Project Instructions –

Data Cleaning and Analysis with the Titanic Dataset

Introduction

In this individual project you will apply your skills in data handling and exploratory data analysis (EDA). You will work with the Titanic dataset from Kaggle, which contains information about passengers on the Titanic and whether they survived or not. The dataset is known to include missing values, different data types and other challenges, making it ideal for practicing data cleaning and analysis.

Objectives

- Load, inspect, and understand a real-world dataset.
- Identify and handle missing values and inconsistencies.
- Perform basic data cleaning and feature engineering.
- Conduct exploratory data analysis (EDA) with visualizations.
- Build a simple model to demonstrate how the cleaned data can be used for prediction.

Data Source

The dataset is available on Kaggle under the name 'Titanic: Machine Learning from Disaster'.

Link:

https://www.kaggle.com/c/titanic/data

Minimum Requirements

1. Inspect and describe the dataset (columns, datatypes, missing values).
2. Handle missing data and explain your choices.
3. Engineer at least 2 new features.
4. Perform EDA with at least 3 different visualizations.
5. Build one simple model of your choice (e.g., logistic regression, decision tree, k-nearest neighbors).
6. Write a short conclusion summarizing your findings.

Example Ideas and Inspiration

These are not requirements, but possible directions you can explore if you need inspiration:

- Feature Engineering: extract passenger titles (Mr, Mrs, Miss, Master) from the Name column.
- Create a FamilySize feature (SibSp + Parch + 1).
- Group Age into categories (child, adult, senior). • Simplify Cabin by using only the first letter (deck).
- EDA: visualize survival rate by gender, ticket class, age group, embarkation port.
- Compare survival between people traveling alone vs. with family.

Example Models

Here are some examples of simple models you could build once your data is cleaned:

- Logistic Regression to predict survival (binary classification).
- Decision Tree Classifier to see which features best split survivors and non survivors.
- K-Nearest Neighbors (KNN) to predict survival based on similarity to other passengers.

Remember: the **main focus** is on data handling and analysis, not on building a perfect predictive model.