**DOCUMENTATION FOR TASK 2**

TASK 2: Perform data cleaning and exploratory data analysis (EDA) on titanic dataset from kaggle. explore the relationships between variables and identify patterns and trends in the data by using r language.

**#### Introduction->**

**TITANIC DATASET:** The Titanic dataset is a famous dataset often used for teaching and practicing data analysis and machine learning techniques. It contains information about the passengers aboard the RMS Titanic, which sank on its maiden voyage in April 1912 after hitting an iceberg. The dataset is often used to predict survival based on various passenger characteristics.

Here are the key variables typically included in the Titanic dataset:

1. **Survived**: Whether the passenger survived (0 = No, 1 = Yes)
2. **Pclass**: Ticket class (1st, 2nd, 3rd)
3. **Name**: Passenger's name
4. **Sex**: Passenger's sex (male or female)
5. **Age**: Passenger's age
6. **SibSp**: Number of siblings/spouses aboard
7. **Parch**: Number of parents/children aboard
8. **Ticket**: Ticket number
9. **Fare**: Fare paid for the ticket
10. **Cabin**: Cabin number
11. **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

**#### Data Cleaning:** Data cleaning, also known as data cleansing or data preprocessing, is the process of identifying and correcting errors, inconsistencies, and missing values in a dataset to improve its quality and prepare it for analysis. It is a crucial step in the data analysis pipeline because raw data often contains imperfections that can lead to inaccurate results or biased interpretations if not properly addressed.

**Key Tasks in Data Cleaning:**

1. **Handling Missing Data**:
   * Identifying missing values (e.g., using functions like is.na() in R).
   * Imputing missing values using techniques such as mean, median, mode imputation, or more sophisticated methods like predictive modeling.
2. **Dealing with Outliers**:
   * Identifying outliers using statistical methods (e.g., Z-score, IQR).
   * Deciding whether to remove outliers or transform them depending on the context of the analysis.
3. **Fixing Incorrect Data Types**:
   * Ensuring variables have the correct data type (e.g., numeric, character, factor).
   * Converting data types to match their intended use (e.g., converting strings representing numbers to numeric).
4. **Standardizing and Normalizing Data**:
   * Bringing data into a common format or scale.
   * Standardizing units of measurement or normalizing distributions to make comparisons meaningful.
5. **Removing Duplicates**:
   * Identifying and removing duplicate rows or entries that skew analysis results.
6. **Handling Inconsistencies**:
   * Resolving inconsistencies in categorical data (e.g., different spellings or formats of the same category).
   * Standardizing formats (e.g., date formats, address formats) to ensure consistency across the dataset.
7. **Feature Engineering**:
   * Creating new variables (features) from existing data to improve model performance.
   * Aggregating or transforming variables to extract more meaningful information.

**#### Exploratory Data Analysis (EDA):**Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often using visual methods. The primary goal of EDA is to understand the data, uncover patterns, relationships, anomalies, and insights that may be hidden within the dataset. It typically involves generating simple summaries of the data and visualizing it in various forms.

**#### Findings and Insights:**

- Survival rates varied significantly based on passenger class, with higher classes having higher survival rates.

- Females had a much higher survival rate compared to males, suggesting adherence to the "women and children first" policy.

- Age played a role in survival, with younger passengers having higher survival rates.

#### Report Insights

1. **Insight 1: Socioeconomic Disparities**:
   * The Titanic dataset highlights socioeconomic disparities in survival rates, with passengers in higher classes having a distinct survival advantage. This finding underscores the importance of socioeconomic status in disaster response and survival outcomes.
2. **Insight 2: Gender Disparity**:
   * The data strongly supports the notion that gender played a significant role in survival, with females having a notably higher survival rate than males. This insight reflects historical maritime evacuation protocols.
3. **Insight 3: Age as a Factor**:

#### ###Identify Trends:

1. **Survival Rates by Passenger Class (Pclass)**:
   * There is a clear trend where passengers in higher classes (1st class) had higher survival rates compared to those in lower classes (3rd class). This suggests a possible correlation between socioeconomic status and survival.
2. **Survival Rates by Sex (Sex)**:
   * Females consistently had higher survival rates compared to males. This trend is evident across all passenger classes and ages, indicating that the "women and children first" policy may have influenced survival outcomes.
3. **Impact of Age (Age)**:
   * There appears to be a higher survival rate among children and younger passengers compared to older adults. This trend might reflect efforts to prioritize the evacuation of younger individuals during the disaster.
4. **Embarked Port (Embarked)**:
   * Survival rates vary slightly based on the port of embarkation (Embarked), with passengers embarking from Cherbourg (C) showing slightly higher survival rates compared to those embarking from Southampton (S) and Queenstown (Q).

**###Conclusion:**

Drawing conclusions from the analysis of the Titanic dataset reveals clear patterns related to socioeconomic status, gender, and age influencing survival outcomes. These insights not only provide historical context but also underscore the complex dynamics of disaster response and emergency evacuation protocols.

By systematically analyzing and interpreting the data, you can formulate hypotheses that align with observed trends and report valuable insights into the factors that influenced survival aboard the Titanic. These conclusions serve as a foundation for further research or discussions on disaster preparedness and response strategies.

**Step 1: Load the Data**

Load the Titanic dataset into R. Make sure to set the correct working directory or provide the full path to the dataset file.

r

Copy code

# Load the dataset

df <- read.csv('titanic.csv', stringsAsFactors = FALSE)

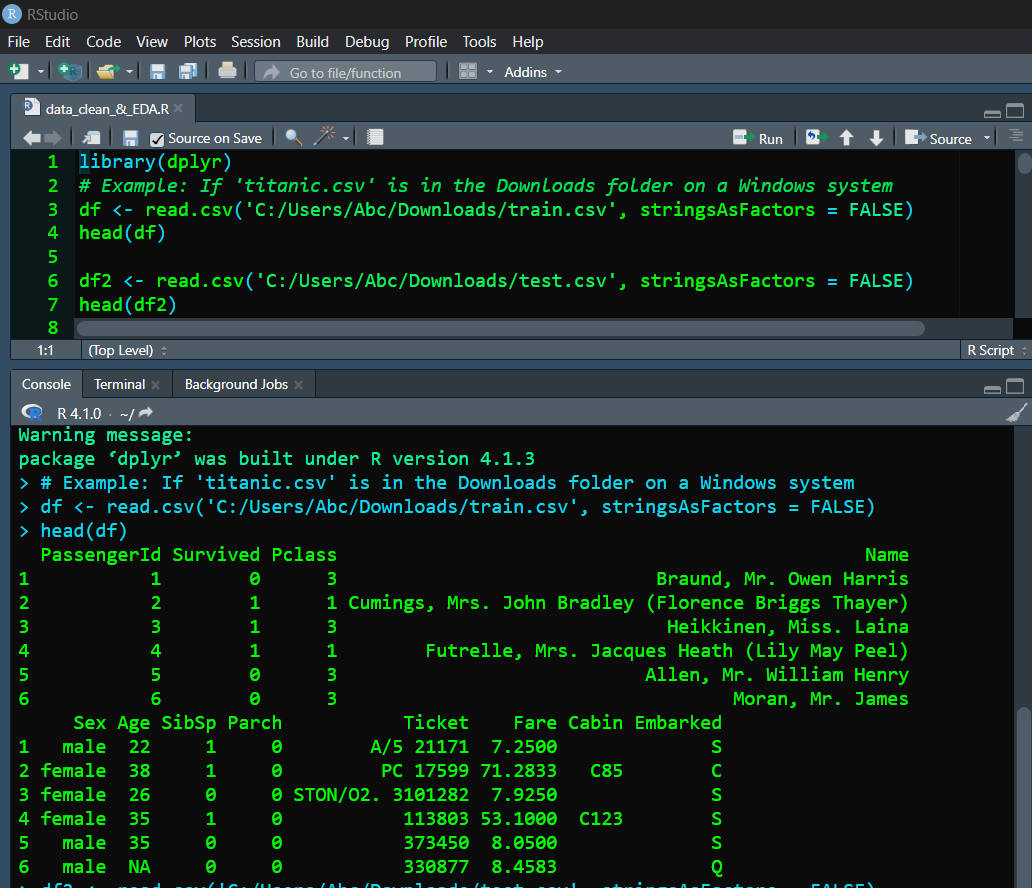
**Step 2: Explore the Data**

* **Preview the dataset**: View the first few rows to understand its structure.

r

Copy code

head(df)



* **Understand the features**: Check the names of columns (features).

r

Copy code

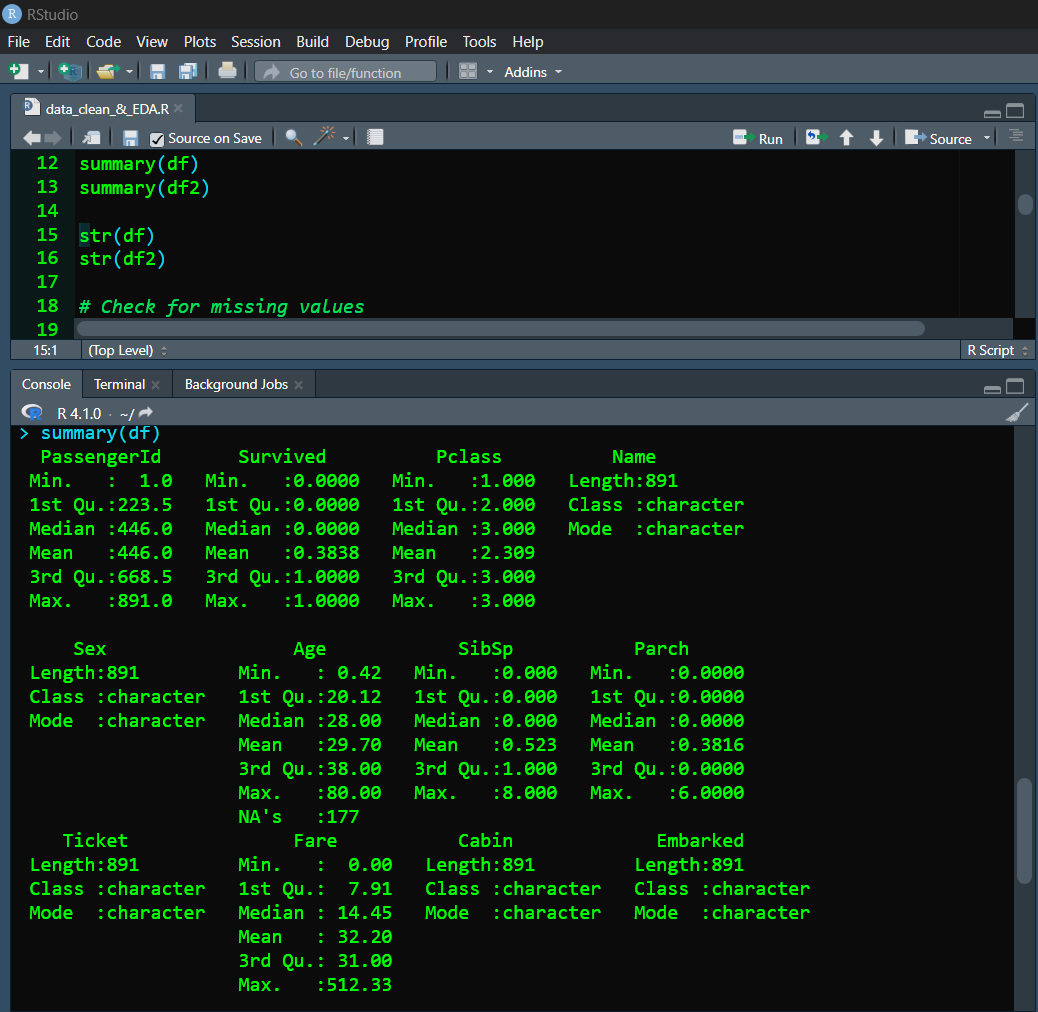
names(df)

* **Summary statistics**: Obtain summary statistics for numerical columns.

r

Copy code

summary(df)



* **Data types and missing values**: Check for missing values and data types of each column.

r

Copy code

str(df)

**Step 3: Data Cleaning**

* **Handle missing values**: Decide how to handle missing values in different columns.

r

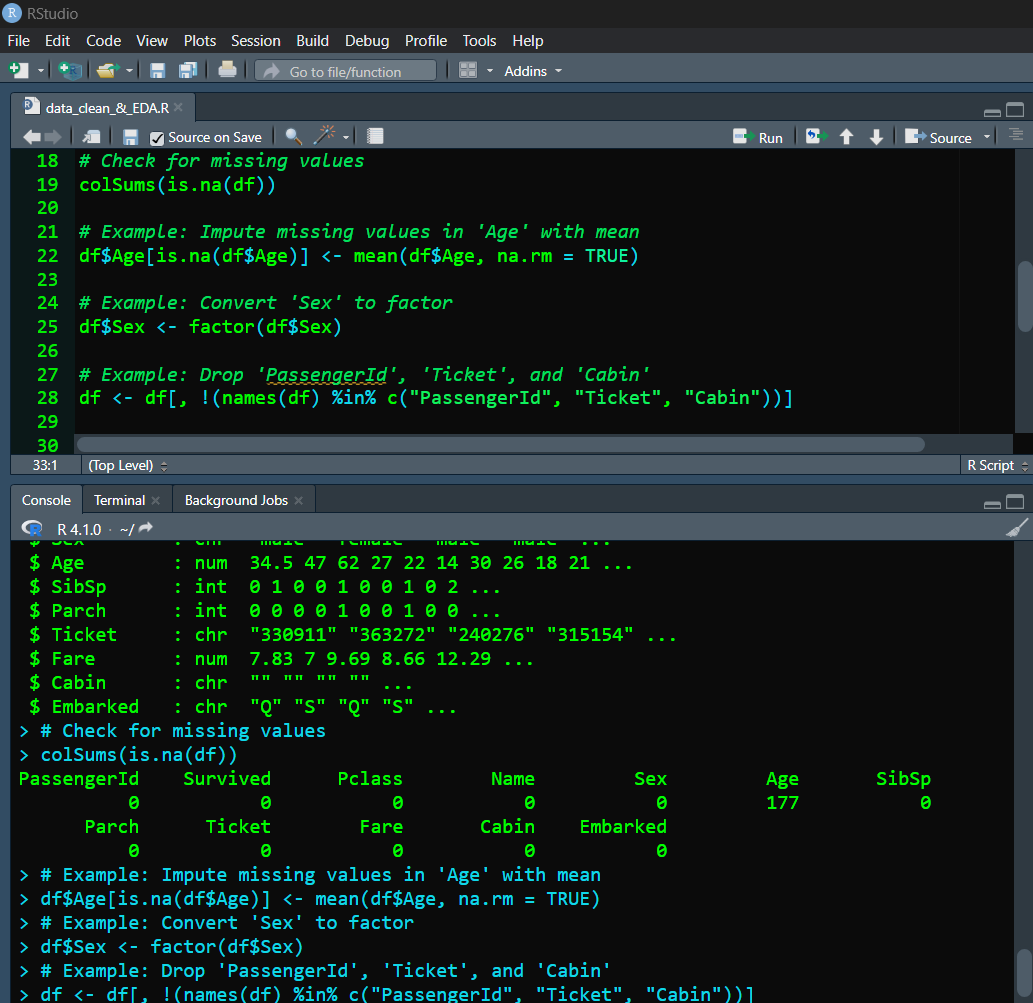
Copy code

# Check for missing values

colSums(is.na(df))

# Example: Impute missing values in 'Age' with mean

df$Age[is.na(df$Age)] <- mean(df$Age, na.rm = TRUE)



* **Correct data formats**: Convert data types if necessary.

r

Copy code

# Example: Convert 'Sex' to factor

df$Sex <- factor(df$Sex)

* **Remove irrelevant columns**: Drop columns that are not useful for analysis.

r

Copy code

# Example: Drop 'PassengerId', 'Ticket', and 'Cabin'

df <- df[, !(names(df) %in% c("PassengerId", "Ticket", "Cabin"))]

**Step 4: Exploratory Data Analysis (EDA)**

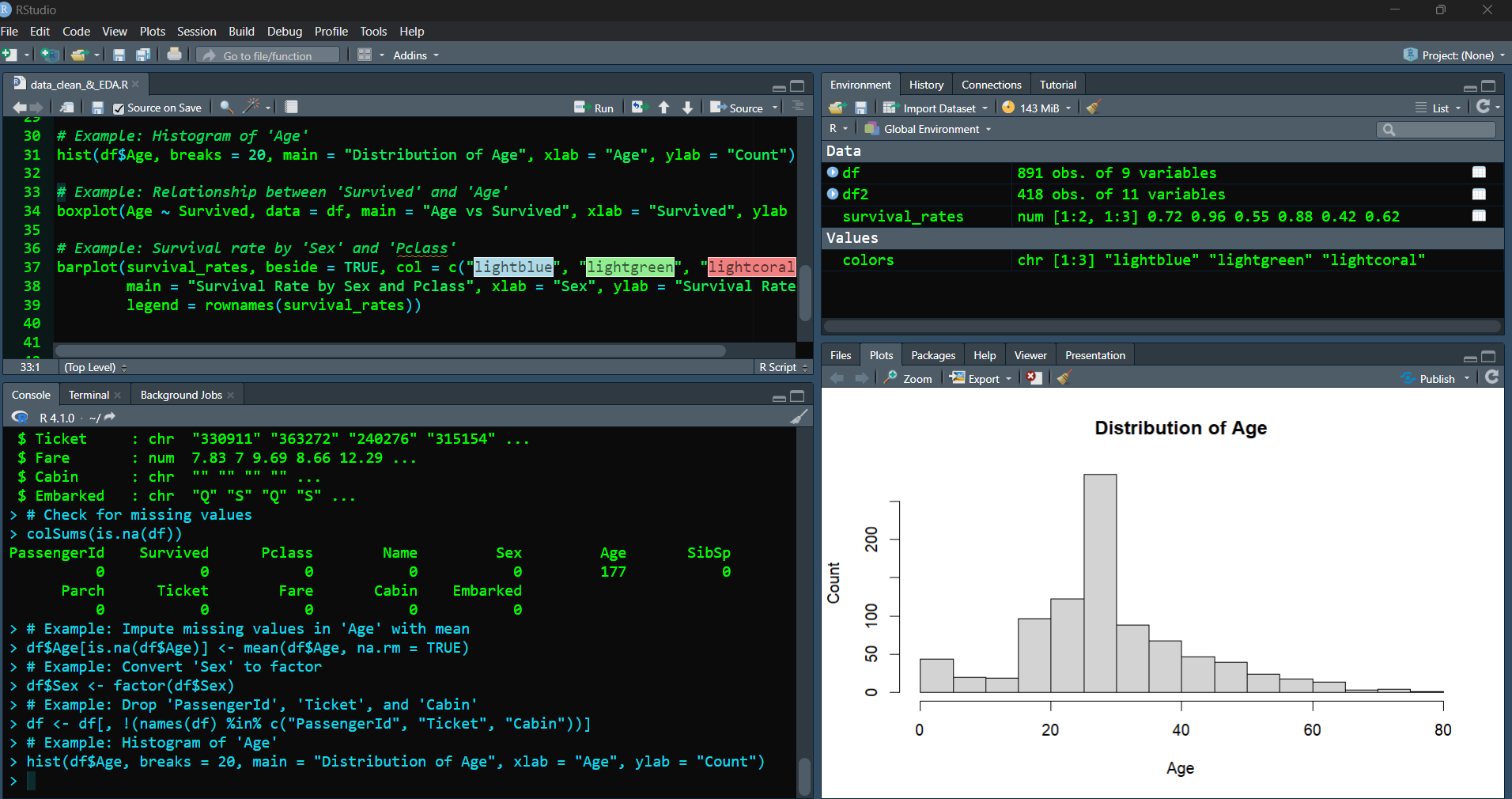
* **Univariate analysis**: Analyze individual variables.

r

Copy code

# Example: Histogram of 'Age'

hist(df$Age, breaks = 20, main = "Distribution of Age", xlab = "Age", ylab = "Count")



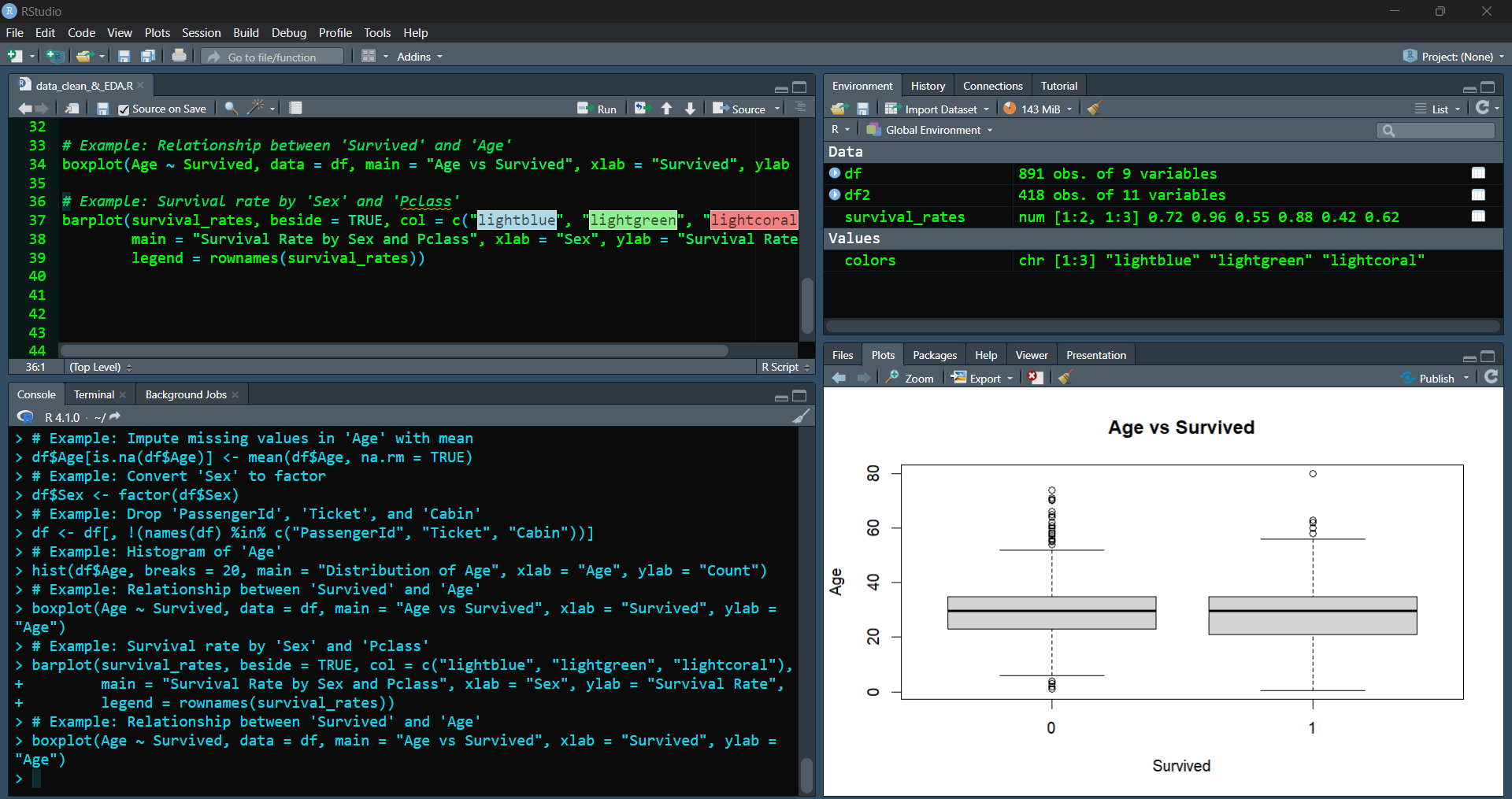
* **Bivariate analysis**: Explore relationships between pairs of variables.

r

Copy code

# Example: Relationship between 'Survived' and 'Age'

boxplot(Age ~ Survived, data = df, main = "Age vs Survived", xlab = "Survived", ylab = "Age")



* **Multivariate analysis**: Look at interactions between multiple variables.

r

Copy code

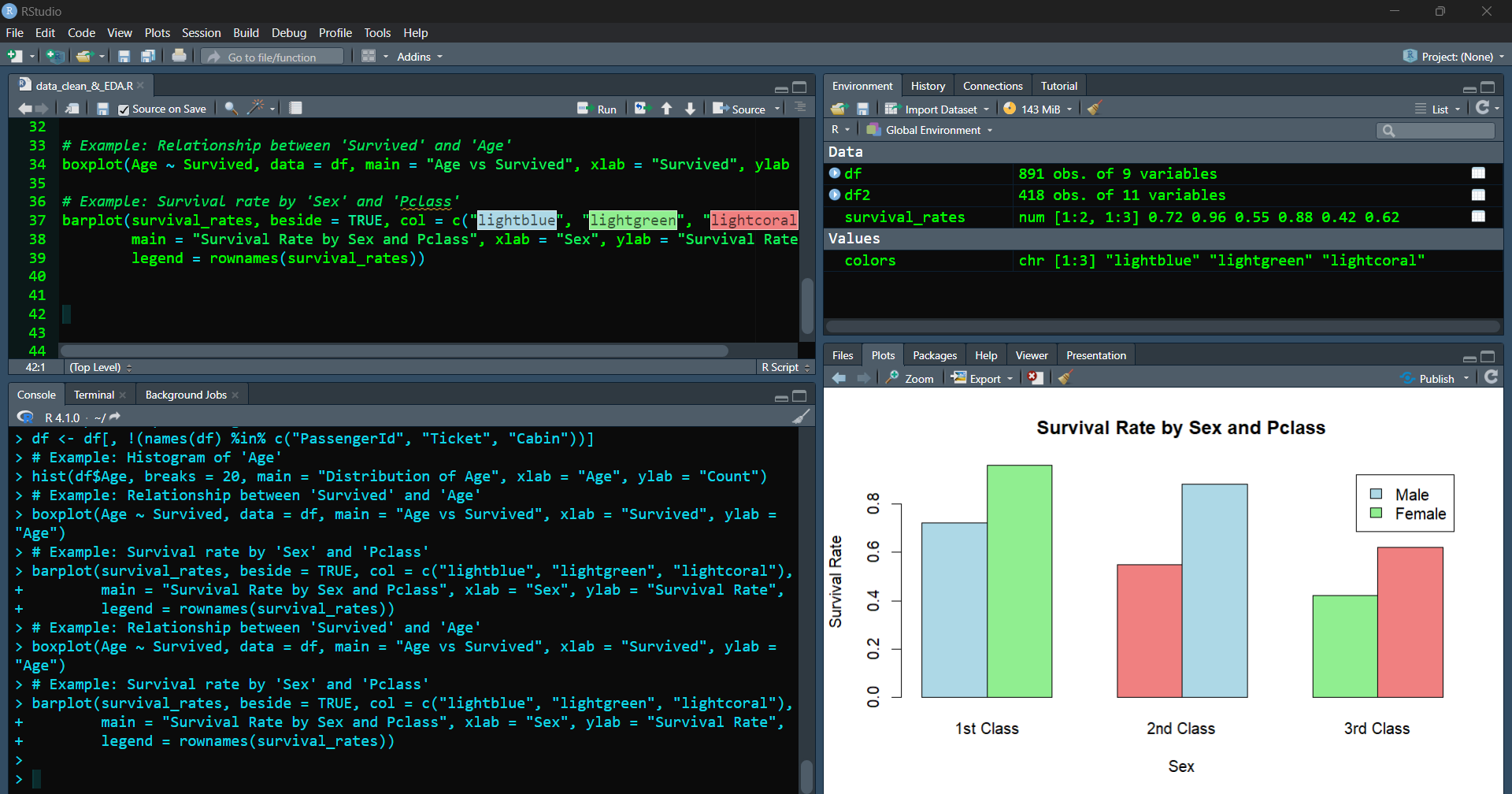
# Example: Survival rate by 'Sex' and 'Pclass'

library(ggplot2)

ggplot(df, aes(x = Sex, y = Survived, fill = factor(Pclass))) +

geom\_bar(stat = "identity", position = "dodge") +

labs(title = "Survival Rate by Sex and Pclass")



### End of Report###