

Understanding the Business Context

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of many passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

This report explores a dataset related to the Titanic disaster to analyze survival rates and identify patterns among different groups of passengers and crew. Through this data exploration, we aim to answer several questions regarding survival probabilities and factors influencing survival. The following sections will present detailed analyses, including statistical summaries and insights drawn from the dataset.

What are these data for?

These data is used for data analysis and machine learning. It contains information about the passengers onboard the Titanic, including features like age, sex, fare, cabin, survival status and many more.

Why do we need this database?

As mentioned, it is a popular database used to understand basics for machine learning. This database can be used to perform comprehensive data exploration and analysis of the Titanic disaster. It allows us to efficiently query and manage detailed passenger and crew data, which is essential for uncovering patterns and insights.

By leveraging this database, we can gain a deeper understanding of survival rates and factors influencing outcomes, providing valuable insights beyond what’s available in other datasets, such as those on Kaggle. This database can also be used to practice and enhance our understanding in machine learning knowledge, refresh our problem-solving and analytical skills.

Where are these data collected?

The dataset, used in this report is a curated version of historical passenger records from the RMS Titanic disaster and those that are available in Kaggle.

Understanding the Technical Context

How are these data collected?

These data are collected probably by extracting information from historical records, such as passenger lists, and sometimes combining it with publicly available databases.

Where are the sources of these data?

The existing database in SQLite is derived from the Titanic records, which were originally collected from historical archives. Sources of these data mostly include archives, historical documents, and sometimes institutional records or online databases that hold Titanic-related information. For the table and field definition, Kaggle was used as a reference.

Is the data coming from surveys, or some computer system? Is it manually input by some data entry personnel or collected by some electronic system?

The Titanic data primarily comes from historical records and archives rather than surveys or electronic systems. It was originally documented manually by shipping line personnel and has been digitized for modern use. Thus, the data is typically collected through archival research and then entered into databases for analysis, rather than through contemporary electronic data collection systems.

What are the systems that touch or use/modify these data?

The Titanic dataset provided may exhibit signs of previous processing, despite being in its raw state. For instance, the data is organized into well-defined columns with consistent headers so it indicates some level of initial formatting. Patterns in missing data could suggest that some records have been removed or that missing values have not yet been addressed.

Standardized values, such as consistent categories for sex(gender) may reflect some degree of normalization. These processing steps can significantly impact the dataset's reliability and usability. While processing can help by correcting errors and making the data easier to work with, it may also alter the original information in ways that could affect the analysis.

What are some of the error sources of this data?

The Titanic dataset can have errors from several sources. Data entry mistakes might have occurred during the manual recording of information from ship logs, leading to inaccuracies. Incomplete records in this dataset is an issue as well as some information might be missing due to gaps in the original documentation. Historical inaccuracies in the original records, such as incorrect ages or misspelled names, could also affect the data. Additionally, issues may arise from the way data was extracted from archives or combined from different sources, leading to inconsistencies.

Is the data complete? Would there be missing pieces of data?

This Titanic dataset is considered to be incomplete in several ways, which can impact analysis significantly. For instance, many records lack age information, which affects the ability to analyze survival rates by age group. Cabin data is often missing or only partially recorded, limiting detailed insights into survival based on cabin location.

What is the impact of incomplete data and what methods can be used to handle missing data during analysis?

Incomplete data, or missing values, can significantly affect a dataset's reliability and accuracy. Missing data can introduce bias and lead to skewed results if not handled properly.

To address missing data, several methods can be used. One approach is to remove records with missing values, though this can lead to substantial data loss. Another approach is to ignore records with missing values during queries or calculations. This means that only the records with complete data are used, effectively excluding any incomplete records from the analysis. While this method is straightforward and can be useful when missing values are rare, it does not address the root cause of missingness and may lead to biased results if the missing data is not randomly distributed.

Other methods for handling missing data include imputation techniques, where missing values are filled in based on observed data, and more sophisticated methods like predictive imputation or multiple imputation, which create several datasets to account for uncertainty. Additionally, some machine learning algorithms can handle missing data directly during analysis, and creating indicator variables for missing data can help models consider the absence of data explicitly.

What are the potential biases in the data?

The Titanic dataset can have several potential biases due to human error and data collection practices. For instance, the dataset might reflect social and economic biases of the time, such as class or gender biases, because it predominantly includes data on passengers and crew from certain social classes or genders, potentially skewing survival analysis. Incomplete or incorrect entries, like missing age or cabin information, could introduce biases in analyzing age-related survival rates or cabin-based survival patterns. Additionally, historical inaccuracies or inconsistencies in recording and digitization can further skew the data, affecting the reliability of conclusions drawn from it. These biases could lead to misleading results, such as overestimating or underestimating the influence of certain factors on survival rates.

Understanding the Tables & Fields

How many tables do we have?

There is only one table named 'passengers'

What are the tables? and what are these tables representing?

The one table available named 'passengers' includes all information about passengers who aboard the Titanic ship on the day the accident happened. Inside the table contained various columns available describing detail information of each passenger, whether they survived or not, their class ticket and many more.

What are the relationships between the tables?

There are no relationship between tables since there's only one table titled 'passengers'

What are the fields in the tables? What is the meaning of each of the field?

There are 12 fields in the 'passengers' table named 'PassengerID', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'

PassengerID: Contained the ID for each passenger

Survived: 1 = Survived, 0 = Did not survive

Pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

Name: Describe the name of each passengers

Sex: Describe the gender of each passengers

i.e. female, male

Age: Describe the age in years. Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

SibSp: The dataset defines family relations in this way.

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

Parch: The dataset defines family relations in this way.

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

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

Ticket: Ticket number of each passenger

Fare: Describe the fare for each passenger

Cabin: Identified the cabin occupied by each passenger

Embarked: Port of Embarkation

C = Cherbourg, Q = Queenstown, S = Southampton

Is the data messy? and how?

There are missing values in the fields such as 'Age' and 'Cabin', which are 'NULL'. Some of the data are also inconsistent and have incorrect data types. For instance, 'Age' field having the text format instead of integer/float. 'Cabin' and 'Ticket' column on the other hand contain mixed data types, such as numbers mixed with text.

Though ticket numbers data are all filled, it may contain errors as some are not listed in the normal ticket number format but as 'LINE' instead. According to the discussion in Kaggle, this probably mean that the workers listed as LINE were staff meant for the SS Philadelphia. Their ship's voyage was canceled due to a workers' strike, so they were onboarded onto the Titanic in Southampton. Their actual ticket number was 370160 and in conclusion, this thus may affect any analysis tied to specific tickets. Possible error in fare data as some actually resulting in '0' can also impact financial analyses or insights related to ticket pricing.

Should I clean the data first? or ignoring those messy columns

First of all, I decided to check if there's any duplicates in this dataset.

```
SELECT Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, COUNT(*) AS count
FROM passengers
GROUP BY Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
HAVING COUNT(*) > 1;
```

Using the query above on SQLite, the result returned with none. Next, I decided to check for the missing values and it turns out that 'Age' column has 177 NULL values and 'Cabin' column has 687 NULL values out of 891 passengers. Since one of the question I wanted to explore involves the 'Age' column, I decided to clean the data with the impute method by filling in missing values with the median. This will be shown later in the 'Free Exploration' section below.

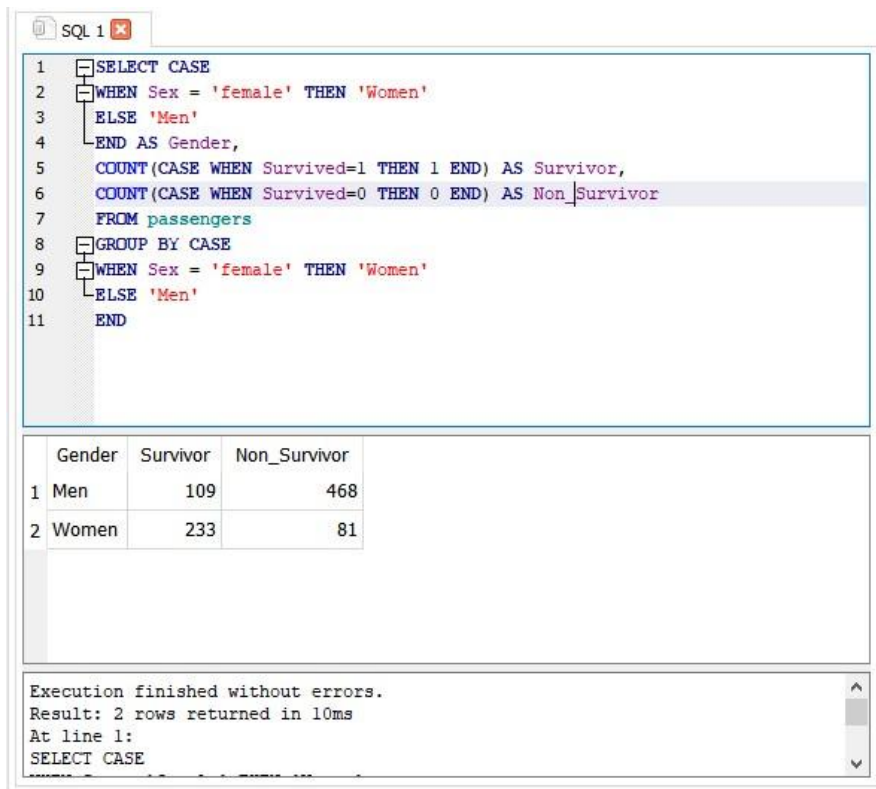
Considering how the 'Cabin' column won't play any important role in my analysis, it was decided that the data in it won't be used so it will be left as is.

Free Exploration

QS 1: How many women survived in Titanic accident compared to men?

SQL Query:

```
SELECT CASE
WHEN Sex = 'female' THEN 'Women'
ELSE 'Men'
END AS Gender,
COUNT(CASE WHEN Survived=1 THEN 1 END) AS Survivor,
COUNT(CASE WHEN Survived=0 THEN 0 END) AS Non_Survivor
FROM passengers
GROUP BY CASE
WHEN Sex = 'female' THEN 'Women'
ELSE 'Men'
END
```



The screenshot shows a SQL query execution window with the following SQL query:

```
1 SELECT CASE
2 WHEN Sex = 'female' THEN 'Women'
3 ELSE 'Men'
4 END AS Gender,
5 COUNT(CASE WHEN Survived=1 THEN 1 END) AS Survivor,
6 COUNT(CASE WHEN Survived=0 THEN 0 END) AS Non_Survivor
7 FROM passengers
8 GROUP BY CASE
9 WHEN Sex = 'female' THEN 'Women'
10 ELSE 'Men'
11 END
```

The results are displayed in a table with 3 columns: Gender, Survivor, and Non_Survivor. The table has 2 rows of data.

	Gender	Survivor	Non_Survivor
1	Men	109	468
2	Women	233	81

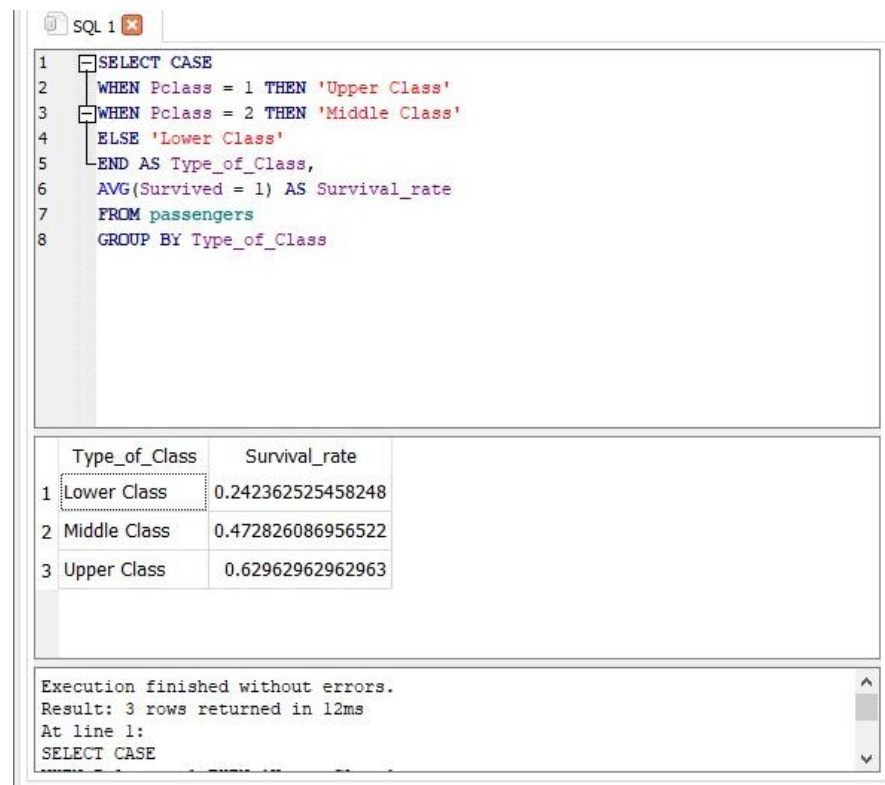
Execution finished without errors.
Result: 2 rows returned in 10ms
At line 1:
SELECT CASE

Conclusion: According to the data provided above, it is clear that the amount of women who survived is higher compared to men. In opposite, it also shows that the amount of men who did not survive is significantly higher compared to women.

QS 2: Does rich people have a higher survival rate because they can get onboard to the rescue boat sooner (like what is shown in the movie)?

SQL Query:

```
SELECT CASE
WHEN Pclass = 1 THEN 'Upper Class'
WHEN Pclass = 2 THEN 'Middle Class'
ELSE 'Lower Class'
END AS Type_of_Class,
AVG(Survived = 1) AS Survival_rate
FROM passengers
GROUP BY
Type_of_Class
```



The screenshot shows a SQL query execution window with the following content:

```
1 SELECT CASE
2   WHEN Pclass = 1 THEN 'Upper Class'
3   WHEN Pclass = 2 THEN 'Middle Class'
4   ELSE 'Lower Class'
5   END AS Type_of_Class,
6   AVG(Survived = 1) AS Survival_rate
7   FROM passengers
8   GROUP BY Type_of_Class
```

	Type_of_Class	Survival_rate
1	Lower Class	0.242362525458248
2	Middle Class	0.472826086956522
3	Upper Class	0.62962962962963

Execution finished without errors.
Result: 3 rows returned in 12ms
At line 1:
SELECT CASE

Conclusion: According to the result shown above, the answer to the proposed question is yes. Rich people as in the upper class in the data does have a higher survival rate which is 62.96% compared to the people in middle class and lower class. In conclusion, first-class passengers had a higher survival rate, indicating socio-economic status played a role in survival chances.

QS 3: Does people that have family relations with other passengers on board have a higher survival rate than others who are not?

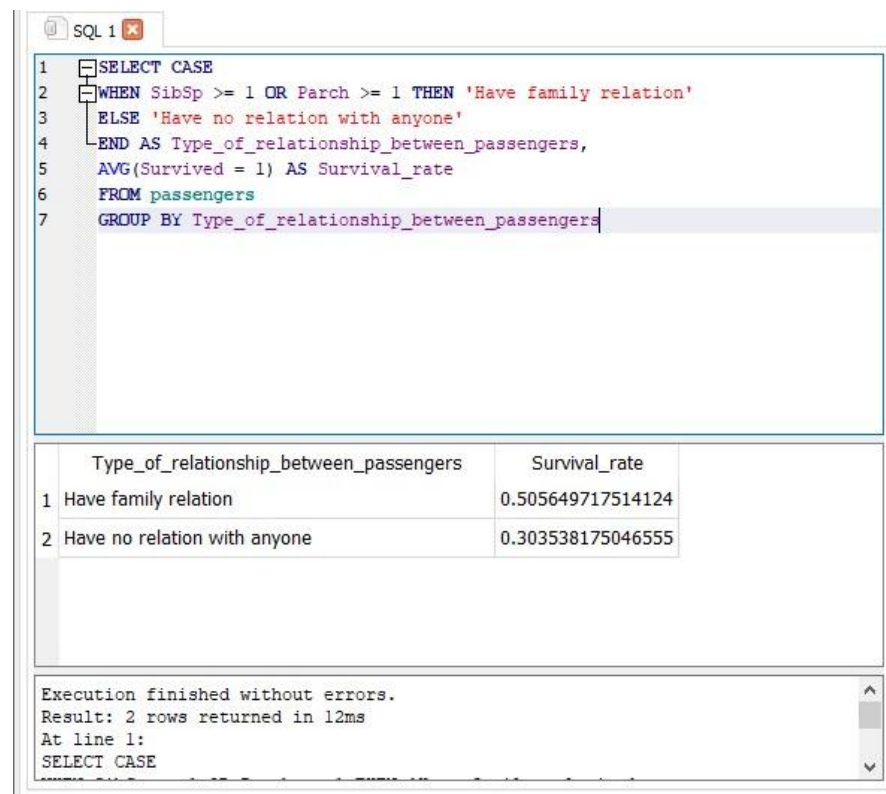
SQL Query:

```
SELECT CASE
WHEN SibSp >= 1 OR Parch >= 1 THEN 'Have family relation'
ELSE 'Have no relation with anyone'
END AS Type_of_relationship_between_passengers,
```

```

AVG(Survived = 1) AS Survival_rate
FROM passengers
GROUP BY Type_of_relationship_between_passengers

```



The screenshot shows a SQL query execution window with the following query:

```

1 SELECT CASE
2   WHEN SibSp >= 1 OR Parch >= 1 THEN 'Have family relation'
3   ELSE 'Have no relation with anyone'
4 END AS Type_of_relationship_between_passengers,
5 AVG(Survived = 1) AS Survival_rate
6 FROM passengers
7 GROUP BY Type_of_relationship_between_passengers

```

The results are displayed in a table with two columns: `Type_of_relationship_between_passengers` and `Survival_rate`.

Type_of_relationship_between_passengers	Survival_rate
1 Have family relation	0.505649717514124
2 Have no relation with anyone	0.303538175046555

Execution finished without errors.
 Result: 2 rows returned in 12ms
 At line 1:
 SELECT CASE

Conclusion:

It was shown that passengers that have family relation with other passengers on board does have a higher survival rate compared to those who have none and are alone.

QS 4: What is the average age of survivors versus non-survivors?

SQL Query:

```

WITH MeanAge AS (SELECT Survived, AVG(Age) as Mean_Age from passengers
WHERE Age IS NOT NULL
GROUP BY Survived)

```

```

UPDATE passengers
SET Age = (SELECT Mean_Age FROM MeanAge
WHERE MeanAge.Survived = passengers.Survived)
WHERE Age IS NULL

```

```

SELECT Survived, AVG(Age) AS Average_Age
FROM passengers
GROUP BY Survived

```


SQL 1		
1	SELECT Survived, AVG(Age) AS Average_Age	
2	FROM passengers	
3	GROUP BY Survived	
	Survived	Average_Age
1	0	30.626179245283
2	1	28.3436896551724
Execution finished without errors. Result: 2 rows returned in 9ms At line 1: SELECT Survived, AVG(Age) AS Average_Age		

Conclusion:

After calculation, it turns out that there are 177 NULL values out of 891 passengers in the 'Age' column so after much consideration, I decided to use the impute method to clean the data. Since 'Age' plays a vital role in answering this question, this approach is better because it would ensure a more complete and accurate analysis.

Based on the result above, it was concluded that the average age for survivors are younger than the non survivors. Rather than indicating that age directly influenced survival chances, it's possible that the younger average age of survivors results from the evacuation prioritization of women and children, socio-economic factors, and from the perspective that the younger ones' physical ability giving them the leverage of reaching the boat faster compared to the older one.

QS 5: What is the minimum and maximum fare for each passenger class?

SQL Query:

```
SELECT Pclass, MIN(Fare) AS minimum_fare, MAX(Fare) AS maximum_fare
FROM passengers
WHERE fare > 0
GROUP BY Pclass
```

SQL 1

```
1 SELECT Pclass, MIN(Fare) AS minimum_fare, MAX(Fare) AS maximum_fare
2 FROM passengers
3 WHERE fare > 0
4 GROUP BY Pclass
```

	Pclass	minimum_fare	maximum_fare
1	1	5	512.3292
2	2	10.5	73.5
3	3	4.0125	69.55

Execution finished without errors.

Result: 3 rows returned in 35ms

At line 1:

```
SELECT Pclass, MIN(Fare) AS minimum_fare, MAX(Fare) AS maximum_fare
```

Conclusion:

Since there's value '0' in 'fare' field , it's assumed that the passengers who had '0' in their fare data are workers and does not have to pay the fare or its assumed as error in the data. Considering all possibilities and also the fact that there are only 15 record of it out of 891 passengers, I decided to exclude them as shown in my SQL calculation.

As shown above, it was concluded that the middle class paid the highest minimum fare compared to other class which is a bit unexpected. Considering everything, it's possible to note that the Titanic's ticket pricing structure featured a range of prices within each class, and second-class tickets might have had a broader range of fares, with some being priced higher than the minimum fares in first and third classes. The distribution of ticket prices could have resulted in higher minimum fares for the middle class compared to the other classes, possibly due to the availability and pricing strategy. Additionally, the minimum fare recorded for second class might be higher due to fewer low-priced tickets being available.