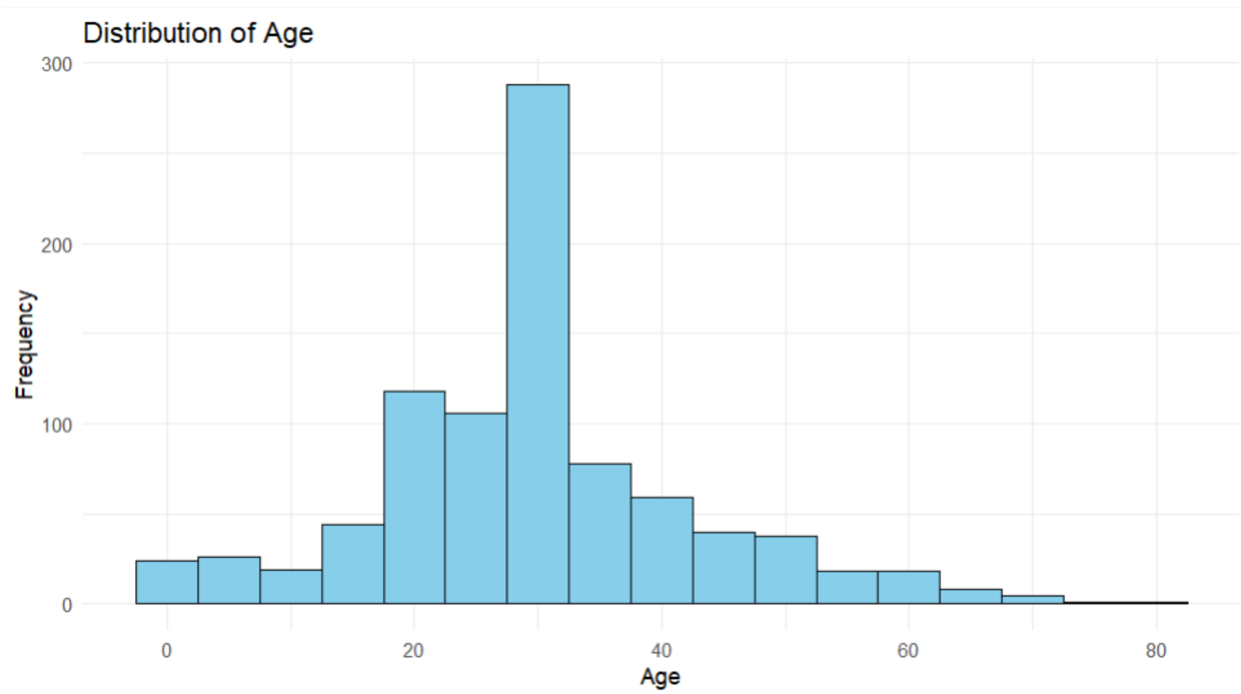
Lab Task 1

Why I choose the Dataset?

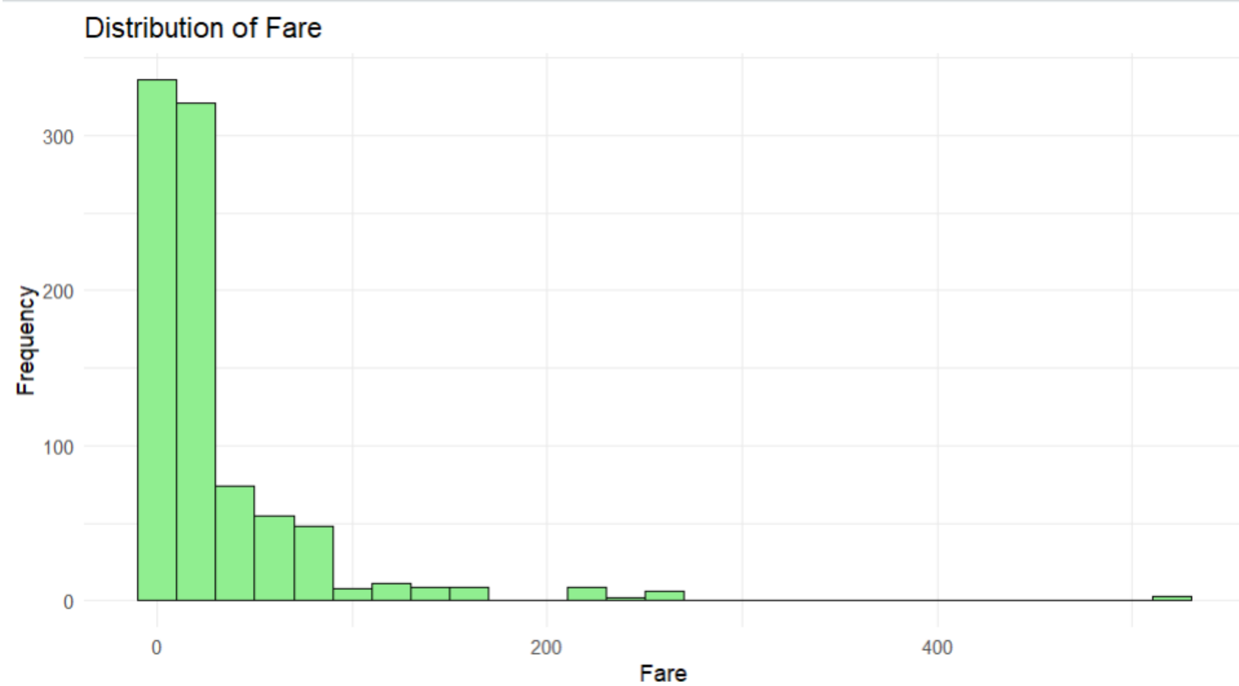
The Titanic dataset was selected from Kaggle as it perfectly aligns with the assignment's requirements, being a readily available labeled dataset with a sufficient number of rows to ensure appropriate chart generation and robust analysis. Crucially, it features a diverse set of more than five columns, encompassing both numerical variables (such as Age, Fare, SibSp, Parch) and categorical variables (including Pclass, Sex, and Embarked), thereby facilitating the application of various visualization and feature selection methods as stipulated.

Dataset Collected from : http://kaggle.com/c/titanic/data?select=train.csv

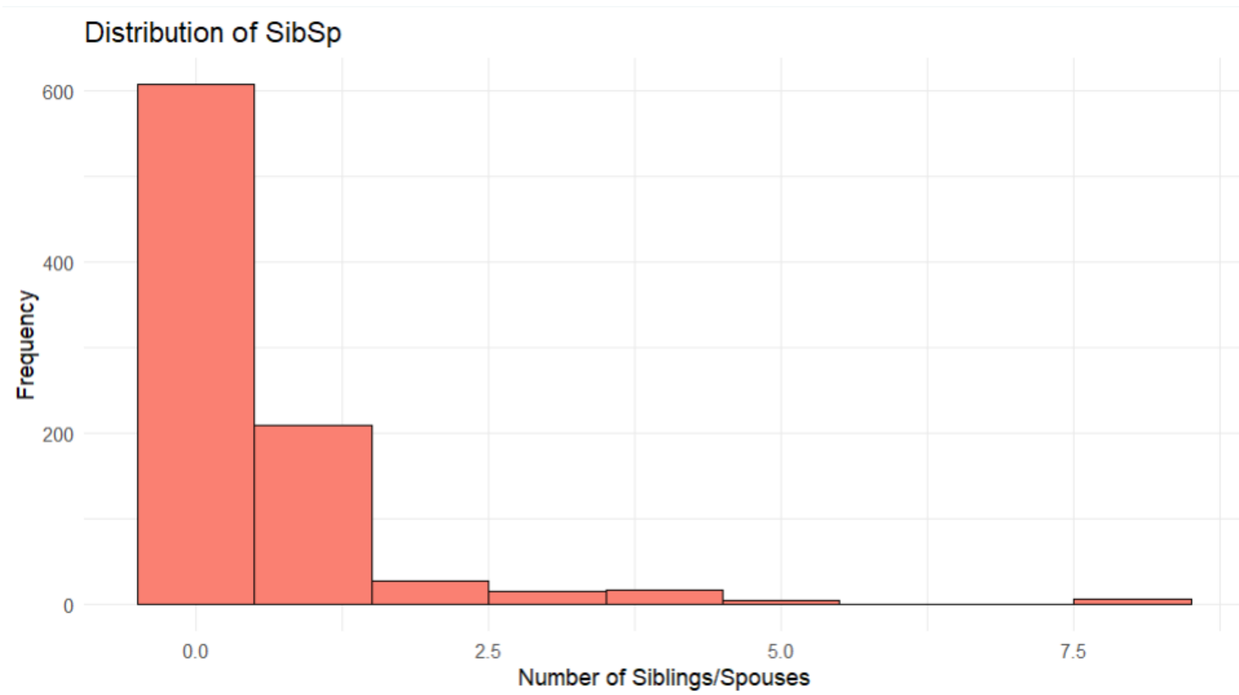
- **Histogram**
  - Histogram for Age Column



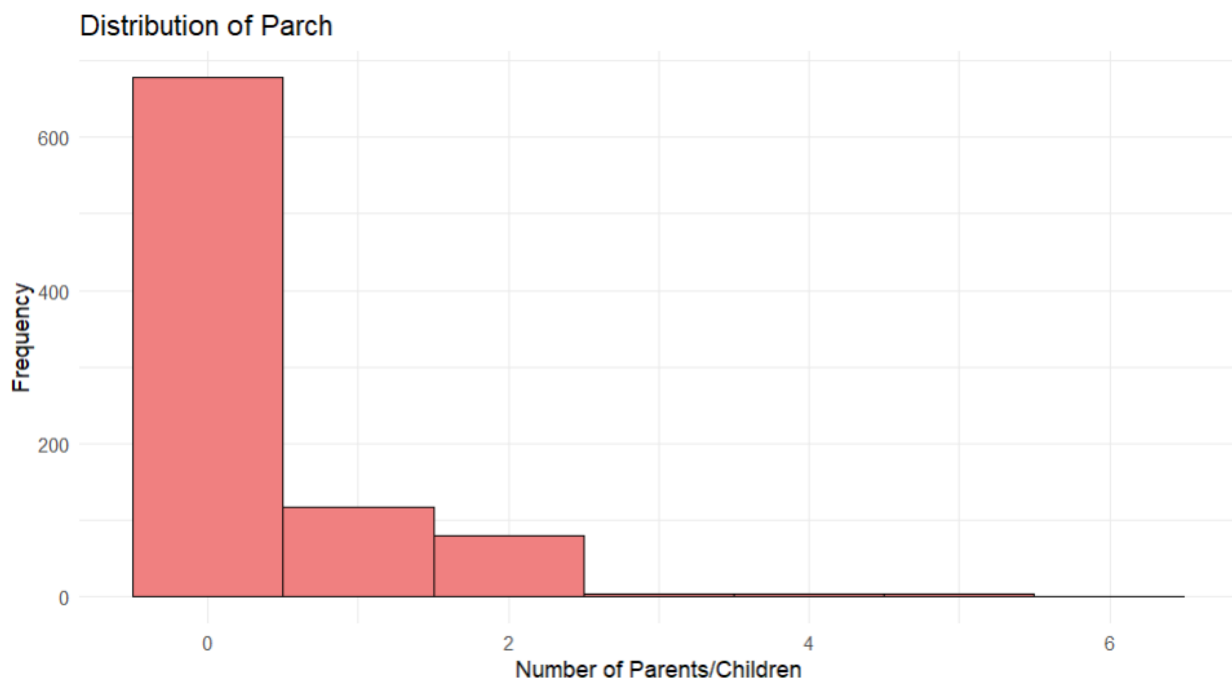
  - Histogram for Fare Column



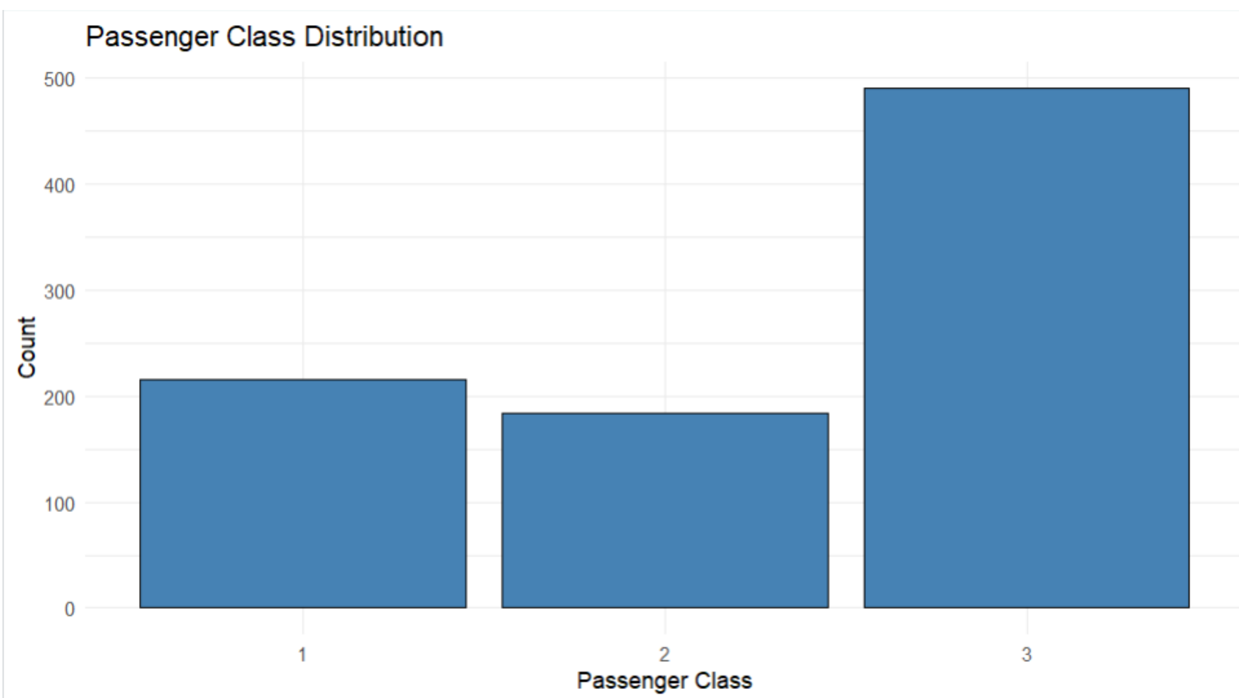
➢ Histogram for SibSp Column



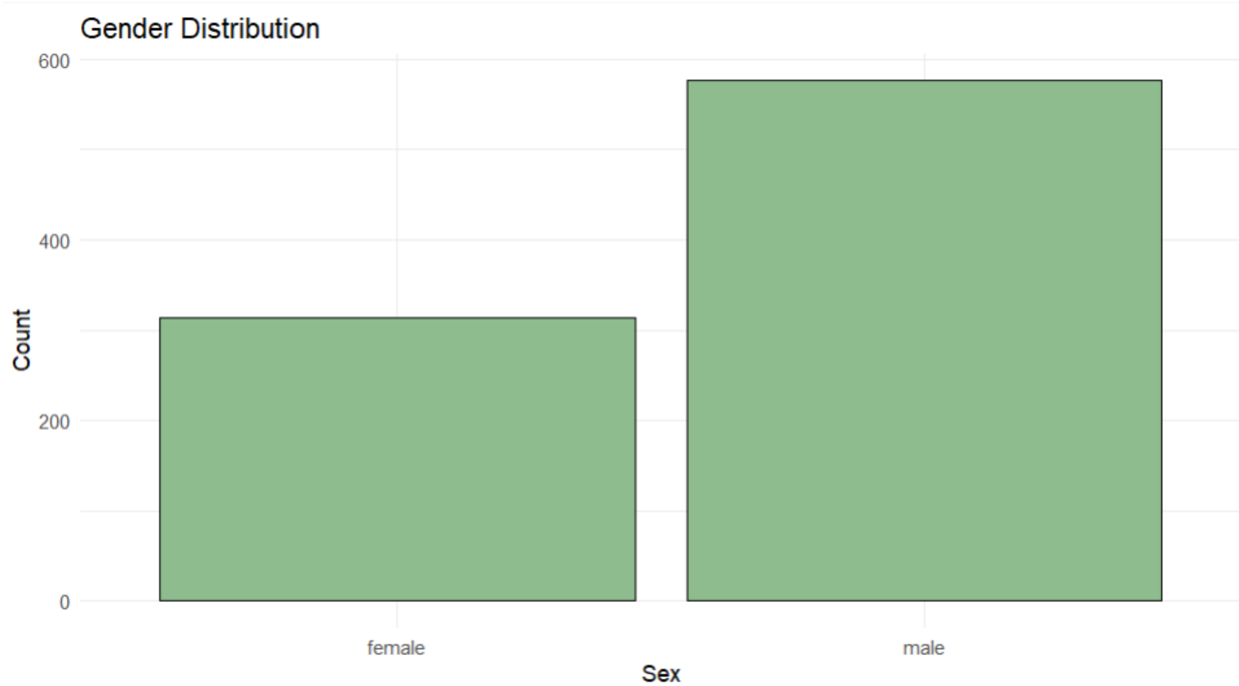
➢ Histogram for Parch Column



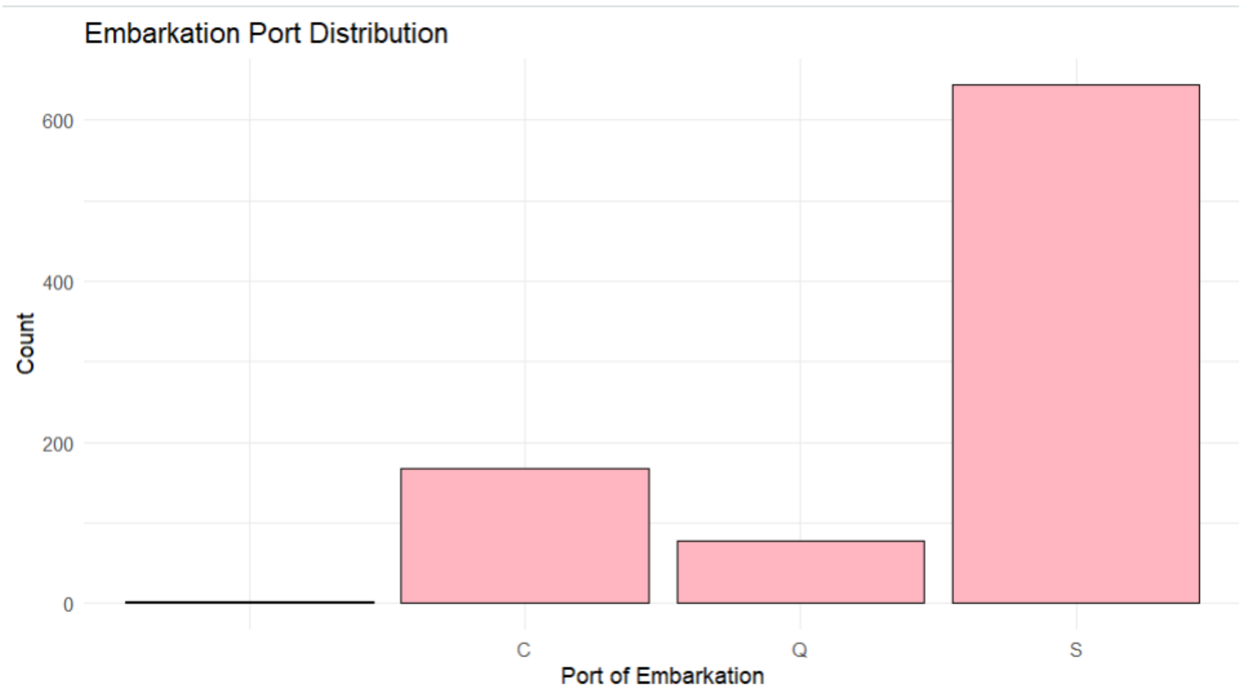
- **Bar Chart**
  - Bar Chart for Pclass Column



  - Bar Chart for Sex Column

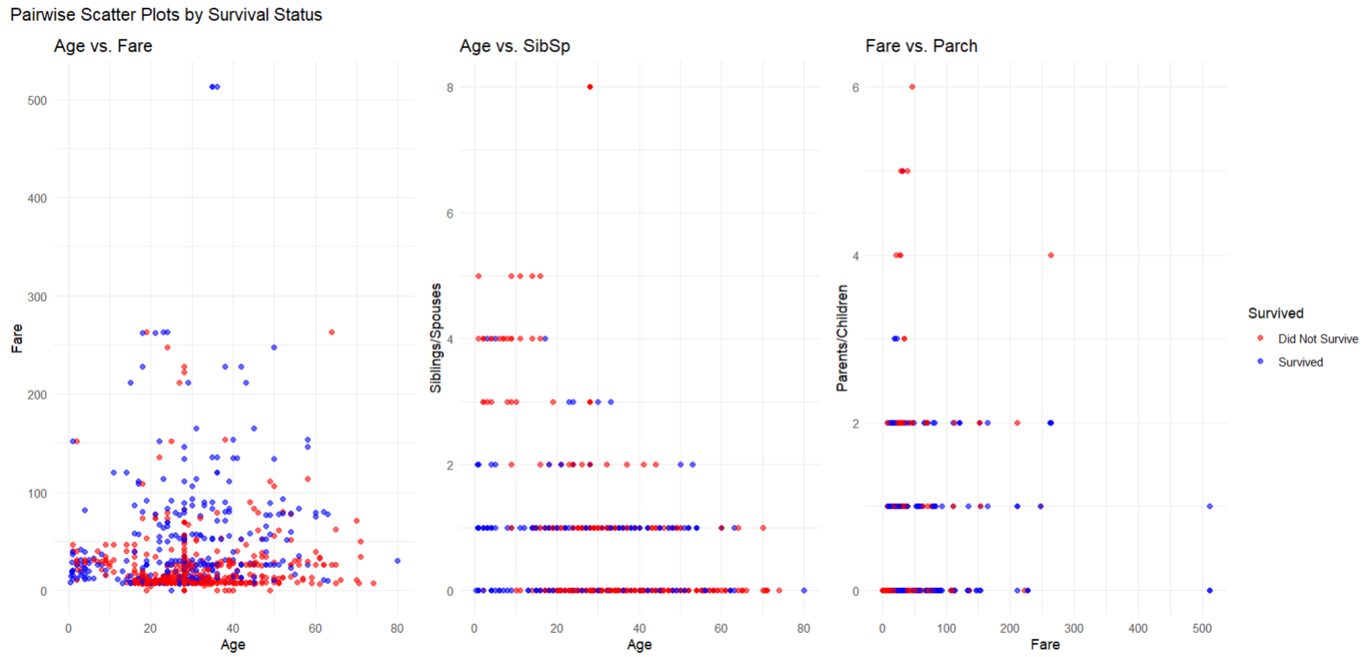


➢ Bar Chart for Embarked Column



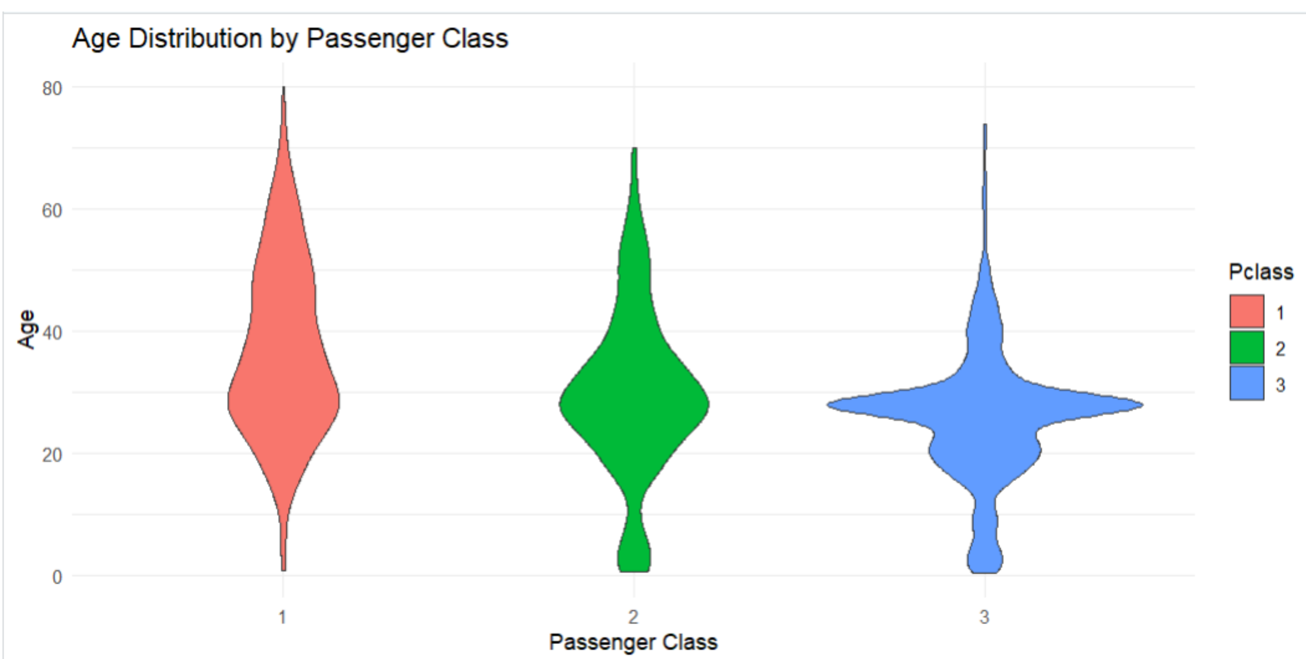
➢ Scatter Plots

- Scatter Plots for Survival status using Age, Fare, SibSp & Parch Column

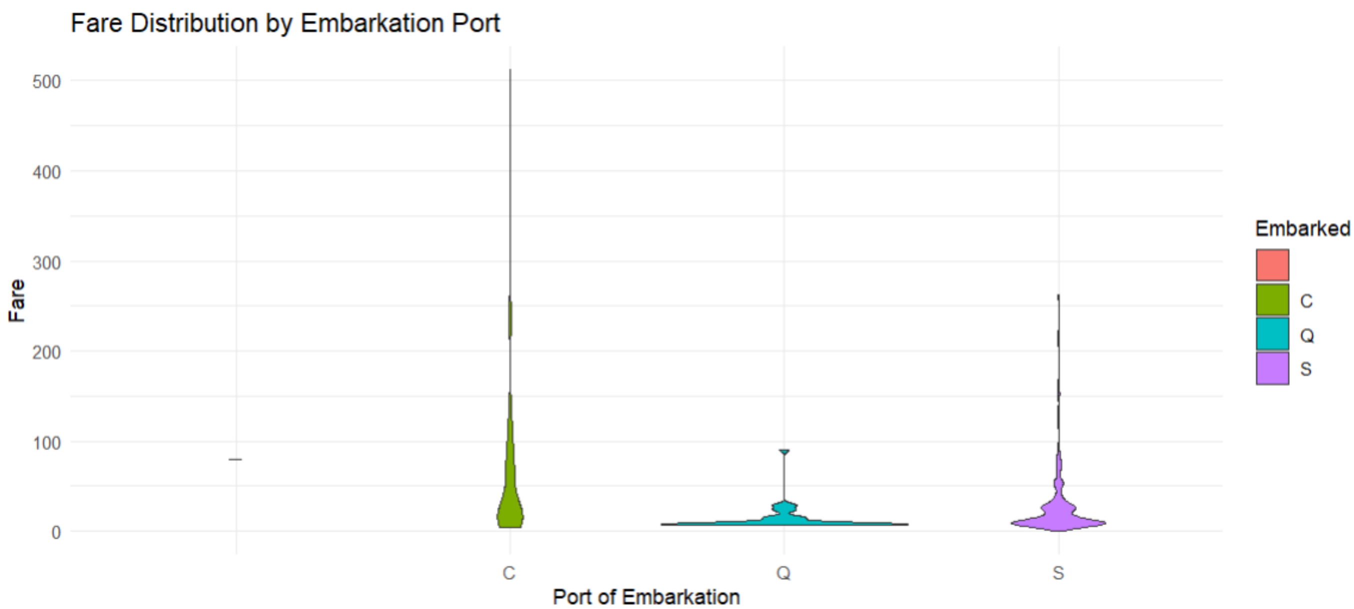


➢ Violin Plot

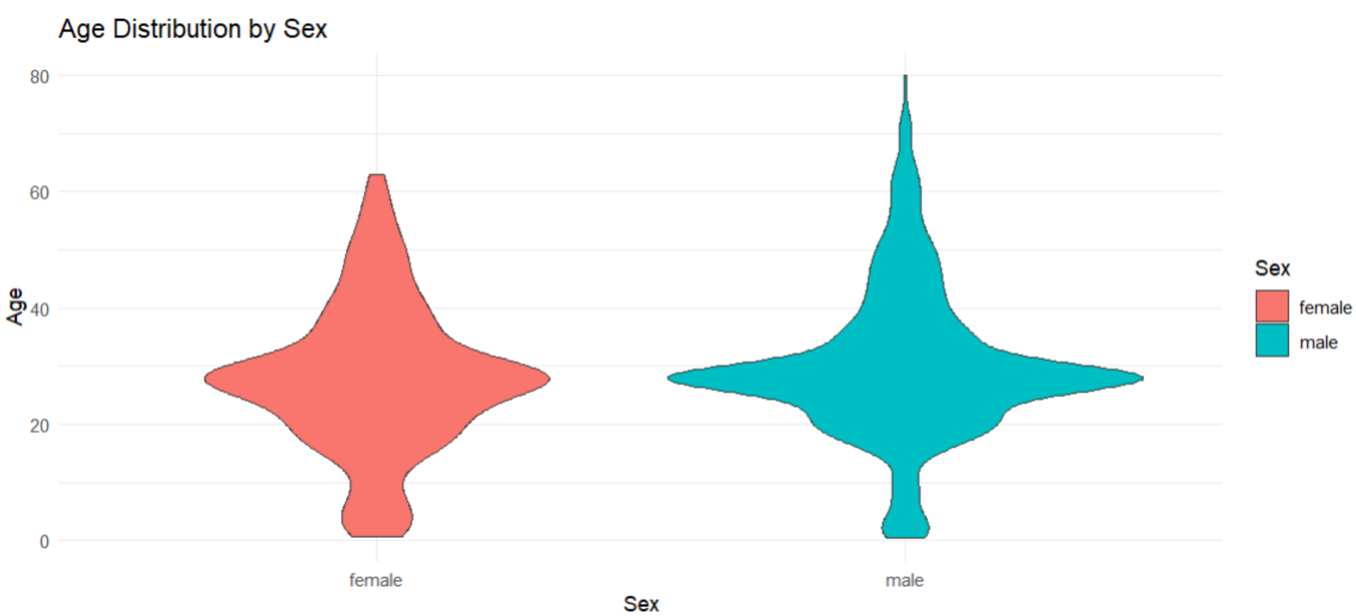
- Violin Plot for Age & PClass Column



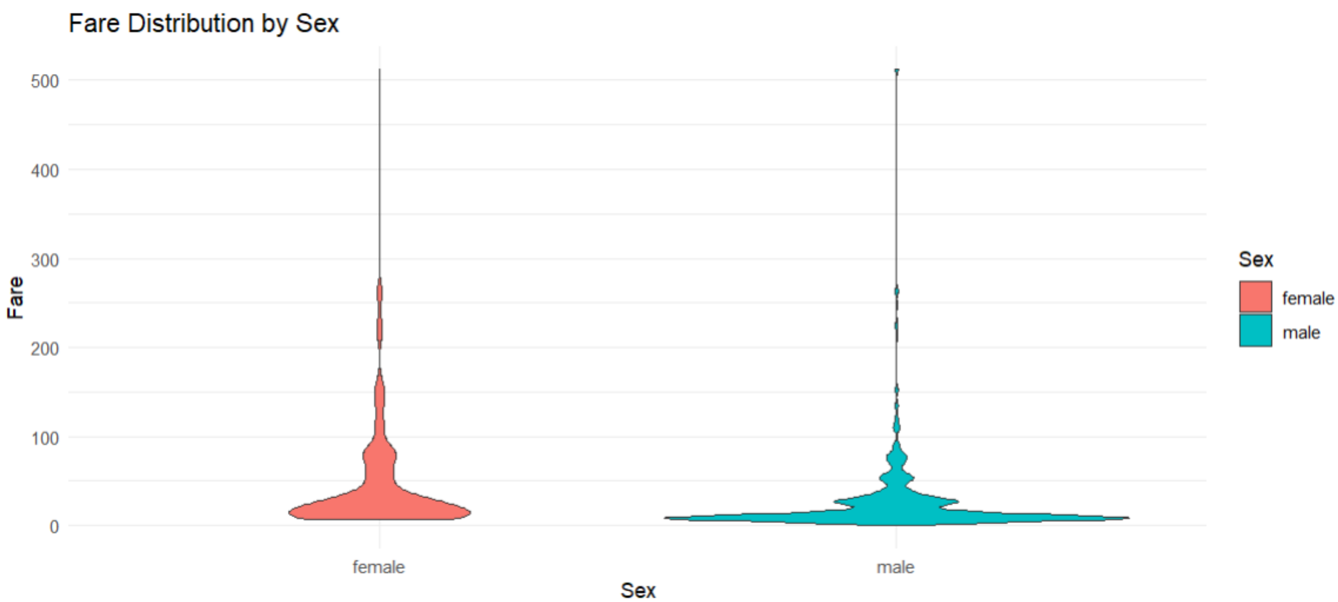
- Violin Plot for Fare & Embark Column



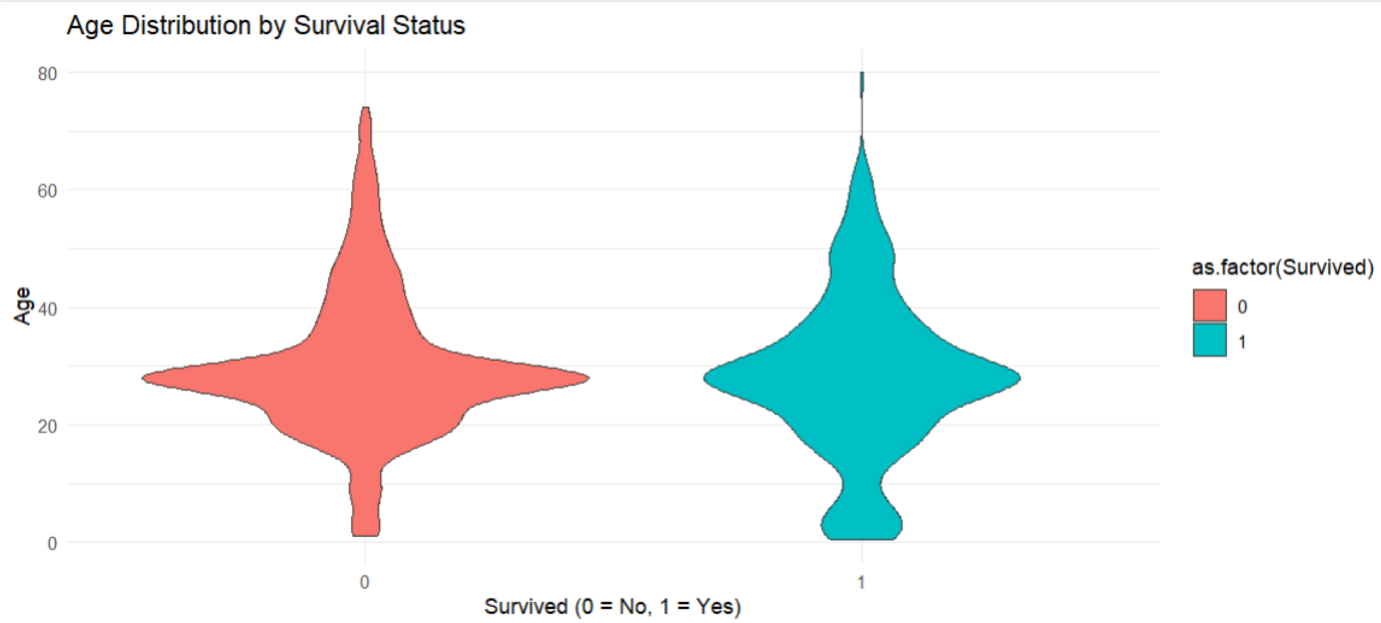
- Violin Plot for Age & Sex Column



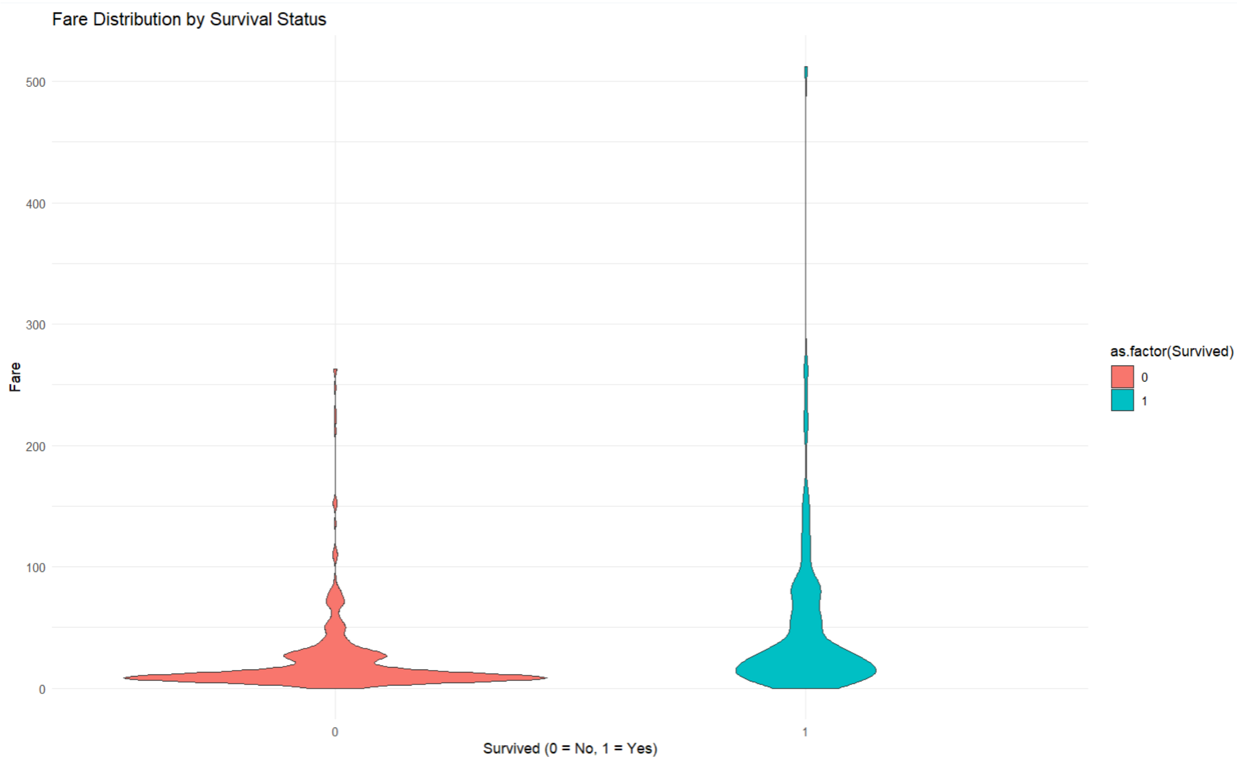
- Violin Plot for Fare & Sex Column



- Violin Plot for Age & Survived Column



- Violin port for Fare & Survived Column


Fare Distribution by Survival Status
500
400
300
200
100
0
Fare
0
1
Survived (0 = No, 1 = Yes)
as.factor(Survived)
0
1

Lab Task 2

Use the same dataset as per instruction.

➢ Output for Pearson correlation coefficient :

```
> correlation_matrix <- cor(numerical_cols_for_corr, use = "pairwise.complete.obs")
> print("Pearson Correlation Matrix:")
[1] "Pearson Correlation Matrix:"
> print(correlation_matrix)
                Age       Fare       SibSp       Parch    Survived
Age       1.00000000 0.09606669 -0.30824676 -0.18911926 -0.07722109
Fare      0.09606669 1.00000000  0.13832879  0.20511888  0.26818862
SibSp    -0.30824676 0.13832879  1.00000000  0.38381986 -0.01735836
Parch    -0.18911926 0.20511888  0.38381986  1.00000000  0.09331701
Survived -0.07722109 0.26818862 -0.01735836  0.09331701  1.00000000
> correlation_with_survived <- as.data.frame(correlation_matrix["Survived", ])
> colnames(correlation_with_survived) <- "Correlation_with_Survived"
> print("Pearson Correlation with Survived:")
[1] "Pearson Correlation with Survived:"
> print(correlation_with_survived %>% arrange(desc(abs(Correlation_with_Survived))))
         Correlation_with_Survived
Survived                1.00000000
Fare                    0.26818862
Parch                   0.09331701
Age                    -0.07722109
SibSp                  -0.01735836
```

➢ ANOVA

- For Age vs Survived

```
> anova_age_survived <- aov(Age ~ Survived, data = titanic_fs_data)
> print("ANOVA for Age vs. Survived:")
[1] "ANOVA for Age vs. Survived:"
> print(summary(anova_age_survived))
             Df Sum Sq Mean Sq F value Pr(>F)
Survived      1    897   897.2   4.271 0.0391 *
Residuals   712 149559   210.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- For Fare vs Survived

```
> anova_fare_survived <- aov(Fare ~ Survived, data = titanic_fs_data)
> print("ANOVA for Fare vs. Survived:")
[1] "ANOVA for Fare vs. Survived:"
> print(summary(anova_fare_survived))
             Df  Sum Sq Mean Sq F value   Pr(>F)
Survived      1  143613  143613   55.18 3.16e-13 ***
Residuals   712 1853082    2603
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- For SibSp vs Survived

```
> anova_sibsp_survived <- aov(SibSp ~ Survived, data = titanic_fs_data)
> print("ANOVA for SibSp vs. Survived:")
[1] "ANOVA for SibSp vs. Survived:"
> print(summary(anova_sibsp_survived))
             Df Sum Sq Mean Sq F value Pr(>F)
Survived      1    0.2  0.1857   0.215  0.643
Residuals   712  616.2  0.8655
```

- For Parch vs Survived

```
> anova_parch_survived <- aov(Parch ~ Survived, data = titanic_fs_data)
> print("ANOVA for Parch vs. Survived:")
[1] "ANOVA for Parch vs. Survived:"
> print(summary(anova_parch_survived))
             Df Sum Sq Mean Sq F value Pr(>F)
Survived      1    4.5   4.521   6.255 0.0126 *
Residuals   712  514.6   0.723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

➢ Chi-squared test

- For Pclass & Survived

```
> chisq_pclass_survived <- chisq.test(titanic_fs_data$Pclass, titanic_fs_data$Survived)
> print("Chi-squared test for Pclass vs. Survived:")
[1] "Chi-squared test for Pclass vs. Survived:"
> print(chisq_pclass_survived)

        Pearson's Chi-squared test

data:  titanic_fs_data$Pclass and titanic_fs_data$Survived
X-squared = 92.901, df = 2, p-value < 2.2e-16

> cat("\nExpected values for Pclass vs. Survived:\n")

Expected values for Pclass vs. Survived:
> print(chisq_pclass_survived$expected)
                      titanic_fs_data$Survived
titanic_fs_data$Pclass        0         1
                     1 110.4538  75.54622
                     2 102.7339  70.26611
                     3 210.8123 144.18768
```

- For Sex & Survived

```
> chisq_sex_survived <- chisq.test(titanic_fs_data$Sex, titanic_fs_data$Survived)
> print("\nChi-squared test for Sex vs. Survived:")
[1] "\nChi-squared test for Sex vs. Survived:"
> print(chisq_sex_survived)

        Pearson's Chi-squared test with Yates' continuity correction

data:  titanic_fs_data$Sex and titanic_fs_data$Survived
X-squared = 205.03, df = 1, p-value < 2.2e-16

> cat("\nExpected values for Sex vs. Survived:\n")

Expected values for Sex vs. Survived:
> print(chisq_sex_survived$expected)
                   titanic_fs_data$Survived
titanic_fs_data$Sex        0        1
             female 154.9916 106.0084
             male   269.0084 183.9916
```

- For Embarked & Survived

```
> temp_chisq_embarked <- chisq.test(titanic_fs_data$Embarked, titanic_fs_data$Survived)
Warning message:
In chisq.test(titanic_fs_data$Embarked, titanic_fs_data$Survived) :
  Chi-squared approximation may be incorrect
> print(temp_chisq_embarked$expected)
                        titanic_fs_data$Survived
titanic_fs_data$Embarked          0           1
                           1.187675   0.8123249
                       C  77.198880  52.8011204
                       Q  16.627451  11.3725490
                       S 328.985994 225.0140056
>
> print("\nFisher's Exact Test for Embarked vs. Survived (Recommended due to low expected counts):")
[1] "\nFisher's Exact Test for Embarked vs. Survived (Recommended due to low expected counts):"
> print(fisher_embarked_survived)

        Fisher's Exact Test for Count Data

data:  titanic_fs_data$Embarked and titanic_fs_data$Survived
p-value = 4.816e-07
alternative hypothesis: two.sided
```

➢ Mutual Information

```
> mi_results <- information.gain(formula_mi, titanic_fs_data)
> print("Mutual Information (Information Gain) with Survived:")
[1] "Mutual Information (Information Gain) with Survived:"
> print(mi_results %>% arrange(desc(attr_importance)))
         attr_importance
Sex           0.14973092
Fare          0.07340473
Pclass        0.05833398
Embarked      0.02178532
Age           0.00000000
SibSp         0.00000000
Parch         0.00000000
```

And more info is added in code.