**Executive ReportSummary**

**This exploratory data analysis (EDA) on the Titanic dataset (891 passengers, 12 columns) investigates factors that influenced survival. Key findings:**

* **Overall survival rate: ≈ 38.38% (342 survivors out of 891).**
* **Strong predictors of survival: Sex (female advantage) and Passenger Class (Pclass) (1st class had higher survival).**
* **Other factors: Fare (higher fares → higher survival), age (children slightly more likely to survive), family size (mixed effect).**
* **Data quality note: Age has ~19.9% missing values; Cabin is missing for ~77% of rows — handle these before modeling.**

**1. Dataset Overview**

* **Rows:** 891
* **Columns:** 12 (PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked)
* **Missing values (major):**
  + Cabin: 687 missing (~77%)
  + Age: 177 missing (~19.9%)
  + Embarked: 2 missing (~0.2%)

**Key numeric summaries (selected):**

* Average age: **~29.7 years**
* Fare mean: **32.20** (right-skewed; large outliers)
* SibSp mean: **0.52**, Parch mean: **0.38**

**2. Methods / Code (what was run)**

The following steps were performed:

1. Load the data with pandas.read\_csv('train.csv').
2. .info(), .describe(), and .value\_counts() for categorical variables.
3. Visualizations: sns.pairplot(), sns.heatmap(), histograms, boxplots, and scatterplots.
4. Grouped statistics: survival rates by Sex, Pclass, Embarked.
5. Saved figures and exported a PDF report with findings.

(See the “Reproducible code” section below for the exact code.)

**3. Findings (Detailed)**

**3.1 Overall survival**

* **Survival rate:** **≈ 38.38%**  
  (A minority of passengers survived; survival is the target variable.)

**3.2 Survival by Sex**

* **Female survival rate:** **~74.2%**
* **Male survival rate:** **~18.9%**  
  **Interpretation:** Sex is a very strong predictor — females were prioritized for rescue.

**3.3 Survival by Passenger Class (Pclass)**

* **Pclass 1 survival:** **~62.96%**
* **Pclass 2 survival:** **~47.28%**
* **Pclass 3 survival:** **~24.24%**  
  **Interpretation:** Higher socio-economic status (1st class) had a large survival advantage.

**3.4 Fare and Survival**

* **Correlation:** Fare positively correlates with survival (higher fare → higher survival probability).
* **Distribution:** Fare is right-skewed with extreme outliers (very wealthy passengers).

**3.5 Age and Survival**

* **Age distribution:** Most passengers between 20–40.
* **Children:** Higher relative survival for very young passengers (consistent with “women and children first”).
* **Overall correlation:** Age shows a weak negative correlation with survival, but patterns appear when combined with Pclass and Sex.

**3.6 Family Size (SibSp + Parch)**

* **Most passengers traveled alone** (SibSp=0 & Parch=0).
* Family size effect is mixed: very large families tended to have lower survival (logistical difficulty) while small families sometimes benefited.

**3.7 Embarked**

* Small differences observed by embarkation port (S, C, Q) — but this is entangled with Pclass and Fare.

**4. Visualizations (descriptions)**

(These visuals were created and inserted into the PDF. Description of what they show.)

* **Pairplot (sample of numeric features):** shows relationships between Survived, Pclass, Age, SibSp, Parch, Fare. Clear separation by Pclass and Fare.
* **Correlation heatmap:** highlights negative correlation between Pclass and Survived (because Pclass=1 is lower number but higher survival), positive correlation between Fare and Survived.
* **Histograms:** Age (peak 20–40), Fare (right skew), SibSp/Parch (most zeros).
* **Boxplots:** Age and Fare vs Survived show survivors cluster at higher fare and slightly lower age.
* **Scatterplot (Age vs Fare colored by Survival):** survivors cluster at higher fares and many young ages.

**5. Conclusions & Recommendations**

* **Conclusion:** Sex (female) and socio-economic status (Pclass, Fare) were the strongest factors associated with survival. Age and family structure contributed but are less decisive alone.
* **Data cleaning needed:** Impute Age (median or model-based imputation), engineer Title from Name, create FamilySize = SibSp + Parch + 1, and extract CabinDeck if possible from Cabin. Consider dropping or carefully encoding Cabin given heavy missingness.
* **For modeling:** Use Sex, Pclass, Fare (log-transform), engineered Title, FamilySize, Age (imputed), and Embarked — these give strong predictive power for classification models.