Project 1: Data Exploration – Titanic

STEP 1: Understanding the Business Context

1) What are these data for?

This data is about passengers aboard the Titanic, a ship which sank on its maiden voyage in 1912 after colliding with an iceberg. This data is commonly used to analyze the survival rates of the passengers and the factors influencing the survival rate.

2) Why do we need this database?

We need this database to understand the factors and demographics that influence the survival rate of the passengers. We can analyze the impact of the socio-economy class, age and gender on the chances of survival.

STEP 2: Understanding the Technical Context

1) Where are the sources of these data?

The data is collected from historical sources such as the passenger manifest, survivor testimonies and other archival documents. The dataset has been compiled and made available by Kaggle.

2) 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?

Since the data is collected from historical records, the data must be manually input by data entry personnel.

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

The data is made available by Kaggle and typically used for analysis using software like SQL, Python, R, Power BI and Tableau.

4) What are some of the error sources of this data?

Some error sources would be the mistake during data entry and inaccurate historical records.

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

The data may not be complete due to unknown details of the passengers, incomplete records and loss of information over time.

STEP 3: Understanding the Tables and Fields

1) How many tables do we have? What are the tables? and what are these tables representing?

There is one table in the dataset which is the passengers table. The table represents the details of the passengers aboard the Titanic ship.

2) What are the fields in the tables? What is the meaning of each of the fields?

- PassengerId: A unique identifier for each passenger
- Survived: Survival status (0 = No, 1 = Yes)
- Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name: Name of the passenger
- Sex: Gender of the passenger
- Age: Age of the passenger
- SibSp: Number of siblings/spouses aboard the Titanic
- Parch: Number of parents/children aboard the Titanic
- Ticket: Ticket number
- Fare: Passenger fare
- Cabin: Cabin number
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Step 4: Free Exploration

Part 1: Understanding the data

The table contains 12 columns:

PassengerId (integer) – passenger id
Survived (integer) – survival (0 = No, 1 = Yes)
Pclass (integer) – ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name (text) - sex
Sex (text) – age in years
SibSp (integer) – no of siblings/spouse aboard
Parch (text) – no of parents/children aboard
Ticket (text) – ticket number
Fare (integer) – passenger fare
Cabin (text) – cabin number
Embarked (text) – port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

To see the overview of the tables:

```
select *
from passengers
limit 10
```

Result:

1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

Part 2 : Cleaning up the data

Check for duplicate:

select PassengerId, count(*) as passenger_count
from passengers
group by PassengerId
having passenger_count > 1

Result: 0 rows returned

Change data type:

Age (text) > Age (integer)
Using modify table in database structure

Check for null:

select count(*) as Age_null
from passengers
where Age is null

Result = 177

Part 3: Data Analysis

Number of passengers:

```
select count(*)
from passengers
```

Result = 891

```
select count(*)
from passengers
where Age is not NULL
```

Result = 714

Survival rate:

```
select avg(Survived) as survival_rate
from passengers
```

Result: 0.383838383838384

Survival rate by gender:

```
select sex, avg(Survived) as survival_rate
from passengers
group by Sex
```

Result:

female	0.74203821656051
male	0.188908145580589

Class analysis:

```
select Pclass, count(*)
from passengers
group by Pclass
```

Result:

1	216
2	184
3	491

Survival rate by class:

```
select Pclass, avg(Survived)
from passengers
group by Pclass
```

Result:

1	0.62962962962963
2	0.472826086956522
3	0.242362525458248

Age analysis:

We will categorize the age range as below:
Child = 0-17 yrs
Adult = 18-59 yrs
Elderly = 60+

Add age_range column and update the data:

```
alter table passengers
add column age_range text;

update passengers
set age_range = CASE
        WHEN Age is null or Age = " " THEN "Unknown"
        WHEN Age >= 60 THEN "Elderly"
        WHEN Age BETWEEN 18 and 59 THEN "Adult"
        WHEN Age BETWEEN 0 and 17 THEN "Child"
END;
```

Count of age range:

```
select age_range, count(*)
from passengers
group by age_range
```

Result:

Adult	575

Child	113
Elderly	26
Unknown	177

Survival rate based on age:

select age_range, avg(Survived)
from passengers
group by age_range

Result:

Adult	0.386086956521739
Child	0.539823008849557
Elderly	0.269230769230769
Unknown	0.293785310734463

Embarkation point analysis:

select Embarked, count(*)
from passengers
where Embarked is not null
group by Embarked

Result:

C	168
Q	77
S	644

Family analysis:

alter table passengers
add column family_size integer;

update passengers
set family_size = SibSp + Parch + 1;

Count of family size:

select family_size, count(*)
from passengers
group by family_size

Result:

1	537
2	161
3	102
4	29
5	15
6	22
7	12
8	6
11	7

Part 4: Conclusion

1) Are children and elderly have a higher survival rate in this accident?

Based on the survival rate based on age, children have a survival rate of 53.98% whereas elderly have a survival rate of 26.92. This shows that children have a higher survival rate, suggesting that rescue for children might have been more prioritized compared to elderly. The analysis also shows that elderly have the lowest survival rate compared to both children and adults.

However, it is important to note that many passengers have missing age data and it might affect the analysis. The “Unknown” category has a survival rate of 29.38% which is relatively close to the overall survival rate of 38.38%.

2) Are females more likely to survive in this incident?

Females have a survival rate of 74.20% and males have a survival rate of 18.89%. The analysis shows that females have a higher survival rate than males. This aligns with the scene in the movie where females can get onboard the rescue boat first.

3) Are rich people have a higher survival rate because they can get onboard to the rescue boat sooner (like what is shown in the movie)?

Based on the analysis, first-class passengers have a survival rate of 62.96%, second-class passengers have a survival rate of 47.28% and third-class passengers have a survival rate of 24.24%. The result supports the hypothesis that rich people have a higher survival rate.