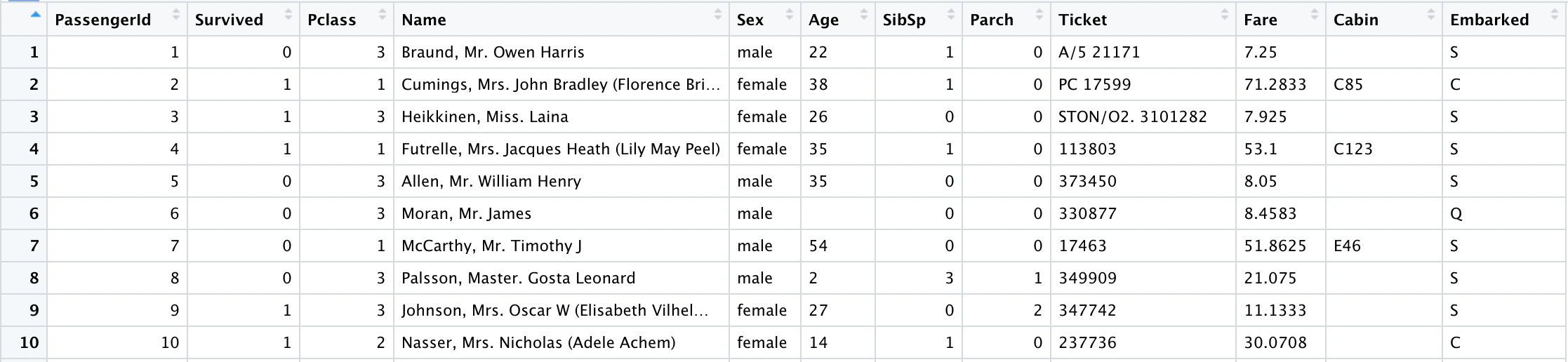
STA 5224: Final Project - Titanic Dataset

# I. EDA:

## 1.1 Overview about the data set:



Train dataset

Titanic dataset includes 11 features as predictors, which helps predict wheather a person would survived through the Titanic disaster.

cat("Number of observations: ", nrow(train.org), "\n ",  
 "Number of duplicated row: ", anyDuplicated(train.org), "\n ")

## Number of observations: 891   
## Number of duplicated row: 0   
##

At first, it seems that there is no missing value in this dataset, but in fact, there are some cells having value of empty string, and containing no information, those are considered as missing value.

Empty.string <- c()  
for (i in colnames(train.org)){  
 Empty.string <- c(Empty.string, sum(train.org[,i] == ""))  
}  
  
mysummary <- data.frame(NA.value = colSums(is.na(train.org)), Empty.string)  
row.names(mysummary) <- colnames(train.org)  
  
kable(mysummary)

NA.value

Empty.string

PassengerId

0

0

Survived

0

0

Pclass

0

0

Name

0

0

Sex

0

0

Age

0

177

SibSp

0

0

Parch

0

0

Ticket

0

0

Fare

0

0

Cabin

0

687

Embarked

0

2