

Titanic Database: Data Exploration Framework Report

By

Lester Wee Yi Cheng

Contents

Data Exploration and its significance	3
Questions and Justifications for Analysis	4
SQL Query Used.....	6
Discussion of Results and Conclusions	10

Data Exploration and its Significance

Data exploration is a critical framework in the field of data analysis that involves examining and understanding data sets to uncover underlying patterns, relationships, and insights. It serves as a foundational step for any analytical process, enabling analysts to formulate hypotheses, test assumptions, and ultimately derive meaningful conclusions. In the case of the Titanic dataset, which comprises various passenger attributes, the significance of data exploration is magnified as it allows us to explore the factors contributing to passenger survival during the infamous maritime disaster.

In this analysis, I will provide a clear explanation of each data field within the Titanic database to establish a fundamental understanding of the data. Following this, I will outline the methodology employed to extract results using SQL queries, demonstrating how each question was approached and analyzed. Finally, I will discuss the outcomes of the analysis and provide conclusions based on the findings. This structured approach will help elucidate the critical role of data exploration in drawing valuable insights from complex datasets.

Questions and Justifications for Analysis

Prior to forming questions, it is crucial to establish a basic understanding of every single data field in Titanic database objectively. There are 12 data fields in total:

PassengerID	Unique Identifiers assigned to each passengers on board, Primary Keys of the titanic database.
Survived	Passenger who survived from the incident, 0 = No while 1 = Yes
PClass	Passenger's Ticket Class: 1 indicates passenger are of upper echelon 2 implies middle class 3 implies lower tier It serves a s proxy for socio-economic status (SES)
Name	Passengers' name
Sex	Gender
Age	Age as the name suggested, age is fractional if less than 1. If the age is estimated, it is in the form of XX.05.
SibSp	SibSp is the acronym for Sibling and Spouse. (without took into account of mistresses and fiancés)
Parch	Parch is the acronym for Parent and Child, the dataset defines family relation by grouping mother and father as parent, whereas daughter, son, stepdaughter and stepson as child, some children traveled only with a nanny, therefore parch=0 for them.
Ticket	Passengers' Ticket Number
Fare	Passenger Fare

Cabin	Cabin Number
Embarked	Port of Embarkation, indicated by 3 letters: C = Cherbourg Q= Queenstown S= Southampton

Given that passenger survival is the primary focus and there are many fatalities, survival can be considered the dependent variable, while the remaining 10 data fields serve as independent variables that may influence survival outcomes.

Question 1: Which Class of Passengers had a Higher Chance of Survival?

Disregarding factors such as sex and age, the focus is on determining which social class of passengers had a greater chance of survival in the event of the Titanic capsizing, assuming all other factors are held constant. This led to an investigation into the differences in survival rates across different social classes.

While both Pclass and fare could be used to represent social status, Pclass was chosen because fare is highly inconsistent. More importantly, Pclass serves as a more reliable indicator of a passenger's background and their access to resources on the Titanic. Based on this, the hypothesis suggests that passengers from higher social classes had a greater chance of survival compared to those in the other two classes.

Question 2: Do Passengers with Larger Family size have a Lower Chance of Survival?

At first glance, a hypothesis can be formed suggesting that passengers with fewer or no dependents might have a better chance of survival. This is because they wouldn't need to prioritize the safety of dependents, which could potentially lead to impulsive actions and missed opportunities to escape from the sinking ship.

Question 3: Which Port of Embarkation had the Highest Survival Rate?

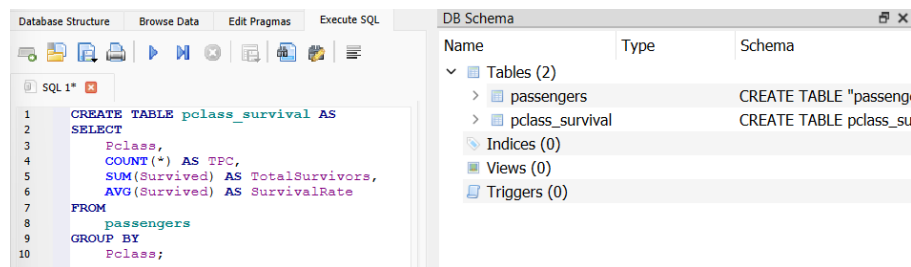
This question is worth exploring because the port of embarkation may be correlated with survival rates, potentially revealing patterns based on where passengers boarded the Titanic and uncover more underlying patterns that warrant further investigation.

SQL Query Used

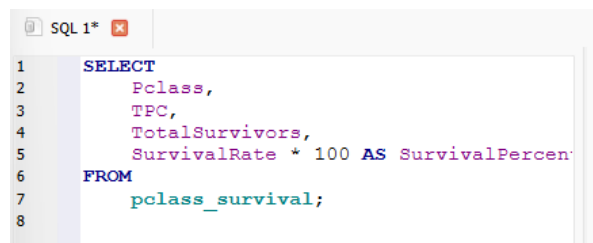
This section showcases and documents the SQL queries used for data retrieval and analysis for each aforementioned question.

Question 1: Which Class of Passengers had a Higher Chance of Survival?

- a. First of all, a new table will be formed based on survived (survived or not) and Pclass (Passenger Class)



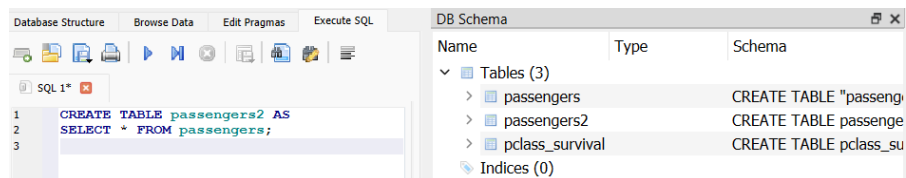
- b. TPC stands for total passenger counts, which is achieved through COUNT(*) when Pclass is selected from passengers table. SUM(Survived) gives the total number of survivors of each class. AVG(Survived) gives the average survival rate for each class in percentage form. The results will then be grouped by passenger class to aggregate the data.
- c. Once the table is created, a comprehensive table showing TotalSurvivors and SurvivalPercentage according to each Pclass can be queried. The SQL Query to obtain the result table is as shown below along with the table generated.



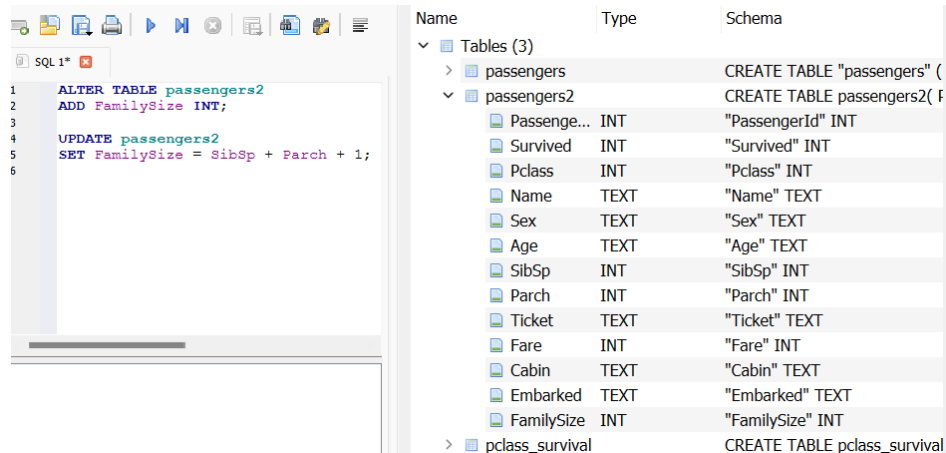
	Pclass	TPC	TotalSurvivors	SurvivalPercentage
1	1	216	136	62.962962962963
2	2	184	87	47.2826086956522
3	3	491	119	24.2362525458248

Question 2: Do Passengers with Larger Family size have a Lower Chance of Survival?

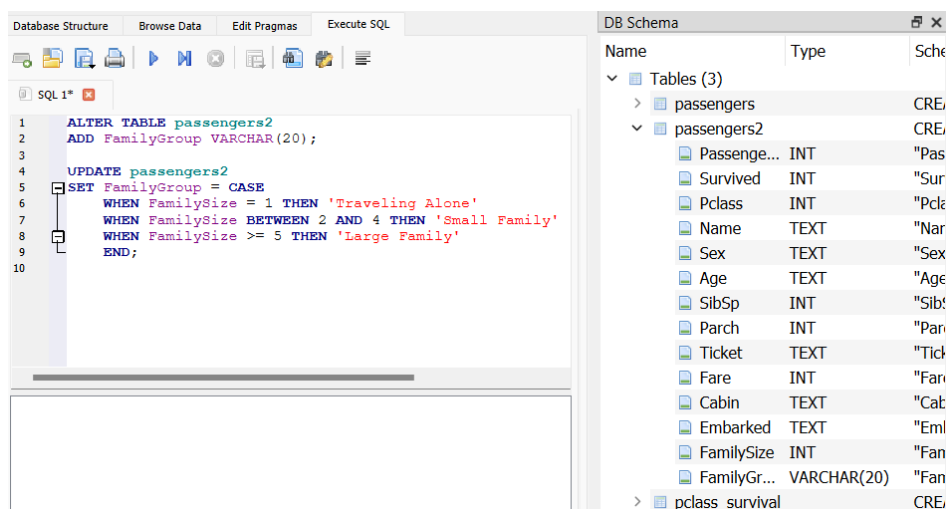
- a. For convenience purpose, the original table passengers will be duplicated and the SQL Query to do so is as below:



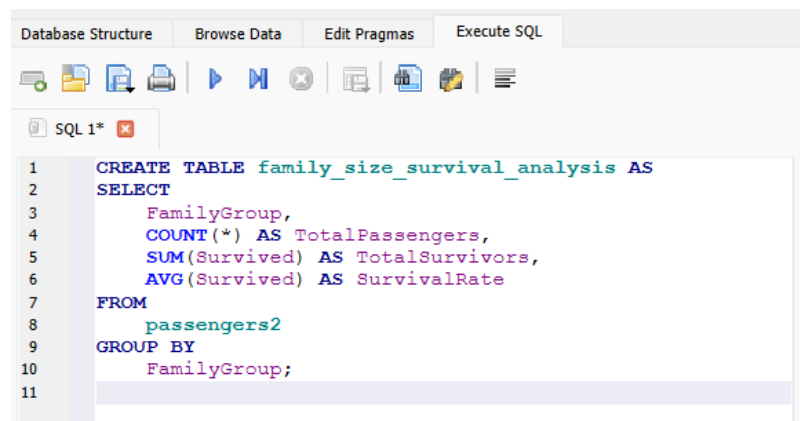
- b. Next step involved calculating family size in passenger2.



- c. Now, categorize the passengers based on their family size, if FamilySize is exactly 1, it falls into the category of “Traveling Alone”, for FamilySize between 2 and 4 that would belongs to “Small Family”, when FamilySize is 5 or larger than that would falls into the category of “Large Family”.



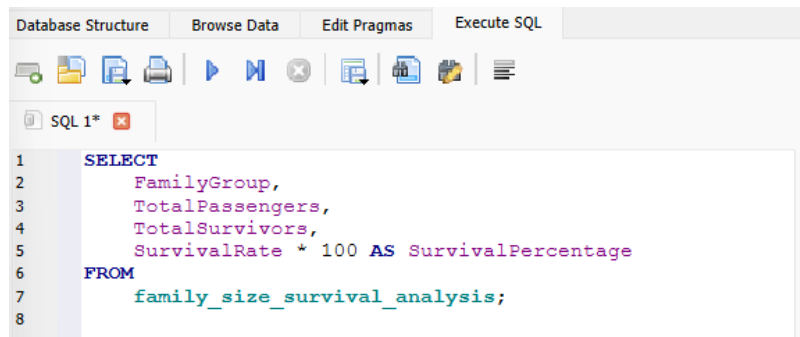
- d. Once the previous step is completed, a new table can be created for grouping passengers by how big the family which dictated by the previous SQL Query.



The screenshot shows a SQL IDE interface with tabs for 'Database Structure', 'Browse Data', 'Edit Pragmas', and 'Execute SQL'. The 'Execute SQL' tab is active, displaying a SQL query in a text editor. The query is as follows:

```
1 CREATE TABLE family_size_survival_analysis AS
2 SELECT
3     FamilyGroup,
4     COUNT(*) AS TotalPassengers,
5     SUM(Survived) AS TotalSurvivors,
6     AVG(Survived) AS SurvivalRate
7 FROM
8     passengers2
9 GROUP BY
10    FamilyGroup;
```

- e. Lastly, SQL Query for retrieving the survival rates for each type of family group and the result is as shown below:



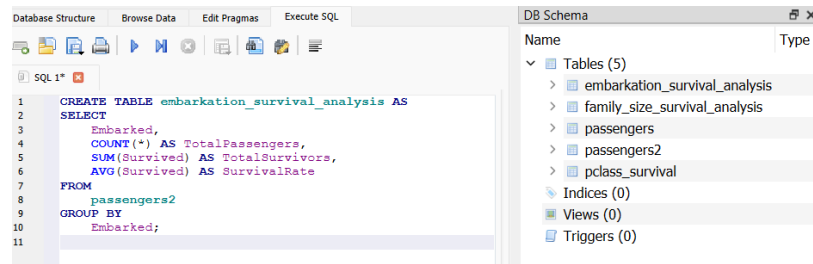
The screenshot shows the same SQL IDE interface. The 'Execute SQL' tab is active, displaying a new SQL query:

```
1 SELECT
2     FamilyGroup,
3     TotalPassengers,
4     TotalSurvivors,
5     SurvivalRate * 100 AS SurvivalPercentage
6 FROM
7     family_size_survival_analysis;
```

	FamilyGroup	TotalPassengers	TotalSurvivors	SurvivalPercentage
1	Large Family	62	10	16.1290322580645
2	Small Family	292	169	57.8767123287671
3	Traveling Alone	537	163	30.3538175046555

Question 3: Which Port of Embarkation had the Highest Survival Rate?

- a. Data field Embarked and Survived will be selected to study the relationship between survival rates and the port of embarkation, a new table is created to using the SQL Query below:



- b. Once step above is completed, next step proceeds to query result table that showcase TotalPassengers, TotalSurvivors and also SurvivalPercentage based on where the passengers embarked. The SQL Query goes like this:

The screenshot shows a SQL IDE with a 'SQL 1*' editor. The query is as follows:

```
1 SELECT
2     Embarked,
3     TotalPassengers,
4     TotalSurvivors,
5     SurvivalRate * 100 AS SurvivalPercentage
6 FROM
7     embarkation_survival_analysis;
```

Below the query editor, the results are displayed in a table:

	Embarked	TotalPassengers	TotalSurvivors	SurvivalPercentage
1	NULL	2	2	100.0
2	C	168	93	55.3571428571429
3	Q	77	30	38.961038961039
4	S	644	217	33.695652173913

Discussion of Results and Conclusions

Based on the generated outcomes, the result for Question 1 indicates that wealthier passengers 1st PClass had the highest chance of survival at about 63%, while the survival chance for 2nd PClass Passengers and 3rd PClass Passengers are 47.3% and 24.24% respectively.

	Pclass	TPC	TotalSurvivors	SurvivalPercentage
1	1	216	136	62.962962962963
2	2	184	87	47.2826086956522
3	3	491	119	24.2362525458248

As for Question 2, the outcome indicates actually showed that Passenger with Small Family actually have a drastically higher chance of survival percentage at almost 58%, solo traveler denoted by Travelling Alone comes in at second with a survival percentage of only 30.35% whereas Passenger with Large Family only stood a chance of 16% at surviving the boat capsize.

	FamilyGroup	TotalPassengers	TotalSurvivors	SurvivalPercentage
1	Large Family	62	10	16.1290322580645
2	Small Family	292	169	57.8767123287671
3	Traveling Alone	537	163	30.3538175046555

The result of Question 3 indicates Passengers who boarded at Cherbourg has the highest chance of survival at 55.4%, coming in close at second is Passengers that boarded from Queenstown at 38% of survival chance, the Passenger that boarded Titanic at Southampton had the lowest chance of survival in comparison to the other 2 ports, at only 33.7%.

	Embarked	TotalPassengers	TotalSurvivors	SurvivalPercentage
1	NULL	2	2	100.0
2	C	168	93	55.3571428571429
3	Q	77	30	38.961038961039
4	S	644	217	33.695652173913

As conclusion, since the objective of this exercise is to evaluate our skills in using SQLite, there's no need to justify or interpret the outcomes, as such interpretations can be highly subjective. In retrospect, I have successfully synthesized certain data and obtained figures that reveal intriguing trends. This achieves the goals of Exploratory Data Analysis by testing hypotheses and examining all

variables and their relationships. This process helps build a comprehensive understanding of the context surrounding the data and ensures that no potential insights are overlooked. It also involves confirming or disproving hypotheses.

While some relationships may be insignificant, the process guarantees that every angle is considered. A prime example is the relationship between survival rates and port of embarkation, which may lead to further inquiries about whether other factors are involved. For instance, passengers from one port might have been seated in different sections of the ship, raising questions about whether lifeboat deployment varied by area, ultimately contributing to a more complete understanding of survival predictions.