Titanic Dataset Cleaning Project Overview

This project uses the Titanic passenger dataset from Kaggle, which includes 891 rows of data on passengers f

Key columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

Main issues addressed: Missing values (e.g., 177 in Age, 687 in Cabin), potential duplicates, inconsistent forr

Goal: Produce a clean dataset ready for analysis, demonstrating skills in data imputation, standardization, an

Issue	**Before**	**After**
Missing Age	177 blanks	Filled with median (28)
Missing Cabin	687 blanks	Column dropped
Missing Embarked	2 blanks	Filled with 'S'
Duplicates	0	0
Outliers in Fare	Some >$200	Capped at upper IQR bound (~$65)
Formats	Mixed case in Sex/Embarked	Standardized (lowercase Sex, uppercase Embarked)

rom the 1912 disaster.

atting, and outliers in Age/Fare.
d validation using Excel.

Method Used
Median imputation
Too sparse for use
Mode (most common)
Checked and none found
IQR method
Formulas like LOWER/UPPER