

PROJECT 1

EXPLORING TITANIC DATABASE





UNDERSTANDING THE BUSINESS CONTEXT




BUSINESS CONTEXT

- **What are these data for?**
 - Study, examine, and compile the most important data we can pull from the Titanic dataset.
- **Why do we need this database?**
 - To estimate the number of Titanic passengers who survived the shipwreck and to identify the Titanic passenger survival rate.
- **Where are these data collected?**
 - The Encyclopedia Titanica (<https://www.encyclopedia-titanica.org/>) was the primary source of information used to compile this Titanic report. However, a description of this dataset was collected on the Kaggle website (<https://www.kaggle.com/c/titanic/data>).



UNDERSTANDING THE TECHNICAL CONTEXT

- 
- **How are these data collected?**
 - The datasets were compiled by a range of primary sources including newspapers, photos, and other records that were created at the time the event occurred.
 - **Where are the sources of these data?**
 - These information were gathered from newspaper stories about the Titanic disaster as well as official investigations conducted in the UK and the USA.
 - **What are some of the error sources of this data?**
 - As the incident occurred in the early 1900s, a time when technology were still in their development, therefore it was challenging to gather accurate data for analysis.
 - **Is the data complete? Would there be missing pieces of data?**
 - There were a few columns in the dataset that were invalid for the analysis because they contained missing values.



UNDERSTANDING THE TABLES AND FIELDS



UNDERSTANDING TABLE AND FIELDS

- **How many tables do we have?**
 - There is 1 table in the database, which is the “passengers” table.
- **What are the tables? And what are these tables representing?**
 - The “passengers” table represents 12 data fields which are the passengers’ id, survival status, class, name, sex, age, their siblings or spouse, parent or children, ticket, fare, cabin, and embarked.

UNDERSTANDING TABLE AND FIELDS

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

Fields	Meaning
PassengerId	The passengers' id.
Survived	The survival status of the passenger.
Pclass	The passengers' social class.
Name	The passengers' name.
Sex	The passengers' sex.
Age	The passengers' age.
SibSp	Siblings or spouse that aboard the ship with the passenger.
Parch	Parents or children that aboard the ship with the passenger.
Ticket	The passengers' type of ticket.
Fare	The amount which the passengers paid.
Cabin	The cabin assigned to the passenger.
Embarked	The port where the passengers board the ship.



UNDERSTANDING TABLE AND FIELDS

- **Is the data messy? And how?**
 - Yes, the data is messy.
 - For instance, the cabin, age, and embarked features contain a number of null values.
 - The 'Cabin' column had 687 missing values. The column 'Embarked' that shows a boarding place had a total of 2 missing values. The property 'Age' had 177 missing values.

UNDERSTANDING TABLE AND FIELDS

- **Should I clean the data first? Or ignoring those messy columns.**

- Yes, the data should be clean first.

- Excluding the null values using:

```
SELECT * FROM passengers WHERE column IS NOT NULL;
```

- Check missing data using:

```
SELECT count(*) FROM passengers WHERE column IS NULL;
```



QUESTIONS

QUESTIONS

- In the movie, children, elderlies and females can get onboarded to rescue boat first, so,
 - Are children and elderlies have a higher survival rate in this accident?
 - Assuming children as under 18 years old and elder as more than 50 years old,

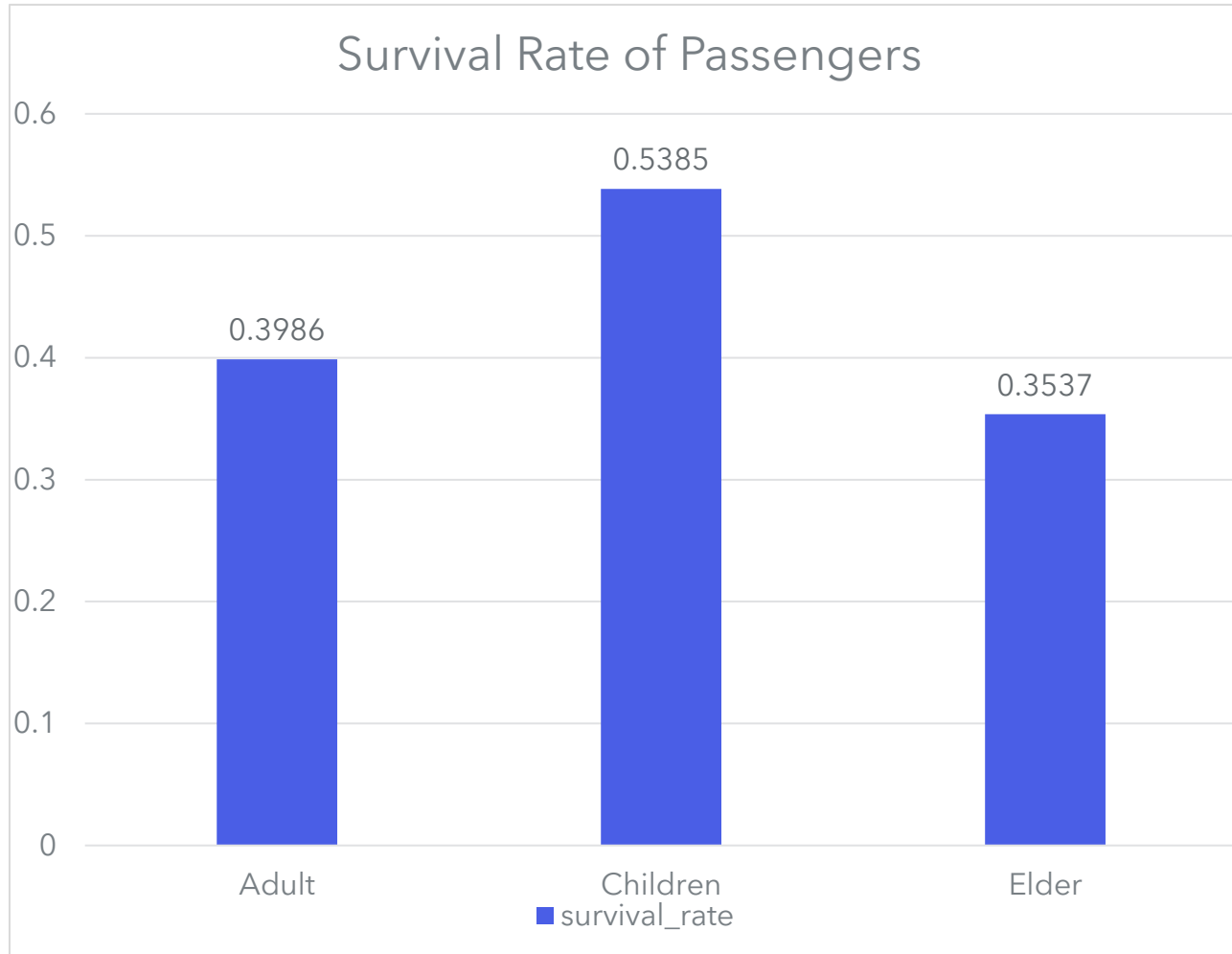
```
SELECT
CASE WHEN Age < 18 THEN 'Children'
      WHEN Age > 50 THEN 'Elder'
      ELSE 'Adult'
END AS age_group,
min(Age) as min_age,
sum(Survived) as total_survived,
count(*) as total_passengers,
round(1.0 * sum(Survived)/count(*),4) as survival_rate
FROM passengers
WHERE Age IS NOT NULL
GROUP BY age_group
```

QUESTIONS

- We Obtained:

age_group	min_age	total_survived	total_passengers	survival_rate
Adult	18	226	567	0.3986
Children	0.42	35	65	0.5385
Elder	51	29	82	0.3537

QUESTIONS



According to the bar graph, **children have a higher chance of surviving this accident** than adults and elders do.

The survival percentage for children is 53.8 percent, or more than half of all passengers. The rate for the elderly is the lowest at 35.37%, which is 5% less than the rate for adults, which is 39.86%.

Thus, it is clear that the children has the highest survival rate compared to the adults and elders in this incident.

QUESTIONS

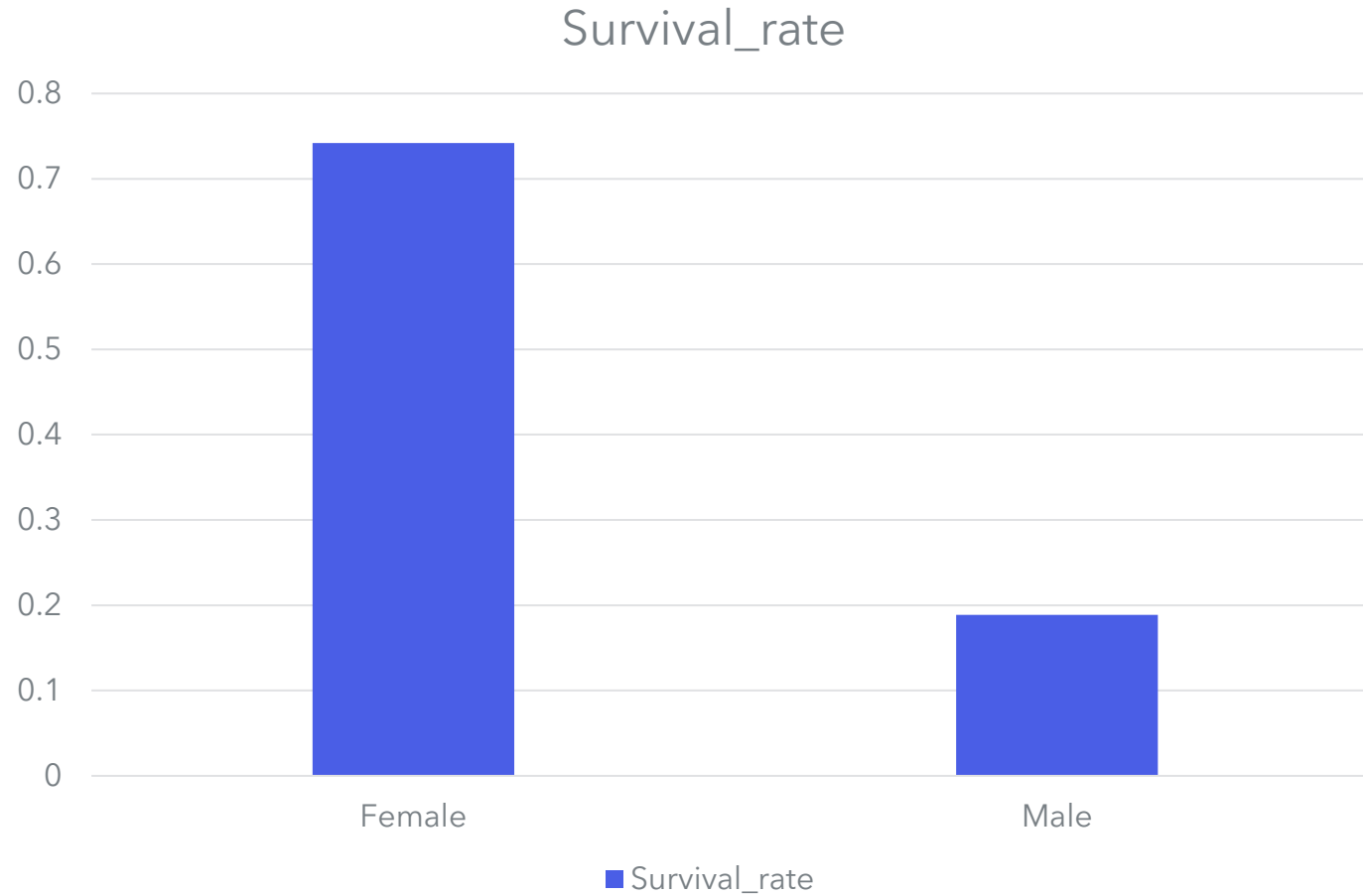
- In the movie, children, elderlies and females can get onboarded to rescue boat first, so
 - Are females more likely to survive in this incident?

```
SELECT
    Sex,
    sum(survived) AS survivors,
    count(*) AS total_passengers,
    round(1.0 * sum(Survived) / count(*),4) AS survival_rate
FROM passengers
GROUP BY Sex
```

- **We obtained:**

Sex	survivors	Total_passengers	Survival_rate
female	233	314	0.742
male	109	577	0.1889

QUESTIONS



Even though there were more male passengers on board, it is evident from the above bar chart that more female passengers survived, with a survival rate of 74.2%, compared to only 18.89% of male passengers.

Therefore, it can be concluded that females are indeed more likely to survive in this incident.

QUESTIONS

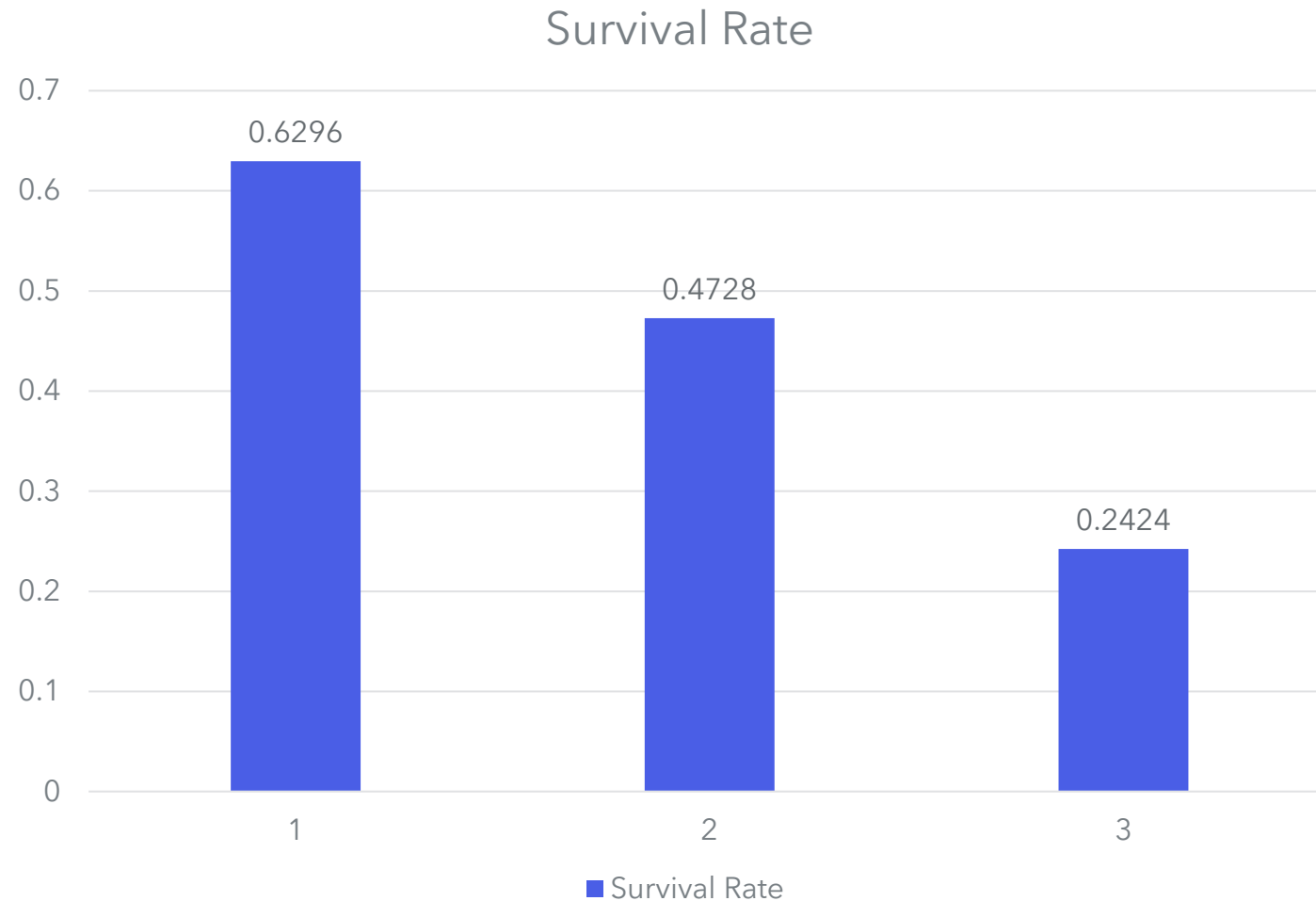
- Are rich people have a higher survival rate because they can get onboard to the rescue boat sooner (like what is shown in the movie)?

```
SELECT Pclass,  
       sum(Survived) AS survivors,  
       count(*) AS class_passengers,  
       round(1.0 * sum(Survived)/count(*),4) AS survival_rate  
FROM passengers  
GROUP BY Pclass
```

- We obtained

Pclass	survivors	Class_passengers	Survival_rate
1	136	216	0.6296
2	87	184	0.4728
3	119	491	0.2424

QUESTIONS



Among the three classes, first class has the highest survival rate, at 63% of all first class passengers.

The third class, which contains over half (47%) of all second class passengers, has the lowest survival percentage when compared to first and second classes. Out of the 491 passengers who make up the third class, 24% of them survive.

As a result, it is clear that the likelihood of survival increases with higher ticket class.



**FREE
EXPLORATION**

FREE EXPLORATION

- What is the total number of passengers?

```
SELECT COUNT(*) AS total_passengers  
FROM passengers;
```

Total_passengers
891

- How many passengers survived?

```
SELECT COUNT(*) AS total_survived  
FROM passengers  
WHERE Survived = 1;
```

Total_survived
342

FREE EXPLORATION

- How many passengers that do not survived?

```
SELECT COUNT(*) AS total_death FROM passengers  
WHERE Survived = 0;
```

Total_death
549

- How many passengers brought their siblings and spouse together?

```
SELECT COUNT(*) AS Sib_Sp FROM passengers  
WHERE SibSp = 1;
```

Sib_Sp
209

209 passengers boarded the ship with their siblings or spouses.