American International University – Bangladesh
Faculty of Computer Science
Department of CSE
Course Name: Introduction to Data Science
Session: Summer 2024-25
Section: D
Group No: 01

Report's Title: Understanding and Preparing Titanic Data for Analysis Using R.

Course Teacher: Dr. Ashraf Uddin

Grouped By

SL No	Name	ID
1	MAHIN SARKAR ATUL	22-47372-2
2	MOHAMMAD RAKIBUL HASAN RIFAT	22-47371-2
3	SUNIPUN SEEMANTA	22-47547-2
4	MAHBUBUR RAHMAN	22-48153-2

1. Introduction

The Titanic dataset is one of the most well-known datasets in data science, originating from the tragic sinking of the RMS Titanic in 1912. It contains information on passengers such as demographic details, ticket class, fare, and survival status.

The primary goal of this analysis was to explore factors influencing passenger survival, focusing on demographic (e.g., age, sex), socio-economic (e.g., passenger class, fare), and situational variables (e.g., family size, embarkation port). The analysis aimed to:

- Identify and address data quality issues.
- Explore patterns and relationships between variables.
- Generate insights that can help explain survival probabilities.

Dataset description:

Col Name	Description
PassengerId	Unique identifier assigned to each passenger. Serves as a row index in the dataset.
Survived	Survival indicator (0 = Did not survive, 1 = Survived).
Pclass	Passenger ticket class (1 = First class, 2 = Second class, 3 = Third class). Acts as a proxy for socio-economic status.
Name	Full name of the passenger, often including titles (e.g., Mr., Mrs., Miss, Master) which can give insights into gender, age, and social standing.
Sex	Biological sex of the passenger (male, female).
Age	Age of the passenger in years (can include fractions for infants). Missing for some passengers.
SibSp	Number of siblings or spouses aboard the Titanic with the passenger.
Parch	Number of parents or children aboard the Titanic with the passenger.
Ticket	Ticket number (alphanumeric). May contain duplicates if passengers traveled together.
Fare	Price of the passenger's ticket in British pounds.
Cabin	Cabin number where the passenger stayed. Missing for most entries. First letter often indicates the deck.
Embarked	Port of embarkation: C = Cherbourg, Q = Queenstown, S = Southampton.

2. Data Cleaning Process

2.1 Missingness & invalid values (raw scan) : The missingness bar chart for df_raw revealed heavy gaps in *Cabin*, notable gaps in *Age*, and minimal gaps in *Embarked*, while a numeric correlation heatmap gave a quick overview of numeric relationships before cleaning. We also flagged unrealistic values—such as non-positive fares, implausible ages (≤0 or >100), and negative SibSp/Parch—producing a tidy summary table with example cases.

2.2 Standardizing categories: We normalized case: Sex → lower, Embarked → upper; coerced invalid Embarked values to NA, then imputed with **mode** (the most frequent port).

2.3 Handling Missing Values: Embarked was standardized to uppercase with missing values filled by mode ("S"), Fare gaps or zeros were replaced with class-specific median fares, and missing or unrealistic ages were imputed with median age by class and sex.

2.4 Outlier Treatment & transformations : We winsorized Fare and Age using the IQR rule (Fare_w, Age_w), applied log1p to Fare for skew reduction (Fare_log), and scaled Age for comparability and visualization (Age_scaled).

2.5 Feature Engineering : New variables were created to enhance analysis: New features include FamilySize (SibSp + Parch + 1), IsAlone (1 if alone, else 0), FarePerPerson (Fare ÷

FamilySize), Title (extracted from names to indicate social role), and AgeGroup (binned into Child, Teen, Young Adult, Adult, Senior).
2.6 Encoding & Scaling: Using fastDummies::dummy_cols(), we created one-hot columns for **Sex, Embarked, Pclass, Title, AgeGroup** with remove_first_dummy=TRUE (to avoid perfect multicollinearity). We appended z-score versions of selected numeric columns (Age_z, Fare_z, Age_w_z, Fare_w_z, FamilySize_z, FarePerPerson_z) to facilitate correlation and pair plots.

```
 [1] "Survived"          "Age"               "SibSp"             "Parch"             "Fare"
 [6] "Fare_w"            "Age_w"             "FamilySize"        "IsAlone"           "FarePerPerson"
[11] "Sex_male"          "Embarked_Q"        "Embarked_S"        "Pclass_2"          "Pclass_3"
[16] "Title_Miss"        "Title_Mr"          "Title_Mrs"         "Title_Nobleman"    "Title_Noblewoman"
[21] "Title_Officer"     "AgeGroup_Teen"     "AgeGroup_YoungAdult" "AgeGroup_Adult"  "AgeGroup_Senior"
[26] "Age_z"             "Fare_z"            "Fare_w_z"          "Age_w_z"           "FamilySize_z"
[31] "FarePerPerson_z"
```

These are the columns after encoding with feature engineering

3. Exploratory Data Analysis (EDA) & Results : We ran two EDA passes:
(A) On the raw dataset (df_raw) to visualize real-world messiness
(B) On the cleaned/engineered dataset (df / df_encoded) for clearer patterns.

3.1 Raw EDA (before cleaning) :
Missing/Empty Counts (bar chart): Cabin dominates missingness; Age has notable NAs; Embarked a few NAs/empties. ***(Figure-1)***
Age histogram (binwidth=5): Most passengers are **20–40**; tails include children and seniors. ***(Figure-2)***
Fare histogram (bins=20): Heavily right-skewed; most tickets cost **< £50**, with a long high-fare tail. ***(Figure-3)***
Survival by Sex (bar): Females show visibly higher survival counts than males. ***(Figure-4)***
Fare by Pclass (boxplot): Class-based fare separation is extreme (1st ≫ 2nd > 3rd). (***Figure-5)***
Raw numeric correlation heatmap: **Pclass** has a moderate negative correlation with both Fare (-0.56) and Age (-0.41), meaning higher-class passengers tend to pay more and be older. **Fare** has a slight positive correlation with Survived (0.23), indicating higher fares are linked to better survival chances. **SibSp** and Parch are moderately correlated (0.37), showing families often traveled together. ***(Figure-6)***
Survival rate matrices (tiles): Embarked × Sex: Rates differ by port; Cherbourg (C) + Female tends to be higher, consistent with more 1st-class passengers boarding there. ***(Figure-7)***

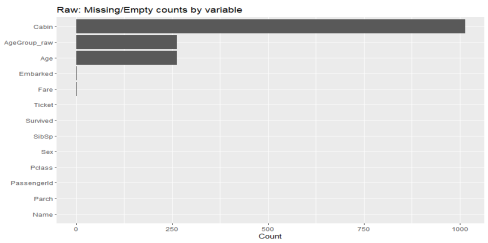


Figure-1

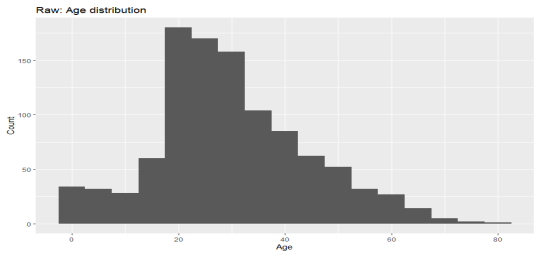


Figure-2

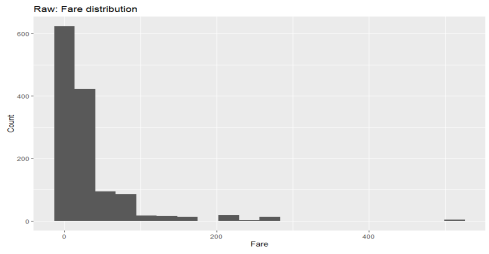


Figure-3

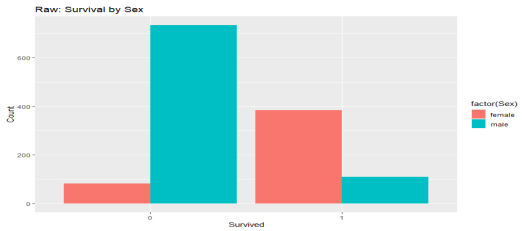


Figure-4

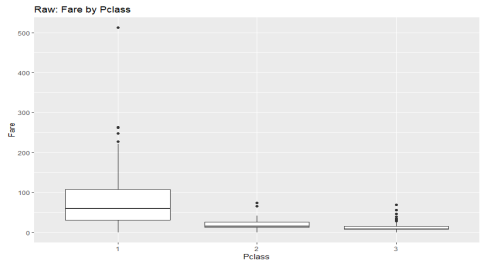


Figure-5

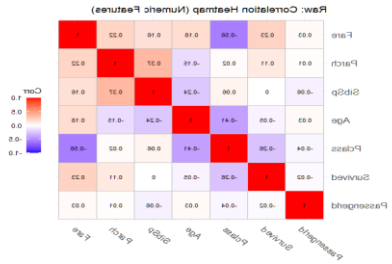


Figure-6

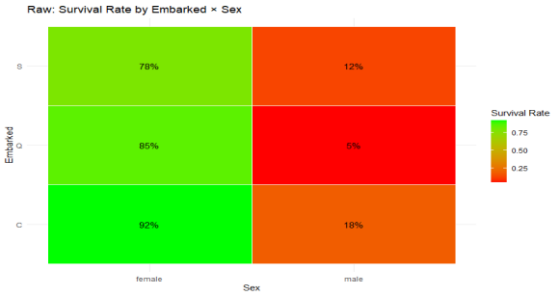


Figure-7

3.2 Cleaned/Engineered EDA (after preprocessing)

A) Distributions & outliers (after winsorization/transform)

- **Boxplots (before vs after)** for **Age** and **Fare** show that winsorization reduces extreme whiskers while preserving central tendencies. ***(Figure-8)***

B) Survival relationships

Survival by Sex (proportional bar): Females' survival odds exceed males' substantially across the full sample. ***(Figure-9)***

Survival by IsAlone (proportional bar): Solo travelers (IsAlone=1) show lower survival proportion than those traveling with family. (***Figure-10)***

Survival by Title (proportional bar): Highest survival rates: *Noblewoman*, *Mrs*, and *Miss* — the majority of these passengers survived. Interpretation: Title captures gender/age/social role signals embedded in names. ***(Figure-11)***

C) Continuous vs survival (density)

Age density by survival: Survivors slightly skew younger to mid-age, but with overlap — age alone is not decisive. ***(Figure-12)***

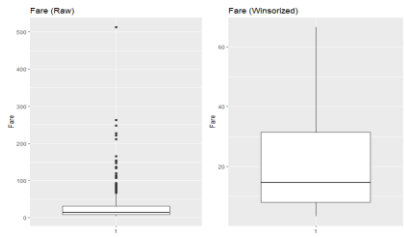
Fare density by survival: Survivors are shifted toward higher fares — an economic advantage signal***. (Figure-13)***

D) Correlations & multivariate

Correlation heatmap (numeric, encoded): Many features show near-perfect correlations with their scaled/weighted versions (e.g., Age–Age_z, Fare–FarePerPerson), with FamilySize strongly linked to SibSp, mostly strong positives, rare strong negatives, and redundancy suggesting potential dimensionality reduction. ***(Figure-14)***

Bubble Graph (survival chances on the Titanic varied greatly by passenger class and age group): The chart shows clear survival patterns on the Titanic, with children in Second Class having the highest survival rate, young adults in Third Class being numerous but with low survival, seniors in lower classes faring worst, First Class adults doing better than lower classes yet below children in some cases, and an overall trend of decreasing survival from Class 1 to 3 across most age groups. ***(Figure-15)***

E) Survival matrices (cleaned) : Overall, younger passengers and higher classes had better survival chances, with Class 1 outperforming others across ages, Class 1 teens (81%) and Class 2 children (92%) leading, and seniors in Classes 2–3 having the lowest survival (17%). ***(Figure-16)***



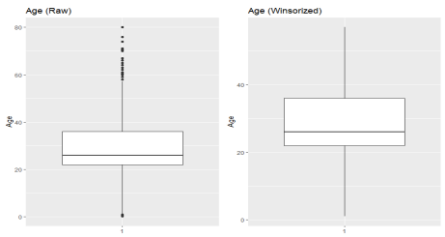


Figure - 8

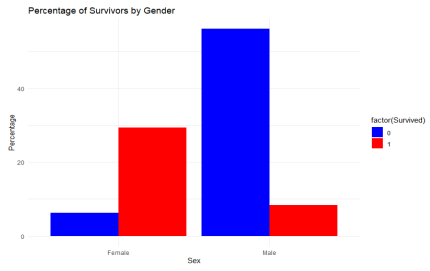


Figure - 9

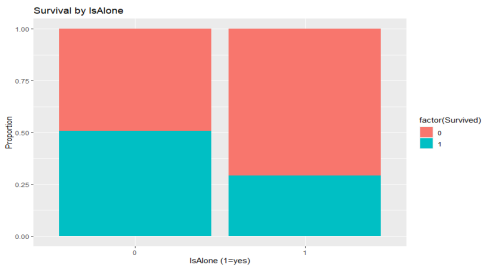


Figure - 10

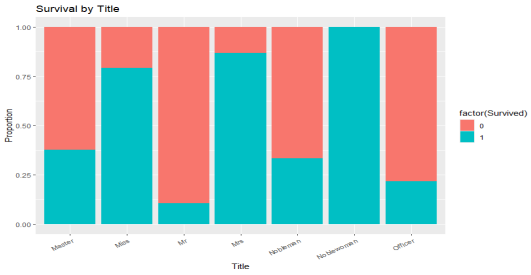


Figure - 11

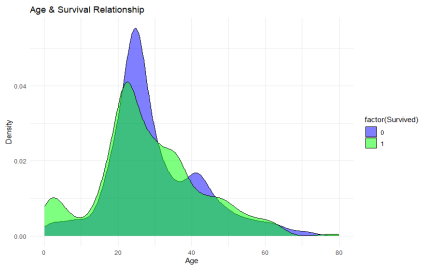


Figure - 12

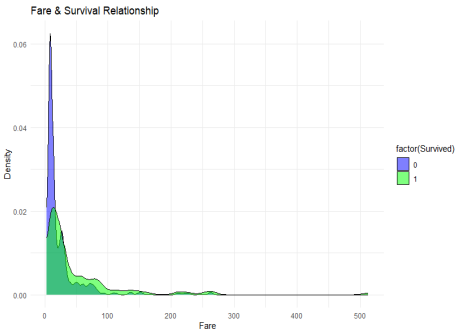


Figure - 13

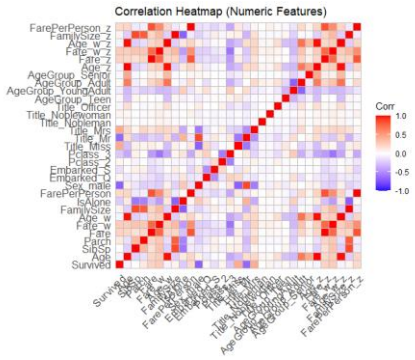


Figure - 14

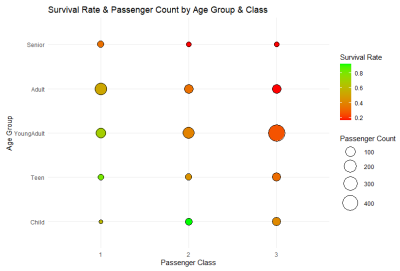


Figure - 15

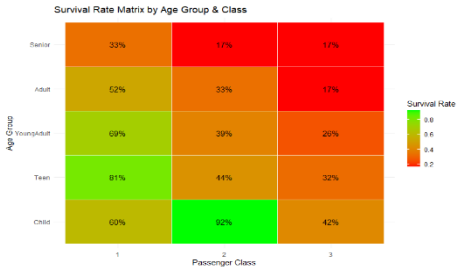


Figure - 16

4. Conclusion

In conclusion, Titanic survival was strongly influenced by socio-demographic and travel-related factors. Women, children, and higher-class passengers often paying higher fares had greater survival chances. Titles and companionship also played a significant role, with those traveling alone generally faring worse.

Limitations:

- The dataset lacks some potentially relevant details (e.g., exact location on ship, lifeboat assignments).
- The large number of missing cabin values limited the ability to analyze cabin location impact.
- Correlations do not imply causation — survival factors are influenced by historical and social context.