ANA-522-OL1 Spring 2022

Mod03 Week05 Lab: Data Preparation

Due: Friday February 11th 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 the Lab exercise is to prepare and clean the dataset to be handy for data analysis tasks.

The Titanic dataset is readily available to be loaded on ANA522 JupyterLab with the filename:

"/home/ANA522/Titanic.csv"

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
### Loading the Titanic Dataset
titanic = pd.read_csv('/home/ANA522/Titanic.csv', sep=',')
```

Q01: Use a pandas function to overview the Titanic dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Q02: Find a pandas attribute to display data type of each column (all pokemon dataset properties.)

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

Q03: Display all entries whose Age attribute value is missing (NA).

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	NaN	C
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	NaN	C
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792	NaN	Q
...	...	...	...	...	...	...	...	...	...	...	...	...
859	860	0	3	Razi, Mr. Raihed	male	NaN	0	0	2629	7.2292	NaN	C
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
868	869	0	3	van Melkebeke, Mr. Philemon	male	NaN	0	0	345777	9.5000	NaN	S
878	879	0	3	Laleff, Mr. Kristo	male	NaN	0	0	349217	7.8958	NaN	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S

177 rows × 12 columns

Q04: Show how many entries whose Embarked attribute value is missing (NA).

There are 2 entries whose Embarked values is NA

Q05: Display all entries whose Age attribute value or Cabin attribute value is missing (NA).

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

706 rows × 12 columns

Q06: List the indices of all entries whose Age attribute value or Cabin attribute value is missing (NA).

```
Int64Index([  0,   2,   4,   5,   7,   8,   9,  12,  13,  14,
            ...
            878, 880, 881, 882, 883, 884, 885, 886, 888, 890],
           dtype='int64', length=706)
```

Q07: Show how many entries are there both Age and Cabin attribute values are missing (NA).

There are 158 entries with NA in both Age and Cabin.

Q08: Sample any 15 entries to be sorted by PassengerId that both Age and Cabin attributes are missing (NA).

The following is a sample of 15 entries sorted by PassengerId.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
36	37	1	3	Mamee, Mr. Hanna	male	NaN	0	0	2677	7.2292	NaN	C
48	49	0	3	Samaan, Mr. Youssef	male	NaN	2	0	2662	21.6792	NaN	C
77	78	0	3	Moutal, Mr. Rahamin Haim	male	NaN	0	0	374746	8.0500	NaN	S
235	236	0	3	Harknett, Miss. Alice Phoebe	female	NaN	0	0	W./C. 6609	7.5500	NaN	S
256	257	1	1	Thorne, Mrs. Gertrude Maybelle	female	NaN	0	0	PC 17585	79.2000	NaN	C
359	360	1	3	Mockler, Miss. Helen Mary "Ellie"	female	NaN	0	0	330980	7.8792	NaN	Q
409	410	0	3	Lefebre, Miss. Ida	female	NaN	3	1	4133	25.4667	NaN	S
454	455	0	3	Peduzzi, Mr. Joseph	male	NaN	0	0	A/5 2817	8.0500	NaN	S
481	482	0	2	Frost, Mr. Anthony Wood "Archie"	male	NaN	0	0	239854	0.0000	NaN	S
495	496	0	3	Yousseff, Mr. Gerious	male	NaN	0	0	2627	14.4583	NaN	C
593	594	0	3	Bourke, Miss. Mary	female	NaN	0	2	364848	7.7500	NaN	Q
612	613	1	3	Murphy, Miss. Margaret Jane	female	NaN	1	0	367230	15.5000	NaN	Q
667	668	0	3	Rommetvedt, Mr. Knud Paust	male	NaN	0	0	312993	7.7750	NaN	S
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
878	879	0	3	Laleff, Mr. Kristo	male	NaN	0	0	349217	7.8958	NaN	S

Q09: Display and show how many entries in the dataset have no missing value (NA) in all Age, Cabin, and Embarked columns.

There are 183 entries has no NA value in any column.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
...	...	...	...	...	...	...	...	...	...	...	...	...
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S
872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.0000	B51 B53 B55	S
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

Q10: Create a new DataFrame, without modify the original, to hold all entries withtout any NA value of any attribute from the original Titanic dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
...	...	...	...	...	...	...	...	...	...	...	...	...
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S
872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.0000	B51 B53 B55	S
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

183 rows × 12 columns

Q11: Replace all missing values (NA) with 0 without overwriting the original dataset by createing/saving in a new DataFrame. Display the overview of the new dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	0	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	0	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	0	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	0	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0.0	1	2	W./C. 6607	23.4500	0	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	0	Q

891 rows × 12 columns

Q12: Replace all missing values (NA) in Age column with 0.0, and in Cabin with 'AXX' without overwriting the original dataset by createing/saving in a new DataFrame. Display the overview of the new dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	AXX	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	AXX	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	AXX	S
...	...	...	...	...	...	...	...	...	...	...	...	...

Q13: Replace all missing values (NA) using forward filling method without overwriting the original dataset, by createing/saving in a new DataFrame. Display the overview of the new dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	C85	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	C123	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	C50	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	19.0	1	2	W./C. 6607	23.4500	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	C148	Q

891 rows × 12 columns

Q14: Exame the previous new DataFrame resulted from forward filling method, and see if there are still missing (NA) values

```
PassengerId  Survived  Pclass  Name   Sex    Age    SibSp  Parch  Ticket  Fare   Cabin  Embarked
False        False     False   False  False  False  False  False  False   False  False  False       890
                                                                                 True   False         1

dtype: int64
```

Q15: Port of Embarkation is abbreviated noted in the Embarked column with either 'C','Q',or 'S' if the value is not missing. Please write an execution statement to verify the value categorical contents in column Embarked

```
Total unique values from Embarked column:
S C Q nan
```

Q16: Use the dictionary mapping to create a new column named PortName, to accommodate the full names of the Embarked.

```
port_to_fullname = {'C': 'Cherbourg', 'Q': 'Queenstown', 'S':'Southampton'}
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	PortName
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Southampton
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Cherbourg
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Southampton
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Southampton
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Southampton
...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	Southampton
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	Southampton
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	Southampton
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	Cherbourg
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	Queenstown

891 rows × 13 columns

In [ ]:
```
port_to_fullname = {'C': 'Cherbourg', 'Q': 'Queenstown', 'S':'Southampton'}
```

Q17: Print out the index and PassengerId of the first five entries from the original Titanic dataset.

```
0    1
1    2
2    3
3    4
4    5
Name: PassengerId, dtype: int64
```

Q18: Rename row indices to be aligned with PassengerId in place.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	PortName
1	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Southampton
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	Cherbourg
3	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Southampton
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	Southampton
5	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Southampton
...	...	...	...	...	...	...	...	...	...	...	...	...	...
887	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	Southampton
888	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	Southampton
889	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	Southampton
890	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	Cherbourg
891	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	Queenstown

891 rows × 13 columns

Q19: Display the age distributions of all passengers in the dataset Using the age bins in the list.

```
ages = [12, 18, 25, 35, 60, 80]

(25, 35]    196
(35, 60]    195
(18, 25]    162
(12, 18]     70
(60, 80]     22
Name: Age, dtype: int64
```

In [ ]:
```
ages = [12, 18, 25, 35, 60, 80]
```

Q20: Display the age ranks of all passengers in the dataset using the categories where the whole age range of all passengers is divided into equal-length.

```
group_names = ['Youth', 'YoungAdult', 'MiddleAged', 'Senior']

1      YoungAdult
2      YoungAdult
3      YoungAdult
4      YoungAdult
5      YoungAdult
          ...
887    YoungAdult
888         Youth
889           NaN
890    YoungAdult
891    YoungAdult
Name: Age, Length: 891, dtype: category
Categories (4, object): ['Youth' < 'YoungAdult' < 'MiddleAged' < 'Senior']
YoungAdult    385
Youth         179
MiddleAged    128
Senior         22
Name: Age, dtype: int64
```

In [ ]:
```
group_names = ['Youth', 'YoungAdult', 'MiddleAged', 'Senior']
```

Q21: Display the age categories in quartiles of all passengers in the Titanic dataset.

```
(20.125, 28.0]     183
(0.419, 20.125]    179
(38.0, 80.0]       177
(28.0, 38.0]       175
Name: Age, dtype: int64
```

Q22: Detect Fare outliners, assuming there shouldn't be free ($0) ticket.

There are 15 zero dolloar tickets.

Q23: Convert categorical data on Embarked column into a dummy mattrix with C, Q, S as indicators

	C	Q	S
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	1
5	0	0	1
...	...	...	...
887	0	0	1
888	0	0	1
889	0	0	1
890	1	0	0
891	0	1	0

891 rows × 3 columns

Q24: Convert age distributions of all passengers in the Titanic dataset into a dummy mattrix with the age bins in the list.

```
ages = [12, 18, 25, 35, 60, 80]
```

	(12, 18]	(18, 25]	(25, 35]	(35, 60]	(60, 80]
1	0	1	0	0	0
2	0	0	0	1	0
3	0	0	1	0	0
4	0	0	1	0	0
5	0	0	1	0	0
...	...	...	...	...	...
887	0	0	1	0	0
888	0	1	0	0	0
889	0	0	0	0	0
890	0	0	1	0	0
891	0	0	1	0	0

891 rows × 5 columns

In [ ]:
```
ages = [12, 18, 25, 35, 60, 80]
```