# **Analysis of Titanic Data**

The objective of this project is to leverage the open-source Titanic dataset to develop a comprehensive data exploration report. I will utilize SQL for exploratory data analysis, given that the data is structured in a database format.

I will implement a five-step framework that encompasses:

* Understanding the business context
* Understanding the technical context
* Analysing the tables and fields
* Creating research questions
* Addressing the research questions through data analysis

# **Understanding the Business Context**

When you get a dataset, it is important to understand the context and background of the database before performing any analysis. Otherwise, you will not able to comprehend and interpret the data correctly.

Consider answering the following questions below:

* what are these data for?
* why do we need this database?
* where are these data collected?

# **Understanding the technical context**

After understanding the context of the dataset, the next step is to examine the technical aspects of the dataset. Familiarity with these technical details will aid in interpreting the data and assessing its accuracy and reliability.

Consider answering the following questions below:

* how are these data collected?
* where are the sources of these data?
* what are the systems that use to modify these data?
* what are the error sources of this data?
* is the data complete? is there any missing pieces of data?

The data is sourced from Kaggle, a renowned repository for data science and free datasets. The titanic dataset on Kaggle is designed for users to apply machine learning techniques to develop predictive models determining which passengers will likely survive the Titanic shipwreck.

# **Analysing the Tables & Fields**

This step aims to identify the available data tables within the database and examine the types of data contained within them.

Some questions to consider are:

* how many tables do we have
* what are the tables representing
* what are the fields in the tables? what is the meaning of the field
* should I clean the data or ignore the missing columns

## Description of Data

## Variables descriptions:

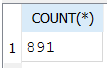
* sex — male or female
* survived — 1= yes; 0 = no
* ticket — ticket number
* fare — passenger fare
* age — age in years
* pclass — ticket class, 1= 1st class; 2 = 2nd class; 3 = 3rd class
* parch — number of parents/children aboard the Titanic
* sibsp — number of siblings/spouses abroad the Titanic
* name — name of the passengers
* passengerId — passenger ID
* cabin — cabin number
* embarked — ports of embarkation, C=Cherbourg; Q= Queenstown, S=Southampton

**Data inspecting & cleaning**

By utilizing the SQL COUNT function, we can ascertain the number of rows present in the ***passengers*** table.

SELECT COUNT(\*)

FROM passengers

****

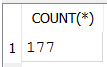
Upon reviewing the data, it is observed that the table contains a total of 891 rows, with several instances of missing data identified within the table. Among the various columns, there are 177 missing entries in the Age column, 687 missing entries in the Cabin column, and 2 missing entries in the Embarked column.

**Missing Entries – Age (177 missing values)**

SELECT COUNT(\*)

FROM passengers

WHERE Age IS NULL

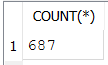


**Missing Entries – Cabin (687 missing values)**

SELECT COUNT(\*)

FROM passengers

WHERE Cabin IS NULL

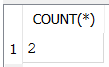


**Missing Entries – Embarked (2 missing values)**

SELECT COUNT(\*)

FROM passengers

WHERE Embarked IS NULL



# **Creating a Research Question**

This phase of the data analysis process involves generating insightful questions that can be answered using the available data.

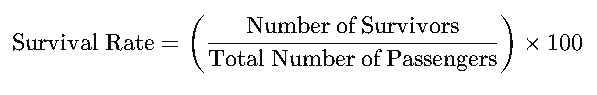
Upon reviewing the Titanic database, several intriguing questions have arisen:

1. What is the overall survival rate for the titanic passenger?
2. Are females more likely to survive this incident?
3. Are children and elderlies have a higher survival rate?
4. Do rich people have a higher survival rate because they get onboard to rescue boats sooner?
5. Which cabin has the highest survival rate?
6. What is the survival rate for each embarkation?
7. What is the survival rate for a person without family onboard?

Having formulated the questions, I will utilize SQL to provide answers. Several SQL aggregation commands will be particularly useful for this project.

**Question 1: What is the overall survival rate for the titanic passenger?**

To determine the survival rate, we can employ the following formula:



This formula expresses the survival rate as a percentage.

SELECT

COUNT(CASE WHEN survived = 1 THEN 1 END) AS Total\_Passenger\_Survived,

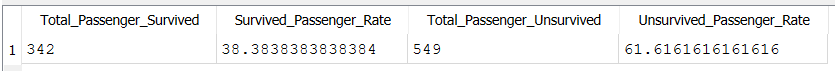
(COUNT(CASE WHEN survived = 1 THEN 1 END) \* 100.0 / COUNT(\*)) AS Survived\_Passenger\_Rate,

COUNT(CASE WHEN survived = 0 THEN 1 END) AS Total\_Passenger\_Unsurvived,

(COUNT(CASE WHEN survived = 0 THEN 1 END) \* 100.0 / COUNT(\*)) AS Unsurvived\_Passenger\_Rate

FROM

passengers;



The number of passengers who survived is 342, which corresponds to a survival rate of 38.38%. This statistic suggests that the majority of passengers did not survive, highlighting the severity of the situation. Understanding this rate can help in analysing the factors that contributed to survival, such as demographics or access to lifeboats.

**Question 2: Are females more likely to survive this incident?**

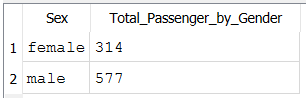
Next, we will investigate whether female passengers had a greater likelihood of survival during the Titanic disaster. Historical data indicates that women and children were typically prioritized during lifeboat evacuations, making it crucial to analyze the survival rates specifically for female passengers.

To conduct this analysis, we will calculate the total number of each gender.

SELECT Sex, COUNT(\*) AS Total\_Passenger\_by\_Gender

FROM passengers

GROUP BY Sex;

****

Based on the result above, among the passengers onboard, there were 314 females and 577 males. By using the formula for survival rate, we can determine if the percentage of women who survived is significantly higher than that of their male counterparts.

**Survival and Unsurvival Rate for Male**

SELECT

COUNT(CASE WHEN survived = 1 THEN 1 END) AS Total\_Male\_Survived,

(COUNT(CASE WHEN survived = 1 THEN 1 END) \* 100.0 / COUNT(\*)) AS Male\_Survival\_Rate,

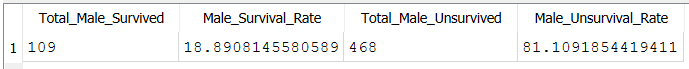
COUNT(CASE WHEN survived = 0 THEN 1 END) AS Total\_Male\_Unsurvived,

(COUNT(CASE WHEN survived = 0 THEN 1 END) \* 100.0 / COUNT(\*)) AS Male\_Unsurvival\_Rate

FROM passengers

WHERE

sex = 'male';



**Survival and Unsurvival Rate for Female**

SELECT

COUNT(CASE WHEN survived = 1 THEN 1 END) AS Total\_Female\_Survived,

(COUNT(CASE WHEN survived = 1 THEN 1 END) \* 100.0 / COUNT(\*)) AS Female\_Survival\_Rate,

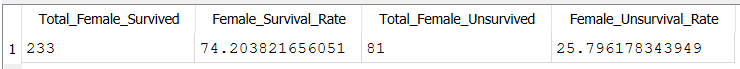
COUNT(CASE WHEN survived = 0 THEN 1 END) AS Total\_Female\_Unsurvived,

(COUNT(CASE WHEN survived = 0 THEN 1 END) \* 100.0 / COUNT(\*)) AS Female\_Unsurvival\_Rate

FROM passengers

WHERE

sex = 'female';



Among the total passengers on board, 233 females with a survival rate of 74.2% survived the Titanic disaster, while 109 males with a survival rate of 18.89% managed to escape. Based on this result, it shows that female passengers are more likely to survive in this tragedy.

These statistics reveal a stark disparity in survival rates between genders. The higher survival rate among females suggests that the evacuation protocols, which prioritized women and children, were somewhat effective in this instance. Notably, a greater percentage of women survived compared to men, indicating that societal norms at the time may have influenced who was allowed access to lifeboats.

**Question 3: Are children and elderlies have a higher survival rate in this accident?**

Upon exploring the relationship between survival and age, there is a notable correlation. We categorize individuals as follows: children are defined as those aged between 0 and 17, the elderly are those over 59 (not including 59), adults are classified as those aged between 18 and 59, and any missing age data is labelled as "NULL."

SELECT

CASE

WHEN age IS NULL THEN 'NULL'

WHEN age BETWEEN 0 AND 17 THEN 'Children'

WHEN age > 59 THEN 'Elderlies'

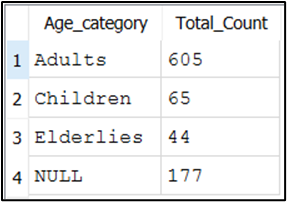
WHEN age BETWEEN 18 AND 59 THEN 'Adults'

ELSE 'Unknown'

END AS Age\_category,

COUNT(\*) AS Total\_Count

FROM passengers GROUP BY Age\_category;



The Titanic dataset includes 605 adults, 65 children, 44 elderly individuals, and 177 entries with missing values in the Age column.

SELECT

CASE

WHEN age IS NULL THEN 'NULL'

WHEN age BETWEEN 0 AND 17 THEN 'Children'

WHEN age > 59 THEN 'Elderlies'

WHEN age BETWEEN 18 AND 59 THEN 'Adults'

ELSE 'Unknown'

END AS Age\_category,

SUM(survived) AS Total\_Survived,

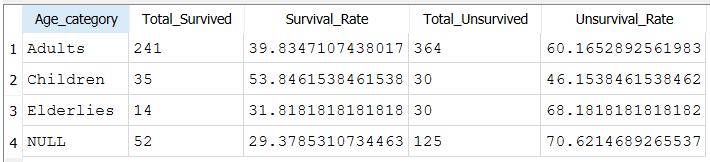
(SUM(survived) \* 100.0 / COUNT(\*)) AS Survival\_Rate,

(COUNT(\*) - SUM(survived)) AS Total\_Unsurvived,

((COUNT(\*) - SUM(survived)) \* 100.0 / COUNT(\*)) AS Unsurvival\_Rate

FROM passengers

GROUP BY Age\_category;



Based on the result, it shows that as age increased, the survival rate decreased. Children exhibited the highest survival rates, followed by the adults. This trend can likely be attributed to the prioritization of women and children during evacuations, the physical fitness of younger individuals, and the increased vulnerability of older adults to cold and exposure.

## Question 4: Do rich people have a higher survival rate because they get on board to rescue boats sooner?

In the class column, the number indicates a different class: -

* 1 = Upper class
* 2 = Middle class
* 3 = Lower class

SELECT

Pclass AS Passenger\_Class,

COUNT(\*) AS Total\_Passenger

FROM

passengers

GROUP BY

Pclass;

## 

## The data reveals that the majority of passengers belonged to Third Class, with a total of 491 passengers, followed by First Class with 216 passengers, and finally Second Class with 184 passengers. Now, let’s calculate each survival rate based on their *Pclass*.

SELECT

Pclass AS Passenger\_Class,

COUNT(CASE WHEN Survived = 1 THEN 1 END) AS Total\_Survived,

AVG(Survived) \* 100 AS Survival\_Rate,

COUNT(CASE WHEN Survived = 0 THEN 1 END) AS Total\_Unsurvived,

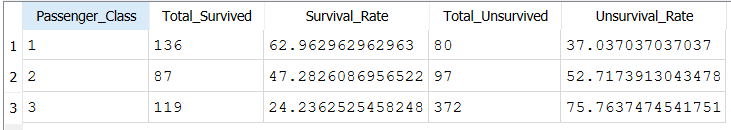
(COUNT(CASE WHEN Survived = 0 THEN 1 END) \* 100.0 / COUNT(\*)) AS Unsurvival\_Rate

FROM

passengers

GROUP BY

Pclass;



Based on the result above, the survival rate for each class is as follows:

* 1st class — 62.96%
* 2nd class — 47.28%
* 3rd class — 24.24%

These statistics indicate that 1st Class passengers had the highest survival rate, with approximately 62.96% of them surviving the disaster. This can be attributed to several factors, including their proximity to lifeboats and the prioritization of women and children, which may have disproportionately benefited those in higher classes.

In contrast, 2nd Class passengers had a survival rate of around 47.28%, while only 24.24% of 3rd Class passengers survived. The significantly lower survival rate among 3rd Class passengers suggests that they faced considerable barriers during the evacuation process. Many were located farther from the lifeboats and may have encountered additional obstacles that hindered their escape.

**Question 5: Which cabin has highest survival rate?**

Let us examine if there are any correlation between cabin deck and survival rate. We will begin by classifying the cabins as "A" if the first character of the cabin identifier is 'A' and so on.

WITH CabinDecks AS (

SELECT

SUBSTR(Cabin, 1, 1) AS CabinDeck,

Pclass,

Survived

FROM

passengers

WHERE

Cabin IS NOT NULL

)

SELECT

CabinDeck,

COUNT(CASE WHEN Pclass = 1 THEN 1 END) AS Total\_First\_Class,

COUNT(CASE WHEN Pclass = 2 THEN 1 END) AS Total\_Second\_Class,

COUNT(CASE WHEN Pclass = 3 THEN 1 END) AS Total\_Third\_Class,

COUNT(CASE WHEN Survived = 1 THEN 1 END) AS Total\_Survived,

AVG(Survived) \* 100 AS Survival\_Rate,

COUNT(CASE WHEN Survived = 0 THEN 1 END) AS Total\_Unsurvived,

(COUNT(CASE WHEN Survived = 0 THEN 1 END) \* 100.0 / COUNT(\*)) AS Unsurvival\_Rate

FROM

CabinDecks

GROUP BY

CabinDeck

ORDER BY

CabinDeck;

## 

## Based on the result, it shows that the survival rate if passengers board on cabin B, D and E is almost the same which is around 75%. Several factors could contribute to this uniform survival rate, such as the location of these cabins on the ship, which may have affected accessibility to lifeboats. Additionally, social factors, including the presence of women and children in these cabins, might have influenced priority during the evacuation.

## However, more investigation is needed as there are 687 missing values in the cabin dataset that could impact the analysis. A thorough examination of these missing data points could provide a more nuanced understanding of how location of the cabin could influence survival during this tragic event.

**Question 6: What is the survival rate for each embarkation?**

There are 3 ports of embarkation: Cherbourg(C), Queenstown(Q), and Southampton(S). There are 2 missing values in this dataset.

SELECT

CASE

WHEN Embarked IS NULL THEN 'Unknown'

ELSE Embarked

END AS Embarkation\_Port,

COUNT(\*) AS Passenger\_Count

FROM

passengers

GROUP BY

CASE

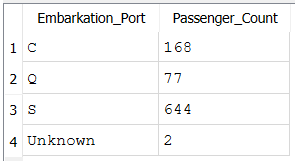
WHEN Embarked IS NULL THEN 'Unknown'

ELSE Embarked

END

ORDER BY

Embarkation\_Port;



Let’s compare the survival rate of each passenger by each of the embarkation port. In this code, we exclude the NULL values.

SELECT

Embarked,

SUM(CASE WHEN Pclass = 1 THEN 1 ELSE 0 END) AS First\_Class\_Count,

SUM(CASE WHEN Pclass = 2 THEN 1 ELSE 0 END) AS Second\_Class\_Count,

SUM(CASE WHEN Pclass = 3 THEN 1 ELSE 0 END) AS Third\_Class\_Count,

SUM(Survived) AS Total\_Survived,

AVG(Survived) \* 100 AS Survival\_Rate,

(COUNT(\*) - SUM(Survived)) AS Total\_Unsurvived,

(100.0 \* (COUNT(\*) - SUM(Survived)) / COUNT(\*)) AS Unsurvival\_Rate

FROM

passengers

WHERE

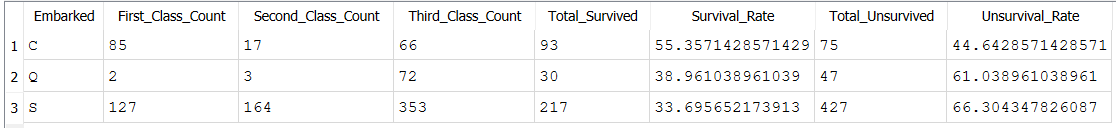
Embarked IN ('C', 'Q', 'S')

GROUP BY

Embarked

ORDER BY

Embarked;



The data reveals that passengers from Cherbourg had the highest survival rate at 55.36%, indicating that a significant proportion of those who boarded the Titanic at this port were able to survive the disaster. Following Cherbourg, Queenstown passengers exhibited a survival rate of 38.96%.

While this rate is notably lower than that of Cherbourg, it still reflects a relatively substantial number of survivors compared to Southampton.

In contrast, Southampton had the lowest survival rate at 33.7%. This may suggest that the composition of passengers from this port included a larger number of individuals in third class, who traditionally had lower survival rates due to their distance from lifeboats and the chaos during the evacuation.

**Question 7: What is the survival rate for a person without family onboard?**

SELECT

(Parch + SibSp) AS FamilySize,

COUNT(\*) AS Total\_Passengers,

SUM(CASE WHEN Survived = 1 THEN 1 ELSE 0 END) AS Total\_Survived,

SUM(CASE WHEN Survived = 0 THEN 1 ELSE 0 END) AS Total\_Unsurvived,

AVG(Survived) \* 100 AS Survival\_Rate,

(SUM(CASE WHEN Survived = 0 THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*)) AS Unsurvival\_Rate

FROM

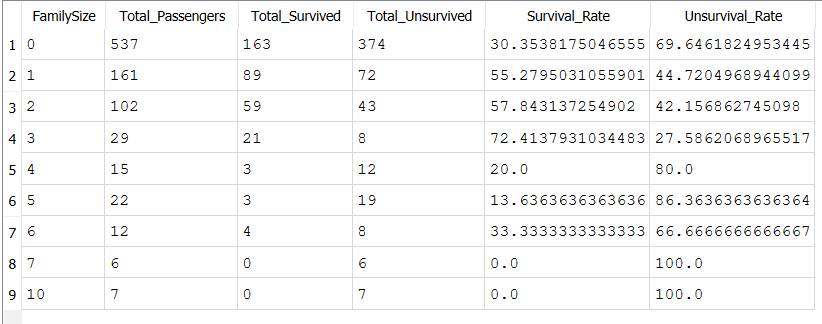
passengers

GROUP BY

FamilySize

ORDER BY

FamilySize;



There are lots of passengers who travelled without any family onboard. The passenger with families of 3, or 4 people had a higher chance of survival than single people or people with larger number of family members.

**Conclusion**

* 38.38% of passengers survived in titanic
* Children exhibited the highest survival rates, followed by the adults
* Female passengers are more likely to survive than male passengers
* The upper class is more likely to survive
* The survival rate if passengers board on cabin B, D and E is almost the same which is around 75%
* Port Cherbourg passengers have the highest surviving rate — 55.36%

**Limitations**

* The analysis has not taken into account the duplication of data. It has been assumed that each entry is unique.
* Rows with missing values in columns (Age, Cabin, Embarked) have been removed for the analysis of corresponding columns. This reduces the fidelity of the analysis based on available data.
* The dataset is only a subset of total original passengers onboard and the factors such as survival rate could vary.