# Day 2 – EDA on Titanic Dataset (Case Study Practice)

**Step 1: Load Required Libraries and the Dataset**

# Load required libraries

library(ggplot2) # For data visualization

library(dplyr) # For data manipulation

library(tidyr) # For data tidying

# Load the Titanic dataset (assuming it's in your working directory)

titanic <- read.csv("path…..titanic.csv")

**Step 2: Explore the Dataset**

Let's start by getting a sense of what the dataset looks like:

# Display the first few rows of the dataset

head(titanic)

# Get an overview of the dataset

summary(titanic)

# Check the structure of the dataset (data types)

str(titanic)

# Check for missing values

sapply(titanic, function(x) sum(is.na(x)))

**Step 3: Visualize Data**

EDA often involves creating visualizations to better understand the data. Here are some common visualizations:

# Histogram of passenger ages

ggplot(titanic, aes(x = Age)) +

geom\_histogram(binwidth = 5, fill = "skyblue", color = "black") +

labs(title = "Distribution of Passenger Ages", x = "Age")

# Bar chart of passenger classes

ggplot(titanic, aes(x = factor(Pclass))) +

geom\_bar(fill = "steelblue") +

labs(title = "Passenger Class Distribution", x = "Class")

# Survival rate by gender

ggplot(titanic, aes(x = Sex, fill = factor(Survived))) +

geom\_bar(position = "fill") +

labs(title = "Survival Rate by Gender", x = "Gender", y = "Proportion") +

scale\_fill\_manual(values = c("red", "green"))

# Scatterplot of age vs. fare

ggplot(titanic, aes(x = Age, y = Fare, color = factor(Survived))) +

geom\_point() +

labs(title = "Age vs. Fare by Survival", x = "Age", y = "Fare") +

scale\_color\_manual(values = c("red", "green"))

**Interpretation of Summary:**

1. Passenger Class (Pclass):
   * This section provides information about the passenger class.
   * It tells us the number of passengers in each class (1st, 2nd, and 3rd class).
   * It also provides the percentage of passengers in each class.
2. Sex of Passengers (Sex):
   * This section shows the number of male and female passengers.
   * It also provides the percentage of male and female passengers.
3. Age of Passengers (Age):
   * This section provides summary statistics for passenger ages.
   * It shows the minimum and maximum ages, as well as the median (50th percentile) age.
   * The 1st quartile (25th percentile) and 3rd quartile (75th percentile) are also shown.
   * The number of missing values in the "Age" column is displayed.
4. Number of Siblings/Spouses (SibSp):
   * This section summarizes the number of siblings or spouses each passenger had on board.
   * It shows the minimum and maximum values, as well as the median, 1st quartile, and 3rd quartile.
   * The number of missing values in the "SibSp" column is displayed.
5. Number of Parents/Children (Parch):
   * This section summarizes the number of parents or children each passenger had on board.
   * Similar to the "SibSp" section, it displays minimum, maximum, median, 1st quartile, 3rd quartile, and the count of missing values.
6. Ticket Fare (Fare):
   * This section provides summary statistics for the ticket fares.
   * It displays the minimum, maximum, median, 1st quartile, and 3rd quartile values.
   * The number of missing values in the "Fare" column is shown.
7. Cabin Information (Cabin):
   * This section summarizes information about the cabins.
   * It displays the number of unique cabin values (cabins that passengers were in).
   * Since "Cabin" likely has a lot of missing values, it also shows the number of missing values.
8. Port of Embarkation (Embarked):
   * This section summarizes information about the port of embarkation.
   * It displays the number of passengers who embarked from each port (C = Cherbourg, Q = Queenstown, S = Southampton).
   * It also shows the number of missing values.
9. Survival Information (Survived):
   * This section provides information about whether passengers survived or not.
   * It shows the number of passengers who survived (1) and the number who did not survive (0).
   * It also displays the percentage of survivors among the passengers.

**Interpretation of Histogram:**

1. **ggplot(titanic, aes(x = Age))**: This line sets up the foundation for creating a plot using the **ggplot2** package in R. It specifies that you want to use the "Age" variable from the "titanic" dataset as the x-axis variable.
2. **geom\_histogram(binwidth = 5, fill = "skyblue", color = "black")**: This line adds a histogram layer to the plot. A histogram is a way to visualize the distribution of a continuous variable (in this case, age) by dividing it into bins and counting how many data points fall into each bin.
   * **binwidth = 5** specifies that each bin in the histogram should cover a range of 5 years. This means that ages are grouped into intervals of 5 years (e.g., 0-5, 5-10, 10-15, and so on).
   * **fill = "skyblue"** and **color = "black"** set the color of the bars in the histogram. The bars will be filled with a sky-blue color and outlined in black.
3. **labs(title = "Distribution of Passenger Ages", x = "Age")**: This line adds labels to the plot. The title of the plot is set to "Distribution of Passenger Ages," and the x-axis label is set to "Age."

The histogram shows the distribution of passenger ages on the Titanic. Each bar represents a range of ages, and the height of the bar indicates how many passengers fall within that age range.

* Most of the passengers appear to be between the ages of 20 and 30, as this range has the tallest bar.
* There are relatively fewer passengers who are very young (0-5 years) or very old (65-70 years and above).
* The distribution is somewhat **right skewed**, meaning there are more passengers in the younger age groups than in the older age groups.

This histogram provides a visual summary of the age distribution of Titanic passengers, which can help us understand the demographics of the people on board.

**Interpretation of Bar Chart of Passenger Classes**

The resulting bar chart visually represents the distribution of passenger classes on the Titanic. In this context, passenger class refers to the socio-economic class of the passengers, with three possible values: 1st class, 2nd class, and 3rd class.

* The x-axis of the chart represents the different passenger classes (1st, 2nd, and 3rd class).
* The y-axis represents the count or frequency of passengers in each class.

1. **1st Class**: You will see a bar corresponding to the 1st class. This bar's height represents the number of passengers who were in the 1st class.
2. **2nd Class**: Similarly, there will be another bar for the 2nd class, and its height represents the number of passengers in the 2nd class.
3. **3rd Class**: Lastly, there will be a bar for the 3rd class, showing the number of passengers in the 3rd class.

By looking at the chart, you can quickly assess the distribution of passengers across the different classes. This information could be useful for various analyses and insights, such as understanding the socio-economic composition of the Titanic's passengers or exploring how class might have been related to survival rates during the Titanic disaster.

**Interpretation of Survival Rate by Gender**

The code generates a stacked bar chart with two bars, one for each gender (male and female). Each bar is divided into two segments: one in red and one in green.

* **Red Segment**: The red segment represents the proportion of passengers in each gender who did not survive (i.e., the ones who perished in the Titanic disaster).
* **Green Segment**: The green segment represents the proportion of passengers in each gender who survived the disaster.

By looking at this chart, you can quickly see the survival rate by gender. For example, if the green segment is larger than the red segment for a particular gender, it means that a higher proportion of passengers of that gender survived. Conversely, if the red segment is larger, it means a higher proportion did not survive.

**Interpretation of Scatterplot of age vs. fare**

The scatterplot shows a visual representation of the Titanic dataset. Here's what we can observe from it:

* On the x-axis, we have the passenger's age, which ranges from young to old.
* On the y-axis, we have the fare that each passenger paid for their ticket, which varies widely.
* The points on the scatterplot represent individual passengers. Each point's color indicates whether that passenger survived (green) or did not survive (red).

Now, let's make some observations:

1. **Age vs. Fare**: There is no clear linear relationship between age and fare. Passengers of various ages paid different fares, and there is no strong correlation visible from the scatterplot.
2. **Survival**: The use of red and green colors makes it easy to distinguish between passengers who survived (green) and those who did not (red).
3. **Distribution**: We can see that both survivors and non-survivors are spread across different age groups and fare ranges. This suggests that survival on the Titanic did not appear to be determined solely by age or fare paid.

In summary, this scatterplot provides a visual overview of the Titanic dataset, allowing us to explore the relationship between age, fare, and survival. It suggests that survival was not strongly correlated with either age or fare, and other factors likely played a role in determining who survived the Titanic disaster.