**About Dataset**

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

This work is describing how much can I get the information from Titanic dataset (this dataset contains 1309 rows, this is largest as I can found in the internet). I will try to show why some people stayed alive and will try to provide facts which are make affect on their surviving. Also I will show regularity.

I will process the data with help of SQL queries. My workbench will be MSSQL. My target is also to show how much SQL can help in this research.

In the beginning, I will join all tables to see the whole data. I have selected top 10 rows.

-- Joining all tables

SELECT TOP 10 PT1.\*, TI2.fare, DT3.embarked, DT3.home\_dest, CI4.pclass

FROM Titanik..Passengers\_titanik PT1

JOIN Titanik..Ticket\_info TI2

ON PT1.ticket = TI2.ticket

JOIN Titanik..Destination\_titanik DT3

ON PT1.personal\_id = DT3.ticket\_id

JOIN Titanik..Cabine\_info CI4

ON PT1.ticket = CI4.ticket



Let`s start with the analysis of the percentage of those who survived.

**1. Survival rate**

-- Percentage of people who survived vs died

SELECT if\_survived, COUNT(if\_survived) AS survival\_rate, CONCAT(CAST(CAST(COUNT(if\_survived) AS DECIMAL (10,2))

/CAST((SELECT COUNT(if\_survived) FROM Titanik..Passengers\_titanik) AS DECIMAL (10,2))AS DECIMAL (10,2))\*100, ' %') AS percentage,

(SELECT COUNT(if\_survived) FROM Titanik..Passengers\_titanik WHERE if\_survived <> 'Unknown') AS total

FROM Titanik..Passengers\_titanik

WHERE if\_survived <> 'Unknown'

GROUP BY if\_survived

ORDER BY survival\_rate;



As we can see from 1309 passengers survived only 500 of them it`s 38% percent.

Let`s deep into the data and try to understand who survived.

I will analyze the difference between gender and then age.

**2. Gender**

-- Survived(The difference between gender)

WITH total\_table1 AS

(

SELECT gender, COUNT(personal\_id) AS total\_passengers,

CONCAT(CAST(CAST(COUNT(personal\_id)AS DECIMAL (10,2))/(SELECT COUNT(personal\_id) FROM Titanik..Passengers\_titanik) \*100 AS DECIMAL(10,2)), ' %') AS gender\_rate

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL

GROUP BY gender

), total\_table2 AS

(

SELECT gender,COUNT(if\_survived) AS count\_suevived

FROM Titanik..Passengers\_titanik

WHERE if\_survived = 'Yes'

GROUP BY gender

)

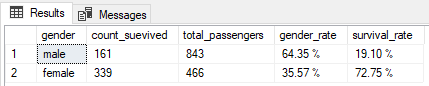
SELECT tab1.gender , tab2.count\_suevived, tab1.total\_passengers, tab1.gender\_rate,

CONCAT(CAST((CAST(tab2.count\_suevived AS DECIMAL(10,2))/tab1.total\_passengers)\*100 AS DECIMAL(10,2)), ' %') AS survival\_rate

FROM total\_table1 tab1

LEFT JOIN total\_table2 tab2

ON tab1.gender = tab2.gender



On board of the Titanic were 1317 passengers and 908 crew members, a total amount were 2225 people.

According to Great Britain law on the ship should have been 20 boats, which could handle 1178 passengers.

Captain understand that this not enough for all members, so he gave order to save females and children first.

As we can see on the analyze females have more survival rate than males, 72% vs 19%.

According to this, females were in twice less than males, 35% vs 64%.

So, we can be sure that the main sign of survival was gender.

Let`s check the age variations, to make sure that the children were saved first.

**3. Age**

-- Survived(The difference between age\_variations)

WITH total\_table3 AS

(

SELECT gender, variations, COUNT(if\_survived) AS count\_survived

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL AND if\_survived = 'Yes'

GROUP BY gender, variations

), total\_table4 AS

(

SELECT gender, variations, COUNT(personal\_id) AS count\_total

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL

GROUP BY gender, variations

)

SELECT tab3.gender, tab3.variations, tab3.count\_survived, tab4.count\_total,

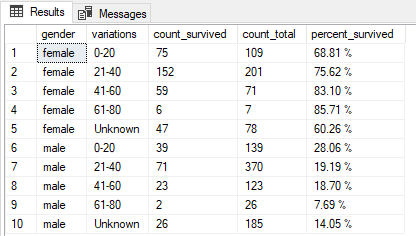
CONCAT(CAST((CAST(tab3.count\_survived AS DECIMAL(10,2))/tab4.count\_total)\*100 AS DECIMAL(10,2)), ' %') AS percent\_survived

FROM total\_table3 tab3

RIGHT JOIN total\_table4 tab4

ON tab3.variations = tab4.variations AND tab3.gender = tab4.gender

ORDER BY gender, variations



Earlier I made the variations by age cause the range was huge, so only 5 variations.

Splitting the age through the groups we can see that all females have more survived rate in any age category than males.

The highest survival rate had the males in category “0-20” than others, so the children were prioritized.

Also, we can analyze what is the survival rate by the title name of people.

Let`s filter by the title name and find how many people have survived.

**4. Title**

--Survived(The difference between title name)

WITH total\_table5 AS

(

SELECT title, COUNT(CASE WHEN if\_survived = 'Yes' THEN 1 ELSE NULL END) AS count\_survived

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL AND title <> ' Martin (Elizabeth L'

GROUP BY title

),total\_table6 AS

(

SELECT title, COUNT(personal\_id) AS count\_total

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL AND title <> ' Martin (Elizabeth L'

GROUP BY title

)

SELECT tab6.title, tab5.count\_survived, tab6.count\_total, CONCAT(CAST((CAST(tab5.count\_survived AS DECIMAL(10,2))/tab6.count\_total)\*100 AS DECIMAL(10,2)), ' %') AS percent\_survived

FROM total\_table5 tab5

RIGHT JOIN total\_table6 tab6

ON tab5.title = tab6.title

--Another variant of the previous query

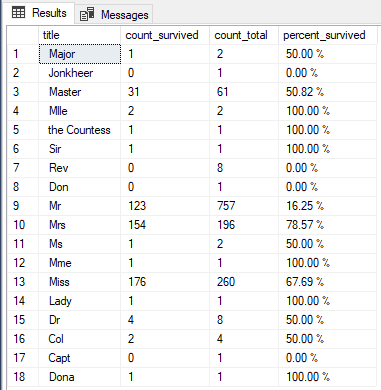
SELECT title, COUNT(CASE WHEN if\_survived = 'Yes' THEN 1 ELSE NULL END) AS count\_survived ,COUNT(personal\_id) AS count\_total,

CONCAT(CAST((CAST(COUNT(CASE WHEN if\_survived = 'Yes' THEN 1 ELSE NULL END) AS DECIMAL(10,2))/COUNT(personal\_id))\*100 AS DECIMAL(10,2)), ' %') AS percent\_survived

FROM Titanik..Passengers\_titanik

WHERE gender IS NOT NULL AND title <> ' Martin (Elizabeth L'

GROUP BY title



In total we have 18 title names, and as we can see a huge amount is situated in title names ‘Master’(children), ‘Mr’(men), ‘Mrs’(not married woman), ‘Miss’(married woman).

So, we can unite all data by this 4 main title names.

--Grouping title names

WITH table\_tem1 AS

(

SELECT

CASE

WHEN TRIM(title) IN ('Dr', 'Rev', 'Col', 'Major', 'Jonkheer', 'Sir', 'Don', 'Capt') AND gender = 'Male' THEN 'Mr'

WHEN TRIM(title) IN ('Mlle', 'Ms') AND gender = 'Female' THEN 'Miss'

WHEN TRIM(title) IN ('Dr', 'the Countess', 'Mme', 'Lady', 'Dona','Martin (Elizabeth L') AND gender = 'Female' THEN 'Mrs'

ELSE TRIM(title)

END AS grouping\_title\_names,

(CASE WHEN if\_survived = 'Yes' THEN 1 ELSE NULL END) AS count\_survived, personal\_id

FROM Titanik..Passengers\_titanik

WHERE title IS NOT NULL

)

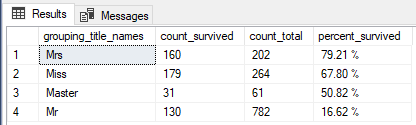
SELECT grouping\_title\_names, COUNT(count\_survived) AS count\_survived, COUNT(personal\_id) AS count\_total,

CONCAT(CAST((CAST(COUNT(count\_survived)AS DECIMAL(10,2))/COUNT(personal\_id))\*100 AS DECIMAL(10,2)), ' %') AS percent\_survived

FROM table\_tem1

GROUP BY grouping\_title\_names

HAVING COUNT(count\_survived) <> 0



As we can see females(79% and 67%) and children(50%) have the highest survival rate.

So, the decision to save females and children was prioritized.

The next analyze we can make though the cabin class and boarding place.

The Titanic`s cabins were divided into three classes

1. First class was closest to the deck of Titanic
2. Second class was in the middle part of Titanic
3. Third class was closest to the bottom of the ship

**5. Cabin class**

--Survived(The difference between cabin class)

SELECT CI2.pclass, COUNT(CASE WHEN PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END) AS count\_survived, COUNT(PT1.personal\_id) AS count\_total,

CONCAT(CAST((CAST(COUNT(CASE WHEN PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END) AS DECIMAL(10,2))/COUNT(PT1.personal\_id)\*100)AS DECIMAL(10,2)), ' %') AS percent\_survived

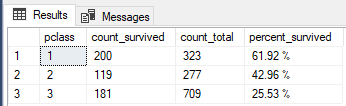
FROM Titanik..Passengers\_titanik PT1

JOIN Titanik..Cabine\_info CI2

ON PT1.ticket = CI2.ticket

GROUP BY CI2.pclass

ORDER BY CI2.pclass



It is possible to see a trend of decreasing survival rate with decrease in the level of cabins 61% for first class, 42% for second class, 25% for third class and the overwhelming majority of passengers were in third class cabins.

Now let`s analyze how many males and females were in these classes.

--Survived(The difference between cabin class and gender)

SELECT CI2.pclass, COUNT(PT1.personal\_id) AS count\_total,

COUNT(CASE WHEN PT1.gender = 'Male' THEN 1 ELSE NULL END) AS male\_count\_total, COUNT(CASE WHEN PT1.gender = 'Female' THEN 1 ELSE NULL END) AS female\_count\_total,

CONCAT(CAST((CAST(COUNT(CASE WHEN PT1.gender = 'Male' AND PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END)AS DECIMAL(10,2))/COUNT(CASE WHEN PT1.gender = 'Male' THEN 1 ELSE NULL END))\*100 AS DECIMAL(10,2)), ' %') AS male\_survived\_rate,

CONCAT(CAST((CAST(COUNT(CASE WHEN PT1.gender = 'Female' AND PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END)AS DECIMAL(10,2))/COUNT(CASE WHEN PT1.gender = 'Female' THEN 1 ELSE NULL END))\*100 AS DECIMAL(10,2)), ' %') AS female\_survived\_rate,

COUNT(CASE WHEN PT1.gender = 'Male'AND PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END) AS male\_count\_survived, COUNT(CASE WHEN PT1.gender = 'Female' AND PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END) AS female\_count\_survived,

COUNT(CASE WHEN PT1.if\_survived = 'Yes' THEN 1 ELSE NULL END) AS count\_survived

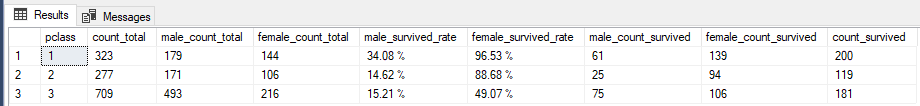
FROM Titanik..Passengers\_titanik PT1

JOIN Titanik..Cabine\_info CI2

ON PT1.ticket = CI2.ticket

GROUP BY CI2.pclass

ORDER BY CI2.pclass



As we can see on the analyze females have more survival rate than males.

The last thing that I will analyze will be the survival by the amount of families. I wanna split all data by the variations like ‘alone’(solo passenger), ‘one\_family’(X amount of passengers from same family), ‘mix\_family’(X amount of passengers what have the same ticket number and the same cabin, so they were together)

**6. Family type**

--The difference between family type

SELECT

CASE

WHEN f.family\_count = 1 AND f.count\_passengers = 1 THEN 'alone'

WHEN f.family\_count = 1 AND f.count\_passengers > 1 THEN 'one\_falmily'

WHEN f.family\_count > 1 AND f.count\_passengers > 1 THEN 'mix\_falmily'

ELSE '0'

END AS type\_falmily, COUNT(f.ticket) AS count\_ticket, SUM(count\_passengers) AS total\_passengers

FROM

(

SELECT ticket, COUNT(DISTINCT(family)) AS family\_count,

COUNT(personal\_id) AS count\_passengers

FROM Titanik..Passengers\_titanik

GROUP BY ticket

HAVING COUNT(DISTINCT(family)) <> 0

)f

GROUP BY CASE

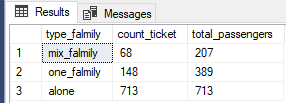
WHEN f.family\_count = 1 AND f.count\_passengers = 1 THEN 'alone'

WHEN f.family\_count = 1 AND f.count\_passengers > 1 THEN 'one\_falmily'

WHEN f.family\_count > 1 AND f.count\_passengers > 1 THEN 'mix\_falmily'

ELSE '0'

END



Here we can see how many passengers traveled alone, with family or have mix family.

Now I will these families by gender.

--Survived(The difference between family type by gender)

SELECT

CASE

WHEN f.family\_count = 1 AND f.count\_passengers = 1 THEN 'alone'

WHEN f.family\_count = 1 AND f.count\_passengers > 1 THEN 'one\_falmily'

WHEN f.family\_count > 1 AND f.count\_passengers > 1 THEN 'mix\_falmily'

ELSE '0'

END AS type\_falmily, COUNT(f.ticket) AS count\_ticket, SUM(f.male\_count) AS male\_count\_survived,

SUM(f.female\_count) AS female\_count\_survived, SUM(count\_passengers) AS count\_passengers\_total , SUM(survival\_count) AS survival\_count,

CONCAT(CAST((CAST(SUM(f.male\_count)AS DECIMAL(10,2))/SUM(survival\_count))\*100 AS DECIMAL (10,2)), ' %') AS survival\_rate\_male,

CONCAT(CAST((CAST(SUM(f.female\_count)AS DECIMAL(10,2))/SUM(survival\_count))\*100 AS DECIMAL (10,2)), ' %') AS survival\_rate\_female

FROM

(

SELECT ticket, COUNT(DISTINCT(family)) AS family\_count, COUNT(CASE WHEN gender = 'Male' AND if\_survived = 'Yes' THEN 1 ELSE NULL END) AS male\_count,

COUNT(CASE WHEN gender = 'Female' AND if\_survived = 'Yes' THEN 1 ELSE NULL END) AS female\_count,

COUNT(personal\_id) AS count\_passengers , COUNT(CASE WHEN if\_survived = 'Yes' THEN 1 ELSE NULL END) AS survival\_count

FROM Titanik..Passengers\_titanik

GROUP BY ticket

HAVING COUNT(DISTINCT(family)) <> 0

)f

GROUP BY CASE

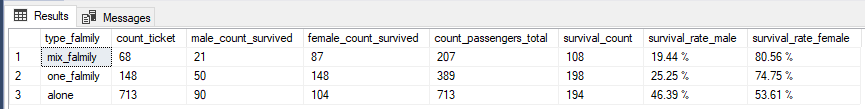
WHEN f.family\_count = 1 AND f.count\_passengers = 1 THEN 'alone'

WHEN f.family\_count = 1 AND f.count\_passengers > 1 THEN 'one\_falmily'

WHEN f.family\_count > 1 AND f.count\_passengers > 1 THEN 'mix\_falmily'

ELSE '0'

END



As we can see on the analyze females have more survival rate than males in all family types.

At the conclusion, the females and children have the most survival rate. First class have the most survival rate.

Hypothesis: The Captain gave order to save females and children first, confirmed.