# ETL Pipeline Report for Titanic Dataset

## Introduction

The Titanic dataset contains information about the passengers on the Titanic, including their demographics and survival status. This report outlines the steps taken to extract, transform, and load (ETL) this dataset, resulting in a clean and transformed data file ready for analysis.

## Steps Performed

### 1. Extract

The dataset was read from a CSV file named `titanic.csv`. The following columns were present in the dataset:

- PassengerId

- Survived

- Pclass

- Name

- Sex

- Age

- SibSp

- Parch

- Ticket

- Fare

- Cabin

- Embarked

### 2. Transform

The data transformation process involved the following steps:

1. Handling Missing Values:

- Age: Filled missing values with the median age of the dataset.

- Embarked: Filled missing values with the mode (most common value).

2. Converting Categorical Data:

- Sex: Converted 'Sex' to numeric form where 'male' = 1 and 'female' = 0.

- Embarked: Applied one-hot encoding to convert 'Embarked' into numeric columns 'Embarked\_C', 'Embarked\_Q', and 'Embarked\_S'.

3. Dropping Irrelevant Columns:

- Deck: Dropped due to a high number of missing values.

### 3. Load

The transformed dataset was saved to a new CSV file named `transformed\_titanic.csv`. This file now contains the following columns:

- PassengerId

- Survived

- Pclass

- Sex

- Age

- SibSp

- Parch

- Fare

- Embarked\_C

- Embarked\_Q

- Embarked\_S

## Result

The ETL process successfully cleaned and transformed the Titanic dataset. The resulting `transformed\_titanic.csv` file is ready for further analysis or machine learning tasks. The transformations applied ensure that the dataset is now consistent, with no missing values in critical columns, and all necessary categorical data has been converted to numeric form.

## Conclusion

This ETL pipeline demonstrates a fundamental data engineering process, showcasing how to handle missing values, convert categorical data, and prepare a dataset for analysis. The transformed Titanic dataset is now more structured and suitable for various analytical purposes.