STUDENT NAME: NESTOR ROMERO LEON

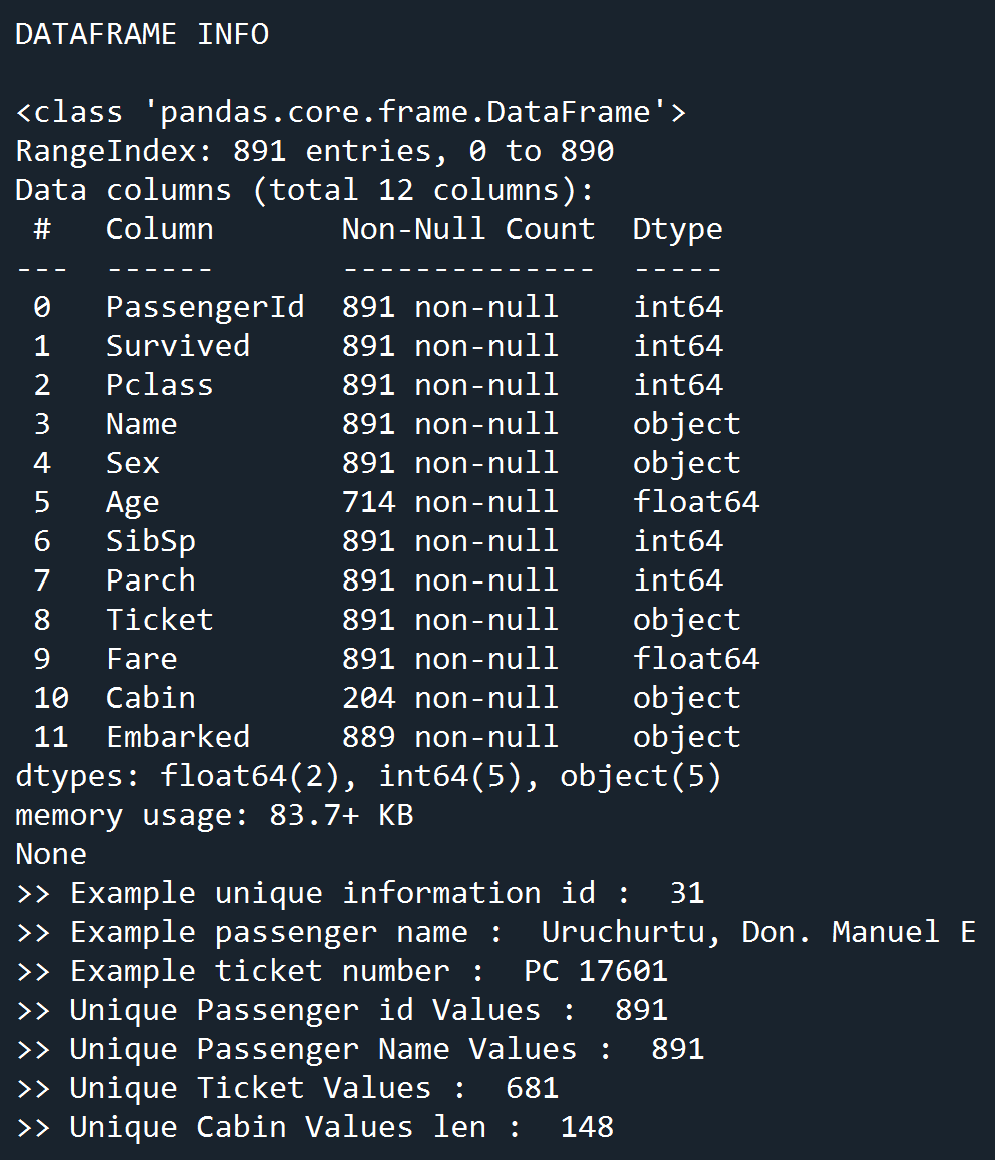
STUDENT ID: 301133331

**INTRODUCTION TO ARTIFICIAL INTELLIGENCE**

**LOGISTIC REGRESSION – ASSIGNMENT**

**EXERCISE 1**

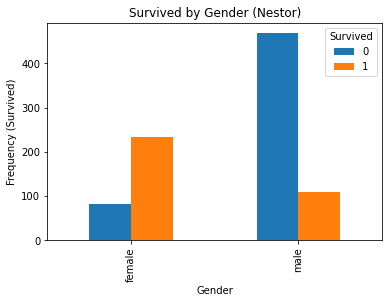
**Part b. Numeral 4 – Columns to Remove**

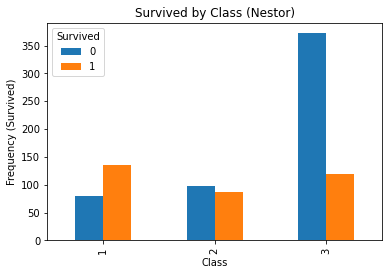


According to the data frame information, the columns to remove will be:

PassengerId, Name, Ticket, and Cabin

The first 3 variables have mostly unique values for each record and the Cabin variable has too many missing values.

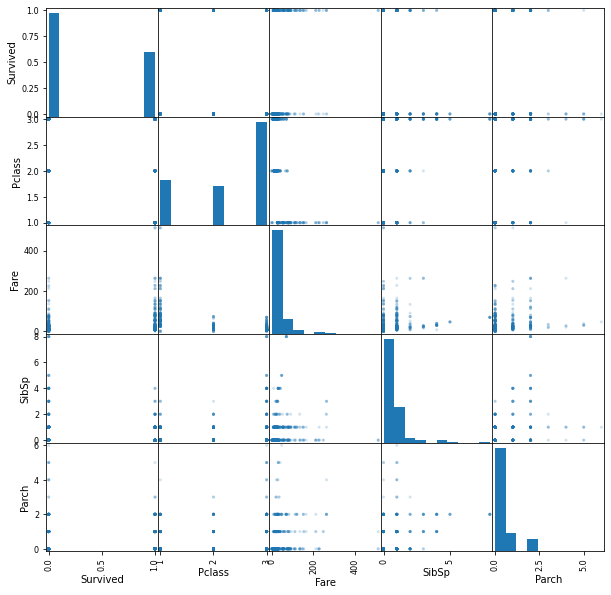
**Part c. Data Visualization**



In the first plot (Survived by Class) the number of survivors is somewhat similar across all classes and around 100. However, the most deaths occurred in the 3rd class in towering numbers compared to classes 1 and 2.

The second plot (Survived by Gender) shows that the number of women that survived was roughly 2.5 times the male survivor count. On the other hand, in terms of deaths, there were nearly 5 times more deaths for men than for women.

**SCATTER MATRIX**

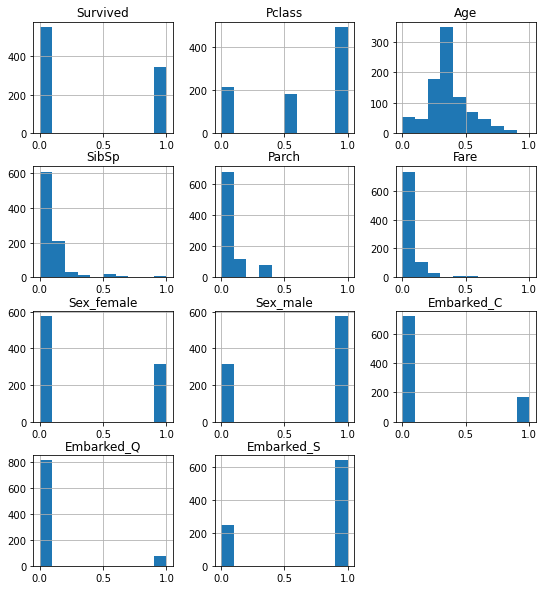


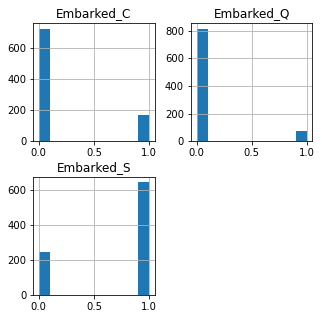
The main diagonal of the matrix that compares each variable vs itself, shows the distribution of the data for each variable. The plots cells comparing variables with discrete values show the data grouping either in columns or rows depending on the comparison being done above or under the diagonal as expected with the pivoted axis. On this particular view, the variables fare, siblings/spouses and parent/children show a similar shape. However, there is not a particularly visible “correlation” of variables in the plot

The plot is helpful in determining not only the approximate shape of the variables but also at detecting ranges for the values and spotting possible outlier values. This plot also helps to raise questions regarding preconceptions regarding the data, for instance, the appearance of very low fare passengers on first class, in fact, all three classes appear to have comparable values when you would expect more clear “buckets” of fares.

**Part c. Data Transformation**

**HISTOGRAMS FOR ALL COLUMNS**



**HISTOGRAMS FOR PORT OF EMBARKMENT VARIABLES**

With the created histograms after the data normalization process, we can conclude that most of the passengers embarked in Southampton