Project 1 - Exploring Titanic Database

Muhamad Firdaus Jamaludin

UNDERSTANDING THE BUSINESS CONTEXT

- What are these data for?
  - To analyse the survivability of passengers after the Titanic sinking tragedy took place.
- Why do we need this database?
  - This database is needed to know the details information of all passenger on board in order to study their survival factors in the shipwreck.
- Where are these data collected?
  - Titanic - Machine Learning from Disaster | Kaggle

UNDERSTANDING THE TECHNICAL CONTEXT

- How are these data collected?
  - The data was compiled by kaggle website.
- Where are the sources of these data?
  - The sources of this data is from kaggle Titanic - Machine Learning from Disaster competition that split into two part which are training set and test set.
- Is the data coming from surveys, or some computer system? Is it manually input by some data entry personnel or collected by some electronic system?
  - Based on the sources of the data, it is manually input and compiled into the excel file format for the competition purposed.

UNDERSTANDING THE TECHNICAL CONTEXT

- What are the systems that touch or use/modify these data?
  - This data was compiled in excel format hence it can be modify.
- What are some of the error sources of this data?
  - Any source that depend on surviving passengers and crew memories that experienced the event is fallible. In many cases, several witnesses have given different account of the same incident.
- Is the data complete? Would there be missing pieces of data?
  - The data is split into two groups which area training set and test set. We will be using the training set data for this report. However these data might not updated due to it was used for competition purposes.

UNDERSTANDING THE TABLES AND FIELDS

- How many tables do we have?
  - There are only one table from this data which is passengers table.
- What are this tables representing?
  - The table represent the passengers detail informations.
- What are the fields in the tables?
  - The passengers table represent detail information consist of PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

UNDERSTANDING THE TABLES AND FIELDS

- What is the meaning of each of the field?

PassengerId	- Id of passenger for this table
Survived	- Indicate survival (0 = No, 1 = Yes)
Pclass	- Indicate ticket class as proxy for social economy status (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	- Name of passenger
Sex	- Gender
Age	- Age in Year
SibSp	- Sibling or Spouse
Parch	- Parent or Children
Ticket	- Ticket Number
Fare	- Passenger ticket fare
Cabin	- Cabin number
Embarked	- Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

UNDERSTANDING THE TABLES AND FIELDS

- Is the data messy?
  - The data is messy because there is a missing data in the “Age”, “Cabin” and “Embarked” columns. Also, it was compiled randomly and not sort into any field such as alphabetic order, sex, age, socio-economy status and etc.
- How to clean the messy data?
  - Check how many missing data by:
    - select count(*) from table where column_x is null or column_x = 0
  - Finding:
    - Total passenger in the data is 891.
    - “Age” column has 177 missing data.
    - “Cabin” column has 687 missing data.
    - “Embarked” column has 2 missing data.

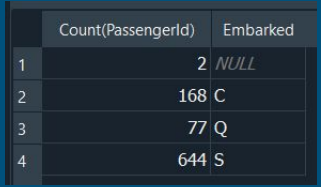
UNDERSTANDING THE TABLES AND FIELDS

Cont.

- Solution:
  - “Age” column has 177 missing data which is approximately 20% from the total data. This variable is important in categorized the passenger age range such as children, adult and elderlies which are significant as the survival factor. Thus, the missing data will be ignored and this column shall not be neglected.

  - “Cabin” column has 687 missing data which is approximately 77% from the total data. The percentage is more than 50% to be considered hence this column shall be neglected as the survival factor.

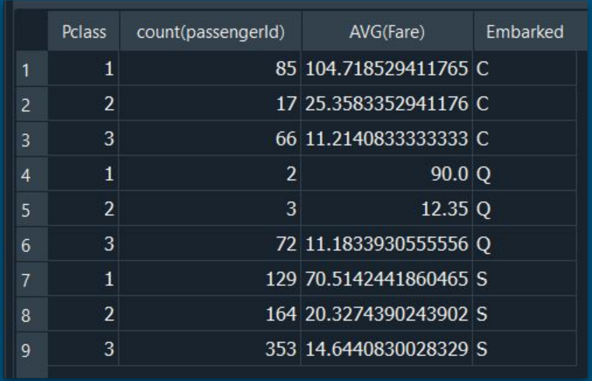
  - “Embarked” column has 2 missing data which is insignificant to be neglected so we will assume the missing data would be embarked on the most populated port which is Port Southampton (“S”) with approximately 72% from the total data by using :
    - SELECT COUNT(PassengerId), Embarked FROM passengers GROUP By Embarked



	Count(PassengerId)	Embarked
1	2	*NULL*
2	168	C
3	77	Q
4	644	S

FREE EXPLORATION

- What is the average fare rate for each passenger class from each port of embarkation?
  - UPDATE passengers SET Embarked = 'S' WHERE Embarked IS NULL
  - SELECT Pclass,count(passengerId), AVG(Fare), Embarked FROM passengers GROUP BY Embarked, Pclass
  - Based on table 1, it shows that the highest average fare is from class 1 that embarked from port C and the lowest average fare is from class 3 that embarked from port Q.

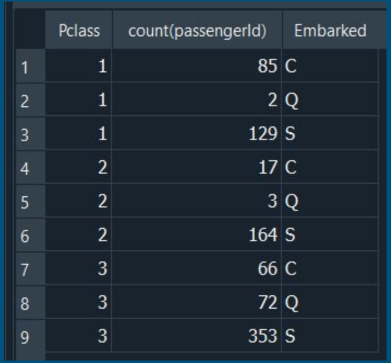


	Pclass	count(passengerId)	AVG(Fare)	Embarked
1	1	85	104.718529411765	C
2	2	17	25.3583352941176	C
3	3	66	11.2140833333333	C
4	1	2	90.0	Q
5	2	3	12.35	Q
6	3	72	11.1833930555556	Q
7	1	129	70.5142441860465	S
8	2	164	20.3274390243902	S
9	3	353	14.6440830028329	S

Table 1 : Average Fare

FREE EXPLORATION

- What is the social economy status (SES) class from each port of embarkation?
    - SELECT Pclass,count(passengerId), Embarked FROM passengers GROUP BY Embarked, Pclass
    - Table 2 indicate that the most passenger was embarked from port S with approximately 70%.
    - Approximately 60%, 39%, and 1% of class 1 embarked from port S, C and Q respectively.
    - Approximately 89%, 9%, and 2% of class 2 embarked from port S, C and Q respectively.
    - Approximately 72%, 15%, and 13% of class 3 embarked from port S, C and Q respectively.

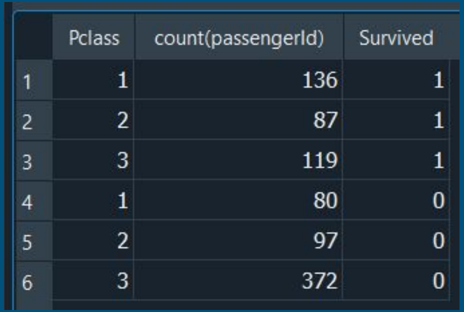


	Pclass	count(passengerId)	Embarked
1	1	85	C
2	1	2	Q
3	1	129	S
4	2	17	C
5	2	3	Q
6	2	164	S
7	3	66	C
8	3	72	Q
9	3	353	S

Table 2: Passenger SES from Embarkation Port

FREE EXPLORATION

- How many percentage of survivor rate from each SES class?
    - SELECT Pclass,count(passengerId), Survived FROM passengers GROUP BY Survived, Pclass ORDER BY Survived DESC
    - Based on table 3, its indicate that approximately 38% of the passenger data was survived while approximately 62% is the victims of the accident.
    - The percentage of survivor from class 1, 2 and 3 respectively are approximately 15%, 10% and 13% from the total passenger data.
    - The most victims come from class 3 that indicate approximately 42%.

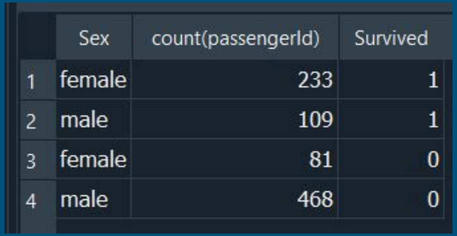


	Pclass	count(passengerId)	Survived
1	1	136	1
2	2	87	1
3	3	119	1
4	1	80	0
5	2	97	0
6	3	372	0

Table 3: SES Survival Rate

FREE EXPLORATION

- How many percentage from each gender that survived the shipwreck?
    - SELECT Sex,count(passengerId), Survived FROM passengers GROUP BY Survived, Sex ORDER BY Survived DESC
    - Table 4 indicated that the survival rate of female is higher compare to male
    - Total of 342 passenger survived the accident while the other 549 passenger is the victims of the shipwreck.
    - 68% of the survivors is female while 32% is male.
    - Unfortunately, 85% of lost passenger is male in this shipwreck compare to female passenger which is only 15%.
    - The survival rate of female is double compare to male passenger.



	Sex	count(passengerId)	Survived
1	female	233	1
2	male	109	1
3	female	81	0
4	male	468	0

Table 4: Gender Survivor Rate

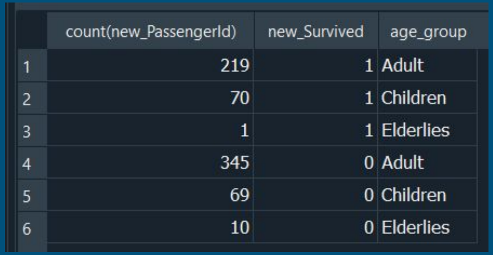
FREE EXPLORATION

- Are children and elderlies have a higher survival rate in this accident?
  - Since the “Age” column is a TEXT datatype and ALTER COLUMN not available in SQLite,
    We have to create a new table and column for new data type. After that insert the data from the old table.
  - Assuming children age less than 19 years old, adult age between 19 to 64 years old and elderlies age between 65 and above

  -
    ```
    CREATE TABLE new_passengers (new_PassengerId INTEGER, new_Survived INTEGER, new_Age NUMERIC)
    ```

  -
    ```
    INSERT INTO new_passengers (new_PassengerId, new_Survived, new_Age)
    SELECT PassengerId, Survived, Age FROM passengers
    ```

  -
    ```
    SELECT count(new_PassengerId), new_Survived,
              CASE
                        WHEN new_Age <= 18 THEN 'Children'
                        WHEN new_Age BETWEEN 18 AND 64 THEN 'Adult'
                        WHEN new_Age >= 65 THEN 'Elderlies'
              END AS age_group FROM new_passengers
    WHERE age_group IS NOT NULL
    GROUP By new_Survived, age_group
    ORDER By new_Survived DESC
    ```



	count(new_PassengerId)	new_Survived	age_group
1	219	1	Adult
2	70	1	Children
3	1	1	Elderlies
4	345	0	Adult
5	69	0	Children
6	10	0	Elderlies

Table 5:Age Group Survival Rate

FREE EXPLORATION

Cont.

- Based on Table 5, the total passenger that survived the shipwreck is approximately 41%. Highest survival rate of the total passenger is 31% which is from adult age group.
- Children that survived from the shipwreck is approximately 10% meanwhile, elderlies that survived is approximately 0.1% from the total passenger data.
- Hence, the survival rate of children and elderlies is low compared to survival rate of adult.