Assignment # 02

Data Mining

Submitted to : Dr Muzamil
Submitted by : Abdullah Aslam(Team Lead)01-134172-005
Ahsan Goheer 01-134172-008
Section : BSCS – 7A
Date : October 25, 2020

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

Contributions:

Mutual Contributions

- Selection of dataset.
- Decision of attribute types for the given data.
- Formatting of assignment document.
- Getting a description of the dataset.

Team Lead Contributions (Abdullah Aslam)

- Compiled the assignment document.
- Generated bar chart of survivors in the dataset.
- Generated bar chart of passengers based on their gender.
- Generated bar chart of siblings/spouses.
- Generated bar chart of passengers based on their embarkment location.
- Generated bar chart for passengers who survived based on the number of siblings/spouses on board.
- Generated histogram for distribution of fare amongst passengers.
- Generated scatter plot showing fare with respect to age.

Team Member Contributions (Ahsan Goheer)

- Editing the assignment document.
- Generated bar chart of passengers based on their class.
- Generated histogram of passenger ages in the data.
- Generated the five number summary for the passenger ages.
- Generated the box plot for passenger age.
- Generated the bar chart for passengers with parents/children.
- Generated bar chart to plot the passenger survival with respect to their gender.
- Generated bar chart for the Number of Passengers with Parents and Children who Survived.

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

TABLE OF CONTENTS

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

LIST OF FIGURES

LIST OF TABLES

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

1 INTRODUCTION

This dataset is associated with the world-famous titanic incident that took place back in April 1912. This dataset is available at Kaggle as an open competition [1] in the form of 2 files (train and test) for Kagglers to apply machine learning to predict the survival of a person. As the test file has one less column (prediction class label of *survived*) than the train file, so we have decided to use only train.csv for this assignment. This dataset has 891 rows and 12 columns.

2 DESCRIPTION OF DATASET

Column name	Attribute type	Description
Passenger	Nominal	This column has the passenger ID.
Survived	Ordinal	This column has binary values depicting survival.
Pclass	Ordinal	Passengers in Titanic were divided into 3 classes. This column stores the values from 1 to 3 of passenger class.
Name	String	This column has the full name of each passenger.
Sex	Ordinal	This column has gender values (male/female).
Age	Ordinal	Age of each passenger is stored in this column.
SibSp	Discrete ordinal	The number of siblings or spouses aboard are stored in this column.
Parch	Discrete ordinal	This column contains the number of parents or children aboard on titanic.
Ticket	Nominal	Ticket number of each passenger is stored here.
Fare	Ratio scale	Fare paid by passenger is saved in this column.
Cabin	Nominal	This column contains the cabin number of each passenger.
Embarked	Nominal	There were 3 ports to embark Titanic. Their initials are stored in this column. C - Cherbourg Q - Queenstown S - Southampton

2.1 INFORMATION OF DATA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

Figure 1 Information regarding dataset

2.2 INFORMATION REGARDING MISSING VALUES

```
Missing Values in Age: 177
Missing Values in Cabin: 687
Missing Values in Embarked: 2
```

Figure 2 Information regarding missing value columns

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

2.3 SUMMARY OF ATTRIBUTES WITH NUMERIC VALUES

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Figure 3 Summary of all numeric attributes

2.4 FIRST AND LAST 5 ROWS OF DATASET

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Figure 4 First 5 rows of dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

Figure 5 Last 5 rows of dataset

3 GRAPHS FOR DATA VISUALIZATION

3.1 BAR CHARTS

3.1.1 Survived passengers

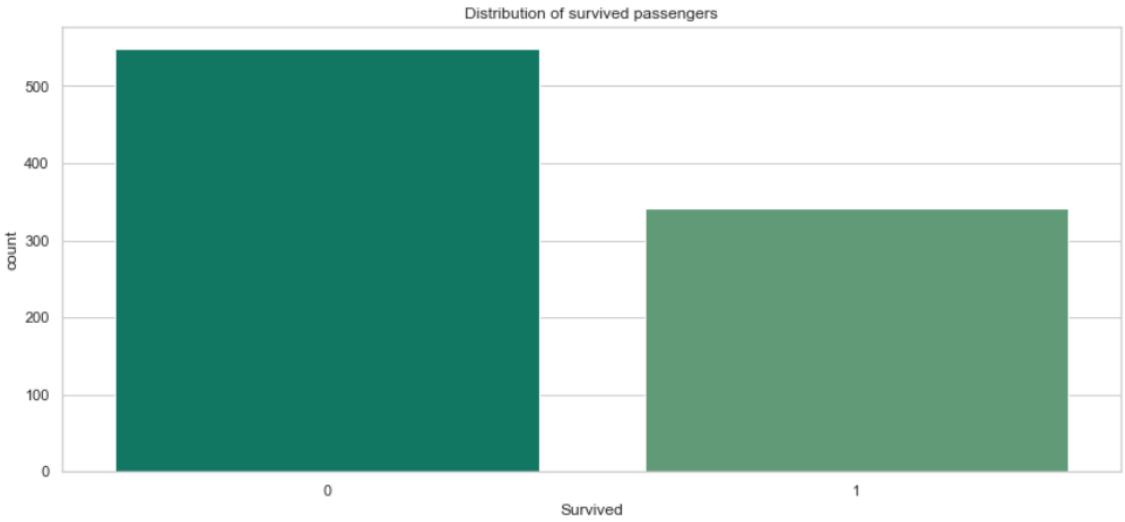

Figure 6 Bar chart of survival of passengers

Out of 891 Passengers in the data, 342 survived and 549 did not survive

3.1.2 Passengers' gender

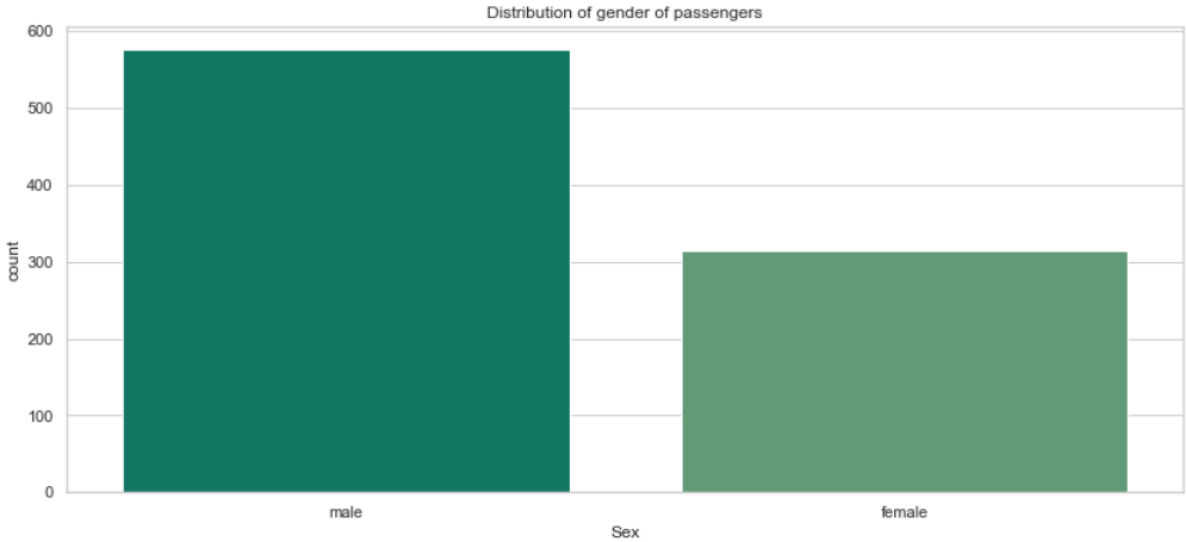

Figure 7 Bar chart of gender of passengers

Out of 891 Passengers in the data, 577 are males and 314 are females.

3.1.3 Passenger class (Pclass attribute of dataset)

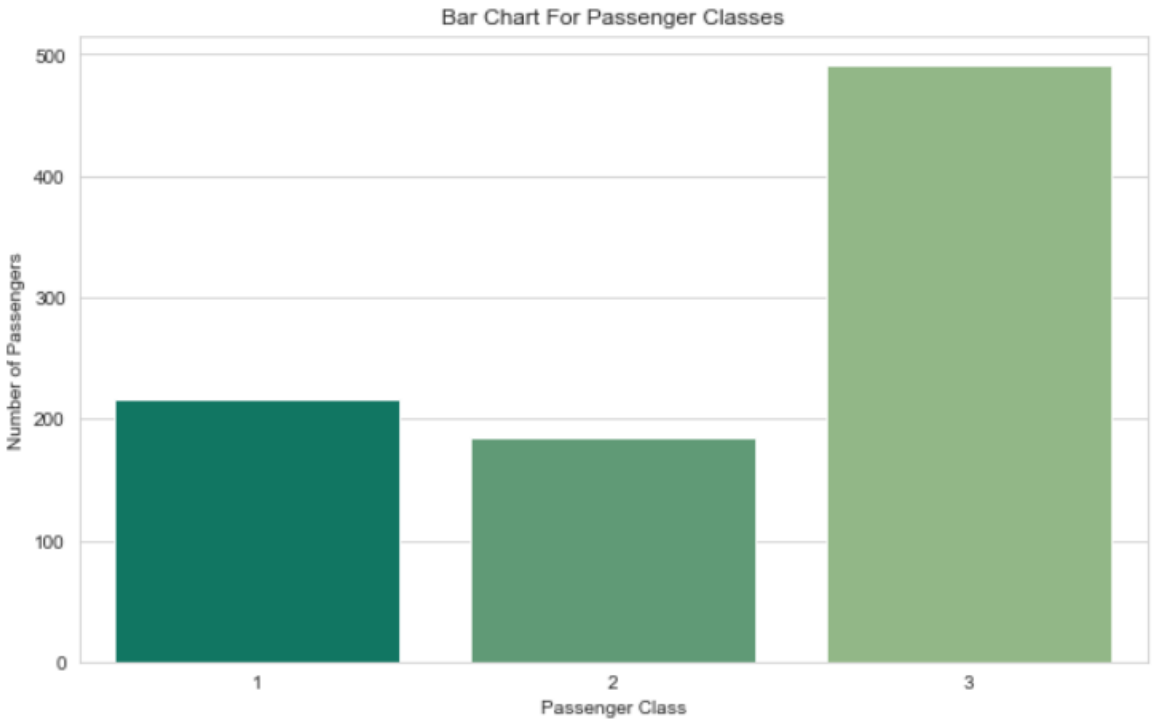

Figure 8 Bar chart of passenger class

Distribution of passengers based on their class:

Upper Class = 216
Middle Class = 184
Lower Class = 491

3.1.4 Siblings or Spouses

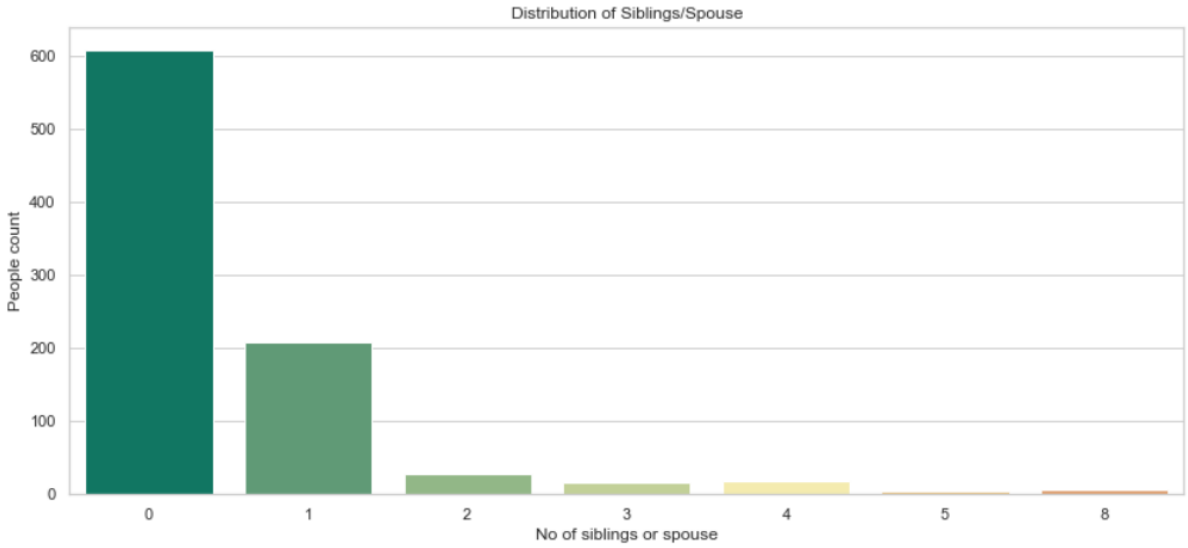

Figure 9 Bar chart of number of siblings or spouses

Breakdown of siblings and spouses of passengers in the data:

608 passengers have 0 siblings/spouses.
209 passengers have 1 siblings/spouses.
28 passengers have 2 siblings/spouses.
18 passengers have 3 siblings/spouses.
16 passengers have 4 siblings/spouses.
7 passengers have 5 siblings/spouses.
5 passengers have 6 siblings/spouses.

3.1.5 Embarked location

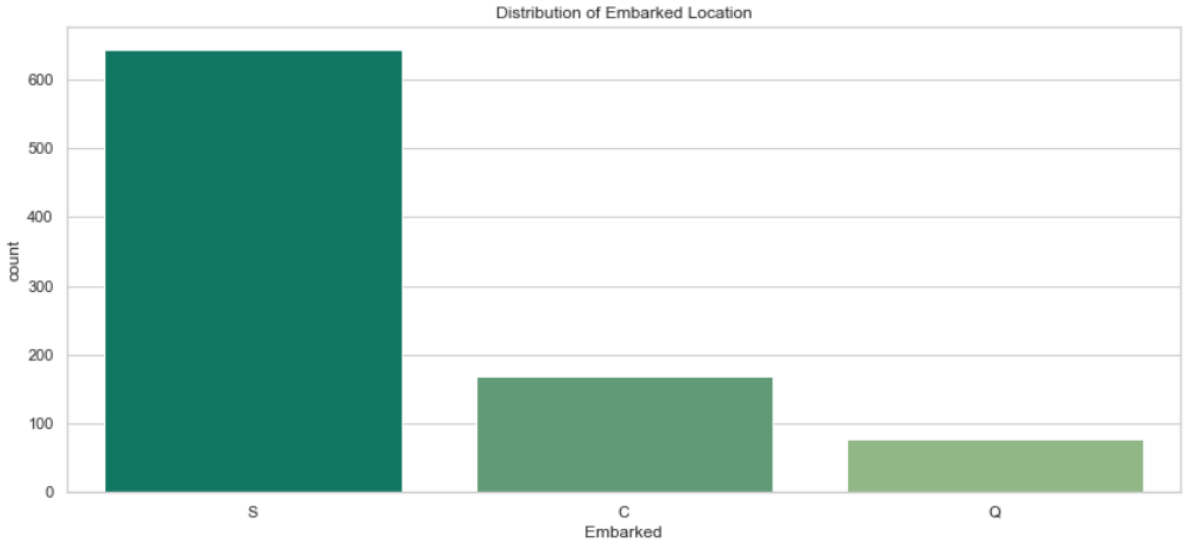

Figure 10 Bar chart of embarked locations

Number of passengers from Southampton : 644
Number of passengers from Cherbourg : 168
Number of passengers from Queenstown : 77

3.1.6 Passengers with corresponding number of parents and children

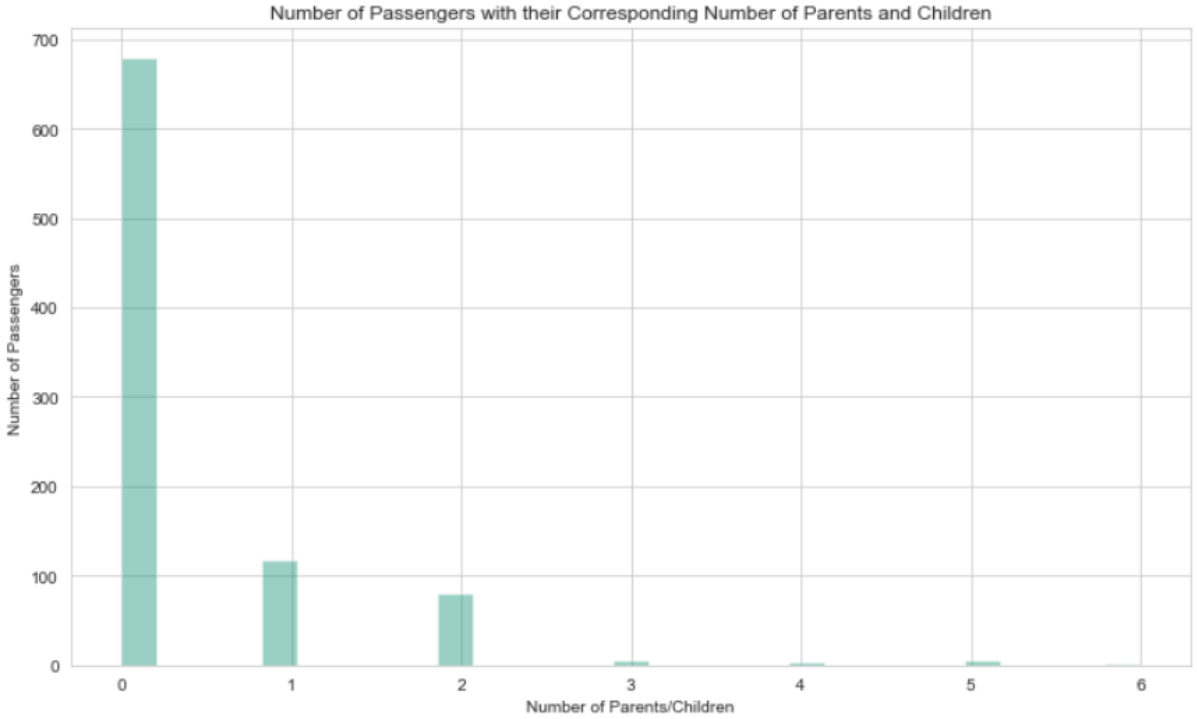

Figure 11 Bar chart of passengers with number of parents and children

Breakdown of parents and children of passengers in the data:

678 Passengers have 0 parents/children
118 Passengers have 1 parents/children
80 Passengers have 2 parents/children
5 Passengers have 3 parents/children
5 Passengers have 4 parents/children
4 Passengers have 5 parents/children
1 Passengers have 6 parents/children

3.1.7 Survival of passenger based on gender

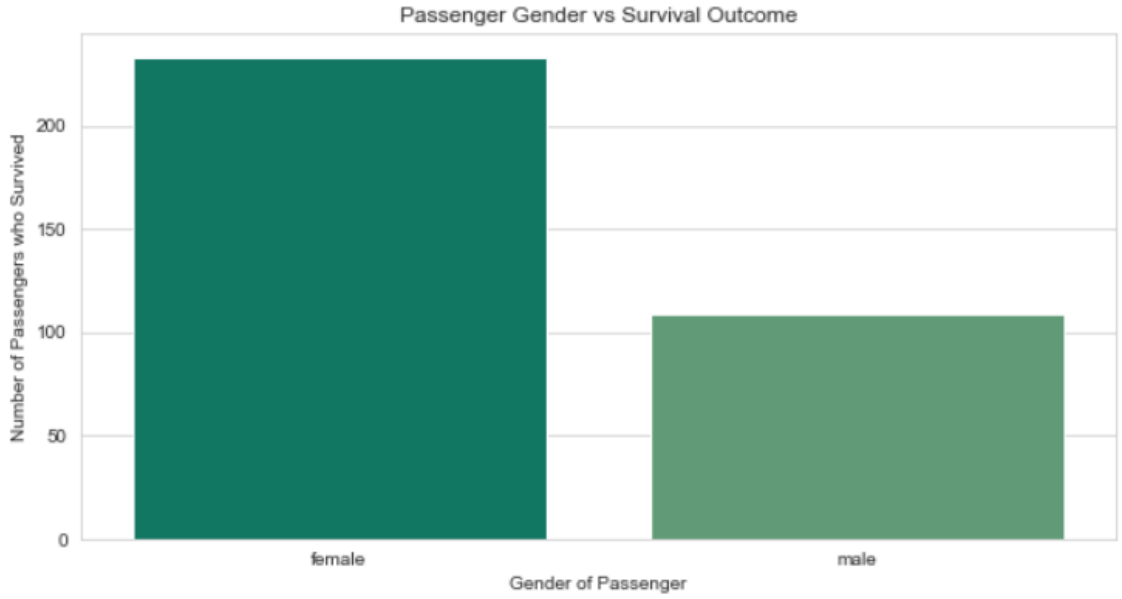

Figure 12 Bar chart of survival based on gender

Number of females who survived: 233
Number of males who survived: 109

3.1.8 Surviving passengers along with parents and children

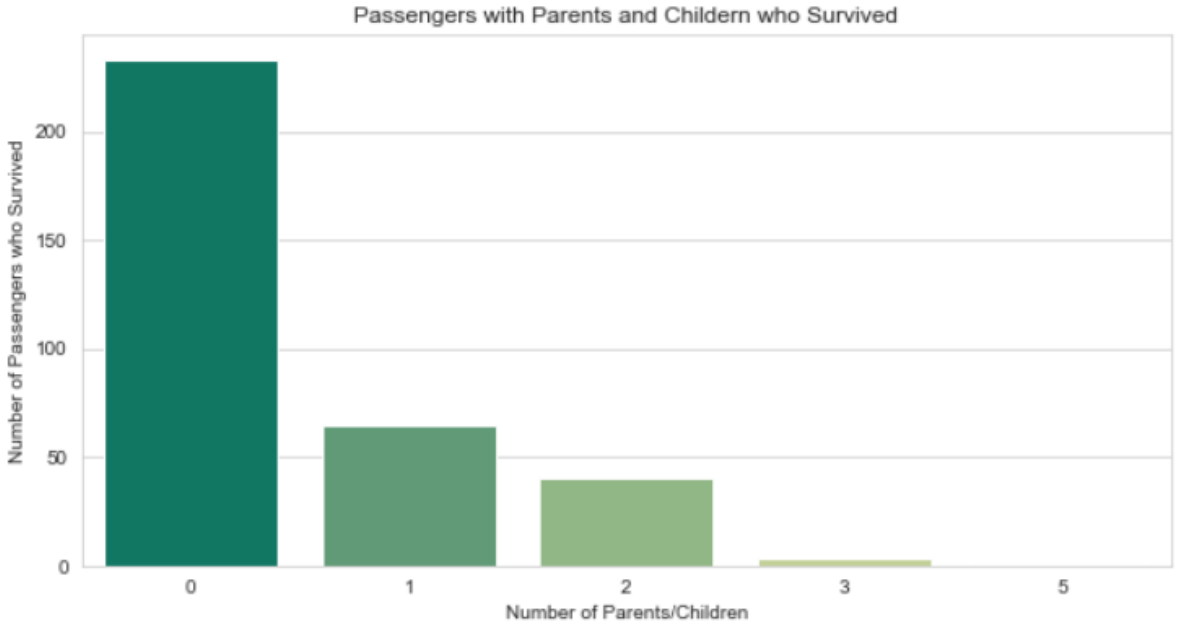

Figure 13 Bar chart of number of parent and number of children

Number of Passengers with 0 parents/children who survived =233
Number of Passengers with 1 parents/children who survived =65
Number of Passengers with 2 parents/children who survived =40
Number of Passengers with 3 parents/children who survived =3
Number of Passengers with 4 parents/children who survived = 1

3.2 HISTOGRAMS

3.2.1 Distribution of age amongst passengers

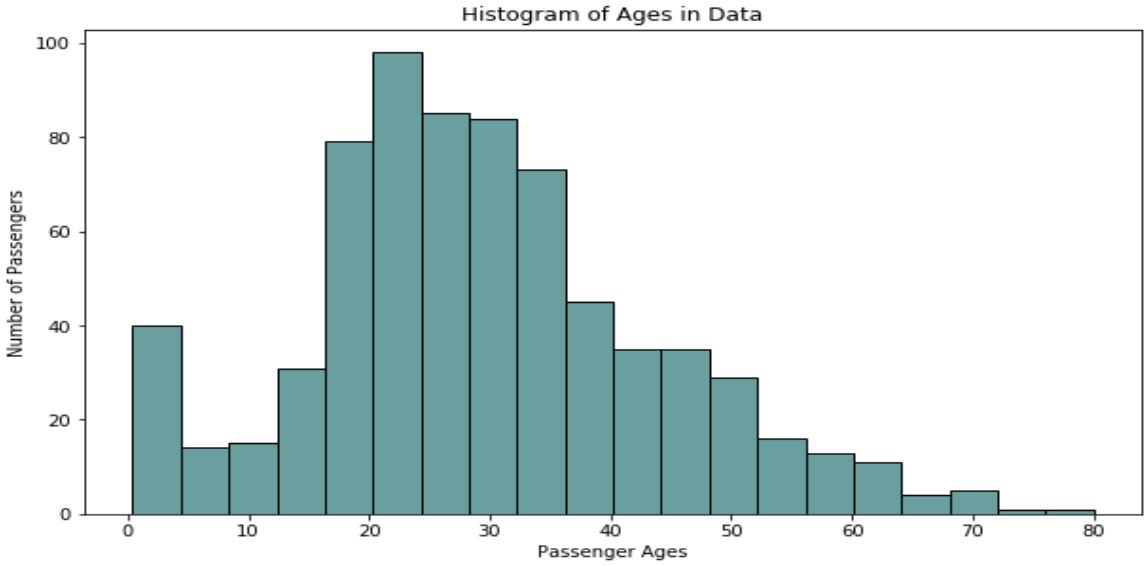

Figure 14 Histogram of ages of passengers

3.2.2 Distribution of fare amongst passengers

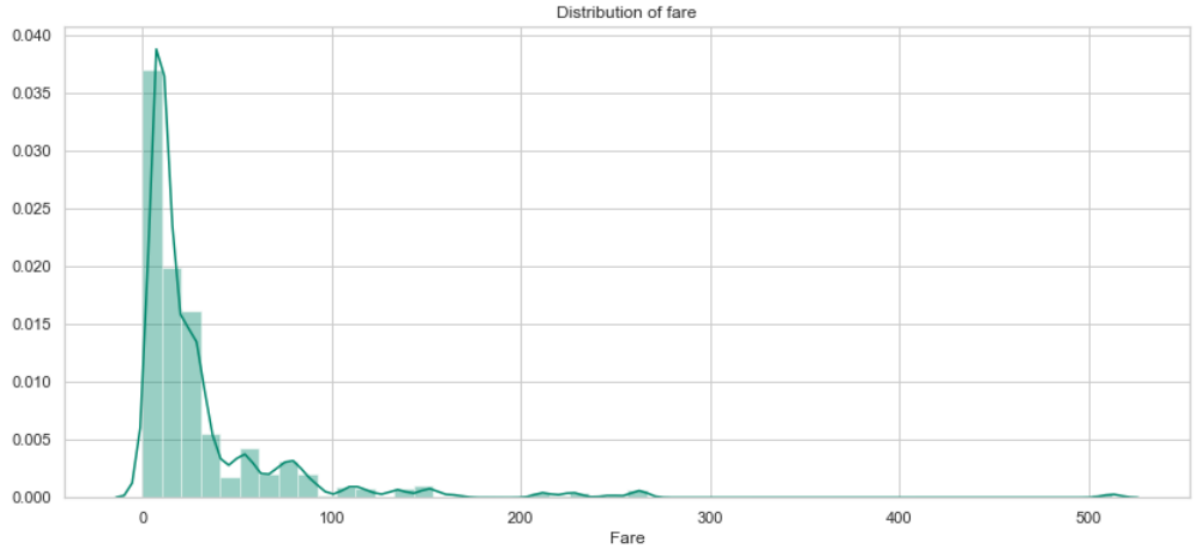

Figure 15 Histogram of fare of passengers

3.3 MULTIPLE BAR CHARTS

3.3.1 Passengers belonging to each class based on their gender

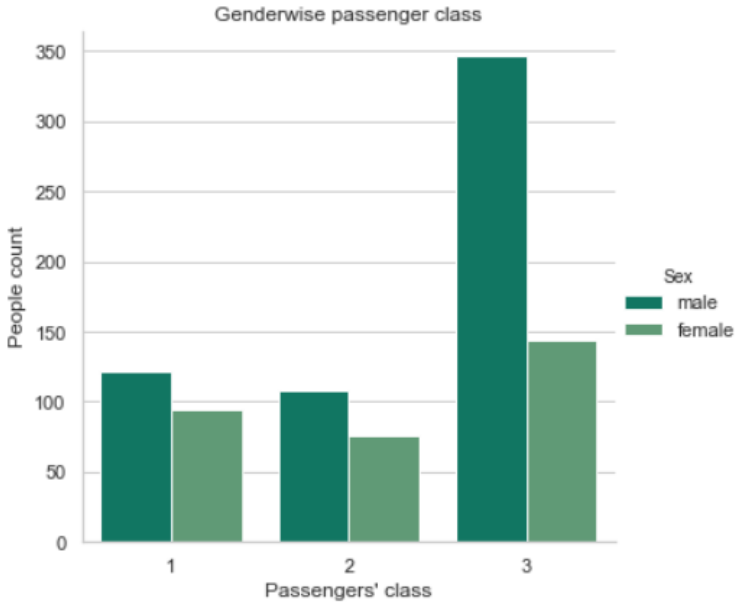

Figure 16 Multiple bar chart of passenger classes based on gender

3.3.2 Count of embarkment based on passenger's class

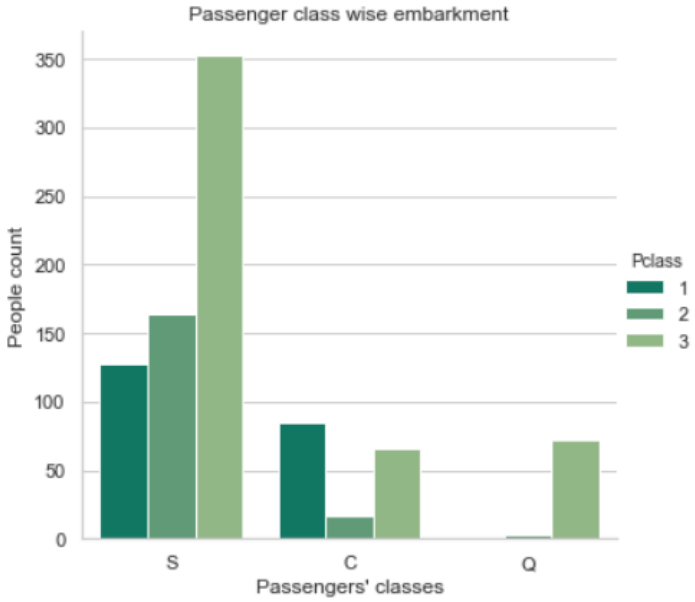

Figure 17 Multiple bar chart of embankment based on passenger class

3.3.3 Survival based on port of embarkment

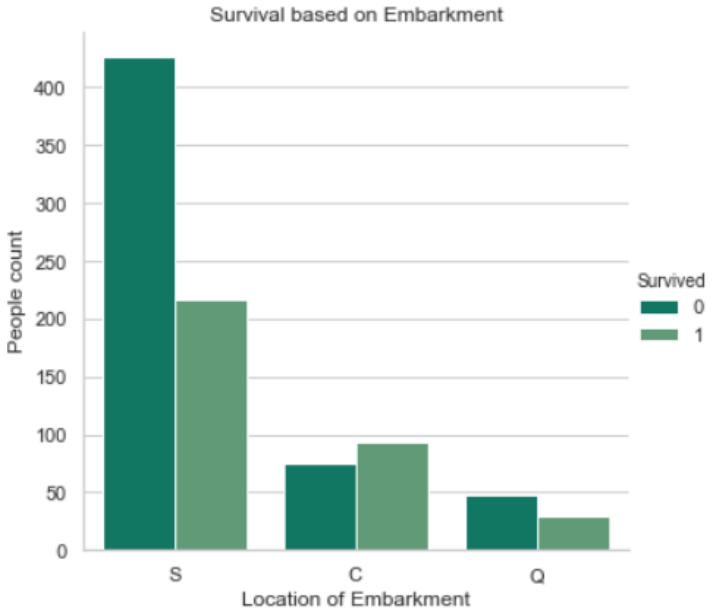

Figure 18 Multiple bar chart of embankment based on survival

3.3.4 Survival based on the number of siblings or spouse

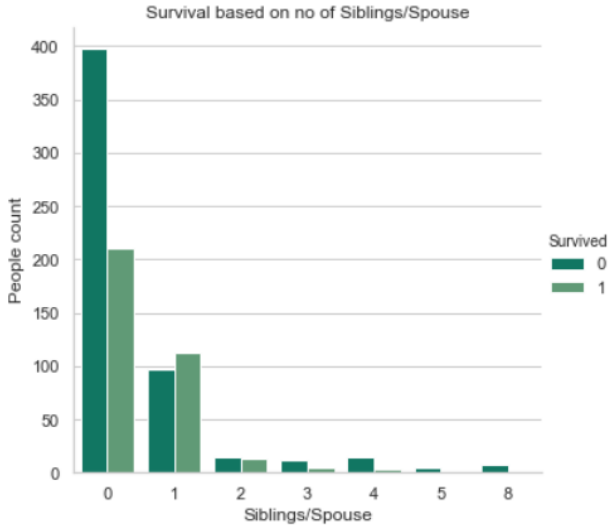

Figure 19 Multiple bar chart of siblings or spouse based on survival

3.3.5 Number of people embarking each port based on their gender

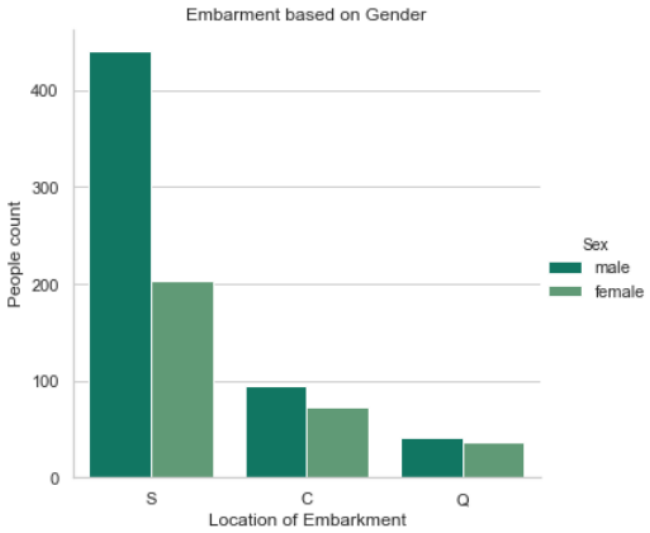

Figure 20 Multiple bar chart of embankment based on gender

3.4 BOX PLOTS

3.4.1 Passengers' age

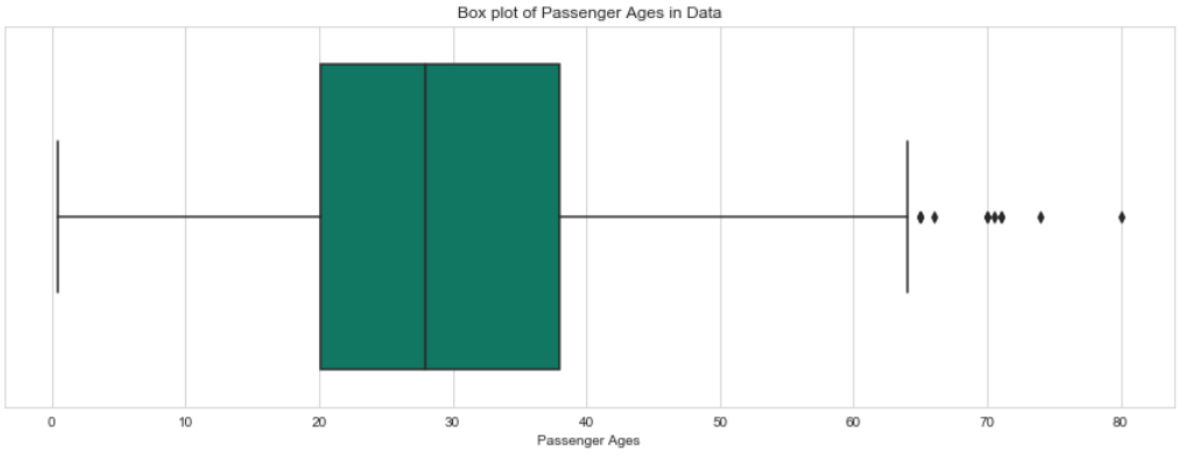

Figure 21 Box plot of passengers' ages

MEASURES OF CENTRAL TENDENCY:
Mean Passenger Age : 29

FIVE NUMBER SUMMARY OF THE AGE DATA:
Minimum Age in Data : 0.42
First Quartile (Q1) : 20.125
Median (Q2) : 28.0
Third Quartile (Q3) : 38.0
Maximum Age in Data : 80.0

3.5 SCATTER PLOTS

3.5.1 Scatter plot showing fare with respect to age

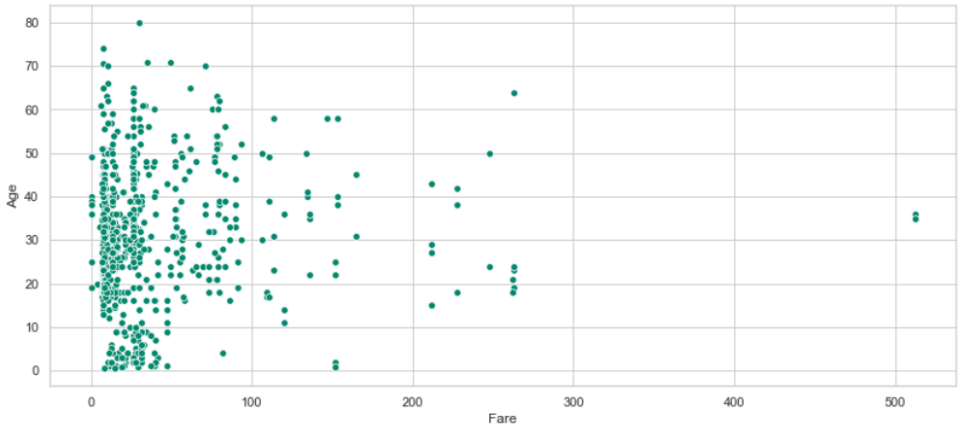

Figure 22 Scatter plot of Fare versus Age

Abdullah Aslam 01-134172-005 ShAbdullahAslam@gmail.com
Ahsan Goheer 01-134172-008 AhsanGoheer@live.com

4 REFERENCES.

[1] Dataset link : https://www.kaggle.com/c/titanic/data