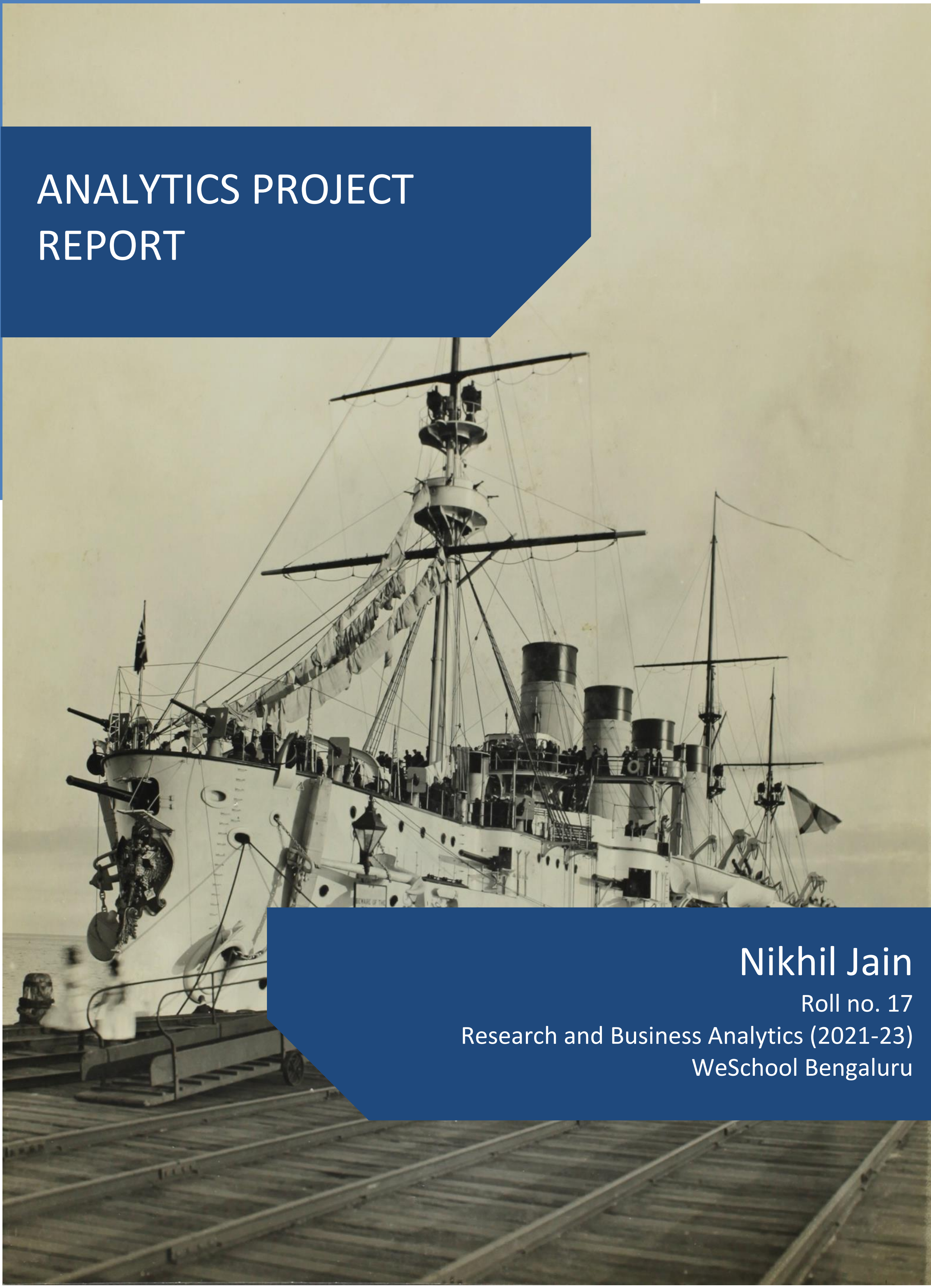
ANALYTICS PROJECT REPORT



Nikhil Jain
Roll no. 17
Research and Business Analytics (2021-23)
WeSchool Bengaluru

Table of Contents

Introduction

The RMS Titanic sank in the early morning hours of 15 April 1912 in the North Atlantic Ocean, four days into her maiden voyage from Southampton to New York City.

Titanic had an estimated 2,224 people on board when she struck an iceberg at around 23:40 (ship's time) on Sunday, 14 April 1912.Her sinking two hours and forty minutes later at 02:20 (ship's time; 05:18 GMT) on Monday, 15 April, resulted in the deaths of more than 1,500 people, making it one of the deadliest peacetime maritime disasters in history.

> *“I thought her unsinkable and I based my opinion on the best expert advice. "*
>
> *Phillip Franklin, White Star Line Vice President*

Data Source: https://www.kaggle.com/hesh97/titanicdataset-traincsv

1. Project Objective

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg.

Unfortunately, there weren’t enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. The objective of the project involves answering the question “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).
The Project is about Exploratory Data Analysis (EDA) with the help of R Language or R Studios as a Tool to identify the answers.

2. Assumptions

It has been assumed that the fare is in $.

3. Exploratory Data Analysis

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

In R-Studios, the following libraries are used to explore the data.

Library	What it is used for?
lattice	It attempts to improve on base R graphics by providing better defaults and the ability to easily display multivariate relationships.
e1071	It provides functions for statistic and probabilistic algorithms like a fuzzy classifier, naive Bayes classifier, etc. Functions such as skewness and kurtosis are used from this library.
ggplot2	It is used for statistical computing and data representation using data visualization. Histogram, Distribution charts are made with the help of this.
esquisse	Esquisse package helps to explore and visualize your data interactively. It is a Shiny gadget to create ggplot charts interactively with drag-and-drop to map your variables

3.1.2 Set up working Directory

Setting a working directory at starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

The working directory in this project is -

```
setwd("E:/Welingkar Tri 2/R_We/R/WE_ASN_WD")
```

Please refer to Appendix A for Source Code.

3.1.3 Import and Read the Dataset

The given dataset is in .csv format with the name “train.csv”. Hence, the command ‘read.csv’ is used for importing the file.

A variable “df” is created to read all the data in this variable. df variable will be used in the complete project to call the dataset.

```
df = read.csv("train.csv")
```

Please refer Appendix A for Source Code.

3.2 Variable Identification

3.2.1 Variable Identification

Defining the variables what does it mean. Defining the data type of each variable from the output of R function str(df). Classifying the type of the variable being continuous, discrete, dependent, and independent. Key represents the notation of numeric values of variables data of 0,1,2,3. The representation of such values in the literal meaning.

Variable	Definition	Variable Identification	Data Type	Key
survival	Survival	Dependent, Discrete	int	0 = No 1 = Yes
pclass	Ticket class	Independent, Discrete	int	1 = 1st 2 = 2nd 3 = 3rd
sex	Sex	Independent, Discrete	Chr	
Age	Age in years	Independent, Continuous	int	
sibsp	# of siblings/spouses aboard the Titanic	Independent, Discrete	int	
parch	# of parents/children aboard the Titanic	Independent, Discrete	int	
ticket	Ticket number	Independent, Discrete	Chr	
fare	Passenger fare	Independent, Continuous	num	
cabin	Cabin number	Independent, Discrete	Chr	

chr = character
int = integer
num = number

3.2.2 Summary Functions

List of functions used to explore the data for a better understanding of the data. The output obtained and the inference of such output.

Function Name	Output	Inference
View(df)	Display the dataset in table format in a new window.	Helps in viewing the data in a smooth table format.
dim(df)	891 11	There are 891 observations and 11 variables in the dataset.
head(df)	Display the top 5 rows with all columns.	Helps to quick overview the data.
tail(df)	Display the bottom 5 rows with all columns.	Helps to quick overview the data from last.
str(df)	data.frame': 891 obs. of 11 variables: $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ... $ Survived : int 0 1 1 1 0 0 0 0 1 1 ... $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ... $ Name : chr "... $ Sex : chr "male" "female" "female" $ Age : num 22 38 26 35 35 NA 54 2 $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ... $ Parch : int 0 0 0 0 0 0 0 1 2 0 ... $ Ticket : chr "A/5 21171" "PC 17599" $ Fare : num 7.25 71.28 7.92 53.1 8.05 ... $ Cabin : chr "" "C85" "" "C123" ...	We can understand the structure of the data with this function. We can understand the data types of each variable. We can understand that what type of graphical representation we can perform by using which variable to better understand the data. For example: Survived is int and Sex is chr, so we can plot a bar graph by count.
names(df)	"PassengerId" "Survived" "Pclass" "Name" "Sex" "Age" "SibSp" "Parch" "Ticket" "Fare" "Cabin"	We can see all the names of all the variables. It helps in understanding all the variables present in the dataset at a glance. It helps to quickly refer to names while writing the code.

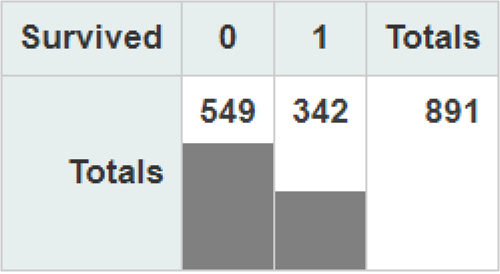
summary(Fare)
summary(Age)
summary(Sex)
	> summary(Fare)
	Min. 1st Qu. Median Mean 3rd Qu. Max.
	0.00 7.91 14.45 32.20 31.00 512.33
	> summary(Age)
	Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
	0.42 20.12 28.00 29.70 38.00 80.00 177
	> summary(Sex)
	Length Class Mode
	891 character character
		Summary function provides the 5 Point summary of quantitative variable along with some other insights too.
		Summary of fare helps us to understand that mean is 32.20 while median is 14.45, which means there should be outliers in the fare. The range is 512.33 - 0 = 512, which means the data is very skewed.
		Summary of Age provides the insights that the data is centred between somewhat 30 Age. There are 177 NA’s in the Age column.

3.3 Uni-variate Analysis

Univariate data consists of **only one variable**. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

SURVIVED

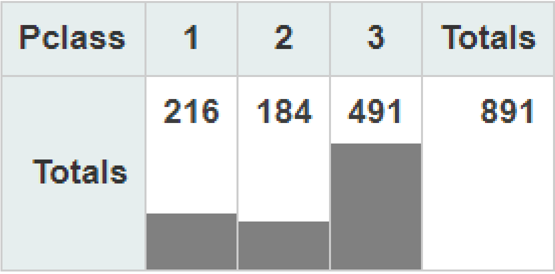
We have a total 891 observations. We know that the values of 0 and 1 represents people survived and not survived. Simple table representation shows us that the people not survived are much higher than survived.



Survived	0	1	Totals
Totals	549	342	891

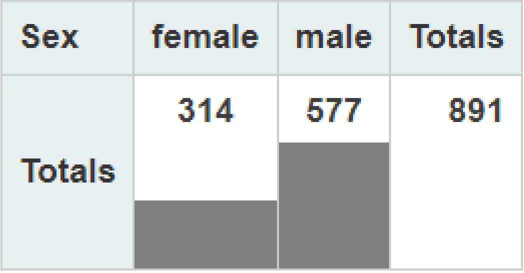
Pclass (Passenger Class)

In Passenger class there are 3 different class. Highest number of passengers are in class 3 followed by class 1 and lowest being class 2. There are almost Double the passengers in class 3 then class 1 and 2. Surprisingly, more people in Pclass 1 than 2. The possible reason of higher number of passengers in class 3 could be the low Fair of class 3.



Pclass	1	2	3	Totals
Totals	216	184	491	891

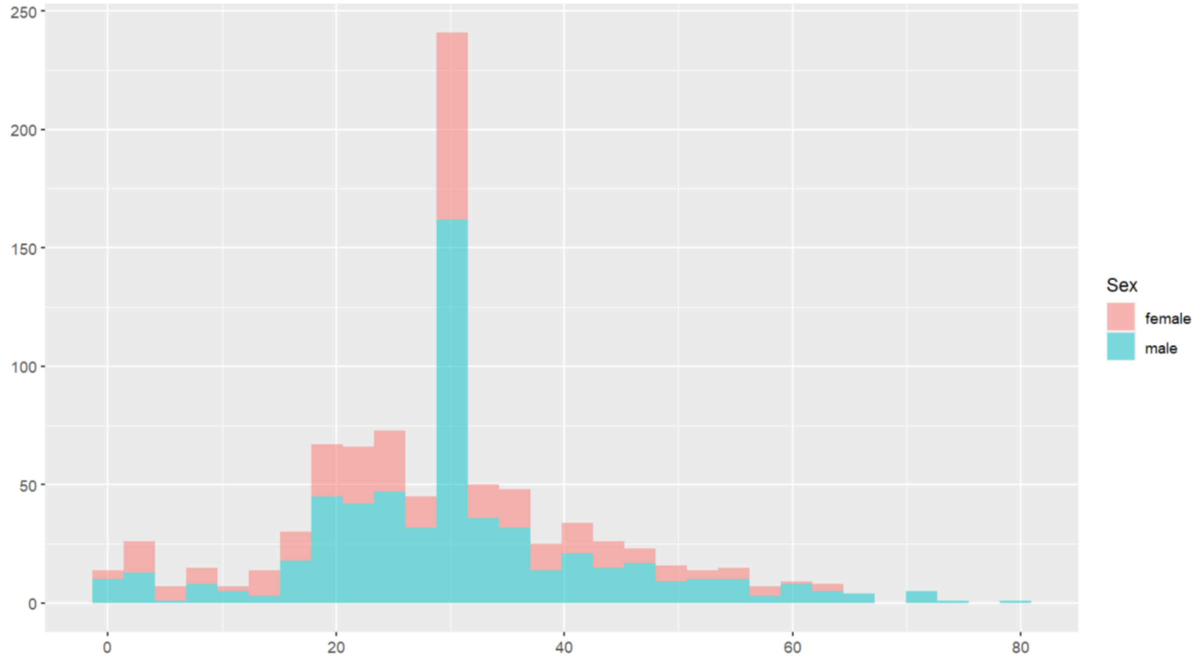
Sex (Gender) – There are 577 males and 314 females out of total passengers. The ration of males is nearly double than females.



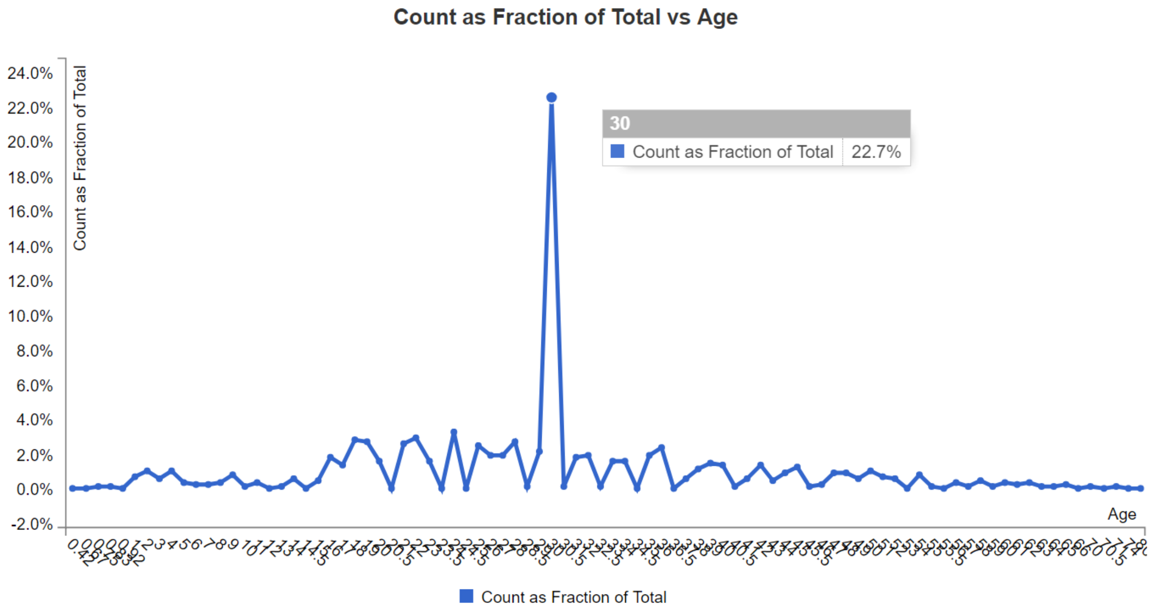
Sex	female	male	Totals
Totals	314	577	891

Age – Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Most of the passengers belongs to 20-40 Age Interval. There are very less passengers above 60.



Age (Line Chart) – At the age of 30 there is 22.7 % of all age groups. There was more young passengers in the ship.

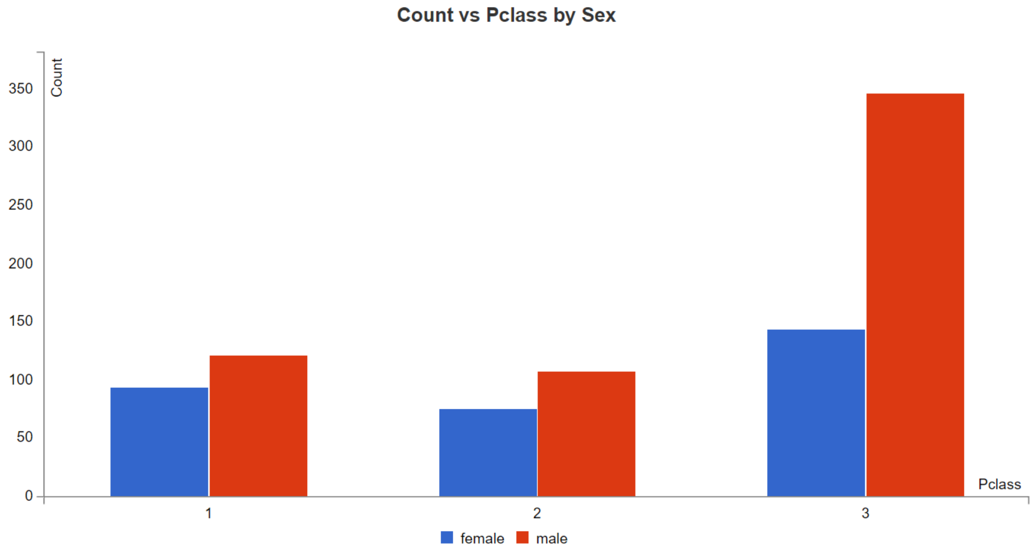


3.4 Bi-Variate Analysis

Bi- variate data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.

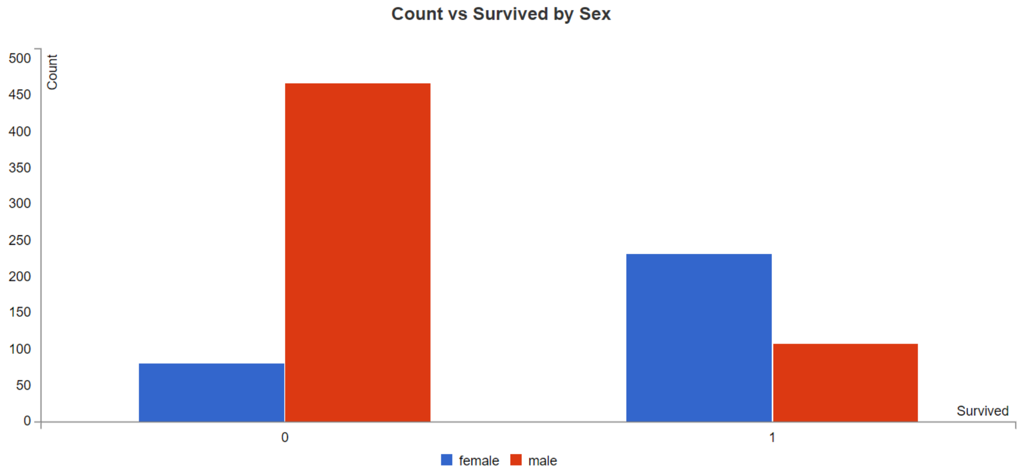
Pclass vs Sex

There are more men in all Pclass. But the ratio of men to women is a lot higher in Pclass 3.



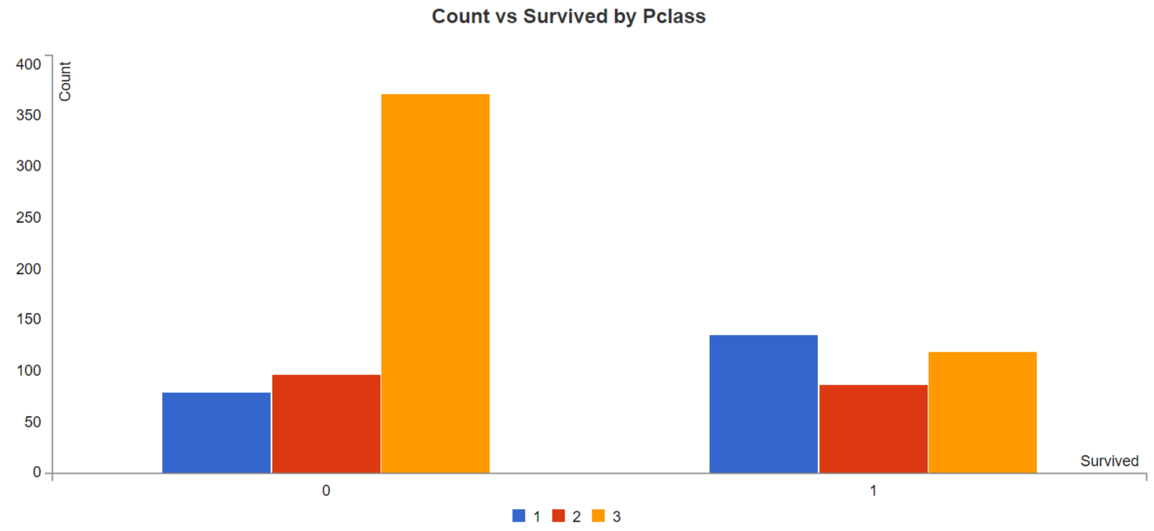
Survived vs Sex

There are more deaths of men’s and the survival of females are higher than men. In above graphs we saw that there are more men’s in class 3 than any other class.



Survived vs Pclass

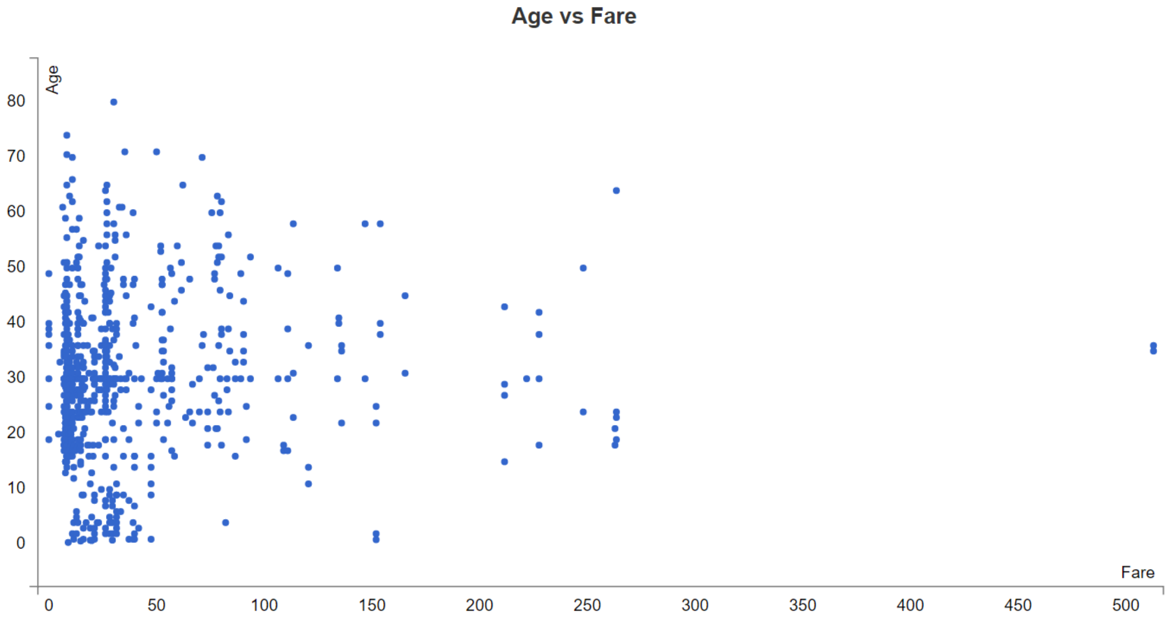
Those who paid more prioritized for rescue. People in Pclass 1 had much better chances of survival than other classes.



Pclass	Survived	Died	Total	Proportion Survived
1	136	80	216	0.63
2	87	97	184	0.47
3	119	372	491	0.24

Age vs Fare (Scatter Plot)

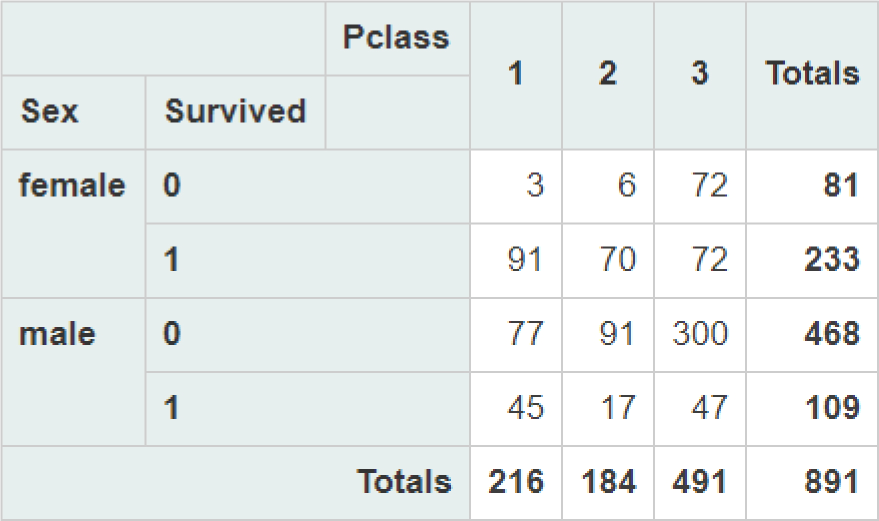
A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present. We see that most of the passengers opt for fare between 0-50 irrespective of their age.



3.4 Multi -Variate Analysis

Pclass – Sex – Survived

Multi variate analysis depicts the relationship between more than two variables. There are only 3 deaths of female from Class 1



		Pclass	1	2	3	Totals
Sex	Survived					
female	0		3	6	72	**81**
	1		91	70	72	**233**
male	0		77	91	300	**468**
	1		45	17	47	**109**
Totals			**216**	**184**	**491**	**891**

3.5 Missing Value Identification

Functions to Find If there is NA/Blanks in the columns.

Output of console

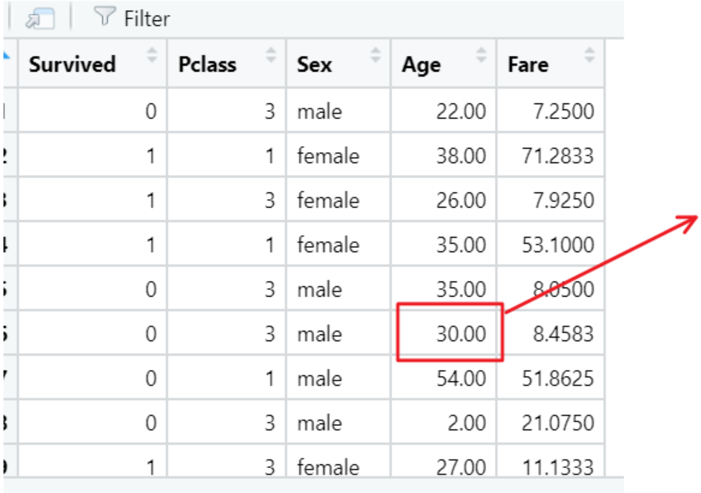
```
> sum(Survived!="")
[1] 891
> sum(Pclass!="")
[1] 891
> sum(Sex!="")
[1] 891
> sum(Age!="")
[1] NA
```

```
> sum(is.na(Age))
[1] 177
> sum(Ticket!="")
[1] 891
> sum(Fare!="")
[1] 891
> sum(Cabin!="")
[1] 204
```

Here we can see that CABIN has only 177 Values and remaining 891-204 = 687 values are missing, Hence 77% values are missing, Thus it is much better to drop this column for EDA purpose.

In age 263 values are missing, this is 19.86% of the dataset. Since Age as a Variable is a important for EDA as per my general opinion hence we will use mean method to fill all the NA's.

With summary function we can see that the MEAN Age is 29.70, So we will take 30 as our MEAN age to fill all the NA.



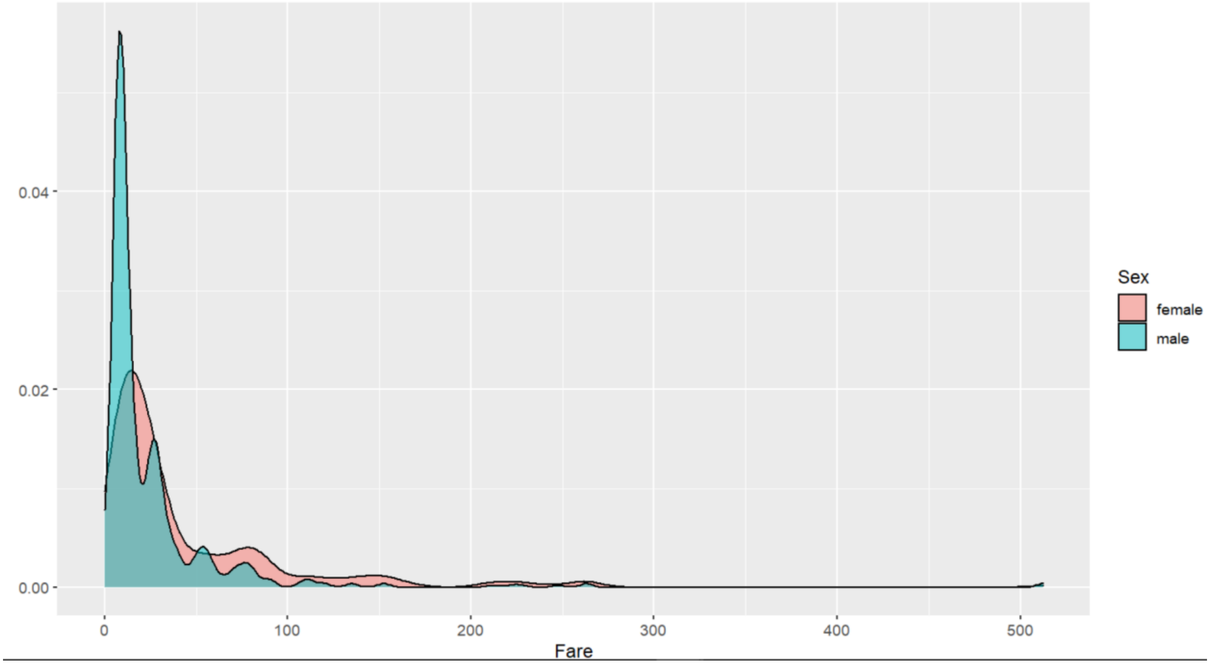
Survived	Pclass	Sex	Age	Fare
0	3	male	22.00	7.2500
1	1	female	38.00	71.2833
1	3	female	26.00	7.9250
1	1	female	35.00	53.1000
0	3	male	35.00	8.0500
0	3	male	30.00	8.4583
0	1	male	54.00	51.8625
0	3	male	2.00	21.0750
1	3	female	27.00	11.1333

Filled NA with 30 as Mean Age

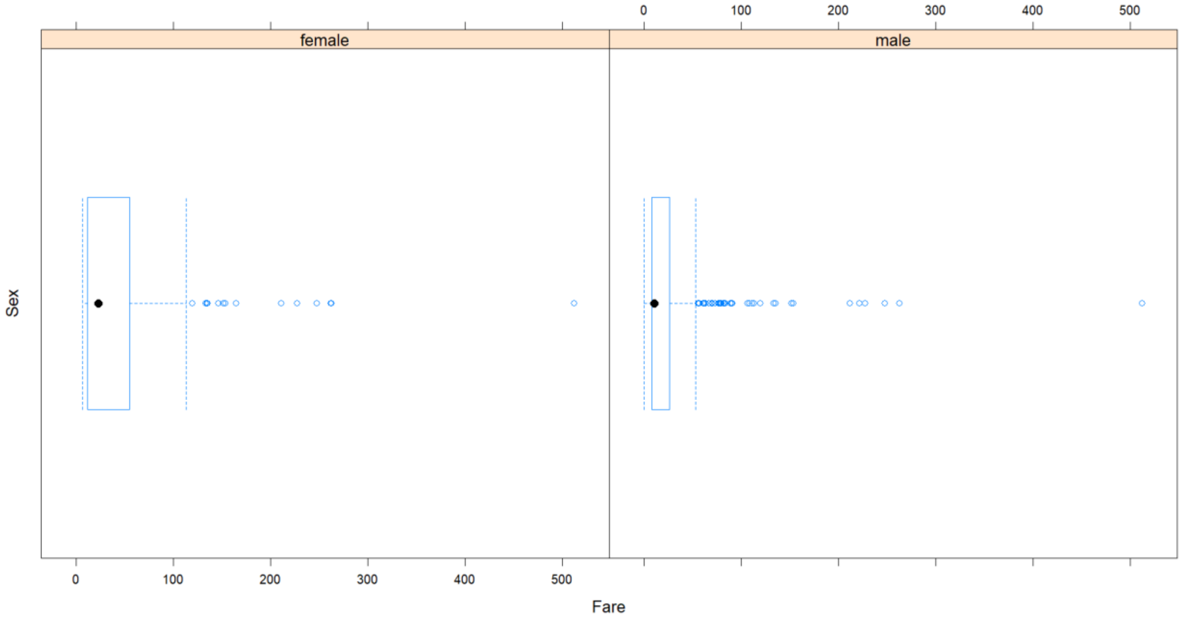
3.6 Outlier Identification

With skewness function from e1071 library, we can see that the skewness in fare is 4.77.

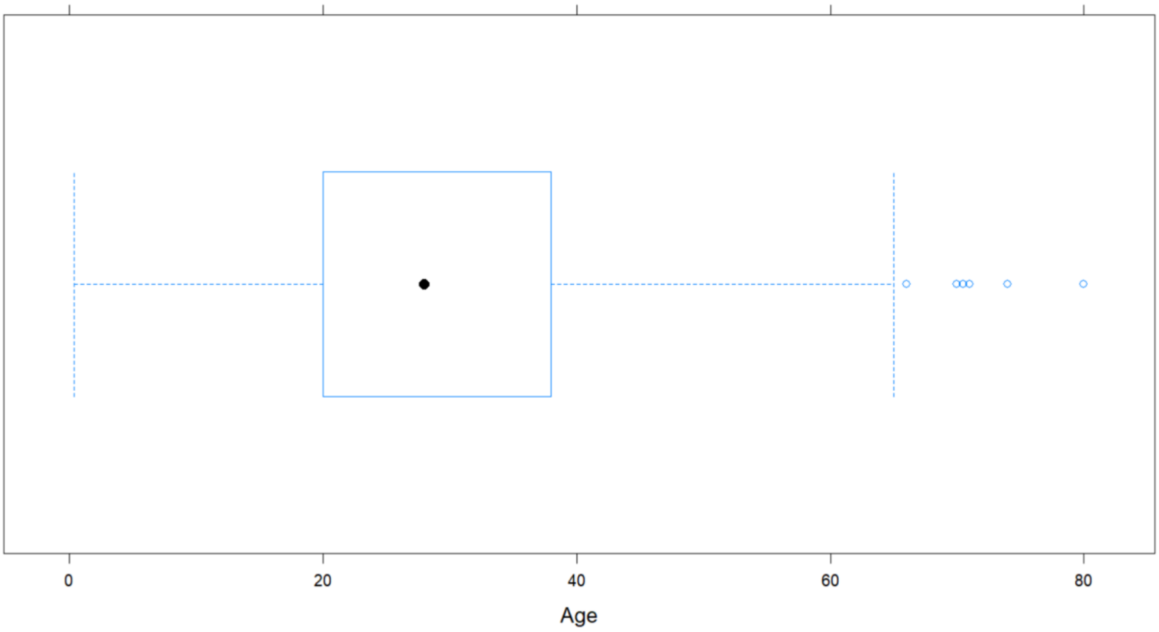
The fare is highly right skewed and the number of outliers are many in fare. We can also infer that the Fare of males is more skewed than female.



In the below graph we can see the number of outliers present in fare in female and male fairs. The Dot represent the outliers in the data.



We have filled all the NA's in Age. After NA fill , there are some outliers present in Age also, which represents there are few people of age that were not in common among other members. The people above 65 (approx.) are considered as outliers.



4 Conclusion

- If you were a male in the Titanic, the best chances of survival would be if you were:
  - A child not in third class
  - A first class young adult
  - A first class middle aged adult
- The best chances for survival overall are if you were a female not in third class.
- The Majority of people belong to 20-40 Age group.
- Men survived less as compared to woman.
- The higher the class number (the cheaper the fare), the more people died.

5 Appendix A

```
setwd("E:/Welingkar Tri 2/R_We/R/WE_ASN_WD")


library(rpivotTable)
library(lattice)
library(e1071)
library(ggplot2)
library(esquisse)


esquisser(viewer = "browser")

df = read.csv("train.csv")
View(df)
attach(df)

dim(df)
head(df)
tail(df)
str(df)

#EDA
names(df)
summary(df)
#which(is.na(Cabin))

#Find If there is NA/Blanks in the columns

#Code to find how many non empty cells are there
sum(df$PassengerId!="")
sum(Survived!="")
sum(Pclass!="")
sum(Sex!="")
sum(Age!="")
sum(is.na(Age))
sum(Ticket!="")
sum(Fare!="")
sum(Cabin!="")

#Here we can see that CABIN has only 177 Values and remaining 891-204 = 687 values are
missing, Hence 77% values are missing, Thus
#it is much better to drop this column for EDA purpose

df = subset(df,select = -c(Cabin,Name,SibSp,Parch,PassengerId,Ticket))
View(df)
```

```
#Since Age as a Variable is a important for EDA as per my general opinion and 19.86% Values are missing, hence we will use mean method
#to fill the NA Values

summary(Age)
#We can see that the MEAN Age is 29.70, So we will take 30 as our MEAN age to fill all the NA.
df[is.na(df)] <- 30

df2 = df
View(df)

#measures of central Tendency - Mean, Median
mean(Age)
median(df$Age)

#measures of dispersion
#range - gives min and max values - base
range(Fare)

#std deviation -
sd(Fare)

#variance
var(Fare)

#quartile deviation - quantile function is equal to quartile
quantile(Fare)

# measures of shape

skewness(df$Fare)
range(Fare)

skewness(df$Age)
kurtosis(Fare)



# Univariate analysis
table(Survived)
table(Pclass)



#Bivariate analysis
table(Survived,Sex)
table(Pclass,Survived)
table(Pclass,Sex)


##Multivariate analysis
table(Pclass, Survived,Sex)

#Lattice Package

histogram(Age)
histogram(Fare)
histogram(~Age|Sex, data = df)


bwplot(~Fare)

#Boxplot by Categories
bwplot(~Fare|Sex, data = df, xlab = "Fare", ylab = "Sex")
bwplot(~Age)

#Rpivot Table for Visualization
rpivotTable(df)


#kernel density plot
qplot(Fare, data = df, geom = "density", fill=Sex, alpha=I(.5))
qplot(Age, data = df, geom = "histogram", fill=Sex, alpha=I(.5))
```

```
#======================================================================  #
```