Understanding the Business Context

For this project, I was given the Titanic Database which is a popular database amongst data science enthusiast and professionals. The database is commonly used to describe the survival rates of passengers on the infamous R.M.S. Titanic which sank on April 15th 1912 after it collided with an iceberg. In Kaggle.com, there is a competition called "Machine Learning from Disaster" where competitors use machine learning models to predict which passengers survived the Titanic shipwreck. Based on the overview of the Titanic data by Kaggle, I found out that the R.M.S. Titanic had limited life boats which could have contributed to 1502 deaths out of the total 2224 passengers. It also appears that some groups of individuals have greater chances of survival than others. For the current project, I will be using SQL to analyse the survival rates between different types of passengers. I would also be exploring others aspects of the passengers such as their port of embarkation, and fare.

Understanding the Technical Context

The Titanic database can be download from Kaggle.com while the primary source for this data originated from Encyclopedia Titanica. The data was created by numerous researchers who dedicated many years to study the information regarding passengers of the ship. Thus, it can be assumed the data is trustworthy for this project though there are many missing values that can limit my analysis.

Understanding the Tables and Fields

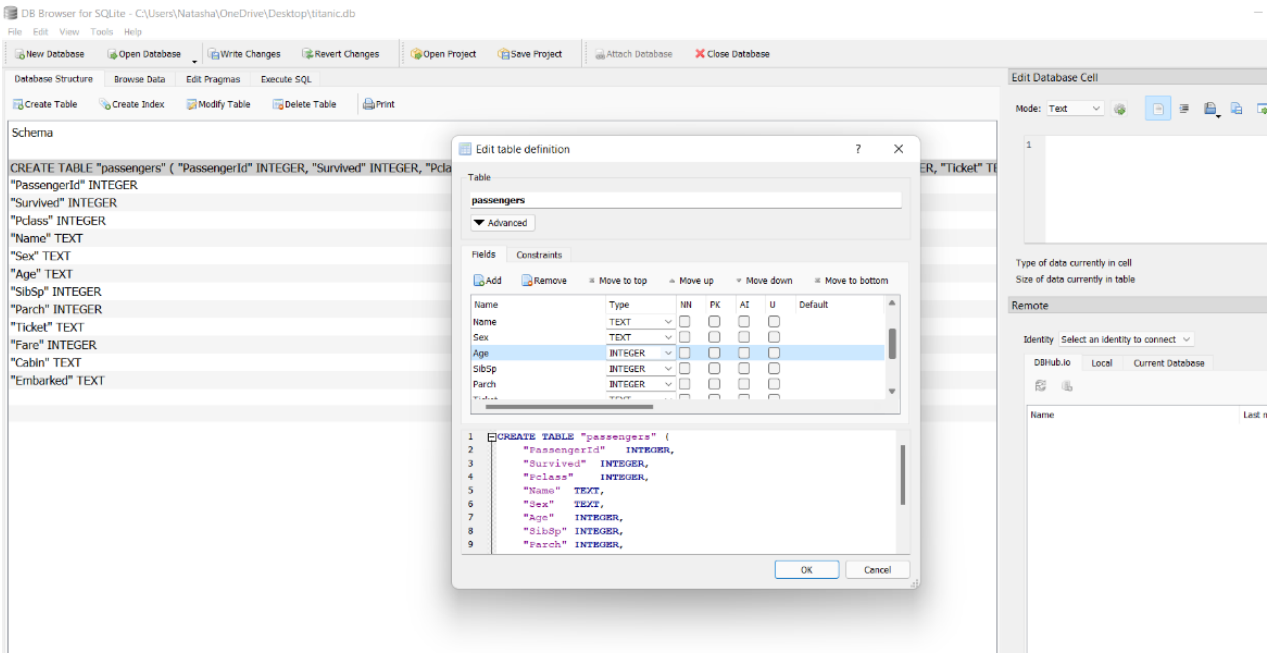
For this database there is only one table called *passengers* which contains information on 891 passengers of the R.M.S Titanic. There are 12 fields in this table which are described below:

Fields	Data type	Definition
PassengerId	Integer	Passengers identification number
Survived	Integer	Either the passenger survived the shipwreck or not. 1 being "Yes" and 2 being "No"
Pclass	Integer	Ticket class. Proxy for socio-economic status. 1 is upper, 2 is middle, and 3 is lower
Name	Text	Passenger name
Sex	Text	Passenger's gender
Age	Text	Passenger's age
Sibsp	Integer	Number of siblings or spouses on board the Titanic Siblings include brother, sister, stepbrother, stepsister. Spouses include husband and wife but not mistress nor fiancés.
Parch	Integer	Number of parents or children aboard the Titanic. Parents include mother and father. Children include daughter, son, stepdaughter, stepson. Children who travelled with their nanny are classified as Parch equal to zero.
Ticket	Text	Ticket name
Fare	Integer	Passenger fare
Cabin	Text	Cabin number
Embarked	Text	Port of embarkation C= Cherbourg, Q = Queenstown, S = Southampton

There are certain issues with regards to the table. Firstly, the cabin field has lot of NULL values thus if I decided to remove rows with null values, it will greatly impact my analysis. Fortunately, some questions that need to be answered does not require the cabin field thus I left the rows with null values as it is. Secondly, although the field of *Sibsp* and *Parch* contains number of passengers travel companions it does not include their actual relationship to the passenger. For example, it cannot be certain whether *Sibsp* for some passengers represent their spouses or siblings. Thus, I can only speculate their relationship by referring back to passengers' age. Thirdly, the *Age* field is in text datatype and this can lead to inaccurate analysis related to passenger's age. To correct this error, I needed to switch datatype from text to integer prior to analysis. Finally, it is uncertain whether the Sibsp and Parch are counted among the 891 passengers.

Free Exploration

After outlining the business context, technical context, and table of the Titanic database, I begin my exploration by modifying the table. I converted the *Age* and *fare* to integer and numeric datatypes respectively. The photo below demonstrates how I modified the datatype. Additionally, since *fare* has decimal points, I think it is appropriate to convert it to a numeric data type just for good measure.



I began exploring the data by determining the total number of passengers in Titanic table. The result of the query is the total passengers in the table being 891.

```
SELECT COUNT(*) AS total_passengers
FROM passengers
```

total_passengers
891

Then, I classified the number of passengers based on ticket class. Note that 1 is upper class, 2 is middle class, and 3 being the low class.

```
WITH
num_passenger_perclass AS
(
SELECT COUNT(*) AS total_passengers,
Pclass AS passenger_ticket_class
FROM passengers
GROUP BY Pclass
ORDER BY total_passengers DESC
)

SELECT passenger_ticket_class,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM
passengers),2) AS percent
FROM num_passenger_perclass
```

passenger_ticket_class	total_passengers	percent
3	491	55.11
1	216	24.24
2	184	20.65

From here we can imply that the majority of the passengers are of the low-income status at 55.11%.

I moved on to identifying total number of male and female passengers

```
WITH
gender_table AS
(
SELECT COUNT(*) AS total_passengers,
sex
FROM passengers
GROUP BY sex
ORDER BY total_passengers DESC
)

SELECT sex AS gender,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM
passengers),2) AS percent
FROM gender_table
```

gender	total_passengers	percent
male	577	64.76
female	314	35.24

Based on the result of this query, 64.78% of the passengers are of male gender hence there are more male passengers as compared to female passengers.

On the site note, I also explored on the number of passengers based on port of embarkation. Note that C is Cherbourg, Q is Queenstown, and S is Southampton.

```
SELECT COUNT(*) AS total_passengers,
Embarked AS port__of_embarkation
FROM passengers
GROUP BY Embarked
ORDER BY total_passengers DESC
```

total_passengers	port_of_embarkation
644	S
168	C
77	Q
2	NULL

Majority of the passengers embarked from Southampton port.

Then I determined the percentage of survival status of passenger from each port of embarkation. Note that for *Survived* 1 is Yes and 0 is No

```
WITH
embark_table AS
(
SELECT COUNT(*) AS total_passengers,
Embarked AS port_of_embarkation,
Survived
FROM passengers
GROUP BY Embarked, Survived
)
SELECT port_of_embarkation,
Survived AS survival_status,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM
passengers),2) AS percent
FROM embark_table
ORDER BY total_passengers DESC
```

port_of_embarkation	survival_status	total_passengers	percent
S	0	427	47.92
S	1	217	24.35
C	1	93	10.44
C	0	75	8.42
Q	0	47	5.27
Q	1	30	3.37
NULL	1	2	0.22

Upon further examination, majority of casualties (47.92%) and survived (24.35%) are those embarked from Southampton port.

Next, I check on the total of passengers that travel with either siblings or spouses

```
WITH
siblings_spouse_table AS
(
SELECT COUNT(*) AS total_passengers_sbsp
FROM passengers
WHERE Sibsp >= 1
)

SELECT total_passengers_sbsp,
ROUND(total_passengers_sbsp*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers),2) AS percent_passengers_with_sbsp
FROM siblings_spouse_table
```

total_passengers_sbsp	percent_passengers_with_sbsp
283	31.76

There are total of 283 passenger that travel with siblings or spouses which is 31.76% from the total passengers.

Then, I checked for the total number of passengers that travel parents or children

```
WITH
parent_children_table AS
(
SELECT COUNT(*) AS total_passengers_with_parch
FROM passengers
WHERE Parch >=1
)
SELECT total_passengers_with_parch,
ROUND(total_passengers_with_parch*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers),2) AS percent_passengers_with_parch
FROM parent_children_table
```

total_passengers_with_parch	percent_passengers_with_parch
213	23.91

Result of this query showed 213 passengers which is 23.91% from overall passengers that travelled with parents or children.

Using CASE statement for Sibsp, and Parch to determine if travelling with family members results with higher fare rate.

```
SELECT passengerId, fare, Pclass, cabin,
CASE
WHEN Sibsp >= 1 THEN 'tarvelling with sibling or spouse'
ELSE 'not with sibling or spouse'
END AS sibling_spouse,
CASE
WHEN Parch >= 1 THEN 'travellihg with parent or children'
ELSE 'not with parent or children'
END AS parent_chilren
FROM passengers
ORDER BY Fare DESC
```

PassengerId	Fare	Pclass	Cabin	sibling_spouse	parent_chilren
259	512.3292	1		not with sibling or spouse	not with parent or children
680	512.3292	1	B51 B53 B55	not with sibling or spouse	travellihg with parent or children
738	512.3292	1	B101	not with sibling or spouse	not with parent or children
28	263	1	C23 C25 C27	tarvelling with sibling or spouse	travellihg with parent or children
89	263	1	C23 C25 C27	tarvelling with sibling or spouse	travellihg with parent or children
342	263	1	C23 C25 C27	tarvelling with sibling or spouse	travellihg with parent or children
439	263	1	C23 C25 C27	tarvelling with sibling or spouse	travellihg with parent or children
312	262.375	1	B57 B59 B63 B66	tarvelling with sibling or spouse	travellihg with parent or children
743	262.375	1	B57 B59 B63 B66	tarvelling with sibling or spouse	travellihg with parent or children
119	247.5208	1	B58 B60	not with sibling or spouse	travellihg with parent or children

The table above is a snapshot of the query results. Based on the result, having travelling companion can indirectly increase the fare rate since passengers require additional cabins. It is also implied that different types of cabins have different charges as well.

Now, I began to focus on passengers' survival rate. I first determine the number of survivors and number of deaths in the table.

```
WITH
survival_status AS
(
SELECT COUNT(*) AS total_passengers,
Survived
FROM passengers
GROUP BY Survived
)
SELECT Survived,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM
passengers),2) AS percent
FROM survival_status
```

Survived	total_passengers	percent
0	549	61.62
1	342	38.38

Based on the result, 61.62% of passengers perished in the shipwreck while the remaining 38.38% of passengers survived.

I then identified the maximum, minimum, and average age of passengers for groups that either survived or perished

```
SELECT MAX(age), MIN(age), ROUND(AVG(age), 2),
survived
FROM passengers
GROUP BY survived
```

max_age	min_age	avg_age	Survived
74	1	30.63	0
80	0.42(5 months old)	28.34	1

However, the results showed that there is not much of any difference in age with regards to survival status.

To the answer the question if children and elderlies have a higher survival rate in the shipwreck, I created a temporary table to calculate the total number of passengers that survived based on ages groups which are 'children group', 'elderly group', and 'other age group'. Then, I calculate the percentage of survivors for each age groups. As a reference;

Children: 0-12 years old
Adolescence: 13 -18 years old
Adult: 19 – 59 years old
Elderly: 60 years old and above

```
WITH
age_survivor AS
(
 SELECT COUNT(*) AS total_passengers,
 CASE
 WHEN age <= 12 THEN 'children group'
 WHEN age >= 60 THEN 'elderly group'
 WHEN age BETWEEN 13 AND 59 THEN 'other age group'
 END AS age_groups
 FROM passengers
 WHERE Survived = 1
 GROUP By age_groups
)
SELECT age_groups,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM
passengers WHERE Survived =1),2) AS percent
FROM age_survivor
```

age_groups	total_passengers	percent
NULL	52	15.2
children group	40	11.7
elderly group	7	2.05
other age group	243	71.05

If we exclude passengers with null age values, children (11.7%) and elderly (2.05%) passengers collectively have lower survival rate as compared to passengers between ages 13 to 59 which make about 71.05% of total survivors.

The next question is if females more likely to survive the incident. The query is as follows;

```
WITH
gender_survival As
(
SELECT survived AS survival_status,
sex AS gender,
COUNT(*) AS total_passengers
FROM passengers
GROUP BY survived, gender
)
SELECT gender,
survival_status,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers),2) AS
percent
FROM gender_survival
```

gender	survival_status	total_passengers	percent
female	0	81	9.09
male	0	468	52.53
female	1	233	26.15
male	1	109	12.23

Around 26.15% of total passengers comprised of female passengers while on 12.23 of total passengers are survival who are male. On the other hand, only 9.09% of total passengers are female casualties, a stark contrast to the 52.53% which are male casualties. Thus, I can conclude that female passengers are more likely to survive the incident as compared male passengers.

I then decided to perform a detailed analysis by comparing between genders and ticket class type amongst passengers who survived.

```
WITH
gender_survival As
(
SELECT survived AS survival_status,
sex AS gender,
pclass AS ticket_class,
COUNT(*) AS total_passengers
FROM passengers
WHERE survived = 1
GROUP BY  gender, pclass
)
SELECT
gender,
ticket_class,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers WHERE
Survived =1),2) AS percent
FROM gender_survival
ORDER BY percent
```

gender	ticket_class	total_passengers	percent
female	1	91	26.61
female	3	72	21.05
female	2	70	20.47
male	3	47	13.74
male	1	45	13.16
male	2	17	4.97

From this analysis, the most survivors are those of upper income class female passengers with 26.61% while middle and income income class female passengers have slight lower rate of survival at 20.47% and 21.05% respectively. The least survived passengers are those of male middle income class with only 4.97%.

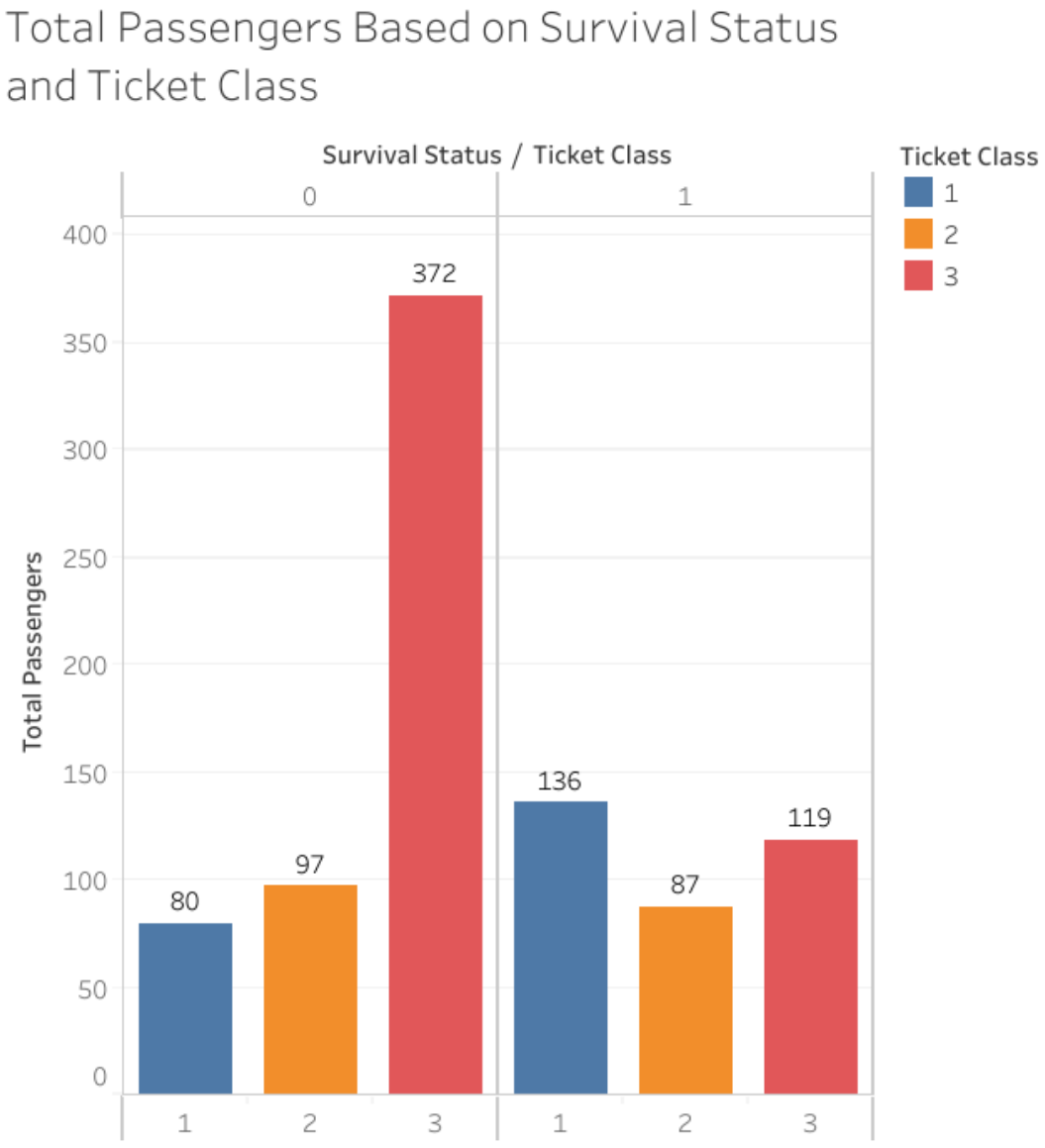
Afterwards, I determine whether rich people have higher survival rate because they can get onboard the rescue boat sooner. To do this, I first had to determine the mortality and survival rate between ticket class groups.

```
WITH
passenger_survival As
(
SELECT survived AS survival_status,
pclass AS ticket_class,
COUNT(*) AS total_passengers
FROM passengers
WHERE survived = 1
GROUP BY pclass
)
SELECT
ticket_class,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers WHERE Survived =1),2) AS percent
FROM passenger_survival
ORDER BY percent DESC
```

ticket_class	total_passengers	percent
1	136	39.77
3	119	34.8
2	87	25.44

```
WITH
passenger_survival As
(
SELECT survived AS survival_status,
pclass AS ticket_class,
COUNT(*) AS total_passengers
FROM passengers
GROUP BY pclass, Survived
)
SELECT survival_status,
ticket_class,
total_passengers,
ROUND(total_passengers*100.0/(SELECT COUNT(*) AS total_passengers FROM passengers),2) AS percent
FROM passenger_survival
```

survival_status	ticket_class	total_passengers	percent
0	1	80	8.98
1	1	136	15.26
0	2	97	10.89
1	2	87	9.76
0	3	372	41.75
1	3	119	13.36



I created a bar chart visualize the results. Based on the chart, the first-class ticket passengers have the highest survival rate at 15.26%. and lowest mortality rate at 8.98%. However, there is 41.75% of mortality rate amongst the third-class passengers which is considerably high which is a stark contrast to their survival rate at only 13.36%. Based on this, it seems that first class passenger had easier access to the rescue boats as compared second and thir- class passengers.

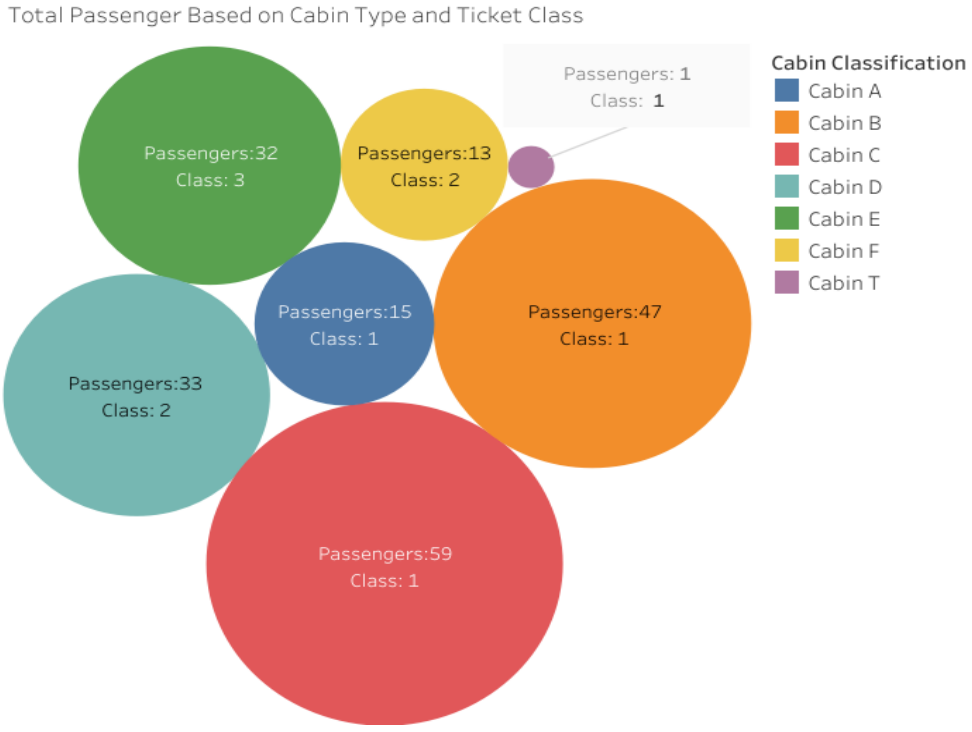
I then considered to explore on the cabin types as I suspect that different ticket classes meant different cabin types and indirectly, the locations of the passengers on board the ship. Passenger locations may impact on how they fast they can get to rescue boat as the ship was sinking. The following query and results are as below.

```
WITH
cabin_classification_table AS (
SELECT *,
CASE
WHEN Cabin LIKE 'A%' THEN 'Cabin A'
WHEN Cabin LIKE 'B%' THEN 'Cabin B'
WHEN Cabin LIKE 'C%' THEN 'Cabin C'
WHEN Cabin LIKE 'D%' THEN 'Cabin D'
WHEN Cabin LIKE 'E%' THEN 'Cabin E'
WHEN Cabin LIKE 'F%' THEN 'Cabin F'
WHEN Cabin = 'T' THEN 'Cabin T'
ELSE 'cabin unknown'
END AS cabin_classification
FROM passengers
)

SELECT COUNT(*) AS total_passengers,
cabin_classification,
Pclass AS ticket_class,
FROM cabin_classification_table
GROUP BY cabin_classification, ticket_class
ORDER BY total_passengers DESC
```

Using the cabin_classsification table, I checked on the total number of based on their cabin type and ticket class.

total_passengers	cabin_classification	ticket_class
483	cabin unknown	3
168	cabin unknown	2
59	Cabin C	1
47	Cabin B	1
40	cabin unknown	1
29	Cabin D	1
25	Cabin E	1
15	Cabin A	1
8	Cabin F	2
5	Cabin F	3
4	Cabin D	2
4	Cabin E	2
3	Cabin E	3
1	Cabin T	1



According to the query results, first ticket class passengers typically stay in Cabin A, B, and C. Second class passengers typically stay in D and F cabins. It is unknown where the majority of the third-class passengers stay and only those with their cabin records shown to have stayed in E type cabins. Based on this, it could be speculated that since third class passengers represents those on low economic status, the majority of them stayed in lower decks of E and F. As the ship was sinking the third-class passengers in lower decks struggled to get the rescue boats on time due to longer distance and large crowds. On the other hand, since the route to the rescue boats is shorter and small number of people, majority of the first-class passengers and almost have of the second-class passengers were able to make to the boats on time.

Conclusion

Out of 891 total passengers in the *passengers* only 342 passengers survived the shipwreck which is a considerable large number of casualties. Majority, of these casualties constitutes male and lower ticket class passengers. Children and elderly have collectively lower survival rate as compared to the other age group members. Furthermore, upper-ticket class passenger shown to have higher survival rate due to being able to reach the rescue boats sooner. The number of casualties could have been reduced by implementing more life boats on the ship, having stronger measures to guide all passengers into the deck in time to escape, and stricter rules on to allow certain type of passengers (children, elderly, and female) to enter the life boats first regardless of socio-economic status.