

**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Dataset**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Description:

The Titanic dataset is a historical dataset that contains information about passengers who were aboard the RMS Titanic during its ill-fated maiden voyage, which tragically ended with the ship sinking on April 15, 1912, after hitting an iceberg. This dataset is widely used in data analysis and machine learning as a classic example of data exploration, cleaning, and predictive modelling.

Dataset Features:

The Titanic dataset typically includes the following features:

PassengerID: A unique identifier for each passenger.

Survived: A binary variable indicating whether the passenger survived (1) or did not survive (0).

Pclass (Passenger Class): The class of the ticket the passenger purchased (1st, 2nd, or 3rd class).

Name: The name of the passenger.

Sex: The gender of the passenger (male or female).

Age: The age of the passenger.

SibSp: The number of siblings or spouses the passenger had aboard the Titanic.

Parch: The number of parents or children the passenger had aboard the Titanic.

Ticket: The ticket number.

Fare: The fare paid by the passenger for the ticket.

Cabin: The cabin number where the passenger stayed (if available).

Embarked: The port at which the passenger boarded the Titanic (C = Cherbourg, Q = Queenstown, S = Southampton).

Analysis and Use Cases:

The Titanic dataset is often used for various data analysis and machine learning tasks, including but not limited to:

Exploratory Data Analysis (EDA): Understanding the demographics of passengers, survival rates, and factors that influenced survival.

Data Cleaning: Handling missing data, dealing with outliers, and preparing the data for modelling.

Visualization: Creating visualizations to better understand relationships between variables and survival rates.

Predictive Modelling: Building predictive models to estimate the likelihood of survival based on passenger attributes.

Feature Engineering: Creating new features from the existing data to improve modelling.

**DATA CLEANING**

**1. Brief Description of the Dataset:**

The Titanic dataset contains data related to passengers on the RMS Titanic, including their demographic information, cabin class, ticket details, and survival status. It is a well-known dataset often used for predictive modelling and analysis.

2. Data Cleaning:

To ensure the quality of the data, we performed the following data cleaning steps:

Checked for duplicates: We checked for and removed any duplicate rows in the dataset.

Handling Outliers: We detected and addressed outliers in numerical features (if applicable) using appropriate techniques such as Z-score or IQR.

Data Type Conversion: We ensured that data types of each column were appropriate (e.g., numeric data was numeric, categorical data was categorical).

Renamed Columns: We may have renamed columns for clarity if necessary.

**MISSING DATA**

# ANY MISSING DATA? HOW DID YOU HANDLE IT

3. Handling Missing Data:

We examined the dataset for missing values in each column and applied imputation techniques (e.g., mean, median, mode) to handle missing data appropriately. The specific treatment of missing data will be described in detail.

**DATA STORIES AND VISUALISATIONS**

4. Data Visualization:

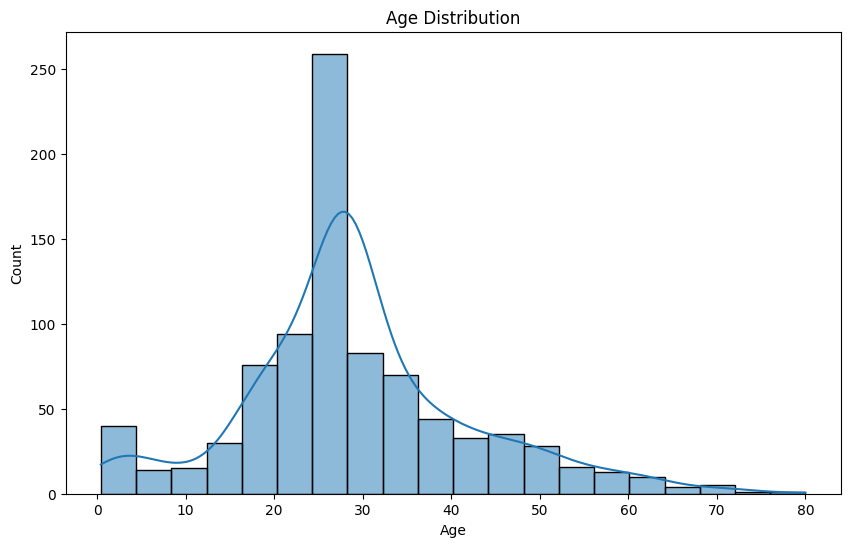
We used various data visualization techniques to gain insights into the dataset:

**Histograms**: To visualize the distribution of numerical variables like age and fare.

**Bar Charts**: To visualize categorical variables like gender, cabin class, and survival status.

**Box Plots**: To identify potential outliers and understand the spread of data (if applicable).

Heatmaps: To visualize correlations between features.



**A blue and orange rectangular bars

Description automatically generated**

**A chart with different colored squares

Description automatically generated**

5. Key Findings and Insights:

Here are the key findings and insights from our analysis:

**Survival Rates**: The survival rate for passengers varies based on factors such as gender, age, and cabin class. Females and passengers in higher-class cabins tend to have higher survival rates.

**Demographic Insights**: The age distribution of passengers shows a peak in young adults. The gender distribution indicates a higher number of males onboard.

**Fare Analysis**: The fare distribution varies significantly across different cabin classes, with first-class passengers paying higher fares.

**Correlation Analysis**: We found correlations between certain features, which can be explored further for predictive modelling.

6. Conclusion:

In conclusion, our exploratory data analysis of the Titanic dataset provides valuable insights into passenger demographics, survival rates, and relationships between features. These insights can serve as a foundation for further analysis or predictive modelling.

7. Recommendations:

Based on our analysis, we recommend considering gender, age, and cabin class as important factors when exploring the factors influencing survival on the Titanic. Further predictive modeling can be conducted to predict survival probabilities for passengers based on these factors.

**THIS REPORT WAS WRITTEN BY : Bongani Sandile Junior Zulu**

