SQL PROJECT 1: EXPLORING TITANIC DATABASE

Understanding the Business Context

On April 15th, 1912, the famous RMS Titanic that had 2,435 passengers and 892 crew members was on its way to New York when it collided with an iceberg and sank, leaving over 700 survivors. Whether or not there were insufficient number of lifeboats that contributed to the low surviving rate, some groups have an element of luck in surviving the sinking.

The data collected in the Titanic Database gives us insights on what sorts of people are likely to survive based on their gender, age, ticket class and their relationship with other fellow boarders who board the ship with them. These data are used to make future predictions on who are likely to survive in similar situations.

Understanding the Technical Context

The Titanic Database used in this project is exclusively found from the Kaggle Titanic machine learning competition. Based on my research, Thomas Cason was the person who updated and improved the Titanic Dataset by referring to the Encyclopaedia Titanica and formed a titanic dataset from it called titanic3. I assumed the titanic dataset was from Thomas as it has a similar table and fields with the one in Kaggle's Titanic machine learning dataset. The data is not complete as there were a lot of "null" after reviewing the dataset.

Understanding the Tables and Fields

The Titanic Database consists of only one table called "passengers". The table contains all the data collected of a single passenger that boards the Titanic. The data stored in the "passengers" table are organised in a row and column format where the rows represent each passenger's unique records while the columns represent a field the record. There are twelve fields in the table namely, "PassengerID", "Survived", "Pclass", "Name", "Sex", "Age", "SibSp", "Parch", "Ticket", "Fare", "Cabin" and "Embarked".

The "PassengerID" field shows each passenger's unique ID that differs from the rest. Their ID helps us to find their data more effectively. The "Survived" field shows us the data of how many passengers survived the sink where the number "1" indicates that the passenger has survived while "0" indicates that they didn't. The "Pclass" field represents their ticket class such as first class ticket, second class ticket and third class ticket. Passengers from the first class are usually those from a high-income family while second class are those from a middle-income family and third class passengers are from the lower-income family. The "Name", "Sex" and "Age" field is the full name, gender and age of the passengers respectively while the "SibSp" field stands for Siblings or Spouse, denoting the number of siblings or spouses aboarding the Titanic with each passenger. Similarly, the "Parch" field stands for Parents or Children, denoting the number of parents or children boarding the titanic with each of the passenger. The "Ticket" field on the other hand contains data of each passenger's ticket number while the "Fare" field shows the passenger's ticket fare. While the

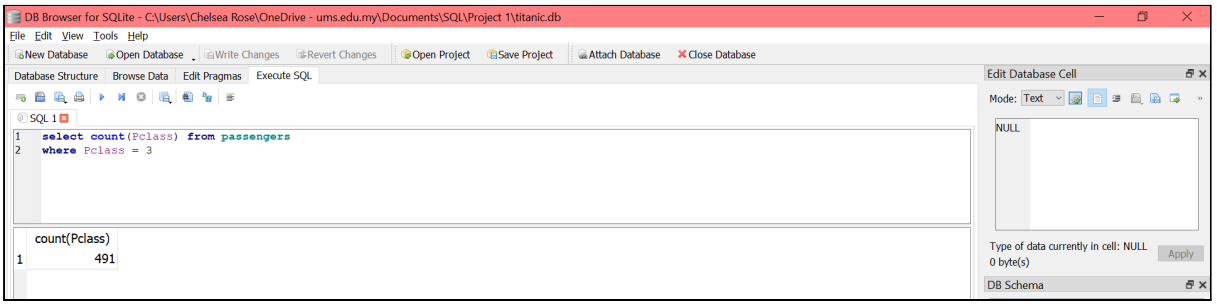
“Cabin” field shows each passenger’s cabin number, the “Embarked” field shows the port in which the passenger embarked their journey from.

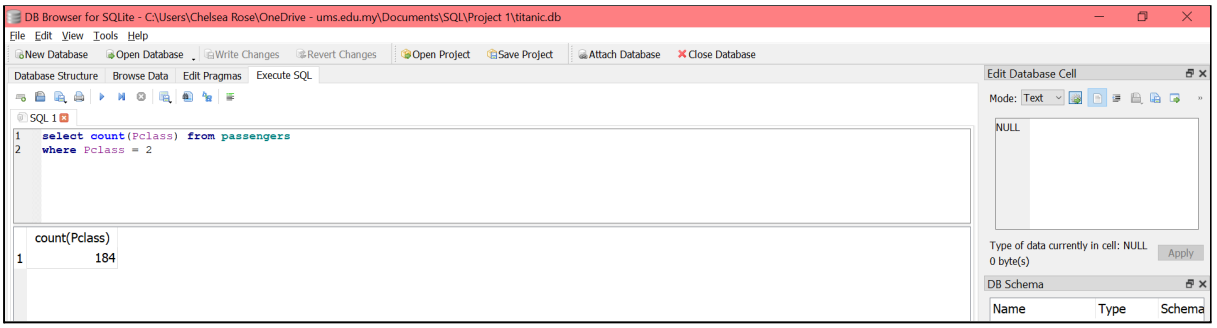
After analysing the whole table, I found out that the dataset was messy for a few reasons. One reason was that the data had a lot of missing data, especially at the “Cabin” field and the “Age” field. Other than that, in the “Fare” field, the data was not standardised. For example, some numbers are in 4 decimal places while some are in 1 or 2 decimal places. If the first fare number was rounded off to 1 decimal places, then the rest of the number should also be rounded off to 1 decimal places. This way, it will look clean. Besides that, the data in the “Ticket” field is also not standardised. Some data contains just numbers while some contains numbers and letters that indicate a location. For example, one passenger has a ticker number “35851”, while another person has a ticket number “C.A. 33595” where C.A stands for California.

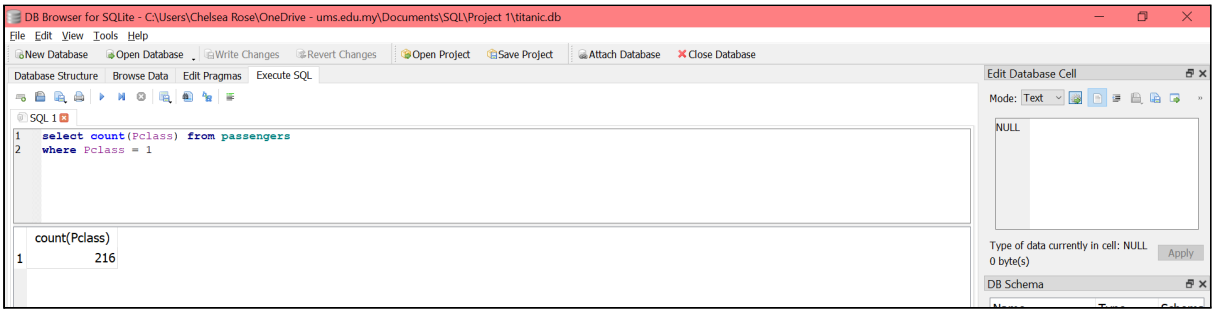
Free Exploration

Question 1:
Which passengers among the three ticket classes have more survival rate?

Firstly, I will count the total number of passengers from each ticket class.

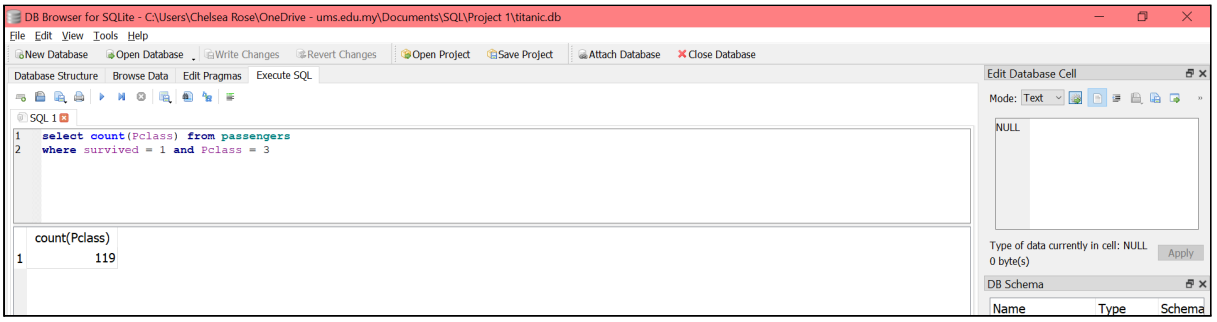




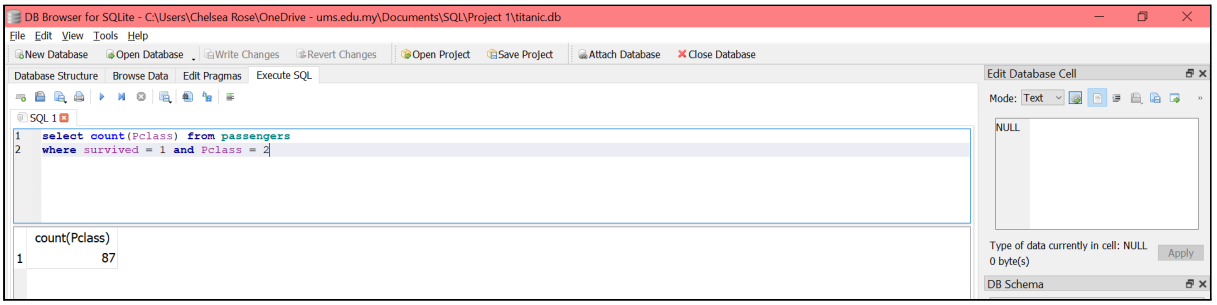


The third class ticket has a total of 491 passengers while the second class ticket has 184 passengers and the first class ticket has 216 passengers.

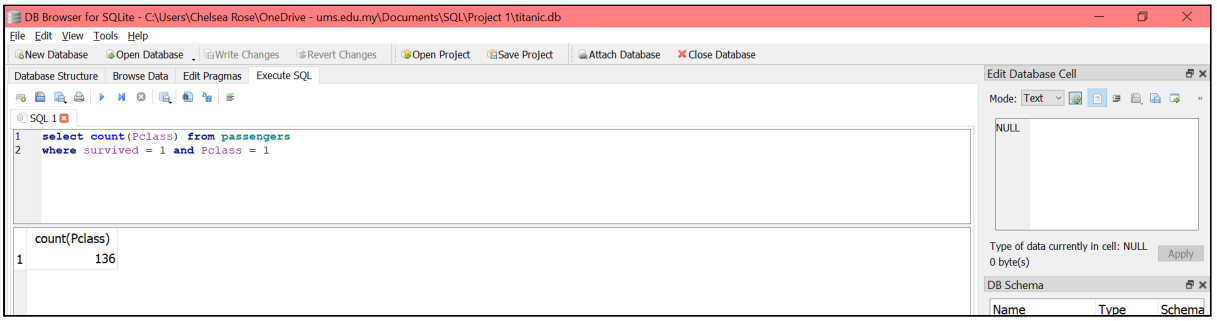
Next, I will calculate the survivors from each ticket class, which means survivors = 1.



The third ticket class has a total of **119** survivors out of **491** passengers. (24%)



As for the second ticket class, there are **87** survivors out of **184** passengers. (47%)

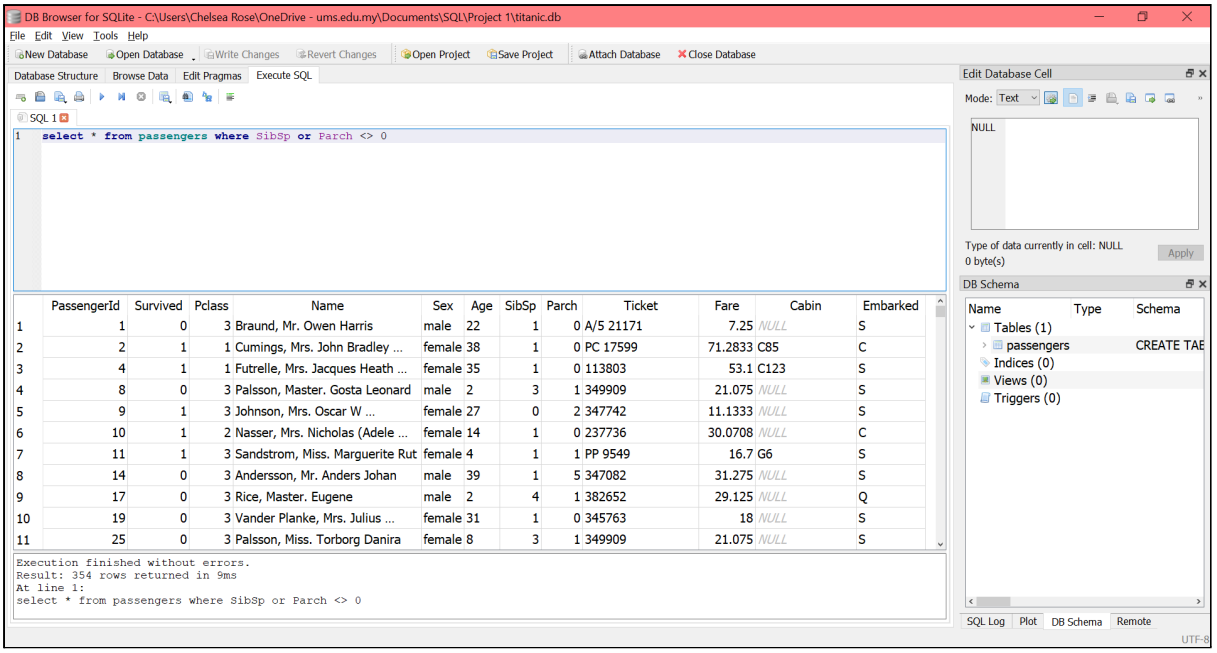


The first ticket class has a total of **136** survivors out of **216** passengers. (63%)

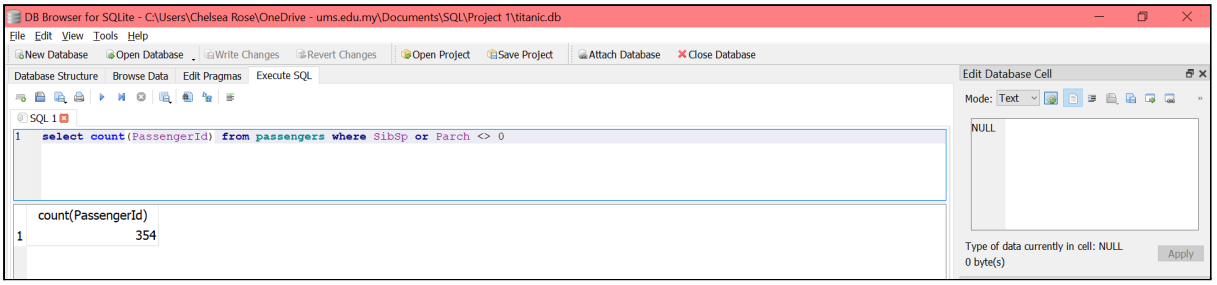
In conclusion, the passengers from the first ticket class have the highest survival rate, followed by the second ticket class and lastly, the third ticket class.

Question 2:
Those passengers with families with them, how many of them manage to survive?

First, I will filter out passengers that have at least 1 or more value in the “SibSp” or “Parch” fields by removing the passengers with 0 values in either one of the fields.

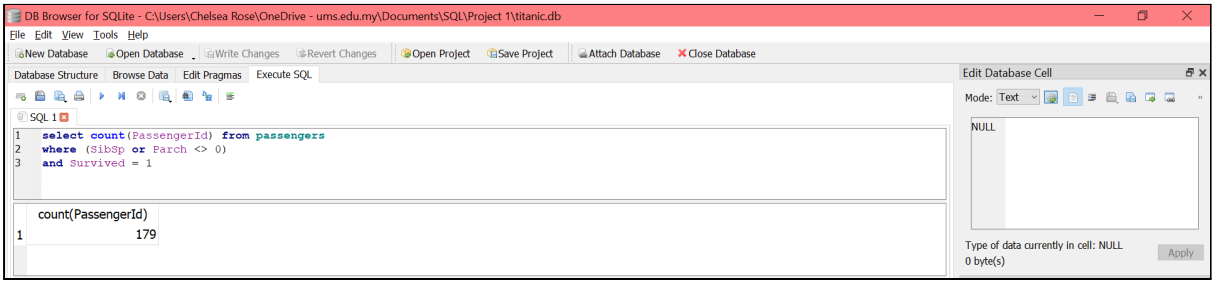


Next, I will first calculate the total number of passengers with families.



There are a total of 354 passengers with families.

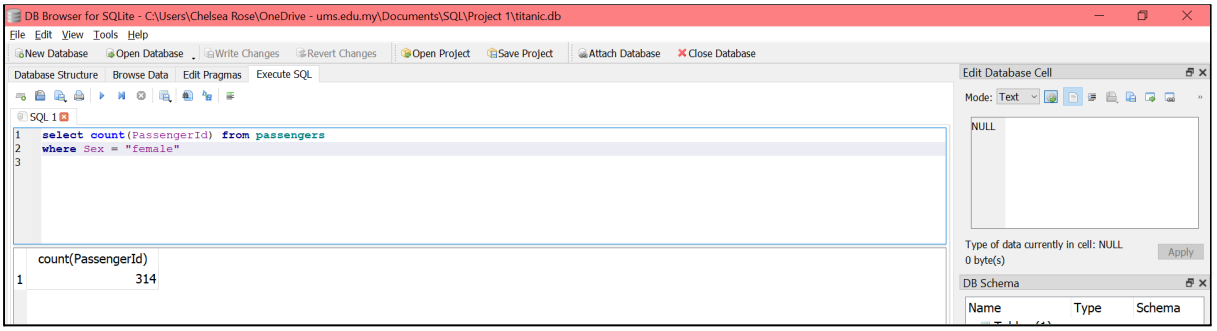
Now I will calculate the total number of passengers with families who survived.

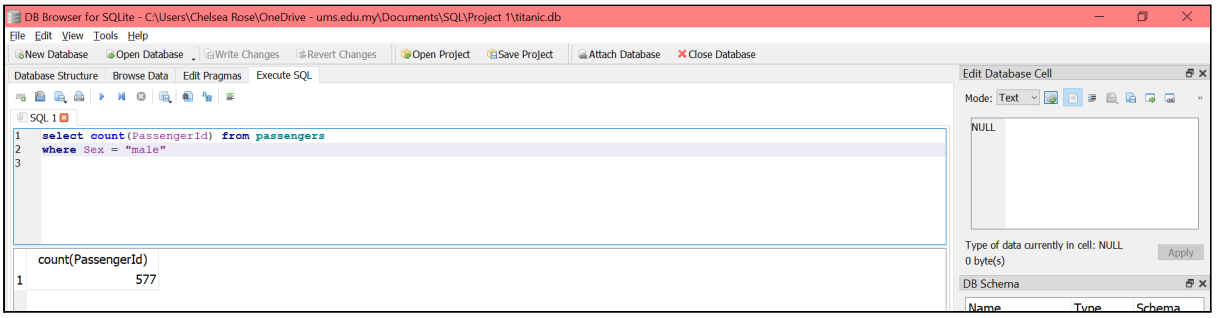


There are a total of 179 survivors out of 354 who had families boarding with them in the Titanic.

Question 3:
Between females and males, who were more likely to survive?

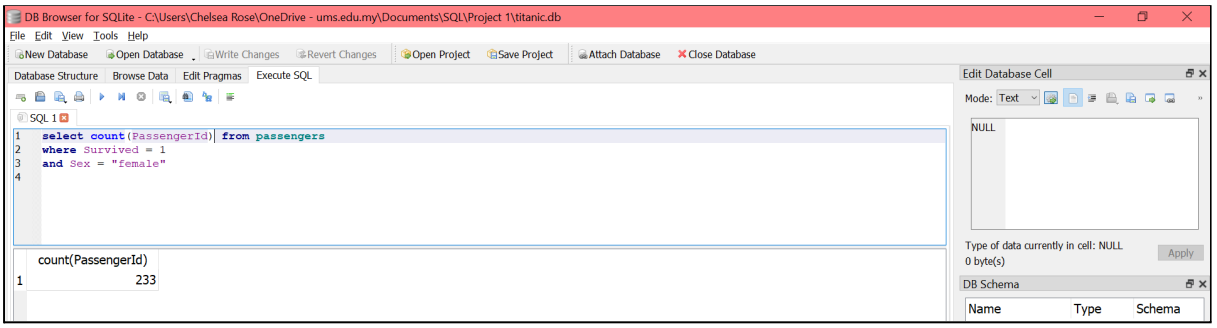
Firstly, I will calculate the total number of each gender who boarded the ship

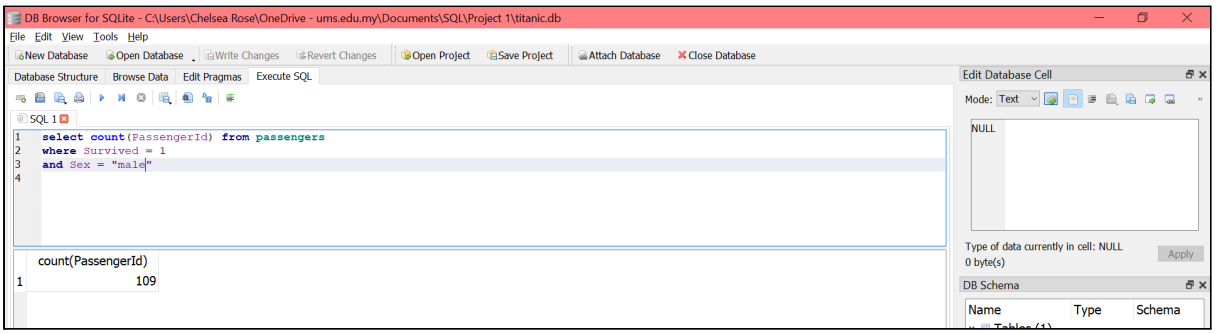




There are a total of 314 female passengers while there are 577 male passengers who boarded the ship.

Then, I'll calculate the total number of each gender who survived the sink.





There are 233 out of 314 (74%) female passengers while there are 109 out of 577 male passengers (19%) who survived the sinking. To conclude, there are more female survivors than the males.