Final Project

AUTHOR
Chelsea Rodrigues

Introduction

This analysis explores the Titanic dataset, aiming to provide insights into passenger demographics and survival rates. The dataset was obtained from GitHub - Awesome Public Datasets

Part 1: Analysis

Load Libraries and Data

```
# Load necessary libraries
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 dataset
titanic <- read.csv("C:/Users/chels/Downloads/R final project/titanic.csv")
```

Data Exploration

Summary Statistics

```
# Display summary statistics
summary(titanic)
```

```
  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                                        
```

The dataset contains information on 891 passengers. The ‘Survived’ variable indicates that around 38.38% of passengers survived. The majority of passengers were in the 2nd and 3rd passenger classes. Age data is available for 714 passengers, with a mean age of approximately 29.7 years

Missing Values

```
# Count missing values
missing_values <- colSums(is.na(titanic))
missing_values
```

```
PassengerId    Survived      Pclass        Name         Sex         Age 
          0           0           0           0           0         177 
      SibSp       Parch      Ticket        Fare       Cabin    Embarked 
          0           0           0           0           0           0 
```

Data Cleaning

Remove NA Rows

```
# Remove rows with missing values
titanic_clean <- na.omit(titanic)
head(titanic)
```

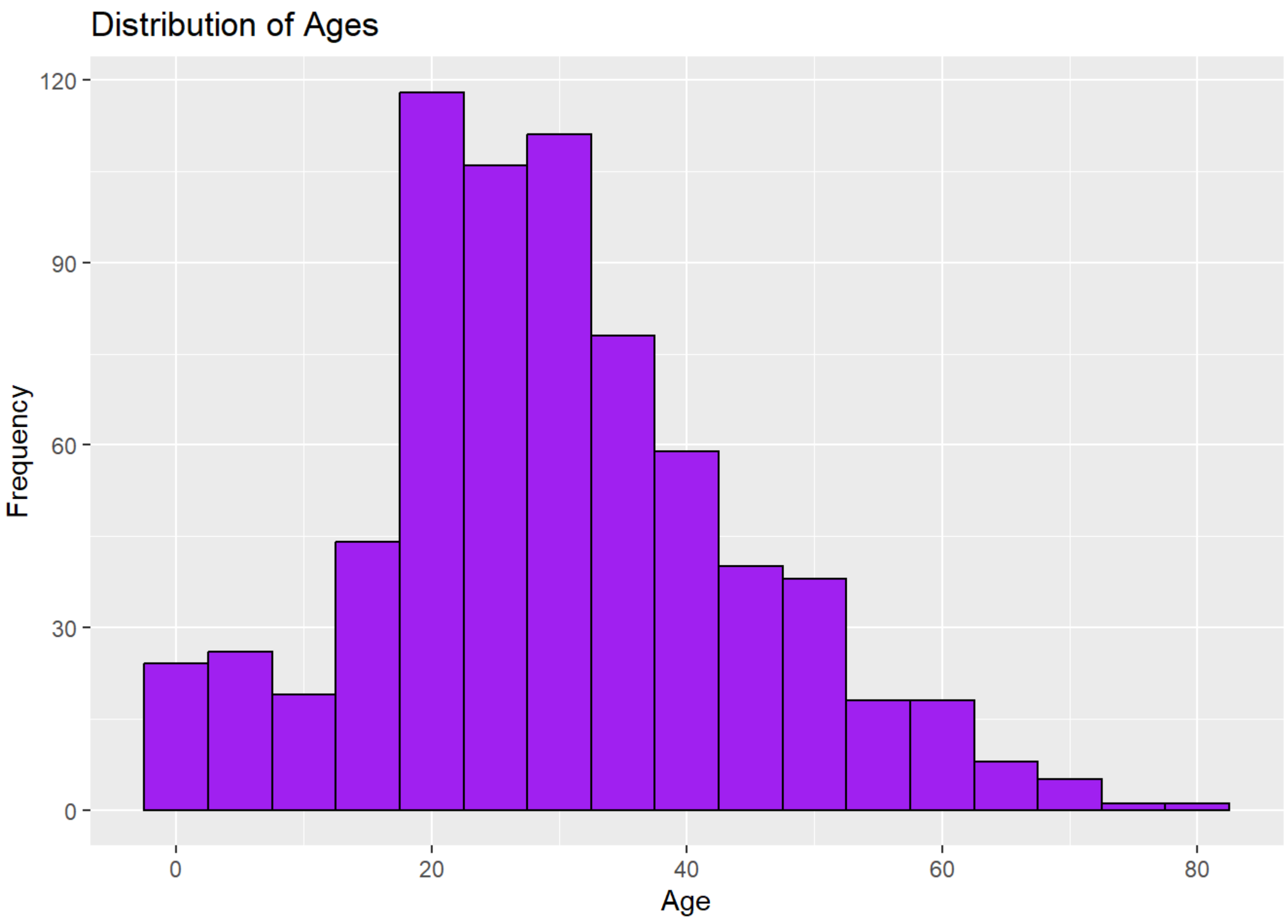
```
  PassengerId Survived Pclass
1           1        0      3
2           2        1      1
3           3        1      3
4           4        1      1
5           5        0      3
6           6        0      3
                                                 Name    Sex Age SibSp Parch
```

```
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
```

Exploratory Data Analysis

Passenger Age Distribution

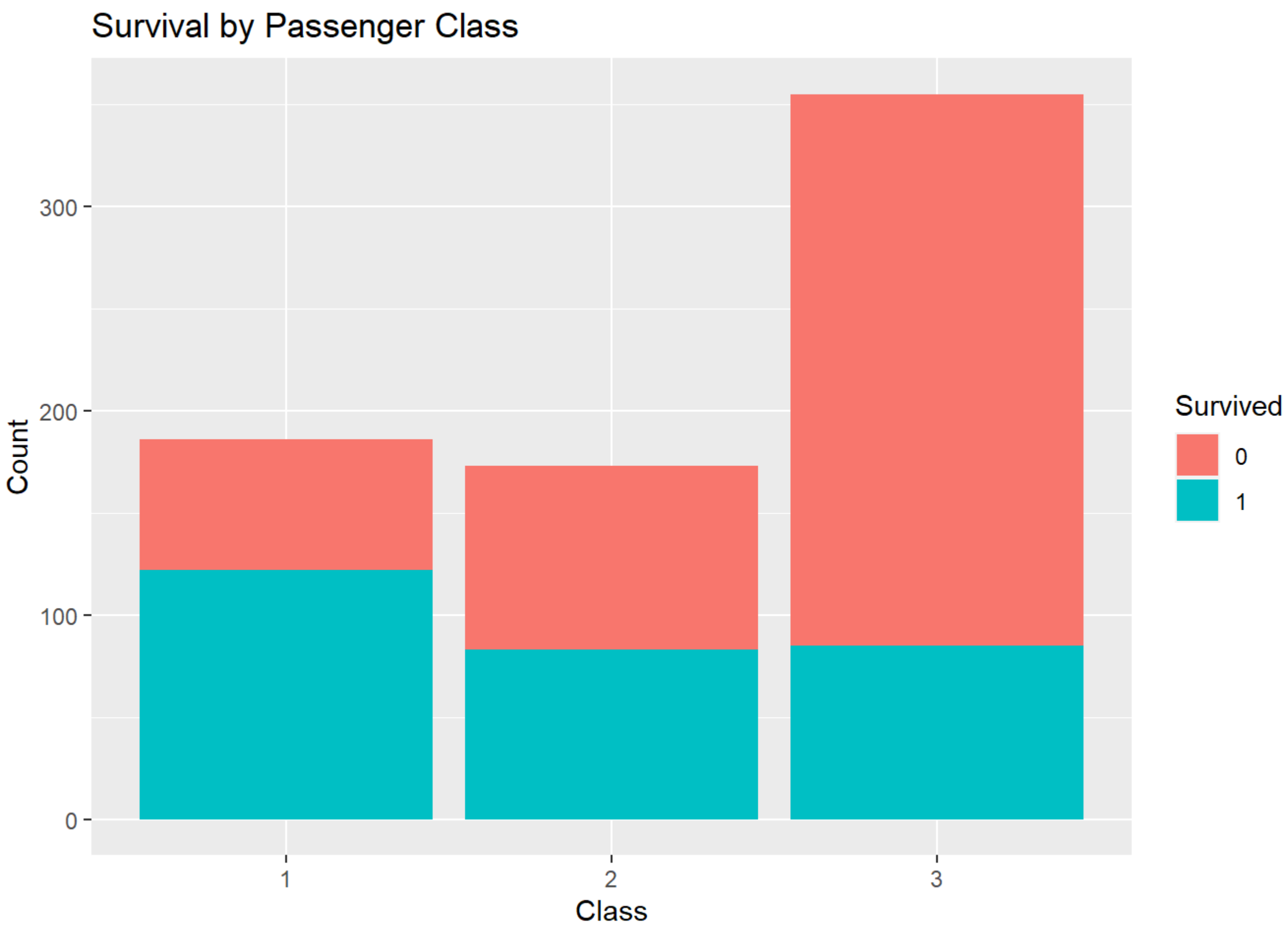
```
# Plot histogram of ages
ggplot(titanic_clean, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "black") +
  labs(title = "Distribution of Ages", x = "Age", y = "Frequency")
```



The histogram of ages illustrates a diverse age distribution among passengers, with a peak in the early 20s.

Survival by Passenger Class

```
# Plot survival by passenger class
ggplot(titanic_clean, aes(x = factor(Pclass), fill = factor(Survived))) +
  geom_bar(position = "stack") +
  labs(title = "Survival by Passenger Class", x = "Class", y = "Count", fill = "Survived")
```



The survival bar chart by passenger class indicates that a higher percentage of 1st-class passengers survived compared to those in the 2nd and 3rd classes.

Conclusion part 1

In conclusion, this preliminary analysis provides valuable insights into the Titanic dataset. Further investigations could include detailed demographic analyses and survival predictions based on various factors.

Part 2: R Package - Tidyverse

Introduction

In this section, we will explore the `tidyverse` package, which is a collection of packages for data manipulation and visualization. The purpose of this demonstration is to showcase some of the key functionality provided by the `tidyverse`.

```r
# Load tidyverse
library(tidyverse)
```

```
Warning: package 'tibble' was built under R version 4.3.2

Warning: package 'purrr' was built under R version 4.3.2

Warning: package 'lubridate' was built under R version 4.3.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ lubridate 1.9.3     ✔ tibble    3.2.1
✔ purrr     1.0.2     ✔ tidyr     1.3.0
✔ readr     2.1.4     
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

Data Preparation

For this demonstration, we will use the Titanic dataset introduced in Part 1.

Demonstrating Functionality

1. Data Manipulation with dplyr

The dplyr package within tidyverse provides a set of functions for data manipulation. Let’s use it to filter passengers who survived and calculate the average age.

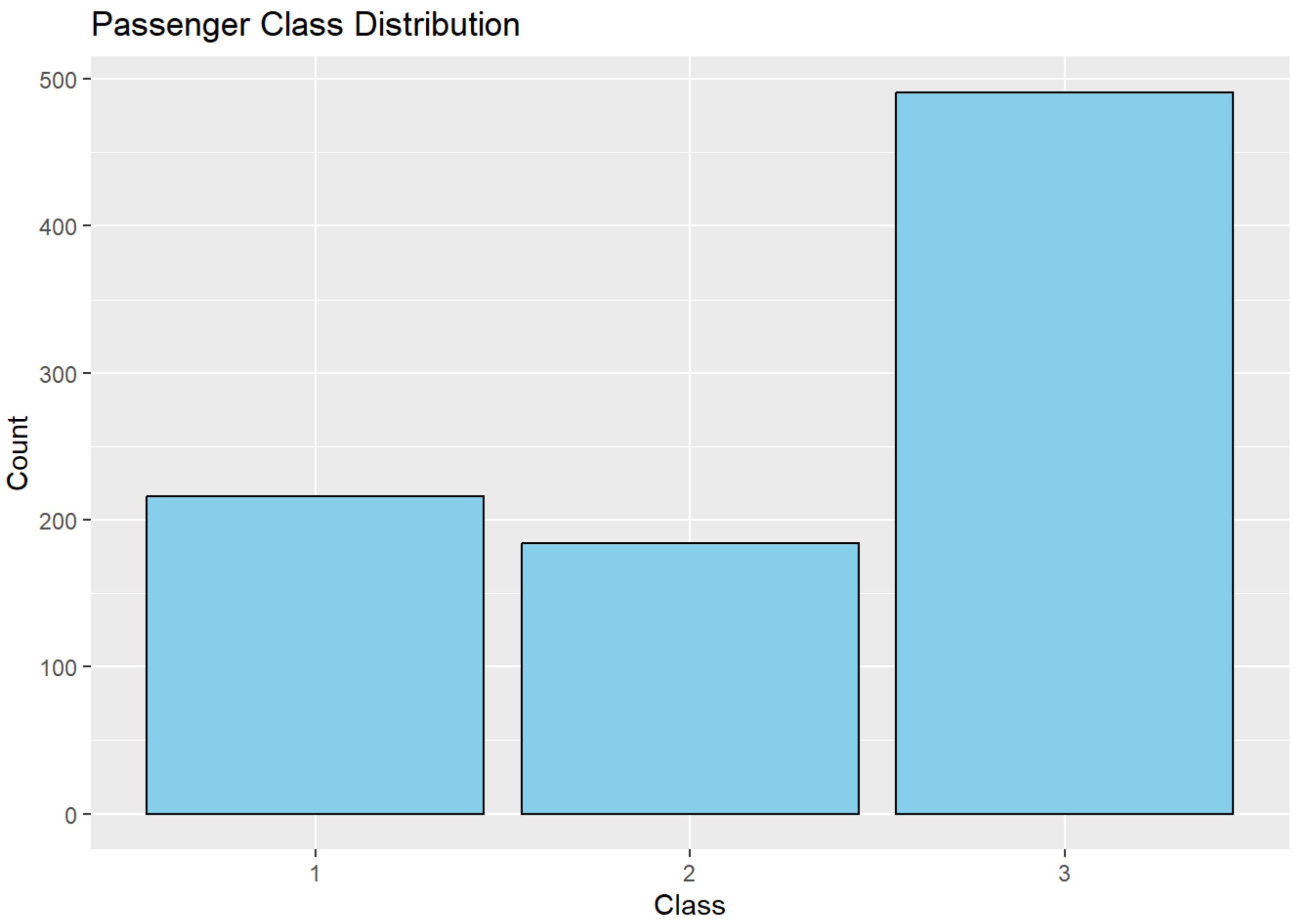
```r
# Filter survivors and calculate average age
survivor_stats <- titanic %>%
  filter(Survived == 1) %>%
  summarise(Avg_Age = mean(Age, na.rm = TRUE))

print(survivor_stats)
```

```
   Avg_Age
1 28.34369
```

2. Data Visualization with ggplot2

The ggplot2 package in tidyverse is a powerful tool for creating visualizations. Let’s use it to create a bar chart of passenger class distribution.

```r
# Plot passenger class distribution
ggplot(titanic, aes(x = factor(Pclass))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Passenger Class Distribution", x = "Class", y = "Count")
```



3. Data Wrangling with tidyr

The tidyr package within tidyverse is used for data reshaping. Let’s use it to gather and spread data.

```
# Gather and spread data
gathered_data <- titanic %>%
  gather(key = "Variable", value = "Value", -PassengerId, -Name)

spread_data <- gathered_data %>%
  spread(key = "Variable", value = "Value")

head(gathered_data)
```

```
  PassengerId                                                Name Variable
1           1                             Braund, Mr. Owen Harris Survived
2           2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) Survived
3           3                              Heikkinen, Miss. Laina Survived
4           4        Futrelle, Mrs. Jacques Heath (Lily May Peel) Survived
5           5                            Allen, Mr. William Henry Survived
6           6                                    Moran, Mr. James Survived
  Value
1     0
```

```
2     1
3     1
4     1
5     0
6     0
```

```
head(spread_data)
```

```
  PassengerId                                                Name  Age Cabin
1           1                             Braund, Mr. Owen Harris   22      
2           2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)   38   C85
3           3                              Heikkinen, Miss. Laina   26      
4           4        Futrelle, Mrs. Jacques Heath (Lily May Peel)   35  C123
5           5                            Allen, Mr. William Henry   35      
6           6                                    Moran, Mr. James <NA>      
  Embarked    Fare Parch Pclass    Sex SibSp Survived           Ticket
1        S    7.25     0      3   male     1        0        A/5 21171
2        C 71.2833     0      1 female     1        1         PC 17599
3        S   7.925     0      3 female     0        1 STON/O2. 3101282
4        S    53.1     0      1 female     1        1           113803
5        S    8.05     0      3   male     0        0           373450
6        Q  8.4583     0      3   male     0        0           330877
```

Conclusion Part 2

The tidyverse package provides a comprehensive set of tools for data manipulation and visualization in R. This demonstration covered only a small portion of its functionality. For more details and advanced usage, refer to the package documentation.

Part 3: Functions/Programming - Age-Fare Correlation

Introduction

In this section, we will create an R function to analyze the correlation between passenger age and fare in the Titanic dataset. The function will output an S3 class object named `AgeFareCorrelation`. We will implement appropriate print, summary, and plot methods for this class.

Function Definition

```
# Function to calculate age-fare correlation
calculate_age_fare_correlation <- function(data) {
  # Filter non-missing age and fare values
  filtered_data <- data %>%
    filter(!is.na(Age) & !is.na(Fare))

  # Calculate correlation
  correlation <- cor(filtered_data$Age, filtered_data$Fare)
  # Create AgeFareCorrelation object
  result <- list(
    correlation = correlation,
    data = filtered_data
  )

  class(result) <- "AgeFareCorrelation"

  return(result)
}
```

S3 Class Definition

```
# Define S3 class
AgeFareCorrelation <- function(correlation, data) {
  obj <- list(
    correlation = correlation,
    data = data
  )
  class(obj) <- "AgeFareCorrelation"
  return(obj)
}
```

Summary Method

```
# Define summary method for AgeFareCorrelation
summary.AgeFareCorrelation <- function(object, ...) {
  cat("Age-Fare Correlation:", object$correlation, "\n")
}
```

Print Method

```
# Define print method for AgeFareCorrelation
print.AgeFareCorrelation <- function(object, ...) {
  cat("Age-Fare Correlation Analysis\n")
  cat("-----------------------------\n")
  cat("Correlation:", object$correlation, "\n")
  cat("-----------------------------\n")
}
```

Plot Method

```
# Define plot method for AgeFareCorrelation
plot.AgeFareCorrelation <- function(object, ...) {
  # Scatterplot of Age vs Fare
  ggplot(object$data, aes(x = Age, y = Fare)) +
    geom_point(color = "pink") +
    labs(title = "Scatterplot of Age vs Fare", x = "Age", y = "Fare")
}
```

Example Usage

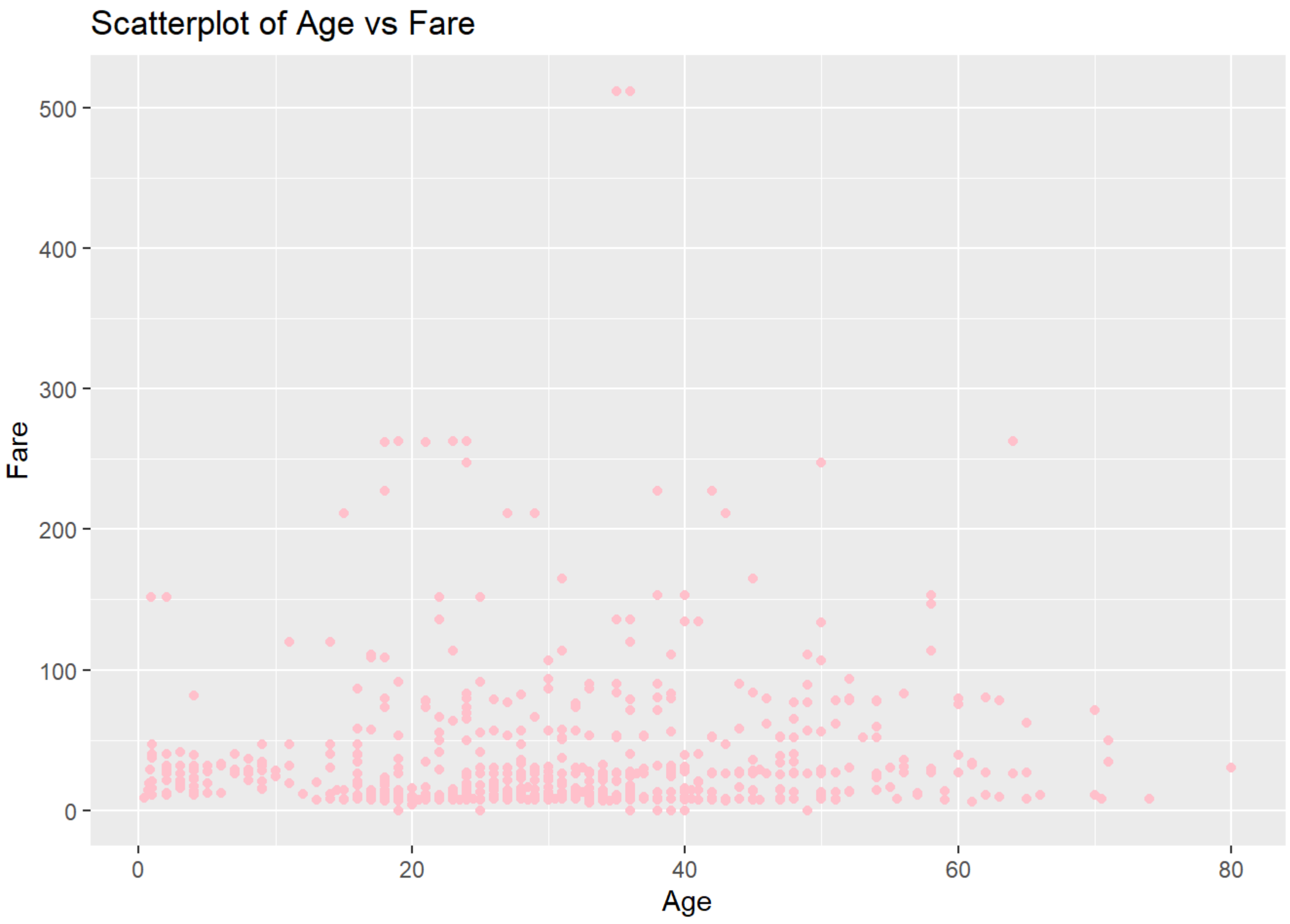
```
#Usage of summary, rpint and plot function
result <- calculate_age_fare_correlation(titanic)
summary(result)
```

```
Age-Fare Correlation: 0.09606669
```

```
print(result)
```

```
Age-Fare Correlation Analysis
----------------------------
Correlation: 0.09606669
----------------------------
```

```
plot(result)
```



Reference

awesomedata (2014). awesome-public-datasets/Datasets/titanic.csv.zip at master · awesomedata/awesome-public-datasets. [online] GitHub. Available at: https://github.com/awesomedata/awesome-public-datasets/blob/master/Datasets/titanic.csv.zip