Understanding the Business Context

The purpose for these datasets is to make a descriptive analysis of what happened to the passengers after the Titanic had sunk. The data will provide us with the insights of the passengers that had sailed on Titanic the day it was sunk, their names, age, their passengers' class, the cabin that they have located in, and the number of passengers that had survived the incident.

The data sets were downloaded from https://www.kaggle.com/competitions/titanic/data for learning purposes.

Understanding the Technical Context

The data sourced from https://www.kaggle.com/competitions/titanic/data with the purpose of predicting the survival of the passengers on Titanic. The data sets were downloaded as an Excel file from the website. The data sets were analysed on SQLite and queries were executed to analyse the survival rate based on the 4 hypotheses that has been made prior pursuing this project. After exploring the data sets, we figured that some the columns consist of NULL values and the details are as the following:

Column Name	Percentage of Null Values
Age	19.9%
Cabin	77.1%
Embarked	0.2%

Since we did not have any sources to fill in the NULL values, those data from the Cabin column will be excluded from the analysis due to high percentages of NULL values. We also figured that there are no duplicate rows inside the database.

Understanding the tables and fields

The Titanic database consist of 1 table with 12 columns and 891 rows of data tabulated in it. The table named as “passengers” indicating the details of the passengers of the Titanic. The columns are :

Variable	Definition	Key
Survived	Survival status	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

As a note, the Passenger Class were assigned as Class 1 – top class, Class 2 – middle class, Class 3 – the bottom class. The **SibSp** field only include brother, sister, stepbrother, stepsister, husband, and wife. The fiancés and mistresses were not included.

The **Parch** field is defined as below:

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them

Free Exploration

Coming up with research questions

Introduction

About 1,500 passengers and crew members perished when the British luxury passenger liner Titanic, officially known as the Royal Mail Ship (RMS) Titanic, sank on April 14 and 15, 1912, while on its journey from Southampton, England, to New York City.

Hence, for this project, we are going to dive deep into the database of the passengers of the Titanic to understand the survival rate of the passengers according to the 4 hypotheses that have been made prior pursuing this project. The four hypotheses are:

1. **Which age class has the highest survival rate?** the age classes were divided into three categories :
    i. Children (age below 18)
    ii. Adults (from 18 to 59 years old)
    iii. Elderlies (above 60 years old)
2. **Do female passengers have higher survival rate than the male?**
3. **Do the passengers' class status influence their survival rate?**
4. **Which of the three ports of embark has the highest survival rate? And which passengers class dominates the specific port of embark?**

To make the analysis more effective, we are going to make two analyses for each hypothesis. The first analysis will indicate the situation before the titanic sank, and the second analysis will indicate the situation after the titanic sank, the survival rate of passengers and the factors influencing the results of the analyses were also discussed.

Pre-analysis

In this section we are going to discuss about the queries used for the pre-analysis of the passengers table inside the Titanic database.

i. Finding NULL Values

Since only column “Age”, “Cabin” and “Embarked” were the columns that consist of NULL Values, the results from the other columns will be excluded for this section.

<table>
<tr><th>Query</th><th>Result</th></tr>
<tr><td>SELECT
SUM(CASE WHEN Age is NULL
THEN 1 ELSE 0 END)
AS "Age Null Values",
COUNT(Age) AS "Age Non-Null Values"
FROM passengers</td><td><table><tr><th>Age Null Values</th><th>Age Non-Null Values</th></tr><tr><td>177</td><td>714</td></tr></table>
$\frac{177}{891}\ x\ 100\% = 19.9\%\ of\ NULL\ Values$</td></tr>
<tr><td>SELECT
SUM(CASE WHEN Cabin is NULL
THEN 1 ELSE 0 END)
AS "Cabin Null Values",
COUNT(Cabin) AS "Cabin Non-Null Values"
FROM passengers</td><td><table><tr><th>Cabin Null Values</th><th>Number Of Non-Null Values</th></tr><tr><td>687</td><td>204</td></tr></table>
$\frac{687}{891}\ x\ 100\% = 77.1\%\ of\ NULL\ Values$</td></tr>
<tr><td>SELECT
SUM(CASE WHEN Embarked is null
THEN 1 ELSE 0 END)
AS "Embarked Null Values" ,
COUNT(Embarked) AS "Embarked Non-Null Values"
FROM passengers</td><td><table><tr><th>Embarked Null Values</th><th>Embarked Non-Null Values</th></tr><tr><td>2</td><td>889</td></tr></table>
$\frac{2}{891} x\ 100\% = 0.2\%\ of\ NULL\ Values$</td></tr>
</table>

ii. Finding duplicate rows

Query	Result
SELECT PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, count(*) FROM passengers GROUP BY PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked HAVING COUNT(*) > 1	0 rows returned Hence, there is no duplicate rows.

Answering the research questions with SQL

Question 1: Which age class has the highest survival rate?

1. The total number of passengers in Titanic group by age classification:

Query	Results
SELECT COUNT(Name) AS total_passengers, CASE WHEN age < 18 THEN "children" WHEN age >= 60 THEN "elderlies" WHEN age >= 18 AND AGE < 60 THEN "adult" END AS age_classification FROM passengers GROUP BY age_classification	**total_passengers** / **age_classification** 177 / (blank) 608 / adult 65 / children 41 / elderlies

2. Total number of passengers **survived the Titanic incident**, classified by **age**:

Query	Results
SELECT COUNT(Survived) AS total_survivors, CASE WHEN Age >= 60 THEN "elderlies" WHEN Age < 18 THEN "children" WHEN Age >= 18 AND Age < 60 THEN "adults" END AS survivor_age_class FROM passengers WHERE Survived = 1 GROUP BY survivor_age_class	**total_survivors** / **survivor_age_class** 52 / (blank) 243 / adults 35 / children 12 / elderlies

Table 1 and Table 2 show the number of passengers on Titanic before and after the incident happened, classified by age.

Adults were those in between 18 to 59 years old, **Children** were those below 18 years old and **Elderlies** were those 60 years old and above. (blank space in age class indicates that there is no data about the passengers' age, hence no age class applicable to the group).

Both tables illustrates that:

- Less than half of adults survived. (40%)
- More than half children were survived. (53.8%)
- Only three out of ten elderlies were survived. (29.3%)

From these numbers, we can conclude that children have higher survival rate in comparison to the other age groups as rescue team will prioritize in rescuing the children in any emergency. The elderlies have the lowest rate of survival, which most probably due to their physical condition.

Question 2 : Do the female passengers have higher survival rate than male?

3. Total number of females and males on Titanic:

Query	Results
SELECT Sex AS Gender, COUNT(Name) AS Total_passengers FROM passengers GROUP BY Sex	Gender: female, Total_passengers: 314 Gender: male, Total_passengers: 577

4. Total number of females and males survived **after the Titanic incident**:

Query	Results
SELECT Sex AS Gender, COUNT(Survived) AS Total_survivors FROM passengers WHERE Survived = 1 GROUP BY Sex	Gender: female, Total_survivors: 233 Gender: male, Total_survivors: 109

Table 3 and Table 4 indicate the number of passengers before and after the Titanic incident grouped by gender.

Both tables shows that:

- More than 2/3 of the total female passengers survived. (74.2%)
- Less than 1/5 of the total male passengers survived. (18.9%)

These numbers most probably because, in any emergency, women and children are often rescued first, followed by elders and adult men. Hence, we can conclude that women have higher survival rate than men in any emergency.

Question 3 : Do the passengers' Class status influence their survival rate?

5. Total number of passengers based on Class:

Query	Results
SELECT Pclass AS Passengers_class, COUNT(Pclass) AS Total_passengers FROM passengers GROUP BY Pclass	<table><tr><th>Passengers_class</th><th>Total_passengers</th></tr><tr><td>1</td><td>216</td></tr><tr><td>2</td><td>184</td></tr><tr><td>3</td><td>491</td></tr></table>

6. Total number of passengers **survived the Titanic incident** based on **Class**:

Query	Results
SELECT Pclass AS Passengers_class COUNT(Pclass) AS Total_survivors FROM passengers WHERE Survived = 1 GROUP BY Pclass	<table><tr><th>Passengers_class</th><th>Total_survivors</th></tr><tr><td>1</td><td>136</td></tr><tr><td>2</td><td>87</td></tr><tr><td>3</td><td>119</td></tr></table>

Based on Table 5 and Table 6 above, we can see the total number of passengers classified by class before, and after the titanic incident.

Both tables depicted that:

- More than half of passengers from Class 1 were survived. (70%)
- Less than half of the passengers from Class 2 were survived. (47.3%)
- Less than a quarter of passengers from Class 3 were survived. (24.2%)

The inference that can be done from these data was that the passengers from a higher Class got the access to rescue faster than the passengers from the middle and lower Class. We can conclude that passengers from the higher class have higher survival rate in comparison to the lower class.

Question 4 : Which of the three ports of embark has the highest survival rate? And which passengers class dominated the specific port of embark?

7. Total passengers based on their port of embark and their passengers' class:

Query	Results

```
SELECT
    Embarked AS "Port of Embark", Pclass,
    COUNT(Embarked)
    AS "Total Passengers"
FROM passengers
GROUP by Embarked, Pclass
```

Port of Embark	Pclass	Total Passengers
C	1	85
C	2	17
C	3	66
Q	1	2
Q	2	3
Q	3	72
S	1	127
S	2	164
S	3	353

8. Total number of survivors based on their port of embark and passengers class:

Query	Results

```
SELECT
    Embarked AS "Port of Embark", Pclass,
    COUNT(Embarked) AS "Number of
    Survivors"
FROM passengers
WHERE Survived = 1
GROUP by Embarked, Pclass
```

Port of Embark	Pclass	Number of Survivors
C	1	59
C	2	9
C	3	25
Q	1	1
Q	2	2
Q	3	27
S	1	74
S	2	76
S	3	67

Table 7 and 8 shows that 50.6% of the passengers who embarked at Cherbourg port were from Class 1, and 5.4% and 37.9% were from Class 2 and Class three respectively. Class 1 dominated the Cherbourg port.

Meanwhile at the Queenstown port, 2.6% was from Class 1, 3.9% from Class 2 and 97.4% from Class 3. Passengers of Class 3 dominated this port. On the other hand, at the Southampton port, 19.7% were from Class 1, 24.5% from Class 2 and 54.8% from Class 3. Passengers of Class 3 dominated the Southampton port.

From table 7 and 8 we know that:

- 69.4% of the passengers from Class 1 who embarked from Cherbourg survived.
- 52.9% of the passengers from Class 2 who embarked from Cherbourg survived.
- 37.9% of the passengers from Class 3 who embarked from Cherbourg survived.
- 50% of the passengers from Class 1 who embarked from Queenstown survived.
- 66.7% of the passengers from Class 2 who embarked from Queenstown survived.
- 37.5% of the passengers from Class 3 who embarked from Queenstown survived.
- 58.3% of the passengers from Class 1 who embarked from Southampton survived.
- 46.3% of the passengers from Class 2 who embarked from Southampton survived.
- 19% of the passengers from Class 3 who embarked from Southampton survived.

As a whole, we can see that :

- **Port C** – Survived : 93, Total passengers : 168, **Survival Rate : 55.4%**
- **Port Q** – Survived : 30, Total passengers : 77, **Survival Rate : 38.9%**
- **Port S** – Survived : 217, Total passengers : 644, **Survival Rate : 33.7%**