http://tinyurl.com/2dzbea59

2 - 3:15PM

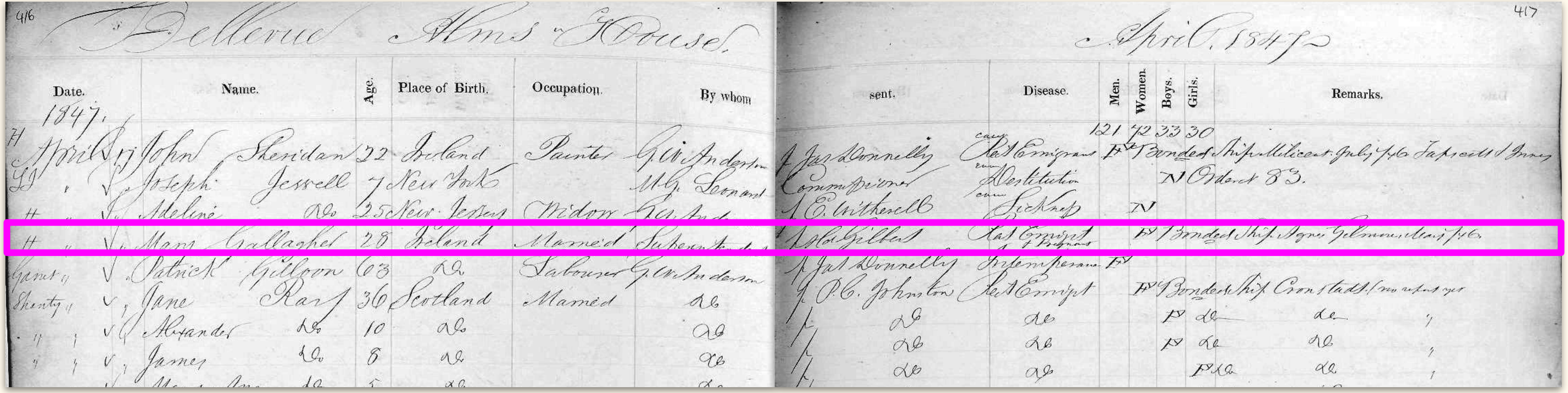
http://tinyurl.com/46czd2ex

5 - 6:15PM

- Activity 2
- Bellevue Almshouse dataset
  - Loading a dataset
  - Description of the dataset
    - .info(), .head(*n*), .sample(*n*)
    - Summary statistics
  - Dealing with duplicates
  - Frequency: Most common item in a column

Bellevue Almshouse Dataset

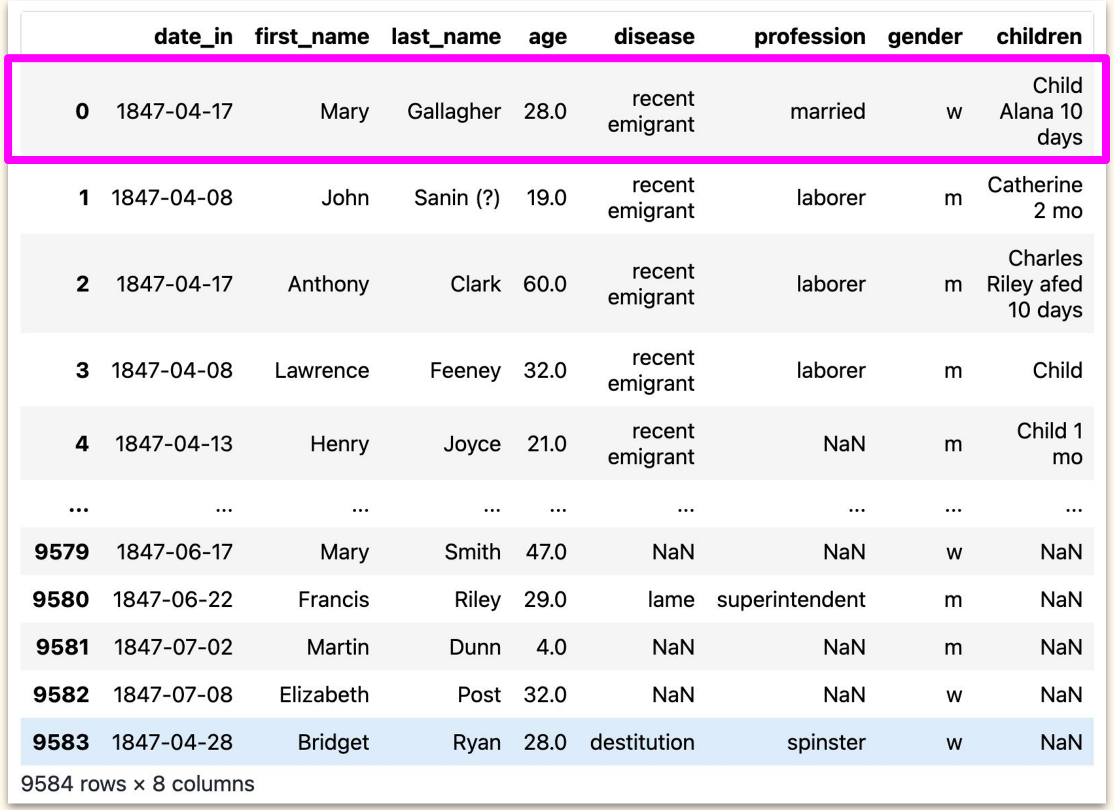
Bellevue Almshouse admission ledger



416 Bellevue Alms House. April 1847 417

Date.	Name.		Age.	Place of Birth.	Occupation.	By whom	sent.	Disease.	Men.	Women.	Boys.	Girls.	Remarks.
1847.									121	72	33	30	
H April 17	John	Sheridan	22	Ireland	Painter	G. W. Anderson	Jas Donnelly	cas Sick Emigrant	F				Bonded Ship Millicent: July 7/46 Tapscott & Innes
SI ″	Joseph	Jessell	7	New York		M. G. Leonard	Commissioner	Destitution			N		Ordered 83.
H ″	Adeline	Do	25	New Jersey	Widow	Geo. And	E. Witherell	Sickness		N			
H ″	Mary	Gallagher	28	Ireland	Married	Superintendent	Ho Gilbert	Sick Emigrant & Pregnant		F			Bonded Ship Agnes Gilmour Dec 7/46
Grant ″	Patrick	Gilloon	63	Do	Labourer	G. W. Anderson	Jas Donnelly	Intemperance	F				
Shenty ″	Jane	Ray	36	Scotland	Married	Do	P. C. Johnston	Sick Emigrant		F			Bonded Ship Cronstadt (no report yet
″	Alexander	Do	10	Do		Do	Do	Do			F		Do Do
″	James	Do	8	Do		Do	Do	Do			F		Do Do
	Mary Ann	Do	5	Do		Do	Do	Do				F	Do Do

Digitizing the Bellevue Almshouse admission ledger



	date_in	first_name	last_name	age	disease	profession	gender	children
0	1847-04-17	Mary	Gallagher	28.0	recent emigrant	married	w	Child Alana 10 days
1	1847-04-08	John	Sanin (?)	19.0	recent emigrant	laborer	m	Catherine 2 mo
2	1847-04-17	Anthony	Clark	60.0	recent emigrant	laborer	m	Charles Riley afed 10 days
3	1847-04-08	Lawrence	Feeney	32.0	recent emigrant	laborer	m	Child
4	1847-04-13	Henry	Joyce	21.0	recent emigrant	NaN	m	Child 1 mo
...	...	...	...	...	...	...	...	...
9579	1847-06-17	Mary	Smith	47.0	NaN	NaN	w	NaN
9580	1847-06-22	Francis	Riley	29.0	lame	superintendent	m	NaN
9581	1847-07-02	Martin	Dunn	4.0	NaN	NaN	m	NaN
9582	1847-07-08	Elizabeth	Post	32.0	NaN	NaN	w	NaN
9583	1847-04-28	Bridget	Ryan	28.0	destitution	spinster	w	NaN

9584 rows × 8 columns

Loading a dataset

- import pandas as pd
- pd.read_csv(‘filepath’)

Description of the dataset

- .info():
  - Displays the total count of non-N/A, non-blank items, and the datatype of each column
- .head(n):
  - Provides the first *n* of rows
- .sample(n)
  - Provides a random *n* of rows
- .describe(include = ‘all’)
  - Provides the summary statistics of all the variables in the dataframe

Summary statistics

- Measures central tendency
  - Mean, median, mode
  - A value that represents the middle or centre of its distribution
- Measures spread of distribution
  - How far are the values spread from the smallest value to the largest value

Dealing with duplicates

- .duplicated(keep = ‘first’/’last’/False):
    - Creates a True/False dataframe to check which rows in the original dataframe are duplicated
    - keep
        - first: considers the first entry in the dataframe as the unique entry
        - last: considers the last entry in the dataframe as the unique entry
        - False: considers all entry as duplicates
    - Default argument: keep = ‘first’

Dealing with duplicates

- df[df.duplicated(keep=False)]
  - Compares the duplicated rows of True/False with the original dataframe and displays all the duplicated rows
- .drop_duplicate(keep = ‘first’/’last’/False):
  - Drops all the duplicated rows and keeps the first entry, last entry, or none of the entries
  - Default argument: keep = ‘first’

Frequency: Most common items in a column

- df[“column_name”].value_counts()
  - To count the number of unique values in a column