DATA CLEANING & PREPROCCESSING

- The following columns have been dropped: *'boat'*, *'name'*, *'ticket'*, *'cabin'*, *'body'*, *'home.dest'*.
- Missing data in the *'age'* column have been filled-in with the average male passenger age for males and with the average female passenger age for females.
- Missing port embarkation data in the *'embarked'* column have been replaced using backfill method.
- The missing value in the *'fare'* column have been replaced with an average fare cost.
- Data visualization have been performed to investigate possible trends further.
- The *'sex'* column has been encoded to represent males as zero (0) and females as one (1).
- The embarkation port letters in the *'embarked'* column have been encoded as follows: C = 1, Q = 2, S = 3.
- The cleaned dataset has been split into training set and testing set with a test size of 5%.
- The following features have been scaled for training and testing sets independently: *passenger age, number of siblings and spouse, number of children and parents, and cost of fare*.

THE MODEL

Out of 8 trained models, the K-Nearest Neighbor model demonstrated the highest accuracy score of 86.36%. The k number with the highest prediction accuracy has been located between 29 – 33 neighbors (see *"Value of k for KNN Model"* plot). Based on the grid search, KNN model parameter for the *p* value has been set to 1. The training data score for this model resulted in 80.21%. As demonstrated by the confusion matrix plot, the model produced 4.55% of Type I error and 9.09% of Type II error.