🔍 Report: Titanic Dataset Analysis – Elevate Labs Task 5

📁 Dataset Used

- **Source:** train.csv (Titanic dataset)
- **Purpose:** Explore passenger data to understand survival patterns and prepare for potential modeling

---

📊 1. Data Exploration

- The dataset was loaded using **pandas**, and key structural details were examined:
    - .info() revealed the data types and presence of null values.
    - .describe() provided summary statistics of numeric columns.
    - .isnull().sum() helped identify columns with missing data (e.g., Age, Cabin, Embarked).

🧼 2. Data Preprocessing

- Basic preprocessing steps were observed:
    - Handling missing values using methods such as .fillna()
    - Likely dropped or imputed non-numeric or incomplete columns
    - Potential encoding of categorical data using functions like pd.get_dummies() or LabelEncoder (though no detailed transformation was documented)

📈 3. Data Visualization

- Visualization libraries **Seaborn** and **Matplotlib** were used to uncover trends:
    - Plots likely included survival rates segmented by features such as:
        - **Sex** (e.g., higher survival rate for females)
        - **Pclass** (e.g., 1st class had better outcomes)
        - **Age** and **Fare** distributions
    - Possibly used bar plots, histograms, and heatmaps for correlation or missing data patterns

---

❌ 4. Missing Elements

- **No Machine Learning Models** were implemented:
    - No classifiers (e.g., Logistic Regression, Random Forest)
    - No train-test splitting or performance evaluation
- The notebook is focused on **exploratory data analysis (EDA)** only

---

✅ Recommendations for Next Steps

1. **Model Building**:
    - Try Logistic Regression or Random Forest to predict survival
    - Use `train_test_split` and `accuracy_score` for evaluation
2. **Feature Engineering**:
    - Extract titles from names, group family members, create age categories

   - Handle missing values in Cabin creatively (e.g., presence vs. absence)

3. **Cross-Validation & Hyperparameter Tuning**:

   - Use GridSearchCV to optimize models

📊 Plot Summaries

1. Survival Distribution
python
CopyEdit
```
sns.countplot(data=df, x='Survived')
```

- **Purpose:** Shows the overall count of survivors (1) and non-survivors (0).
- **Insight:** More passengers did not survive than those who did.

---

2. Age Distribution
python
CopyEdit
```
sns.histplot(df['Age'].dropna(), kde=True)
```

- **Purpose:** Displays the age distribution of passengers.
- **Insight:** Most passengers were between 20–40 years old, with a smooth density curve overlay.

---

3. Survival Count by Gender
python
CopyEdit
```
sns.countplot(x='Sex', hue='Survived', data=df)
```

- **Purpose:** Compares survival rates between male and female passengers.
- **Insight:** Females had a significantly higher survival rate than males.

---

4. Survival by Passenger Class
python
CopyEdit
```
sns.countplot(x='Pclass', hue='Survived', data=df)
```

- **Purpose:** Examines survival differences across passenger classes (1st, 2nd, 3rd).
- **Insight:** Passengers in 1st class had the highest survival rate, while 3rd class had the lowest.

---

5. Correlation Heatmap
python
CopyEdit
```
sns.heatmap(df[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']].corr())
```

- **Purpose:** Displays correlation between numeric variables.
- **Insight:** `Fare` and `Pclass` show the strongest relationship to `Survived`.

---

6. Boxplot of Age by Survival
python
CopyEdit
```
sns.boxplot(x='Survived', y='Age', data=df)
```

- **Purpose:** Visualizes the age spread for survivors vs. non-survivors.
- **Insight:** Survivors tended to be slightly younger; median age was lower for survivors.

---

7. Survival Rate by Embarked Port
python
CopyEdit
```
sns.barplot(x='Embarked', y='Survived', data=df)
```

- **Purpose:** Shows average survival rate per embarkation port (C, Q, S).
- **Insight:** Passengers embarking from port C had the highest survival rate.

---

8. Fare vs. Age by Survival
python
CopyEdit
```
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
```

- **Purpose:** Scatterplot of fare paid vs. age, color-coded by survival.
- **Insight:** Survivors generally paid higher fares, often associated with younger and middle-aged groups.

---

9. Survival Rate by Age Group
python
CopyEdit
```
sns.barplot(x='AgeGroup', y='Survived', data=df)
```

- **Purpose:** Groups passengers into age brackets and shows average survival.
- **Insight:** Children had the highest survival rate, seniors had the lowest.

---

10. Survival by Class and Embarked Port
python
CopyEdit
```
sns.heatmap(df.pivot_table(index='Pclass', columns='Embarked', values='Survived'))
```

- **Purpose:** Heatmap of survival rates broken down by class and embarkation port.
- **Insight:** 1st class passengers from port C had the highest survival rates.

---

11. Survival by Siblings/Spouses Aboard

python
CopyEdit

```
sns.countplot(x='SibSp', hue='Survived', data=df)
```

- **Purpose:** Compares survival by the number of siblings/spouses aboard.
- **Insight:** Solo travelers and those with 1 companion had better chances than those with many companions.