Analytical Project

TITANIC DATA ANALYSIS

Assylbek Bugybay

ABSTRACT

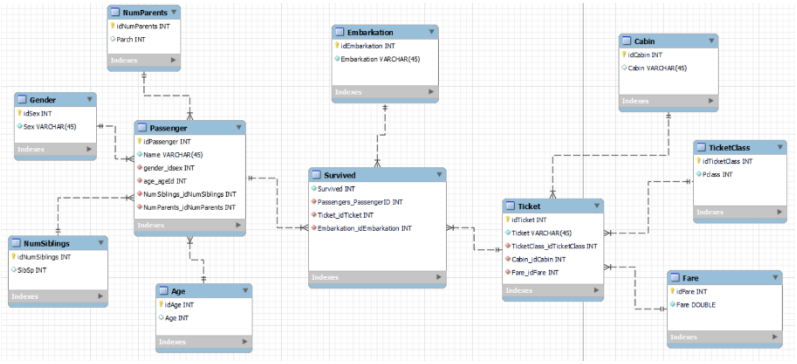
Titanic was supposed to be unsinkable. Despite that Titanic sank on its first voyage. The greatest tragedy of all may be that there were not enough lifeboats for everyone on board.
It seems that some groups of people were more likely to survive than others, having data of passengers it is possible to predict what sort of people had more chance to survive.

1. CONCEPT FOR PROBLEM

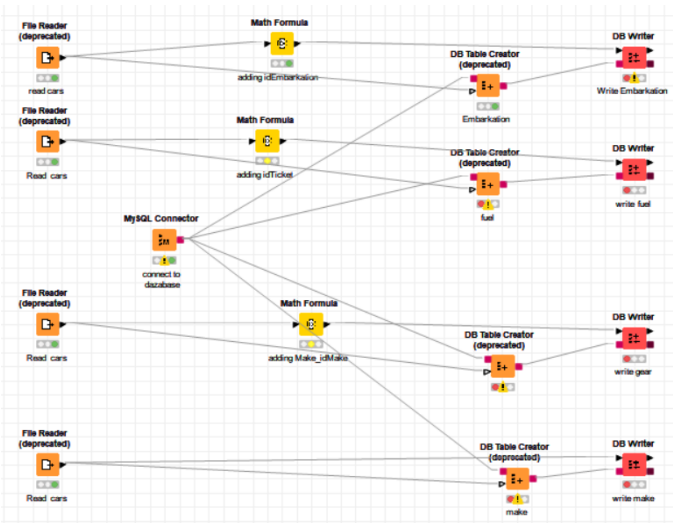
I used the following tools to conduct our research:
1) MySQL Workbench – I created the Snowflake schema with forwards engineering and solved queries in order to find information and make analysis
2) Excel – in order to create Pivot tables
3)Power BI- in order to visualize the results
4) Knime Analytics Platform – I used Knime for Data Insertion

2. SNOWFLAKE AND DATA MART SCHEMA

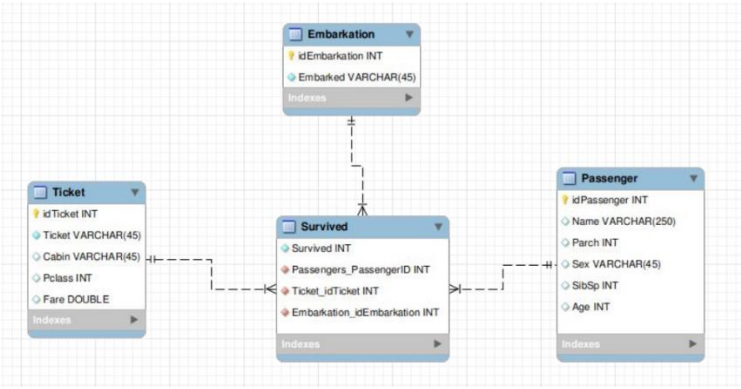
First of all, I modelled ER model in MySQl Workbench. Fact table includes data column Survived as well as foreign keys of Dimension tables. All the other columns are in Dimension tables.



Then I used Knime Analytical Platform and created a workflow to import data from csv file into the Database, which is represented in Snowflake schema.



Next step was deriving a Data Mart in order to make reporting and analysis of data easier. I used SQL queries to further adjust the data needed to report.



Last step is to provide visualization in my report. For that I used Excel to create pivot tables and Power BI to visualize graphs.

3. ANALYSIS

Firstly the SQl queries were written for each analysis question.

```
SELECT IFNULL(t.Pclass, "All classes") AS "Pclass", IFNULL(s.Survived, "Total of 1st class passenger") AS "Survived",
IFNULL(p.Sex, "All genders") AS "Sex", COUNT(*) as "NumberSurvived"
FROM Passenger p JOIN Survived s ON s.Passengers_PassergerID = p.idPassenger
JOIN Ticket t ON s.Ticket_idTicket = t.idTicket
WHERE Pclass = 1
GROUP BY t.Pclass, s.Survived, p.Sex
WITH ROLLUP

UNION

SELECT IFNULL(t.Pclass, "All classes") AS "Pclass", IFNULL(s.Survived, "All 2 class passengers"),
IFNULL(p.Sex, "All genders"),
COUNT(*) as "NumberSurvived"
FROM Passenger p  JOIN Survived s ON s.Passengers_PassergerID = p.idPassenger
JOIN Ticket t ON s.Ticket_idTicket = t.idTicket
WHERE Pclass = 2
GROUP BY t.Pclass, s.Survived, p.Sex
WITH ROLLUP

UNION

SELECT IFNULL(t.Pclass, "All classes") AS "Pclass", IFNULL(s.Survived, "All 3 class passengers"),
IFNULL(p.Sex, "All genders"),
COUNT(*) as "NumberSurvived"
FROM Passenger p JOIN Survived s ON s.Passengers_PassergerID = p.idPassenger
JOIN Ticket t ON s.Ticket_idTicket = t.idTicket
WHERE Pclass = 3
GROUP BY t.Pclass, s.Survived, p.Sex
WITH ROLLUP;
```

The second step was to create pivot tables to see the results in a more structured way and for visualization purposes charts were created for these pivot tables. This pivot table was created using Excel.

Sum of NumberSurvived	Column Labels		
Ticket class	**female**	**male**	**Grand Total**
1	**85**	**101**	**186**
0	3	61	64
1	82	40	122
2	**74**	**99**	**173**
0	6	84	90
1	68	15	83
3	**102**	**253**	**355**
0	55	215	270
1	47	38	85
Grand Total	**261**	**453**	**714**

1- survived
0- not survived

Above you can see the number of people survived depending on Ticket Class and Gender. Below diagrams show the unequal distribution, that percentage of people with 1st and 2nd class ticket survived 2-3 times more compared to those with 3rd class tickets. Percentage of survived passengers from 1st class compared to those from 2nd class is also higher.

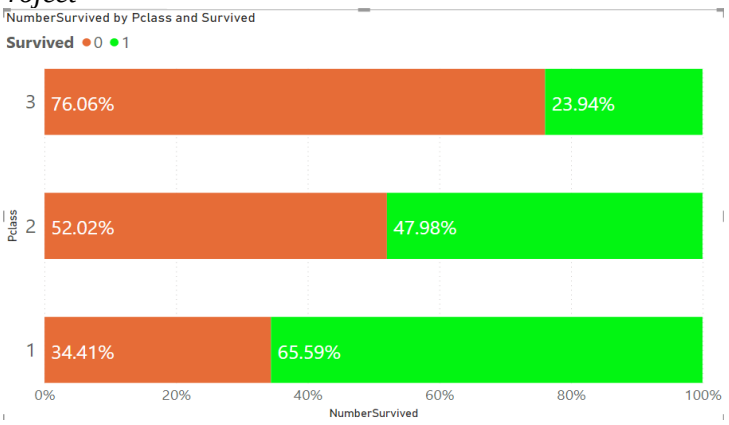
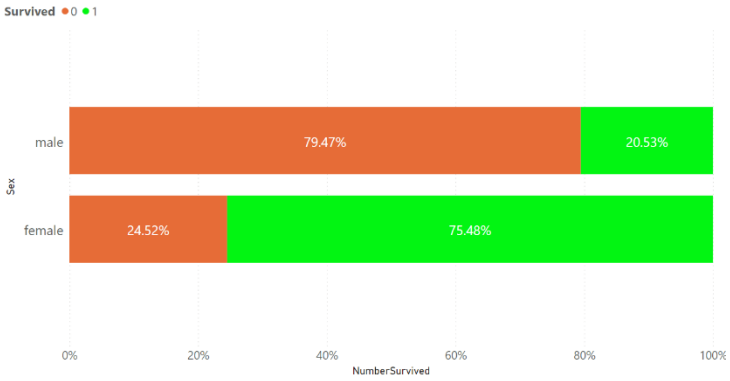
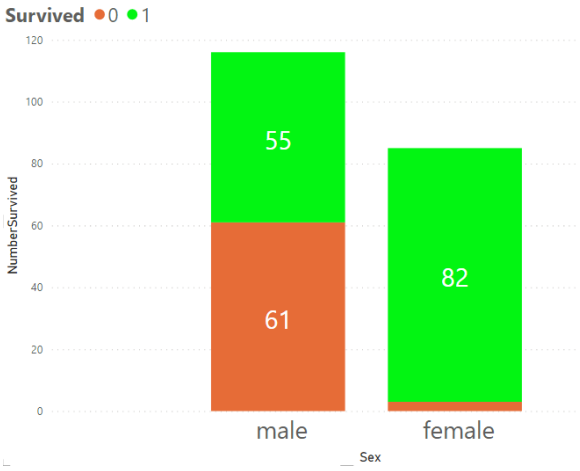


Diagram below shows percentage of Female and Male survived (only gender is taken into account).

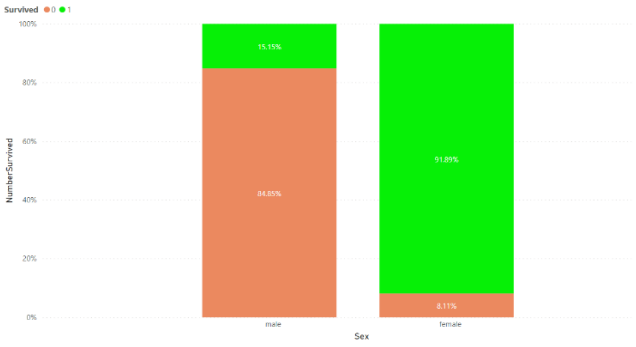


It should be also noticeable that, between 1st, 2nd and 3rd classes among Male and Female passengers were different survival rates. Next 3 diagrams illustrate these tendencies.

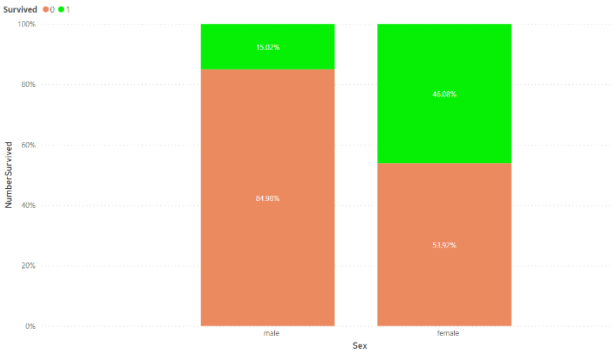
1st class:



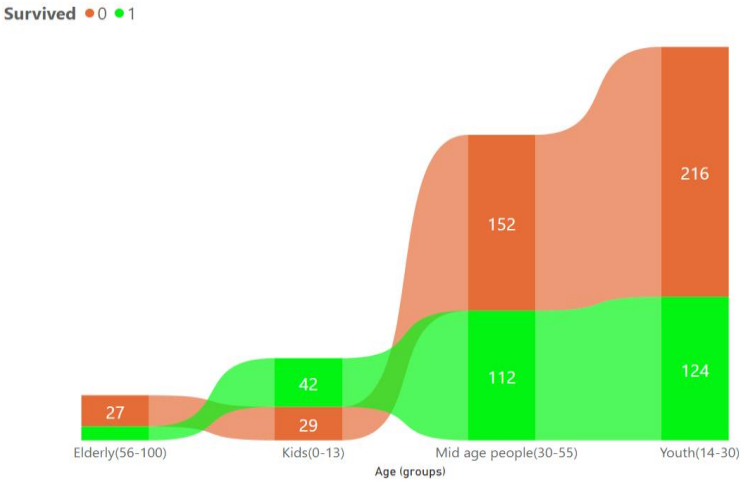
2nd class (male- first column, female- second column):



3rd class:



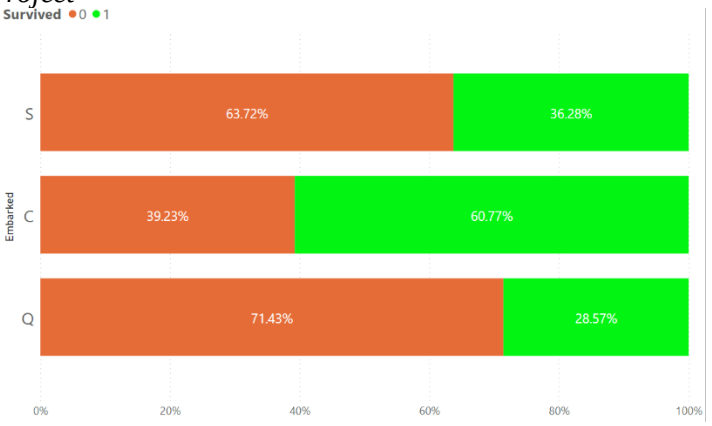
3.1 Which age group tend to be in advantage to survive?



I classified the ages into four groups, as a result we can see Kids (0-13) have an advantage to be rescued. But there is a huge drop in numbers for Mid Age people (30-55) and Youth(14-30) as a most of them could not survive.

3.2 Was the port of embarkation a criteria to influence survival rates in Titanic?

S – stands for Southampton, C = Cherbourg, Q = Queenstown. More than 60% of passengers from Cherbourg survived. Furthermore, there is rapid drop on the percentage of passengers survived, who embarked in Southampton and Queenstown.



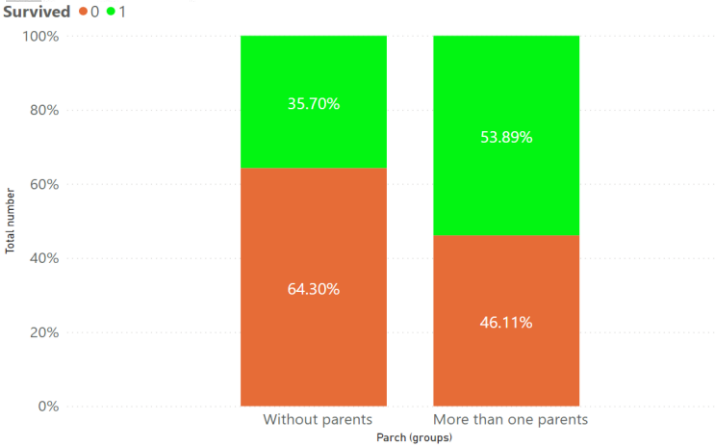
So, it is clear to see port of embarkation played important role on the probability distribution.

3.3 How important was to have a parent or other family members on board?

First using SQL I wrote the query to find out number of people on board with parents or family members. Result of this query the table below.

#	Parch	Survived	Total number
1	0	0	335
2	0	1	186
3	1	0	49
4	1	1	61
5	2	0	29
6	2	1	39
7	3	0	2
8	3	1	3
9	4	0	4
10	5	0	4
11	5	1	1
12	6	0	1

I analyzed this data further. Using Power BI I grouped rows showing Parch more than or equal to 1 (having at least one parent or family member). After thatI derived the following diagram:



4. CONCLUSION

As a recap, with Data Analyze I analyzed the following points:

- Most crucial criteria to survive was the class of the ticket and gender of the passengers.
- Furthermore kids (0-13 years old) and kids with parents had also more chances to survive.
- Another important point is that almost all Females from 1st and 2nd class survived, whereas in 3rd class females had much less possibility to survive.