**Data Exploration: Titanic**

**Understanding the Business Context**

The shipwreck of Titanic happened over 100 years ago where it had collided with an iceberg. Due to the lack of lifeboats onboard, it had resulted in the death of a number passengers. These data give us information about the passengers such as their age, number of siblings or spouse onboard, ticket class, sex and whether they survived the shipwreck. These datasets allow people to understand the factors that might affect the passengers’ survivability. These data were collected in the USA.

**Understanding the Technical Context**

This dataset was obtained from Kaggle and the sources for the datasets from Kaggle are unknown. However, it was found that the datasets were based on the data available as of 2nd August 1999. In terms of error sources of this data, one of them would be human error which was transcription error where data was not recorded correctly. In this case, some of the values for the titanic dataset was not complete.

**Understanding the Tables and Fields**

Upon opening the database in SQLite, there is only one table which represents the details of the passengers that were onboard the Titanic. There are 12 fields where it contains the passenger’s name, sex, age, number of siblings or spouses aboard, number of parents or children abord, ticket number, passenger’s fare, cabin number, ticket class, port of embarkation and whether they survived the shipwreck.

For the ‘survival’ field, 0 represents no and 1 represents yes. For the ‘pclass’ field, 1 represents 1st class, 2 represents 2nd class and 3 represents 3rd class. For ‘embarked’ field, C represents Cherbourg, Q represents Queenstown and S represents Southampton. For ‘parch’ field, it represents the number of parents or children onboard. Some children travelled with nannies; therefore, ‘parch’ field would be 0.

Some of the information from the dataset are missing such as for age, cabin and embarked. There are 177 missing values for age, 687 missing values for cabin and 2 missing values for embarked. Data should be cleaned before proceeding to do other analysis but right now data fields that are null for age will be ignored. Other fields such as embarked and cabin will be ignored as well as they will not be used to answer the research questions.

Before proceeding, data wrangling was done whereby the data type for ‘age’ was changed into ‘INTEGER’ as it was ‘TEXT’.

**Free Exploration**

Upon studying the titanic database, there are a few research questions that I would like to propose which are:

1. Did more women survive the shipwreck?

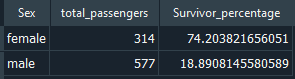
2. Are the survivors younger compared to those who perished?

3. Does a person’s social economic status play a role in ensuring their survivability?

4. Do children accompanied by parents have a higher chance of survivability compared to those accompanied by nannies?

5. What is the youngest and oldest age to survive?

*Question 1: Did more women survive the shipwreck?*



It seems like there were more male passengers onboard compared to women. In terms of survival rate, women had a higher survival rate, 74.20% compared to men which was 18.89%. Therefore, more women survived the shipwreck compared to men.

Query Used:

SELECT sex, COUNT(passengerID) as total\_passengers, SUM(Survived)\* 100 / CAST(COUNT(\*) AS FLOAT) AS Survivor\_percentage

FROM passengers

GROUP BY sex

*Question 2: Are the survivors younger compared to those who perished?*



The average age for those who had survived is 28 so it can be said that the survivors were younger compared to those who had perished.

Query Used:

SELECT survived, AVG(age) AS average\_age

FROM passengers

GROUP BY survived

*Question 3: Does a person’s social economic status play a role in ensuring their survivability?*



Based on the output, those who were in the first class had 62.96% of surviving, followed by those in 2nd class which was 47.28% and followed by those in the 3rd class which was 24.24%. There is a downward trend whereby the chances of survivability decrease as the classes decreases. Therefore, it can be said that social economic status does play a role in ensuring survivability.

Query Used:

SELECT pclass, SUM(survived)\*100 / CAST(COUNT(\*) AS FLOAT) AS survivors\_percentage

FROM passengers

GROUP BY pclass

*Question 4: Do children accompanied by parents have a higher chance of survivability compared to those accompanied by nannies?*



 Based on the two outputs above, it can be seen that children who were accompanied by parents (parch > 1) had a higher survival rate which was 55.56% compared to children who were accompanied by nannies (parch = 0), 50%. This can be due to the average age of children with parents being younger (6 years old) compared to those with nannies (15 years old) as seen in the output below:

Query Used:

**Children accompanied by parents:**

SELECT parch, COUNT(\*) as total\_children, SUM(survived)\*100 / CAST(COUNT(\*) AS FLOAT) AS survivors

FROM passengers

WHERE parch > 0 and age < 18

**Average age of children accompanied by parents:**

SELECT COUNT(age) as total\_children, AVG(age) as average\_age\_accompanied\_parents

FROM passengers

WHERE survived = 1 and parch > 1 and age < 18

**Children accompanied by nannies:**

SELECT parch, age, COUNT(\*) as total\_children, SUM(survived)\*100 / CAST(COUNT(\*) AS FLOAT) AS survivors

FROM passengers

WHERE parch = 0 and age < 18

**Average age of children accompanied by nannies:**

SELECT COUNT(age) as total\_children, AVG(age) as average\_age\_accompanied\_nannies

FROM passengers

WHERE survived = 1 and parch = 0 and age < 18

*Question 5: What is the youngest and oldest age to survive?*



Based on the output, the youngest to survive is 0.42 years old and the oldest to survive is 80 years old.

Query Used:

SELECT MAX(age), MIN(age)

FROM passengers

WHERE survived = 1