PROJECT 1-EXPLORING TITANIC DATABASE

1. Understanding the Business Context

RMS Titanic sank in the North Atlantic Ocean on 15 April 1912 after striking an iceberg during its maiden voyage from England to New York City. This tragedy cost many lives of the passengers and crews and only a few survived.

This project will perform an exploratory data analysis of the titanic dataset to know what sorts of people were more likely to survive. What factors that contribute to the survivability of the passengers. This project will be using Titanic database that comes from the Kaggle website. The database contains data and records of each passenger on the ship that will be used in finding the pattern of the data and make conclusion of the factors that may contribute to the survival of the passengers.

2. Understanding the Technical Context

The data use in this project come from Kaggle website. Most of the Kaggle data available are publicly available datasets, datasets posted by Companies, its partners, such as Google, Zillow, and Microsoft, or datasets that are shared by individuals so others can try their hands on and share their ideas on working on the data. Most of the column in the data set have complete records except for column Age, Cabin and Embarked. Data cleaning need to be done before exploration starts to ensure the accuracy of the analysis.

3. Understanding the Tables and Fields

The titanic dataset only contains one table called “passengers”. There are 12 columns in this table representing the records for each passenger. This table does not establish any relationship with other table since it is the only table exist and no primary key and foreign key was determined in this table. Table below shows the data field, data type of each field and metadata of each field.

Data Field	Definition	Key
PassengerID - INTEGER	ID of each passenger	
Survived - INTEGER	Survival status of passenger	0 = No, 1 = Yes
Pclass - INTEGER	A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower	1 = 1st, 2 = 2nd, 3 = 3rd
Name - TEXT	Passenger’s name	

Sex - TEXT	Passenger’s gender	
Age - TEXT	Age in years Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5	
SibSp - INTEGER	# of siblings / spouses aboard the Titanic Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)	
Parch - INTEGER	# of parents / children aboard the Titanic Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.	
Ticket - TEXT	Ticket number	
Fare - INTEGER	Passenger fare	
Cabin - TEXT	Cabin number	
Embarked - TEXT	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

There are some missing records in this dataset. A query was executed to find the percentage of missing data in every column.

```
SELECT
        100 * SUM(CASE WHEN PassengerId IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS id_missing,
        100 * SUM(CASE WHEN Survived IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS survived_missing,
        100 * SUM(CASE WHEN Pclass IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS pclass_missing,
        100 * SUM(CASE WHEN Name IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS name_missing,
        100 * SUM(CASE WHEN Sex IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS sex_missing,
        100 * SUM(CASE WHEN Age IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS age_missing,
        100 * SUM(CASE WHEN SibSp IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS sibsp_missing,
        100 * SUM(CASE WHEN Parch IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS parch_missing,
        100 * SUM(CASE WHEN Ticket IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS ticket_missing,
        100 * SUM(CASE WHEN Fare IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS fare_missing,
        100 * SUM(CASE WHEN Cabin IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS cabin_missing,
        100 * SUM(CASE WHEN Embarked IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS embarked_missing
FROM passengers
```

id_missing	survived_missing	pclass_missing	name_missing	sex_missing	age_missing	sibs_missing	parch_missing	ticket_missing	fare_missing	cabin_missing	embarked_missing
0	0	0	0	0	19	0	0	0	0	77	0

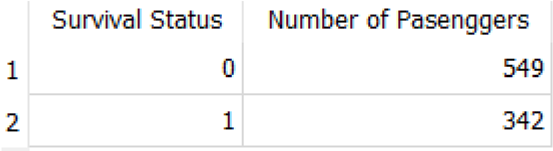
From the result, “Cabin” has 77% missing, with this many empty values, removing this variable from any further exploration data analysis seems reasonable. “Age” also has missing records but still low percentage and acceptable.

4. Free Exploration

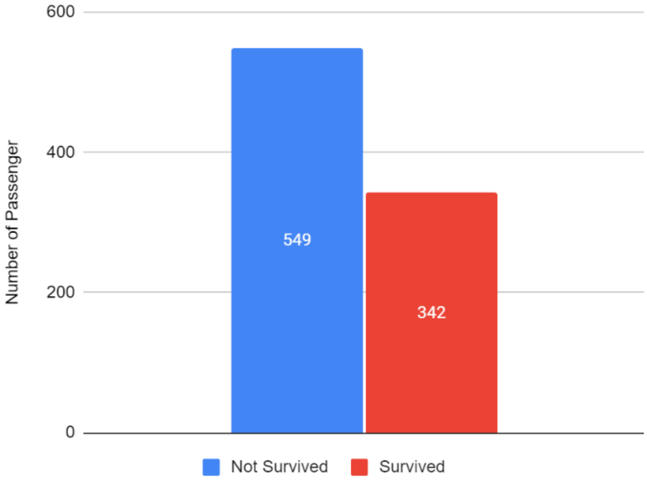
Overall survival

- Find the number of passengers survived or not survived.

```
SELECT Survived AS "Survival Status", count(Survived) AS "Number of Pasenggers"
FROM passengers
WHERE Survived = 1
GROUP BY Survived
UNION
SELECT Survived AS "Survival Status", count(Survived) AS "Number of Pasenggers"
FROM passengers
WHERE Survived = 0
GROUP BY Survived
```



	Survival Status	Number of Pasenggers
1	0	549
2	1	342

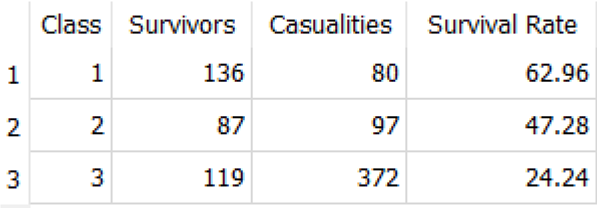


This graph shows that the number of people died is more than people that survived.

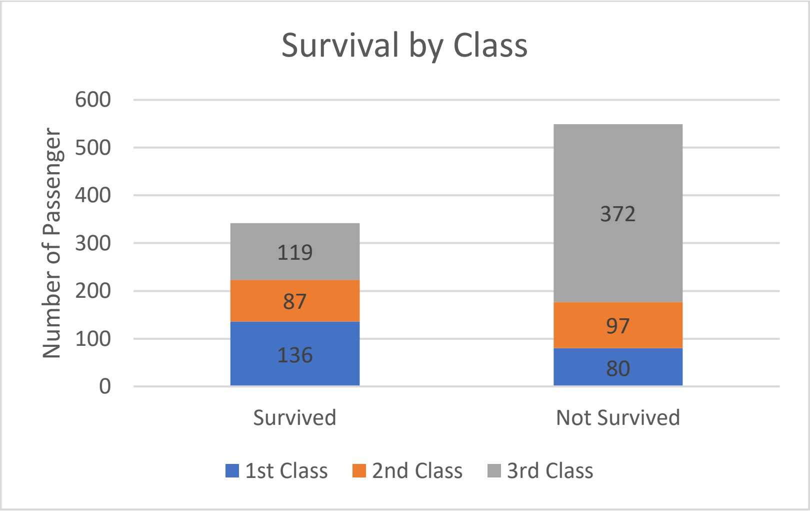
Survival by Class (Socio-economic)

- Does people in higher class are more likely to survive as they can access the rescue boat earlier?
- Find the number of passengers survive, not survived, and survival rate in every class.

```
SELECT s.Pclass AS Class , Survivors, Casualities, round(CAST(Survivors AS FLOAT) / (Survivors + Casualities) * 100,2) AS "Survival Rate"
FROM (
        SELECT Pclass, count(Pclass) AS Survivors
        FROM passengers
        WHERE Survived = 1
        GROUP BY Pclass
)s
JOIN(
        SELECT Pclass, count(Pclass) AS Casualities
        FROM passengers
        WHERE Survived = 0
        GROUP BY Pclass
)c
ON s.Pclass = c.Pclass
```



	Class	Survivors	Casualities	Survival Rate
1	1	136	80	62.96
2	2	87	97	47.28
3	3	119	372	24.24

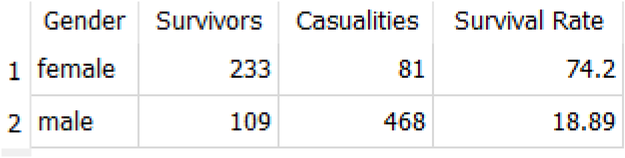


Conclusion: The graph shows that passenger in higher class is more likely to survive with more than 50% passenger in class 1 survive while most of passengers in class 3 died with the lowest survival rate.

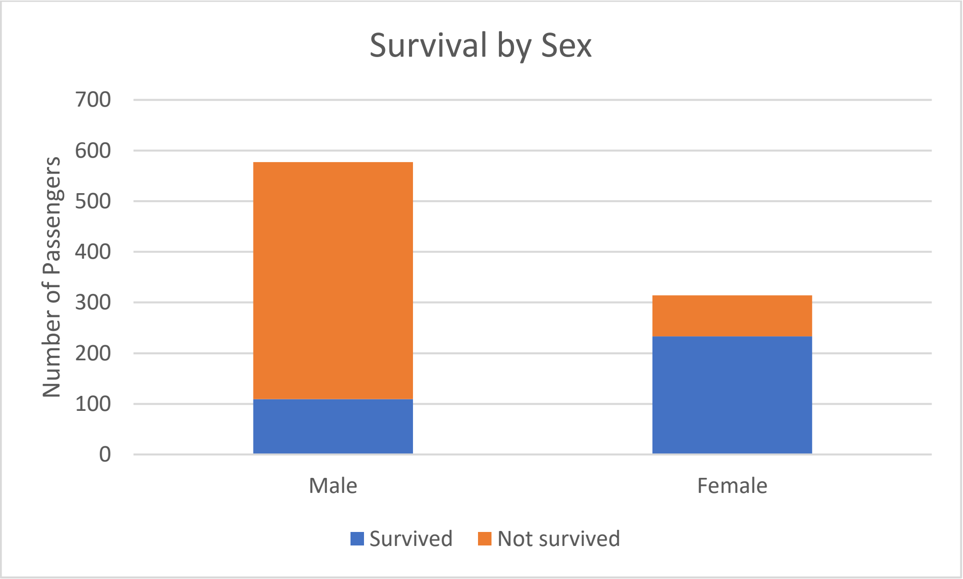
Survival by Sex

- Are females have higher survival rate in this incident?
- Find the number of passengers survive, not survived, and survival rate for male and female.

```
SELECT s.Sex AS Gender , Survivors, Casualities, round(CAST(Survivors AS FLOAT) / (Survivors +
Casualities) * 100,2) AS "Survival Rate"
FROM (
        SELECT Sex, count(Sex) AS Survivors
        FROM passengers
        WHERE Survived = 1
        GROUP BY Sex
)s
JOIN(
        SELECT Sex, count(Sex) AS Casualities
        FROM passengers
        WHERE Survived = 0
        GROUP BY Sex
)c
ON s.Sex = c.Sex
```



	Gender	Survivors	Casualities	Survival Rate
1	female	233	81	74.2
2	male	109	468	18.89

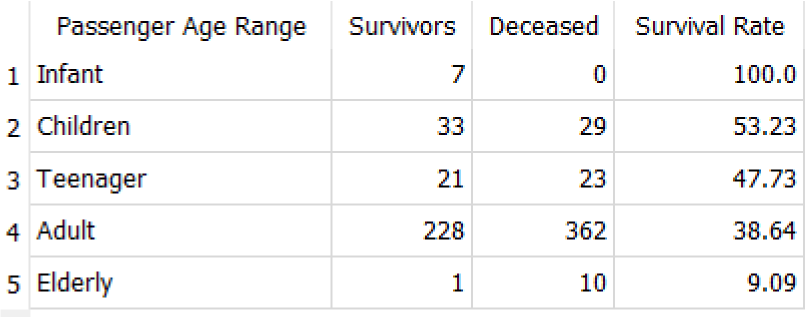


Conclusion: There are many male passengers than female on this ship. However, the data shows that more than half of the female survive while only a quarter of male survive. It can be concluded that female have a higher chance of survive probably they can board the rescue boat first.

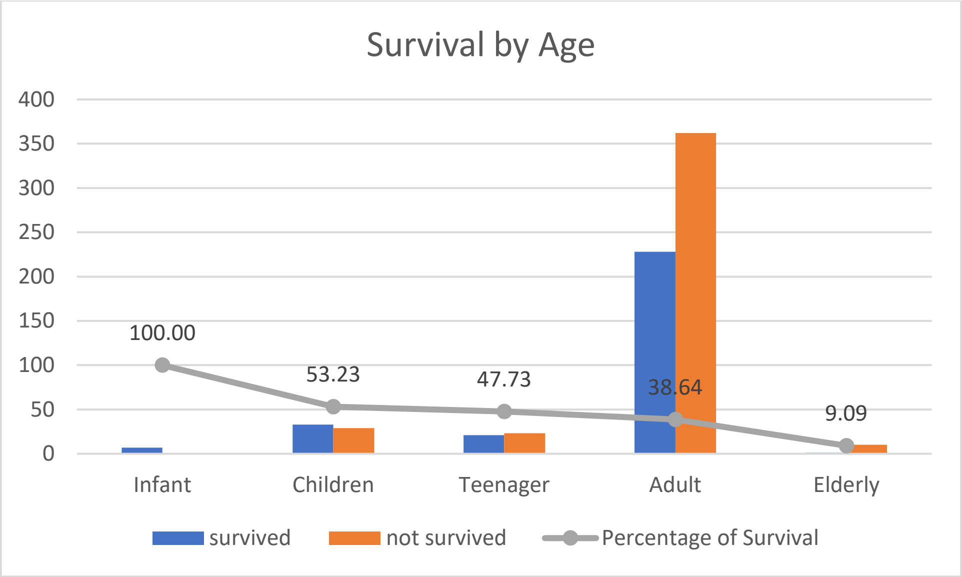
Survival by Age

- What is the range of age that more likely to survive this incident?
- Find the number of passengers survived, not survived and survival rate for each group of age.

```
SELECT s.age_range AS "Passenger Age Range", Survivors, ifnull(Dead,0) AS Deceased,
round(CAST(Survivors AS FLOAT) / (Survivors + ifnull(Dead,0)) * 100,2) AS "Survival Rate"
FROM (
        SELECT CASE WHEN CAST(Age AS FLOAT) < 1 THEN 'Infant'
                    WHEN CAST(Age AS FLOAT) BETWEEN 1 AND 12 THEN 'Children'
                    WHEN CAST(Age AS FLOAT) BETWEEN 13 AND 17 THEN 'Teenager'
                    WHEN CAST(Age AS FLOAT) BETWEEN 18 AND 64 THEN 'Adult'
                    WHEN CAST(Age AS FLOAT) > 64 THEN 'Elderly'
                END AS age_range,
                count(*) AS Survivors
        FROM passengers
        WHERE Survived = 1 AND Age IS NOT NULL
        GROUP BY age_range
        ORDER BY CASE WHEN age_range = "Infant" THEN 1
                      WHEN age_range = "Children" THEN 2
                      WHEN age_range = "Teenager" THEN 3
                       WHEN age_range = "Adult" THEN 4
                      WHEN age_range = "Elderly" THEN 5
                    END
)s
LEFT OUTER JOIN(
        SELECT CASE WHEN CAST(Age AS FLOAT) < 1 THEN 'Infant'
                    WHEN CAST(Age AS FLOAT) BETWEEN 1 AND 12 THEN 'Children'
                    WHEN CAST(Age AS FLOAT) BETWEEN 13 AND 17 THEN 'Teenager'
                    WHEN CAST(Age AS FLOAT) BETWEEN 18 AND 64 THEN 'Adult'
                    WHEN CAST(Age AS FLOAT) > 64 THEN 'Elderly'
                END AS age_range,
                count(*) AS Dead
        FROM passengers
        WHERE Survived = 0 AND Age IS NOT NULL
        GROUP BY age_range
)c
ON s.age_range = c.age_range
```



	Passenger Age Range	Survivors	Deceased	Survival Rate
1	Infant	7	0	100.0
2	Children	33	29	53.23
3	Teenager	21	23	47.73
4	Adult	228	362	38.64
5	Elderly	1	10	9.09



Conclusion: Most of the passengers were adult. However, younger group of age have a higher survival rate with 100% of the infant survive the incident followed by children, teenager, adult and elderly respectively.