Titanic EDA:Visualization and Statistical Exploration

Earnest Ouma Andele

2025-07-29

Table of Contents

[1 📝 Overview. 1](#_Toc17796)

[2 📥 Loading the Necessary Libraries. 2](#_Toc16497)

[3 📥 Loading and inspecting the data. 2](#_Toc16825)

[3.1 Load the dataset. 2](#_Toc30184)

[3.2 Preview structure. 2](#_Toc18102)

[3.3 Summary Statistics. 2](#_Toc3048)

[4 🧩 Missing Values Analysis 3](#_Toc22128)

[4.1 Visualize missing values 3](#_Toc9648)

[5 📊 Grouped Bar Plots. 5](#_Toc29657)

[5.1 Survival by Passenger Class and Gender 5](#_Toc15440)

[6 🌐 Scatterplots. 5](#_Toc22022)

[6.1 Age vs Fare colored by Survival. 5](#_Toc19400)

[7 🔍 Correlation Matrix (Numerical Variables Only) 6](#_Toc3396)

[7.1 Prepare and Visualize Correlation. 6](#_Toc24118)

[8 📌 Conclusion 7](#_Toc29217)

# 1 📝 Overview.

This project analyzes the Titanic dataset, using visualizations and statistics to reveal survival patterns linked to *class*, *gender*, *age* and *fare*.

# 2 📥 Loading the Necessary Libraries.

library(tidyverse)  
library(ggplot2)  
library(readr)  
library(dplyr)  
library(corrplot)  
library(ggthemes)  
library(VIM)  
library(RColorBrewer)

# 3 📥 Loading and inspecting the data.

## 3.1 Load the dataset.

titanic\_data <- read.csv("C:/Users/Earnest/Desktop/School of Statistician/week9\_titanic\_eda\_data.csv")

## 3.2 Preview structure.

glimpse(titanic\_data)

## Rows: 5,000  
## Columns: 12  
## $ PassengerID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,…  
## $ Survived <int> 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0…  
## $ Pclass <int> 3, 3, 3, 3, 3, 1, 3, 3, 1, 1, 1, 1, 1, 1, 3, 3, 2, 3, 3, 1…  
## $ Name <chr> "Name\_1", "Name\_2", "Name\_3", "Name\_4", "Name\_5", "Name\_6"…  
## $ Sex <chr> "female", "male", "female", "female", "female", "female", …  
## $ Age <dbl> 35.6, 19.5, 26.1, 10.4, 28.5, 45.6, 39.6, 49.6, 25.8, 48.8…  
## $ SibSp <int> 2, 3, 3, 3, 1, 0, 0, 2, 3, 3, 4, 5, 1, 1, 2, 0, 3, 5, 2, 1…  
## $ Parch <int> 1, 3, 4, 0, 4, 4, 1, 3, 4, 4, 2, 0, 3, 0, 1, 4, 4, 5, 0, 1…  
## $ Ticket <int> 65207, 67643, 50593, 57762, 26461, 41574, 62842, 68180, 80…  
## $ Fare <dbl> 71.42, 57.08, 152.16, 124.40, 117.36, 38.37, 43.43, 63.54,…  
## $ Cabin <chr> "C123", "D33", "", "G6", "D33", "G6", "E44", "G6", "C123",…  
## $ Embarked <chr> "C", "Q", "Q", "Q", "C", "C", "S", "Q", "Q", "C", "C", "Q"…

## 3.3 Summary Statistics.

summary(titanic\_data)

## PassengerID Survived Pclass Name   
## Min. : 1 Min. :0.0000 Min. :1.000 Length:5000   
## 1st Qu.:1251 1st Qu.:0.0000 1st Qu.:1.000 Class :character   
## Median :2500 Median :0.0000 Median :2.000 Mode :character   
## Mean :2500 Mean :0.4992 Mean :2.001   
## 3rd Qu.:3750 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :5000 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:5000 Min. :-24.90 Min. :0.000 Min. :0.000   
## Class :character 1st Qu.: 20.57 1st Qu.:1.000 1st Qu.:1.000   
## Mode :character Median : 29.80 Median :2.000 Median :3.000   
## Mean : 29.94 Mean :2.462 Mean :2.564   
## 3rd Qu.: 39.40 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. : 79.40 Max. :5.000 Max. :5.000   
## NA's :300   
## Ticket Fare Cabin Embarked   
## Min. :10021 Min. : 10.04 Length:5000 Length:5000   
## 1st Qu.:32547 1st Qu.: 58.34 Class :character Class :character   
## Median :54537 Median :105.56 Mode :character Mode :character   
## Mean :54803 Mean :105.85   
## 3rd Qu.:77298 3rd Qu.:154.06   
## Max. :99998 Max. :199.98   
##

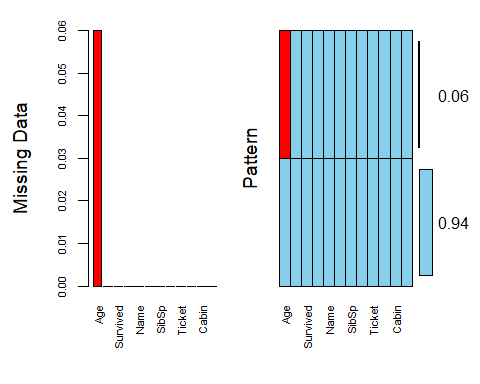
**Explanation**

loaded the Titanic dataset and viewed its structure and summary. This helps in knowing the types of variables(categorical, numerical) and checks for missing values.

# 4 🧩 Missing Values Analysis

## 4.1 Visualize missing values

VIM::aggr(titanic\_data, numbers = TRUE, sortVars = TRUE,   
 labels = names(titanic\_data), cex.axis = 0.7,  
 gap = 3, ylab = c("Missing Data", "Pattern"))



##   
## Variables sorted by number of missings:   
## Variable Count  
## Age 0.06  
## PassengerID 0.00  
## Survived 0.00  
## Pclass 0.00  
## Name 0.00  
## Sex 0.00  
## SibSp 0.00  
## Parch 0.00  
## Ticket 0.00  
## Fare 0.00  
## Cabin 0.00  
## Embarked 0.00

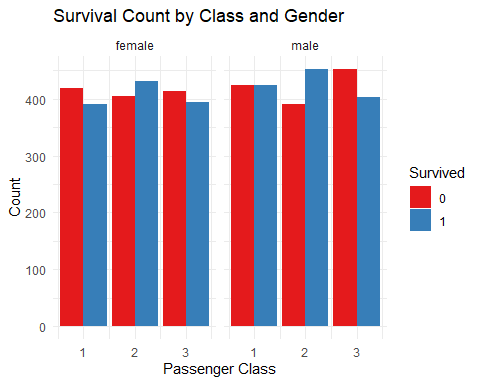
**Explanation:**

This bar chart identifies variables with missing data. It helps us decide whether to impute, remove, or model around them.

# 5 📊 Grouped Bar Plots.

## 5.1 Survival by Passenger Class and Gender

ggplot(titanic\_data, aes(x = Pclass, fill = factor(Survived))) +  
 geom\_bar(position = "dodge") +  
 facet\_wrap(~Sex) +  
 scale\_fill\_brewer(palette = "Set1", name = "Survived") +  
 labs(title = "Survival Count by Class and Gender",  
 x = "Passenger Class",  
 y = "Count") +  
 theme\_minimal()

 **Explanation:**

This grouped bar plot shows how survival varied by *passenger class* and *gender*. We observe that:

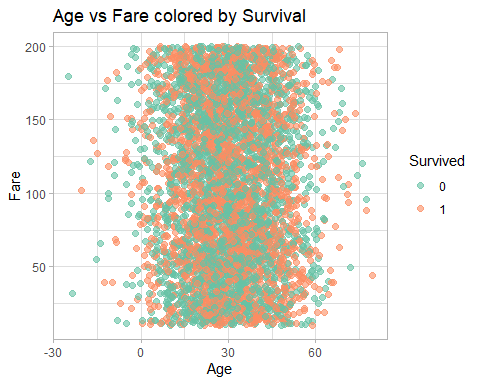
Females had higher survival rates.

Higher classes (1st class) had better survival chances than lower classes.

# 6 🌐 Scatterplots.

## 6.1 Age vs Fare colored by Survival.

ggplot(titanic\_data, aes(x = Age, y = Fare, color = factor(Survived))) +  
 geom\_point(alpha = 0.6, size = 2) +  
 scale\_color\_brewer(palette = "Set2", name = "Survived") +  
 labs(title = "Age vs Fare colored by Survival",  
 x = "Age",  
 y = "Fare") +  
 theme\_light()

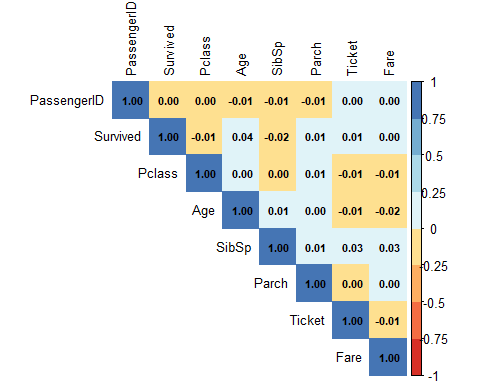
 **Explanation**

This scatterplot reveals cluster of passengers. Those who paid higher fares and were younger had a slightly better chance of survival.

# 7 🔍 Correlation Matrix (Numerical Variables Only)

## 7.1 Prepare and Visualize Correlation.

# Select only numeric variables  
numeric\_data <- titanic\_data %>%  
 select\_if(is.numeric) %>%  
 na.omit()  
  
# Compute correlation matrix  
cor\_matrix <- cor(numeric\_data)  
  
# Plot correlation  
corrplot(cor\_matrix, method = "color", type = "upper",   
 col = brewer.pal(n = 8, name = "RdYlBu"),  
 addCoef.col = "black", tl.col = "black",   
 tl.cex = 0.8, number.cex = 0.7)



**Explation**

This correlation matrix shows the strength and direction of linear relationship between numeric variables:

Positive correlation (blue), negative correlation (red).

Age and Fare show weak correlation.

SibSp and Parch are moderately correlated, as expected.

# 8 📌 Conclusion

This visual EDA enhances our understanding of key survival patterns and relationships in the Titanic dataset. Future steps may include modeling using logistic regression or tree-based methods.

Prepared by **Earnest Ouma Andele**, Week 10 EDA Project, 2025-07-29