A Comprehensive Analysis of the Titanic Dataset

Jan 2024

Contents

LIST OF TABLES

LIST OF FIGURES

Introduction

The Titanic dataset, a widely utilized resource in machine learning tutorials and competitions, comprises details about the Titanic's passengers, encompassing their names, ages, genders, social classes, and their survival status. This dataset is segmented into two sections: a training set, which encompasses information about passengers who either survived or perished, and a test set, containing data about passengers with unconfirmed survival statuses. This dataset not only offers a glimpse into a tragic historical event but also serves as a versatile platform for delving into a wide range of data analysis and machine learning inquiries.

Key Applications in Machine Learning:

- **Classification:** Predicting passenger survival.
- **Regression:** Estimating the probability of survival.

- **Clustering:** Grouping passengers based on characteristics.
- **Anomaly Detection:** Identifying unexpected survival outcomes.

Beyond its richness, the dataset introduces challenges and opportunities for addressing critical data-related questions:

- **Handling Missing Data:** Common in real-world data, addressing missing values through imputation or exclusion is essential for robust analysis.
- **Handling Realistic Data:** Reflecting real-world nuances, the dataset contains variations, missing information, and inconsistencies, providing valuable experience in dealing with imperfections.
- **Survival Rate:** Determining the overall survival rate among Titanic passengers sets the stage for more in-depth analyses.
- **Survival Rates by Age:** Exploring age-related survival patterns, including the "women and children first" principle.
- **Survival Rate by Gender:** Analyzing how gender influenced survival chances sheds light on societal norms.
- **Survival Rate by Class of Ticket:** Investigating the relationship between socio-economic status and survival outcomes.
- **Fare Distribution by Embarkation Point:** Understanding fare distribution by embarkation point reveals socio-economic disparities among passengers.
- **Survival Rate by Class of Ticket and Gender:** Examining survival rates by both ticket class and gender unveils potential interaction effects.
- **Distribution of Passenger Ages:** An overview of age distribution aids in understanding the demographics of Titanic passengers.
- **Survival Rates by Age, Segmented by Gender and Class of Ticket:** Segmenting survival rates by age, gender, and ticket class provides a nuanced perspective on factors influencing survival probabilities.

- **Survival Rates by Fare and Age:** Investigating survival rates by fare and age adds another layer to understanding survival dynamics.

By addressing these questions and challenges within the Titanic dataset, data analysts and machine learning practitioners can hone their skills, extract valuable insights, and contribute to our understanding of this historical event.

Literature Review:

Data visualization is a tool for deciphering complex datasets, and its significance becomes particularly pronounced when scrutinizing historical events such as the tragic Titanic. This literature review aims to explore various research studies on data visualization techniques, specifically pertinent to the analysis of Titanic survival.

A seminal work in the realm of data visualization is Edward Tufte's groundbreaking book, "The Visual Display of Quantitative Information" (Tufte 1983). Tufte's principles serve as a cornerstone, emphasizing the paramount importance of clarity, precision, and effectiveness in visually presenting data. These principles provide a robust foundation for constructing visualizations that accurately convey information and resonate with and captivate the audience.

Integral to contemporary data visualization is the ggplot2 package, a brainchild of Hadley Wickham (Wickham 2016). Wickham's work is deeply rooted in Wilkinson's "Grammar of Graphics" (Wilkinson 2005), offering a systematic framework for constructing visualizations. The ggplot2 package, celebrated for its simplicity and flexibility, emerges as a vital tool in creating a diverse range of visualizations. Its relevance becomes particularly apparent when navigating the multifaceted dimensions of Titanic survival data.

The significance of interactivity in data visualization is explored in the research conducted by Heer and Shneiderman (2012). Their work underscores the empowering nature of interactive visualizations, allowing users to dynamically explore and comprehend intricate datasets. This facet becomes indispensable when unravelling the complexities of Titanic passenger information, enabling a nuanced exploration of the factors influencing survival.

Uncertainty visualization takes centre stage in the study by Kay et al. (2016), delving into techniques for effectively communicating uncertainty in predictive systems. Given the inherent incompleteness of historical data, understanding and conveying uncertainty become paramount in Titanic survival analysis. Techniques such as probabilistic visualizations and uncertainty intervals contribute significantly to a more nuanced dataset interpretation.

The concept of narrative visualization, introduced by Segel and Heer (2010), assumes relevance when aiming to weave a coherent story through data. The diverse variables and socio-economic factors inherent in Titanic survival data make it an ideal candidate for narrative visualization techniques. The study underscores the pivotal role of storytelling in guiding audiences through the intricacies of the dataset, rendering the analysis more accessible and engaging.

Considering the cognitive aspects of visualization, Ware's work (Ware 2012) on information visualization provides valuable insights into how humans perceive and interpret visual information. Applying principles of visual perception to the design of Titanic survival visualizations proves instrumental in enhancing their effectiveness. Considerations such as judicious colour choices, thoughtfully crafted visual hierarchies, and strategic attention allocation become pivotal for creating visualizations that resonate with the cognitive processes of the audience.

In summary, the body of work on data visualization techniques provides a strong basis for the intricate examination of Titanic survival data. Drawing inspiration from Tufte's focus on clarity, embracing the interactive features highlighted by Heer and Shneiderman, and leveraging the flexibility of ggplot2, scholars can craft visualizations that not only accurately convey information but also encourage a more profound comprehension of the diverse factors impacting survival aboard the Titanic.

1.Project Objective

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. The objective of the project involves answering the question “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc) (Lord, 1955).

The Project is about Exploratory Data Analysis (EDA) with the help of R Language or R Studios as a Tool to identify the answers.

2.Assumptions

It has been assumed that the fare is in $.

3. Exploratory Data Analysis

Titanic dataset's Exploratory Data Analysis (EDA) employs a systematic approach to unveil patterns and insights, enabling researchers to understand its structure and derive preliminary observations. This vital phase facilitates data scientists' familiarity with the dataset, forming the foundation for in-depth analysis (Tukey, 1977).

3.1 Data Loading

Initiate the project by installing R and RStudio. Download R for statistical computing and install RStudio, an IDE enhancing the R experience with user-friendly features and project management tools (Grolemund and Wickham, 2016).

3.2 Variable Identification

Define variables by determining their data type using the R function str(TitanicData). Classify variables as continuous, discrete, dependent, or independent. Numeric values like 0, 1, 2, 3 hold specific literal meanings in the data.

Variable	Definition	Data Type	Key
survived	Survival	int	0 = No 1 = Yes
pclass	Ticket class	int	1 = 1st 2 = 2nd 3 = 3rd
sex	Male/Female	chr	
age	Age in years	int	
embarked	port of embarkation for each passenger	chr	C: Cherbourg Q: Queenstown S: Southampton
sibsp	# of siblings/spouses aboard the Titanic	int	
parch	# of parents/children aboard the Titanic	int	
ticket	Ticket number	chr	
fare	Passenger fare	float	
cabin	Cabin number	chr	

Table 1: Variable table

3.3 Univariate Analysis

Univariate data consists of only one variable. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

3.3.1 SURVIVED

With 1308 observations, a binary representation (0 and 1) distinguishes between survivors and survivors. A simple table illustrates a higher count of non-survivors.

survived	Count	Percentage	Total
0	808	62	1308
1	500	38	

Table 2: Titanic Survival Rate.

The finalized chart, titled "Titanic Survival Rates," visually conveys the distribution of survival cases, enhancing clarity in passenger outcomes (Eaton and Haas, 1995).

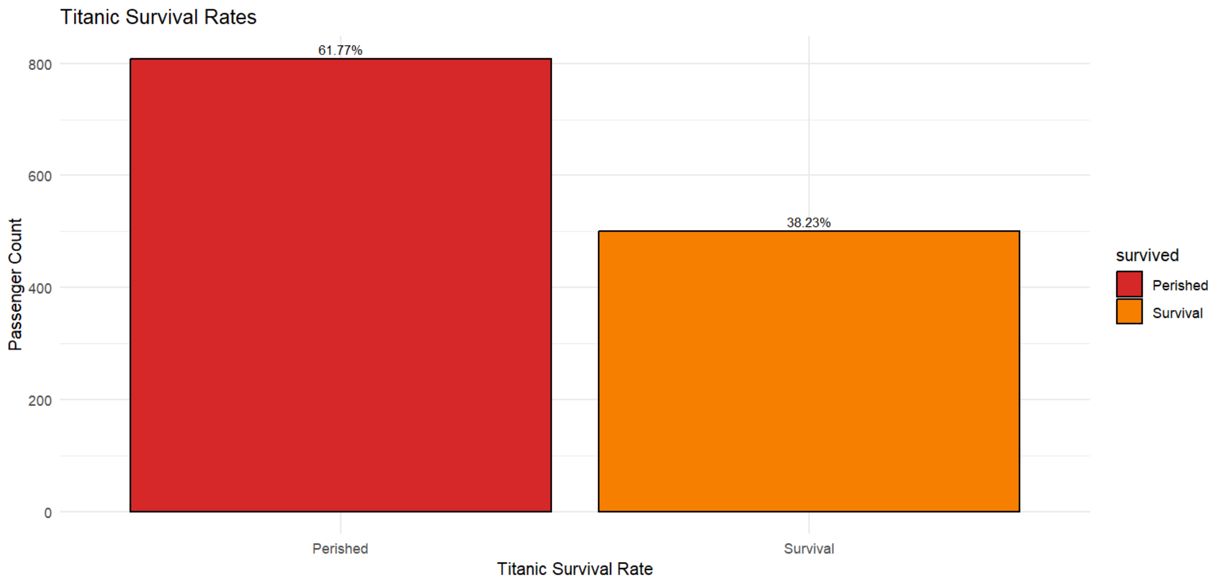

Figure 1: Titanic Survival Rate

3.3.2 Passenger Class Distribution

Passenger Class (pclass) includes three classes, with class 3 having the highest number, almost double the sum of classes 1 and 2. Notably, Pclass 1 has more passengers than Pclass 2. The higher prevalence in class 3 may be linked to its lower fare, offering insight into passenger distribution based on class and associated fares.

pclass	Count	Total
1	323	1308
2	277	
3	708	

Table 3: Number of Passengers in Each Class.

The plot is finalized with axis labels ("Passenger Class" and "Passenger Count"), a title ("Number of Passengers in Each Class"). The x-axis represents the passenger class, the y-axis represents the count of passengers in each class (Lord, 1955).

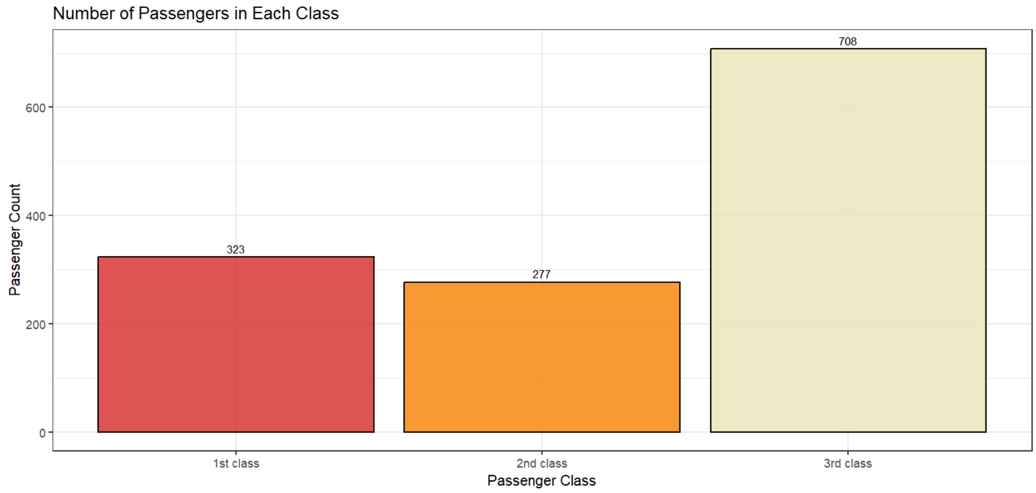

Figure 2: Number of Passengers in Each Class.

3.3.3 Passenger Distribution by Gender

Out of the entire passenger count, there are 842 males and 466 females, with the male count approximately twice that of the female count.

Sex	Count	Total
male	842	1308
female	466	

Table 4: Gender Counts in Titanic.

A bar plot visually displays gender counts with labelled bars, titled "Gender Counts in Titanic," using distinct colours for male and female. This analysis offers a clear overview of passenger distribution, providing valuable insights into Titanic's demographics.

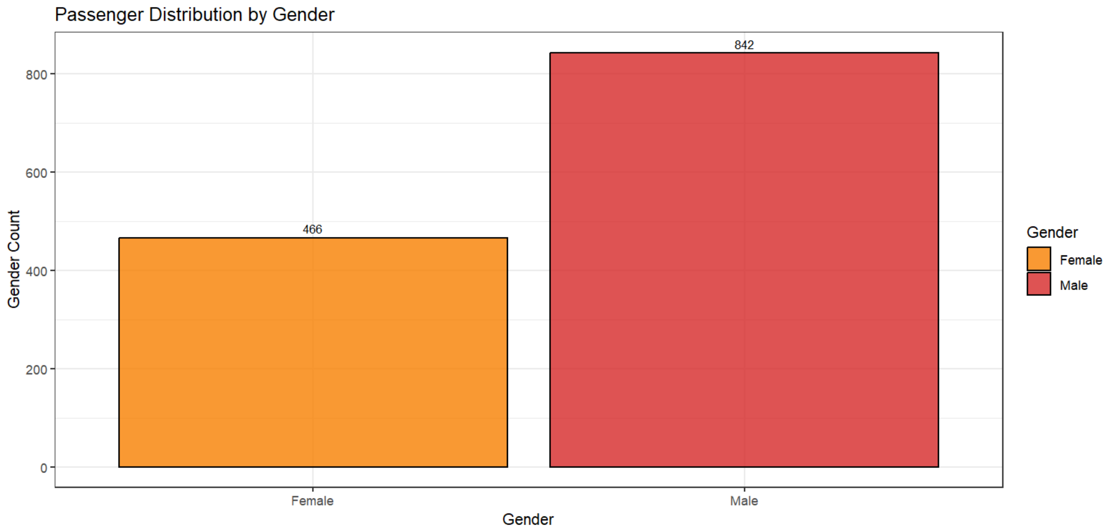

Figure 3: Gender Counts in Titanic.

3.3.5 Age Categories

The Titanic dataset's age distribution is right skewed, with a median age of 29.5 years. The majority falls between 20 and 30 years, and elderly passengers are scarce, comprising only 1%. Distinct age dynamics exist among passenger classes, with first-class passengers generally older. Survival outcomes reveal age's crucial role, as children and young adults have a higher likelihood of survival compared to older individuals, emphasizing age as a significant factor in Titanic's survival rates (Hart, 2012).

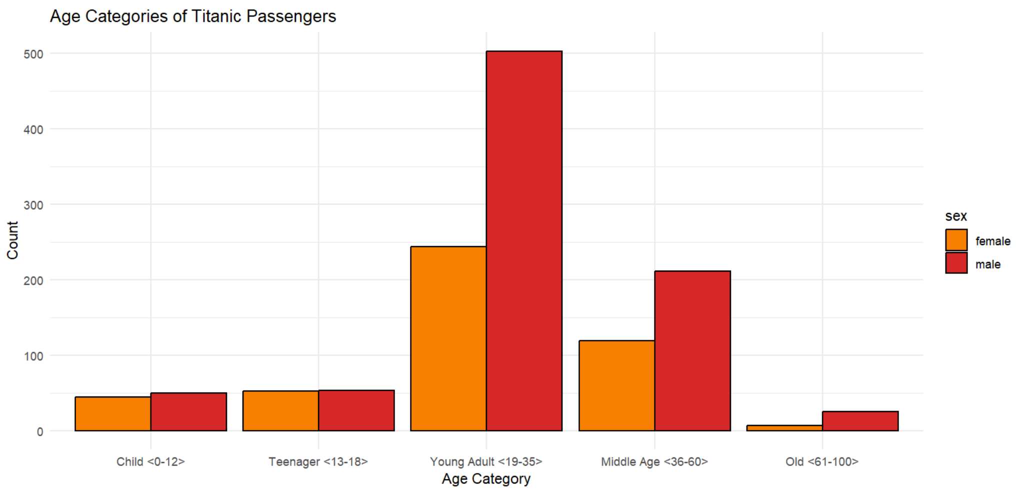

Figure 4: Distribution of Passenger Ages in Titanic.

3.3.6 Embarkation Point

Most Titanic passengers (70%) boarded in Southampton (S), followed by Cherbourg (C) at 21% and Queenstown (Q) at 9%. This distribution reflects various factors, including port locations, shipping company popularity, and passenger socioeconomic status. Southampton's proximity to London likely made it appealing for UK passengers. Cherbourg and Queenstown, in France and Ireland, may have been favoured by passengers from those countries. Southampton's status as the home port for the White Star Line, the Titanic's operator, may have contributed to its high embarkation rate. Socioeconomically, those affording higher fares and seeking luxurious accommodations likely chose Southampton, closer to London's commerce hub (Boyer, 1998).

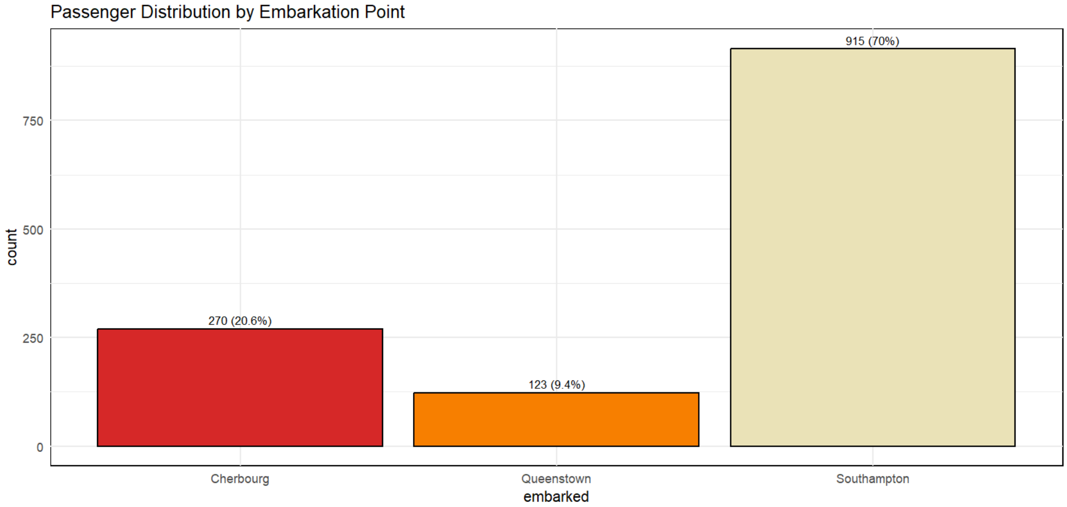

Figure 5: Passenger Distribution by Embarkation Point.

3.4 Bivariate Analysis

Bivariate data encompasses two distinct variables. Analysing such data involves exploring causes and relationships, seeking to uncover connections between the two variables.

3.4.1 Age Distribution by Gender

The box plot examines age distribution among male and female Titanic passengers, exploring potential survival rate patterns by age and gender. Data is separated by gender (x = sex), depicting age distribution (y = age) with filled boxes indicating the interquartile range. Median lines indicate central tendencies, and whiskers extend to show the overall age range. The plot reveals that males outnumber females, and, on average, females appear younger. This analysis provides insights into age distribution by gender, suggesting further exploration of survival rates by age and gender differences to deepen understanding of factors influencing passenger survival (Davies and Beveridge, 2012).

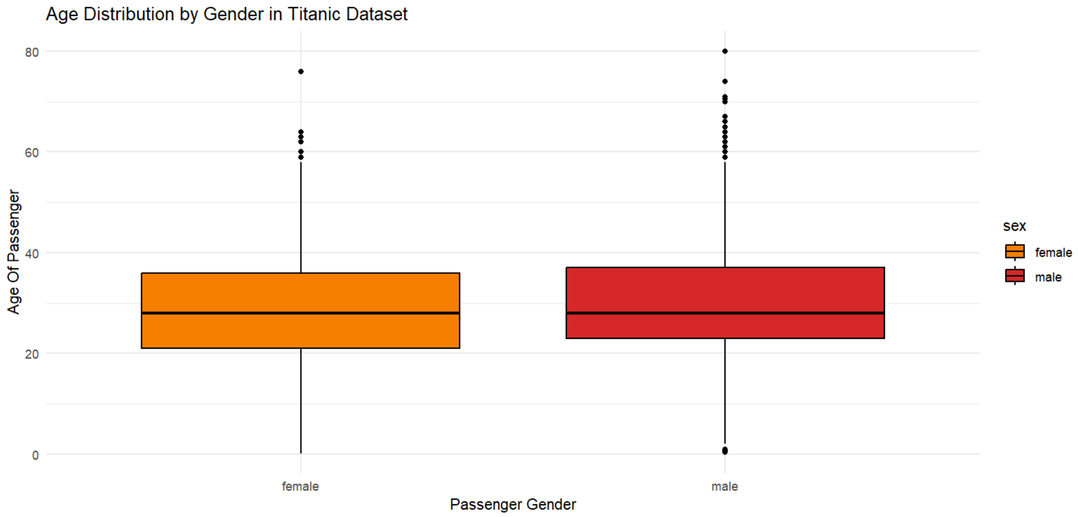

Figure 6: Age Distribution by Gender in Titanic Dataset.

3.4.2 pclass vs Gender

Men outnumber women in all passenger classes. However, the gender ratio is significantly higher in the 3rd class.

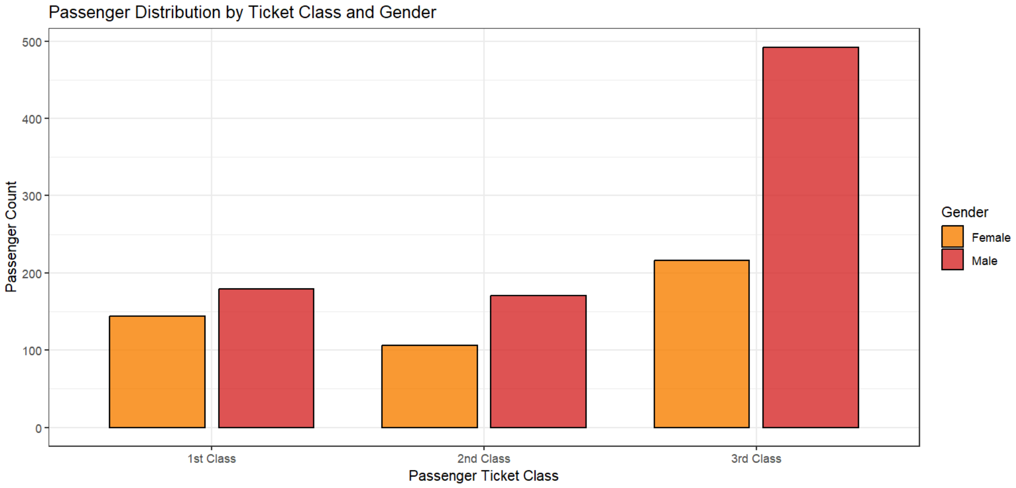

Figure 7: Passenger Distribution by Ticket Class and Gender.

3.4.3 Survived vs Gender

There is a higher incidence of male fatalities, while females exhibit a greater survival rate than males. The previous visualizations indicate a higher male presence in the third class compared to other classes.

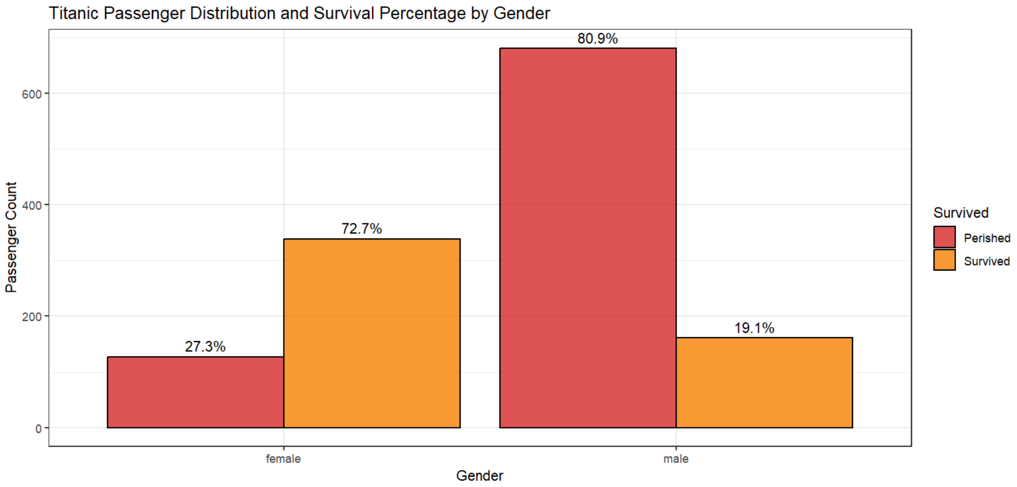

Figure 8: Passenger Distribution and Survival Percentage by Gender.

3.4.4 Survived vs pclass

Those who paid higher fares were given preference in the rescue efforts. Passengers in 1st class had significantly higher survival probabilities compared to those in other classes.

pclass	Survived	Perished	Total	Proportion Survived
1st class	200	123	323	0.62
2nd class	119	158	277	0.43
3rd class	181	527	708	0.25

Table 5: Survival Rate by Passenger Ticket Class.

This suggests that passengers in higher pclasses were more likely to survive the sinking of the Titanic. This may be due to several factors, such as having access to better lifeboats, having more experience with swimming, or being in better physical condition.

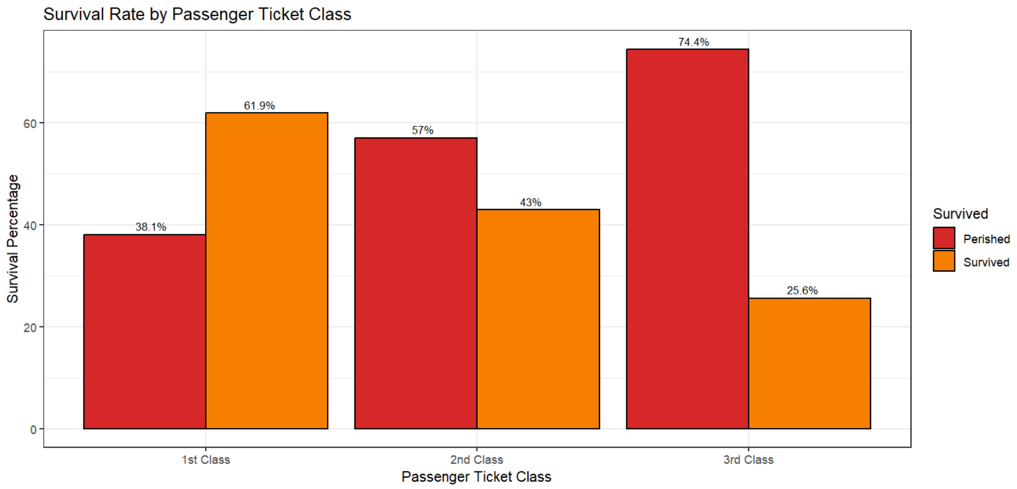

Figure 9: Survival Rate by Passenger Ticket Class.

3.5 Multi variate Analysis

Multivariate analysis explores relationships among multiple variables simultaneously, offering a nuanced understanding of their complex interactions within a dataset. This approach allows for a thorough investigation of the intricate interplay between various elements, providing valuable insights into the multidimensional nature of the data (Hair et al., 2010).

3.5.1 Age vs Fare by Survival Status

A chart plotting values of two variables on two axes, with the arrangement of points indicating any existing correlation. It is observed that most passengers choose fares ranging from 0 to 50, regardless of their age.

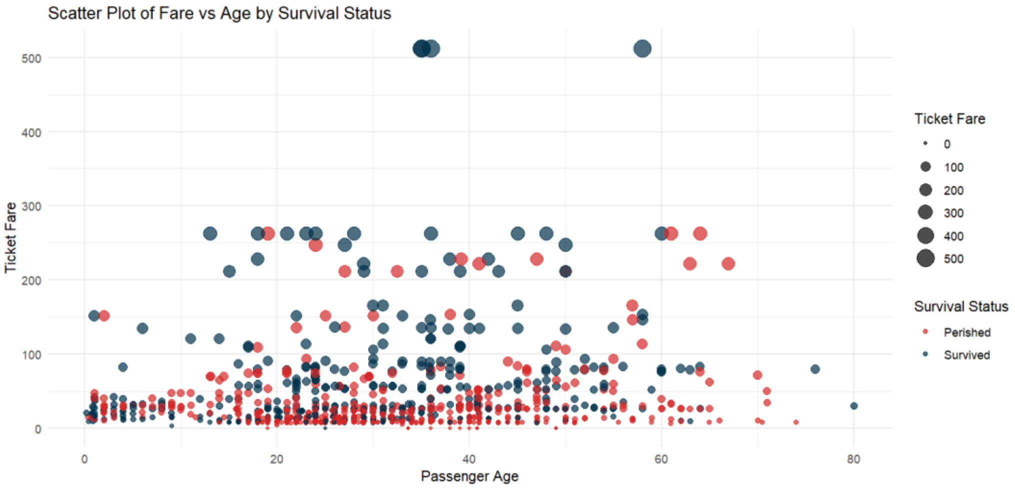

Figure 10: Scatter Plot of Fare vs Age by Survival Status.

3.5.2 pclass vs sex by survival Status

Multivariate analysis depicts the relationship between more than two variables. The below table provides the total 1308 passengers, 466 females and 842 males are listed. Survival among females is significantly higher in comparison to males across all classes.

For females:

139 out of 339 females survived, with the highest survival rates in the first class, followed by more than half in the second class, and around one in ten in the third class. 127 females did not survive, with the majority being from the third class.

For males:

161 out of 842 males survived. The survival rate for males was highest in the first class (about half), significantly lower in the second class, and minimal in the third class. A large majority of males did not survive, with the highest number of non-survivors in the third class.

		pclass	**1st class**	**2nd class**	**3rd class**	**Total**
sex	**survived**					
Female	**0**		5	12	110	**127**
	1		139	94	106	**339**
Male	**0**		118	146	417	**681**
	1		61	25	75	**161**
Total			**323**	**277**	**708**	**1308**

Table 6: Multivariate analysis Survival Status by pclass vs sex.

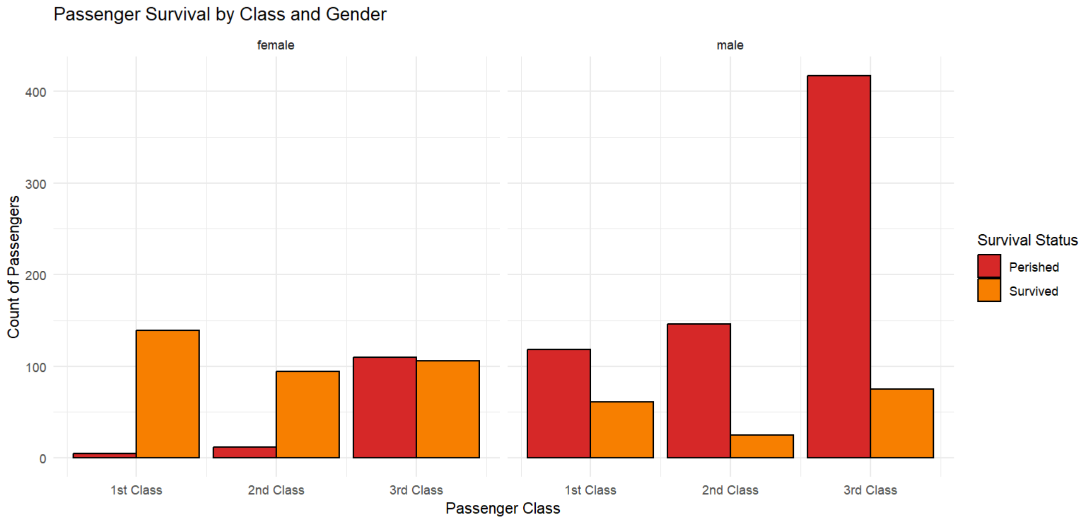

Figure 11 Passenger Survival by Pclass and Gender

This analysis clearly indicates that class and gender were significant factors in survival probabilities. Females had a much higher chance of survival than males, and first-class passengers had a much higher chance of survival than those in the lower classes, reflecting the social hierarchies and evacuation priorities of the time.

3.6 Handling Missing Data

Identifying and handling missing values is a crucial step in the data preprocessing phase. In the context of the Titanic dataset, which captures information about passengers aboard the ship, understanding and addressing missing values is vital for conducting accurate analyses. Here's an expanded explanation of the process:

3.6.1 Missing values on 'embarked' column

In the Titanic dataset, the 'Embarked' column has two missing values. Options for handling these gaps include:

Drop rows with missing values, losing data.

Impute missing values using mean, median, or mode.

Create a new category for missing values in the 'Embarked' column.

Imputation, a widely adopted technique, is employed to manage missing data and ensure the dataset's suitability for analysis. In this instance, missing values in the "embarked" column are imputed with the most prevalent embarkation point, denoted as "S." (Rubin, 1987).

3.6.2 Missing and negative values on 'age' column

There are 263 missing values and 8 negative values in the 'Age' column. These issues can be addressed through appropriate data cleaning and handling techniques. This approach involves imputing missing ages using a regression model, creating a subset for missing data, and correcting negative age values by replacing them with the mean age, ensuring a thorough and documented data-handling process (Little & Rubin, 2002).

3.6.3 Missing and negative values on 'fare' column

The 'Fare' column in the Titanic dataset has a single missing value, associated with a non-survivor, and no negative values. Possible reasons include passengers not prepaying or incomplete payment recording, with data loss during collection also possible. Addressing the missing 'Fare' involves a regression model for prediction. Using a linear regression model, justified by the 'Fare'-'pclass' correlation, allows replacing the missing fare with predicted values, enhancing data completeness (Montgomery et al., 2012).

4 Conclusion

In conclusion, the analysis of the Titanic dataset reveals several key factors influencing survival rates. For males, optimal survival chances were observed among children not in the third class and first-class individuals in the young adult or middle-aged categories. Conversely, the highest overall survival likelihood was identified for females outside the third class. Furthermore, a predominant age group among passengers falls within the 20-40 range. Gender-wise, the data underscores that women had a higher survival rate compared to men. Additionally, a striking correlation emerges between socio-economic class and survival, indicating that as the class number increased (indicating cheaper fares), the likelihood of survival decreased, elucidating the socio-economic disparities in the tragic events aboard the Titanic.

5 Appendix

```
# Libraries
library(tidyverse)
library(dslabs)
library(ggplot2)
library(dplyr)

# Load an external CSV file
TitanicData <- read.csv("./titanic.csv")

#Sorting data
sort(TitanicData$total)
str(TitanicData)

#--------------------Missing values on ‘embarked’ column------------------------

# Check missing embarked data
table(is.na(TitanicData$embarked))

# Filter rows in the Titanic data set where "Embarked" is empty (contains NA values)
TitanicFiltered <- TitanicData %>%
  filter(is.na(embarked))

# Replace empty values with NA in the entire dataset
TitanicData <- TitanicData %>%
  mutate_all(~ifelse(. == "", NA, .))
```

```
# Create a subset without missing values in the
"Embarked" column

TrainTitanic <-
TitanicData[!is.na(TitanicData$embarked), ]


# Find missing values in the "Embarked" column

MissingEmbarked <-
TitanicData[is.na(TitanicData$embarked), ]


# Print the missing values

print(MissingEmbarked)


# Explore the data to see the most common value in the
"embarked" column

table(TitanicData$embarked)


# Impute missing values in the "Embarked" column with
"S"

TitanicData$embarked <- ifelse(

  is.na(TitanicData$embarked),

  "S",

  TitanicData$embarked

)


#-------------------End Missing values on ‘embarked’
column----------------------


#-----------------Missing and negative values on ‘age’
column--------------------


# Print negative age values

NegativeAgeValues <- TitanicData$age[TitanicData$age <
0]
```

```
print("Negative Age Values:")
print(NegativeAgeValues)

# Check missing age data
table(is.na(TitanicData$age))

# Check which rows in the "Age" column have missing values
MissingAge <- is.na(TitanicData$age)

# Create a linear regression model for age imputation
LmModel <- lm(
  age ~ pclass + sibsp + parch + fare + embarked,
  data = TitanicData
)

# Impute Missing Ages Using the Regression Model
# Create a subset of the data with missing ages
MissingAgeData <- TitanicData[MissingAge,]

# Predict missing ages
PredictedAges <- predict(LmModel, newdata = MissingAgeData)

# Replace missing values with predicted ages
TitanicData$age[MissingAge] <- PredictedAges

# To correct negative age values in the Titanic dataset
# Convert 'age' to numeric (if it's not already)
TitanicData$age <- as.numeric(as.character(TitanicData$age))
```

```
# Identify and replace negative ages with the mean age
MeanAge <- mean(TitanicData$age[TitanicData$age >= 0],
na.rm = TRUE)
TitanicData$age[TitanicData$age < 0] <- MeanAge



#--------------End Missing and negative values on ‘fare’
column-----------------


#----------------Missing and negative values on ‘fare’
column-------------------
# Check which rows in the "Fare" column have missing
values
# Check for missing fare values
MissingFare <- is.na(TitanicData$fare)
print("Missing Fare Values:")
print(sum(MissingFare))

# Filter out rows with negative fare values
NegativeFareValues <- TitanicData$fare[TitanicData$fare
< 0]
print("Negative Fare Values:")
print(NegativeFareValues)

# Remove rows with negative fare values
TitanicData <- TitanicData %>%
  filter(fare >= 0)

# Check which rows in the "Fare" column have missing
values
MissingFare <- is.na(TitanicData$fare)
```

```
# Create a linear regression model for fare imputation
LmModelFare <- lm(
  fare ~ pclass + age + sibsp + parch + embarked,
  data = TitanicData
)

# Impute Missing Fares Using the Regression Model
# Create a subset of the data with missing fares
MissingFareData <- TitanicData[MissingFare,]

# Predict missing fares
PredictedFares <- predict(LmModelFare, newdata =
MissingFareData)

# Replace missing values with predicted fares
TitanicData$fare[MissingFare] <- PredictedFares

#--------------End Missing and negative values on ‘fare’
column-----------------

# Check for Missing Data
any(is.na(TitanicData$embarked))
any(is.na(TitanicData$age))
any(is.na(TitanicData$fare))
#-----------------------------------------

#--------------The survival rate?----------------
#1. What was the survival rate?
# Calculate percentages
PercentageData <- TitanicData %>%
  group_by(survived) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

# Custom colours
PerishedColour <- "#D62828"
SurvivalColour <- "#F77F00"

# Convert 'survived' to a factor
PercentageData$survived <- factor(
  PercentageData$survived,
  labels = c("Perished", "Survival")
)

# Create a bar plot for survival rates with custom colours
ggplot(PercentageData, aes(x = survived, y = Count, fill = survived)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = paste0(round(Percentage, 2), "%")), vjust = -0.5, size = 3) +
  scale_fill_manual(values = c(Perished = PerishedColour, Survival = SurvivalColour)) +
  labs(x = "Titanic Survival Rate"
       , y = "Passenger Count"
       , title = "Titanic Survival Rates") +
  theme_minimal()

#-------------End survival rate--------------------

#---------distribution of Passenger Ages------------
```

```
#2. what was the distribution of Passenger Ages in
Titanic?
# Create age categories
# Defines the age intervals for categorization (0-12,
13-18, 19-35, 36-60, 61-100)
AgeBins <- c(0, 12, 18, 35, 60, 100)
AgeLabels <- c("Child", "Teenager", "Young Adult",
"Middle Age", "Old")
TitanicData$AgeCategory <- cut(
  TitanicData$age,
  breaks = AgeBins,
  labels = AgeLabels,
  include.lowest = TRUE
)


# Create a bar plot for age categories with custom
colours
ggplot(TitanicData, aes(x = AgeCategory, fill = sex)) +
  geom_bar(position = "dodge", color = "black",
show.legend = TRUE) +
  labs(
    title = "Age Categories of Titanic Passengers",
    x = "Age Category",
    y = "Count",
    fill = "sex"
  ) +
  scale_fill_manual(values = c("#F77F00", "#D62828")) +
  theme_minimal()


#--------End distribution of Passenger Ages---------


# -----------Survival rates by age?-----------------
```

```
#3. What was the survival rates by age?
# Custom colours
FemaleColour <- "#F77F00"
MaleColour <- "#D62828"

# Create a boxplot for age distribution by gender with
custom colours
ggplot(TitanicData, aes(x = sex, y = age, fill = sex)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c(female = FemaleColour,
male = MaleColour)) +  # Set custom fill colours
  labs(
    title = "Age Distribution by Gender in Titanic
Dataset",
    y = "Age Of Passenger",
    x = "Passenger Gender"
  ) +
  theme_minimal()

#----------End survival rate by age-----------------
# ---------Gender Proportions in Titanic-----------
#4. What was the gender Counts in Titanic?
# Filter the dataset to only include passengers with
known genders
TitanicFiltered <- TitanicData[!is.na(TitanicData$sex),
]
# Calculate gender counts
GenderCounts <- TitanicFiltered %>%
  group_by(sex) %>%
  summarise(GenderCount = n())
# Create a bar plot with count labels
```

```
ggplot(GenderCounts, aes(
  x = factor(sex, labels = c("Female", "Male")),
  y = GenderCount
)) +
  geom_bar(
    stat = "identity",
    aes(fill = sex),
    alpha = 0.8,
    color = "black"
  ) +
  geom_text(
    aes(label = GenderCount),
    vjust = -0.5,
    size = 3
  ) +  # Add count labels
  labs(
    x = "Gender",
    y = "Gender Count",
    title = "Gender Counts in Titanic"
  ) +
  scale_fill_manual(
    values = c("#F77F00", "#D62828"),
    name = "Gender"
  ) +  # Specify colours for male and female
  theme_bw()


#---------End gender Proportions in Titanic---------

#--------survival rate by gender?--------------------
```

```
#5. What was the survival rate by gender?
# Filter the dataset to only include passengers with known genders
TitanicData <- TitanicData %>%
  filter(!is.na(sex))  # 'Sex' is used instead of 'sex'
# Calculate the number of passengers and survival percentage for each gender
GenderData <- TitanicData %>%
  group_by(sex, survived) %>%
  summarise(
    PassengerCount = n()
  ) %>%
  group_by(sex) %>%
  mutate(SurvivalPercentage = (PassengerCount / sum(PassengerCount)) * 100)

# Create a bar scatter plot with passenger count and survival percentage
ggplot(GenderData, aes(
  x = sex,
  y = PassengerCount,
  fill = factor(survived)
)) +
  geom_col(
    position = "dodge",
    color = "black",
    alpha = 0.8
  ) +
  geom_text(
    aes(label = paste0(round(SurvivalPercentage, 1), "%")),
    position = position_dodge(width = 0.8),
    vjust = -0.5
  ) +
  labs(
    x = "Gender",
    y = "Passenger Count",
    title = "Titanic Passenger Distribution and Survival
Percentage by Gender"
  ) +
  scale_fill_manual(
    values = c("#D62828", "#F77F00"),
    name = "Survived",
    labels = c("Perished", "Survived")
  ) +
  theme_bw()


#----------End survival rate by gender--------------

#----------- ticket Class Proportions---------------

#6. What was the ticket Class Proportions?
# Filter the dataset to only include passengers with
known classes
# Calculate passenger counts
PassengerCounts <- table(TitanicData$pclass)
# Convert counts to a data frame
PassengerCountsDf <- as.data.frame(PassengerCounts)
names(PassengerCountsDf) <- c("Passenger_Class",
"Count")
# Create a bar plot
ggplot(PassengerCountsDf, aes(
  x = factor(Passenger_Class, labels = c("1st Class",
"2nd Class", "3rd Class")),
  y = Count
)) +
  geom_bar(
    stat = "identity",
    fill = c("#D62828", "#F77F00", "#EAE2B7"),
    color = "black",
    alpha = 0.8
  ) +
  geom_text(
    aes(label = Count),
    vjust = -0.5,
    size = 3
  ) +  # Add count labels
  labs(
    x = "Passenger Class",
    y = "Passenger Count",
    title = "Number of Passengers in Each Class"
  ) +
  theme_bw()

#-----------End ticket Class Proportions-----------
#--Passenger Distribution by Ticket Class and Gender-
#7. what was the Passenger Distribution by Ticket Class
and Gender
ggplot(TitanicData, aes(
  x = factor(pclass, labels = c("1st Class", "2nd
Class", "3rd Class")),
  fill = sex
)) +
  geom_bar(
    position = position_dodge(0.8),
    stat = "count",
    width = 0.7,
    color = "black",
    alpha = 0.8
  ) +
  labs(
    x = "Passenger Ticket Class",
    y = "Passenger Count",
    title = "Passenger Distribution by Ticket Class and Gender"
  ) +
  scale_fill_manual(
    values = c("#F77F00", "#D62828"),
    name = "Gender",
    labels = c("Female", "Male")
  ) +
  theme_bw()
# Create a contingency table for pclass, sex, and survived
ContingencyTable <- table(
  TitanicData$pclass,
  TitanicData$sex,
  TitanicData$survived
)
# Display the contingency table
print(ContingencyTable)
#---End Passenger Distribution by Ticket Class and Gender
# Check missing fare data
```

```
table(is.na(TitanicData$fare))
# Bar plot of 'Fare' by 'Embarked'
# Passenger Count and Percentage by Embarkation Point
# Calculate percentages by embarkation point
EmbarkedPercentages <- TitanicData %>%
  filter(!is.na(embarked)) %>%
  group_by(embarked) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)
# Custom colours
CherbourgColour <- "#D62828"
QueenstownColour <- "#F77F00"
SouthamptonColour <- "#EAE2B7"
# Create a bar plot with percentages and counts
ggplot(EmbarkedPercentages, aes(
  x = embarked,
  y = count,
  fill = embarked
)) +
  geom_col(
    position = "dodge",
    fill = c(CherbourgColour, QueenstownColour,
SouthamptonColour)
  ) +
  geom_text(
    aes(
      label = paste0(count, " (", round(percentage, 1),
"%)")
    ),
    position = position_dodge(width = 0.8),
    vjust = -0.5,
    size = 3
  ) +
  labs(
    title = "Passenger Distribution by Embarkation
Point"
  ) +
  scale_x_discrete(labels = c("Cherbourg"
                            , "Queenstown"
                            , "Southampton"))
#---End passenger Count, Percentage by Embarkation Point
#-------The survival rate by class of ticket------
#9. What was the survival rate by class of ticket?
# Calculate survival percentages by ticket class
SurvivalByClass <- TitanicData %>%
  group_by(pclass, survived) %>%
  summarise(count = n()) %>%
  group_by(pclass) %>%
  mutate(percentage = count / sum(count) * 100)
# Create a bar chart with percentages
ggplot(SurvivalByClass, aes(
  x = factor(pclass, labels = c("1st Class", "2nd
Class", "3rd Class")),
  y = percentage,
  fill = factor(survived)
)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(
    aes(label = paste0(round(percentage, 1), "%")),
    position = position_dodge(width = 0.8),
    vjust = -0.5,
    size = 3
  ) +
  labs(
    title = "Survival Rate by Passenger Ticket Class",
    x = "Passenger Ticket Class",
    y = "Survival Percentage",
    fill = "Survived"
  ) +
  scale_fill_manual(
    values = c("#D62828", "#F77F00"),
    name = "Survived",
    labels = c("Perished", "Survived")
  ) +
  theme_bw()
#--------End the survival rate by class of ticket----

#--------Survival rates by fare and age-------------
#10. What was the survival rates by fare and age?
# Custom colours
SurvivedColours <- c("#D62828", "#003049")  # Adjust as needed
# Create a scatter plot with custom colours
ggplot(
  TitanicData,
  aes(
    x = age,
    y = fare,
    color = factor(survived),
    size = fare
  )
) +
  geom_point(alpha = 0.7) +
  scale_color_manual(values = SurvivedColours) +  # Set custom colours
  labs(
    title = "Scatter Plot of Fare vs. Age by Survival Status",
    x = "Passenger Age",
    y = "Ticket Fare",
    color = "Survival Status",
    size = "Ticket Fare"
  ) +
  theme_minimal()
#-------End survival rates by fare and age------

#---Multi-variate analysis pclass – sex – survived---
#11. Multi-variate analysis pclass – sex – survived
# Create a summary using tidyverse
ResultSummary <- TitanicData %>%
  group_by(pclass, sex, survived) %>%
  summarise(Count = n()) %>%
  pivot_wider(names_from = survived, values_from = Count, names_prefix = "Survived_")

# Display the result
print(ResultSummary)

#------------End Multi-variate analysis------------
# Export data frame to a CSV file
write.csv(TitanicData, "Train-Titanic.csv", row.names = FALSE)
```

References:

Tufte, E. R. 1983. The Visual Display of Quantitative Information. Graphics Press.

Tukey, J. W. (1977) 'Exploratory Data Analysis', Reading, MA: Addison-Wesley.

Lord, W. (1955) 'The Sinking of the Titanic', New York: Holt, Rinehart and Winston.

Grolemund, G., & Wickham, H. (2016) R for Data Science. Sebastopol, CA: O'Reilly Media.

Davies, M.H. and Beveridge, B. (2012) The Titanic Passenger List.

Eaton, J.P., & Haas, C.A. (1995) Titanic: A Journey into History.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer.

Boyer, P.S. (1998) Encyclopedia Titanica: The Comprehensive Reference to the History of the RMS Titanic.

Hart, E. (2012) The Titanic Disaster: The Full Story of the Sinking of the Unsinkable Ship.

Wilkinson, L. 2005. The Grammar of Graphics. Springer.

Heer, J., & Shneiderman, B. 2012. Interactive Dynamics for Visual Analysis. ACM Queue, 10(2), 30–53.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010) Multivariate Data Analysis (7th ed.). Pearson Prentice Hall.

Rubin, D. B. (1987). Multiple imputation for missing data. Journal of the American Statistical Association, 81(403), 267-280.

Kay, M., Kola, T., Hullman, J., & Munson, S. A. 2016. When (ish) is My Bus? User-Centric Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '16), 5092–5103.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley.

Segel, E., & Heer, J. 2010. Narrative Visualization: Telling Stories with Data. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1139–1148.

Ware, C. 2012. Information Visualization: Perception for Design. Morgan Kaufmann.