SC HW1

R26131044 JHI-CHI CHANG

2025-02-28

Dataset Titanic

```
cat("Dataset Titanic:\n")
```

```
Dataset Titanic:
```

```
str(titanic_df)
```

```
'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 T
 $ 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" ...
```

Dataset Structure

```r
cat("Dataset Structure:\n")
```

```
Dataset Structure:
```

```r
str(titanic_df)
```

```
'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 T
 $ 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" ...
```

Summary Statistics

```r
cat("\nSummary Statistics:\n")
```

```

Summary Statistics:
```

```r
summary(titanic_df)
```

```
  PassengerId       Survived          Pclass          Name          
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891        
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                     
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                     
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                     
```

```
     Sex                 Age            SibSp           Parch
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                    Mean   :29.70   Mean   :0.523   Mean   :0.3816
                    3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000
                    NA's   :177
    Ticket               Fare           Cabin             Embarked
 Length:891         Min.   :  0.00   Length:891         Length:891
 Class :character   1st Qu.:  7.91   Class :character   Class :character
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                    Mean   : 32.20
                    3rd Qu.: 31.00
                    Max.   :512.33
```

Handling Missing Values

```r
titanic_df <- titanic_df %>%
  mutate(Age = ifelse(is.na(Age), median(Age, na.rm = TRUE), Age),
         Embarked = ifelse(is.na(Embarked), "S", Embarked))
```

Total Passengers

```r
total_passengers <- nrow(titanic_df)
cat("\nTotal Passengers:", total_passengers, "\n")
```

```
Total Passengers: 891
```

Titanic Dataset Variable Description

```r
variables_table <- data.frame(
  Variable = c("PassengerId", "Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "
  Type = c("Numeric", "Categorical", "Numeric", "Categorical", "Numeric", "Numeric", "Numeri
  Description = c("Passenger ID", "Survival Status", "Passenger Class", "Sex", "Age", "Numbe
  Possible_Values = c("Integer", "0, 1", "1, 2, 3", "Male, Female", "Real Number", "Integer"
)

kable(variables_table, caption = "Key Variables in Titanic Dataset")
```

Table 1: Key Variables in Titanic Dataset

Variable	Type	Description	Possible_Values
PassengerId	Numeric	Passenger ID	Integer
Survived	Categorical	Survival Status	0, 1
Pclass	Numeric	Passenger Class	1, 2, 3
Sex	Categorical	Sex	Male, Female
Age	Numeric	Age	Real Number
SibSp	Numeric	Number of Siblings/Spouse	Integer
Parch	Numeric	Number of Parents/Children	Integer
Fare	Numeric	Fare	Real Number
Embarked	Categorical	Embarkation Port	C, Q, S

Survival Rate Analysis

```r
survival_summary <- titanic_df %>%
  group_by(Survived) %>%
  summarise(Total = n()) %>%
  mutate(Percentage = round(Total / total_passengers * 100, 2))

cat("\nSurvival Rate Statistics:\n")
```

```
Survival Rate Statistics:
```

```r
print(survival_summary)
```

```
# A tibble: 2 x 3
  Survived Total Percentage
     <int> <int>      <dbl>
1        0   549       61.6
2        1   342       38.4
```

Survival Rate by Class

```
class_survival <- titanic_df %>%
  group_by(Pclass, Survived) %>%
  summarise(Total = n(), .groups = 'drop') %>%
  mutate(Percentage = round(Total / sum(Total) * 100, 2))

cat("\nSurvival Rate by Class:\n")
```

```
Survival Rate by Class:
```

```
print(class_survival)
```

```
# A tibble: 6 x 4
  Pclass Survived Total Percentage
   <int>    <int> <int>      <dbl>
1      1        0    80       8.98
2      1        1   136      15.3
3      2        0    97      10.9
4      2        1    87       9.76
5      3        0   372      41.8
6      3        1   119      13.4
```

Survival Rate by Gender

```
sex_survival <- titanic_df %>%
  group_by(Sex, Survived) %>%
  summarise(Total = n(), .groups = 'drop') %>%
  mutate(Percentage = round(Total / sum(Total) * 100, 2))
```

```r
cat("\nSurvival Rate by Gender:\n")
```

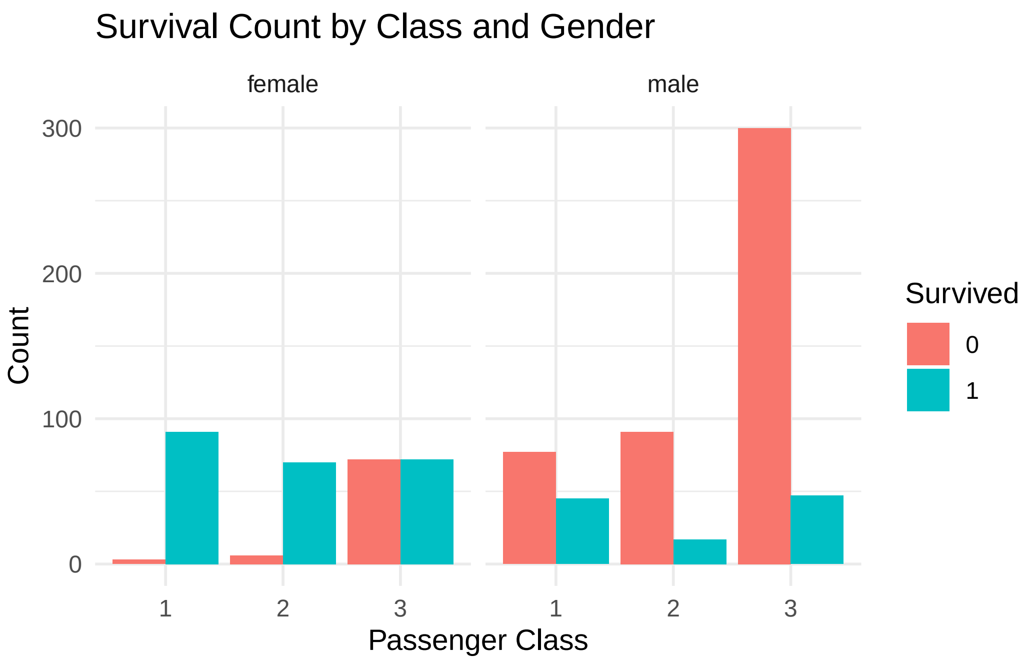
```

Survival Rate by Gender:
```

```r
print(sex_survival)
```

```
# A tibble: 4 x 4
  Sex    Survived Total Percentage
  <chr>     <int> <int>      <dbl>
1 female        0    81       9.09
2 female        1   233      26.2
3 male          0   468      52.5
4 male          1   109      12.2
```

Visualization: Survival by Class and Gender

```r
ggplot(titanic_df, aes(x = factor(Pclass), fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  facet_wrap(~ Sex) +
  labs(title = "Survival Count by Class and Gender", x = "Passenger Class", y = "Count", fil
  theme_minimal()
```



Survival Rate by Fare

```r
fare_survival <- titanic_df %>%
  group_by(Survived) %>%
  summarise(
    Mean_Fare = mean(Fare, na.rm = TRUE),
    Median_Fare = median(Fare, na.rm = TRUE),
    Fare_SD = sd(Fare, na.rm = TRUE),
    Total = n()
  )

print(fare_survival)
```

```
# A tibble: 2 x 5
  Survived Mean_Fare Median_Fare Fare_SD Total
     <int>     <dbl>       <dbl>   <dbl> <int>
1        0      22.1        10.5    31.4   549
2        1      48.4        26      66.6   342
```

Survival Rate by Embarked

```r
embarked_survival <- titanic_df %>%
  group_by(Embarked, Survived) %>%
  summarise(Total = n(), .groups = 'drop') %>%
  mutate(Percentage = round(Total / sum(Total) * 100, 2))

print(embarked_survival)
```

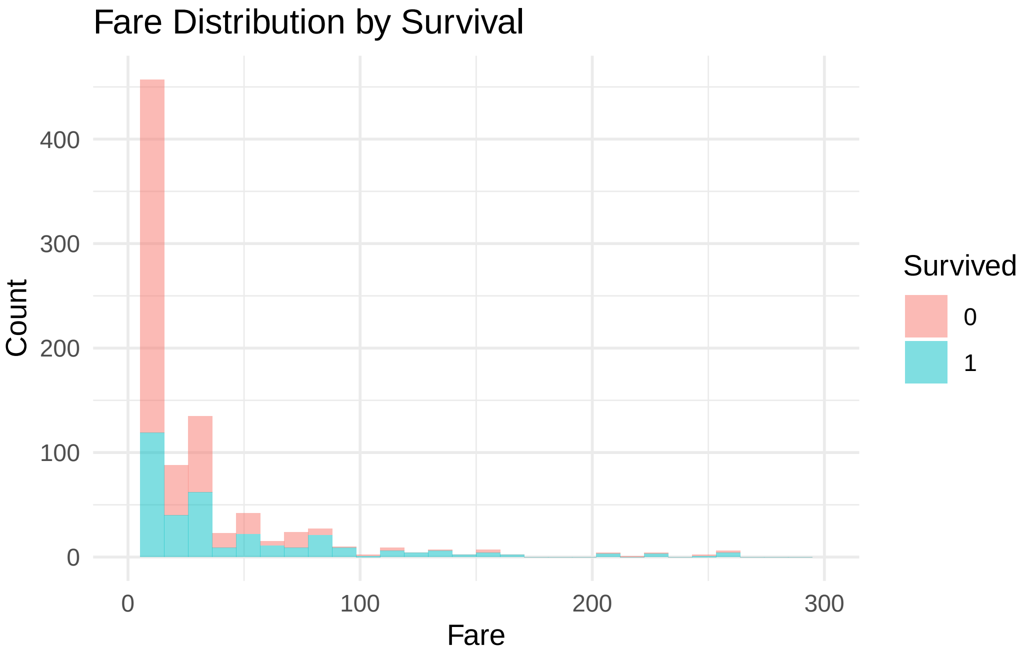
```
# A tibble: 7 x 4
  Embarked Survived Total Percentage
  <chr>       <int> <int>      <dbl>
1 ""              1     2       0.22
2 "C"             0    75       8.42
3 "C"             1    93      10.4 
4 "Q"             0    47       5.27
5 "Q"             1    30       3.37
6 "S"             0   427      47.9 
7 "S"             1   217      24.4 
```

Visualization: Survival by Fare

```r
ggplot(titanic_df, aes(x = Fare, fill = factor(Survived))) +
  geom_histogram(bins = 30, alpha = 0.5, position = "stack") +  #
  scale_x_continuous(limits = c(0, 300)) +  #
  labs(
    title = "Fare Distribution by Survival",
    x = "Fare",
    y = "Count",
    fill = "Survived"
  ) +
  theme_minimal()
```

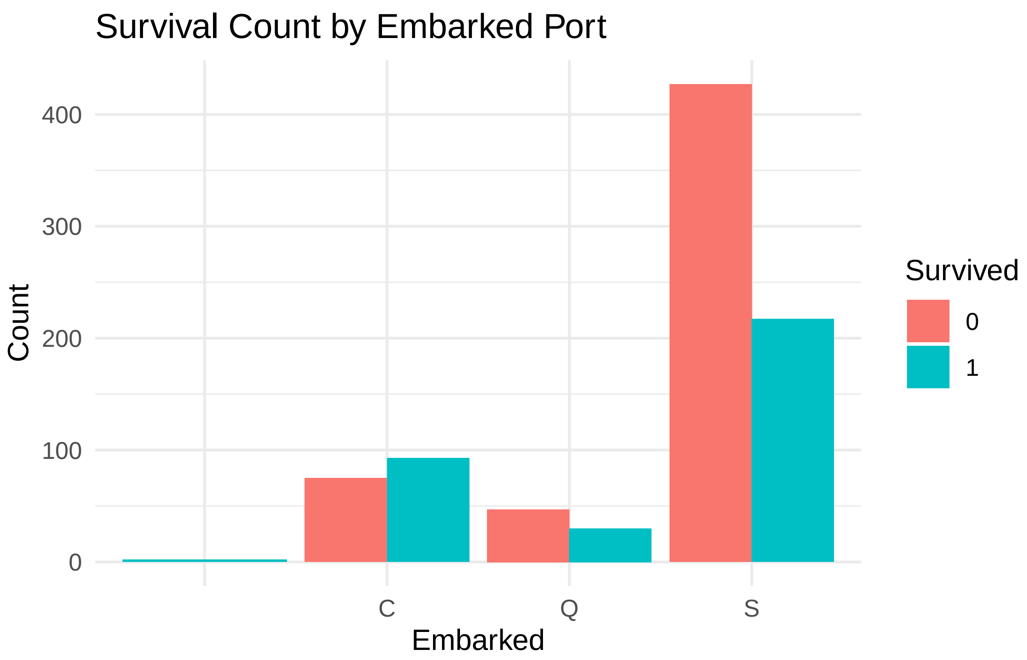
```
Warning: Removed 3 rows containing non-finite outside the scale range
(`stat_bin()`).
```

```
Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



Visualization: Survival by Embarked

```
ggplot(titanic_df, aes(x = Embarked, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  labs(title = "Survival Count by Embarked Port", x = "Embarked", y = "Count", fill = "Survi
  theme_minimal()
```


Survival Count by Embarked Port
Count
400
300
200
100
0
C
Q
S
Embarked
Survived
0
1

Conclusion

1.There are missing values in the dataset, which have been handled appropriately based on logical imputation.

2.According to the data, less than 40% of passengers survived, which is undoubtedly a tragic outcome.

3.The majority of passengers were in third class, followed by first class, and then second class. However, third-class passengers experienced a higher fatality rate. In contrast, first-class passengers had a relatively higher survival rate, while second-class passengers showed no significant difference.

4.There was a noticeable gender imbalance among the passengers, and reliable data suggests that females had a higher likelihood of survival.

5.The average fare of surviving passengers was higher than that of non-survivors.

6.Over 70% of passengers boarded from port S, but unfortunately, this port also accounted for more than 60% of the casualties. This is undoubtedly distressing news for residents of S port and its surrounding towns.