Class Activities -- Pandas

**Activity 1). Selecting Data Using Pandas**

1). What is the average age of all passengers on board?

Confirm that it’s 29.7

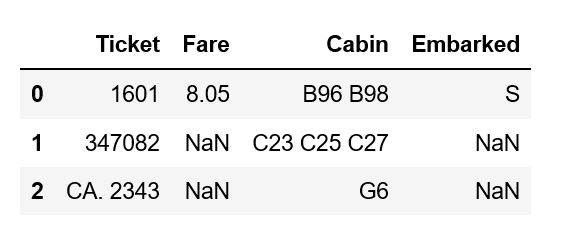
2). In one line, find the median values of the Fare and SibSp columns

Confirm that the answer is 14.45 and 0.00

3). In one line, find the median values of the Fare and SibSp columns *in the first 100 rows*.

Confirm that the answer is 15.675, and 0.00, respectively

4). Using the .iloc command, grab the modal value for the last 4 columns in the dataset. This is the result that you should get:



5). Using the .iloc command, grab the mean value of the first 250 rows of the first 3 columns in the dataset. The answer should be:

* PassengerId: 125.5
* Survived: 0.344
* Pclass: 2.416

**Activity #2). Selecting Data in Pandas using conditions**

* How many females were on board the Titanic? Men?
  + Confirm the answers are 314, 577
* What was the survival rate for females on the Titanic?
  + Confirm the answer is 0.74
* What was the survival rate for males?
  + Confirm the answer is 0.19
* What was the survival rate for passengers in either Pclass1 or 2?
  + Confirm the answer is 0.55
* What was the survival rate if you were female and had at least 1 sibling on board?
  + Confirm the answer is 0.686
* What was the survival rate for females that had no siblings on board?
  + Confirm the answer is 0.79

**Now, answer the following questions applying some of the methods used listed on the bottom of the page.**

Among women, how many unique names were there?

What was the most common age among passengers in class 3?

How many men departed from the ports C, S, and Q, respectively?

Who were the 10 oldest men in passenger class 2?

Can you sort the women in the titanic dataset by their passenger class and age, in descending value?

**Activity #3).** Create a new column called ‘Total\_Family\_Size’ that’s the combination of the SibSP and Parch columns. This column indicates the total amount of company everyone had on board with them.

**Activity #4).** Create the following columns:

* A column called ‘Is\_Alone’, that’s either True or False depending on whether or not Total\_Family\_Size is > 0 or not.
* Create a column called ‘Gender\_Status’ that returns the following values:
  + ‘F-High’ if passenger is female and passenger class is 1.
  + ‘F-Low’ if passenger is female and passenger class is 2 or 3.
  + ‘M-High’ if passenger is male and passenger class is 1.
  + ‘M-Low’ if passenger is male and passenger class is 2 or 3.
* To verify you did your selections correctly, value counts for each category are as follows:
  + M-Low: 455
  + F-Low: 220
  + M-High: 122
  + F-High: 94

## Popular Methods

You don’t have to memorize all of these, but here are frequently used methods in pandas to keep in handy when viewing and organizing your data:

* describe() – returns descriptive stats
* info() – pulls info about data type, null values, memory usage
* dtypes – pulls info about data types for a dataframe
* isnull() – can check if a column has null values or not
  + How could you check the total value of null values in your dataset?
* duplicated() – checks to see if you have duplicate values in your data set
  + How could you check the total value of duplicate rows in your dataset?
* fillna() – fills in empty values in your dataset
* nlargest/nsmallest – pull a certain number of the smallest and largest values of a column
* sort\_values – sort values by a particular column
* groupby() – Creates a grouping, similar to a pivot table in Excel or a GROUPBY operation in SQL

nunique()/unique() – Get the number of unique values in a column

* value\_counts() – Get the number of times a values appears in a column