Analysis of Titanic Data

The objective of this project is to leverage the open-source Titanic dataset to develop a comprehensive data exploration report. I will utilize SQL for exploratory data analysis, given that the data is structured in a database format.

I will implement a five-step framework that encompasses:

- Understanding the business context
- Understanding the technical context
- Analysing the tables and fields
- Creating research questions
- Addressing the research questions through data analysis

Understanding the Business Context

When you get a dataset, it is important to understand the context and background of the database before performing any analysis. Otherwise, you will not able to comprehend and interpret the data correctly.

Consider answering the following questions below:

- what are these data for?
- why do we need this database?
- where are these data collected?

Understanding the technical context

After understanding the context of the dataset, the next step is to examine the technical aspects of the dataset. Familiarity with these technical details will aid in interpreting the data and assessing its accuracy and reliability.

Consider answering the following questions below:

- how are these data collected?
- where are the sources of these data?
- what are the systems that use to modify these data?
- what are the error sources of this data?
- is the data complete? is there any missing pieces of data?

The data is sourced from Kaggle, a renowned repository for data science and free datasets. The titanic dataset on Kaggle is designed for users to apply machine learning techniques to develop predictive models determining which passengers will likely survive the Titanic shipwreck.

Analysing the Tables & Fields

This step aims to identify the available data tables within the database and examine the types of data contained within them.

Some questions to consider are:

- how many tables do we have
- what are the tables representing
- what are the fields in the tables? what is the meaning of the field
- should I clean the data or ignore the missing columns

Description of Data

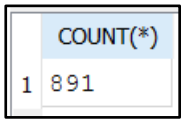
Variables descriptions:

- sex — male or female
- survived — 1= yes; 0 = no
- ticket — ticket number
- fare — passenger fare
- age — age in years
- pclass — ticket class, 1= 1st class; 2 = 2nd class; 3 = 3rd class
- parch — number of parents/children aboard the Titanic
- sibsp — number of siblings/spouses abroad the Titanic
- name — name of the passengers
- passengerId — passenger ID
- cabin — cabin number
- embarked — ports of embarkation, C=Cherbourg; Q= Queenstown, S=Southampton

Data inspecting & cleaning

By utilizing the SQL COUNT function, we can ascertain the number of rows present in the ***passengers*** table.

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

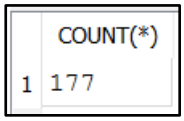


	COUNT(*)
1	891

Upon reviewing the data, it is observed that the table contains a total of 891 rows, with several instances of missing data identified within the table. Among the various columns, there are 177 missing entries in the Age column, 687 missing entries in the Cabin column, and 2 missing entries in the Embarked column.

Missing Entries – Age (177 missing values)

```
SELECT COUNT(*)
FROM passengers
WHERE Age IS NULL
```



	COUNT(*)
1	177

Missing Entries – Cabin (687 missing values)

```
SELECT COUNT(*)
FROM passengers
WHERE Cabin IS NULL
```

	COUNT(*)
1	687

Missing Entries – Embarked (2 missing values)

```
SELECT COUNT(*)
FROM passengers
WHERE Embarked IS NULL
```

	COUNT(*)
1	2

Creating a Research Question

This phase of the data analysis process involves generating insightful questions that can be answered using the available data.

Upon reviewing the Titanic database, several intriguing questions have arisen:

1. What is the overall survival rate for the titanic passenger?
2. Are females more likely to survive this incident?
3. Are children and elderlies have a higher survival rate?

4. Do rich people have a higher survival rate because they get onboard to rescue boats sooner?
5. Which cabin has the highest survival rate?
6. What is the survival rate for each embarkation?
7. What is the survival rate for a person without family onboard?

Having formulated the questions, I will utilize SQL to provide answers. Several SQL aggregation commands will be particularly useful for this project.

Question 1: What is the overall survival rate for the titanic passenger?

To determine the survival rate, we can employ the following formula:

$$\text{Survival Rate} = \left(\frac{\text{Number of Survivors}}{\text{Total Number of Passengers}} \right) \times 100$$

This formula expresses the survival rate as a percentage.

```
SELECT

    COUNT(CASE WHEN survived = 1 THEN 1 END) AS
Total_Passenger_Survived,

      (COUNT(CASE WHEN survived = 1 THEN 1 END) * 100.0 / COUNT(*))
AS Survived_Passenger_Rate,

    COUNT(CASE WHEN survived = 0 THEN 1 END) AS
Total_Passenger_Unsurvived,

    (COUNT(CASE WHEN survived = 0 THEN 1 END) * 100.0 / COUNT(*))
AS Unsurvived_Passenger_Rate

FROM

    passengers;
```

	Total_Passenger_Survived	Survived_Passenger_Rate	Total_Passenger_Unsurvived	Unsurvived_Passenger_Rate
1	342	38.3838383838384	549	61.6161616161616

The number of passengers who survived is 342, which corresponds to a survival rate of 38.38%. This statistic suggests that the majority of passengers did not survive, highlighting the severity of the situation. Understanding this rate can help in analysing the factors that contributed to survival, such as demographics or access to lifeboats.

Question 2: Are females more likely to survive this incident?

Next, we will investigate whether female passengers had a greater likelihood of survival during the Titanic disaster. Historical data indicates that women and children were typically prioritized during lifeboat evacuations, making it crucial to analyze the survival rates specifically for female passengers.

To conduct this analysis, we will calculate the total number of each gender.

```
SELECT Sex, COUNT(*) AS Total_Passenger_by_Gender
FROM passengers
GROUP BY Sex;
```

	Sex	Total_Passenger_by_Gender
1	female	314
2	male	577

Based on the result above, among the passengers onboard, there were 314 females and 577 males. By using the formula for survival rate, we can determine if the percentage of women who survived is significantly higher than that of their male counterparts.

Survival and Unsurvival Rate for Male

```
SELECT
    COUNT(CASE WHEN survived = 1 THEN 1 END) AS
Total_Male_Survived,
     (COUNT(CASE WHEN survived = 1 THEN 1 END) * 100.0 / COUNT(*))
AS Male_Survival_Rate,
    COUNT(CASE WHEN survived = 0 THEN 1 END) AS
Total_Male_Unsurvived,
    (COUNT(CASE WHEN survived = 0 THEN 1 END) * 100.0 / COUNT(*))
AS Male_Unsurvival_Rate
FROM passengers
WHERE
    sex = 'male';
```

	Total_Male_Survived	Male_Survival_Rate	Total_Male_Unsurvived	Male_Unsurvival_Rate
1	109	18.8908145580589	468	81.1091854419411

Survival and Unsurvival Rate for Female

```
SELECT
    COUNT(CASE WHEN survived = 1 THEN 1 END) AS
Total_Female_Survived,
     (COUNT(CASE WHEN survived = 1 THEN 1 END) * 100.0 / COUNT(*))
AS Female_Survival_Rate,
    COUNT(CASE WHEN survived = 0 THEN 1 END) AS
Total_Female_Unsurvived,
    (COUNT(CASE WHEN survived = 0 THEN 1 END) * 100.0 / COUNT(*))
AS Female_Unsurvival_Rate
FROM passengers
WHERE
    sex = 'female';
```

	Total_Female_Survived	Female_Survival_Rate	Total_Female_Unsurvived	Female_Unsurvival_Rate
1	233	74.203821656051	81	25.796178343949

Among the total passengers on board, 233 females with a survival rate of 74.2% survived the Titanic disaster, while 109 males with a survival rate of 18.89% managed to escape. Based on this result, it shows that female passengers are more likely to survive in this tragedy.

These statistics reveal a stark disparity in survival rates between genders. The higher survival rate among females suggests that the evacuation protocols, which prioritized women and children, were somewhat effective in this instance. Notably, a greater percentage of women survived compared to men, indicating that societal norms at the time may have influenced who was allowed access to lifeboats.

Question 3: Are children and elderlies have a higher survival rate in this accident?

Upon exploring the relationship between survival and age, there is a notable correlation. We categorize individuals as follows: children are defined as those aged between 0 and 17, the elderly are those over 59 (not including 59), adults are classified as those aged between 18 and 59, and any missing age data is labelled as "NULL."

```
SELECT
    CASE
        WHEN age IS NULL THEN 'NULL'
        WHEN age BETWEEN 0 AND 17 THEN 'Children'
        WHEN age > 59 THEN 'Elderlies'
        WHEN age BETWEEN 18 AND 59 THEN 'Adults'
        ELSE 'Unknown'
    END AS Age_category,
    COUNT(*) AS Total_Count
FROM passengers GROUP BY Age_category;
```

	Age_category	Total_Count
1	Adults	605
2	Children	65
3	Elderlies	44
4	NULL	177

The Titanic dataset includes 605 adults, 65 children, 44 elderly individuals, and 177 entries with missing values in the Age column.

```
SELECT
    CASE
        WHEN age IS NULL THEN 'NULL'
        WHEN age BETWEEN 0 AND 17 THEN 'Children'
        WHEN age > 59 THEN 'Elderlies'
        WHEN age BETWEEN 18 AND 59 THEN 'Adults'
        ELSE 'Unknown'
    END AS Age_category,
    SUM(survived) AS Total_Survived,
    (SUM(survived) * 100.0 / COUNT(*)) AS Survival_Rate,
    (COUNT(*) - SUM(survived)) AS Total_Unsurvived,
    ((COUNT(*) - SUM(survived)) * 100.0 / COUNT(*)) AS
Unsurvival_Rate
FROM passengers
GROUP BY Age_category;
```

	Age_category	Total_Survived	Survival_Rate	Total_Unsurvived	Unsurvival_Rate
1	Adults	241	39.8347107438017	364	60.1652892561983
2	Children	35	53.8461538461538	30	46.1538461538462
3	Elderlies	14	31.8181818181818	30	68.1818181818182
4	NULL	52	29.3785310734463	125	70.6214689265537

Based on the result, it shows that as age increased, the survival rate decreased. Children exhibited the highest survival rates, followed by the adults. This trend can likely be attributed

to the prioritization of women and children during evacuations, the physical fitness of younger individuals, and the increased vulnerability of older adults to cold and exposure.

Question 4: Do rich people have a higher survival rate because they get on board to rescue boats sooner?

In the class column, the number indicates a different class: -

- 1 = Upper class
- 2 = Middle class
- 3 = Lower class

```
SELECT
    Pclass AS Passenger_Class,
    COUNT(*) AS Total_Passenger
FROM
    passengers
GROUP BY
    Pclass;
```

	Passenger_Class	Total_Passenger
1	1	216
2	2	184
3	3	491

The data reveals that the majority of passengers belonged to Third Class, with a total of 491 passengers, followed by First Class with 216 passengers, and finally Second Class with 184 passengers. Now, let’s calculate each survival rate based on their *Pclass*.

```
SELECT
    Pclass AS Passenger_Class,
    COUNT(CASE WHEN Survived = 1 THEN 1 END) AS
Total_Survived,
     AVG(Survived) * 100 AS Survival_Rate,
    COUNT(CASE WHEN Survived = 0 THEN 1 END) AS
Total_Unsurvived,
    (COUNT(CASE WHEN Survived = 0 THEN 1 END) * 100.0 /
COUNT(*)) AS Unsurvival_Rate
FROM
    passengers
GROUP BY
    Pclass;
```

	Passenger_Class	Total_Survived	Survival_Rate	Total_Unsurvived	Unsurvival_Rate
1	1	136	62.962962962963	80	37.037037037037
2	2	87	47.2826086956522	97	52.7173913043478
3	3	119	24.2362525458248	372	75.7637474541751

Based on the result above, the survival rate for each class is as follows:

- 1st class — 62.96%
- 2nd class — 47.28%
- 3rd class — 24.24%

These statistics indicate that 1st Class passengers had the highest survival rate, with approximately 62.96% of them surviving the disaster. This can be attributed to several factors, including their proximity to lifeboats and the prioritization of women and children, which may have disproportionately benefited those in higher classes.

In contrast, 2nd Class passengers had a survival rate of around 47.28%, while only 24.24% of 3rd Class passengers survived. The significantly lower survival rate among 3rd Class passengers suggests that they faced considerable barriers during the evacuation process. Many were located farther from the lifeboats and may have encountered additional obstacles that hindered their escape.

Question 5: Which cabin has highest survival rate?

Let us examine if there are any correlation between cabin deck and survival rate. We will begin by classifying the cabins as "A" if the first character of the cabin identifier is 'A' and so on.

```
WITH CabinDecks AS (
    SELECT
        SUBSTR(Cabin, 1, 1) AS CabinDeck,
        Pclass,
        Survived
    FROM
        passengers
    WHERE
        Cabin IS NOT NULL
)

SELECT
    CabinDeck,
    COUNT(CASE WHEN Pclass = 1 THEN 1 END) AS Total_First_Class,
    COUNT(CASE WHEN Pclass = 2 THEN 1 END) AS Total_Second_Class,
    COUNT(CASE WHEN Pclass = 3 THEN 1 END) AS Total_Third_Class,
    COUNT(CASE WHEN Survived = 1 THEN 1 END) AS Total_Survived,
     AVG(Survived) * 100 AS Survival_Rate,
    COUNT(CASE WHEN Survived = 0 THEN 1 END) AS Total_Unsurvived,
    (COUNT(CASE WHEN Survived = 0 THEN 1 END) * 100.0 / COUNT(*))
AS Unsurvival_Rate
FROM
    CabinDecks
GROUP BY
    CabinDeck
ORDER BY
    CabinDeck;
```

	CabinDeck	Total_First_Class	Total_Second_Class	Total_Third_Class	Total_Survived	Survival_Rate	Total_Unsurvived	Unsurvival_Rate
1	A	15	0	0	7	46.6666666666667	8	53.3333333333333
2	B	47	0	0	35	74.468085106383	12	25.531914893617
3	C	59	0	0	35	59.3220338983051	24	40.6779661016949
4	D	29	4	0	25	75.7575757575758	8	24.2424242424242
5	E	25	4	3	24	75.0	8	25.0
6	F	0	8	5	8	61.5384615384615	5	38.4615384615385
7	G	0	0	4	2	50.0	2	50.0
8	T	1	0	0	0	0.0	1	100.0

Based on the result, it shows that the survival rate if passengers board on cabin B, D and E is almost the same which is around 75%. Several factors could contribute to this uniform survival rate, such as the location of these cabins on the ship, which may have affected accessibility to lifeboats. Additionally, social factors, including the presence of women and children in these cabins, might have influenced priority during the evacuation.

However, more investigation is needed as there are 687 missing values in the cabin dataset that could impact the analysis. A thorough examination of these missing data points could provide a more nuanced understanding of how location of the cabin could influence survival during this tragic event.

Question 6: What is the survival rate for each embarkation?

There are 3 ports of embarkation: Cherbourg(C), Queenstown(Q), and Southampton(S). There are 2 missing values in this dataset.

```
SELECT
    CASE
        WHEN Embarked IS NULL THEN 'Unknown'
        ELSE Embarked
    END AS Embarkation_Port,
    COUNT(*) AS Passenger_Count
FROM
    passengers
GROUP BY
    CASE
        WHEN Embarked IS NULL THEN 'Unknown'
        ELSE Embarked
    END
ORDER BY
    Embarkation Port;
```

	Embarkation_Port	Passenger_Count
1	C	168
2	Q	77
3	S	644
4	Unknown	2

Let’s compare the survival rate of each passenger by each of the embarkation port. In this code, we exclude the NULL values.

```
SELECT
    Embarked,
    SUM(CASE WHEN Pclass = 1 THEN 1 ELSE 0 END) AS
First_Class_Count,
    SUM(CASE WHEN Pclass = 2 THEN 1 ELSE 0 END) AS
Second_Class_Count,
    SUM(CASE WHEN Pclass = 3 THEN 1 ELSE 0 END) AS
Third_Class_Count,
    SUM(Survived) AS Total_Survived,
    AVG(Survived) * 100 AS Survival_Rate,
    (COUNT(*) - SUM(Survived)) AS Total_Unsurvived,
    (100.0 * (COUNT(*) - SUM(Survived)) / COUNT(*)) AS
Unsurvival_Rate
FROM
    passengers
WHERE
    Embarked IN ('C', 'Q', 'S')
GROUP BY
    Embarked
ORDER BY
    Embarked;
```

	Embarked	First_Class_Count	Second_Class_Count	Third_Class_Count	Total_Survived	Survival_Rate	Total_Unsurvived	Unsurvival_Rate
1	C	85	17	66	93	55.3571428571429	75	44.6428571428571
2	Q	2	3	72	30	38.961038961039	47	61.038961038961
3	S	127	164	353	217	33.695652173913	427	66.304347826087

The data reveals that passengers from Cherbourg had the highest survival rate at 55.36%, indicating that a significant proportion of those who boarded the Titanic at this port were able to survive the disaster. Following Cherbourg, Queenstown passengers exhibited a survival rate of 38.96%.

While this rate is notably lower than that of Cherbourg, it still reflects a relatively substantial number of survivors compared to Southampton.

In contrast, Southampton had the lowest survival rate at 33.7%. This may suggest that the composition of passengers from this port included a larger number of individuals in third class, who traditionally had lower survival rates due to their distance from lifeboats and the chaos during the evacuation.

Question 7: What is the survival rate for a person without family onboard?

```
SELECT
    (Parch + SibSp) AS FamilySize,
    COUNT(*) AS Total_Passengers,
    SUM(CASE WHEN Survived = 1 THEN 1 ELSE 0 END) AS
Total_Survived,
    SUM(CASE WHEN Survived = 0 THEN 1 ELSE 0 END) AS
Total_Unsurvived,
    AVG(Survived) * 100 AS Survival_Rate,
    (SUM(CASE WHEN Survived = 0 THEN 1 ELSE 0 END) * 100.0 /
COUNT(*)) AS Unsurvival_Rate
FROM
    passengers
GROUP BY
    FamilySize
ORDER BY
    FamilySize;
```

	FamilySize	Total_Passengers	Total_Survived	Total_Unsurvived	Survival_Rate	Unsurvival_Rate
1	0	537	163	374	30.3538175046555	69.6461824953445
2	1	161	89	72	55.2795031055901	44.7204968944099
3	2	102	59	43	57.843137254902	42.156862745098
4	3	29	21	8	72.4137931034483	27.5862068965517
5	4	15	3	12	20.0	80.0
6	5	22	3	19	13.6363636363636	86.3636363636364
7	6	12	4	8	33.3333333333333	66.6666666666667
8	7	6	0	6	0.0	100.0
9	10	7	0	7	0.0	100.0

There are lots of passengers who travelled without any family onboard. The passenger with families of 3, or 4 people had a higher chance of survival than single people or people with larger number of family members.

Conclusion

- 38.38% of passengers survived in titanic

- Children exhibited the highest survival rates, followed by the adults
- Female passengers are more likely to survive than male passengers
- The upper class is more likely to survive
- The survival rate if passengers board on cabin B, D and E is almost the same which is around 75%
- Port Cherbourg passengers have the highest surviving rate — 55.36%

Limitations

- The analysis has not taken into account the duplication of data. It has been assumed that each entry is unique.
- Rows with missing values in columns (Age, Cabin, Embarked) have been removed for the analysis of corresponding columns. This reduces the fidelity of the analysis based on available data.
- The dataset is only a subset of total original passengers onboard and the factors such as survival rate could vary.