Titanic Survival Analysis Report

Using machine learning concepts to analyse the Titanic dataset and predict survival rate.

Motive:

To build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e., name, age, gender, pclass, ticket, fare etc).

Problem understanding and definition:

Predict whether passenger will survive or not.

Model evaluation: Which one is the best model?

K-Fold Cross Validation

K-Fold Cross Validation randomly splits the training data into **K subsets called folds**. We split our data into 4 folds (K = 4). The random forest model would be trained and validated 4 times, using a different fold for validation every time, while it would be trained on the remaining 3 folds.

Using 4 folds (K = 4). Every row represents one training + validation process. In the first row, the model is trained on the second, third and fourth subsets and validated on the first subset. In the second row, the model is trained on the first, third and fourth subsets and validated on the second subset. K-Fold Cross Validation repeats this process until every fold acted once as an evaluation fold.

The result of our K-Fold Cross Validation example would be an array that contains 4 different scores. We then need to compute the mean and the standard deviation for these scores.

Random Forest

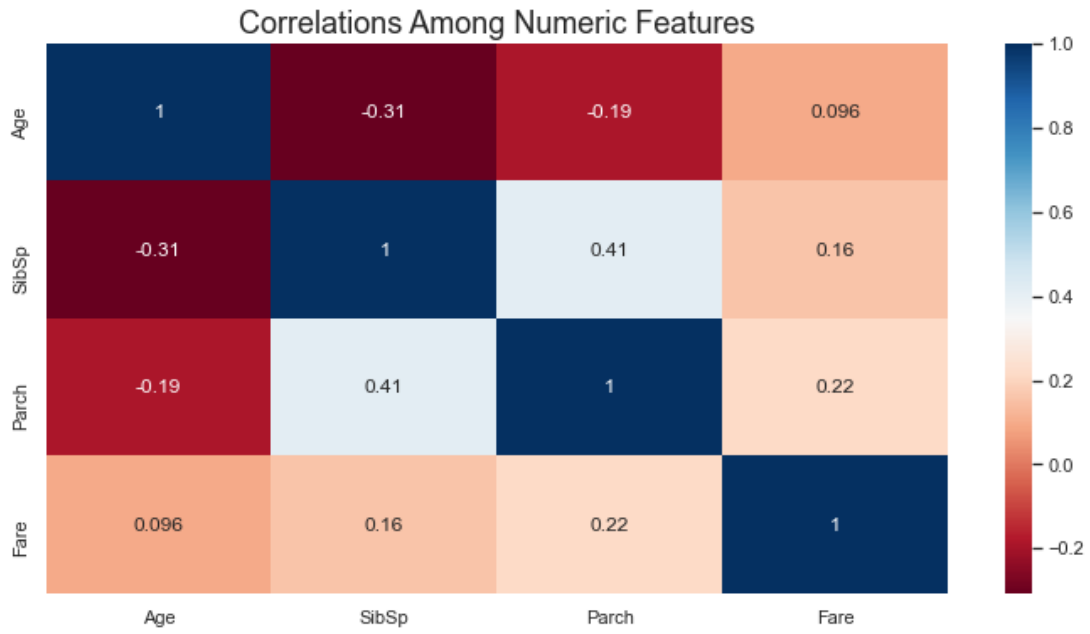
Random Forest is a supervised learning algorithm. It works by building multiple decision trees and merging them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. With a few exceptions a random-forest classifier has all the hyperparameters of a decision-tree classifier and also all the hyperparameters of a bagging classifier, to control the ensemble itself.

Feature importance

Sklearn is able to measure the importance of a features by looking at how much the tree nodes that are used by that particular feature reduce impurity on average across all trees in the forest. It computes this score automatically for each feature after training, and scales the results so that the sum of all importance is equal to 1.

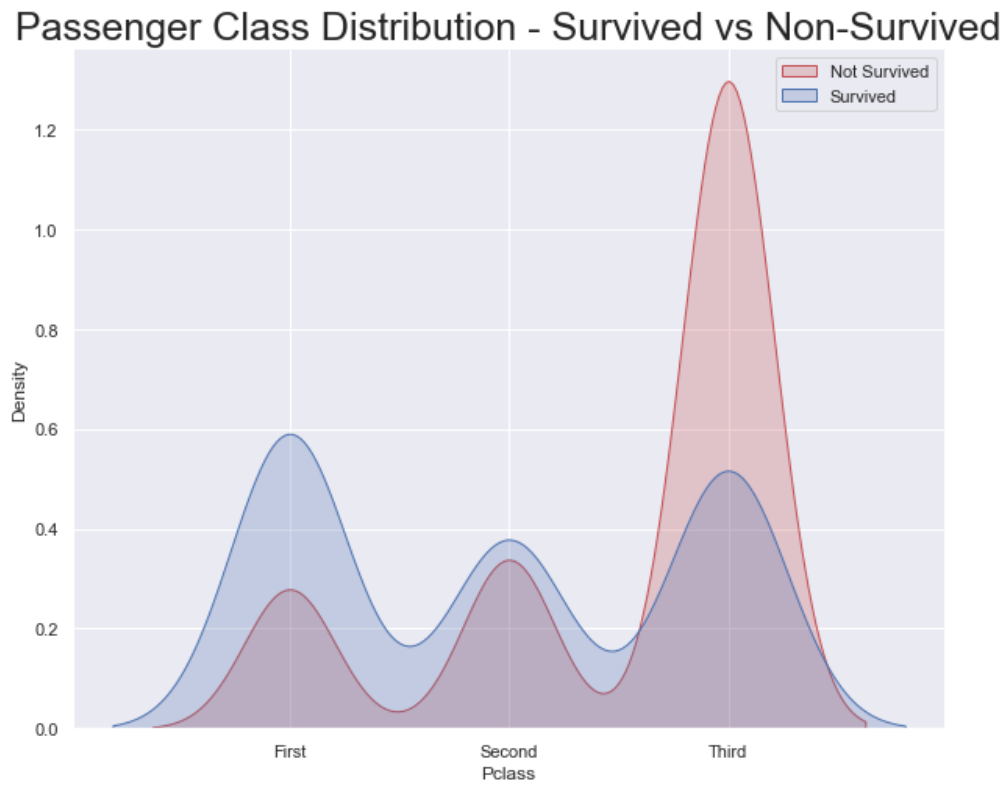
Correlation Matrix and Heatmap



We notice from the heatmap above that:

- Parents and sibling like to travel together
- Age has a high negative correlation with number of siblings

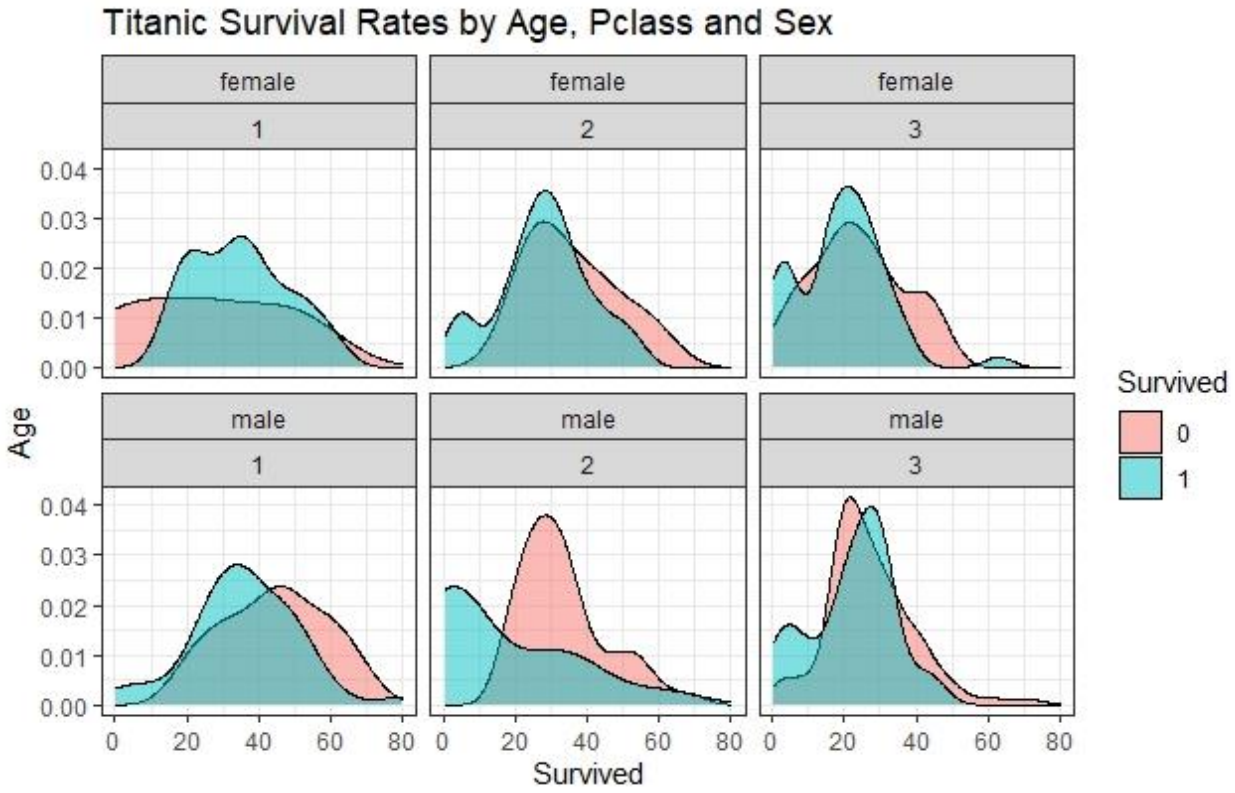
Passenger class distribution; Survived vs Non-Survived



The graphs above clearly shows that economic status (Pclass) played an important role regarding the potential survival of the Titanic passengers. First class passengers had a much higher chance of survival than passengers in the 3rd class. We note that:

- 63% of the 1st class passengers survived the Titanic wreck
- 48% of the 2nd class passengers survived
- Only 24% of the 3rd class passengers survived

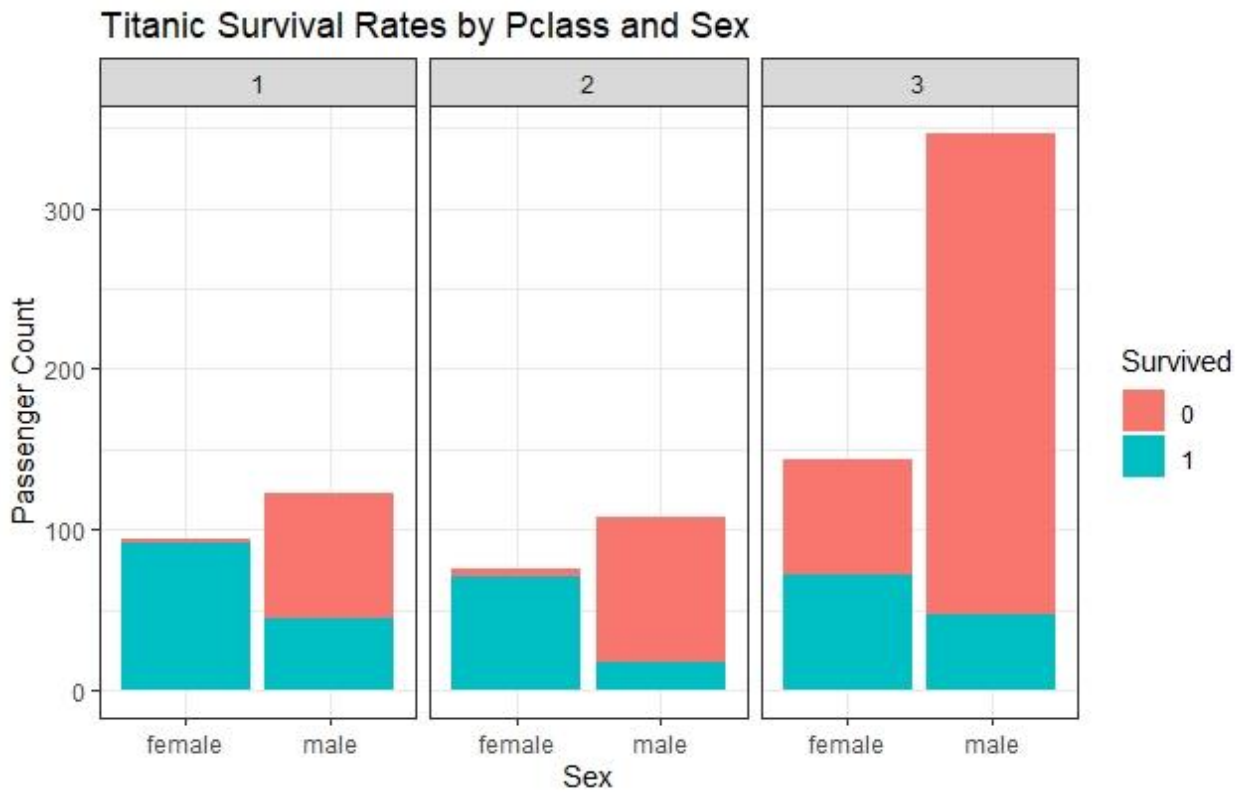
Age and Sex distributions



We can see that men have a higher probability of survival when they are between 18 and 35 years old. For women, the survival chances are higher between 15 and 40 years old.

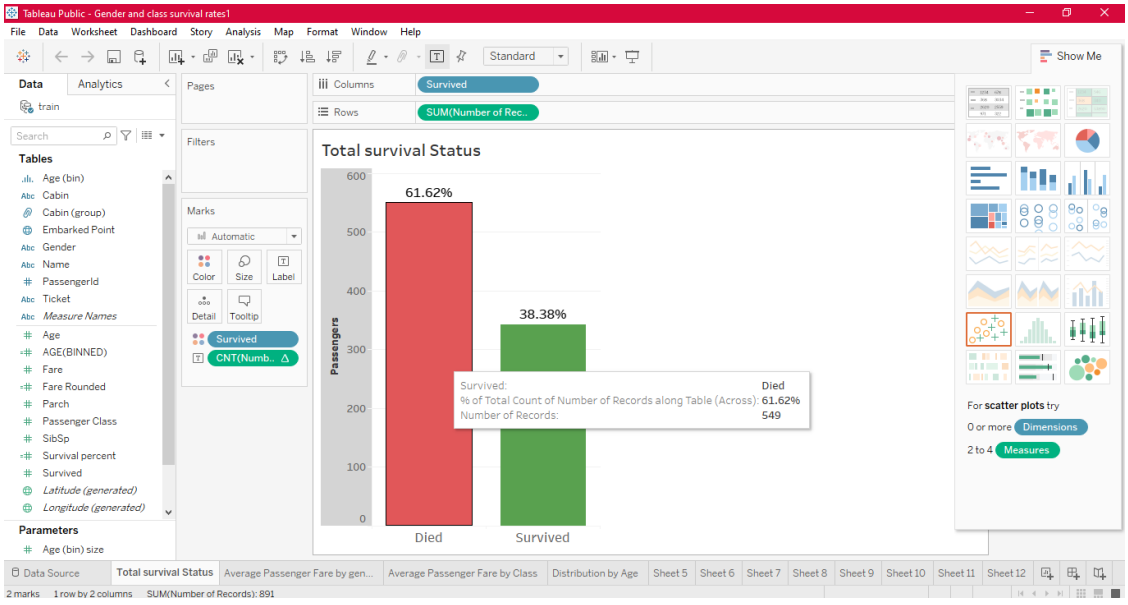
For men the probability of survival is very low between the ages of 5 and 18, and after 35, but that isn't true for women. Another thing to note is that infants have a higher probability of survival.

Distribution of Pclass and Survived

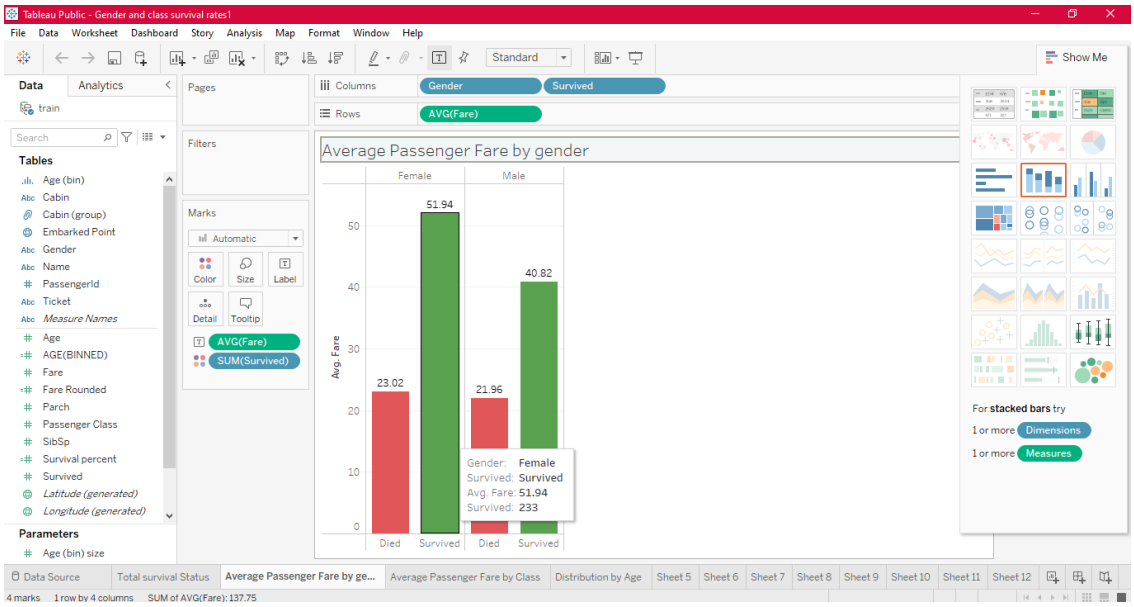


Women are much more likely to survive than men. 74% of the women survived, while only 18% of men survived.

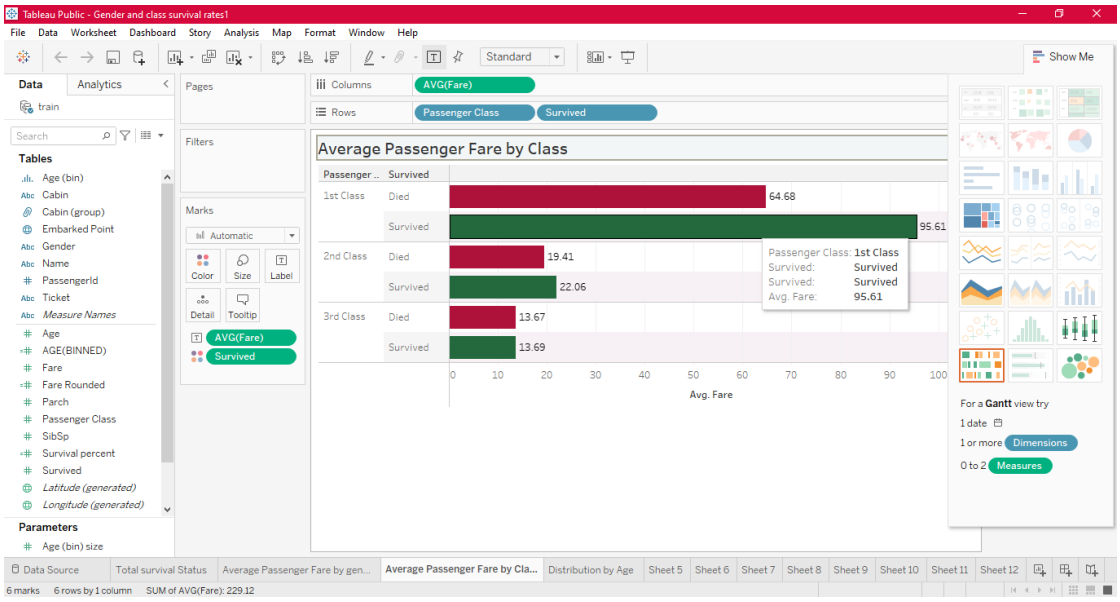
Total Survival Status



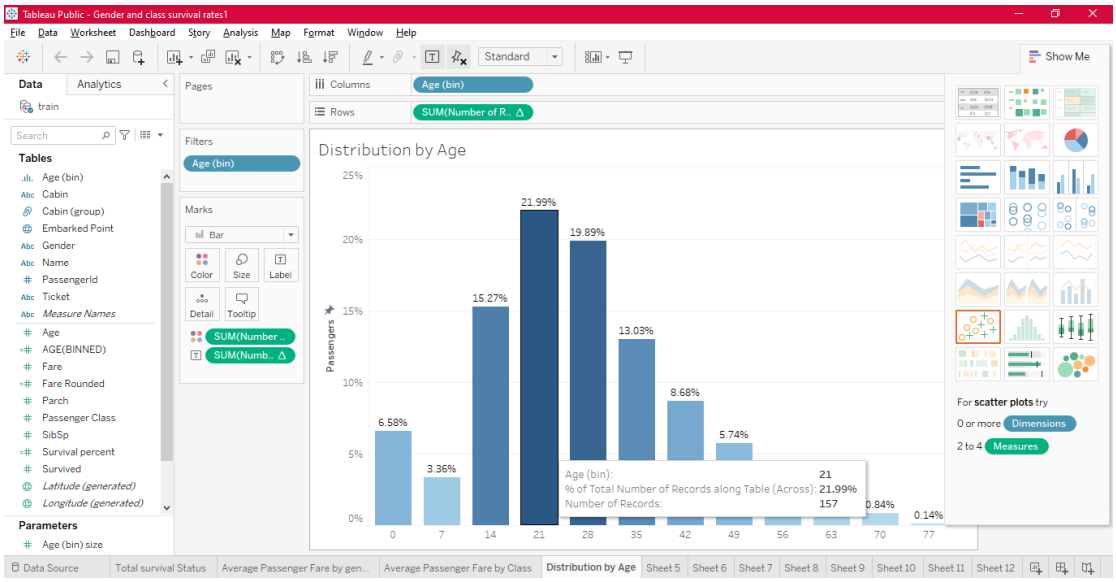
Average Passenger Fare by Gender



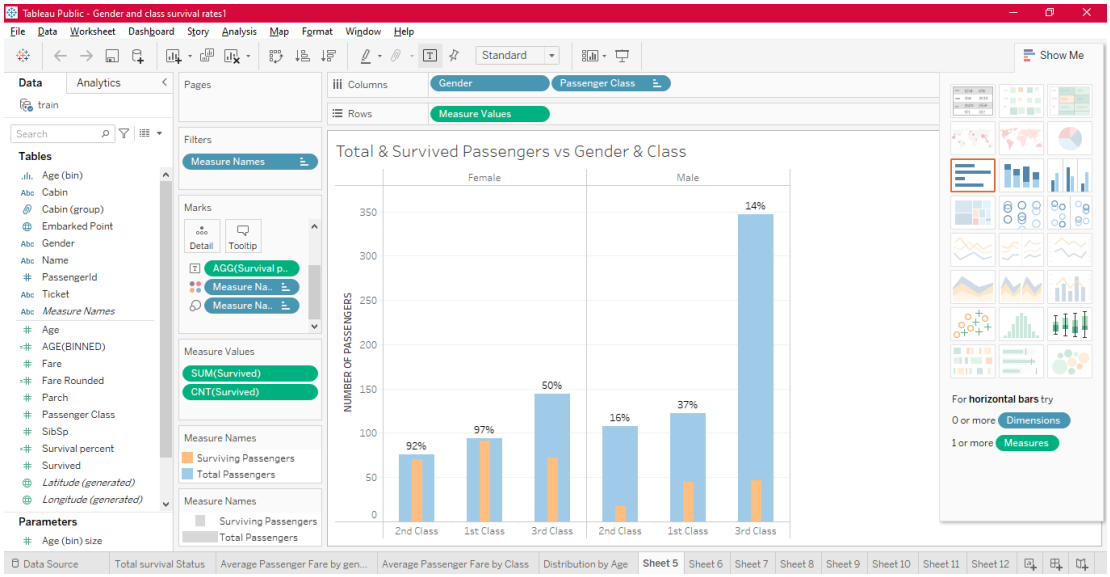
Average Passenger Fare by class



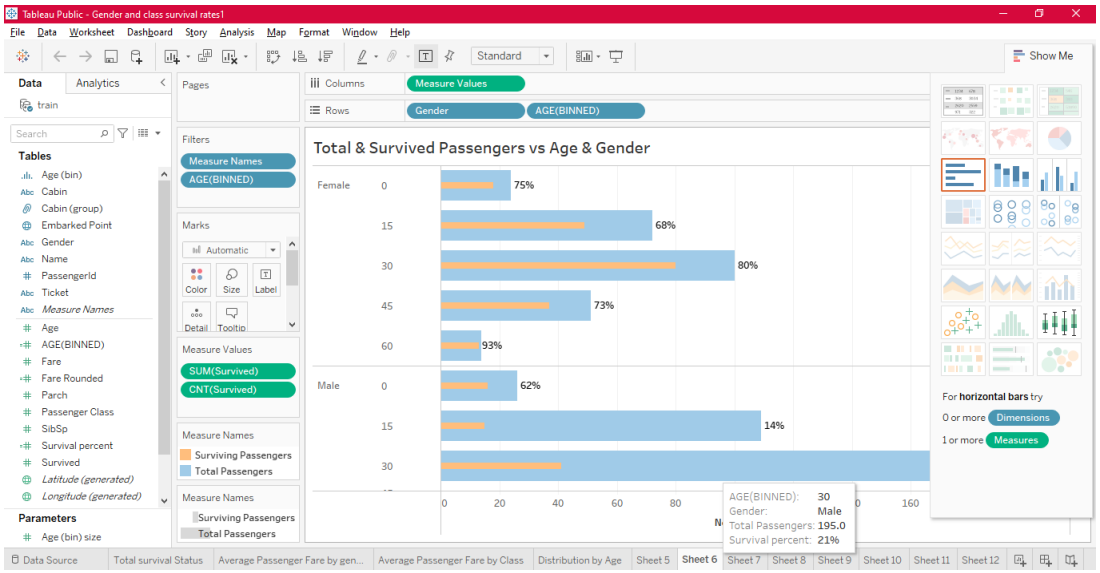
Distribution by Age



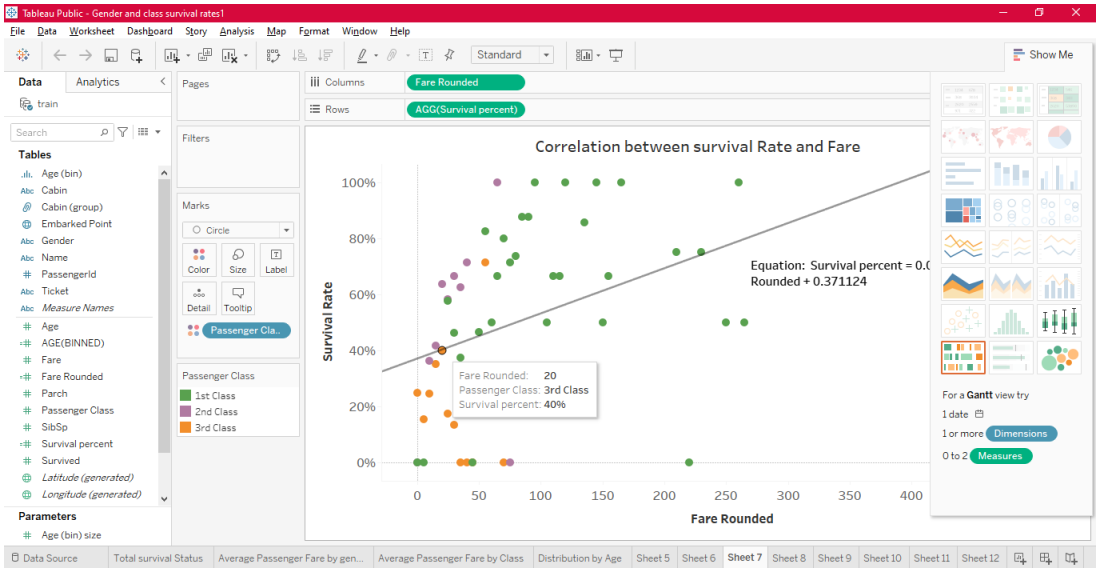
Total vs Survived Passengers, Class and Gender-wise



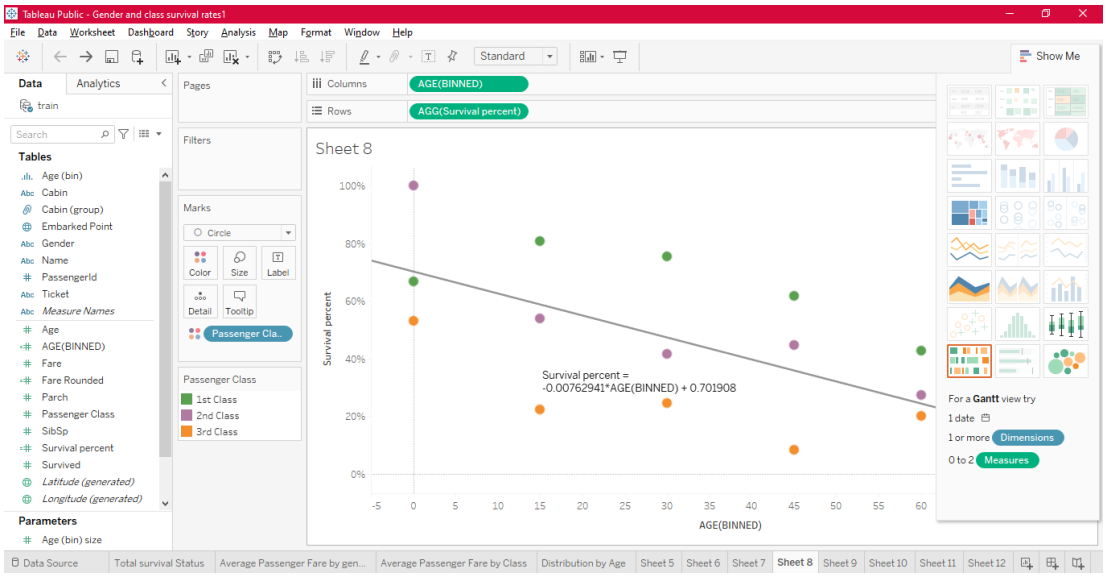
Total vs Survived Passengers, Age and Gender wise



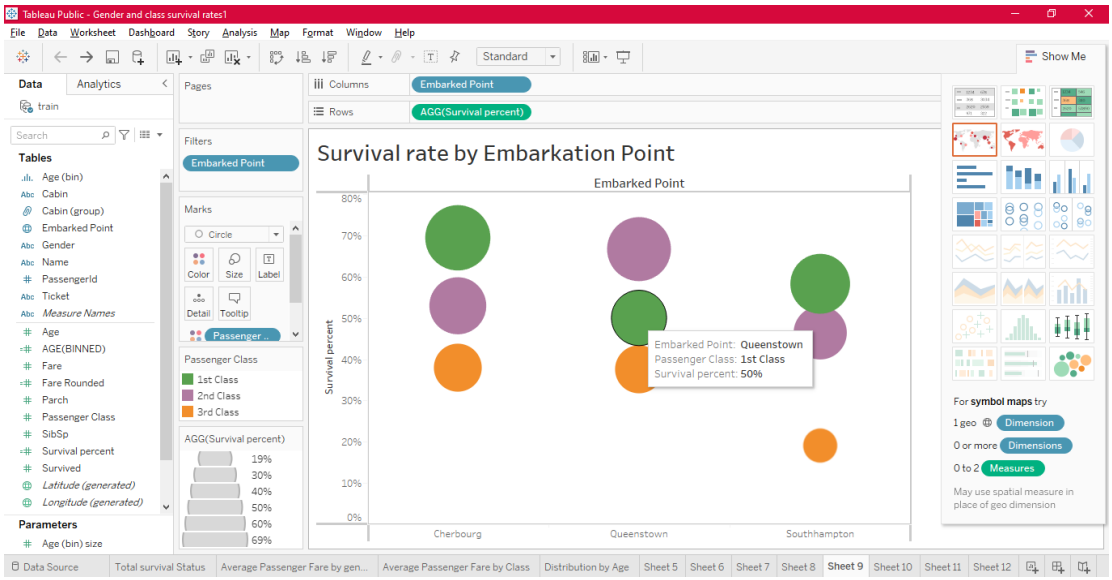
Correlation between survival rate & Fare



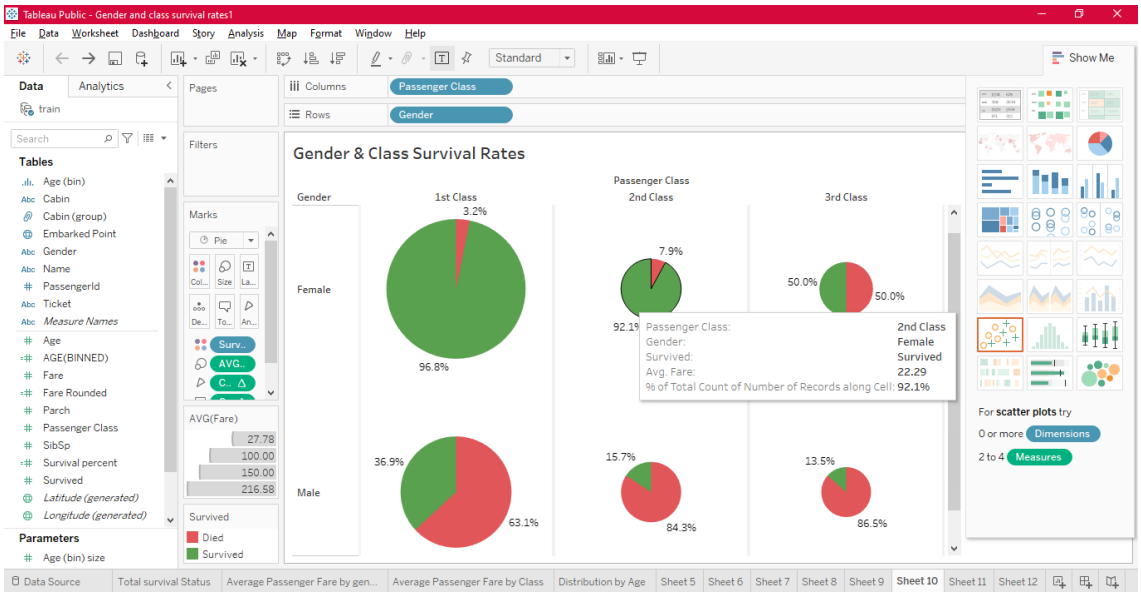
Correlation between survival rate & Age



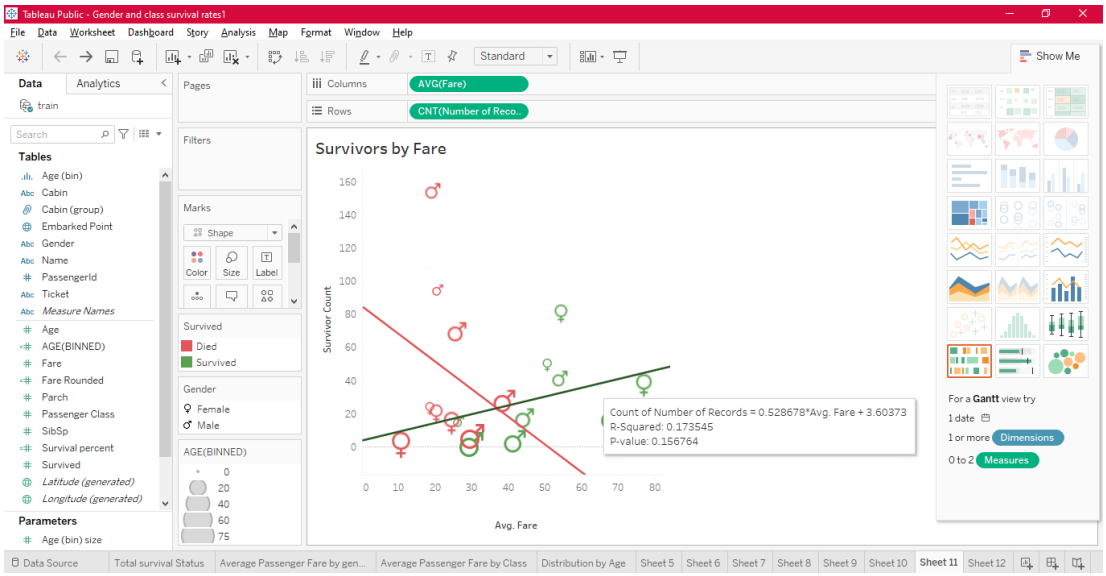
Survival Rate by Point of Embarkation



Survival Rate by Gender & Class



Male and Female Survivors by Average Fare



Cabin wise Survivors