American International University-Bangladesh (AIUB)

Department of Computer Science
Faculty of Science & Technology (FST)

Data Science Midterm Project Report

A Data Science Project Submitted By

Semester: Summer_23_24		**Section:**
SN	Student Name	Student ID
2	Khan, Tahfim Ibn	21-45719-3

Introduction:

In this project, we are using a modified version of the Titanic dataset, which contains information about the passengers on the Titanic, including their gender, age, sibsp, parch, fare, embarked, class, who, alone, and survival status. The dataset is prepared for analysis using various data preparation and descriptive statistics techniques.

Data Preparation Steps:

1. Library Installation and Loading: Installed and loaded necessary libraries (‘dplyr’, ‘tibble’, ‘ggplot2’, ‘ROSE’).

```
install.packages("dplyr")
install.packages("tibble")
install.packages("ggplot2")
install.packages("ROSE")

library(dplyr)
library(tibble)
library(ggplot2)
library(ROSE)
```

2. Loaded the dataset and checked its structure, data type and missing values.

```
titanic_Dataset <- read.csv("/Users/tahfimibnkhan/Desktop/Midterm_Project_Dataset_section(
str(titanic_Dataset)
colSums(is.na(titanic_Dataset))
```

```
> titanic_Dataset <- read.csv("/Users/tahfimibnkhan/Desktop/Midterm_Project_Dataset_section(B).c
sv")
> str(titanic_Dataset)
'data.frame':	105 obs. of  10 variables:
 $ Gender  : chr  "female" "female" "male" "male" ...
 $ age     : int  24 17 21 35 37 16 NA 33 40 28 ...
 $ sibsp   : int  0 0 0 0 0 0 1 0 0 0 ...
 $ parch   : int  0 0 0 0 0 0 0 2 0 0 ...
 $ fare    : chr  "7.7958" "8.6625" "7.75" "7.6292" ...
 $ embarked: chr  "S" "S" "Q" "Q" ...
 $ class   : chr  "Third" "Third" "Third" "Third" ...
 $ who     : chr  "mannn" "man" "woman" "woman" ...
 $ alone   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ survived: int  0 0 0 0 0 1 0 0 1 0 ...
> colSums(is.na(titanic_Dataset))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived 
       0       14        2        2        0        0        0        0        2        0 
>                                  labels = c(1,2))
```

3. Replaced empty values in categorical columns with ‘NA’. And checked using is.na() function.

```
titanic_Dataset$Gender[titanic_Dataset$Gender == ""] <- NA

titanic_Dataset$fare <- as.numeric(titanic_Dataset$fare)

titanic_Dataset$embarked[titanic_Dataset$embarked == ""] <- NA

titanic_Dataset$class[titanic_Dataset$class == ""] <- NA

titanic_Dataset$who[titanic_Dataset$who == ""] <- NA
colSums(is.na(titanic_Dataset))
```

```
> titanic_Dataset$Gender[titanic_Dataset$Gender == ""] <- NA
> titanic_Dataset$fare <- as.numeric(titanic_Dataset$fare)
Warning message:
NAs introduced by coercion
> titanic_Dataset$embarked[titanic_Dataset$embarked == ""] <- NA
> titanic_Dataset$class[titanic_Dataset$class == ""] <- NA
> titanic_Dataset$who[titanic_Dataset$who == ""] <- NA
> colSums(is.na(titanic_Dataset))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived
       7       14        2        2        7        2        6        2        2        0
>                                   labels = c(1,2))
```

4. Declared a function to remove rows with more than 2 ‘NA’ values.

```
count_nas <- function(row) {
  sum(is.na(row))
}
titanic_Dataset <- titanic_Dataset[apply(titanic_Dataset, 1, count_nas) <= 2, ]
colSums(is.na(titanic_Dataset))
```

```
> count_nas <- function(row) {
+   sum(is.na(row))
+ }
> titanic_Dataset <- titanic_Dataset[apply(titanic_Dataset, 1, count_nas) <= 2, ]
> colSums(is.na(titanic_Dataset))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived
       5       12        0        0        5        0        4        0        0        0
>                                   labels = c(1,2))
```

5. Declaring two functions **replace_na_with_mean** and **replace_na_with_mode** to replace the missing values of numeric columns with mean and categorical columns with most frequent value.

```
replace_na_with_mean <- function(x) {
  replace(x, is.na(x), mean(x, na.rm = TRUE))
}

replace_na_with_mode <- function(x) {
  replace(x, is.na(x), names(sort(table(x), decreasing = TRUE))[1])
}
```

```
> replace_na_with_mean <- function(x) {
+   replace(x, is.na(x), mean(x, na.rm = TRUE))
+ }
> replace_na_with_mode <- function(x) {
+   replace(x, is.na(x), names(sort(table(x), decreasing = TRUE))[1])
+ }
>                         labels = c(1,2))
```

6. Filled missing values in numeric columns ('age', 'sibsp', 'parch') with their mean and in categorical columns ('Gender', 'embarked', 'class', 'who', 'alone') with their mode. And checked the overall missing values using is.na() function.

```
titanic_Dataset$age <- replace_na_with_mean(titanic_Dataset$age)
titanic_Dataset$fare <- replace_na_with_mean(titanic_Dataset$fare)
titanic_Dataset$sibsp <- replace_na_with_mean(titanic_Dataset$sibsp)
titanic_Dataset$parch <- replace_na_with_mean(titanic_Dataset$parch)

titanic_Dataset$Gender <- replace_na_with_mode(titanic_Dataset$Gender)
titanic_Dataset$embarked <- replace_na_with_mode(titanic_Dataset$embarked)
titanic_Dataset$class <- replace_na_with_mode(titanic_Dataset$class)
titanic_Dataset$who <- replace_na_with_mode(titanic_Dataset$who)
titanic_Dataset$alone <- replace_na_with_mode(titanic_Dataset$alone)

colSums(is.na(titanic_Dataset))
```

```
> titanic_Dataset$age <- replace_na_with_mean(titanic_Dataset$age)
> titanic_Dataset$fare <- replace_na_with_mean(titanic_Dataset$fare)
> titanic_Dataset$sibsp <- replace_na_with_mean(titanic_Dataset$sibsp)
> titanic_Dataset$parch <- replace_na_with_mean(titanic_Dataset$parch)
> titanic_Dataset$Gender <- replace_na_with_mode(titanic_Dataset$Gender)
> titanic_Dataset$embarked <- replace_na_with_mode(titanic_Dataset$embarked)
> titanic_Dataset$class <- replace_na_with_mode(titanic_Dataset$class)
> titanic_Dataset$who <- replace_na_with_mode(titanic_Dataset$who)
> titanic_Dataset$alone <- replace_na_with_mode(titanic_Dataset$alone)
> colSums(is.na(titanic_Dataset))
  Gender      age    sibsp    parch     fare embarked    class      who    alone survived 
       0        0        0        0        0        0        0        0        0        0 
```

7. Corrected invalid value errors in the `who` column.

```
unique(titanic_Dataset$who)
titanic_Dataset <- titanic_Dataset %>%
  mutate(who = recode(who, "mannn" = "man"))
unique(titanic_Dataset$who)
```

```
> unique(titanic_Dataset$who)
[1] "mannn" "man"    "woman" "child"
> titanic_Dataset <- titanic_Dataset %>%
+   mutate(who = recode(who, "mannn" = "man"))
> unique(titanic_Dataset$who)
[1] "man"    "woman" "child"
>                                    labels = c(1,2))
```

8. Converted categorical variables to factors with appropriate labels.

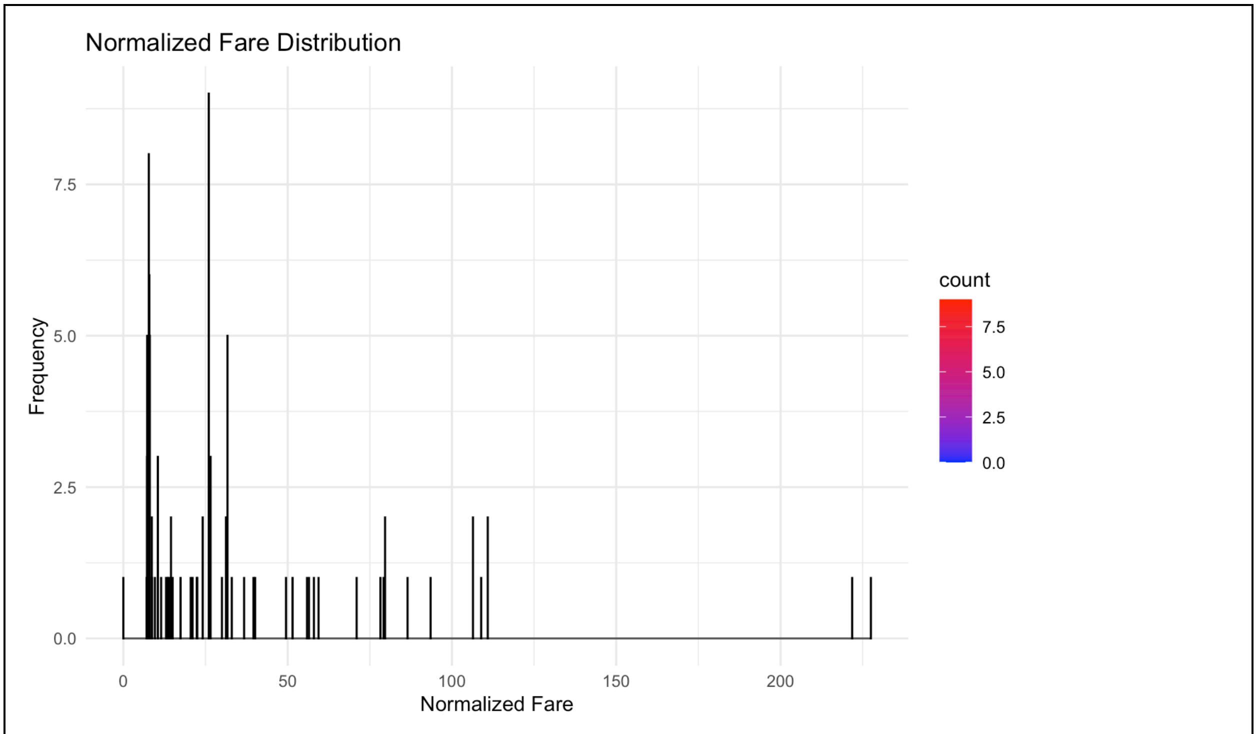
```
unique(titanic_Dataset$Gender)
titanic_Dataset$Gender <- factor(titanic_Dataset$Gender,
                  levels = c("female", "male"),
                  labels = c(1,2))

unique(titanic_Dataset$class)
titanic_Dataset$class <- factor(titanic_Dataset$class,
                                 levels = c("Third","First","Second"),
                                 labels = c(1,2,3))

unique(titanic_Dataset$who)
titanic_Dataset$who <- factor(titanic_Dataset$who,
                               levels = c("man","woman","child"),
                               labels = c(1,2,3))

unique(titanic_Dataset$alone)
titanic_Dataset$alone <- factor(titanic_Dataset$alone,
                             levels = c("TRUE", "FALSE"),
                             labels = c(1,2))
titanic_Dataset
```

```
> unique(titanic_Dataset$Gender)
[1] "female" "male"
> titanic_Dataset$Gender <- factor(titanic_Dataset$Gender,
+                   levels = c("female", "male"),
+                   labels = c(1,2))
> unique(titanic_Dataset$embarked)
[1] "S" "Q" "C"
> titanic_Dataset$embarked <- factor(titanic_Dataset$embarked,
+                                  levels = c("S", "Q", "C"),
+                                  labels = c(1,2,3))
> unique(titanic_Dataset$class)
[1] "Third"  "First"  "Second"
> titanic_Dataset$class <- factor(titanic_Dataset$class,
+                                    levels = c("Third","First","Second"),
+                                    labels = c(1,2,3))
> unique(titanic_Dataset$who)
[1] "man"   "woman" "child"
> titanic_Dataset$who <- factor(titanic_Dataset$who,
+                                 levels = c("man","woman","child"),
+                                 labels = c(1,2,3))

> unique(titanic_Dataset$alone)
[1] "TRUE"  "FALSE"
> titanic_Dataset$alone <- factor(titanic_Dataset$alone,
+                                  levels = c("TRUE", "FALSE"),
+                                  labels = c(1,2))
> titanic_Dataset
    Gender       age sibsp parch      fare embarked class who alone survived
1        1  24.00000     0     0   7.79580        1     1   1     1        0
2        1  17.00000     0     0   8.66250        1     1   1     1        0
3        2  21.00000     0     0   7.75000        2     1   2     1        0
4        2  35.00000     0     0   7.62920        2     1   2     1        0
5        2  37.00000     0     0   9.58750        1     1   2     1        0
6        2  16.00000     0     0  86.50000        1     2   2     1        1
7        1  36.25275     1     0 108.90000        3     2   1     2        0
8        2  33.00000     0     2  31.72296        1     3   2     2        0
9        1  40.00000     0     0  26.55000        1     2   1     1        1
10       1  28.00000     0     0  22.52500        1     1   1     1        0
```

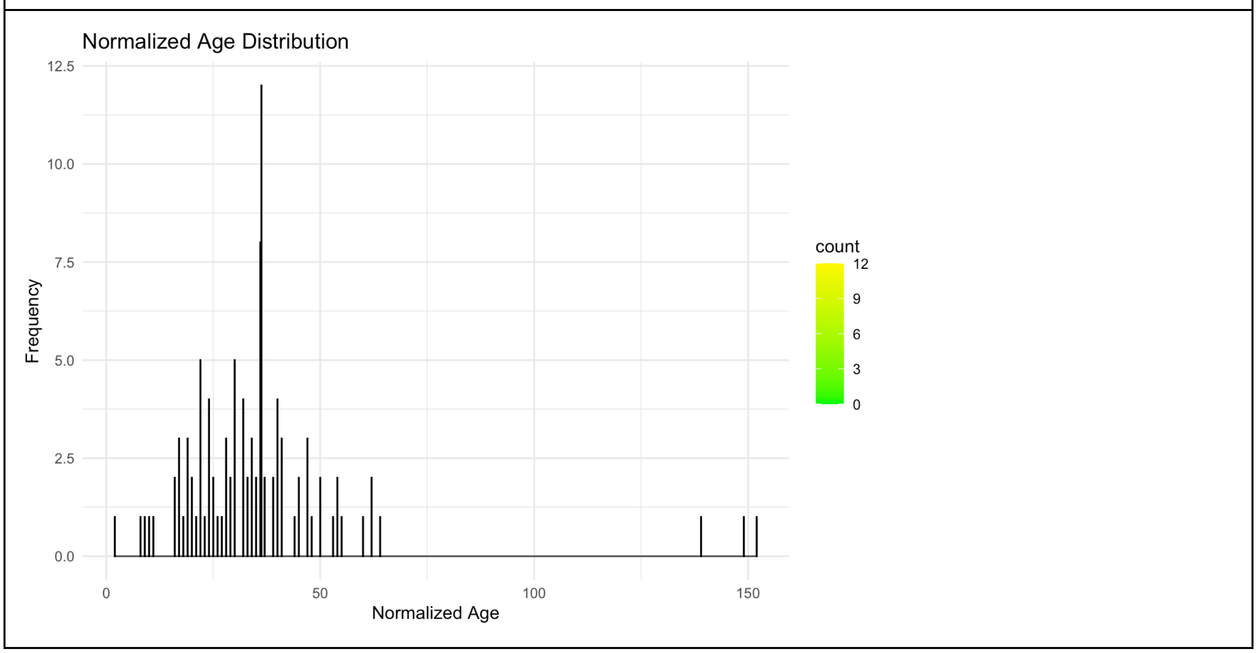
9. Calculating mean of age and fare, and plotting these columns using histogram to detect outliers.

```r
mean(titanic_Dataset$age)
mean(titanic_Dataset$fare)
summary(titanic_Dataset$age)
summary(titanic_Dataset$fare)
ggplot(titanic_Dataset, aes(x = fare, fill = ..count..)) +
  geom_histogram(binwidth = 0.05, color = "black", alpha = 0.7) +
  ggtitle("Normalized Fare Distribution") +
  xlab("Normalized Fare") +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_gradient(low = "blue", high = "red")

ggplot(titanic_Dataset, aes(x = age, fill = ..count..)) +
  geom_histogram(binwidth = 0.05, color = "black", alpha = 0.7) +
  ggtitle("Normalized Age Distribution") +
  xlab("Normalized Age") +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_gradient(low = "green", high = "yellow")
```

```
> mean(titanic_Dataset$age)
[1] 36.25275
> mean(titanic_Dataset$fare)
[1] 31.72296
> summary(titanic_Dataset$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   24.00   35.00   36.25   40.00  152.00 
> summary(titanic_Dataset$fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   7.896  21.075  31.723  31.723 227.525 
```



Normalized Fare Distribution
Frequency
7.5
5.0
2.5
0.0
0
50
100
150
200
Normalized Fare
count
7.5
5.0
2.5
0.0



Normalized Age Distribution
12.5
10.0
7.5
5.0
2.5
0.0
Frequency
0
50
100
150
Normalized Age
count
12
9
6
3
0

10. Outliers were detected.
To remove these outliers, we applied a function to remove outliers based on IQR.

```
remove_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  x[x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)] <- NA
  return(x)
}
titanic_Dataset$fare <- remove_outliers(titanic_Dataset$fare)
titanic_Dataset$age <- remove_outliers(titanic_Dataset$age)
```

```
> remove_outliers <- function(x) {
+   Q1 <- quantile(x, 0.25)
+   Q3 <- quantile(x, 0.75)
+   IQR <- Q3 - Q1
+   x[x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)] <- NA
+   return(x)
+ }
> titanic_Dataset$fare <- remove_outliers(titanic_Dataset$fare)
> titanic_Dataset$age <- remove_outliers(titanic_Dataset$age)
```

11. Removing rows with NA values caused by outlier removal.

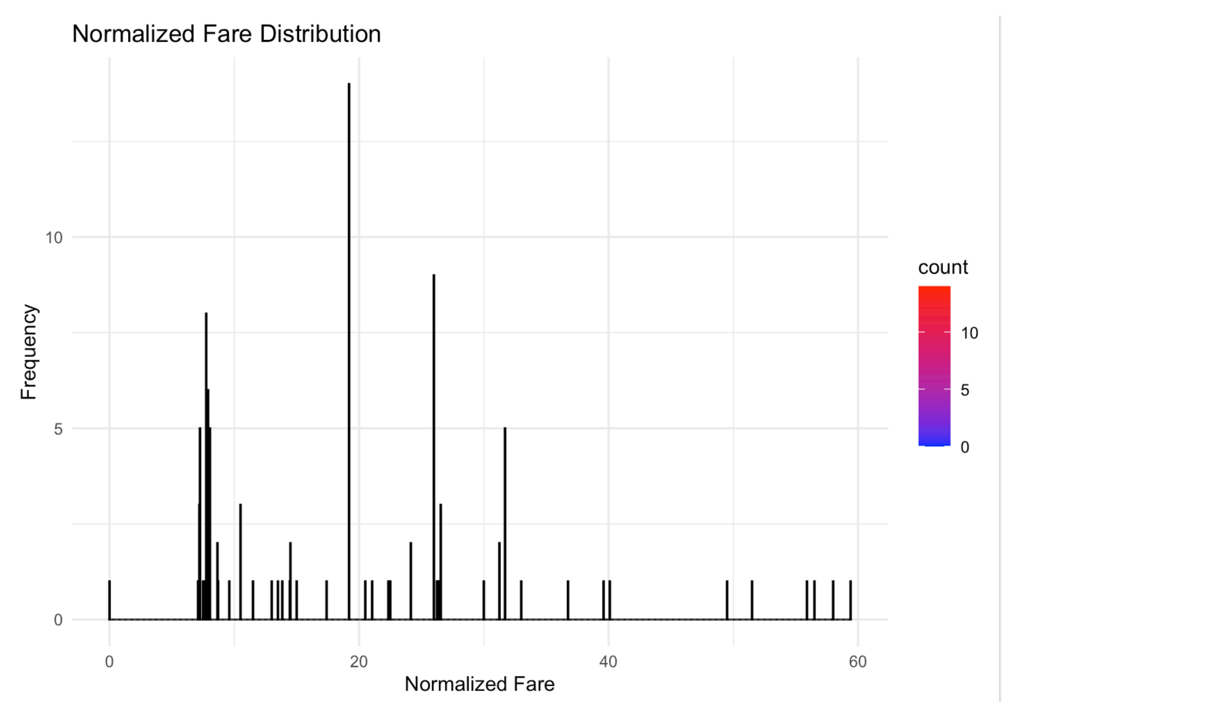
```
titanic_Dataset$fare <- replace_na_with_mean(titanic_Dataset$fare)
titanic_Dataset$age <- replace_na_with_mean(titanic_Dataset$age)
```

```
> titanic_Dataset$fare <- replace_na_with_mean(titanic_Dataset$fare)
> titanic_Dataset$age <- replace_na_with_mean(titanic_Dataset$age)
>                          labels = c(1,2))
```

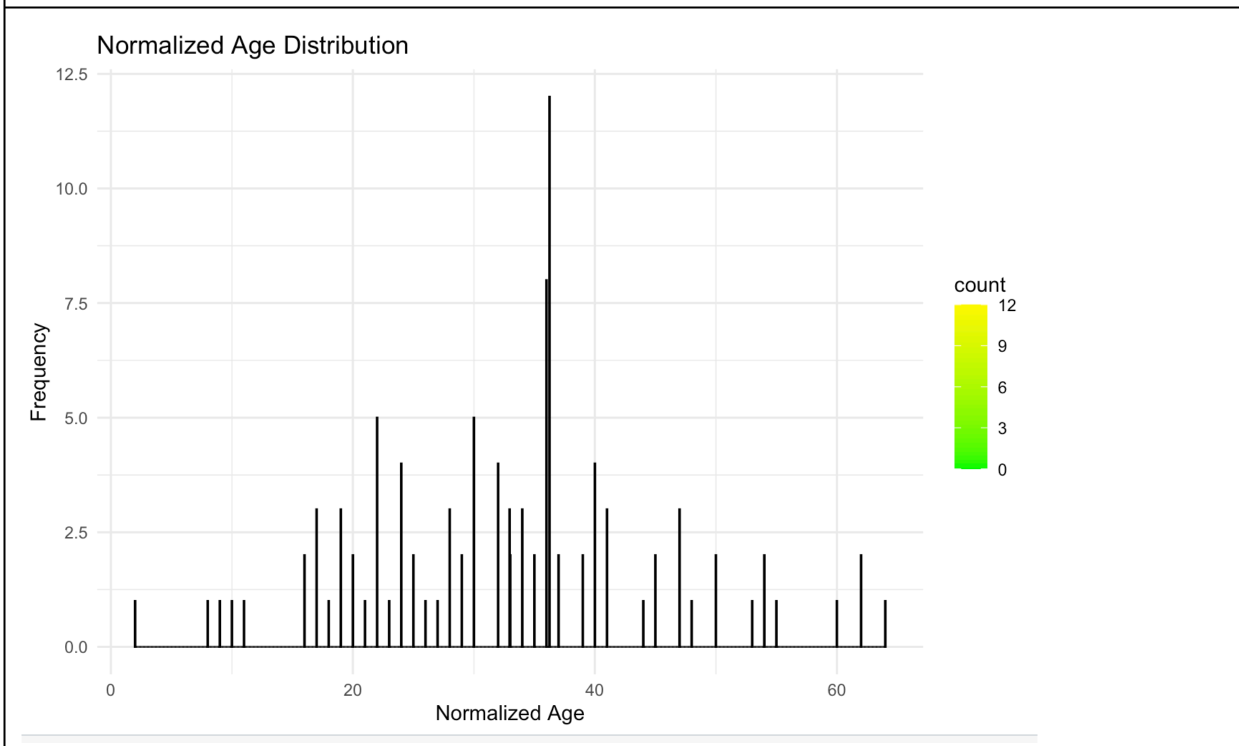
12. Then again did histogram and mean calculation to check the removal of outliers.

```r
mean(titanic_Dataset$age)
mean(titanic_Dataset$fare)
summary(titanic_Dataset$age)
summary(titanic_Dataset$fare)
ggplot(titanic_Dataset, aes(x = fare, fill = ..count..)) +
  geom_histogram(binwidth = 0.05, color = "black", alpha = 0.7) +
  ggtitle("Normalized Fare Distribution") +
  xlab("Normalized Fare") +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_gradient(low = "blue", high = "red")

ggplot(titanic_Dataset, aes(x = age, fill = ..count..)) +
  geom_histogram(binwidth = 0.05, color = "black", alpha = 0.7) +
  ggtitle("Normalized Age Distribution") +
  xlab("Normalized Age") +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_gradient(low = "green", high = "yellow")
```

```
> mean(titanic_Dataset$age)
[1] 32.94033
> mean(titanic_Dataset$fare)
[1] 19.1784
> summary(titanic_Dataset$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   24.00   34.00   32.94   38.00   64.00 
> summary(titanic_Dataset$fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   7.896  19.178  19.178  26.000  59.400 
```



Normalized Fare Distribution
Frequency
Normalized Fare
count
0
5
10
0
20
40
60



Normalized Age Distribution
Frequency
Normalized Age
count
0.0
2.5
5.0
7.5
10.0
12.5
0
3
6
9
12
0
20
40
60

8. Normalizing Continuous Attribute 'age'

```
summary(titanic_Dataset$age)
titanic_Dataset$age <- (titanic_Dataset$age - min(titanic_Dataset$age)) / (max(titanic_Dataset$age) - min(titanic_Dataset$age))
summary(titanic_Dataset$age)
```

```
> summary(titanic_Dataset$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   24.00   34.00   32.94   38.00   64.00 
> titanic_Dataset$age <- (titanic_Dataset$age - min(titanic_Dataset$age)) / (max(titanic_Dataset$age) - min(titanic_Dataset$age))
> summary(titanic_Dataset$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.3548  0.5161  0.4990  0.5806  1.0000 
>                               labels = c(1,2))
```

Descriptive Statistics:

1. Calculated summary statistics for continuous attributes ('age', 'sibsp', 'parch', 'fare').

```
gender_summary <- summary(titanic_Dataset$Gender)
age_summary <- summary(titanic_Dataset$age)
sibsp_summary <- summary(titanic_Dataset$sibsp)
parch_summary <- summary(titanic_Dataset$parch)
fare_summary <- summary(titanic_Dataset$fare)

list(gender = gender_summary, age = age_summary, sibsp = sibsp_summary, parch = parch_summary, fare = fare_summary)
```

```
> gender_summary <- summary(titanic_Dataset$Gender)
> age_summary <- summary(titanic_Dataset$age)
> sibsp_summary <- summary(titanic_Dataset$sibsp)
> parch_summary <- summary(titanic_Dataset$parch)
> fare_summary <- summary(titanic_Dataset$fare)
> list(gender = gender_summary, age = age_summary, sibsp = sibsp_summary, parch = parch_summary, fare = fare_summary)
$gender
 1  2 
66 37 

$age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.3548  0.5161  0.4990  0.5806  1.0000 

$sibsp
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3495  1.0000  4.0000 

$parch
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3398  0.0000  4.0000 

$fare
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   7.896  19.178  19.178  26.000  59.400 
```

Data Balancing:

1. Checked the distribution of the `survived` variable to identify class imbalance.

```
table(titanic_Dataset$survived)
table(titanic_Dataset$Gender)
```

```
> table(titanic_Dataset$survived)

 0  1
65 38
```

2. It is found, imbalanced.
Used the `ROSE` package to balance the dataset through oversampling and undersampling.

```
balanced_dataset <- ovun.sample(survived ~ ., data = titanic_Dataset, method = "both", p = 0.5, seed = 1)$data
table(balanced_dataset$survived)
```

```
> balanced_dataset <- ovun.sample(survived ~ ., data = titanic_Dataset, method = "both", p = 0.5, seed = 1)$data
> table(balanced_dataset$survived)

 0  1
54 49
```

Display of prepared DataSet:

```
titanic_Dataset
balanced_dataset
```

```
> titanic_Dataset
   Gender        age sibsp parch     fare embarked class who alone survived
1       1 0.35483871     0     0  7.79580        1     1   1     1        0
2       1 0.24193548     0     0  8.66250        1     1   1     1        0
3       2 0.30645161     0     0  7.75000        2     1   2     1        0
4       2 0.53225806     0     0  7.62920        2     1   2     1        0
5       2 0.56451613     0     0  9.58750        1     1   2     1        0
6       2 0.22580645     0     0 19.17840        1     2   2     1        1
7       1 0.55246367     1     0 19.17840        3     2   1     2        0
8       2 0.50000000     0     2 31.72296        1     3   2     2        0
9       1 0.61290323     0     0 26.55000        1     2   1     1        1
10      1 0.41935484     0     0 22.52500        1     1   1     1        0
> balanced_dataset
   Gender        age sibsp parch     fare embarked class who alone survived
1       1 1.00000000     0     0 26.00000        1     2   1     1        0
2       1 0.27419355     0     0  7.89580        1     1   1     1        0
3       1 0.62903226     0     0  7.75000        2     1   1     1        0
4       2 0.14516129     4     2 31.27500        1     1   3     2        0
5       1 0.61290323     0     0  7.22500        3     1   1     1        0
6       1 0.22580645     0     0  8.05000        1     1   1     1        0
7       1 0.49903758     0     0  0.00000        1     1   1     1        0
8       1 0.72580645     0     0  7.25000        1     1   1     1        0
9       1 0.48387097     0     0  7.92500        1     1   1     1        0
10      1 0.50000000     1     1 20.52500        1     1   1     2        0
```

Conclusion:

In this project, we successfully applied various data preparation steps, calculated descriptive statistics and balanced the dataset to understand the relationships between continuous variables. These steps are crucial for preparing data for further analysis and modeling.