**Task Two Work Description**

**Introduction**

In this task, I conducted an Exploratory Data Analysis (EDA) on the Titanic dataset provided on Kaggle. The goal was to understand the structure of the data, clean it, generate useful statistical insights, and visualize important patterns, especially those related to passenger survival. The tools used were Python, Pandas for data manipulation, and Seaborn/Matplotlib for visualization.

**Uploading and Understanding the Data**

I started by uploading the **train.csv** file, which contains information about passengers such as their age, sex, class, and survival status. The initial exploration included displaying the first 10 rows of the dataset to get a basic understanding of its structure. Then, I used df.info() and df.describe(include='all') to learn about data types, the presence of missing values, and descriptive statistics for each column. This step is essential because it helps identify which columns need cleaning and which ones can be used directly in analysis.

**Data Cleaning**

After identifying missing values with df.isnull().sum(), I proceeded to clean the data:

* **Age:** This column had 177 missing values. Since age is a numeric feature important for analysis, I filled the missing values using the median age to avoid skewing the data.
* **Embarked:** Only two records were missing the embarkation port. I filled these with the mode (the most common value), which is a standard practice when the missing rate is very low.
* **Cabin:** This column had a large number of missing values (687). Instead of dropping it entirely, I engineered a new feature HasCabin, which indicates whether a passenger had a cabin or not (1 if present, 0 if missing). I then dropped the original Cabin column since its detailed values were not useful in this form.

**Generating Summary Statistics and Group-Based Insights**

With a clean dataset, I generated summary statistics again to verify our data. Next, I focused on group-based insights:

* I analyzed survival rates based on **gender** and found that females had a significantly higher survival rate compared to males.
* I analyzed survival rates based on **passenger class (Pclass)** and observed that first-class passengers had the highest survival rates, while third-class passengers had the lowest.
* I further combined **gender and class** to see more detailed patterns, using a pivot table to visualize how these two factors interact in relation to survival.

**Data Visualization**

Visualizing the data helps to uncover patterns that are not immediately obvious from statistics alone. I created several visualizations:

* A **correlation heatmap** was generated using only numeric columns to explore relationships between variables. This showed which features had the strongest correlation with survival.
* A **count plot** of survival by gender clearly illustrated that females had a much higher survival rate.
* A **count plot** of survival by passenger class confirmed that higher classes had better survival outcomes.
* An **age distribution histogram** with survival overlay helped visualize how survival chances varied across different age groups.

**Bonus Visualizations**

I also completed the bonus section by adding additional visual insights:

* A **bar plot** showing survival rates across combinations of passenger class and gender provided a more granular look at the interaction of these two factors.
* A **heatmap** of survival rates based on gender and embarkation port highlighted how the port of embarkation influenced survival rates for different genders.

**Conclusion**

In conclusion, I successfully followed all the steps required in the task. I explored the Titanic dataset, cleaned it thoroughly, generated comprehensive statistical summaries, and visualized key patterns and correlations. The analysis clearly showed that gender, passenger class, and embarkation port were important factors influencing survival. The bonus visualizations further deepened our understanding of how these factors interacted. This EDA serves as a strong foundation for any further predictive modeling or detailed analysis on this dataset.