the reason of choosing this visualization techniques.

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.

A signature on a white background

Description automatically generated

Group member’s signature:

Group member’s name: Siah Kah Chuan

Group member’s ID: 1191100577

Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Grayson Goh Jin Yi

Student ID: 1191101340

For the purpose of completing this assignment, I have performed the following tasks:

Dara Cleaning, Data Filling, Data dictionary

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.



Group member’s signature:

Group member’s name: Grayson Goh Jin Yi

Group member’s ID: 1191101340

Date: 21/8/2023