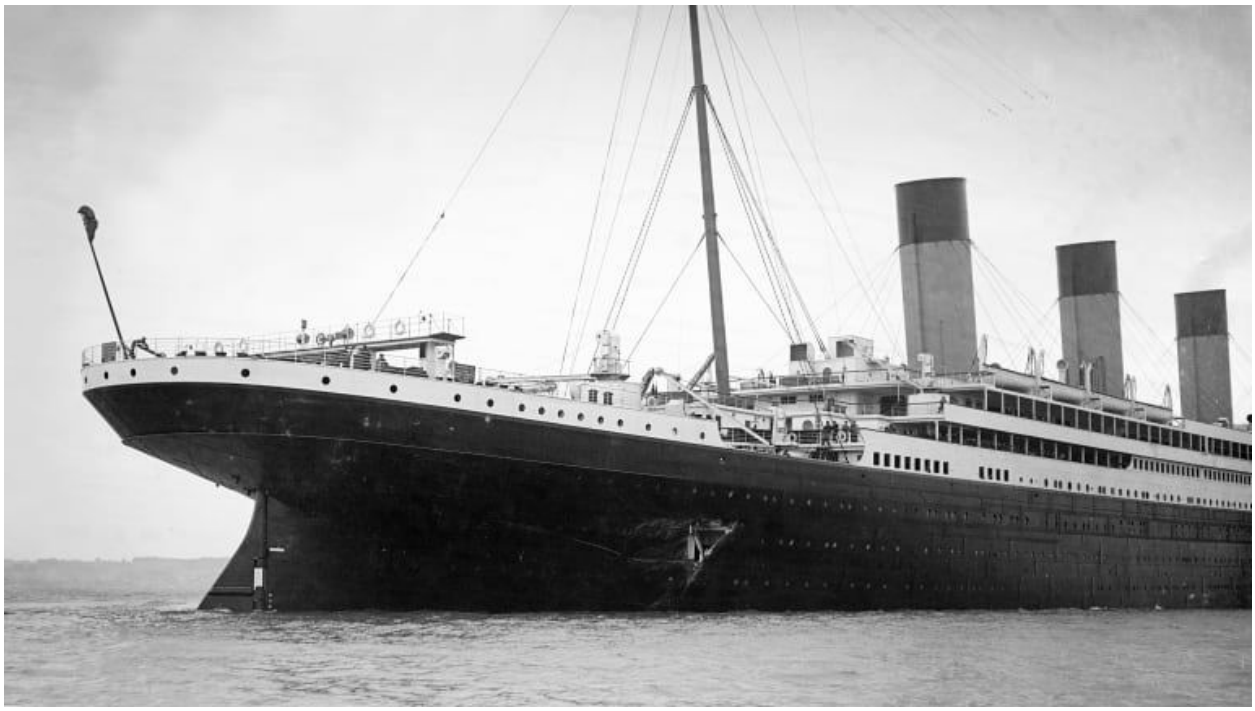
Data Mining

Titanic Dataset Report



Data Cleaning Stage:

1. While analyzing the **Age** column we found 0 values which didn't make sense!

   So we Categoriezed the data by the **SEX** Column and replace the **Male** age from

   (0 -> 31) as the average of the males ages, and did the same thing with **females**

   and replace the 0 values to 28 which is the average.

2. While analyzing the **Fare** column we found 0 values also, so we solved it by categorizing the data by the **Pclass** column and got the average of the fares for each class and replaced it:

   -Class 3 converted to the average which is 13.6755501

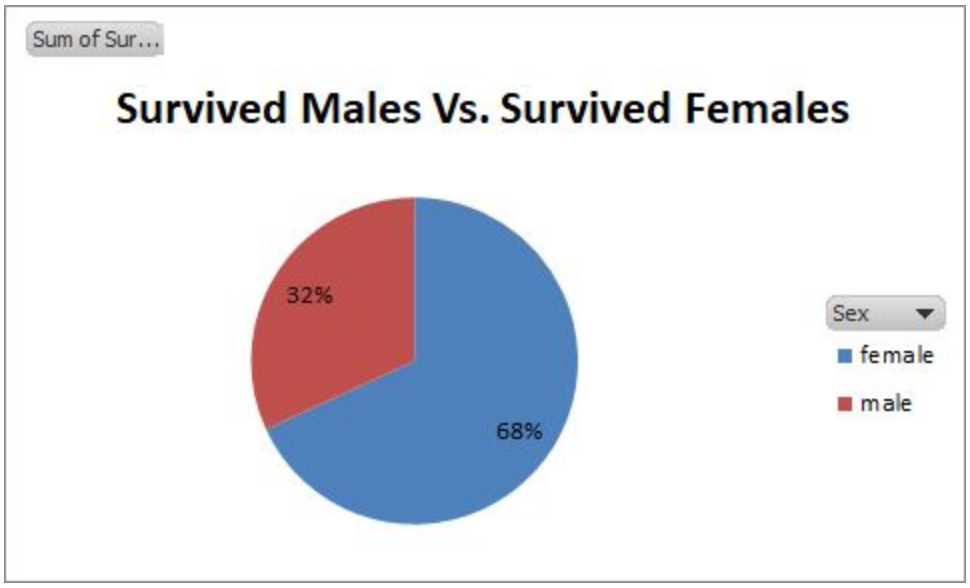
   -Class 2 converted to the average which is 20.66

-Class 1 converted to the average which is 84.1546875

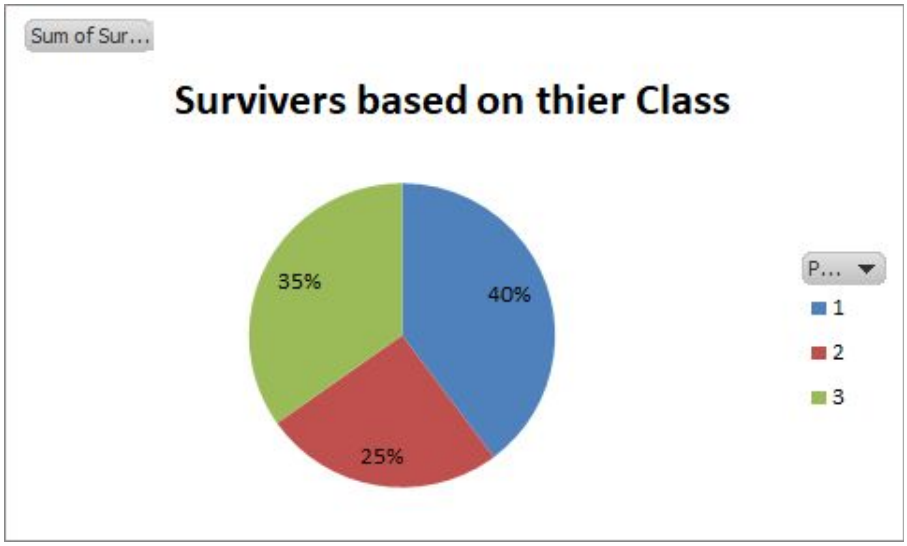
3. During analyzing the **Cabinet** column we found that more than half of the data is missing, so we deleted the entire column also we won't use it further.
4. While analyzing the **Emabrked** column we found two cells with null values and we replaced it with **"S"** value as the most common value for the **Pclass Females**.
5. We deleted the **Name** and the **PassengerID** because they aren't useful for our analysis.

Data Analysis Stage:

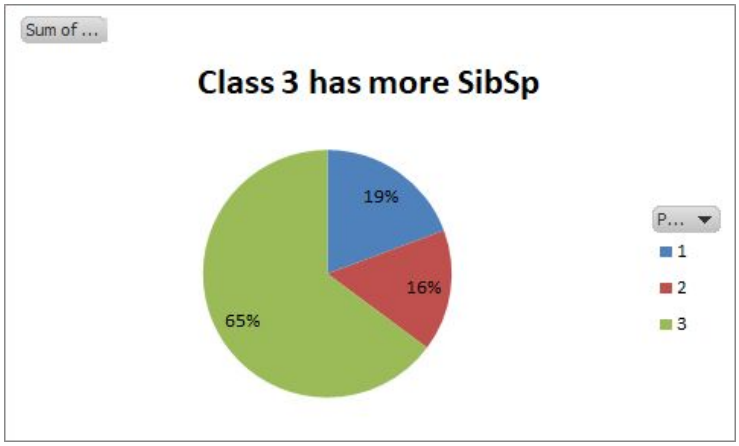
1. The **females** were more lucky than **males**!!

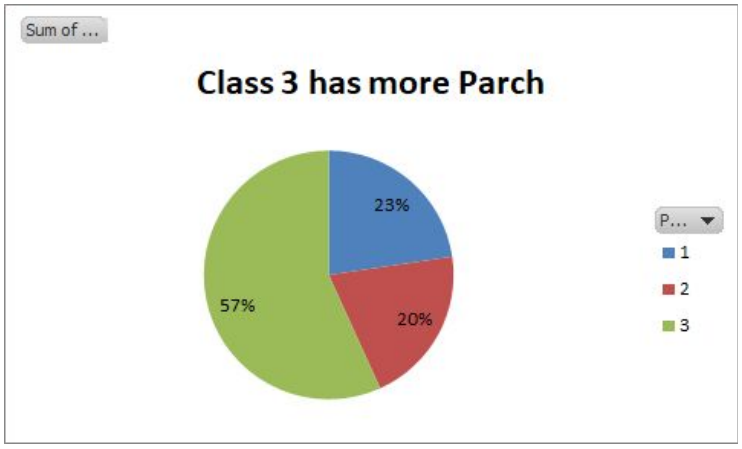


2. The first class of passengers were survived first then the rest of passengers

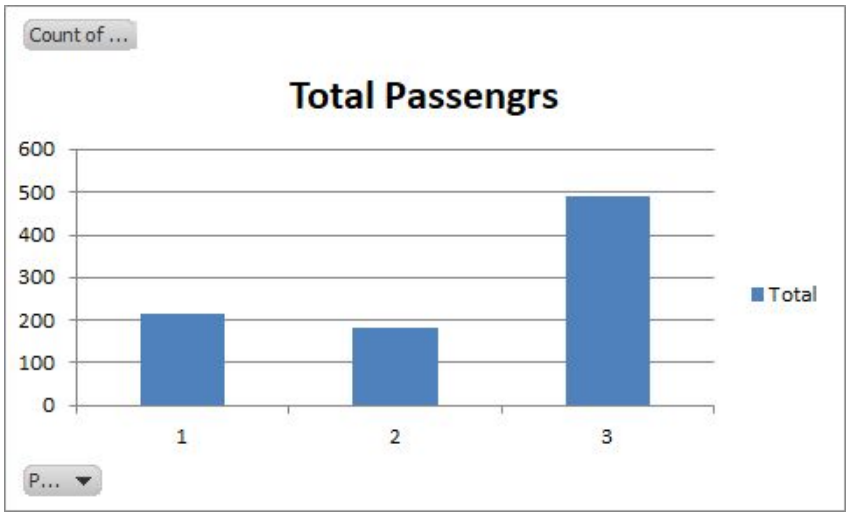


3. Class 3 has more SibSp and Parch

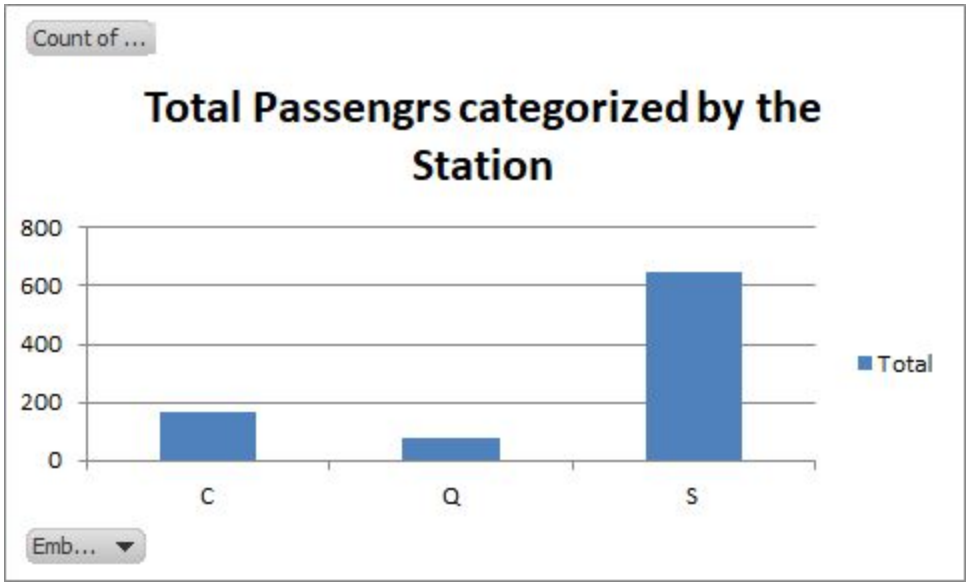




4. From counting the **Ticket** column we found that the most passengers were from the **third class**.



5. The top station(Embarked) was “S”



5. The passengers which had more SibSp and Parch didn't survive, it might be because they sacrificed themselves for their SibSp and Parch.

6. We found that the kids with 5 years survived than the other kids.