ANA-522-OL1 Spring 2022
Mod03 Week06 Lab: Data Munging
Due: Friday February 18th at midnight

Titanic Dataset

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. With the availabilities of data analytic and machine learning technologies these days, people are researching in finding clues of survival from hidden data patterns.

Our purpose in this Lab exercise is to transform the dataset with proper multiple indexing and reshaping so as to easily slice information from categorical perspectives.

The Titanic dataset is readily available to be loaded on ANA522 JupyterLab with the filename: "/home/ANA522/Titanic.csv"

[ ]:
```python
import numpy as np
import pandas as pd
```

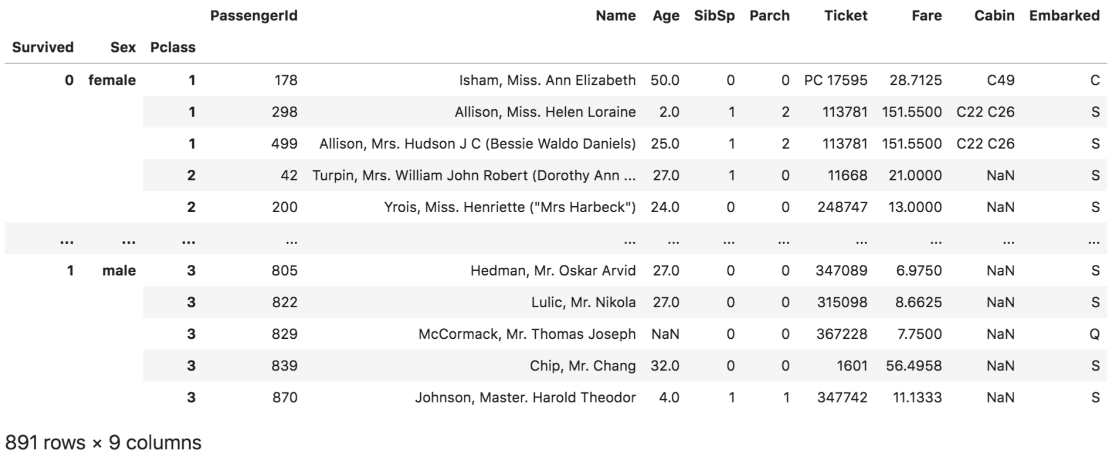
[ ]:
```python
## DO NOT OVERWRITE "titanic_src" throughout the Notebook.

titanic_src = pd.read_csv('/home/ANA522/Titanic.csv', sep=',')
titanic_src.shape
```

[ ]:
```python
## Example of using set_index() to transform the Titanic dataset with␣
 ↪hierarchical indexing
## Test setting multiple index using Survived and Sex columns and save in␣
 ↪TDS_Pclass_Sex.
## Sort TDS_Pclass_Sex by row indices, then save in TDS_Pclass_Sex_Sorted

TDS_Pclass_Sex = titanic_src.set_index(['Survived','Sex'])
TDS_Pclass_Sex_Sorted = TDS_Pclass_Sex.sort_index()
TDS_Pclass_Sex_Sorted
```

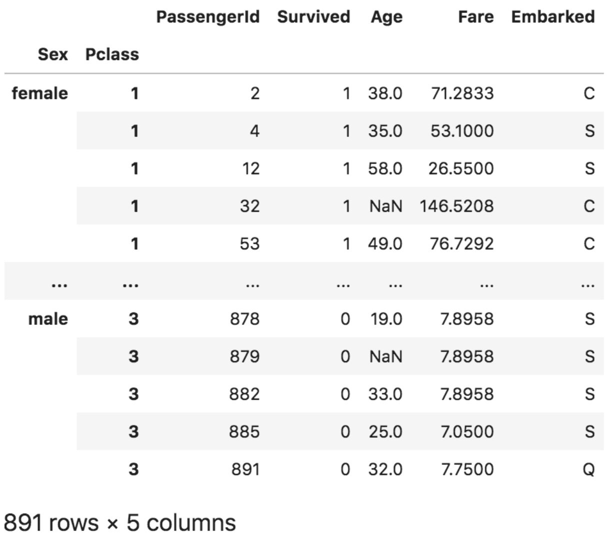
1 Q01: Transform the Titanic dataset with hierarchical indexing by the multiple index in the order of Survived, Sex, and Pclass attributes.



			PassengerId	Name	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Survived	**Sex**	**Pclass**									
0	**female**	**1**	178	Isham, Miss. Ann Elizabeth	50.0	0	0	PC 17595	28.7125	C49	C
		1	298	Allison, Miss. Helen Loraine	2.0	1	2	113781	151.5500	C22 C26	S
		1	499	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	25.0	1	2	113781	151.5500	C22 C26	S
		2	42	Turpin, Mrs. William John Robert (Dorothy Ann ...	27.0	1	0	11668	21.0000	NaN	S
		2	200	Yrois, Miss. Henriette ("Mrs Harbeck")	24.0	0	0	248747	13.0000	NaN	S
...	**...**	**...**	...	...	...	...	...	...	...	...	...
1	**male**	**3**	805	Hedman, Mr. Oskar Arvid	27.0	0	0	347089	6.9750	NaN	S
		3	822	Lulic, Mr. Nikola	27.0	0	0	315098	8.6625	NaN	S
		3	829	McCormack, Mr. Thomas Joseph	NaN	0	0	367228	7.7500	NaN	Q
		3	839	Chip, Mr. Chang	32.0	0	0	1601	56.4958	NaN	S
		3	870	Johnson, Master. Harold Theodor	4.0	1	1	347742	11.1333	NaN	S

891 rows × 9 columns

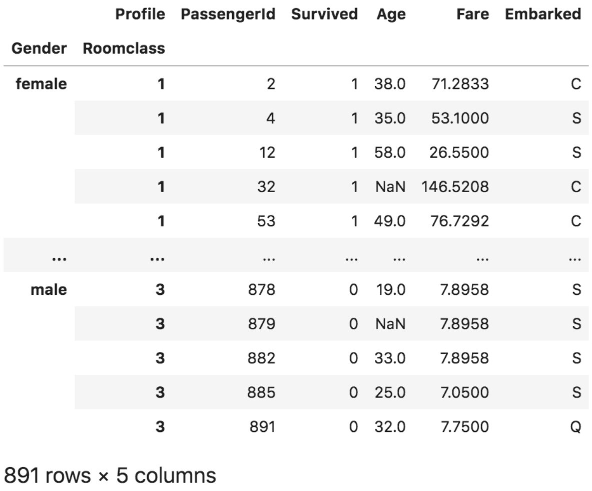
2 Q02: Transform the Titanic dataset with hierarchical indexing by the multiple index in the order of Sex and Pclass columns with only PassengerId, Survived, Age, Fare, and Embarked attributes included. Sort by row indices and name the new dataset as TDS_Sex_Pclass_Sorted to be used in the following questions.



		PassengerId	Survived	Age	Fare	Embarked
Sex	**Pclass**					
female	**1**	2	1	38.0	71.2833	C
	1	4	1	35.0	53.1000	S
	1	12	1	58.0	26.5500	S
	1	32	1	NaN	146.5208	C
	1	53	1	49.0	76.7292	C
...	**...**	...	...	...	...	...
male	**3**	878	0	19.0	7.8958	S
	3	879	0	NaN	7.8958	S
	3	882	0	33.0	7.8958	S
	3	885	0	25.0	7.0500	S
	3	891	0	32.0	7.7500	Q

891 rows × 5 columns

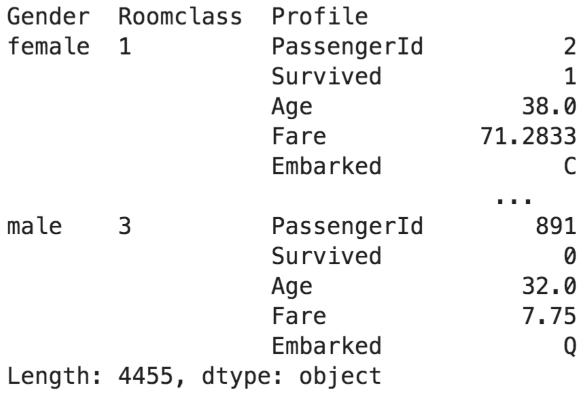
3 Q03: Follow up from Q02. Setup names for index and columns of the multi-index frame for TDS_Sex_Pclass_Sorted. Use “Gender” and “Roomclass” for index names respectively. Use “Profile” for columns’ name.



	Profile	PassengerId	Survived	Age	Fare	Embarked
Gender	Roomclass					
female	1	2	1	38.0	71.2833	C
	1	4	1	35.0	53.1000	S
	1	12	1	58.0	26.5500	S
	1	32	1	NaN	146.5208	C
	1	53	1	49.0	76.7292	C
...	...	...	...	...	...	...
male	3	878	0	19.0	7.8958	S
	3	879	0	NaN	7.8958	S
	3	882	0	33.0	7.8958	S
	3	885	0	25.0	7.0500	S
	3	891	0	32.0	7.7500	Q

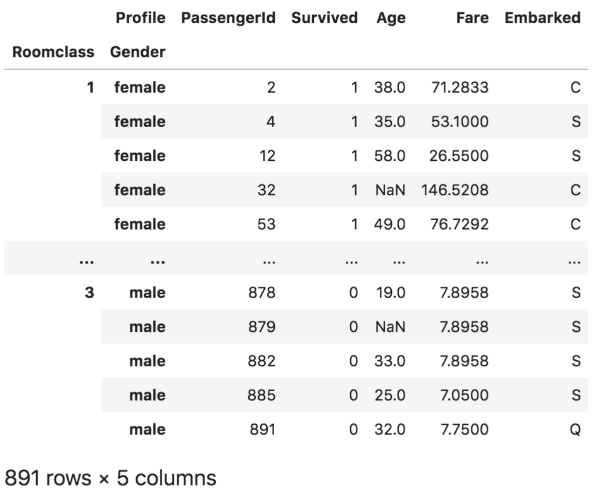
891 rows × 5 columns

4 Q04: Create a Series of data for PassengerId, Survived, Age, Fare, and Embarked for every multi-index combination of Gender(Sex) and Roomclass(Pclass) in TDS_Sex_Pclass_Sorted



```
Gender  Roomclass  Profile
female  1          PassengerId          2
                   Survived             1
                   Age               38.0
                   Fare           71.2833
                   Embarked             C
                                     ...
male    3          PassengerId        891
                   Survived             0
                   Age               32.0
                   Fare              7.75
                   Embarked             Q
Length: 4455, dtype: object
```

5 Q05: Adjust the hierarchical indexing so that the Roomclass is at the first level followed by Gender from TDS_Sex_Pclass_Sorted



	Profile	PassengerId	Survived	Age	Fare	Embarked
Roomclass	Gender					
1	female	2	1	38.0	71.2833	C
	female	4	1	35.0	53.1000	S
	female	12	1	58.0	26.5500	S
	female	32	1	NaN	146.5208	C
	female	53	1	49.0	76.7292	C
...	...	...	...	...	...	...
3	male	878	0	19.0	7.8958	S
	male	879	0	NaN	7.8958	S
	male	882	0	33.0	7.8958	S
	male	885	0	25.0	7.0500	S
	male	891	0	32.0	7.7500	Q

891 rows × 5 columns

6 Q06: Compute total Fare of all passengers by Roomclass using TDS_Sex_Pclass_Sorted

```
Roomclass
1    18177.4125
2     3801.8417
3     6714.6951
Name: Fare, dtype: float64
```

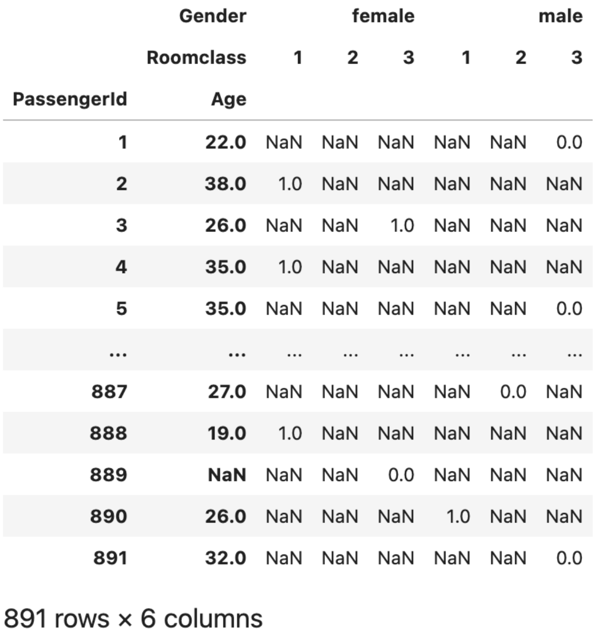
7 Q07 Compute average Fare of all passengers by female and male using TDS_Sex_Pclass_Sorted

```
Gender
female    44.479818
male      25.523893
Name: Fare, dtype: float64
```

8 Q08 Compute average Fare of all passengers from TDS_Sex_Pclass_Sorted up to 2 decimal points.

```
The average Fare for all passengers: 32.20
```

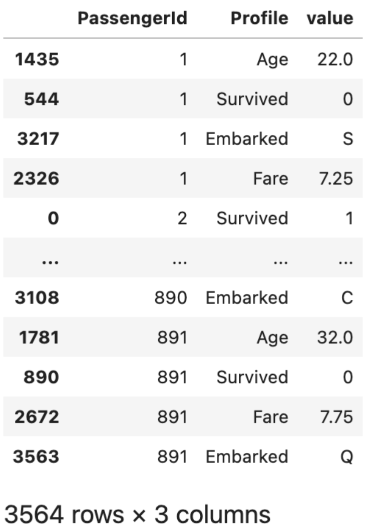
9 Q09 Transform TDS_Sex_Pclass_Sorted so that PassengerId and Age are multiple indices while Gender and Roomclass are columns. Display the result with Survived values only.



	Gender	female			male		
	Roomclass	1	2	3	1	2	3
PassengerId	Age						
1	22.0	NaN	NaN	NaN	NaN	NaN	0.0
2	38.0	1.0	NaN	NaN	NaN	NaN	NaN
3	26.0	NaN	NaN	1.0	NaN	NaN	NaN
4	35.0	1.0	NaN	NaN	NaN	NaN	NaN
5	35.0	NaN	NaN	NaN	NaN	NaN	0.0
...	...	...	...	...	...	...	...
887	27.0	NaN	NaN	NaN	NaN	0.0	NaN
888	19.0	1.0	NaN	NaN	NaN	NaN	NaN
889	NaN	NaN	NaN	0.0	NaN	NaN	NaN
890	26.0	NaN	NaN	NaN	1.0	NaN	NaN
891	32.0	NaN	NaN	NaN	NaN	NaN	0.0

891 rows × 6 columns

10 Q10 Create a dataframe from TDS_Sex_Pclass_Sorted that uses PassengerId as the key index to make all tuples of attribute and value to be data entries in the table.



	PassengerId	Profile	value
1435	1	Age	22.0
544	1	Survived	0
3217	1	Embarked	S
2326	1	Fare	7.25
0	2	Survived	1
...	...	...	...
3108	890	Embarked	C
1781	891	Age	32.0
890	891	Survived	0
2672	891	Fare	7.75
3563	891	Embarked	Q

3564 rows × 3 columns