TALENTLABS - MYSTEP PROGRAM

EXPLORING TITANIC DATABASE

FUNDAMENTALS OF SQL PROJECT 1

DATA EXPLORATION FRAMEWORK

- Understanding Business Context
- Understanding Technical Context
- Understanding the Tables and Fields
- Coming up with research questions
- Answering the questions with data analysis

UNDERSTANDING BUSINESS CONTEXT

Understanding Business Context

Background

About The Dataset

RMS Titanic was a ship which sank after hitting an iceberg in the North Atlantic Ocean on 15 April 1912.

It carried an estimate of 2,2224 passengers and crew aboard. The ship carried some richest people as well as emigrants.

The ship only carried 20 lifeboats and only can carry about half the number of passengers.

This tragedy was the most well-known shipwrecks in history.


UNDERSTANDING
TECHNICAL CONTEXT

Source of Data

The data is collected from Kaggle, a well known data science and free dataset repository website.

The Kaggle's Titanic dataset was aimed for users to use Machine Learning to create a prediction model to predict which passenger will survive the Titanic shipwreck.

UNDERSTANDING THE TABLES AND FIELDS

Understanding the Tables and Fields

Fields Provided

Sex

Ticket

Age

Survived

1 = yes ; 0 = no

Fare

Pclass

ticket class

1 = 1st class

2 = 2nd class

3 = 3rd class

Fields Provided

Parch

# of parents /children abroad the Titanic

Name

Cabin

Embarked

ports of embarkation

C = Cherbourg
Q = Queenstown
S = Southampton

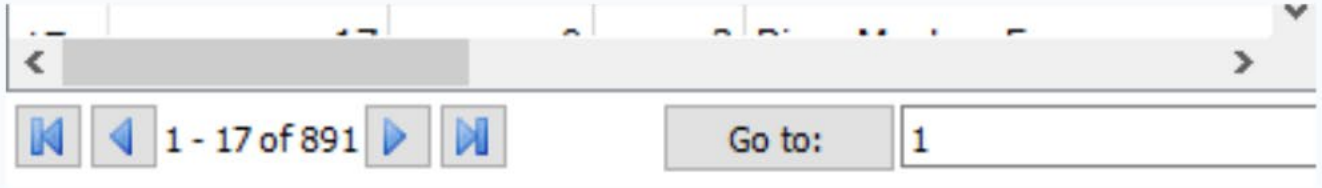
SibSp

# of siblings /spouses abroad the Titanic

PassengerId

Data inspecting & cleaning

There are a total of 891 columns in the table, and few of the missing data found in the table



	2	4	1	382652	29.125	NULL
	NULL	0	0	244373	13	NULL
e	31	1	0	345763	18	NULL
e	NULL	0	0	2649	7.225	NULL

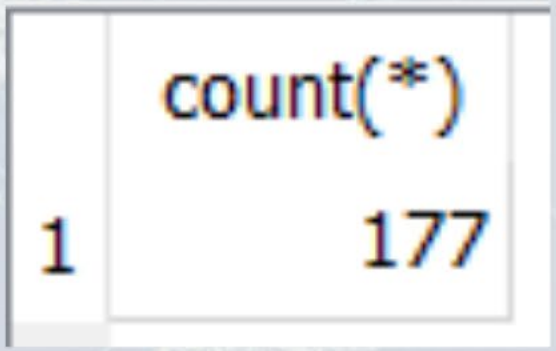
After inspecting the fields, only the data in "Age", "Cabin" and "Embarked" has null value

Age	SibSp	Parch	Ticket	Fare	Cabin
Filter	Filter	Filter	Filter	Filter	Filter
22	1	0	A/5 21171	7.25	NULL
38	1	0	PC 17599	71.2833	C85
26	0	0	STON/O2. 3101282	7.925	NULL
35	1	0	113803	53.1	C123
35	0	0	373450	8.05	NULL
NULL	0	0	330877	8.4583	NULL
54	0	0	17463	51.8625	E46

Data inspecting & cleaning

CHECK HOW MANY MISSING DATA IN "AGE":

```
SELECT COUNT(*)
FROM passengers
WHERE Age IS null OR Age = 0
```



	count(*)
1	177

Data inspecting & cleaning

CHECK HOW MANY MISSING DATA IN "CABIN":

```
SELECT COUNT(*)
FROM passengers
WHERE Cabin IS null OR cabin = 0
```

	count(*)
1	687

Data inspecting & cleaning

CHECK HOW MANY MISSING DATA IN "EMBARKED":

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

	count(*)
1	2

Data inspecting & cleaning

"Age" missing data can be ignored (if "Age" used as a factor in answering the research question) by using the query below:

SELECT * FROM Passengers
WHERE Age IS NOT null

```
Execution finished without errors.
Result: 714 rows returned in 5ms
At line 1:
SELECT *
FROM passengers
WHERE Age IS NOT null
```

A total of 714 rows have no missing data in "Age"


COMING UP WITH
RESEARCH QUESTIONS

Coming up with research questions

QUESTIONS TO TACKLE

WHAT IS THE SURVIVAL RATE OF THE TITANIC?

ARE CHILDREN AND ELDERLIES HAVE A HIGHER SURVIVAL RATE?

ARE FEMALES MORE LIKELY TO SURVIVE IN THIS INCIDENT?

QUESTIONS TO TACKLE

ARE RICH PEOPLE HAVE A HIGHER SURVIVAL RATE BECAUSE THEY GET ONBOARD TO RESCUE BOAT SOONER?

WHAT IS THE SURVIVAL RATE FOR EACH EMBARKATION?

WHAT IS THE SURVIVAL RATE FOR A PERSON WITHOUT FAMILY ONBOARD?

ANSWERING THE QUESTIONS WITH DATA ANALYSIS

Question 1 - What is survival rate of Titanic?

```
SELECT Survived, count(*) as survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as survive_percent
FROM passengers
GROUP BY Survived
```

	Survived	survive_count	survive_percent
1	0	549	61.62
2	1	342	38.38


61.62%
NOT SURVIVED
38.38%
SURVIVED

Question 2 - Are children and elderlies have a higher survival rate?

What are the age considered as children and elderlies?

Children = Under 18

Article 1. For the purposes of the present Convention, a child means **every human being below the age of eighteen years unless under the law applicable to the child, majority is attained earlier**.

https://www.unicef.org › child-rights-convention › conve...

Convention on the Rights of the Child text | UNICEF

Source: UNICEF

Elderlies = Over 60

1 Overview

An older person is defined by the United Nations as a person who is over 60 years of age. However, families and communities often use other socio-cultural referents to define age, including family status (grandparents), physical appearance, or age-related

Source: WHO

Children & Elderlies survival rate

```
SELECT count(*) FROM passengers
WHERE NOT (Age > 18 AND Age < 60)
```

	count(*)
1	132

```
SELECT Survived, count(*) as survive_count,
 round(count(*) * 100.0/sum(count(*)) OVER(),2) as survive_percent
FROM passengers
WHERE NOT (Age > 18 AND Age < 60) AND Age IS NOT NULL
GROUP BY Survived
```

	Survived	survive_count	survive_percent
1	0	76	57.58
2	1	56	42.42

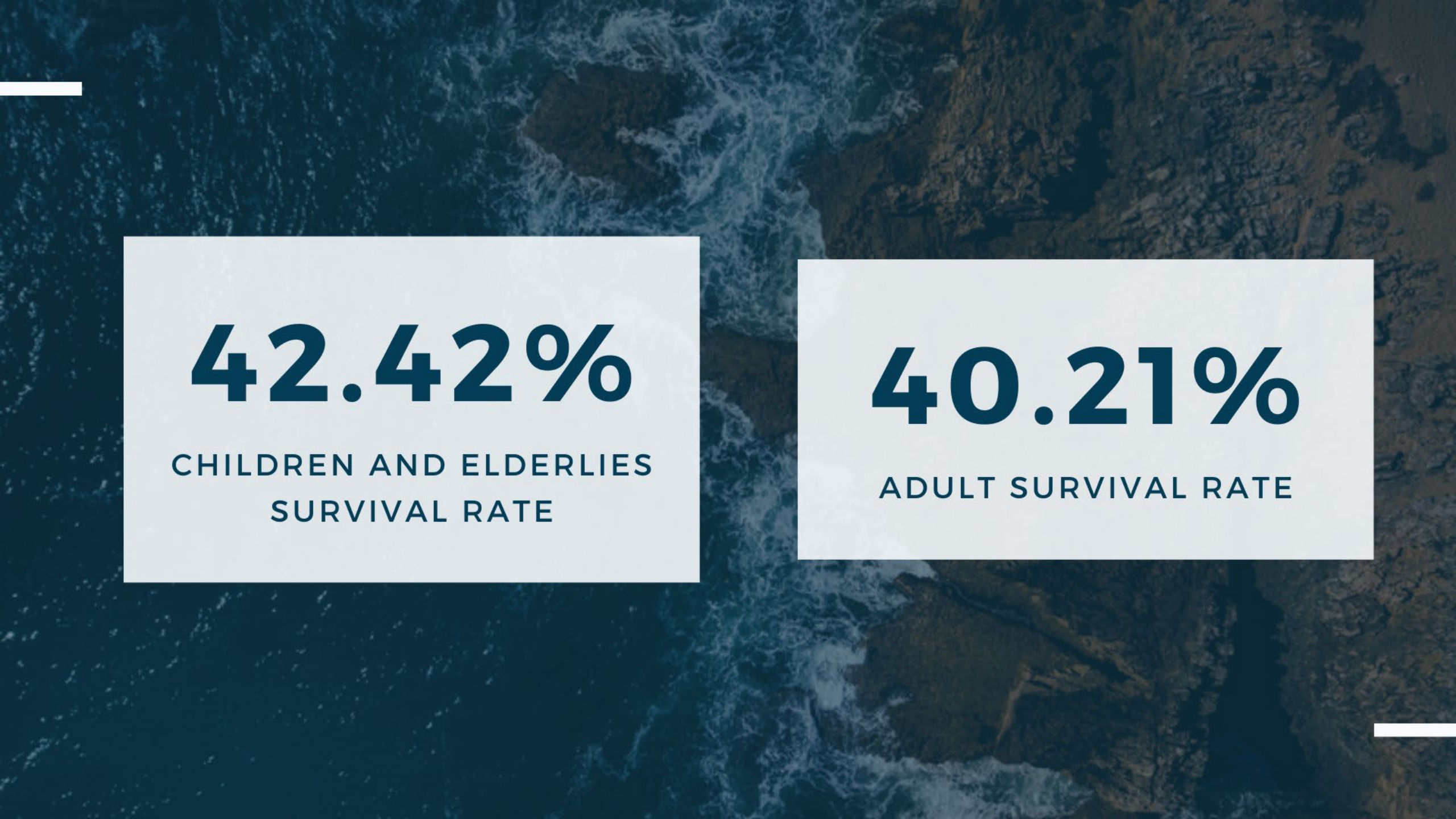
Adult survival rate

```
SELECT count(*) FROM passengers
WHERE (Age > 18 AND Age < 60)
```

	count(*)
1	582

```
SELECT Survived, count(*) as survive_count,
 round(count(*) * 100.0/sum(count(*)) OVER(),2) as survive_percent
FROM passengers
WHERE (Age > 18 AND Age < 60) AND Age IS NOT NULL
GROUP BY Survived
```

	Survived	survive_count	survive_percent
1	0	348	59.79
2	1	234	40.21



42.42%

CHILDREN AND ELDERLIES SURVIVAL RATE

40.21%

ADULT SURVIVAL RATE

Question 3 - Are females more likely to survive in this incident?

Total passengers group by sex

```
SELECT Sex, count(*)
FROM passengers
GROUP BY Sex
```

	Sex	count(*)
1	female	314
2	male	577

Survived

```
SELECT Sex, count(*) as survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as survive_percent
FROM passengers
WHERE Survived = '1'
GROUP BY Sex
```

	Sex	survive_count	survive_percent
1	female	233	68.13
2	male	109	31.87

Not survived

```
SELECT Sex, count(*) as not_survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as not_survive_percent
FROM passengers
WHERE Survived = '0'
GROUP BY Sex
```

	Sex	not_survive_count	not_survive_percent
1	female	81	14.75
2	male	468	85.25


68.13%
FEMALE PASSENGER SURVIVED
31.87%
MALE PASSENGER SURVIVED
14.75%
FEMALE PASSENGER NOT SURVIVED
85.25%
MALE PASSENGER NOT SURVIVED

Question 4 - Are rich people have a higher survival rate because they get onboard to rescue boat sooner?

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

Class count

```
SELECT Pclass, count(*)
FROM passengers
GROUP BY Pclass
```

	Pclass	count(*)
1	1	216
2	2	184
3	3	491

Survived

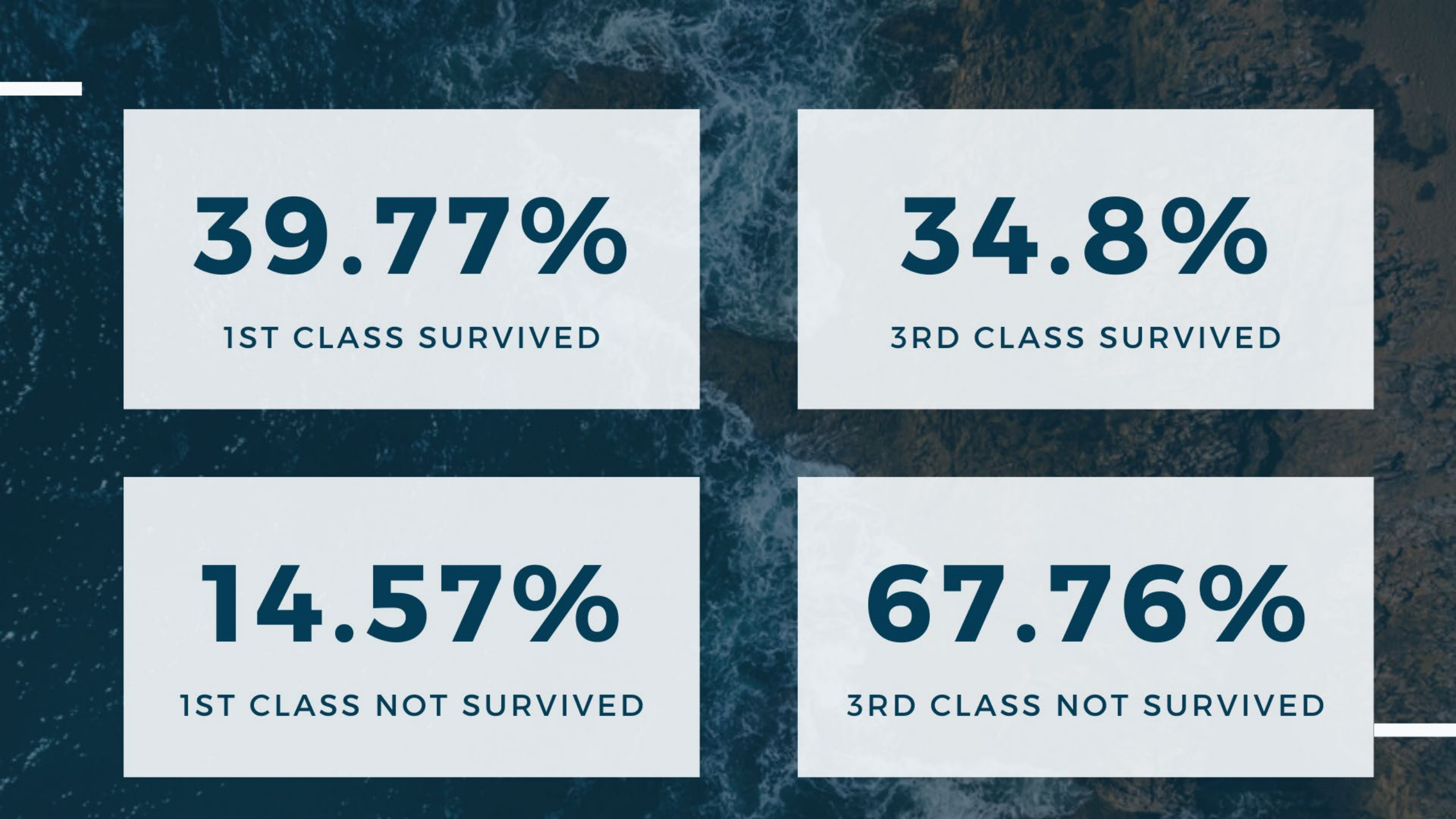
```
SELECT Pclass, count(*) as survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as survive_percent
FROM passengers
WHERE Survived = '1'
GROUP BY Pclass
```

	Pclass	survive_count	survive_percent
1	1	136	39.77
2	2	87	25.44
3	3	119	34.8

Not Survived

```
SELECT Pclass, count(*) as not_survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as not_survive_percent
FROM passengers
WHERE Survived = '0'
GROUP BY Pclass
```

	Pclass	not_survive_count	not_survive_percent
1	1	80	14.57
2	2	97	17.67
3	3	372	67.76

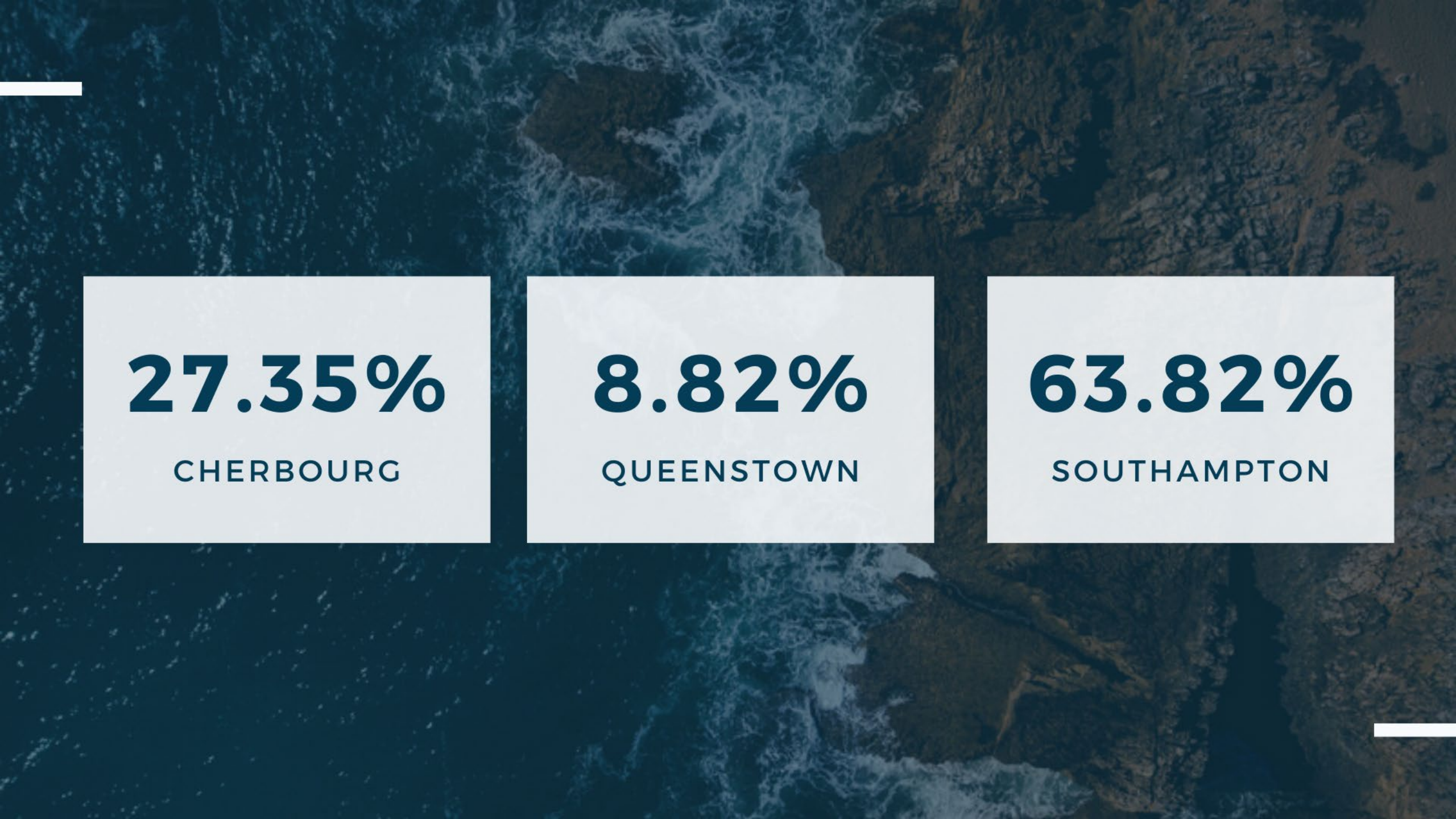

39.77%
1ST CLASS SURVIVED
34.8%
3RD CLASS SURVIVED
14.57%
1ST CLASS NOT SURVIVED
67.76%
3RD CLASS NOT SURVIVED

Question 5 - What is the survival rate for each embarkation?

Survive count & rate for each embark

```
SELECT Embarked, count(*) as survive_count,
 round(count(*) * 100.0/ sum(count(*)) OVER(),2) as survive_percent
FROM passengers
WHERE Survived = '1' AND Embarked IS NOT null
GROUP BY Embarked
```

	Embarked	survive_count	survive_percent
1	C	93	27.35
2	Q	30	8.82
3	S	217	63.82


27.35%
CHERBOURG
8.82%
QUEENSTOWN
63.82%
SOUTHAMPTON

Question 6 - What is the survival rate for a person without family onboard?

```
SELECT Parch, SibSp, count(*)
FROM passengers
GROUP BY Parch, SibSp
```

Parch	SibSp	count(*)
0	0	537
0	1	123
0	2	16
0	3	2
1	0	38
1	1	57
1	2	7
1	3	7
1	4	9
2	0	29
2	1	19
2	2	4
2	3	7
2	4	9
2	5	5
2	8	7
3	0	1
3	1	3
3	2	1
4	0	1
4	1	3
5	0	2
5	1	3
6	1	1

```
SELECT Parch,
        SibSp,
        count(*) as survive_count,
        round(count(*)  *  100.0/sum(count(*))
OVER(),2) as survive_percent
FROM passengers
WHERE Survived = '1'
GROUP BY Parch, SibSp
```

Parch	SibSp	survive_count	survive_percent
0	0	163	47.66
0	1	64	18.71
0	2	4	1.17
0	3	2	0.58
1	0	25	7.31
1	1	34	9.94
1	2	6	1.75
2	0	21	6.14
2	1	12	3.51
2	2	2	0.58
2	3	2	0.58
2	4	3	0.88
3	0	1	0.29
3	1	1	0.29
3	2	1	0.29
5	1	1	0.29

Survival rate

47.66%

PARCH = 0,
SIBSP = 0

CONCLUSIONS

38.38% SURVIVED IN TITANIC

CHILDREN AND ELDERLIES HAVE HIGHER SURVIVAL RATE

FEMALE PASSENGERS ARE MORE LIKELY TO SURVIVE

UPPER CLASS ARE MORE LIKELY TO SURVIVE

CHERBOURG = 27.35 %
QUEENSTOWN = 8.82 %
SOUTHAMPTON = 63.82%

SURVIVAL RATE FOR PERSON WITHOUT FAMILY IS 47.66 %

Sources

TITANIC

Wikipedia article

TITANIC - MACHINE LEARNING FROM DISASTER

Kaggle

SQL - HTTPS://STACKOVERFLOW.COM/QUESTIONS/37303779/SQL-CALCULATE-PERCENTAGE-ON-COUNTCOLUMN

Stack Overflow

HTTPS://WWW.WHO.INT/HEALTH-TOPICS/AGEING#TAB=TAB_1

WHO

HTTPS://WWW.UNICEF.ORG/CHILD-RIGHTS-CONVENTION/CONVENTION-TEXT#:~:TEXT=FOR%20THE%20PURPOSES%20OF%20THE,CHILD%2C%20MAJORITY%20IS%20ATTAINED%20EARLIER.

UNICEF