```
data("titanic_train")
titanic <- titanic_train
```

We use the above code to load the titanic_train dataset to R and make a copy of it into "titanic". We make a copy of the dataset into titanic because "titanic" is easier and fast to write instead of "titanic_train"

```
dim(titanic)
```

We use the above code to show the number of row and column in titanic dataset.

Output:

```
> dim(titanic)
[1] 891  12
```

From the output, we know that the titanic dataset has 12 columns and 891 rows

```
colnames(titanic)
```

We use the above code to show the name of each column.

Output:

```
> colnames(titanic)
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"         "SibSp"
 [8] "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
```

By reading throught the output, we know that the titanic dataset is a dataset that has information about the passenger of titanic such as the class, name, sex, age, cabin, and also provides information on their survival status (whether they survived or did not survive)

```
head(titanic)
tail(titanic)
```

We use **head(titanic)** to show us the first six rows from titanic dataset and we use **tail(titanic)** to show us the last six rows from our titanic dataset.

Output:

```
  PassengerId Survived Pclass                                                 Name    Sex
1           1        0      3                              Braund, Mr. Owen Harris   male
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3           3        1      3                               Heikkinen, Miss. Laina female
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5           5        0      3                             Allen, Mr. William Henry   male
6           6        0      3                                     Moran, Mr. James   male
  Age SibSp Parch            Ticket    Fare Cabin Embarked
1  22     1     0         A/5 21171  7.2500              S
2  38     1     0          PC 17599 71.2833   C85        C
3  26     0     0 STON/O2. 3101282  7.9250              S
4  35     1     0            113803 53.1000  C123        S
5  35     0     0            373450  8.0500              S
6  NA     0     0            330877  8.4583              Q
```

```
> tail(titanic)
    PassengerId Survived Pclass                                      Name    Sex Age SibSp Parch
886         886        0      3     Rice, Mrs. William (Margaret Norton) female  39     0     5
887         887        0      2                    Montvila, Rev. Juozas   male  27     0     0
888         888        1      1             Graham, Miss. Margaret Edith female  19     0     0
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  NA     1     2
890         890        1      1                    Behr, Mr. Karl Howell   male  26     0     0
891         891        0      3                      Dooley, Mr. Patrick   male  32     0     0
        Ticket   Fare Cabin Embarked
886     382652 29.125              Q
887     211536 13.000              S
888     112053 30.000   B42        S
889 W./C. 6607 23.450              S
890     111369 30.000  C148        C
891     370376  7.750              Q
```

By reading through the output, we know that the output briefly informs us that our dataset contains both numeric and character data

str(titanic)

We use the above code to show us the structure of the titanic dataset. It show us the number of row and column, the data types of each column, and also give us the example data for each column/variable.

Output:

```
> str(titanic)
'data.frame':   891 obs. of  12 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  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkin
en, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ 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" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

By reading through it, we know that the variables that has “character/factor” as it’s data type is Name, Sex, Ticket, Cabin, and Embarked. The variables that has “numeric” as it’s data type is PassengerId, Survived, Pclass, SibSp, Parch, and Fare.

View(titanic)

We use the above code to open a new window that show us all the data from titanic dataset.

Output:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	237736	30.0708		C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	PP 9549	16.7000	G6	S
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	113783	26.5500	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	A/5. 2151	8.0500		S
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	347082	31.2750		S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	350406	7.8542		S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	248706	16.0000		S
17	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	382652	29.1250		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	244373	13.0000		S

By reading through the output, we know that the **titanic** dataset contains information about the passengers on the Titanic, including

- **PassengerId** --> Passenger Id
- **Survived** --> survival status (1 = survive, 0 = not survive)
- **Pclass** --> Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Name** --> Passenger Name
- **Sex** --> Passenger Gender
- **Age** --> Passenger Age
- **SibSp** --> Number of siblings aboard
- **Parch** --> Number of parents/children aboard
- **Ticket** --> Passenger Ticket
- **Fare** --> Fare paid
- **Cabin** --> Cabin number
- **Embarked** --> Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

```
table(titanic$Survived)
table(titanic$Pclass)
table(titanic$Sex)
table(titanic$SibSp)
table(titanic$Parch)
table(titanic$Embarked)
```

We use the above code to counts the number of each unique value in each variable

Output:

```
> table(titanic$Survived)

  0   1
549 342
> table(titanic$Pclass)

  1   2   3
216 184 491
> table(titanic$Sex)

female   male
   314    577
> table(titanic$SibSp)

  0   1   2   3   4   5   8
608 209  28  16  18   5   7
> table(titanic$Parch)

  0   1   2   3   4   5   6
678 118  80   5   4   5   1
> table(titanic$Embarked)

      C   Q   S
  2 168  77 644
```

By reading through the output we know that

- **table(titanic$Survived)** counts the number of instances of each unique value in variable/column "Survived", which represents whether a passenger survived or not (1 = Survive, 0 = not survive)
- **table(titanic$Pclass)** counts the number of instances of each unique value in variable/column "Pclass", which represents the passenger class (1 = 1st class, 2 = 2nd class, 3 = 3rd class).
- **table(titanic$Sex)** counts the number of instances of each unique value in variable/column "Sex", which represents the gender of the passenger (male or female).
- **table(titanic$SibSp)** counts the number of instances of each unique value in variable/column "SibSp", which represents the number of siblings aboard

the Titanic (min: 0, max: 8)

- **table(titanic$Parch)** counts the number of instances of each unique value in variable/column "Parch", which represents the number of parents/children aboard the Titanic (min: 0, max: 6)
- **table(titanic$Embarked)** counts the number of instances of each unique value in variable/column "Embarked", which represents the port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

```
summary(titanic)
```

We use the above code to show us the summary from all the variables in titanic dataset.

Output:

```
> summary(titanic)
  PassengerId       Survived          Pclass          Name               Sex                 Age       
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00  
                                                                                        NA's   :177    
     SibSp           Parch           Ticket               Fare            Cabin          
 Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00   Length:891        
 1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character  
 Median :0.000   Median :0.0000   Mode  :character   Median : 14.45   Mode  :character  
 Mean   :0.523   Mean   :0.3816                      Mean   : 32.20                     
 3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00                     
 Max.   :8.000   Max.   :6.0000                      Max.   :512.33                     

   Embarked        
 Length:891         
 Class :character   
 Mode  :character   
```

By reading through the output, we know that all the variables that has numeric as its data type such as Age, Fare, SibSP, and Parch will show us its minimum, 1st quartile, median, mean, 3rd quartile, maximum values, and number of missing values.

For categorical variables like Name, Sex, and Embarked the summary() function will only show us the numbers of data and its data type. I don't want this so we have to convert the variables into factor with the code below.

```
titanic$Survived = as.factor(titanic$Survived)
titanic$Sex <- as.factor(titanic$Sex)
titanic$Embarked <- as.factor(titanic$Embarked)
summary(titanic)
```

These above lines of code convert the data type of certain variables in the titanic dataset to the factor data type. I do this because I want the summary show us the number of class in each variable. Eventhough the Survived variable is numeric, but its only have two class and its represent the survival state (0 = not survive, 1 = survive) so we have to convert it into factor.

Output:

```
> summary(titanic)
  PassengerId     Survived      Pclass          Name               Sex           Age            SibSp
 Min.   :  1.0   0:549    Min.   :1.000   Length:891         female:314   Min.   : 0.42   Min.   :0.000
 1st Qu.:223.5   1:342    1st Qu.:2.000   Class :character   male  :577   1st Qu.:20.12   1st Qu.:0.000
 Median :446.0            Median :3.000   Mode  :character               Median :28.00   Median :0.000
 Mean   :446.0            Mean   :2.309                                  Mean   :29.70   Mean   :0.523
 3rd Qu.:668.5            3rd Qu.:3.000                                  3rd Qu.:38.00   3rd Qu.:1.000
 Max.   :891.0            Max.   :3.000                                  Max.   :80.00   Max.   :8.000
                                                                         NA's   :177
     Parch           Ticket               Fare            Cabin           Embarked
 Min.   :0.0000   Length:891         Min.   :  0.00   Length:891          :  2
 1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   C:168
 Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Q: 77
 Mean   :0.3816                      Mean   : 32.20                      S:644
 3rd Qu.:0.0000                      3rd Qu.: 31.00
 Max.   :6.0000                      Max.   :512.33
```

By reading through the output we know that, the number of passenger that's not survive is more than the number of passenger that's survive. We also know that in the ship there are more male passenger that female passenger. From the Embarked variable we know that there are 2 passenger that doesn't have any port of embarkation.

```
colSums(is.na(titanic))
```

We use **colSums(is.na(titanic))** to count the number of missing values in each column in **titanic** dataset. We use **is.na(titanic)** to make a matrix the same size of the dataset and if the element is NULL then the element will represent by TRUE. We use **colSums()** to sums up all the TRUE element in every column.

Output:

```
> colSums(is.na(titanic))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch
          0           0           0           0           0         177           0           0
     Ticket        Fare       Cabin    Embarked
          0           0           0           0
```

By reading through the output, we know that the titanic dataset has 177 missing values in only one variable, which is variable **Age.**

```
new_titanic = titanic[rowSums(is.na(titanic)) <=0, ]
colSums(is.na(new_titanic))
```

We use the code new_titanic = titanic[rowSums(is.na(titanic)) <=0,] create a new dataset called new_titanic by excluding all rows that contain missing values in titanic dataset. Then we use colSums(is.na(new_titanic)) to count the number of missing values in each column in the new_titanic dataset, we use this code again to check whether we correctly assign all rows in the titanic dataset that doesn't have missing values.

Output:

```
> new_titanic = titanic[rowSums(is.na(titanic)) <=0, ]
> colSums(is.na(new_titanic))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch
          0           0           0           0           0           0           0           0
     Ticket        Fare       Cabin    Embarked
          0           0           0           0
```

By reading through the output, we know that the new_titanic dataset doesn't has any missing value.

```
duplicated_rows <- duplicated(new_titanic)
sum(duplicated_rows)
```

We use the above code to count all the duplicated rows in new_titanic dataset. **Duplicated(new_titanic)** will return logical vector whether the row is duplicated or not then we assigned the vector into **duplicated_rows.** With **sum()** it will give us the number of rows that is duplicated.

Output:

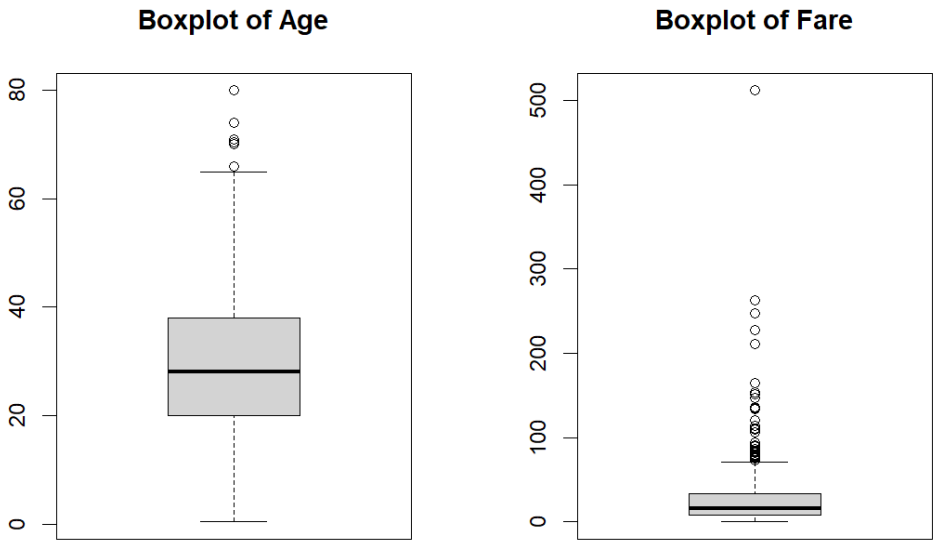
```
> duplicated_rows <- duplicated(new_titanic)
> sum(duplicated_rows)
[1] 0
```

By reading through the output, we know that the titanic dataset doesn't has duplicated rows.

```
var_num <- c("Age", "Fare")

par(mfrow = c(1,2))
for (i in 1:length(var_num)) {
  boxplot(new_titanic[, var_num[i]], main = paste("Boxplot
of", var_num[i]))
}
```

We use the above code to creates boxplots for the "Age" and "Fare" variables in the new_titanic dataset

Output:



From the plot above we can see that there are some outliers in variables Age and Fare. But we can see very well which values/elements is the outliers. To know which data is the outliers of variables Age and Fare we will try to print it using the code below.

```
for (i in 1:length(var_num)) {
```

```
  var <- var_num[i]
  boxplot.stats(new_titanic[, var], coef = 1.5)$out
  print(paste("Outliers     in",     var,     "variable:",
boxplot.stats(new_titanic[, var], coef = 1.5)$out))
}
```

We use the above code to print all the outliers of variables Age and Fare using **boxplot.stats()** function. The coefficient of 1.5 is used to define the whiskers of the boxplot, and any values outside of the whiskers are considered as the outliers.

Output:

```
[1] "Outliers in Age variable: 66, 71, 70.5, 71, 80, 70, 70, 74"
[1] "Outliers in Fare variable: 263, 82.1708, 76.7292, 80, 83.475, 73.5, 263, 7
7.2875, 247.5208, 73.5, 77.2875, 79.2, 146.5208, 113.275, 76.2917, 90, 83.475,
90, 86.5, 512.3292, 79.65, 153.4625, 135.6333, 77.9583, 78.85, 91.0792, 151.55,
247.5208, 151.55, 108.9, 83.1583, 262.375, 164.8667, 134.5, 135.6333, 153.4625,
134.5, 263, 75.25, 135.6333, 211.5, 227.525, 73.5, 120, 113.275, 90, 120, 263,
81.8583, 89.1042, 91.0792, 90, 78.2667, 151.55, 86.5, 108.9, 93.5, 106.425, 10
6.425, 110.8833, 79.65, 110.8833, 79.65, 79.2, 78.2667, 153.4625, 77.9583, 76.7
292, 73.5, 113.275, 133.65, 73.5, 512.3292, 76.7292, 211.3375, 110.8833, 227.52
5, 151.55, 227.525, 211.3375, 512.3292, 78.85, 262.375, 86.5, 120, 77.9583, 21
1.3375, 79.2, 120, 93.5, 80, 83.1583, 164.8667, 83.1583"
```

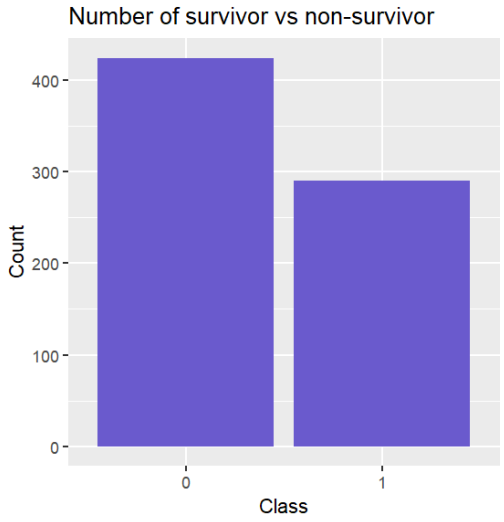
These values are considered outliers because they are outside the upper or lower fence of the boxplot. By reading throught it, we know that variable Age has seven outliers, while variable Fare has many outliers.

```
counts <- table(new_titanic$Survived)

ggplot(data = data.frame(Class = names(counts), Count =
as.vector(counts)), aes(x = Class, y = Count)) +
  geom_bar(stat = "identity", fill = "slateblue") +
  labs(x = "Class", y = "Count", title = "Number of
survivor vs non-survivor")
```

we use this code to create a bar plot using ggplot() to show us the comparison of passengers who survived and the passenger who doesn't survive.

counts <- table(new_titanic$Survived) used to count and assign the number of survivor and non-survivor into a table

Output:



By looking at the output we can see that the number of passenger who doesn't survive is more bigger than the number of passenger that is survived.

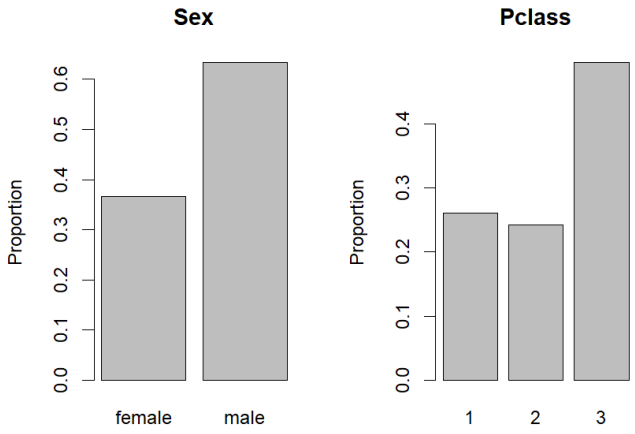
```
columns <- c('Sex', 'Pclass', 'Age', 'SibSp', 'Parch', 'Embarked')

for(i in columns){
  freq_table <- prop.table(table(new_titanic[,i]))
  barplot(freq_table, main=i, ylab="Proportion")
}
```

we use this code to see the distribution of values in column Sex, Pclass, Age, SibSP, Parch, and Embarked. This code will plot the distribution in the form of barplot.

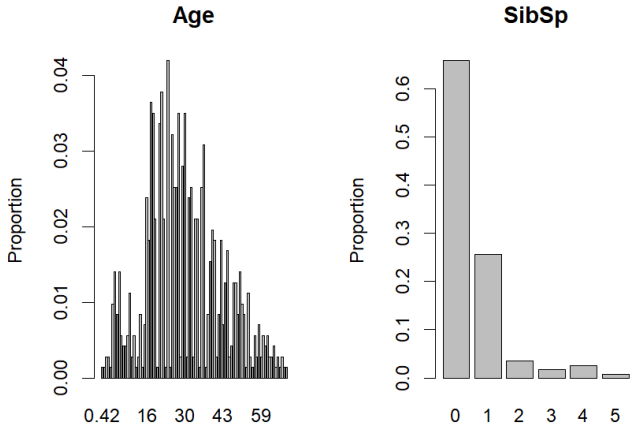
freq_table <- prop.table(table(new_titanic[,i])) used to create a frequency table of each value in column Sex, Pclass, Age, SibSP, Parch, and Embarked

barplot(freq_table, main=i, ylab="Proportion") used to create the barplot using the data in freq_table

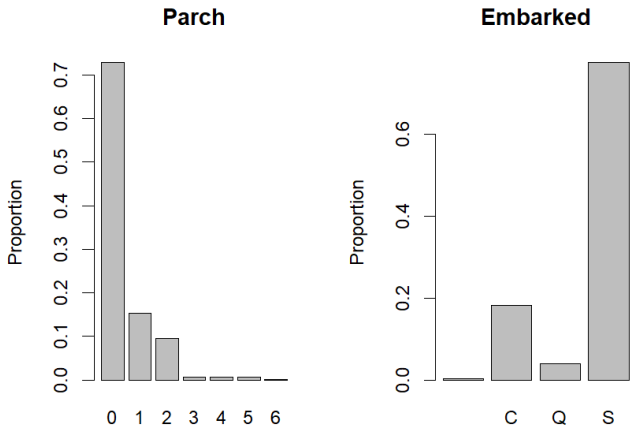
Output:



As we look at the plots above we can see the proposition of each value in variable Sex and Pclass. There are more female passenger than male passenger and there are more 3rd class passenger than 1st class passenger and 2nd class passenger (3rd > 1st > 2nd)



As we look at the plots above, we can see the proposition of each value in variable Age and SibSp. There are more passenger in age 16-43 and there are more passenger who travels alone or with one siblings/spouse using titanic than passenger who travels with > 1 siblings/spuse.



As we observe the plots above, we can see the proportion of each value in variables Parch and Embarked.

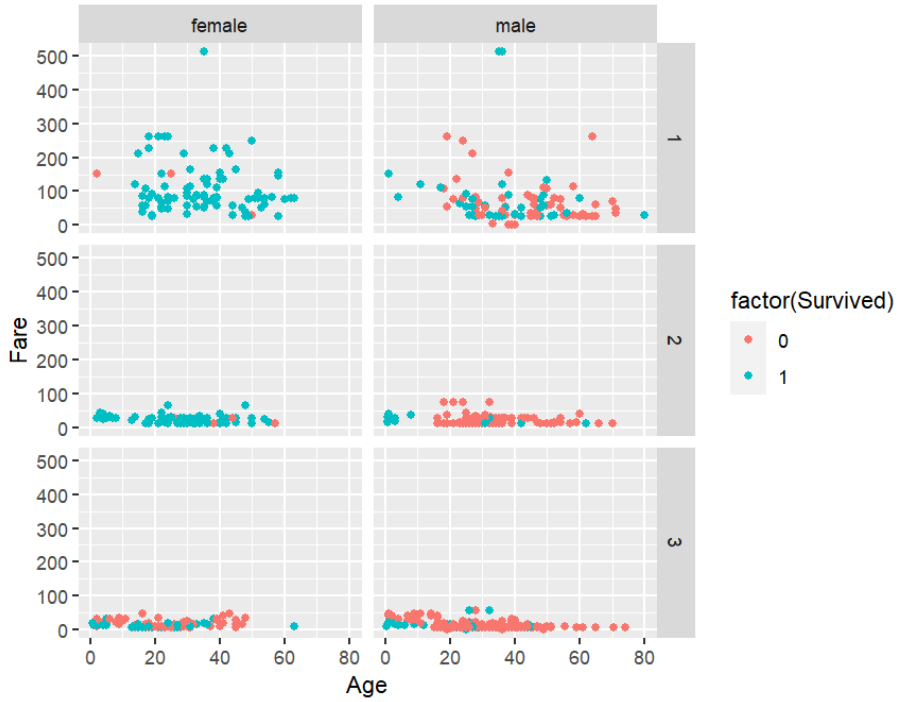
From the Embarked Plot it shows that more passengers embarked from Southampton than from other ports, with the order of embarkation port being S > C > Q. it also indicates that there are some passengers whose embarkation port data is missing.

Furthermore, we can also observe that the number of passengers traveling on the Titanic decreased as the number of parents/children accompanying them increased. Therefore, more passengers traveled alone rather than with their families.

```
ggplot(new_titanic, aes(Age, Fare)) + geom_point(aes(color = factor(Survived))) + facet_grid(Pclass ~ Sex)
```

We use the above code to create a scatter plot based on passenger Gender, Age, Fare, and Class. The color of each point represents whether the passenger survived or not (red = not survived, blue = survived)

Output:



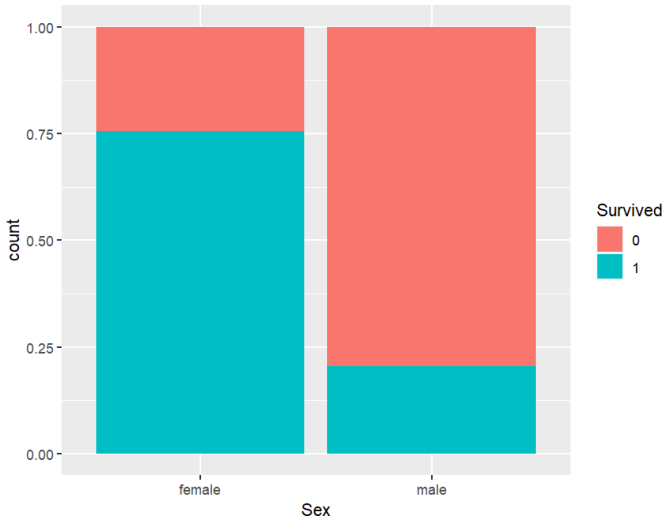
As we observe the plots above, we can see that Passenger Class and Passenger Gender is highly correlated with their survival state. while, Age and Passenger Fare seems doesn't really correlated with their survival state.

Females in 1st class have a higher chance of survival compared to other passenger classes and genders, while males in 3rd class have a higher chance of not surviving

```
ggplot(new_titanic, aes(Sex, fill = Survived)) + geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their gender. The bars are separated by gender (male and female) and stacked by survival state (survived and not survived).

Output:

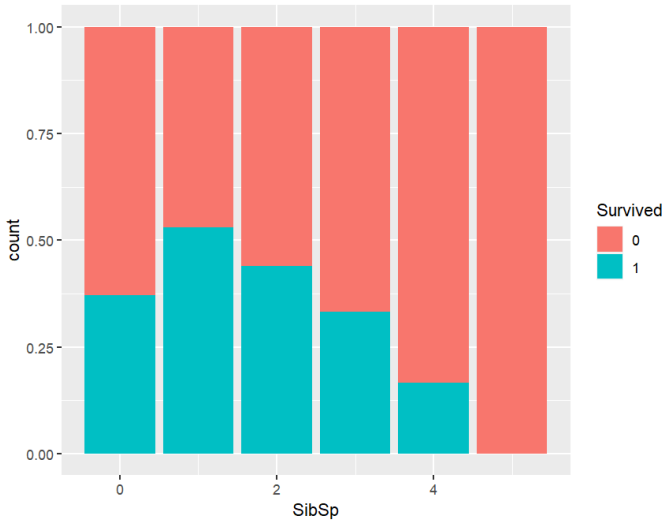


As we observe the plots above, we can see that there were more female passengers who survived than those who did not. We also know that, there were more male passengers who not survive than those who did. Therefore, we can conclude that gender is strongly correlated with their survival status.

```
ggplot(new_titanic, aes(SibSp, fill = Survived)) + geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their numbers of siblings/spouse who travels with them. The bars are separated by number of siblings/spouse who travels with them (0, 1, 2, 3, etc) and stacked by survival state (survived and not survived).

Output:

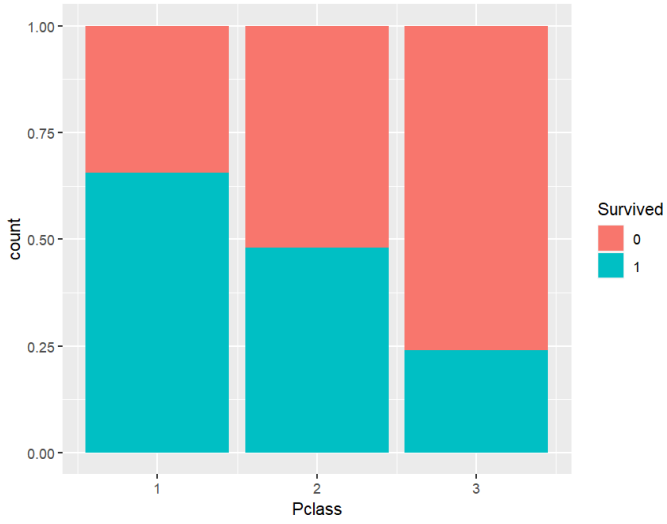


As we observe the plots above, we can see that passenger who travels with 3 or more siblings/spouse have a higher chance of not surviving because they have to help their siblings/spouse first. But passenger who travels with one or 2 siblings have a higher change of survival than passenger who travels alone because the one who travels alone doesn't have someone who help them.

```
ggplot(new_titanic, aes(Pclass, fill = Survived)) + geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their class. The bars are separated by class (1^{st} , 2^{nd} , 3^{rd}) and stacked by survival state (survived and not survived).

Output:

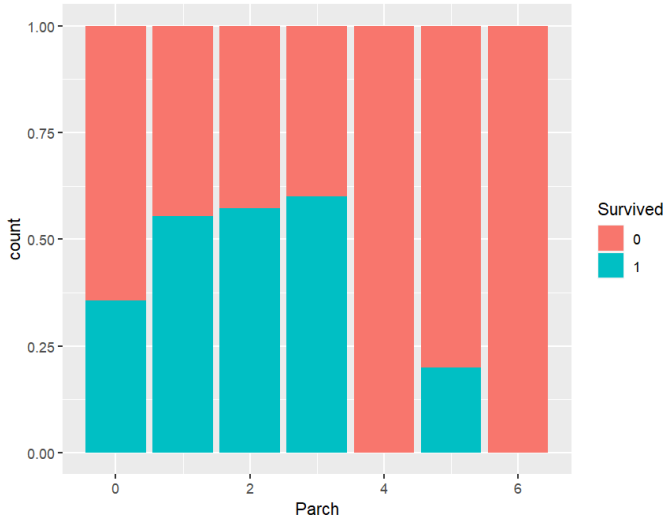


As we observe the plots above, we can see that the more lower your class (3rd class) the more lower your change of survival. Therefore, we can conclude that class is strongly correlated with their survival status. It is possible that first-class passengers were given priority during the evacuation process and they also have better access to lifeboats.

```
ggplot(new_titanic, aes(Parch, fill = Survived)) +
geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their numbers of parents/children who travels with them. The bars are separated by number of parents/children who travels with them (0, 1, 2, 3, etc) and stacked by survival state (survived and not survived).

Output:



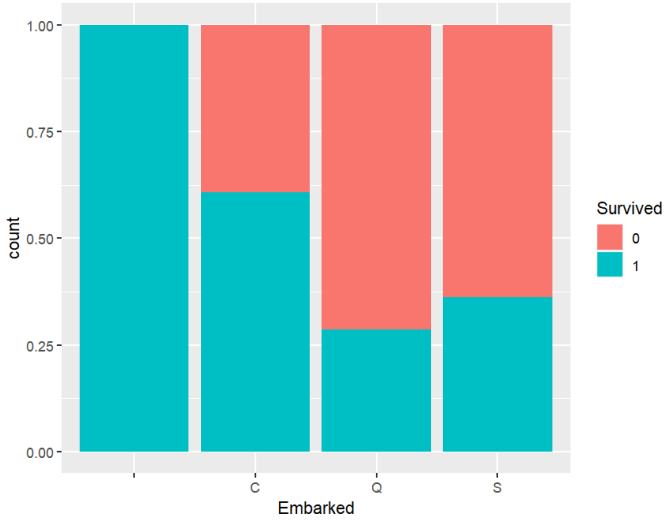
As we observe the plots above, we can see that passenger who travels with 3 or more parents/children have a higher chance of not surviving because they have to help their parents/children first. But passenger who travels with one, two, or 3 parents/children have a higher change of survival than passenger who travels alone

because the one who travels alone doesn't have someone who help them.

```
ggplot(new_titanic, aes(Embarked, fill = Survived)) +
geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their port embarktion. The bars are separated by the port (C = Cherbourg, Q = Queenstown, S = Southampton) and stacked by survival state (survived and not survived).

Output:



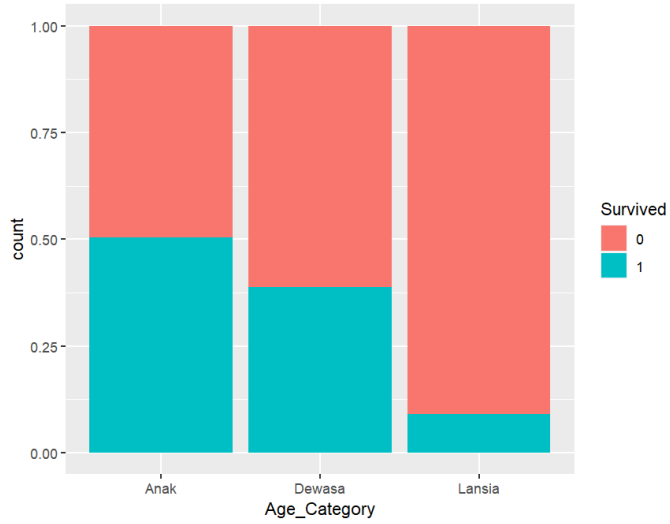
From the Embarked Plot it shows that more passengers embarked from Southampton than from other ports, with the order of embarkation port being S > C > Q. it also indicates that there are some passengers whose embarkation port data is missing.

As we observe the plots above, we can see that there were more passengers embark from Cherbough who survived than those who did not. We also know that, there were more passengers embark from Queenstown and Southampton who not survive than those who did. The order of survival being C > S > Q.

```
new_titanic$Age_Category <- cut(new_titanic$Age,
breaks=c(0, 18, 64, max(new_titanic$Age)),
labels=c("Anak", "Dewasa", "Lansia"))

ggplot(new_titanic, aes(Age_Category, fill = Survived)) +
geom_bar(position = "fill")
```

We use the code above to create a stacked bar plot that shows us propotion of passenger who survived and not survived by their Age. The bars are separated by Age and stacked by survival state (survived and not survived). But before we create a stacked bar, first we group the age into 3 groups. People who has age around 0-18 (Anak), age around 18-64 (Dewasa), and age > 64 (lansia)

Output:

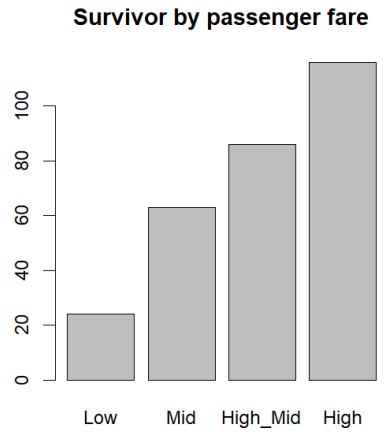


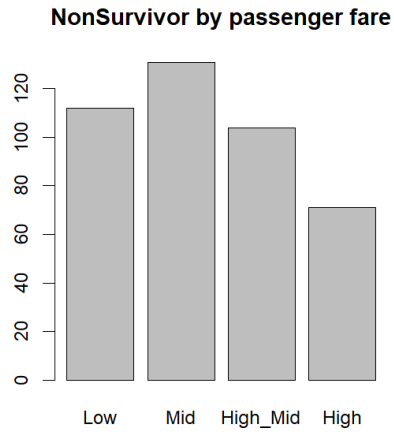
As we observe the plots above, we can see that the more young the passenger they have more change to survive. It is probably because people will try to help children first after that they will try to help themselves, so people who are older than 64> probably doesn't survive because they can't help themselves and they doesn't have someone to help them.

```
new_titanic$Fare_Category <- cut(new_titanic$Fare, breaks=c(0, 7.9, 14.45, 31.28, max(new_titanic$Fare)), labels=c("Low", "Mid", "High_Mid", "High"))

non_survivor = new_titanic[new_titanic$Survived == 0,]
survivor = new_titanic[new_titanic$Survived == 1,]

barplot(table(survivor$Fare_Category), main = "Survivor by passenger fare")
barplot(table(non_survivor$Fare_Category), main = "NonSurvivor by passenger fare")
```

We use the code above to create a bar that shows us the propotion of passenger who survived and not survived by their fare. But before we create the plot, first we assign data from new_titanic into survivor and non survivor and group the fare into Low, Mid, High_Mid, and High.

Output:





As we observe the plots above, we can see that the more high your Fare the more higher your change to survive and the more low your fare the more higher your change not surviving.

SUMMARY

- **titanic** dataset is a dataset that displays about the passenger of titanic such as the **class, name, sex, age, cabin,** and others**. This dataset is used to classify titanic passenger into 2 survival state,** namely **0 = not survive** or **1 = survive** based on the information given in the dataset
- Our dataset consists of 891 rows and 12 columns
- The variables that has "character" as it's data type is Name, Sex, Ticket, Cabin, and Embarked. The variables that has "numeric" as it's data type is PassengerId, Survived, Pclass, SibSp, Parch, and Fare.
- Our dataset has 177 missing value in Age column and doesn't has any duplicated rows
- We have seven outliers in column Age and many in column Fare
- In our dataset, there are more passenger that doesn't survive than the one who survive
- Passenger Class and Passenger Gender is highly correlated with their survival state. while, Age and Passenger Fare seems doesn't really correlated with their survival state.
- Female passenger are more likely to survive
- Young passenger are more likely to survive
- higher class (1^{st}) are more likely to survive
- The higher the fare the more likely to survive