Titanic Dataset Summary

Introduction

This report summarizes the Titanic dataset, offering insights into passenger demographics and survival rates from the tragic sinking of the RMS Titanic in 1912.

Data Loading and Preparation

```
library(ggplot2) ; library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
# Load the Titanic dataset
dt <- read.csv("titanic.csv")

# Convert relevant columns to factors
dt[,c(2:5,7:9,11,12)] <- lapply(dt[,c(2:5,7:9,11,12)], as.factor)
head(dt)
```

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

```
4           113803 53.1000  C123        S
5           373450  8.0500              S
6           330877  8.4583              Q
```

Summary of the Data Set

```
# Display structure and summary of the dataset
str(dt)
```

```
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
 $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417
581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...
 $ Parch      : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133
...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
summary(dt)
```

```
  PassengerId    Survived Pclass                                      Name
 Min.   :  1.0   0:549    1:216   Abbing, Mr. Anthony                   :  1
 1st Qu.:223.5   1:342    2:184   Abbott, Mr. Rossmore Edward           :  1
 Median :446.0            3:491   Abbott, Mrs. Stanton (Rosa Hunt)      :  1
 Mean   :446.0                    Abelson, Mr. Samuel                   :  1
 3rd Qu.:668.5                    Abelson, Mrs. Samuel (Hannah Wizosky):  1
 Max.   :891.0                    Adahl, Mr. Mauritz Nils Martin        :  1
                                  (Other)                               :885
    Sex           Age        SibSp   Parch        Ticket         Fare
 female:314   Min.   : 0.42   0:608   0:678   1601    :  7   Min.   :  0.00
 male  :577   1st Qu.:20.12   1:209   1:118   347082  :  7   1st Qu.:  7.91
              Median :28.00   2: 28   2: 80   CA. 2343:  7   Median : 14.45
              Mean   :29.70   3: 16   3:  5   3101295 :  6   Mean   : 32.20
              3rd Qu.:38.00   4: 18   4:  4   347088  :  6   3rd Qu.: 31.00
              Max.   :80.00   5:  5   5:  5   CA 2144 :  6   Max.   :512.33
              NA's   :177     8:  7   6:  1   (Other) :852
        Cabin     Embarked
           :687    :  2
 B96 B98    :  4   C:168
 C23 C25 C27:  4   Q: 77
 G6         :  4   S:644
 C22 C26    :  3
```

```
D          :  3
(Other)    :186
```

Handling Missing Values

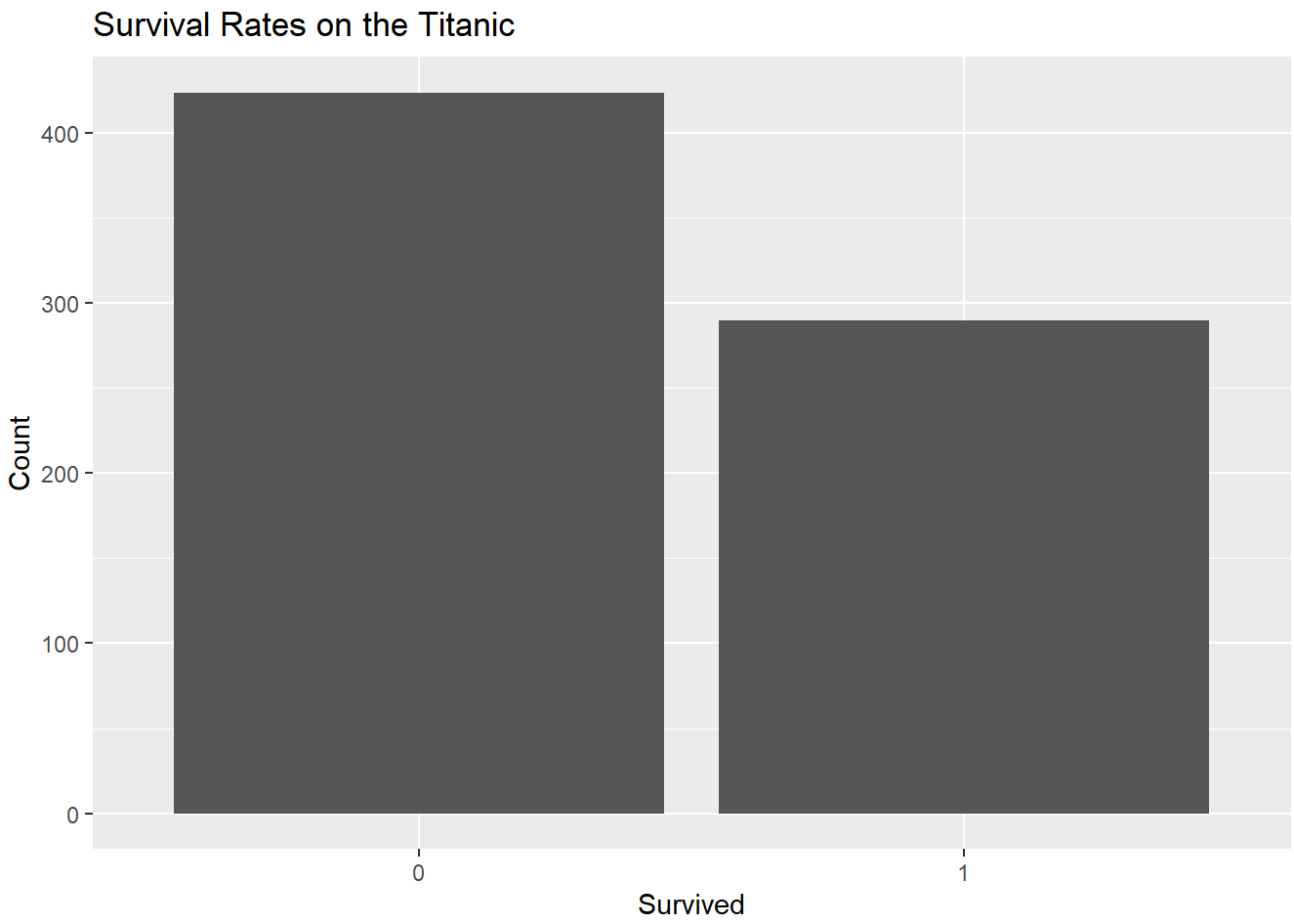
```
# Remove rows with NA values
dt <- na.omit(dt)
summary(dt)
```

```
 PassengerId    Survived Pclass                                      Name
Min.   :  1.0   0:424    1:186   Abbing, Mr. Anthony                   :  1
1st Qu.:222.2   1:290    2:173   Abbott, Mr. Rossmore Edward           :  1
Median :445.0            3:355   Abbott, Mrs. Stanton (Rosa Hunt)      :  1
Mean   :448.6                    Abelson, Mr. Samuel                   :  1
3rd Qu.:677.8                    Abelson, Mrs. Samuel (Hannah Wizosky):  1
Max.   :891.0                    Adahl, Mr. Mauritz Nils Martin        :  1
                                 (Other)                               :708
    Sex           Age        SibSp   Parch            Ticket
female:261   Min.   : 0.42   0:471   0:521   347082      :  7
male  :453   1st Qu.:20.12   1:183   1:110   3101295     :  6
             Median :28.00   2: 25   2: 68   347088      :  6
             Mean   :29.70   3: 12   3:  5   CA 2144     :  6
             3rd Qu.:38.00   4: 18   4:  4   382652      :  5
             Max.   :80.00   5:  5   5:  5   S.O.C. 14879:  5
                             8:  0   6:  1   (Other)     :679
     Fare                Cabin     Embarked
Min.   :  0.00              :529    :  2
1st Qu.:  8.05   B96 B98    :  4   C:130
Median : 15.74   C23 C25 C27:  4   Q: 28
Mean   : 34.69   G6         :  4   S:554
3rd Qu.: 33.38   C22 C26    :  3
Max.   :512.33   D          :  3
                 (Other)    :167
```

Survival Rates:

```
# Calculate overall survival rates
survival_rate <- dt %>% group_by(Survived) %>% summarize(Count = n())

# Plot overall survival rates
ggplot(survival_rate, aes(x = Survived, y = Count)) +
  geom_bar(stat = "identity") +
  labs(x = "Survived", y = "Count", title = "Survival Rates on the Titanic")
```

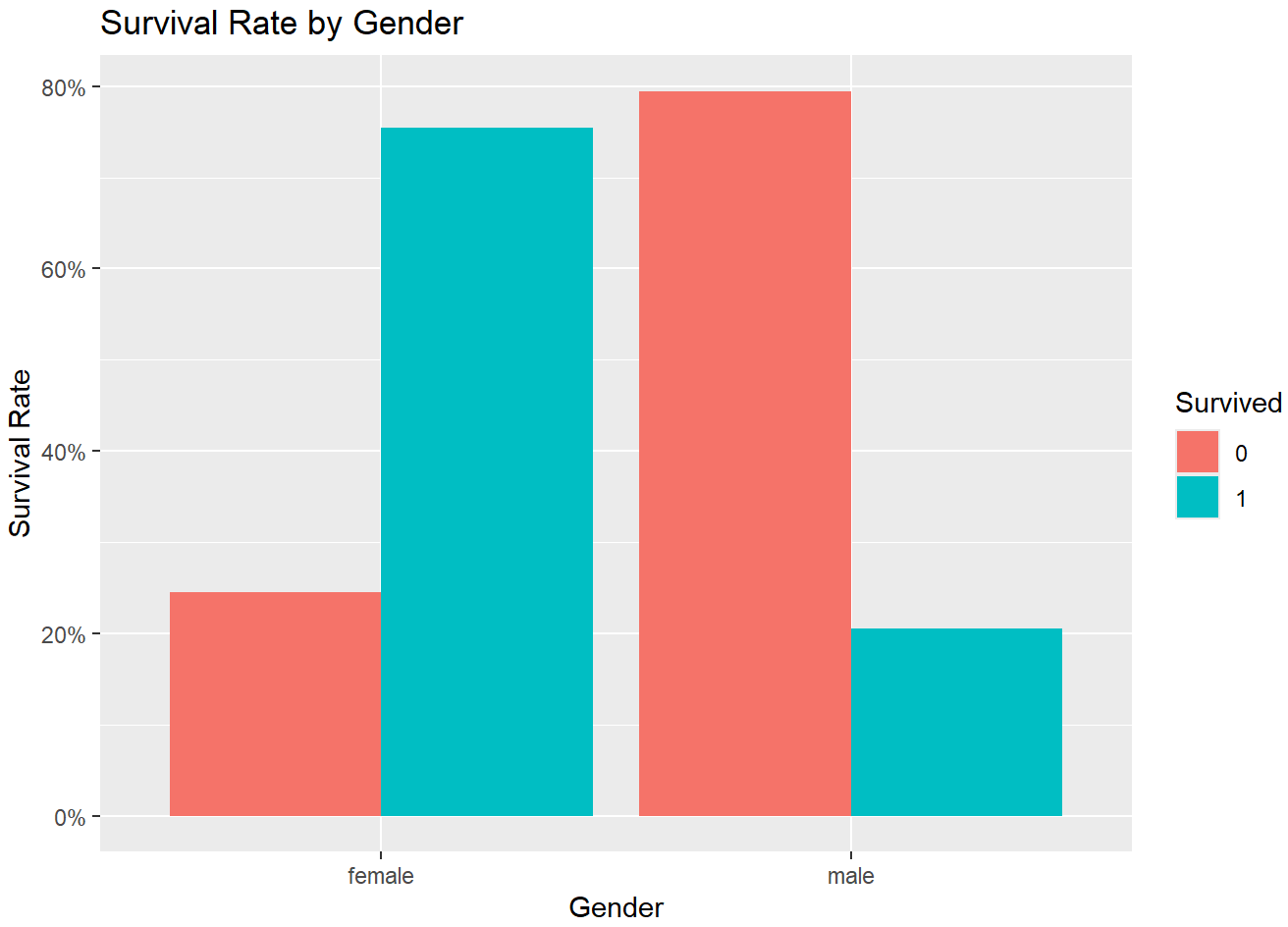


Survival Rate by Gender

```
# Calculate survival rates by gender
gender_survival <- dt %>% group_by(Sex, Survived) %>% summarize(Count = n()) %>%
  mutate(Survival_Rate = Count / sum(Count))
```

```
`summarise()` has grouped output by 'Sex'. You can override using the `.groups`
argument.
```

```
# Plot survival rates by gender
ggplot(gender_survival, aes(x = Sex, y = Survival_Rate, fill = Survived)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Gender", y = "Survival Rate", fill = "Survived", title = "Survival Rate by Gender") +
  scale_y_continuous(labels = scales::percent)
```

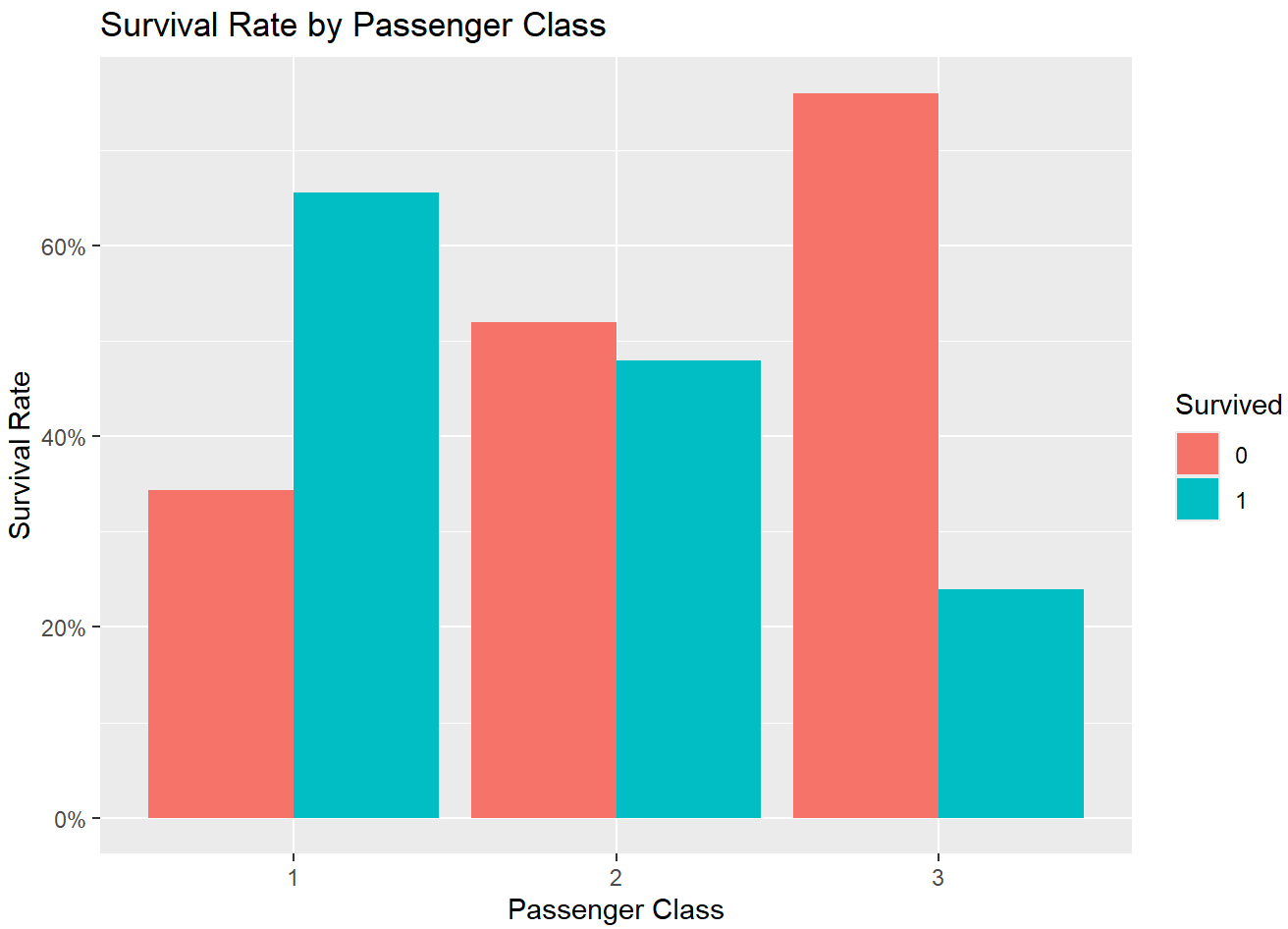


Survival Rate by Passenger Class

```
# Calculate survival rates by passenger class
class_survival <- dt %>% group_by(Pclass, Survived) %>% summarize(Count = n()) %>%
  mutate(Survival_Rate = Count / sum(Count))
```

```
`summarise()` has grouped output by 'Pclass'. You can override using the
`.groups` argument.
```

```
# Plot survival rates by passenger class
ggplot(class_survival, aes(x = Pclass, y = Survival_Rate, fill = Survived)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Passenger Class", y = "Survival Rate", fill = "Survived", title = "Survival Rate by P
  scale_y_continuous(labels = scales::percent)
```



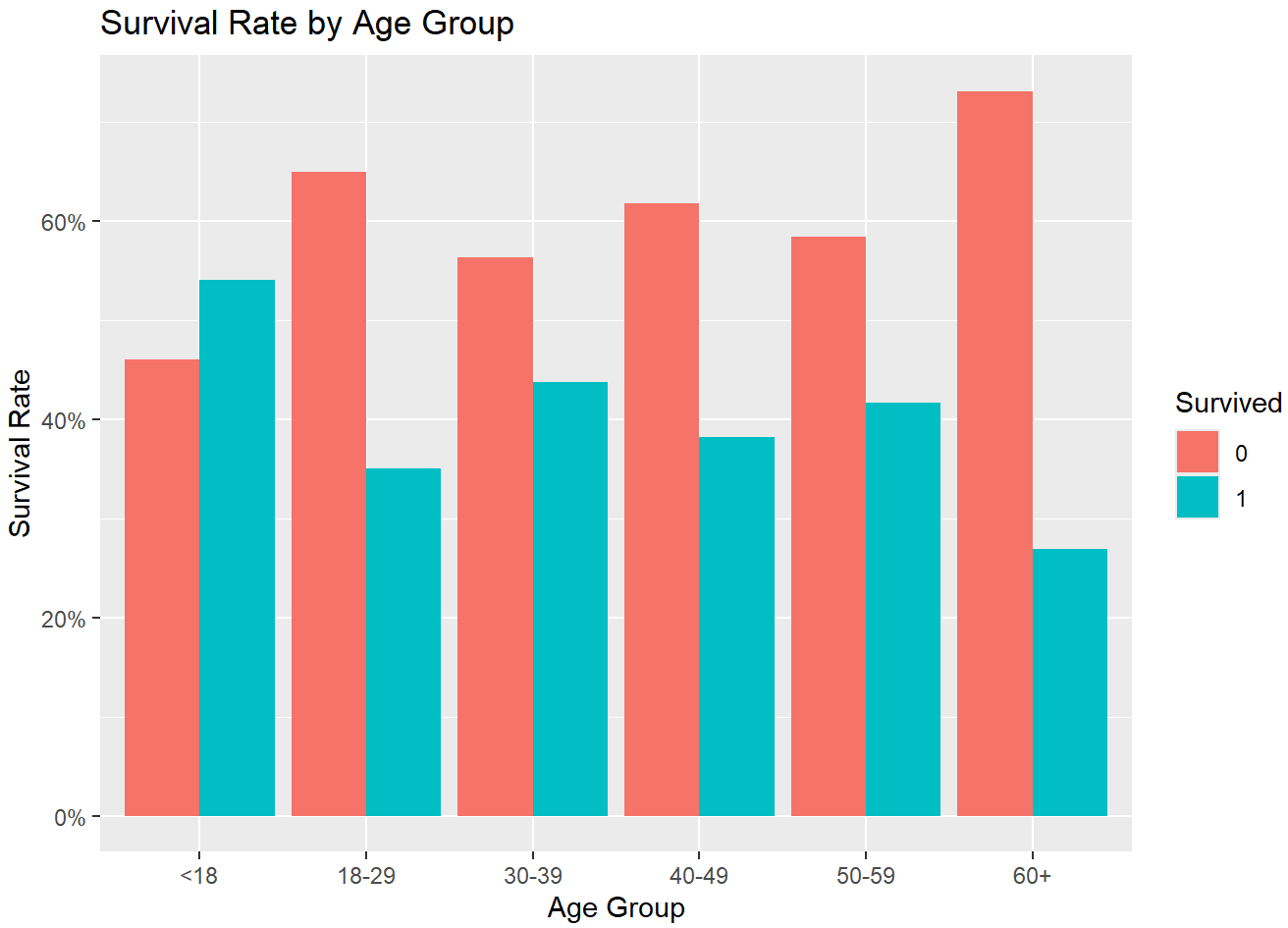
Survival Rate by Age Groups

```
# Create age groups
dt <- dt %>%
  mutate(Age_Group = case_when(
    Age < 18 ~ "<18",
    Age >= 18 & Age < 30 ~ "18-29",
    Age >= 30 & Age < 40 ~ "30-39",
    Age >= 40 & Age < 50 ~ "40-49",
    Age >= 50 & Age < 60 ~ "50-59",
    Age >= 60 ~ "60+",
    TRUE ~ "Unknown"
  ))

# Calculate survival rates by age group
age_survival <- dt %>% group_by(Age_Group, Survived) %>%
  summarize(Count = n()) %>% mutate(Survival_Rate = Count / sum(Count))
```

```
`summarise()` has grouped output by 'Age_Group'. You can override using the
`.groups` argument.
```

```
# Plot survival rates by age group
ggplot(age_survival, aes(x = Age_Group, y = Survival_Rate, fill = Survived)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Age Group", y = "Survival Rate", fill = "Survived", title = "Survival Rate by Age Grou
  scale_y_continuous(labels = scales::percent)
```



Conclusion

This report analyzed the Titanic dataset, revealing significant insights into survival rates based on gender, passenger class, and age groups.

https://github.com/NaomiPang01/Statistical-Consulting/tree/main