Exploring titanic database

Context and background of database

Titanic database was widely used for data analysis tutorials and demonstrations to predict the survival rate from sinking of RMS Titanic incidence. The sinking of the ship was very famous among the shipwreck history. On 15 April 1912, during its first trip, the unexpected accident was occurred which the “unsinkable ship” sank after colliding with an iceberg. Among 2224 passengers and crew, 1502 people was reported death due to insufficient lifebuoy on the ship but some group of people was luckily survived even without the use of lifebuoy. The dataset was containing the actual information about the passenger aboard from the Titanic excluding the crew. The data was collected by variety of researchers and one of the original sources comes from Eaton and Hans (1994) which has been listed in their book which is *Titanic: Triumph and Tragedy*. The details about each passenger also can be read at https://www.encyclopedia-titanica.org/.

Technical context

The Titanic data can be useful for teaching the basic of statistical computing and also for demonstrating many functions including applying some logic and analysis to the data. The data of Titanic passenger list was edited and expanded by Michael A. Findlay with the help of internet community. The data containing all information without missing passenger except the crew also includes the survival status which 1 refer to survive and 0 refer to reported death. However, some of the information of the passenger like age, cabin number and port of embarkation was undetermined. Even though the cabin number and port of embarkation are less important to do the analysis about the survival rate, the age of passenger is still important to analyst the range of age that take the priority when incident happens. Some data type must be change especially for some calculation and determination of the range of age.

Tables and field

This Titanic database only contains 1 table name passenger. The table contains 12 field which are:

PassengerID	:	Unique Id of the passenger
Survived	:	Survival status 0 = No, 1 = Yes
Pclass	:	Ticket Class based on socio-economic status (SES) 1=1st class, 2 = 2nd class, 3= 3rd class
Name	:	Passenger Name

Sex	:	Passenger Gender
Age	:	Passenger Age (Year)
SibSp	:	Number of siblings or spouses aboard in Titanic
Parch	:	Number of parent or children aboard in Titanic
Ticket	:	Ticket Number
Fare	:	Passenger Fare
Cabin	:	Cabin Number
Embarked	:	Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

The data was straightforward containing each detail of the passenger. By doing simple command in SQLite which is

```
SELECT count(*), count(DISTINCT Name) FROM passengers (1)
```

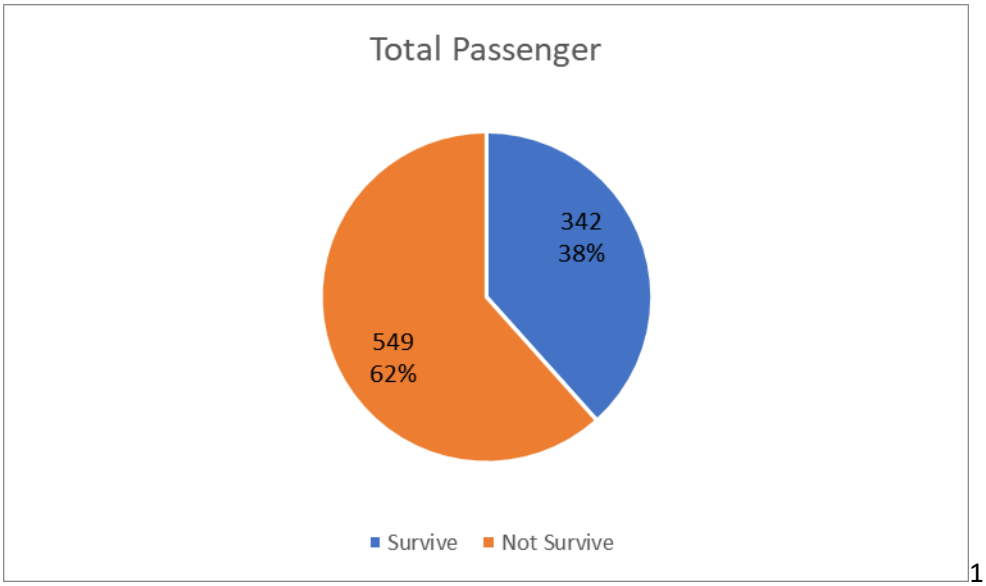
It is shows that there is no duplicate data from this table. To check whether there is an unknown value or not, next line was executed which is

```
SELECT count(*) FROM passengers where name IS NOT NULL (2)
```

From line (2), it is clearly shows that there is no unknown passenger. Therefore, we can use this data to perform the next analysis.

Data analysis

All the data analysis was running using SQLite.

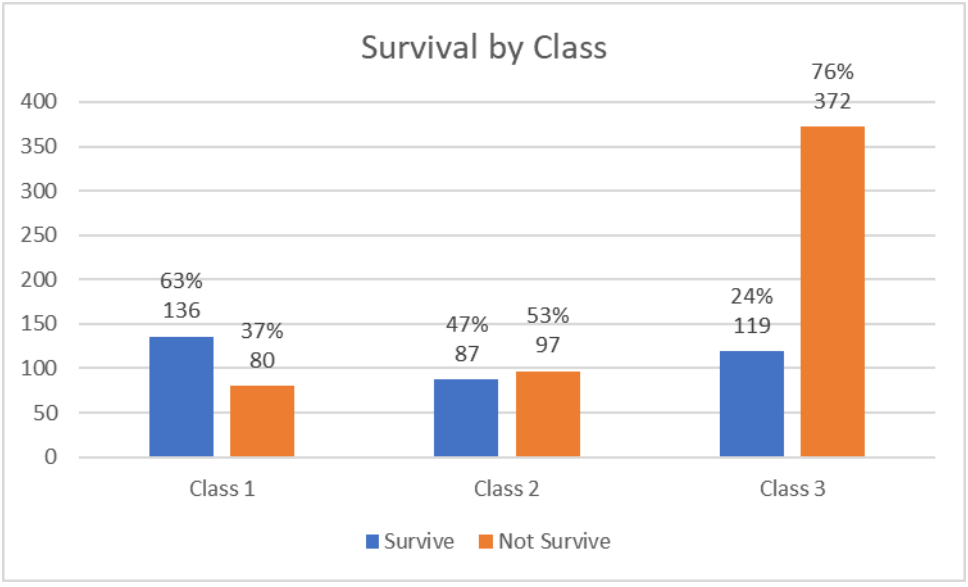


11

SQLite code:

SELECT count(*) AS totalsurvive FROM passengers WHERE Survived=1) (Count survival)

SELECT count(*) AS notsurvive FROM passengers WHERE Survived=0 (Count did not survive)

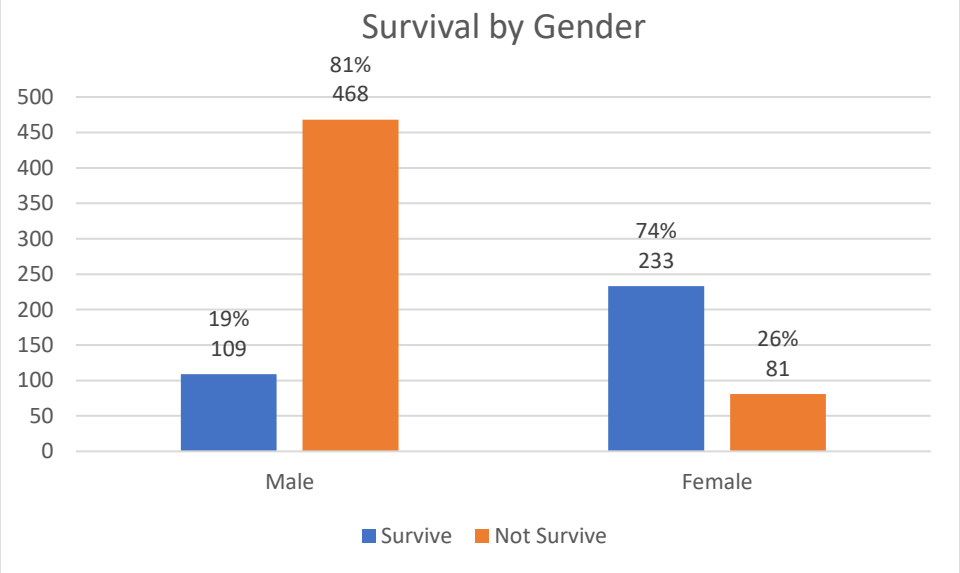


SQLite code:

SELECT count(*) FROM passengers where Pclass = 1 AND Survived=1 (Count survive)

SELECT count(*) FROM passengers where Pclass = 1 AND Survived=0 (Count not survive)

The Pclass was change respectively

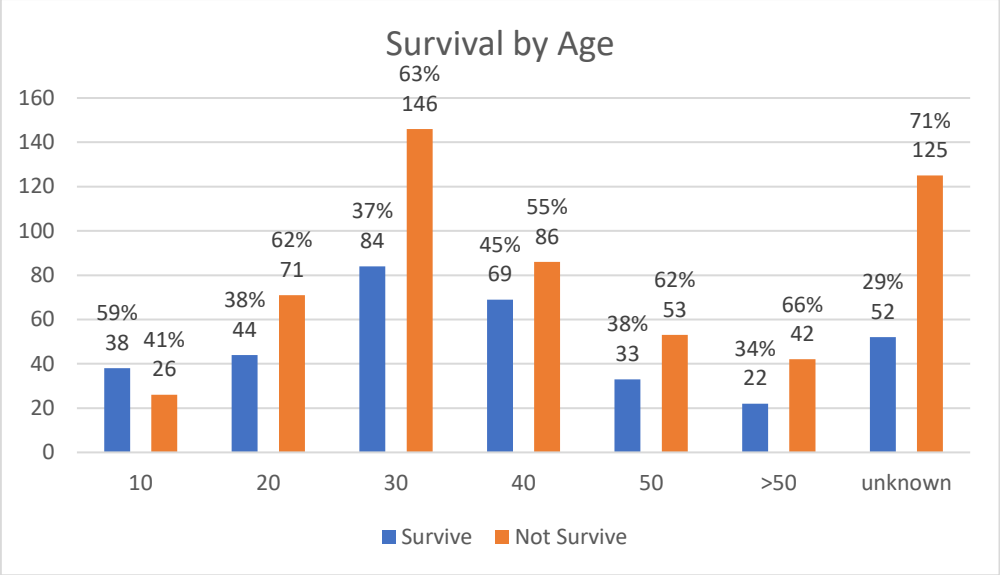


SQLite code:

SELECT count(*) FROM passengers where Sex=”male” AND Survived=1 (Count survive)

SELECT count(*) FROM passengers where Sex=”male” AND Survived=0 (Count not survive)

The Sex was change respectively

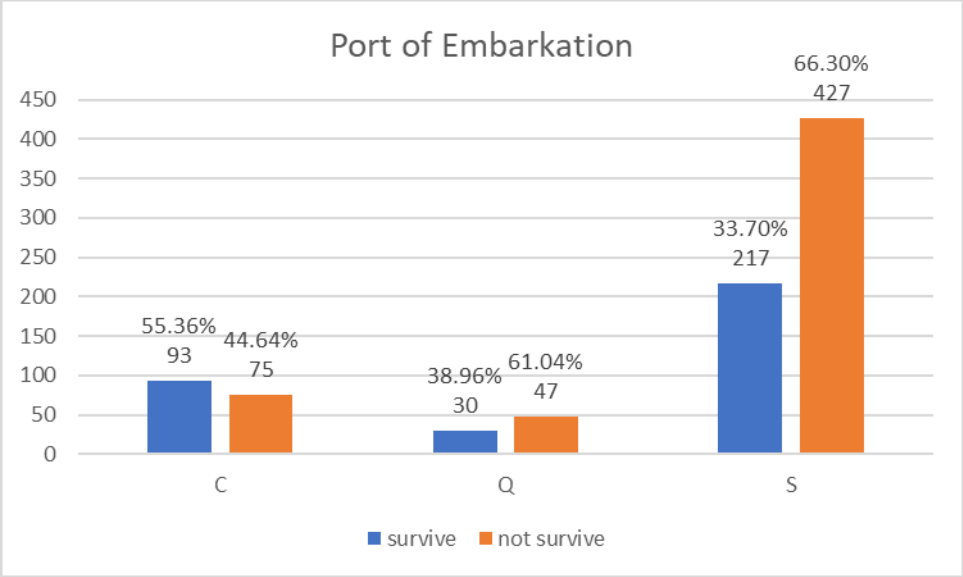


SQLite code:

```
WITH

rangeAge AS

(SELECT Survived,CAST(Age as NUMERIC) as PAge FROM passengers WHERE PAge > 10 AND <=20)

SELECT count(*) FROM rangeAge WHERE Survived=0
```

There was subquery by WITH keyword to make it easier to identify the data. CAST function was use in order to change the data type of the age field which already declare as TEXT to new datatype NUMERIC so that the analysis about the range of age can be perform. Survived value was change to 0 and 1 respectively. The range scale was stated as below

Range	SQL range
0 - 10	PAge <= 10
11 – 20	PAge > 10 AND PAge<= 20
21 - 30	PAge > 20 AND PAge<=30
31 - 40	PAge > 30 AND PAge<=40
41 - 50	PAge > 40 AND PAge<=50
Above 50	PAge > 50
Unknown age	PAge IS NULL



SQLite code:

SELECT count(*) FROM passengers where Embarked = "C" AND Survived=1 (Count survive)

The embarked and survived was change respectively.

Summary

Based on the data analysis, among 891 passengers on the Titanic, most of the passengers was unable to survive which around 62% of the passengers didn't make it. The passenger that have the priority to use the ship facilities was from the first class since the data shows about 63% passengers from the first class was able to survive compare with other class which the percentage was not reach even the halve of the class population. Female passengers also have the advantages which 74% of female population (314 in total) survive the incidence while only 19% of male passengers out of 487 in total was survived. Even though some passenger's age was undetermined, the data shows that the passenger below the age of 10 (59% survived) was highly prioritize followed by passenger in range of 30 years old (45% survived). Some minor aspect that can contributed to passenger survival rate are passenger that embarked from Cherbourg might has higher chance to survive (55% survived) while passenger that embarked from Queenstown and Southampton were only 39% and 33% respectively from their population.

Conclusion

Based on the Titanic dataset that have been analyze, the criteria for passengers to have higher chance to get survive if this tragedy happens again are the passengers must be from the first class, woman and children. This dataset was very useful for people to exploring about the data analytic demonstration and use of computer language to solve some problem.