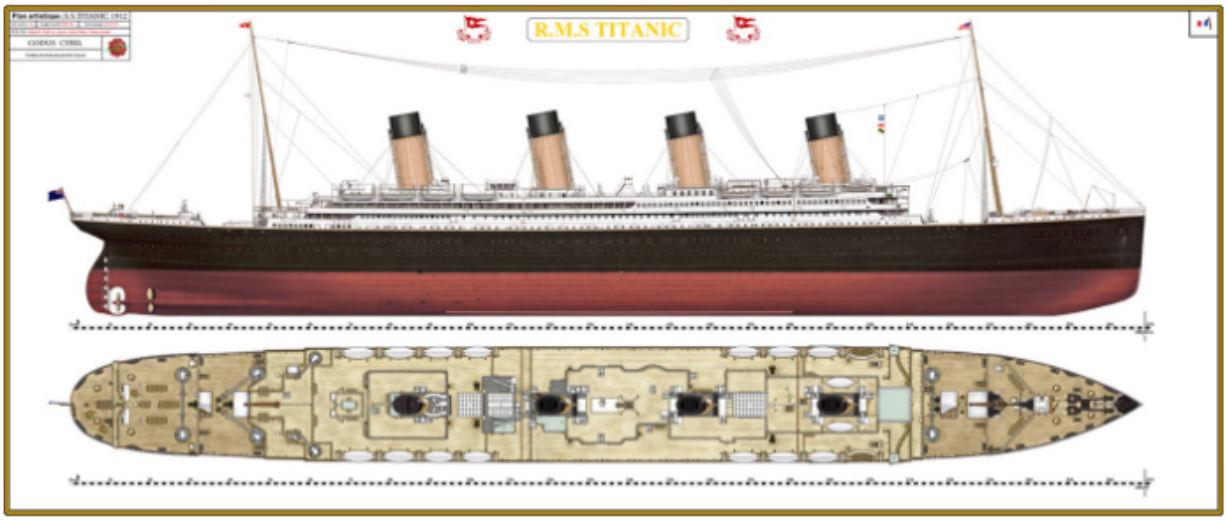
PROJECT 1: EXPLORING TITANIC DATABASE

PART 1: DATA EXPLORATION FRAMEWORK



The RMS Titanic, a luxury British steamship sank after striking an iceberg in April 1912 which lead to 1500 out of 2224 passengers and crew casualties (only 32% survival rate). The data of the passengers are originally retrieved from Encyclopedia Titanica and the titanic datasets are obtained from the Kaggle website. The purpose of this dataset is to predict the survival outcome by using the existing features of passengers onboard.

The data was collected by a collaborative effort from researchers and one of the sources is Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay. According to Encyclopedia Titanica, the information is collected through various mediums such as newspaper articles, books, broadcasts, and recollections of survivors and the families of victims.

The primary error of this data is a human error since one of the sources of data collected is through recollections of survivors and families of the victims who each experience different situations or perspectives of the tragedy, and these researchers sometimes need to rely on human memories of a terrible ordeal regardless. This dataset reflects the state of data available as of 2 August 1999. Some duplicate passengers have been dropped, many errors corrected, many missing ages filled in, and new variables created.

The data is not fully complete and there are some missing pieces of data, especially the independent features/variables. The table below lists the missing data of the features/variables from the dataset.

Column	Total missing data	Percentage of missing data
Cabin	687	77.1%
Age	177	19.9%
Embarked	2	0.2%
Fare	0	0.0%
Ticket	0	0.0%

From the table above, we could see the cabin has the highest percentage of missing data, thus, we will cross out the cabin from our further data exploration in the next chapter.

The dataset retrieved from the Kaggle consists of only one table, passengers. The passenger table is a database subject that contains all data of the dataset organized in terms of rows and columns, whereby rows are known as records and columns as fields. The column holds a set of data values with each having its particular data types (integer, text, date & time, etc.). For the table passengers, it comprises 12 fields or columns such as PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked, and the details are as followed:

Variable/Feature	Definition	Key
PassengerId	The id of the passenger	
Survived	Survival	0 = No 1 = Survived
Pclass	Ticket Class	1 = First class 2 = Second class 3 = Third class
Name	Name of passenger	
Sex	Gender of passenger	
Age	Age of passenger	
SibSp	Number of siblings/spouses aboard the ship	
Parch	Number of parents/children aboard the ship	
Ticket	Ticket number	

Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of embarked	C = Cherbourg Q = Queenstown S = Southampton

The table is slightly messy when the field mostly contains nulls which leads to inaccuracy in analysis. For example, the cabin column. To ensure the data is well-kept, data analysts need to do data cleaning first and foremost before analyzing the data, and there are few ways to clean the data depending on the data type.

PART 3: DATA EXPLORATION

Research Questions:

1) Are males more likely to survive the shipwreck?
2) What category of age group has a higher survival rate in this accident?
3) Are rich people have a higher survival rate?

1) Are males more likely to survive the shipwreck?

For the first research question, the first step would be to figure out how many females and males passengers in the ship, hence:

```
1   SELECT
2   SEX AS Gender,
3   COUNT (NAME) AS Total_Passengers
4   FROM PASSENGERS
5   GROUP BY SEX
```

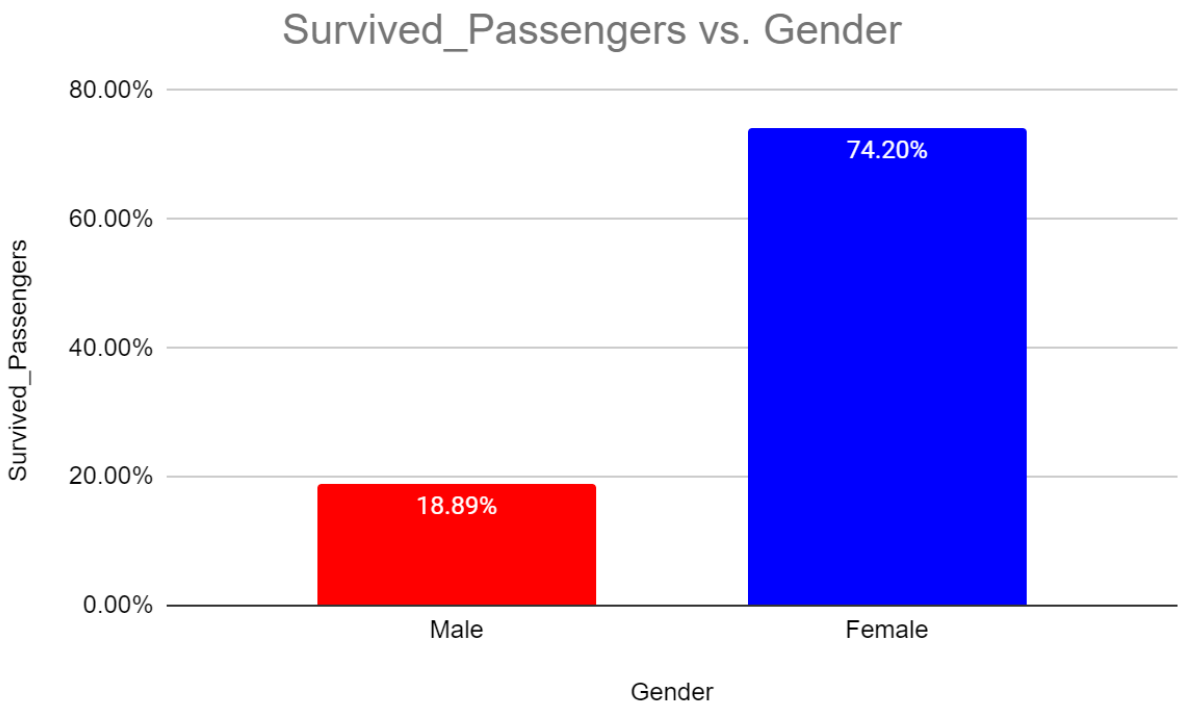
	Gender	Total_Passengers
1	female	314
2	male	577

The second step would be to count the survived passengers according to their gender:

```
1   SELECT
2   SEX AS Gender,
3   COUNT (SURVIVED) AS Survived_Passengers
4   FROM PASSENGERS
5   WHERE SURVIVED = 1
6   GROUP BY SEX
```

	Gender	Survived_Passengers
1	female	233
2	male	109

From the queries above, through the column chart, we will convert it into percentages and it illustrates that:



Thus, females had a very high survival rate at 74.2% and males had at 18.89% and back to the first question, is males more likely to survive the shipwreck? No, it did not, but the female is since they had a higher survival rate than males.

2) What category of age group has a higher survival rate in this accident?

Before we dive into the queries, the age categories are defined as below:

Categories	**Age (years old)**
Children	0 - 14
Youth	15 - 24
Adult	25 - 64
Senior	65 and over

(*Notes.* Adapted from https://www.statcan.gc.ca/en/concepts/definitions/age2. Copyright 2017 by Statistics Canada).

The first query would be to find the total passengers according to the age categories:

```
SELECT
COUNT (NAME) AS Total_Passengers,
CASE
WHEN AGE <= 14 THEN 'Children'
WHEN AGE >= 14.5 AND AGE <=24 THEN 'Youth'
WHEN AGE >= 24.5 AND AGE <=64 THEN 'Adult'
WHEN AGE >=65 THEN 'Senior'
ELSE 'Missing Age'
END AS Age_Categories
FROM PASSENGERS
GROUP BY Age_Categories
```

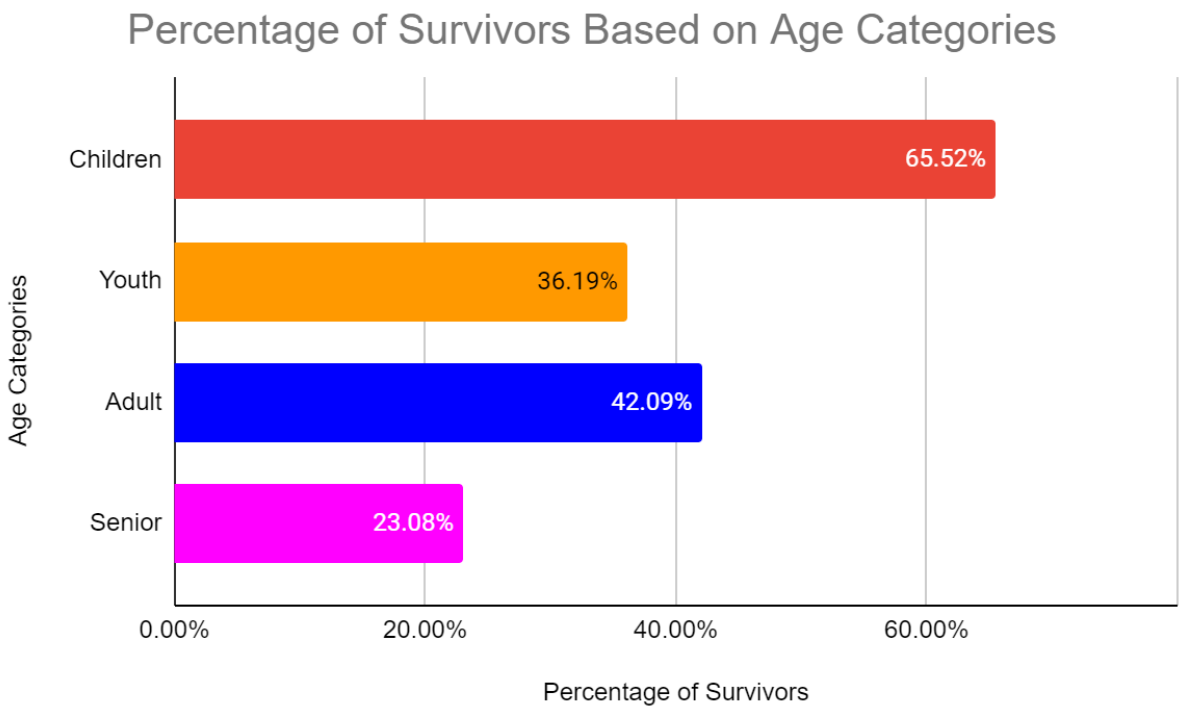
	Total_Passengers	Age_Categories
1	449	Adult
2	29	Children
3	177	Missing Age
4	26	Senior
5	210	Youth

The next query would be to count the survived passengers in their respective age categories:

```
SELECT
COUNT (SURVIVED) AS Total_Survivors,
CASE
WHEN AGE <= 14 THEN 'Children'
WHEN AGE >= 14.5 AND AGE <=24 THEN 'Youth'
WHEN AGE >= 24.5 AND AGE <=64 THEN 'Adult'
WHEN AGE >=65 THEN 'Senior'
ELSE 'Missing Age'
END AS Survivors_Age_Categories
FROM PASSENGERS
WHERE SURVIVED = 1
GROUP BY Survivors_Age_Categories
```

	Total_Survivors	Survivors_Age_Categories
1	189	Adult
2	19	Children
3	52	Missing Age
4	6	Senior
5	76	Youth

From the queries above, we will translate the information into percentages and from the bar chart, it shows that:



The children (age under 14) had a higher chance to survive the tragedy compared to the other age categories. The adult is the second age category that has a better survival rate than youth and is followed by the senior. The missing age data needs to be dropped from the list since the age is unknown.

3) Are rich people have a higher survival rate?

From the question above, we are going to use the ticket class (Pclass) variable to determine whether rich people have a higher survival rate since the Pclass is a proxy of socio-economic-status (SES) as shown below:

1st class = Upper SES
2nd class = Middle SES
3rd class = Lower SES

The first query would be to find the total passengers in their respective pClass:

```
SELECT
PCLASS AS Ticket_Class,
COUNT (NAME) AS Total_Passengers
FROM PASSENGERS
GROUP BY PCLASS
```

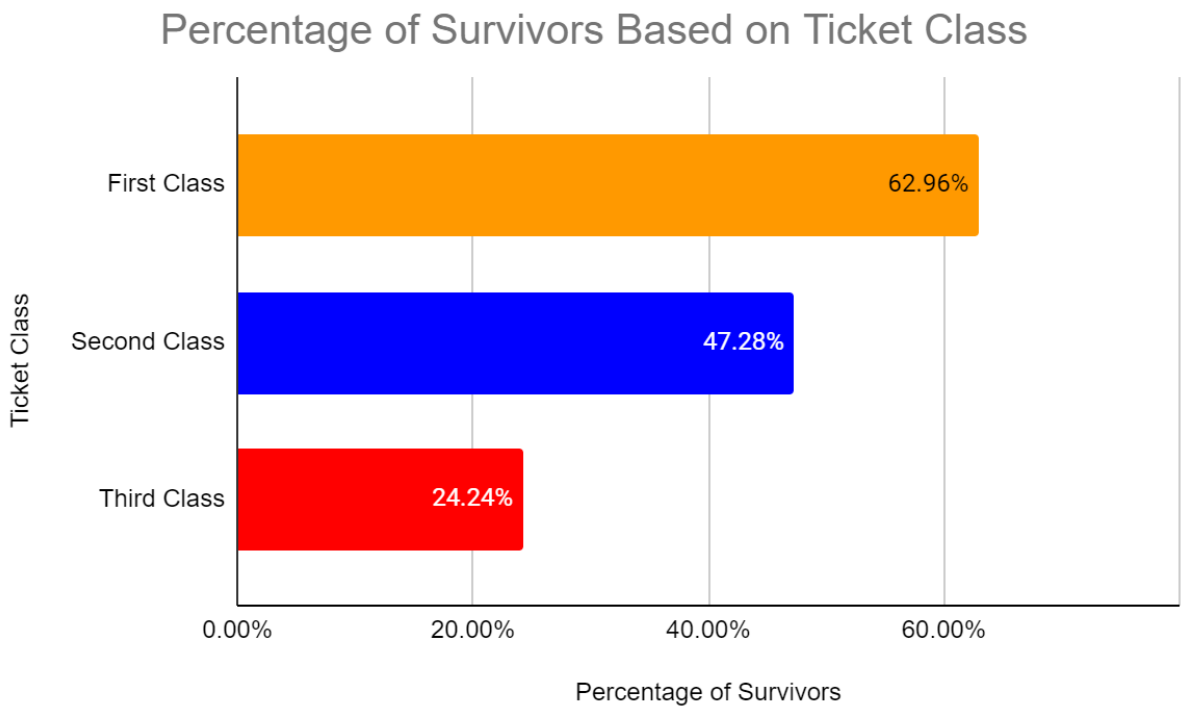
	Ticket_Class	Total_Passengers
1	1	216
2	2	184
3	3	491

The next step would be to find the number of survivors according to their ticket class:

```
SELECT
PCLASS AS Ticket_Class,
COUNT (SURVIVED) AS Survived_Passengers
FROM PASSENGERS
WHERE SURVIVED = 1
GROUP BY PCLASS
```

	Ticket_Class	Survived_Passengers
1	1	136
2	2	87
3	3	119

From the queries above, we could calculate the percentage of survivors based on the ticket class and it illustrates from the bar chart that:



In conclusion, it shows that the ticket class plays its role to determine whether the rich really had a higher survival chance since the first class has a higher percentage of survivors followed by the second class and the third class.

To sum up the research questions above, male is less likely to survive the shipwreck, children had the highest survival chance compared to the other age categories, and last but not least, the rich got to survive the tragedy than the other socio-economic classes.