TITANIC EXPLORATORY DATA ANALYSIS REPORT

By: Blessing Ilesanmi (WITS 24)

Women In Tech Scholarship Data Science Final Project

7th April, 2025

**INTRODUCTION**

The Titanic dataset is a widely recognized historical dataset that contains detailed information about the passengers who were aboard the RMS Titanic during its ill-fated maiden voyage in April 1912. The ship, which was considered unsinkable at the time, tragically struck an iceberg and sank in the North Atlantic Ocean, resulting in the deaths of over 1,500 passengers and crew members. The dataset includes various features such as passenger age, sex, ticket class, fare paid, port of embarkation, and survival status.

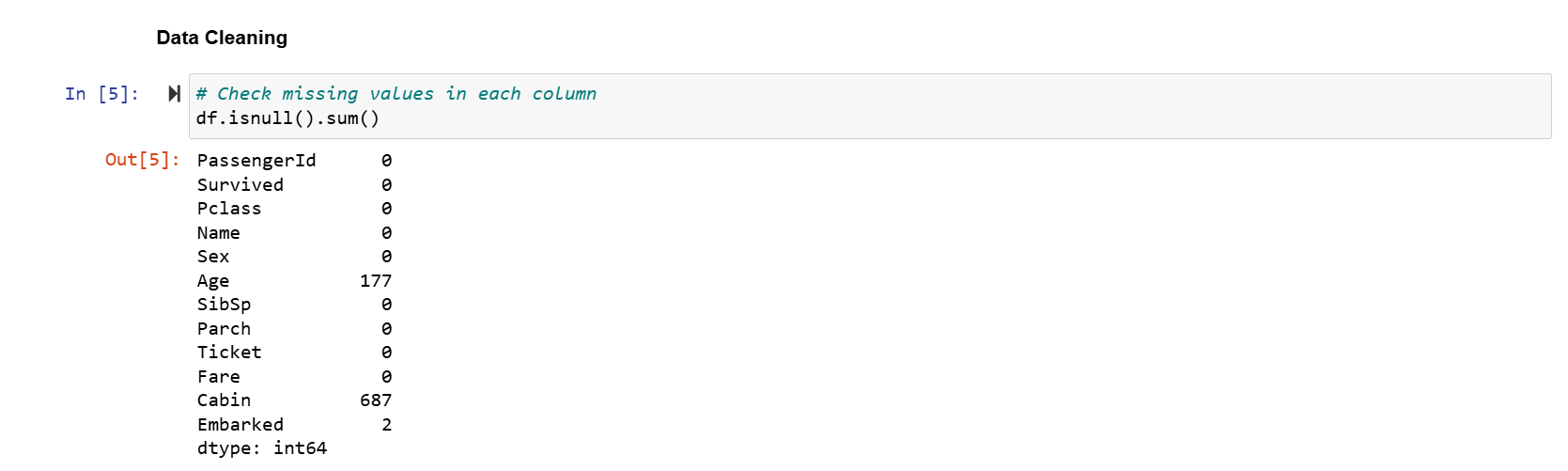
The goal of this analysis is to explore and understand the factors that influenced survival during the disaster. By performing exploratory data analysis (EDA), this report aims to uncover relationships and trends within the dataset — particularly focusing on how attributes like age, gender, and passenger class affected the likelihood of survival. Through statistical summaries and visual representations, this report provides insights into the underlying patterns of the Titanic tragedy, offering a data-driven perspective on one of the most well-known maritime disasters in history.

**DATA LOADING AND CLEANING**

The dataset used for this analysis was obtained from Kaggle and consists of three CSV files: *train.csv*, *test.csv*, and *gender\_submission*.csv. For the purpose of this exploratory analysis, only the *train.csv* file was used, as it contains both the feature variables and the target variable (Survived). Upon loading the data into a Pandas DataFrame, the shape of the dataset was examined. The training set contains 891 row*s*and 12 columns, representing individual passengers and their respective features, such as age, sex, ticket class, fare, and survival status.

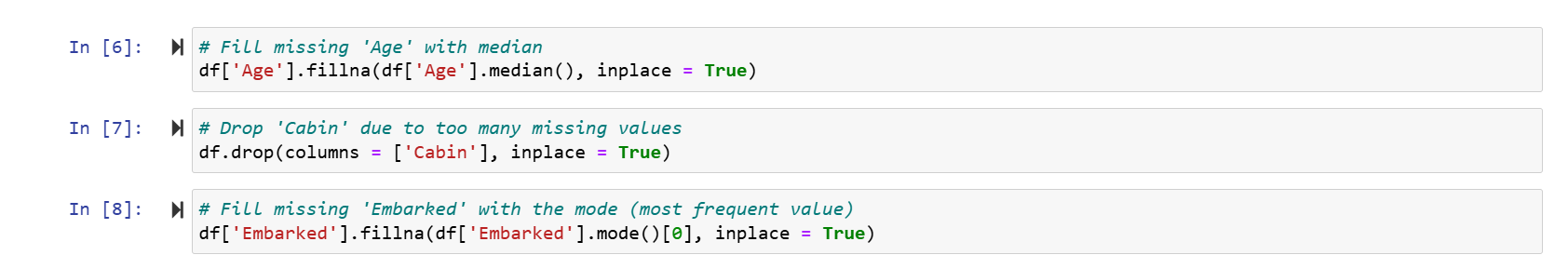


Data cleaning inspection revealed that some columns contained missing values. The Age column had a moderate number of missing entries (177), while the Cabin column had a substantial proportion of missing values (687). The Embarked column was missing only two values.



To address this:

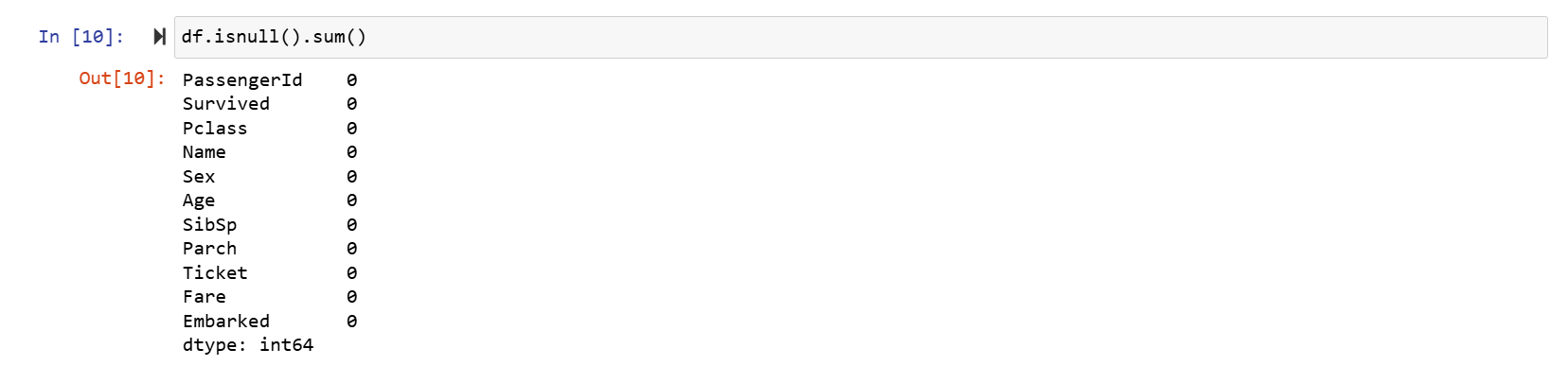
* Missing values in the Age column were filled using the median age to minimize the influence of outliers.
* The Cabin column was dropped due to the high volume of missing data and its limited utility for this analysis.
* Missing values in the Embarked column were imputed with the most frequent port of embarkation (mode).



The dataset was also checked for duplicate entries, and none were found.



After these steps, the dataset was clean, with no missing or duplicate values, and was ready for further exploration and analysis.



**SUMMARY STATISTICS**

The dataset was first analyzed to compute basic summary statistics, which helped in understanding the distribution of values across different columns. The key statistics for numerical columns are as follows:

* **Age**: The average age of passengers was approximately 29.36 years, with a minimum of 0.42 years (an infant) and a maximum of 80 years.
* **Fare**: The average fare paid was about 32.20, but the fare ranged widely, with some passengers paying as much as 512.33, indicating some high-class passengers.
* **Pclass**: The majority of passengers (55.11%) were in the third class, followed by first class (24.24%) and second class (20.65%).

