NAME: SNEHA K

19-05-2025

TASK 5

1. Scatter Plot: Age vs Fare

Observation:

- There is a **cluster** of passengers between ages 20–40 who paid **low to moderate fares** (under 100).
- A few outliers in older age brackets paid significantly high fares.
- **No strong linear relationship** between age and fare is evident.

2. Scatter Plot (Fare vs Age, Colored by Survival)

Observation:

- Survivors (Survived = 1) are more concentrated in the **higher fare** range.
- Non-survivors are scattered across all fare ranges, but dominate the **low fare** region.
- This suggests a **positive association between fare and survival**, i.e., higher paying passengers were more likely to survive.

3. Survival Rate by Gender

- **Observation:**
  - The survival rate of **females** was **much higher** than that of **males**.
  - This confirms the "**women and children first**" protocol followed during the evacuation.

4. Correlation Heatmap

- **Observation:**
  - **Survived** is:
    - Negatively correlated with Pclass (higher class, higher survival).
    - Slightly positively correlated with Fare.
  - **Fare** and Pclass show a strong negative correlation.
  - **Age** has weak correlation with survival.

5. Line Plot: Fare vs Survival Rate

- **Observation:**
  - Survival rate generally **increases with fare**, although some variability exists.
  - Indicates that passengers who paid **higher fares had better chances of survival**.

SUMMARY:

The analysis reveals a strong gender-based survival trend: women had a much higher survival rate (about 74%) compared to men (only about 19%). This reflects the historical policy of prioritizing women during the evacuation.

A clear trend emerges when examining the relationship between **fare and survival**. Passengers who paid higher fares were generally more likely to survive. This pattern is consistent across both the scatter plot and the survival rate line plot, indicating a socioeconomic bias—first-class passengers had better access to lifeboats and resources.

The scatter plot comparing **Age and Fare** shows that most passengers paying lower fares were in the 20–40 age range, with no strong linear trend between age and fare. However, when color-coded by survival, the plot shows that survivors were concentrated in the higher-fare segment, reinforcing the earlier observation.

The **correlation matrix heatmap** supports these findings: survival is negatively correlated with passenger class (Pclass) and positively correlated with fare. Age, on the other hand, shows a weak or negligible correlation with survival, suggesting it was not a major determining factor in survival outcomes.

Overall, the visual data strongly indicates that **fare paid (and hence, passenger class)** and **gender** were significant factors in survival, while **age** played a relatively minor role.

INTERVIEW QUESTIONS :

1. What is EDA and why is it important?

Exploratory Data Analysis (EDA) is the process of exploring and understanding your dataset before applying any machine learning models. It helps you identify patterns, trends, anomalies, and relationships within the data. EDA is crucial because it gives you a deep understanding of your data's structure and quality, helping you make better decisions about cleaning, transforming, and modeling it.

2. Which plots do you use to check correlation?

To check the correlation between variables, you can use visualizations such as heatmaps, scatter plots, and pairplots. A heatmap gives a color-coded view of how strongly variables are related, while scatter plots let you visually examine the relationship between two variables. Pairplots are helpful to view multiple scatter plots together and see all variable combinations at once.

3. How do you handle skewed data?

Skewed data can be managed by applying transformations to make the distribution more symmetrical. Common methods include log, square root, or box-cox transformations. You can also treat extreme values by capping or removing them, or choose models that are less affected by skewness. The goal is to reduce the impact of outliers and improve model accuracy

4. How to detect multicollinearity?

Multicollinearity occurs when two or more features in a dataset are highly related to each other, which can distort model predictions. You can detect it by looking at a correlation matrix to find features that are highly correlated. Another method is checking the Variance Inflation Factor (VIF), where a high value indicates the presence of multicollinearity.

5. What are univariate, bivariate, and multivariate analyses?

Univariate analysis involves examining a single variable at a time to understand its distribution or frequency. Bivariate analysis explores the relationship between two variables, such as age and income. Multivariate analysis looks at three or more variables together to understand complex interactions and patterns across multiple features.

6. Difference between heatmap and pairplot

A heatmap is a grid-like visualization that shows the strength of relationships (correlation) between numerical variables using colors. It's useful for quickly identifying which variables are strongly related. On the other hand, a pairplot displays scatter plots for each pair of variables and also shows their individual distributions. It's more detailed and helps visually explore relationships between multiple variables.

7. How do you summarize your insights?

Insights are summarized by clearly stating what was observed, why it matters, and what actions or interpretations can be made from it. You can break down your findings into key points and explain patterns, unusual values, relationships, or trends. Good summaries are simple, logical, and directly tied to your analysis goals. Always focus on clarity and relevance when communicating results.