Fundamentals of SQL

PROJECT 1

TITANIC DATABASE EXPLORATION

by: Ainnur Maryam

STEP 1: UNDERSTANDING THE BUSINESS CONTEXT

What are these data for?

The Titanic dataset is a collection of historical **records concerning the passengers aboard the RMS Titanic** during its ill-fated maiden **voyage in April 1912**. The data includes various attributes of the passengers, such as their demographic information (age, sex, and ticket class), survival status, and family connections.

This dataset is primarily used for **data analysis, statistical modelling, and machine learning tasks** to explore factors influencing survival rates during the disaster.

Why do we need this database?

This database serves multiple purposes:

Educational Use

It is widely used in data science education and tutorials to teach fundamental concepts such as data cleaning, exploratory data analysis, and predictive modelling.

Historical Insight

It provides insights into social dynamics and survival patterns during a significant historical event, allowing researchers to analyse the impact of socioeconomic factors on survival.

Statistical Analysis

It helps in understanding relationships between various factors (e.g., class, gender, age) and outcomes, facilitating discussions on data-driven decision-making.

Where are these data collected?

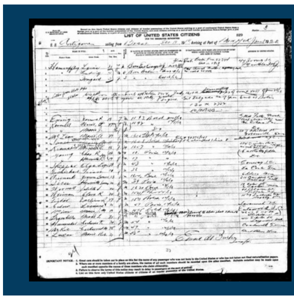
The data were **collected from passenger records maintained by the White Star Line**, the company that operated the Titanic. This information was **derived from historical archives, including ship manifests and other official documents** that recorded the details of passengers who boarded the Titanic. Researchers and data scientists have compiled and curated this data for analysis and educational purposes, making it publicly accessible through platforms like GitHub and Kaggle.

STEP 2:
UNDERSTANDING THE TECHNICAL CONTEXT

How are these data collected?

The data for the Titanic dataset were collected from various **official records maintained by the White Star Line,** including **passenger manifests** and **ticketing information. Historical documents and logs** from the Titanic's maiden voyage were the primary sources of this data.

Where are the sources of these data?



LIST OF UNITED STATES CITIZENS

Passenger Manifests: Official documents that listed all passengers aboard the Titanic, including their details such as name, age, sex, and class

Historical Archives: Records stored in libraries, museums, and online databases that preserve maritime history and passenger information

Is the data coming from surveys, or some computer system? Is it manually input by some data entry personnel or collected by some electronic system?

The data were **originally recorded manually by crew members and ticketing personnel at the time of boarding.** There were no electronic systems in place during the Titanic's voyage in 1912. The information was **later digitised and compiled by researchers and historians**, who transcribed the data into modern formats for analysis.

What are the systems that touch or use/modify these data?

The data can be accessed and modified through various software systems, including:

- **Data Analysis Tools:** Software like Python (with libraries such as Pandas and NumPy) and R for statistical analysis.
- **Database Management Systems:** Tools such as SQL databases or cloud-based platforms that facilitate data storage and querying.
- **Visualisation Tools:** Applications like Tableau for creating visual representations of the data to help communicate findings.

STEP 2:
UNDERSTANDING THE TECHNICAL CONTEXT

What are some of the error sources of this data?

Potential sources of error in the dataset include:

Human Error

Mistakes in data entry when transcribing historical records, such as misspellings or incorrect entries.

Missing Records

Incomplete records due to loss or damage to documents over time, leading to gaps in data.

Ambiguities

Inconsistent naming conventions or abbreviations that could lead to confusion in interpreting certain entries (e.g., variations in how names are recorded).

Is the data complete? Would there be missing pieces of data?

The Titanic dataset is not completely comprehensive. While it includes a significant number of passengers, there are known missing pieces of data, such as:

Missing Age Information

Many entries lack age data, particularly for adults, as it was not always recorded.

Cabin Numbers

Some passengers do not have cabin numbers recorded, which may affect analyses related to their location on the ship.

Embarked Information

In some cases, the port of embarkation may be missing for certain passengers.

STEP 3:
UNDERSTANDING THE TABLES AND FIELDS

How many tables do we have?

There is **one** main table, referred to as the **"passengers"** table. This table contains all the relevant information about the passengers.

What are the tables? And what are these tables representing?

Passengers Table: This single table **represents the passengers aboard the Titanic.** It contains various fields that detail the characteristics of each passenger, their ticket information, and their survival status.

What are the relationships between the tables?

Since the Titanic dataset primarily consists of one table, there are no relationships between multiple tables. However, if we were to merge this dataset with other datasets (e.g., a crew dataset or a lifeboat dataset), we would establish relationships based on common fields such as passenger ID.

What are the fields in the tables? What is the meaning of each field?

Field	Meaning
PassengerId	A unique identifier for each passenger
Survived	Survival status (0 = No, 1 = Yes)
Pclass	Ticket class (1 = 1st class, 2 = 2nd class, 3 = 3rd class)
Name	Full name of the passenger
Sex	Gender of the passenger (male of female)
Age	Age of the passenger
SibSp	Number of siblings or spouses aboard the Titanic
Parch	Number of parents or children aboard the Titanic
Ticket	Ticket number
Fare	Fare paid for the ticket
Cabin	Cabin number
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

STEP 3:
UNDERSTANDING THE TABLES AND FIELDS

Is the data messy? And how?

Yes, the data can be considered "messy" due to several factors:

Missing Values: Some fields, such as Age and Cabin, may have many NULL values, which can complicate analysis.
Inconsistent Formats: The Name field might have variations in naming conventions, including titles (e.g., Mr., Mrs., Miss).
Outliers: Extreme values in the Age or Fare fields may need to be reviewed.
Redundant Information: The Ticket field may contain duplicate ticket numbers, which can lead to ambiguity in analyses.

Should I clean the data first? Or ignore those messy columns?

Whether to clean the data first or ignore messy columns **depends on the specific analysis I will be performing.** If certain columns are **relevant to my questions, I should clean them** to ensure accuracy and reliability in my results. However, if some columns are **not related to my analysis, I may choose to ignore them.**

STEP 4:
FREE EXPLORATION

$$\boxed{Survival\ Rate = \frac{Number\ of\ Survivors}{Total\ Passengers}\ x\ 100\%}$$

What was the overall survival rate of passengers on the Titanic?

```
SELECT
    COUNT(*) AS total_passengers
    SUM(Survived) AS survivors,
    ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS survival_rate
FROM passengers
```

Result:

total_passengers	survivors	survival_rate
891	342	38.38

The overall survival rate was approximately **38.38%.** This indicates that out of all the passengers aboard the Titanic, **slightly more than one-third were able to survive** the tragic sinking of the ship.

STEP 4:
FREE EXPLORATION

Does social-economic standing (according to ticket class and ticket fare) contribute to better survival?

1. Survival Rate by Ticket Class:

```
SELECT Pclass,
    COUNT(*) AS Total_Passengers,
     SUM(Survived) AS Survivors,
     ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS Survival_Rate
FROM passengers
GROUP BY Pclass
```

Result:

Pclass	Total_Passengers	Survivors	Survival_Rate
1	216	136	62.96
2	184	87	47.28
3	491	119	24.24

2. Average Fare by Ticket Class

*This is calculated to reinforce the socio-economic distinctions between the classes. Generally, higher ticket prices correlate with wealthier individuals

```
SELECT Pclass,
    AVG(Fare) AS Average_Fare
FROM passengers
GROUP BY Pclass
```

Result:

Pclass	Average_Fare
1	84.1546875
2	20.66218315217391
3	13.675550101833

Passengers in **1st class** had a notably **higher survival rate** compared to those in 2nd and 3rd class, suggesting that **socio-economic standing played a crucial role in survival**.
Additionally, the **average ticket fare for 1st class was significantly higher**, reinforcing the **correlation between economic status and access to lifeboats**. This data illustrates the impact of social hierarchy during the disaster, highlighting how factors such as class and wealth influenced survival outcomes in a critical situation.

STEP 4:
FREE EXPLORATION

Did women and children really get priority for lifeboats, hence higher survival?

1. Survival Rate by Gender:

```
SELECT Sex,
    COUNT(*) AS Total_Passengers,
     SUM(Survived) AS Survivors,
     ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS Survival_Rate
FROM passengers
GROUP BY Sex
```

Result:

Sex	Total_Passengers	Survivors	Survival_Rate
female	314	233	74.0
male	577	109	18.0

2. Age Grouping - Survival Rates for Children (Under 18) Compared to Adults (18 and Above)

```
SELECT CASE,
        WHEN Age < 18 THEN ‘Child’
        ELSE ‘Adult’
    END AS Age_Group,
     COUNT(*) AS Total_Passengers,
     SUM(Survived) AS Survivors,
     ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS Survival_Rate
FROM passengers
GROUP BY Age_Group
```

Result:

Age_Group	Total_Passengers	Survivors	Survival_Rate
Adult	826	307	37.17
Child	65	35	53.85

(continue on next page)

STEP 4:
FREE EXPLORATION

(cont.)

Did women and children really get priority for lifeboats, hence higher survival?

3. Survival Rate for Women and Children Combined:

```
SELECT CASE,
        WHEN Sex = ‘female’ OR Age < 18 THEN ‘Women & Children’
        ELSE ‘Men & Adults’
    END AS Category,
    COUNT(*) AS Total_Passengers,
    SUM(Survived) AS Survivors,
    ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS Survival_Rate
FROM passengers
GROUP BY Category
```

Result:

Category	Total_Passengers	Survivors	Survival_Rate
Men & Adults	541	97	17.93
Women & Children	350	245	70.0

The analysis reveals that **women and children** indeed experienced **higher survival rates** compared to their male counterparts on the Titanic.

The data shows that **women**, likely prioritized during lifeboat evacuations, had **significantly better survival rates**.

Additionally, **children** also demonstrated a **favourable survival rate**, supporting the notion that they were given preference for lifeboats.

This aligns with historical accounts of the tragedy, which indicate that the **policy of "women and children first" was a key factor in determining survival** during the disaster.

STEP 4:
FREE EXPLORATION

Do people who aboard alone have higher or lower survival than people who aboard with relatives (siblings, spouse, parents, children)?

For this analysis, passengers are categorised into two groups:

1. Passengers who boarded alone
2. Passengers who boarded with at least one relative

```
SELECT CASE,
        WHEN SibSp = 0 AND Parch = 0 THEN 'Aboard Alone'
        ELSE 'Aboard With Relatives'
    END AS Classification,
    COUNT(*) AS Total_Passengers,
    SUM(Survived) AS Survivors,
    ROUND((SUM(Survived) * 100 / COUNT(*)), 2) AS Survival_Rate
FROM passengers
GROUP BY Classification
```

Result:

Classification	Total_Passengers	Survivors	Survival_Rate
Aboard Alone	537	163	30.35
Aboard With Relatives	354	179	50.56

- **Survival Rate for Passengers Aboard Alone: 30.35%**

This indicates that only **about one in three** passengers who boarded alone **survived** the disaster. The **lower survival rate** suggests that individuals without companions may have faced **greater challenges during the evacuation process**, potentially **lacking the support and prioritisation** that comes with traveling in groups.

.

- **Survival Rate for Passengers Aboard with Relatives: 50.56%**

In contrast, the survival rate for those who boarded with relatives is notably **higher**, with **over half** of these passengers **surviving**. This suggests that s**ocial connections played a crucial role during the evacuation**, possibly allowing for **better access to lifeboats** and **greater chances of receiving assistance**.