# 6 - Pandas-IngestData

October 14, 2021

Table of Contents



## 1 Ingest, Inspect & Clean

[35]:
```
import numpy as np
import pandas as pd
from numpy.random import randn


#import os
```

```
#for dirname, _, filenames in os.walk('/kaggle/input'):
#    for filename in filenames:
#        print(os.path.join(dirname, filename))
```

## 1.1 Ingest data

https://pandas.pydata.org/docs/user_guide/io.html

**NOTE** - diabetes1 - 19 columns w/ header - diabetes2 - 19 columns no header

### 1.1.1 Reading csv files

[36]:
```
diabetes1 = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes1.csv')
```

[37]:
```
diabetes1.head()
```

[37]:

| | encounter_id | patient_nbr | race | gender | age | weight |
|---|---|---|---|---|---|---|
| 0 | 2278392 | 8222157 | Caucasian | Female | [0-10) | ? |
| 1 | 149190 | 55629189 | Caucasian | Female | [10-20) | ? |
| 2 | 64410 | 86047875 | AfricanAmerican | Female | [20-30) | ? |
| 3 | 500364 | 82442376 | Caucasian | Male | [30-40) | ? |
| 4 | 16680 | 42519267 | Caucasian | Male | [40-50) | ? |

| | admission_type_id | admission_source_id | time_in_hospital | payer_code |
|---|---|---|---|---|
| 0 | 6 | 1 | 1 | ? |
| 1 | 1 | 7 | 3 | ? |
| 2 | 1 | 7 | 2 | ? |
| 3 | 1 | 7 | 2 | ? |
| 4 | 1 | 7 | 1 | ? |

| | num_lab_procedures | num_procedures | num_medications | diag_1 | A1Cresult |
|---|---|---|---|---|---|
| 0 | 41 | 0 | 1 | 250.83 | None |
| 1 | 59 | 0 | 18 | 276 | None |
| 2 | 11 | 5 | 13 | 648 | None |
| 3 | 44 | 1 | 16 | 8 | None |
| 4 | 51 | 0 | 8 | 197 | None |

| | metformin | miglitol | insulin | readmitted | Date |
|---|---|---|---|---|---|
| 0 | No | No | No | NO | 10/20/2021 |
| 1 | No | No | Up | >30 | 09/25/2021 |
| 2 | No | No | No | NO | 08/29/2021 |
| 3 | No | No | Up | NO | 10/20/2021 |
| 4 | No | No | Steady | NO | 09/25/2021 |

### 1.1.2 read_csv options

[38]:
```
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv')
```

[39]:
```
df.head()
```

[39]:
```
   2278392   8222157        Caucasian  Female   [0-10)  ?  6  1  1.1 ?.1  41  \
0   149190  55629189        Caucasian  Female  [10-20)  ?  1  7    3   ?  59
1    64410  86047875  AfricanAmerican  Female  [20-30)  ?  1  7    2   ?  11
2   500364  82442376        Caucasian    Male  [30-40)  ?  1  7    2   ?  44
3    16680  42519267        Caucasian    Male  [40-50)  ?  1  7    1   ?  51
4    35754  82637451        Caucasian    Male  [50-60)  ?  2  2    3   ?  31

   0  1.2 250.83  None  No No.1    No.2   NO  10/20/21
0  0   18    276  None  No   No      Up  >30   9/25/21
1  5   13    648  None  No   No      No   NO   8/29/21
2  1   16      8  None  No   No      Up   NO  10/20/21
3  0    8    197  None  No   No  Steady   NO   9/25/21
4  6   16    414  None  No   No  Steady  >30   8/29/21
```

[40]:
```
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv',
                 header = None)
df.head()
```

[40]:
```
         0         1                2       3        4  5   6   7   8  9   10  \
0  2278392   8222157        Caucasian  Female   [0-10)  ?   6   1   1  ?  41
1   149190  55629189        Caucasian  Female  [10-20)  ?   1   7   3  ?  59
2    64410  86047875  AfricanAmerican  Female  [20-30)  ?   1   7   2  ?  11
3   500364  82442376        Caucasian    Male  [30-40)  ?   1   7   2  ?  44
4    16680  42519267        Caucasian    Male  [40-50)  ?   1   7   1  ?  51

   11  12      13    14  15  16      17   18        19
0   0   1  250.83  None  No  No      No   NO  10/20/21
1   0  18     276  None  No  No      Up  >30   9/25/21
2   5  13     648  None  No  No      No   NO   8/29/21
3   1  16       8  None  No  No      Up   NO  10/20/21
4   0   8     197  None  No  No  Steady   NO   9/25/21
```

[41]:
```
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv',
                 header = None,
                names = ('EncounterId','b','c','d','e','f','g','h','i','j','k',
                         'l','m','n','o','p','q','r','s','EDT'))
#df.head()
```

[42]:
```
# Override existing column names

diabetes1 = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes1.csv',
                        header = 0,
                        names =␣
↪('EncounterId','b','c','d','e','f','g','h','i','j','k',
                                'l','m','n','o','p','q','r','s','EDT'))
#diabetes1.head()
```

[43]:
```
# Use a column as the index

df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv',
                 header = None,
                names = ('EncounterId','b','c','d','e','f','g','h','i','j','k',
                         'l','m','n','o','p','q','r','s','EDT'),
                index_col = 'EncounterId')
#df.head()
```

[44]:
```
df.loc[64410]
```

[44]:
```
b             86047875
c      AfricanAmerican
d               Female
e              [20-30)
f                    ?
g                    1
h                    7
i                    2
j                    ?
k                   11
l                    5
m                   13
n                  648
o                 None
p                   No
q                   No
r                   No
s                   NO
EDT            8/29/21
Name: 64410, dtype: object
```

[45]:
```
# Only use a subset of the columns

df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv',
```

```
                 header = None,
                names = ('EncounterId','b','c','d','e','EDT'),
                index_col = 'EncounterId',
                usecols = [0,1,2,3,7,19])
#df.head()
```

[46]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 101768 entries, 2278392 to 443797076
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   b       101768 non-null  int64
 1   c       101768 non-null  object
 2   d       101768 non-null  object
 3   e       101768 non-null  int64
 4   EDT     101768 non-null  object
dtypes: int64(2), object(3)
memory usage: 4.7+ MB
```

[47]:
```
# Change data types as the data is read in
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪diabetes2.csv',
                 header = None,
                names = ('EncounterId','b','c','d','e','EDT'),
                index_col = 'EncounterId',
                usecols = [0,1,2,3,7,19],
                dtype = {'e':object }
                )
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 101768 entries, 2278392 to 443797076
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   b       101768 non-null  int64
 1   c       101768 non-null  object
 2   d       101768 non-null  object
 3   e       101768 non-null  object
 4   EDT     101768 non-null  object
dtypes: int64(1), object(4)
memory usage: 4.7+ MB
```

[48]:
```
# Change data types as the data is read in
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪diabetes2.csv',
```

```
                 header = None,
                names = ('EncounterId','b','c','d','e','EDT'),
                index_col = 'EncounterId',
                usecols = [0,1,2,3,7,19],
                dtype = object)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 101768 entries, 2278392 to 443797076
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   b       101768 non-null  object
 1   c       101768 non-null  object
 2   d       101768 non-null  object
 3   e       101768 non-null  object
 4   EDT     101768 non-null  object
dtypes: object(5)
memory usage: 4.7+ MB
```

### 1.1.3 Ingesting Dates

[49]:
```
# Change data types as the data is read in
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪diabetes2.csv',
                 header = None,
                names = ('EncounterId','b','c','d','e','EDT'),
                index_col = 'EncounterId',
                usecols = [0,1,2,3,7,19],
                dtype = {'e':object, },
                parse_dates=['EDT']
                )
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 101768 entries, 2278392 to 443797076
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   b       101768 non-null  int64
 1   c       101768 non-null  object
 2   d       101768 non-null  object
 3   e       101768 non-null  object
 4   EDT     101768 non-null  datetime64[ns]
dtypes: datetime64[ns](1), int64(1), object(3)
memory usage: 4.7+ MB
```

[50]:
```
df
```

[50]:
```
                     b                c       d  e        EDT
EncounterId
2278392        8222157        Caucasian  Female  1 2021-10-20
149190        55629189        Caucasian  Female  7 2021-09-25
64410         86047875  AfricanAmerican  Female  7 2021-08-29
500364        82442376        Caucasian    Male  7 2021-10-20
16680         42519267        Caucasian    Male  7 2021-09-25
…                …               …      … ..         …
443854148     41088789        Caucasian    Male  7 2021-10-20
443857166     31693671        Caucasian  Female  7 2021-09-25
443867222    175429310        Caucasian    Male  7 2021-08-29
62256         49726791  AfricanAmerican  Female  2 2021-10-20
443797076    183766055        Caucasian    Male  1 2021-09-25

[101768 rows x 5 columns]
```

[51]:
```
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
↪diabetes2.csv',
                 header = None,
                 names = ('EncounterId','b','c','d','e','EDT'),
                 index_col = 'EncounterId',
                 usecols = [0,1,2,3,7,19],
                 dtype = {'e':object, },
                 parse_dates=['EDT'],
                 dayfirst=True                    # Why isn't this working?
                )
df
```

[51]:
```
                     b                c       d  e        EDT
EncounterId
2278392        8222157        Caucasian  Female  1 2021-10-20
149190        55629189        Caucasian  Female  7 2021-09-25
64410         86047875  AfricanAmerican  Female  7 2021-08-29
500364        82442376        Caucasian    Male  7 2021-10-20
16680         42519267        Caucasian    Male  7 2021-09-25
…                …               …      … ..         …
443854148     41088789        Caucasian    Male  7 2021-10-20
443857166     31693671        Caucasian  Female  7 2021-09-25
443867222    175429310        Caucasian    Male  7 2021-08-29
62256         49726791  AfricanAmerican  Female  2 2021-10-20
443797076    183766055        Caucasian    Male  1 2021-09-25

[101768 rows x 5 columns]
```

[52]:
```
df.shape
```

[52]:
```
(101768, 5)
```

[53]:
```
# Only bring in X number of rows
df = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪diabetes2.csv',
                 header = None,
                names = ('EncounterId','b','c','d','e'),
                index_col = 'EncounterId',
                usecols = [0,1,2,3,7],
                nrows = 1000)
df.shape
```

[53]:
```
(1000, 4)
```

### 1.1.4 Reading other file types

- excel
- json
- APIs
- database

[54]:
```
# Read in an excel spreadsheet
# I had to install openpyxl in my anaconda environment for this to work.

imm = pd.read_excel('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪immunotherapy.xlsx',
                                sheet_name = 'Immuno1')
imm.head()
```

[54]:

| | sex | age | Time | Number_of_Warts | Type | Area | induration_diameter | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | |
| 2 | 1 | 16 | 10.50 | 2 | 1 | 100 | 25 | |
| 3 | 1 | 27 | 4.50 | 9 | 3 | 80 | 30 | |
| 4 | 1 | 20 | 8.00 | 6 | 1 | 45 | 8 | |

| | Result_of_Treatment |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |

[55]:
```
# Saving a dataframe to json formay
imm.to_json('/Users/jimcody/Documents/2021Python/intropython/data/immjson.json')
```

[56]:
```
# Read in a json file

imm2 = pd.read_json('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪immjson.json')
```

```
imm2.head()
```

[56]:

| | sex | age | Time | Number_of_Warts | Type | Area | induration_diameter | Result_of_Treatment |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | 1 |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | 1 |
| 2 | 1 | 16 | 10.50 | 2 | 1 | 100 | 25 | 1 |
| 3 | 1 | 27 | 4.50 | 9 | 3 | 80 | 30 | 1 |
| 4 | 1 | 20 | 8.00 | 6 | 1 | 45 | 8 | 1 |

### 1.1.5 Using an API as input - - - This will not run on Kaggle

[57]:
```
import requests
```

[58]:
```
# Send the request and receive a response
# Get the content of the response (there are other parts of the response (e.g., header))
# Display the content in json format

#response = requests.get("https://data.cdc.gov/resource/saz5-9hgg.json")
#jsonhold = response.json()
#jsonhold
```

[59]:
```
# Put the content into a DataFrame
# Display the DataFrame

#vaccines = pd.DataFrame(jsonhold)
#vaccines
```

[60]:
```
# Same process
# Just set the url as a variable
# Combined a few other steps


#response = requests.get(url)
#vermont = pd.DataFrame(response.json())
#vermont
```

### 1.1.6 A local example running on my machine

[61]:
```
#import requests
#url = 'http://localhost:8080/rest/applicants'
#response = requests.get(url)
#json1 = response.json()
#json1
# Notice that all the data is wrapped in 'content'
```

[62]:
```
# We need to get the results inside of 'content'.
#df5 = pd.DataFrame(json1['content'])
#df5
```

## 1.2 Exercise - 15 minutes

### 1.2.1 Part 1

- Bring the csv file diabetes_inspect into a DataFrame. Name the dataframe df.
- What is its shape?

### 1.2.2 Part 2

- Read the outbreaks2.csv file into a pdf. Name the df something you will remember.
- The csv file does not have a header row.
- This data is a listing of food born illness outbreaks.
- Do not bring in the 6th or 8th columns. The df should have 10 columns.
- The remaining columns contain the following data. You can decide how to name the columns
  - year
  - month
  - state
  - location
  - food
  - status
  - illnesses
  - hospitalizations
  - fatalities
- What is the shape of this DataFrame?

| A | B | C | D | E | G | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|
| Year | Month | State | Location | Food | Species | Status | Illnesses | Hospitalizati | Fatalities |
| 1998 | January | California | Restaurant | | | | 20 | 0 | 0 |
| 1998 | January | California | | Custard | | | 112 | 0 | 0 |
| 1998 | January | California | Restaurant | | | | 35 | 0 | 0 |
| 1998 | January | California | Restaurant | Fish, Ahi | Scombroid toxin | Confirmed | 4 | 0 | 0 |
| 1998 | January | California | Private Home/Residence | Lasagna, Unspecified; Eggs, Other | Salmonella enterica | Confirmed | 26 | 3 | 0 |
| 1998 | January | California | Restaurant | | Shigella boydii | Confirmed | 25 | 3 | 0 |
| 1998 | January | California | Restaurant | | | | 8 | 0 | 0 |
| 1998 | January | California | Restaurant | Stuffing, Unspecified; Sandwich, Turkey | Salmonella enterica | Confirmed | 4 | 3 | 0 |

[63]:
```
# Part 1
```

```
diabetes_i = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪diabetes_inspect.csv')
diabetes_i.shape
```

[63]: (101766, 33)

[64]:
```
outbreaks = pd.read_csv('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪outbreaks2.csv',
                header = None,
                usecols = [0,1,2,3,4,6,8,9,10,11],
                names = ('year','month','state','location','food','species',
                         'status','illness','hospitalizations','fatalities')
                )
outbreaks.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19119 entries, 0 to 19118
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   year              19119 non-null  int64
 1   month             19119 non-null  object
 2   state             19119 non-null  object
 3   location          16953 non-null  object
 4   food              10156 non-null  object
 5   species           12500 non-null  object
 6   status            12500 non-null  object
 7   illness           19119 non-null  int64
 8   hospitalizations  15494 non-null  float64
 9   fatalities        15518 non-null  float64
dtypes: float64(2), int64(2), object(6)
memory usage: 1.5+ MB
```

# 2 Renaming Preview

## 2.1 Rename using names in read_csv

[65]:
```
imm = pd.read_excel('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪immunotherapy.xlsx',
                                sheet_name = 'Immuno1')
imm.head()
```

[65]:

| | sex | age | Time | Number_of_Warts | Type | Area | induration_diameter | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | |
| 2 | 1 | 16 | 10.50 | 2 | 1 | 100 | 25 | |
| 3 | 1 | 27 | 4.50 | 9 | 3 | 80 | 30 | |

```
4    1   20   8.00                6      1    45                    8

   Result_of_Treatment
0                    1
1                    1
2                    1
3                    1
4                    1
```

[66]:
```
imm = pd.read_excel('/Users/jimcody/Documents/2021Python/intropython/data/
 ↪immunotherapy.xlsx',
                    names = ('a','b','c','d','e','f','g','h'),
                    sheet_name = 'Immuno1')
imm.head()
```

[66]:

| | a | b | c | d | e | f | g | h |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | 1 |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | 1 |
| 2 | 1 | 16 | 10.50 | 2 | 1 | 100 | 25 | 1 |
| 3 | 1 | 27 | 4.50 | 9 | 3 | 80 | 30 | 1 |
| 4 | 1 | 20 | 8.00 | 6 | 1 | 45 | 8 | 1 |

### 2.2 Rename all columns using rename()

[67]:
```
imm.rename(columns = {'a':'a1','b':'b1','c':'c1','d':'d1','e':'e1','f':'f1','g':
 ↪'g1','h':'h1'}, inplace = True)
imm.head()

# An alternative - using the axis=1
# imm.rename({'a':'a1','b':'b1','c':'c1','d':'d1','e':'e1','f':'f1','g':
 ↪'g1','h':'h1'}, axis =1, inplace = True)
```

[67]:

| | a1 | b1 | c1 | d1 | e1 | f1 | g1 | h1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | 1 |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | 1 |
| 2 | 1 | 16 | 10.50 | 2 | 1 | 100 | 25 | 1 |
| 3 | 1 | 27 | 4.50 | 9 | 3 | 80 | 30 | 1 |
| 4 | 1 | 20 | 8.00 | 6 | 1 | 45 | 8 | 1 |

### 2.3 Rename some columns using rename()

[68]:
```
imm.rename(columns = {'b1':'BB','e1':'EE'}, inplace = True)
imm.head()
```

[68]:

| | a1 | BB | c1 | d1 | EE | f1 | g1 | h1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22 | 2.25 | 14 | 3 | 51 | 50 | 1 |
| 1 | 1 | 15 | 3.00 | 2 | 3 | 900 | 70 | 1 |

```
2   1  16  10.50   2   1  100  25   1
3   1  27   4.50   9   3   80  30   1
4   1  20   8.00   6   1   45   8   1
```

### 2.4 Rename using setaxis()

[69]:
```
imm.set_axis(['AA','BB','CC','DD','EE','FF','GG','HH'], axis = 'columns',␣
↪inplace = True)
imm.head()
```

[69]:
```
   AA  BB     CC  DD  EE   FF  GG  HH
0   1  22   2.25  14   3   51  50   1
1   1  15   3.00   2   3  900  70   1
2   1  16  10.50   2   1  100  25   1
3   1  27   4.50   9   3   80  30   1
4   1  20   8.00   6   1   45   8   1
```

### 2.5 Rename using .columns

[70]:
```
imm.columns = ['AaA','BbB','CcC','DdD','EeE','FfF','GgG','HhH']
imm.head()
```

[70]:
```
   AaA  BbB    CcC  DdD  EeE  FfF  GgG  HhH
0    1   22   2.25   14    3   51   50    1
1    1   15   3.00    2    3  900   70    1
2    1   16  10.50    2    1  100   25    1
3    1   27   4.50    9    3   80   30    1
4    1   20   8.00    6    1   45    8    1
```

### 2.6 Rename using str.replace (a string function)

[71]:
```
imm.columns = imm.columns.str.replace('CcC','CCC')
imm.head()
```

[71]:
```
   AaA  BbB    CCC  DdD  EeE  FfF  GgG  HhH
0    1   22   2.25   14    3   51   50    1
1    1   15   3.00    2    3  900   70    1
2    1   16  10.50    2    1  100   25    1
3    1   27   4.50    9    3   80   30    1
4    1   20   8.00    6    1   45    8    1
```