**ProdigyInfotech – Data Science Internship.**

**Completion of task 02**

**DATA UNDERSTANDING**

* The datasets is obtained from Kaggle: Titanic
* The dataset contains 891 rows (entries) and 12 columns The columns are:
* PassengerId: Unique identifier for each passenger.
* Survived: Binary variable indicating survival (1 = Survived, 0 = Did Not Survive).
* Pclass: Ticket class (1st, 2nd, 3rd class).
* Name: Passenger's name.
* Sex: Gender of the passenger.
* Age: Age of the passenger.
* SibSp: Number of siblings/spouses aboard.
* Parch: Number of parents/children aboard.
* Ticket: Ticket number.
* Fare: Passenger fare.
* Cabin: Cabin number.

**Steps for Data Cleaning and EDA on Titanic Dataset**

1. **Load the Data**
   * Import necessary libraries (pandas, numpy, matplotlib, seaborn).
   * Load the Titanic dataset (df = pd.read\_csv('titanic.csv')).
2. **Initial Data Exploration**
   * Check the first few rows of the dataset (df.head()).
   * Understand the summary statistics of numerical variables (df.describe()).
   * Check for data types and missing values (df.info()).
3. **Data Cleaning**
   * Handle missing values:
     + Identify columns with missing values (df.isnull().sum()).
     + Decide on strategies to fill missing values (e.g., median for numerical data, mode for categorical data).
     + Implement filling missing values (df.fillna() or df.drop() as needed).
4. **Exploratory Data Analysis (EDA)**
   * Perform univariate analysis:
     + Visualize distributions of numerical variables (e.g., histogram, boxplot).
     + Count categorical variables (e.g., sns.countplot for categorical variables).
   * Perform bivariate/multivariate analysis:
     + Explore relationships between variables:
       - Compare survival rates across different categories (e.g., sex, passenger class).
       - Investigate correlations between numerical variables (e.g., age vs. fare).
5. **Insights and Conclusion**
   * Summarize findings from EDA:
     + Highlight significant patterns or trends observed (e.g., higher survival rates for certain groups).
     + Discuss any correlations or relationships discovered (e.g., correlation between fare and passenger class).
     + Provide insights into potential factors influencing survival on the Titanic.

By following these steps, you can systematically clean the data, explore its characteristics, and derive meaningful insights through EDA, which is crucial for understanding the Titanic dataset or any similar dataset.

Top of Form

Bottom of Form

**Refer the Link :**

1. <https://colab.research.google.com/drive/1vz0nuWNM5IgAV0sVV9O03-RPBN4Wu-V5#scrollTo=NZD42EctH2Fs>
2. <https://github.com/SANJAY1409/Prodigy_DS_02>