Data Exploration and Insights on Titanic Dataset: A Real-World Case Study

Business Context As part of a data science consultancy firm working with the maritime industry, your team has been tasked with uncovering insights into the survival dynamics of passengers aboard the Titanic. By conducting a thorough **Data Exploration** on the Titanic dataset, your objective is to analyze passenger demographics, ticket classes, and other key attributes to identify factors that may have influenced survival. This exercise will help stakeholders design better safety protocols and risk management strategies in future maritime operations.

Dataset Overview

- **Dataset**: Titanic Passenger Dataset (from Kaggle)
- **Features**: Passenger details including age, gender, ticket class, fare, port of embarkation, and survival status.
- **Link**: https://www.kaggle.com/datasets/brendan45774/test-file

Project Workflow

Phase 1: Data Acquisition and Initial Setup

1. **Objective Definition**
   - Clearly outline the goals of the analysis (e.g., understanding survival patterns).
2. **Environment Setup**
   - Install and import the required libraries: `Numpy`, `Pandas`, `Matplotlib`, `Seaborn`.
   - Load the dataset into a Jupyter Notebook for processing.

Phase 2: Data Understanding and Preprocessing

1. **Dataset Structure and Preview**
   - Examine the dataset's structure using `info()` and `head()`.
   - Check for missing data.
2. **Data Cleaning**
   - Address missing values (e.g., impute missing ages, handle missing `Embarked` values).

Phase 3: Exploratory Data Analysis (EDA)

Univariate Analysis

- Analyze the distribution of **numerical features** like `Age`, `Fare` using histograms and KDE plots.
- Examine **categorical features** (`Pclass`, `Sex`, `Embarked`) using bar plots.
- Understand survival rates through count plots and pie charts.

Bivariate Analysis

- Assess the relationship between **Age** and **Survival** using scatter plots.
- Analyze survival rates by **Gender** and **Class** using grouped bar plots.
- Use box plots to compare **Fare** distributions across different **Pclass** categories.

Multivariate Analysis

- Combine multiple features (e.g., `Pclass`, `Sex`, `Survived`) in a **FacetGrid** to analyze combined survival trends.
- Employ violin plots to compare **Age** and **Survival** within **Pclass** categories.

Correlation Analysis

- Generate a **correlation matrix** to identify relationships among numerical features.
- Visualize correlations using Seaborn's **heatmap** to spot highly correlated variables.

Phase 4: Advanced Data Visualization

- Use Seaborn's **pairplot** to explore feature relationships.
- Leverage **violin plots** and **swarm plots** for detailed survival analysis.
- Apply **categorical heatmaps** to show survival percentages by port of embarkation, gender, and class.

Phase 5: Business Insights and Recommendations

- Summarize findings such as:
  - Higher survival rates for women and children.
  - Class-based survival disparities (e.g., first-class passengers had better survival odds).
  - The impact of fare and age on survival.

Key Learning Outcomes

- Mastery of real-world Data Exploration techniques.
- Hands-on experience with industry-standard libraries like **Pandas**, **Matplotlib**, and **Seaborn**.
- Ability to present data-driven insights and recommendations to non-technical stakeholders.

Deliverables

1. **Interactive Jupyter Notebook** with detailed EDA steps.
2. **Submit** in your respective mentoring WhatsApp group in .ipynb format..