**Title: Titanic Survival Analysis: EDA and Insights**

**Author: RISHABH DHOUNDIYAL Date: 11 Aug. 2025**

1. **Objective**

* To explore the Titanic training dataset, perform basic data cleaning, and generate descriptive insights about survival patterns across key variables such as class, sex, age, fare, and embarkation port.

1. **Dataset Description**

* Source: Titanic training dataset (commonly used in Kaggle’s Titanic competition).
* Rows: 891 passengers.
* Columns:
  + Target: Survived (0 = No, 1 = Yes).
  + Features: Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, PassengerId.

1. **Data Loading and Initial Preview**

* Action: Loaded the dataset into a pandas DataFrame and displayed the first few rows.
* Observation:
  + Typical entries include personal details (Name, Sex, Age), travel class (Pclass), family-related counts (SibSp, Parch), ticket info, fare, cabin (often missing), and embarkation port.
  + Example rows confirm presence of missing ages (e.g., “Moran, Mr. James” has Age = NaN) and missing cabins.

1. **Data Cleaning: Age Handling**

* Action: Converted Age to integers in the displayed sample (NaN replaced by a plausible value in preview: 29 shown for the previously missing age). While the exact imputation logic is not shown, common approaches include:
  + Median imputation overall.
  + Median by Pclass or Sex.
  + Age binning.
* Result:
  + Age appears as integers in the second preview (22, 38, 26, 35, 35, 29).
  + Note: Clearly document your chosen imputation method in the notebook; for reproducibility, prefer median by Sex and Pclass or overall median.

1. **Descriptive Statistics (Numeric Variables)**

* Action: Computed DataFrame describe() for selected numeric columns.
* Key Stats (based on shown output):
  + PassengerId: Mean 446, range 1–891 (index surrogate, not informative for modeling).
  + Survived: Mean 0.384 → 38.4% overall survival rate.
  + Pclass: Mean 2.309 (more passengers in lower classes), range 1–3.
  + Age: Mean 29.43, median 29, IQR approx. 22–35, min 0, max 80.
  + SibSp: Mean 0.523, most passengers traveled with 0–1 siblings/spouses.
  + Parch: Mean 0.382, most passengers with 0 parents/children.
  + Fare: Mean 32.20, median 14.45, IQR 7.91–31, highly skewed with max 512.33.

1. **Target Distribution: Survived**

* Action: Value counts for Survived.
* Result:
  + 0 (did not survive): 549
  + 1 (survived): 342
  + Survival rate ~38.4%.
* Insight:
  + The dataset is moderately imbalanced toward non-survivors.

1. **Passenger Class Distribution: Pclass**

* Action: Value counts for Pclass.
* Result:
  + 3rd class: 491
  + 1st class: 216
  + 2nd class: 184
* Insight:
  + Majority traveled in 3rd class. Class is known to be a strong predictor in Titanic survival analyses, with higher classes typically showing higher survival.

1. **Gender Distribution: Sex**

* Action: Value counts for Sex.
* Result:
  + Male: 577
  + Female: 314
* Insight:
  + More male passengers than female.
  + Historically, “women and children first” influenced outcomes; we expect higher survival among females.

1. **Embarkation Port Distribution: Embarked**

* Action: Value counts for Embarked.
* Result:
  + S (Southampton): 644
  + C (Cherbourg): 168
  + Q (Queenstown): 77
* Insight:
  + Most passengers embarked at Southampton.
  + Embarked may correlate with class and fare, and thus indirectly with survival.

1. **Missing Data Overview**

* Observed Missingness:
  + Age: Missing in original data (e.g., “Moran, Mr. James”). You addressed this in the preview with imputation.
  + Cabin: Frequently missing (NaN appears often).
  + Embarked: Mostly present; counts sum to 889 (644+168+77) indicating 2 missing originally (common in this dataset).
* Implications:
  + Age imputation is necessary for most models.
  + Cabin can be used as a binary flag (HasCabin vs. NoCabin) or omitted due to sparsity.
  + Ensure Embarked missing values are imputed (mode “S” is common) if used in modeling.

1. **Key Findings and Insights**

* Survival Rate: Approximately 38%.
* Class Imbalance: More passengers in 3rd class; anticipate lower survival in 3rd class and higher in 1st class.
* Gender Effect: Expect female survival to be significantly higher than male, consistent with historical accounts.
* Age Distribution: Median age is 29; consider age groups (children, adults, seniors) when analyzing survival.
* Fare Skewness: Fare is skewed with outliers; consider log-transform for modeling and use robust statistics for EDA.
* Embarked: Majority from Southampton; embarkation may correlate with class/fare.

1. **Suggested Next Steps (if building a model)**

* Feature Engineering:
  + Title extraction from Name (Mr, Mrs, Miss, Master) to improve age imputation and predictive power.
  + FamilySize = SibSp + Parch + 1 and IsAlone = (FamilySize == 1).
  + HasCabin flag = ~Cabin.isna().
  + Ticket group size or ticket prefixes (optional).
  + Binning: Age bands (e.g., child < 16), Fare bands or log(Fare+1).
* Encoding:
  + One-hot encode Sex and Embarked; ordinal-encode Pclass if appropriate.
* Imputation:
  + Age: median by Title and Pclass or by Sex and Pclass.
  + Embarked: mode (often “S”).
  + Fare: median by Pclass and Embarked if any missing.
* Modeling:
  + Start with baseline models: Logistic Regression, Random Forest, Gradient Boosting.
  + Evaluate with cross-validation; use accuracy, F1, ROC-AUC.
* Validation:
  + Check class imbalance. Consider class weights or balanced metrics.
* Interpretation:
  + Use feature importance (tree models), coefficients (logistic), and partial dependence to understand drivers.

1. **Limitations**

* The report is based on the visible outputs; ensure the notebook contains the exact preprocessing steps (especially for Age imputation) and code for reproducibility.
* No visual plots included here; adding histograms, bar plots, and survival rate comparisons by Sex, Pclass, and Embarked will strengthen the analysis.

1. **Conclusion**

* The dataset shows a 38% survival rate with strong expected effects of Sex and Pclass.
* There is missingness in Age and Cabin that must be handled carefully.
* Basic EDA supports known Titanic patterns: higher-class and female passengers likely had higher survival rates.
* With targeted feature engineering and appropriate modeling, solid predictive performance is achievable.