

Titanic Survival Analysis



Table of Contents

1. Project Description.....	2
1.1 Business Problem.....	2
1.2 Project Objectives.....	2
2. Data Exploration.....	2
2.1 Data Collection.....	2
2.2 Data Dictionary.....	3
2.3 Data Exploration and Visualisation.....	4
2.3.1 Survival Status (Target Variable)	4
2.3.2 Sex (Input Variable)	5
2.3.3 Passenger Class/Pclass (Input Variable)	5
2.3.4 Age (Input Variable).....	6
2.3.5 Embarked.....	7
2.3.6 FamilySize.....	8
2.3.7 Title (Input Variable)	9
2.3.8 Fare (Input Variable)	10
2.3.9 Deck (Dropped)	10
2.3.10 Ticket (Dropped)	11
2.3.11 Correlation Matrix.....	12
2.4 Challenges Encountered in Data Mining.....	12
3. Data Preprocessing.....	13
3.1 Data Cleaning.....	13
3.2 Feature Engineering.....	14
3.3 Data Transformation.....	14
3.4 Dimensionality Reduction.....	14
4. Appendix	
.....	Error!
Bookmark not defined.	
4.1 Work distribution table.....	Error! Bookmark not defined.

1. Project Description

1.1 Business Problem

The sinking of the RMS Titanic is a well-known historical event that resulted in the loss of many lives. Historical records show that some passengers were more likely to survive than others. The main problem for this analysis is to use the available passenger data to understand these patterns and identify the key factors that influenced a passenger's chance of survival. This initial stage of the project focuses on exploring and preparing the data, which will later be used to build a predictive model. This analysis is not for a commercial purpose but act as a case study in using data to find patterns in a major historical event. The target variable for this problem is Survived, which indicates if a passenger survived (1) or did not survive (0).

1.2 Project Objectives

This project has two main objectives. The first objective, which is also the focus of this report is to conduct a completely data exploration and preparation. This object involves understanding the data's characteristics, identifying any quality issues, cleaning the data, and creating some meaningful features (feature engineering). The goal of this first stage is to prepare a high-quality dataset for predictive modelling. The second objective, which will be to use the prepared data to build and evaluate several predictive models. These future models will predict passenger's survival and help to formally identify which passenger characteristics were the most influential. This report therefore lays the essential groundwork for that future modelling task.

2. Data Exploration

2.1 Data Collection

The public Titanic dataset is used, which contains 12 original attributes (e.g., PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked). This is the standard dataset for binary classification on survival outcomes and aligns with our project scope for exploratory analysis and feature engineering. The CSV file was uploaded to SAS Viya so that downstream and exploration/visualisation could run in-memory:

1. Upload: In SAS Drive/Studio, uploaded TITANIC.csv to our personal area.
2. Load to CAS: Right-click → *Load to CAS* (or use *Import Data*), target caslib: CASUSER (personal session) and table name: TITANIC.

3. Verify types: Confirmed numeric vs character assignments (e.g., Survived, Pclass, Age, SibSp, Parch, Fare numeric; Sex, Ticket, Cabin, Embarked, Name character).
4. Persist (optional): Saved the in-memory table to disk in the same caslib for reuse in the pipeline.

2.2 Data Dictionary

The dataset contains passenger records and a set of original as well as engineered attributes that describe demographic details, ticket information, and travel circumstances. The table (see Table 1 below) provides a summary of each variable, its data type, and its analytical importance. Furthermore, the summary statistics of numerical variables can be found in Table 2 below.

Attribute	Type	Role	Description	Example Value	Analysis
PassengerID	Nominal	Identifier	Unique identifier for each passenger	1	Serves only as a row index; no predictive value.
Survived	Binary	Target	Survival status	"Survived"	Target variable for classification.
Pclass	Ordinal	Input	Socio-economic class	"Upper"	Strong predictor of survival, represents social status.
Name	Text	Rejected	Full passenger name	"Allen, Miss. Elisabeth"	Non-predictive in raw form but used to derive Title.
Sex	Binary	Input	Gender of passenger	"Male"	Highly predictive of survival due to "women first" policy.
Age	Ratio	Input	Passenger's age in years	29	Follows the "women, baby, and old people first" narrative.
SibSp	Ratio	Rejected	Num of siblings/spouses aboard	1	Used with Parch to create FamilySize.
Parch	Ratio	Rejected	Num of parent/children aboard	3	Used with SibSp for FamilySize
Ticket	Nominal	Rejected	Ticket number	"A/5 21171"	Irregular and non-informative, dropped from modelling.
Fare	Ratio	Input	Price of ticket	7.25	Wide range with extreme outliers; transformed by binning.
Cabin	Nominal	Rejected	Cabin identifier	"C85"	Over 70% missing values; excluded from analysis.
Embarked	Nominal	Input	Port of embarkation	"S"	Initially thought to have predictive power but after further investigation, it is due to Pclass.
FamilySize	Ratio	Input	Size of family (including the passenger)	2	Feature engineered. FamilySize = SibSp + Parch + 1

Title	Nominal	Input	Title of the passenger	“Dr”	Feature engineered. High predictive power information about passengers
Deck	Ordinal	Rejected	Deck on the Titanic	“A”	Feature engineered from Cabin. Over 70% missing values; excluded

Table 1. Data dictionary

Interval Variable Moments									
Variable Name	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Relative Variability	Mean plus 2 SD	Mean minus 2 SD
Age	0.1700	80	29.5032	12.9052	0.5410	0.9699	0.4374	55.3137	3.6927
Fare	0	512.3292	33.2811	51.7415	4.3695	27.0497	1.5547	136.7641	-70.2019

Table 2. Summary statistics for numerical features

2.3 Data Exploration and Visualisation

This section presents an initial analysis of individual variables in the dataset to understand their distributions, data types, and summary statistics.

2.3.1 Survival Status (Target Variable)

The overall survival proportion is shown in the donut chart (See Figure 1 below). While the proportion of people who survived/perished is not exactly divide exactly 50-50, the dataset is still considered to be balance with approximately 61.7% of passengers perishing and only 38.3% surviving.

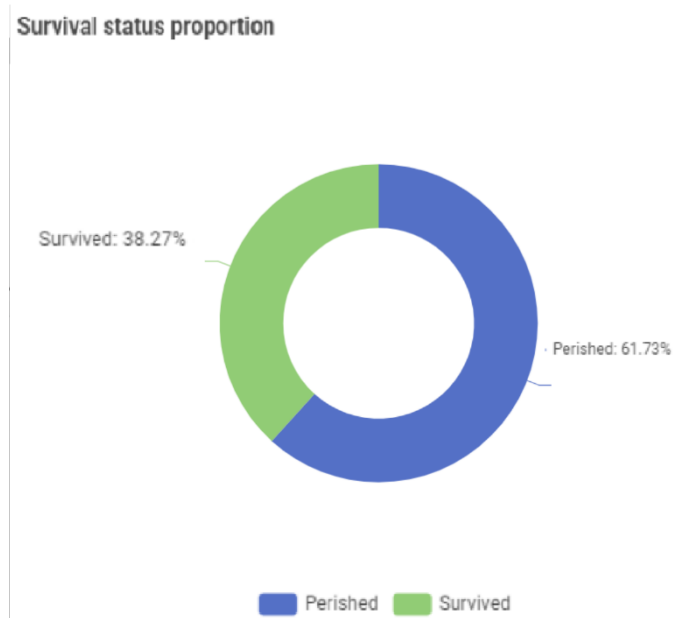


Figure 1. Survival status proportion

2.3.2 Sex (Input Variable)

Most of the passengers on the Titanic were predominantly male, with males making up 64.4% of the passengers compared to 35.6% for females (See Figure 2a below). The heatmap in Figure 2b supports the belief of “ladies first” as it can be clearly seen that most of those who died were male. Statistically speaking, the mortality rate of male is 3 times higher than female.

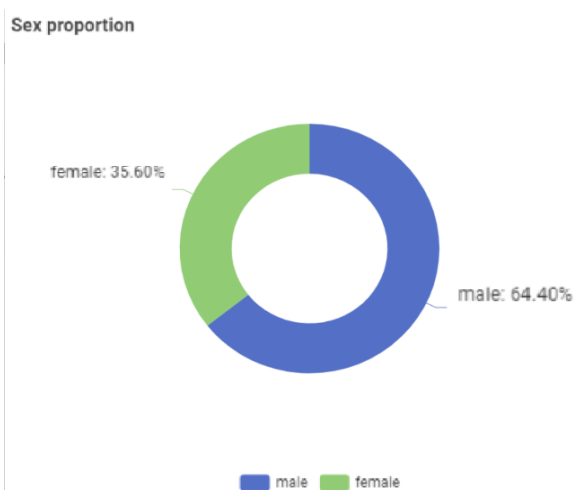


Figure 2a. Sex proportion

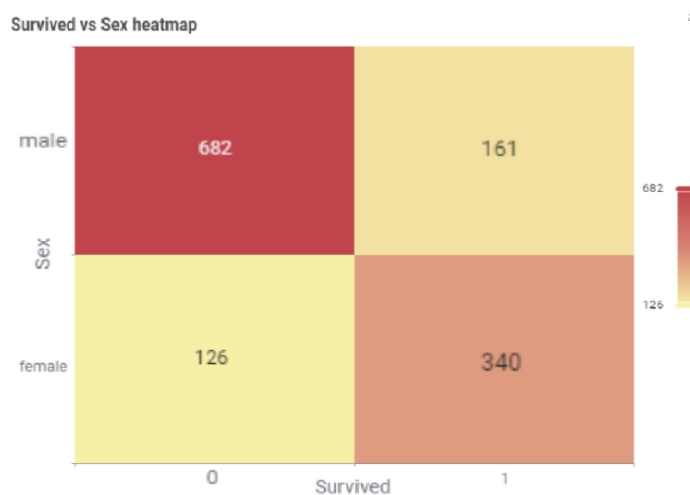


Figure 2b. Sex heatmap

2.3.3 Passenger Class/Pclass (Input Variable)

The pie chart (See Figure 3a below) shows that most passengers (54.2%) were in the Lower class. The Upper and Middle classes are more evenly distributed, at 24.7% and 21.2% respectively. Unsurprisingly, a passenger's class significantly influences their survival rate since high classes are given more priority as shown in the bar chart in Figure 3b below. This indicates that there is a strong social class bias in an individual's survival outcome. On average, an upper-class passenger is 2.5 times more likely to survive compared to their counterparts in the lower-class.

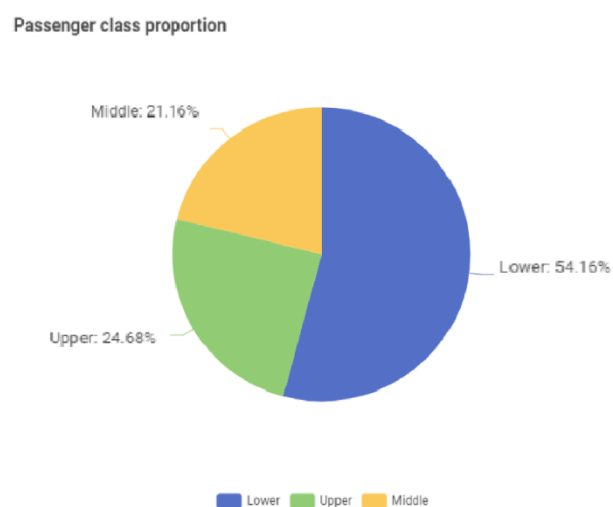


Figure 3a. Passenger classes proportion

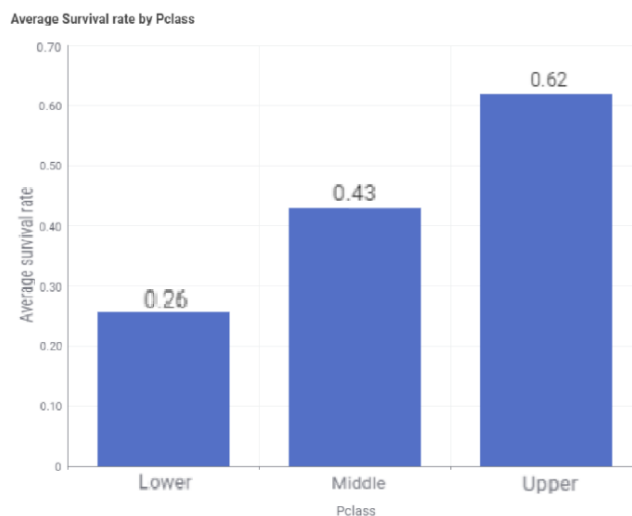


Figure 3b. Survival rate by Passenger class

2.3.4 Age (Input Variable)

The age of a passenger is also an extremely important factor that determines whether they survive as shown in Figure 4c and Figure 4d below. This further support the narrative of “women, children, and old people first” since passengers under 10 are approximately twice as likely to survive compared to other age groups. The second highest age group to survive are those in their 50s. The histogram for passengers’ age shows that the largest group of passengers were young adults between 20 and 30 years old, making up nearly half of all passengers (See Figure 4a below). The box plot confirms this, showing a median age in the late 20s (See Figure 4b below). It also highlights the presence of outliers since there are passengers on both ends of the spectrum as some passengers are as old as 80 years old whereas some are babies only a couple of months old. Usually, outliers are a cause for concern, however, the values of these data points are not wrong and are perfectly reasonable. There are indeed people who are very old (80 years old or older) and there are babies who were just born a few months ago that hasn’t even had their first birthday. As a result, for not perform any outlier imputation for now.

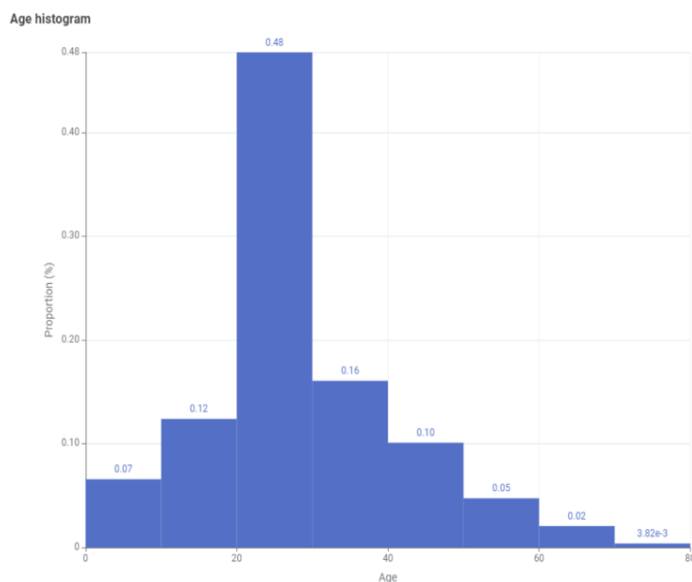


Figure 4a. Age histogram

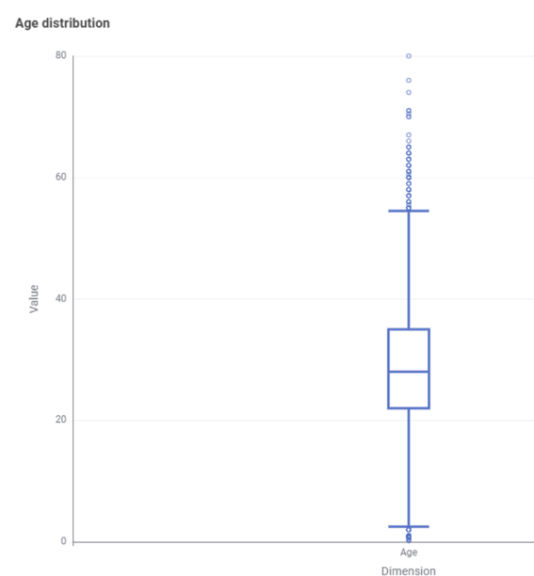


Figure 4b. Age boxplot

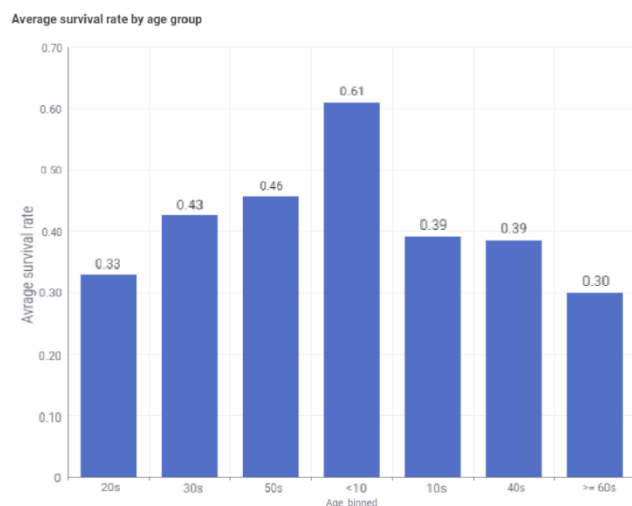


Figure 4c. Survival rate by age group

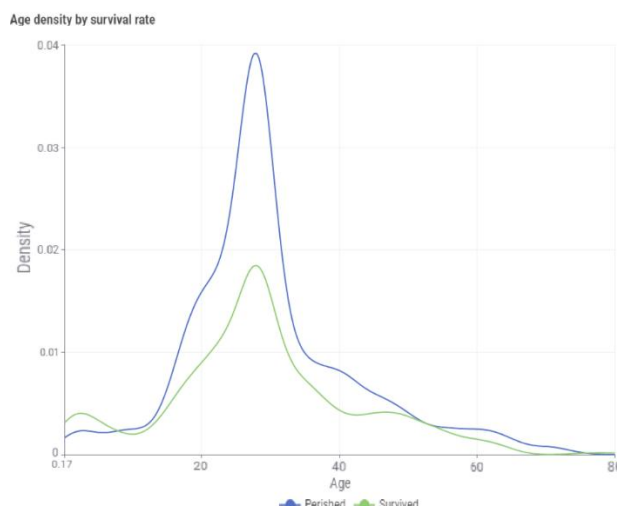


Figure 4d. Age density by survival status

2.3.5 Embarked (Input Variable)

The proportions of ports where the passengers got on the ship are shown in the pie chart (See Figure 5a below). Most passengers, about 70%, embarked at Southampton (S). Cherbourg (C) was the next most common port at around 20%.

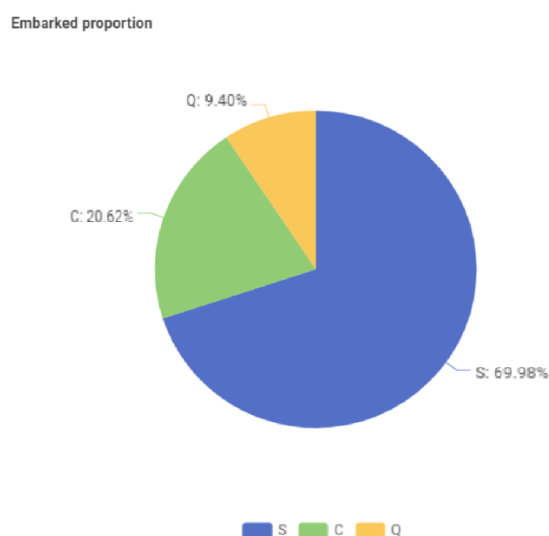


Figure 5a. Embarked proportion

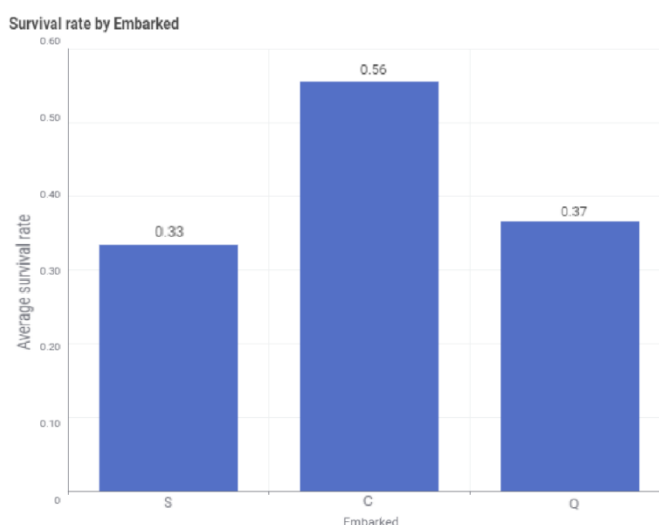


Figure 5b. Survival rate by Embarked

Despite Figure 5b above indicating that those who embarked at Cherbourg have the highest chance of survival and by a huge margin compared to the 2 other ports, we suspected that it shouldn't be the case. To confirm our hypothesis that regardless of which port the passengers embarked at, it should not affect the likelihood of survival, the relationship between Embarked with other features were investigated using a confusion matrix. Finally, Figure 5c below confirms that our hypothesis was reasonable. The reason why Figure 5b indicates that those who embarked at port Southampton have the lowest average survival rate turns out to be because most people embark at port Southampton and most of those people are in the

lower class. As above, the report already uncovered the social bias that influences passenger's survival rate where upper- and middle-class passengers are more likely to survive compared to their lower-class counterparts. Similarly, the reason why those who embarked at Cherbourg have a higher survival rate isn't due to Cherbourg having any special properties, but it is since most of the people who embarked at Cherbourg are in the upper- and middle-class.

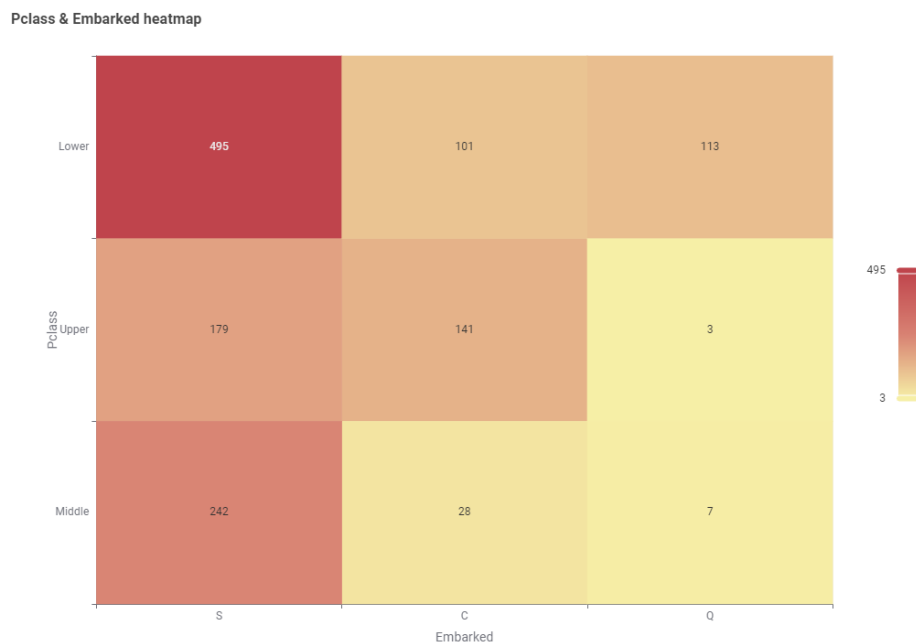


Figure 5c. Relationship between Embarked and Pclass

2.3.6 FamilySize (Input Variable)

Originally, the feature **FamilySize** is not present in the dataset because it is the result after perform some feature engineering by combining **SibSp** and **Parch** into a new variable using the following formula ($FamilySize = SibSp + Parch + 1$). The reason for adding 1 is because we must also be included. For example, if a passenger has $SibSp = 1$, this means they are going with 1 other person who is either their sibling or spouse but after including themselves, there are a total of 2 people. The distribution of family size (See Figure 6a below) shows that most passengers were travelling alone. As expected, the bigger the family size the less frequent it is to occur. Unsurprisingly, passengers with a larger family size can be seen to have a higher survival rate on average, however, this is only true up to a certain limit which seems to be 4. After that, the survival rate plummets. The reason why this seems reasonable is because having a relationship with someone compared to just being alone makes it harder to be left behind. For example, when $FamilySize = 2$ which includes a father with his new born child, then even though we discussed above that males are more likely to be perished compared to females, but since the father is with a baby, there is a very high chance that he will be able to escape onto the limited life boats since leaving the baby alone is very unlikely.

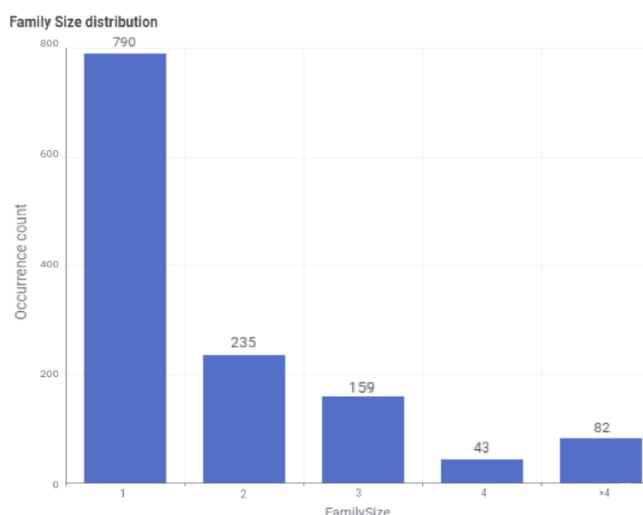


Figure 6a. Family size distribution

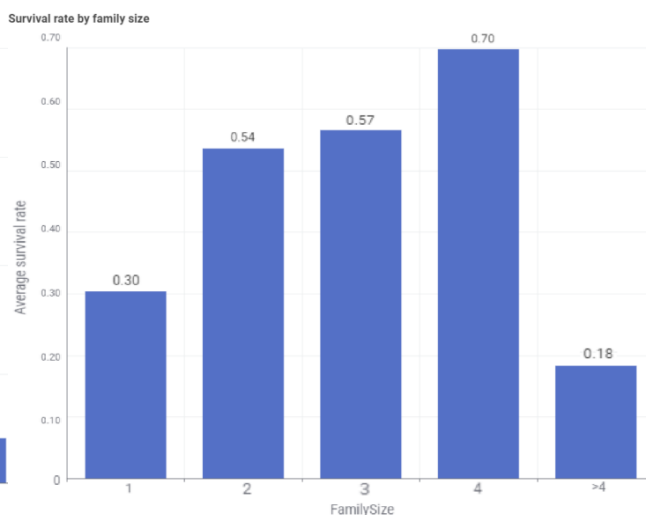


Figure 6b. Survival rate by family size

2.3.7 Title (Input Variable)

Like the **FamilySize** feature, the Title feature is an engineered variable that was created from the Name variable. The justification for this is that a passenger's full name offers little predictive value, but their title (e.g., "Master", "Lady", "Dr") provides important information. A title often indicates both a passenger's sex and their social standing, making it a potentially powerful indicator for survival. The bar chart in Figure 7a shows the distribution of the different titles. It confirms our earlier finding that most passengers were male, as "Mr" is the most frequent title. It is also important to note that very rare titles, such as "Capt" or "the Countess," were grouped into a single "Rare" category to create more stable groups for analysis. The bar chart in Figure 7b confirms the strong relationship between a passenger's title and their survival rate. For example, titles associated with women ("Mrs", "Ms") show very high survival rates, while the "Mr" title has a very low rate. An interesting finding is that no passenger with the "Rev" (Reverend) title survived. It can be speculated that this might be due to the nature of their occupation, as they may have sacrificed their own safety to help others.

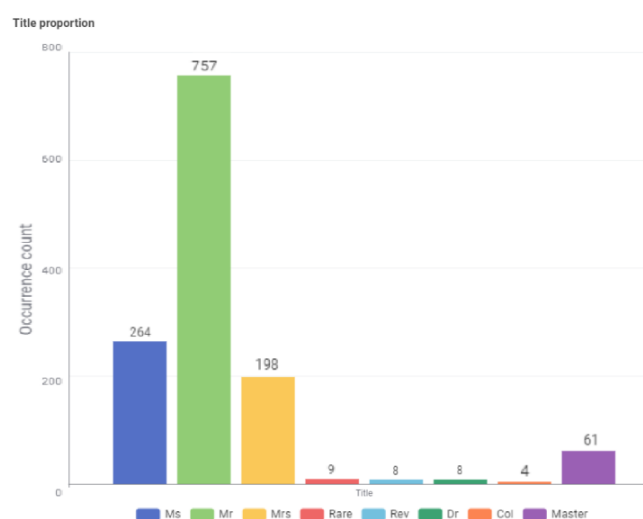


Figure 7a. Title distribution

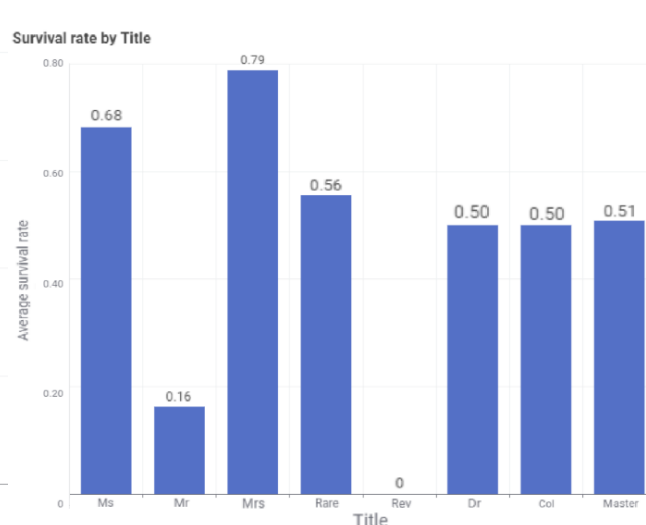


Figure 7b. Survival rate by Title

2.3.8 Fare (Input Variable)

The box plot below (see Figure 8) shows that Fares have a relationship with survival. This is to be expected since **Fare** is usually tied to a person's social standing. An ordinary person is highly unlikely able to pay for a huge fare for expensive cabins/passenger class whereas an influential social standing is highly likely to do so. It can be clearly observed that both the median and maximum fare for passengers who survived is higher than for those who perished. This suggests that passengers who paid more for their tickets had a better chance of survival.

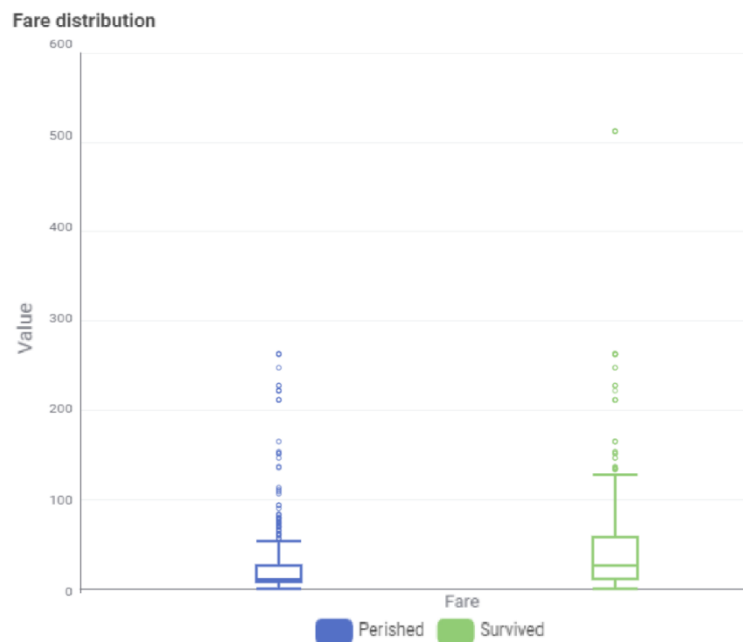


Figure 8. Fare distribution by survival status

2.3.9 Deck (Dropped)

Deck is also one of the features that has been engineered like **Title** or **FamilySize**. Deck is extracted from the **Cabin** feature since the cabin number is a combination of the deck level with the room number. For example, cabin A31 means room 31 in deck A. The deck layout of the RMS Titanic can be found in Figure 9a below. Intuitively, the deck level in which the passengers are in highly affects their survival rate since those in the lower decks or closer to the water surface will have less time to react to the incident compared to those in the upper decks. Unfortunately, this hypothesis cannot be confirmed as the proportion of missing values in the **Cabin** feature is over 70% leading to the **Deck** feature having the same overwhelming number of missing values (see Figure 9b below). With many missing values, imputation is not possible as it will only introduce noises into the dataset making it harder to train the ML models down the line. Replacing the missing values with the mode is also impractical since with such a huge number of missing values; after replacing them with the mode, the proportion of mode will only be further inflated. With the issues above, we justified that it is appropriate to drop/reject this feature for ML modelling.

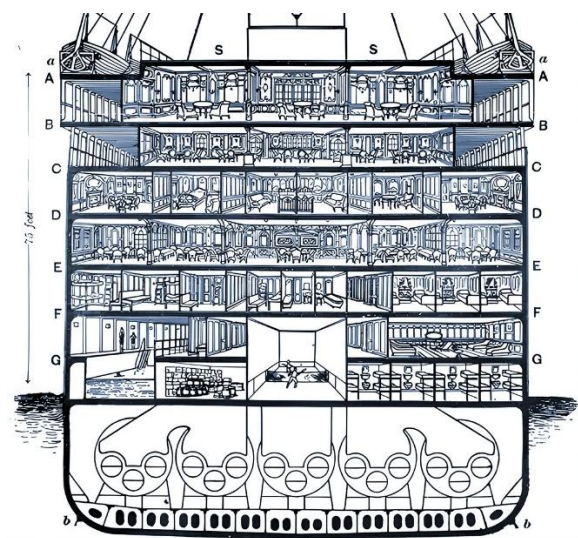


Figure 9a. RMS Titanic deck layout

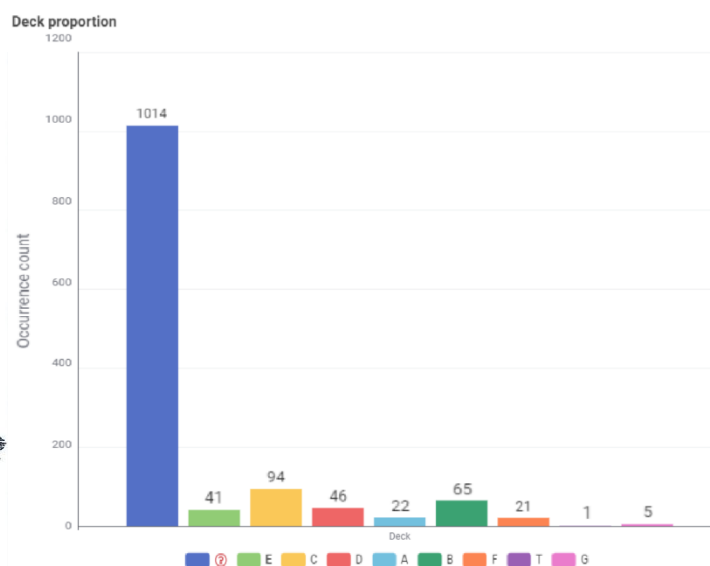


Figure 9b. Deck proportion

2.3.10 Ticket (Dropped)

The Ticket variable contains the ticket numbers for each passenger. This feature is complex because it includes a mix of purely numeric values and alphanumeric codes with various prefixes (e.g., "PC 17599", "A/5 21171"). As shown in the pie chart in **Figure 10** below, an analysis of the ticket prefixes reveals that most tickets (73.1%) are simple numeric strings. A few prefixes like "PC" and "CA" appear frequently, but many others are unique or rare and have been grouped into the "Other" category. Due to the high number of unique values (high cardinality) and the inconsistent format, the Ticket variable is difficult to use directly in a predictive model. While the prefixes might contain information related to passenger location or class, this information is likely already captured more effectively by other variables such as **Pclass** and **Fare**. Therefore, because it adds more noise than predictive value, the Ticket variable was dropped from the final analysis.

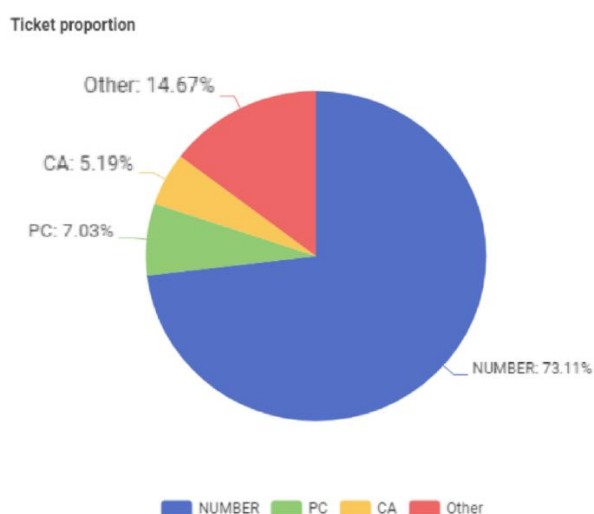


Figure 10. Ticket type proportion

2.3.11 Correlation Matrix

This section explores the relationships between different variables, with a focus on how they relate to the Survived target variable. The correlation matrix (see Figure 11 below) summarizes the relationships between variables. We can see a strong negative correlation (red) between Sex and Survived, and between Pclass and Survived. This confirms that being female (coded as a higher number) and being in a higher passenger class (coded as a lower number) are both associated with a higher chance of survival.

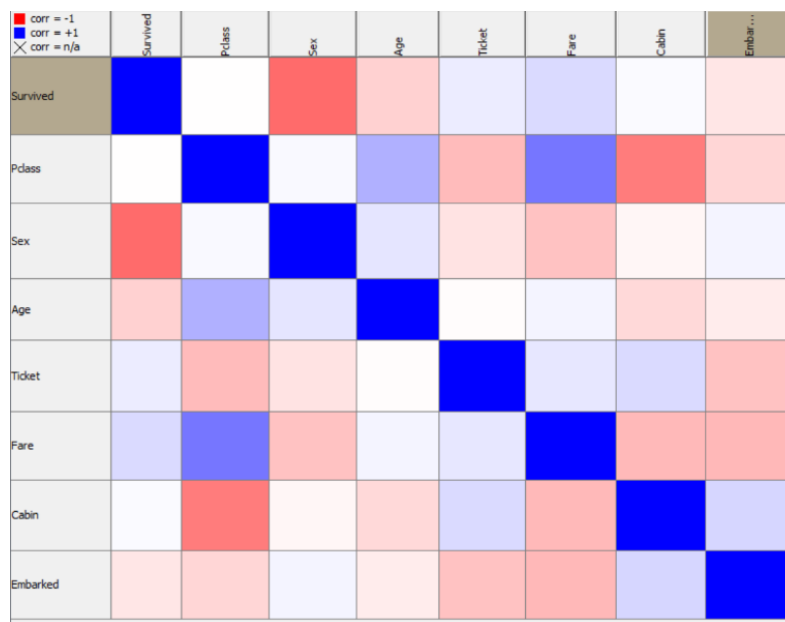


Figure 11. Correlation Matrix between features

2.4 Challenges Encountered in Data Mining

Our exploration surfaced four practical challenges. For each, we quantify the issue, propose the solution, and state the modelling impact to ensure replicability and rigour.

2.4.1 Missing Data

Issue: *Cabin* has >70% missingness and the engineered *Deck* inherits this pattern, making both unreliable for imputation at this sparsity.

Solution: Drop *Cabin/Deck* and keep a simple missingness indicator in exploration (not in final features) to test whether “absence of cabin info” carries signal.

2.4.2 Outliers and Skewness

Issue: *Fare* shows a long right tail with extreme values (>500), while *Age* has legitimate extremes (infants, >80).

Solution: Binned *Age/Fare* for the main, model-agnostic pipeline to stabilise splits and improve

interpretability. Keep *Age* extremes as valid historical cases.

2.4.3 Curse of Dimensionality

Issue: ID-like or weak-signal variables (e.g., PassengerId, Ticket, Cabin/Deck) inflate dimensionality without predictive gain; one-hot encoding also risks creating many sparse dummies.

Solution: Remove PassengerId/Ticket/Cabin/Deck and control dummy expansion via rare-level aggregation (pre-buckets for Title/Embarked).

2.4.4 Feature Engineering Complexity

Issue: Raw Name/SibSp/Parch are either weakly informative or duplicative; however, they embed social and family structure.

Solution: Engineer Title (from Name) and FamilySize ($= \text{SibSp} + \text{Parch} + 1$), then drop the raw sources to prevent redundancy. Verify confounding discovered in Section 2.3 (e.g., Embarked's apparent effect largely reflects Pclass composition) and avoid double-counting such pathways.

3. Data Preprocessing

3.1 Data Cleaning

Firstly, we performed data cleaning on the Embarked feature. While the Embarked feature itself does not have any data issues like missing values or format inconsistencies however, its values contain the feature name as a prefix which is quite redundant. After cleaning the Embarked feature, its values of [EmbarkedC, EmbarkedQ, EmbarkedS] becomes just [C, Q, S]. The most challenging feature to clean is the Ticket feature due to its high cardinality as well as the confusing format consistency. The first step for cleaning the Ticket feature is to convert every numeric ticket label like "110152" into a new label called "NUMBER". This significantly decreases the cardinality, from 929 unique labels down to only 225. This is appropriate because the number on a ticket often only serves as an identifier and does not hold any significance. We now deal with the tickets that consist of a mix of characters and numbers, and as mentioned above, a number on a ticket holds no significance so we just extract the character part of the ticket ("PC 17761" → "PC"). Next we resolve the case inconsistency by making every ticket label to be uppercase. Afterwards, we resolve the format inconsistency by removing every white space, ",", or "/" symbol because there are many instances where 2 tickets are very similar but just formatted differently. For example, ticket "W.E.P" and ticket "WE/P". After all the previous cleaning steps, we are left with only 36 unique labels, which is a huge improvement, but the issue of high cardinality persists. The Ticket feature can be further cleaned since there are many cases where we suspect that there was a human typo error. For example, there is a very high suspicion that the 4 labels "SOTONO2", "SOTONOQ", "STONO2", and "STONOQ" are the same label but is different due to a typo. This is possible because the location of the number "2" and the

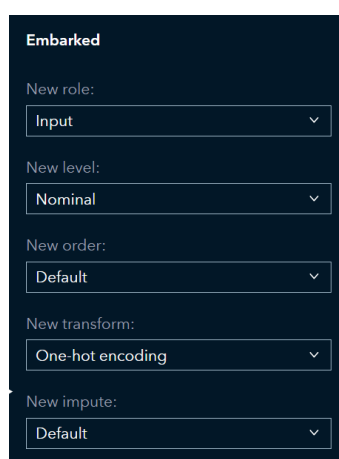
character “Q” on a keyboard are directly above/below each other. However, we do not have historical records to check if these really are due to human errors therefore, we decided to not continue cleaning the Title feature.

3.2 Feature Engineering

While a passenger’s name does not have any predictive power, but the title within a passenger’s name conveys both their sex as well as social standing. As a result, we deemed that it is necessary to extract the title from the passenger’s name for further analysis. Additionally, we decided to combine the features SibSp and Parch into a new feature called FamilySize because both essentially describes the same information but just with different role.

3.3 Data Transformation

Firstly, since ML models do not work with raw labels, for nominal input categorical features, we decided to perform one-hot encoding on to create a separate binary feature for each unique label as seen in Figure 12 below. For ordinal input categorical feature like **Pclass**, we decided to encode them with numerical values corresponding to their rank. Specifically, the key-value pairs encoded for the **Pclass** feature are as follows: *Lower* = 1, *Middle* = 2, *Upper* = 3. For the target feature “**Survived**” we also decided to encode them with numerical values where *Survived* = 1 and *Perished* = 0. While transforming the Survived feature is not necessary since they are our target labels and not input labels, but by assigning them numerical values, it makes it more convenient to extract insights from our dataset by allowing us to calculate the average survival rate. Lastly, for numerical features like Age and Fare, since these features have very big standard deviations as well as the presence of a lot of outliers, we decided to perform binning on them as seen in section 2.3.



The image shows a dark-themed user interface for configuring a feature named 'Embarked'. It contains five vertically stacked dropdown menus, each with a label and a selection box. The selections are as follows:

- New role:** Input
- New level:** Nominal
- New order:** Default
- New transform:** One-hot encoding
- New impute:** Default

Figure 12. One-hot encoding nominal features

3.4 Dimensionality Reduction

After performing exploratory data analysis, we decided to exclude the features Deck and Cabin first due to the overwhelming proportion of missing values (> 70%). Furthermore, after feature engineering, we dropped the features SibSp and Parch since we have already combined those into a single feature called

FamilySize, so they are no longer needed. The same reason applies to the feature Name. Finally, we decided to drop the Ticket feature due to high cardinality despite our best effort in data cleaning.