

Titanic Dataset Analysis - Part1

Dataset Download

Download the Titanic dataset from:

- Kaggle: <https://www.kaggle.com/c/titanic/data>
- Direct CSV link:
<https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv>

Assignment Overview

In this assignment, you will analyze the Titanic passenger dataset to uncover patterns and insights about survival factors during the Titanic disaster. You'll practice data cleaning, feature engineering, and statistical analysis using pandas.

Prerequisites

- Python 3.x
- pandas
- numpy

Assignment Tasks

1. Data Cleaning

a) Handle missing values:

- Fill missing ages using median age based on passenger class
- Handle missing values in 'Cabin' column
- Process missing values in 'Embarked' column

b) Create new age categories:

- Child (0-15)
- Young Adult (16-30)
- Adult (31-45)
- Middle Aged (46-60)
- Senior (60+)

2. Feature Engineering

Create the following new features:

a) FamilySize:

- Combine 'SibSp' and 'Parch'
- Create a new column showing total family members

b) Title from Name:

- Extract titles (Mr, Mrs, Miss, etc.) from passenger names
- Create a new column for titles

c) Ticket Fare Analysis:

- Create fare ranges/bins
- Calculate fare per person for family groups

d) Cabin Information:

- Extract deck information (first letter of cabin)
- Create a new column for deck

Deliverables

1. Python script (.py) or Jupyter notebook (.ipynb)
2. Brief report (PDF) containing:
 - Summary of findings
 - Key insights
 - Methodology explanation
 - Visualizations
 - Conclusions

Note

- All code should be properly documented
- Include error handling where appropriate
- Follow Python PEP 8 style guidelines
- Use efficient pandas operations
- Avoid duplicate code

Resources

- Pandas documentation: <https://pandas.pydata.org/docs/>
- Kaggle Titanic tutorials: <https://www.kaggle.com/c/titanic/tutorials>