Project 1 - Exploring Titanic Database

Step 1: Understanding the Business Context

1. What are these data for?

   Titanic database is used to find the survival rate of the victims of titanic accident based on their categories such as age, gender and status.
2. Why do we need this database?

   These data can be used in future prediction for similar situation to know who have higher possibility to survive the accident.
3. Where are these data collected?

   The data about Titanic passengers is from the Encyclopedia Titanica. The datasets used here were begun by a variety of researchers.

Step 2: Understanding the Technical Context

1. How are these data collected?

   The data were collected through manual inspection from the tragedy and recorded.
2. Where are the sources of these data?

   These data is store in Encyclopedia Titanica.
3. What are the systems that touch or use/modify these data?

   Machine learning systems or manual data change may modify these data.
4. What are some of the error sources of this data?

   The age information for the passenger contains value of post Titanic death age and not the day of disaster age as it does for other survived passengers age attribute values.
5. Is the data complete? Would there be missing pieces of data?

   The data is not complete as there are NULL value in some of the row.

Step 3: Understanding the Tables and Fields

1. How many tables do we have?

   There is only one table.

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

   The table name passengers.

3. What are the relationships between the tables?

   Since there is only one table so there are no relationship to another related tables.

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

Variable	Definition
PassengerID	The primary key or the unique identity of the passengers.
Survived	The survival status of the passengers.
Pclass	Ticket class of the passengers.
Name	Name of the passenger.
Sex	The gender of the passenger.
Age	The age of the passenger.
Sibsp	Number of siblings aboard the titanic.
Parch	Number of parents aboard the titanic.
Ticket	Ticket number.
Fare	Passenger fare.
Cabin	The cabin number of the passenger.
Embarked	The port of Embarkation.

5. Is the data messy? and how?

   The data is messy but we can ignore the NULL value in the row.

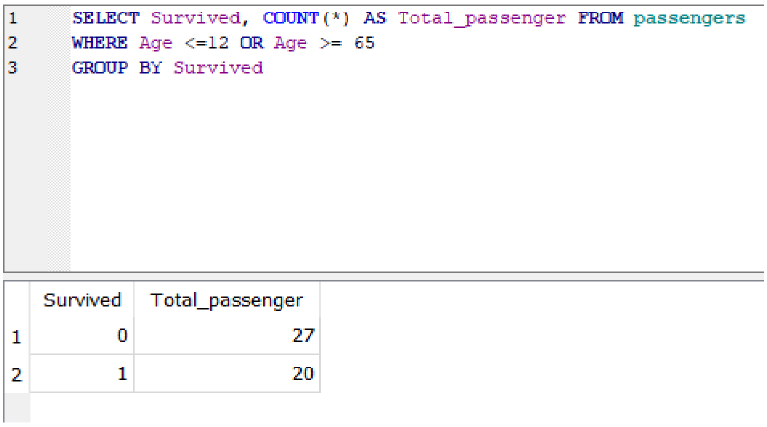
Data Exploration

1. In the movie, children, elderlies and females can get onboarded to rescue boat first, so

- Are children and elderlies have a higher survival rate in this accident?

  Assuming that children are 12 years old and below and elderlies are 65 years old and above, we need to group them together to find the survival rate.

  Firstly, we need to find the number of children and elderlies and categorized it to the one that survive and does not survive.

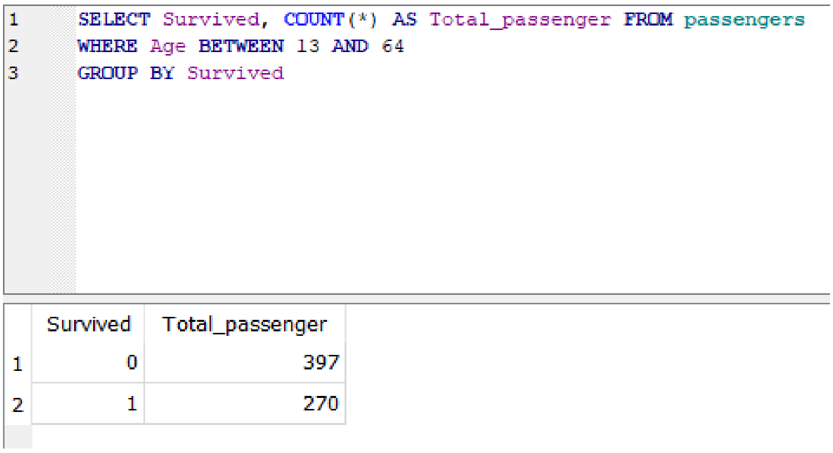


```sql
SELECT Survived, COUNT(*) AS Total_passenger FROM passengers
WHERE Age <=12 OR Age >= 65
GROUP BY Survived
```

```
     Survived   Total_passenger
1           0                27
2           1                20
```

  As we can see from the SQL output there are 20 survivors and 27 children and elderlies that died from the accident. The survival rate for children and elderlies that survive is (42.55%).

  Next, we need to compare with the other group of age between 13 and 64 which managed to survive.



```sql
SELECT Survived, COUNT(*) AS Total_passenger FROM passengers
WHERE Age BETWEEN 13 AND 64
GROUP BY Survived
```

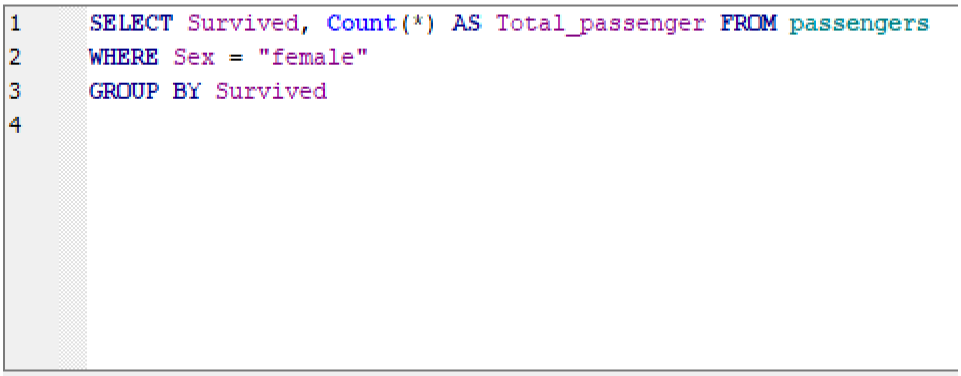
```
     Survived   Total_passenger
1           0               397
2           1               270
```

  From the SQL output, we can see there are 270 survivors and 397 people that died from the accident. The survival rate for the rest of the age group is only (40.48%).

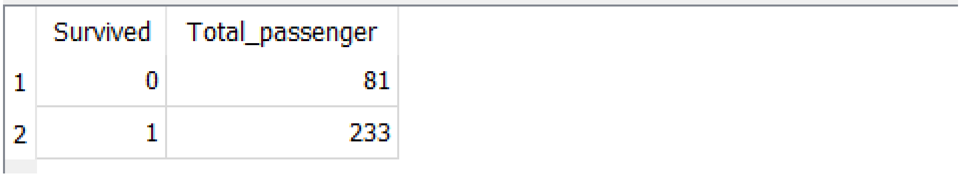
  From this result, we can conclude that children and elderlies have a higher survival rate in this accident.

- Are females more likely to survive in this incident?

Firstly, we need to find the total number of the female survivors and also the female passenger that died in the accident.



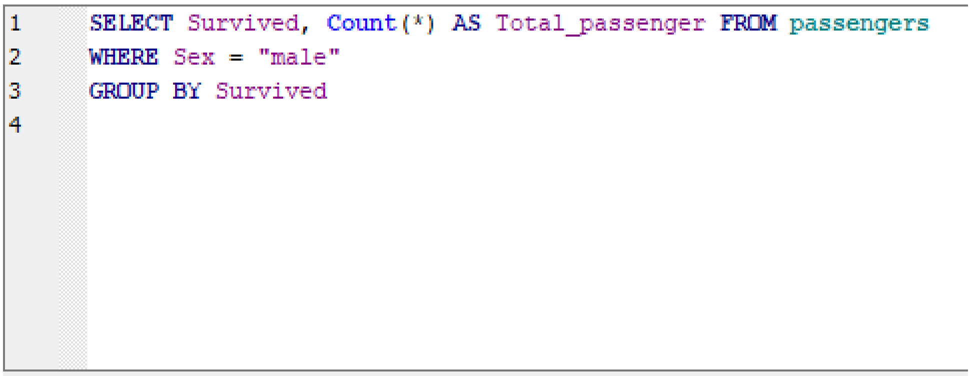
```
SELECT Survived, Count(*) AS Total_passenger FROM passengers
WHERE Sex = "female"
GROUP BY Survived

```



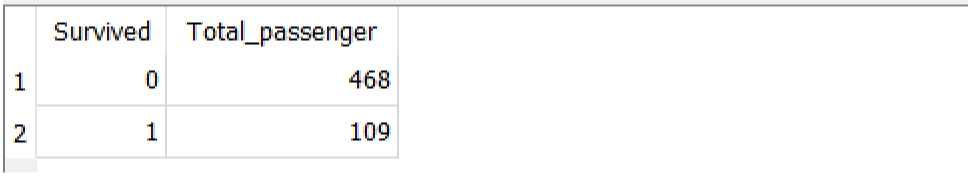
	Survived	Total_passenger
1	0	81
2	1	233

As we can see, the number of female survivors are 233 and 81 female does not survive. The survival rate of female is (74.20%).

Then, we need to find the number of male survivors as well as the male that died in the accident.



```
SELECT Survived, Count(*) AS Total_passenger FROM passengers
WHERE Sex = "male"
GROUP BY Survived

```



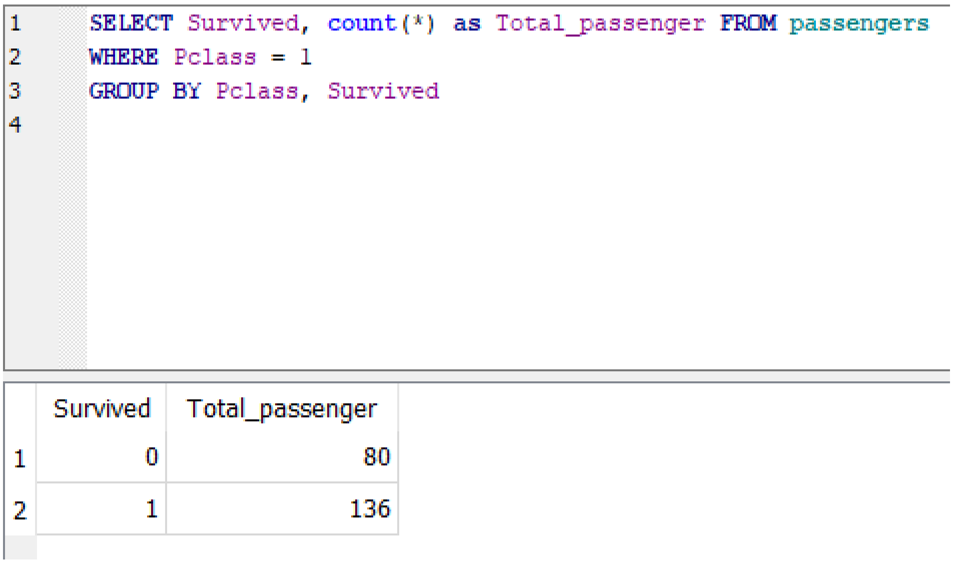
	Survived	Total_passenger
1	0	468
2	1	109

From the output, we can see the male survivors are only 109 while 468 of them did not managed to survive. The survival rate of male is only (18.89%)

As a conclusion, females are more likely to survive in this situation than males.

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

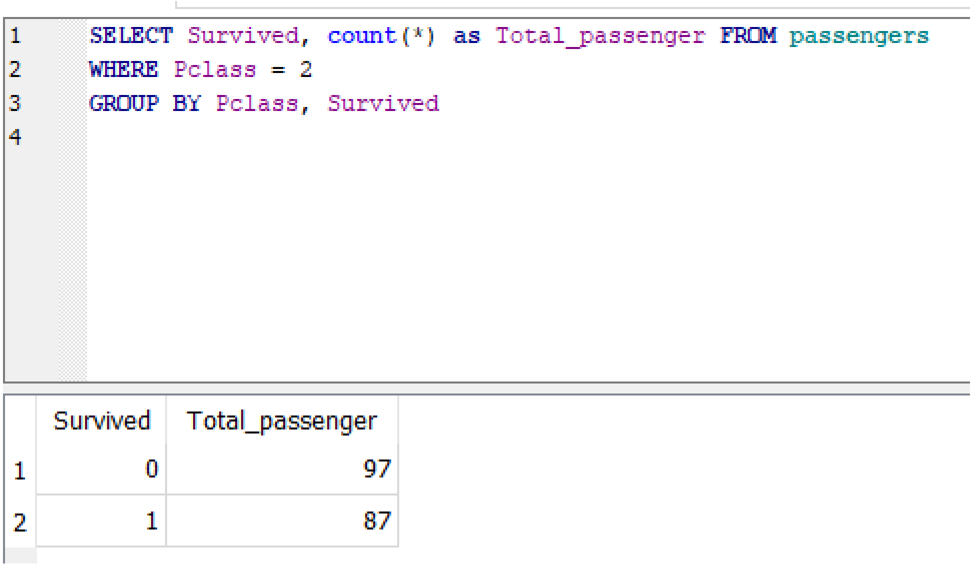
Firstly we need to find the number of passengers that survive and did not survive accordingly based on their class ticket.



```sql
SELECT Survived, count(*) as Total_passenger FROM passengers
WHERE Pclass = 1
GROUP BY Pclass, Survived
```

	Survived	Total_passenger
1	0	80
2	1	136

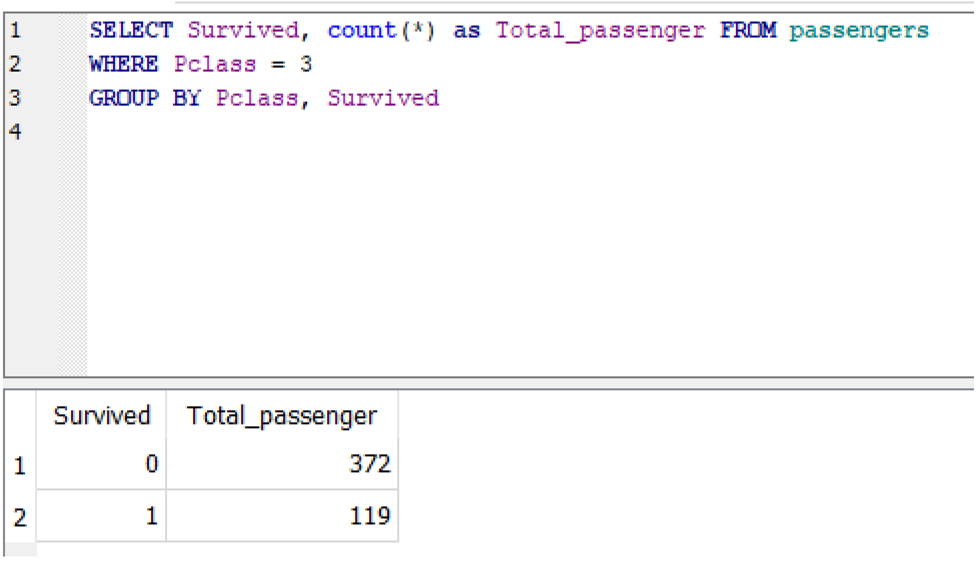
For Class 1, we can see there are 136 survivors and 80 passengers could not make it out the titanic ship. The survival rate of Class 1 is (62.96%).



```sql
SELECT Survived, count(*) as Total_passenger FROM passengers
WHERE Pclass = 2
GROUP BY Pclass, Survived
```

	Survived	Total_passenger
1	0	97
2	1	87

For Class 2, we can see there are 87 survivors and 97 passengers could not make it out the titanic ship. The survival rate of Class 1 is (47.28%).



```sql
SELECT Survived, count(*) as Total_passenger FROM passengers
WHERE Pclass = 3
GROUP BY Pclass, Survived
```

	Survived	Total_passenger
1	0	372
2	1	119

For Class 3, we can see there are 119 survivors and 372 passengers could not make it out the titanic ship. The survival rate of Class 1 is (24.24%).

We can conclude that Class 1 ticket which are rich people have the highest rate of survival in this titanic ship accident.