Final Project Report:
Titanic Survival Analysis

1. Dataset Summary

The dataset chosen for this project is the **Titanic dataset**, which contains information on passengers aboard the RMS Titanic and whether they survived. It is commonly used for binary classification and statistical inference tasks.

- **Source**: Kaggle – Titanic: Machine Learning from Disaster
- **File Used**: train.csv
- **Dataset Dimensions**: 891 rows and 12 columns

Key Variables:

- **Target variable**: Survived (0 = No, 1 = Yes)
- **Predictor variables**:
    - Pclass: Ticket class (1st, 2nd, 3rd)
    - Sex: Gender
    - Age: Passenger's age
    - SibSp: Number of siblings/spouses aboard
    - Parch: Number of parents/children aboard
    - Fare: Ticket fare
    - Embarked: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

---

2. Data Exploration Plan

The goal of this project is to identify patterns that influenced survival. The analysis was structured as follows:

- Assessing distributions and summaries of key features
- Identifying and handling missing values
- Visualizing relationships between features and survival

- Engineering new features
- Preparing the dataset for hypothesis testing and machine learning

---

3. Exploratory Data Analysis (EDA)

A. Gender and Survival

- Observation: A significantly higher percentage of women survived compared to men.
- Visualization: Countplot showing the proportion of survivors by gender.

B. Passenger Class

- Observation: 1st class passengers had the highest survival rate, followed by 2nd, then 3rd.
- Visualization: Countplot of Pclass vs Survived.

C. Age Distribution

- Observation: Children had relatively higher survival rates; adults showed mixed patterns.
- Visualization: Histogram and KDE plot of Age split by survival.

D. Embarkation Port

- Observation: Passengers boarding at Cherbourg (C) had better survival outcomes.
- Visualization: Countplot showing Embarked vs Survived.

E. Correlation Heatmap

- Observation: Variables like Sex, Pclass, Fare, and engineered FamilySize showed moderate correlation with survival.

---

4. Data Cleaning & Feature Engineering

A. Missing Values

- Dropped the Cabin column due to over 75% missing values.

- Filled missing Age values with the median (28).
- Filled missing Embarked values with the most frequent value ('S').

B. Encoding Categorical Variables

- Sex: Encoded as male = 0, female = 1
- Embarked: Encoded as C = 0, Q = 1, S = 2

C. Engineered Features

- FamilySize = SibSp + Parch + 1
- IsAlone = 1 if FamilySize == 1, else 0

---

5. Key Findings & Insights

- Gender had a significant impact on survival — women were much more likely to survive.
- First-class passengers had a clear survival advantage.
- Passengers who traveled alone were less likely to survive.
- Embarkation port (Cherbourg) and fare amount also showed positive correlation with survival.
- These insights align with the historical accounts of the Titanic disaster.

---

6. Hypothesis Formulation

Three hypotheses were developed based on the EDA:

Hypothesis 1: Gender and Survival

- **Null (H_0)**: Survival is independent of gender
- **Alternative (H_1)**: Gender has a significant effect on survival

Hypothesis 2: Passenger Class and Survival

- **Null (H_0)**: Passenger class does not affect survival
- **Alternative (H_1)**: Higher-class passengers were more likely to survive

Hypothesis 3: Age and Survival

- **Null (H_0)**: The average age of survivors and non-survivors is the same
- **Alternative (H_1)**: The average age differs between survivors and non-survivors

---

7. Hypothesis Testing & Significance Analysis

Hypothesis 1: Gender and Survival

Test Used: Chi-Square Test of Independence

A contingency table was created:

Survived	Female	Male
0	81	468
1	233	109

Since the p-value is extremely small, we reject the null hypothesis.

✅ **Gender is significantly associated with survival.**

Hypothesis 2: Passenger Class and Survival

Test Used: Chi-Square Test of Independence

We evaluated the relationship between Pclass and Survived.

Survived	1st Class	2nd Class	3rd Class
0	**80**	**97**	**372**
1	**136**	**87**	**119**

- **Chi-Square Statistic**: ~102.89
- **Degrees of Freedom**: 2
- **p-value**: < 1e-22
- **Significance Level**: α = 0.05

Conclusion: The p-value is extremely small, so we reject the null hypothesis.
✅ **Passenger class is significantly associated with survival.** Higher-class passengers had higher survival rates.

Hypothesis 3: Age and Survival

Test Used: Independent Two-Sample t-test (Welch's t-test)

We tested whether the average age differs between survivors and non-survivors.

- **Mean Age (Survived = 1)**: ~28.34
- **Mean Age (Survived = 0)**: ~30.63
- **t-statistic**: ~2.18
- **p-value**: ~0.029
- **Significance Level**: $\alpha = 0.05$

Conclusion: Since p-value < 0.05, we reject the null hypothesis.
✅ **There is a statistically significant difference in average age between survivors and non-survivors.** Survivors tend to be slightly younger.

---

8. Conclusion & Next Steps

Summary:

This project demonstrated how structured exploratory analysis and hypothesis testing can provide strong insights into survival patterns. We found gender, class, and traveling status to be strong predictors.

Recommended Next Steps:

- Apply logistic regression to model survival probability
- Use advanced feature engineering (e.g., extracting titles from Name)
- Test additional hypotheses using t-tests or ANOVA
- Train and evaluate classifiers like decision trees or random forests

---

Final Thoughts

This analysis showcased the ability to extract and communicate data-driven insights. The dataset is now clean, encoded, and well-prepared for predictive modeling. This work would serve as a foundation for deployment in real-world applications or as a demonstration of data science skills to a senior audience.