

Lorena Karine Sousa dos Santos, ID 100404805

Exploratory Data Analysis Using Titanic Data Set

The Titanic dataset contains information about the passengers who were onboard the Titanic when it sank in April 15, 1912. The data set data like demographics, ticket info and survival status. The purpose of analyzing it is to gain insights of what made a difference in a passenger's survival rate and uncover patterns and relationships in the data that can help us understand the events. Also, it is helpful for beginners in data to practice in how to use all the skills we're learning.

The dataset has 891 rows and 12 columns, including PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The data types of these columns include 7 integer, 2 float, and 3 object columns. The dataset provides information regarding demographics, ticket information, and survival status. Nevertheless, the Age, Cabin, and Embarked columns contain some missing values, which require proper handling during data analysis.

Hypothesis 1

Comparing before and after the accident, we could see that the majority of the members of Titanic were in the class 3, followed by 1 and 2. Still, 62% of the survivors were from class 1, followed by 2 (47%) and 3 (24%). We can assume that the higher the class (first, second, third) the more safe they were.

Hypothesis 2

From the total members of the Titanic, 64% were male and 35% were female. Still, the survival rates indicates that 74% were women while only 18% were men. In the Titanic accident, we could assume that is more likely to survive if one is a woman. More data is needed to assume the whys of that

Hypothesis 3

Using a histogram to check on the ages of the people onboard, it did not show a relationship between the survivals or not according to their age. We can see many children and infants, as well as adults and elder people between the survivals. Age did not seem to play a role in a survival rate during the Titanic accident.