Data Science Lab 1 – Titanic Dataset Analysis

U23AI114
B.Tech AI (SEM V)

Objective

The purpose of this lab is to:

1. Create and analyze frequency tables for categorical variables.
2. Calculate joint, marginal, and conditional probabilities from contingency tables.
3. Understand and compute correlation between numerical variables.

Dataset

We used the Titanic dataset available from Seaborn: `https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv`

This dataset contains demographic and survival information of passengers aboard the Titanic.

Part I: Frequency Table

Passenger Class Distribution:

Class	Absolute Freq	Relative Freq	Cumulative Freq	Cumulative Rel. Freq
First	216	0.2424	216	0.2424
Second	184	0.2065	400	0.4489
Third	491	0.5511	891	1.0000

Observation: More than half of the passengers (55.11%) were traveling in Third Class, making it the largest group. First and Second Class accounted for 24.24% and 20.65% respectively.

Part II: Probability Analysis

Contingency Table (Sex vs Survived)

	Survived=0	Survived=1	Total
Female	81	233	314
Male	468	109	577
Total	549	342	891

Observation: The majority of male passengers did not survive (468 out of 577), while a high proportion of female passengers survived (233 out of 314).

Joint Probability:

P(female, survived=1) = 0.2615 **Interpretation:** Around 26.15% of all passengers were females who survived.

Marginal Probabilities:

$$P(\text{Sex=female}) = 0.3524, \quad P(\text{Survived=1}) = 0.3838$$

Interpretation: About 35.24% of passengers were female, and 38.38% of all passengers survived.

Conditional Probabilities:

$$P(\text{Survived=1} \mid \text{Sex=female}) = 0.7420$$

$$P(\text{Sex=female} \mid \text{Survived=1}) = 0.6813$$

Interpretation: Females had a survival rate of 74.20%, while 68.13% of all survivors were female. This clearly shows that being female greatly increased the chances of survival.

Part III: Correlation Analysis

We selected `Age` and `Fare` as numerical variables. Pearson Correlation Coefficient: $\rho = 0.0961$

Observation: The correlation is weak and positive, meaning there is a slight tendency for older passengers to have paid higher fares, but the relationship is not strong.

Bonus Task: Survival by Class

A stacked bar chart was created (not shown here) which indicated that **First Class** passengers had the highest survival rate, followed by Second Class, with Third Class passengers having the lowest.

Conclusion

From this analysis:

- Third Class passengers were the majority, but they had the lowest survival rates.
- Female passengers had a significantly higher chance of survival than males.
- Age and fare showed only a weak positive correlation.
- Passenger class strongly influenced survival rates, with First Class being the safest.