Data Analysis with Pandas

Data Analysis

Kunal Khurana

2023-12-25

Table of contents

Creating, Reading and Writing

DataFrame and Series

```
import pandas as pd
```

2 core objects- - DataFrame - array of inidividual entries (contains row and column) >keys = 'column names', values = list of entries

`>rows = **Index**`

- Series- sequence of data values > don't have any column name

  row names defined by **index** parameter aswell

```
#DataFrame_integer
pd.DataFrame({'Yes' : [390, 233], 'No' : [1,23]})
```

	Yes	No
0	390	1
1	233	23

```
# DataFrame_Strings
pd.DataFrame({'Suzaine': ['I liked chocolate', 'Lets have some fun'],
              'Marie': ['butterscotch worked fine', 'wow, its raining']},
             index = ['topic_1', 'topic_2'])
```

	Suzaine	Marie
topic_1	I liked chocolate	butterscotch worked fine
topic_2	Lets have some fun	wow, its raining

```
# series
pd.Series([1, 2, 3],
index= ['2014_sales', '2015_sales', '2016_sales'],
name = 'Product A')
```

```
2014_sales    1
2015_sales    2
2016_sales    3
Name: Product A, dtype: int64
```

```
# example
Dinner = pd.Series(['4 cups', '1 cup', '2 large', '1 can'],
         index = ['Flour', 'Milk', 'Eggs', 'Spam'],
         name = 'Dinner')
print(Dinner)
```

```
Flour     4 cups
Milk       1 cup
Eggs     2 large
Spam       1 can
Name: Dinner, dtype: object
```

Writing data files

```
Dinner.to_csv("Dinner.csv")
```

Reading data files

```
reactions = pd.read_csv('Reactions.csv')
print(reactions.shape)
```

```
(25553, 5)
```

```
print(reactions.head())
```

```
   Unnamed: 0                            Content ID  \
0           0  97522e57-d9ab-4bd6-97bf-c24d952602d2
1           1  97522e57-d9ab-4bd6-97bf-c24d952602d2
2           2  97522e57-d9ab-4bd6-97bf-c24d952602d2
3           3  97522e57-d9ab-4bd6-97bf-c24d952602d2
4           4  97522e57-d9ab-4bd6-97bf-c24d952602d2

                                User ID     Type             Datetime
0                                   NaN      NaN  2021-04-22 15:17:15
1  5d454588-283d-459d-915d-c48a2cb4c27f  disgust  2020-11-07 09:43:50
2  92b87fa5-f271-43e0-af66-84fac21052e6  dislike  2021-06-17 12:22:51
3  163daa38-8b77-48c9-9af6-37a6c1447ac2   scared  2021-04-18 05:13:58
4  34e8add9-0206-47fd-a501-037b994650a2  disgust  2021-01-06 19:13:01
```

Indexing, Selecting and Assigning

```
data = pd.read_csv("winemag-data-130k-v2.csv")
pd.set_option('display.max_rows', 5)
print(data.head())
```

```
   Unnamed: 0   country                                        description  \
0           0     Italy  Aromas include tropical fruit, broom, brimston...
1           1  Portugal  This is ripe and fruity, a wine that is smooth...
2           2        US  Tart and snappy, the flavors of lime flesh and...
3           3        US  Pineapple rind, lemon pith and orange blossom ...
4           4        US  Much like the regular bottling from 2012, this...

                      designation  points  price           province  \
```

```
0                        Vulkà Bianco      87    NaN  Sicily & Sardinia
1                            Avidagos      87   15.0              Douro
2                                 NaN      87   14.0             Oregon
3                  Reserve Late Harvest      87   13.0           Michigan
4  Vintner's Reserve Wild Child Block      87   65.0             Oregon

              region_1           region_2         taster_name  \
0                 Etna                NaN       Kerin O'Keefe
1                  NaN                NaN          Roger Voss
2    Willamette Valley  Willamette Valley        Paul Gregutt
3  Lake Michigan Shore                NaN  Alexander Peartree
4    Willamette Valley  Willamette Valley        Paul Gregutt

  taster_twitter_handle                                          title  \
0          @kerinokeefe                 Nicosia 2013 Vulkà Bianco  (Etna)
1            @vossroger      Quinta dos Avidagos 2011 Avidagos Red (Douro)
2           @paulgwine       Rainstorm 2013 Pinot Gris (Willamette Valley)
3                   NaN  St. Julian 2013 Reserve Late Harvest Riesling ...
4           @paulgwine   Sweet Cheeks 2012 Vintner's Reserve Wild Child...

          variety               winery
0     White Blend              Nicosia
1  Portuguese Red  Quinta dos Avidagos
2      Pinot Gris            Rainstorm
3        Riesling           St. Julian
4      Pinot Noir         Sweet Cheeks
```

```
print(data.columns)
```

```
Index(['Unnamed: 0', 'country', 'description', 'designation', 'points',
       'price', 'province', 'region_1', 'region_2', 'taster_name',
       'taster_twitter_handle', 'title', 'variety', 'winery'],
      dtype='object')
```

```
print(data.country)
```

```
0            Italy
1         Portugal
             ...
129969      France
```

```
129970      France
Name: country, Length: 129971, dtype: object
```

```
print(data['country'])  #handles reserved characters
```

```
0            Italy
1         Portugal
            ...
129969      France
129970      France
Name: country, Length: 129971, dtype: object
```

```
print(data['country'][4])
```

```
US
```

Indexing

index based or numerical position based (.iloc operator used)

- python's std. library appraoch (0:10 selects 0, 1, ...9)

label based or value based (.loc operator used)

-indexes inclusively. So 0:10 will select entries 0,...,10

```
# selecting first row
data.iloc[0]
```

```
Unnamed: 0              0
country             Italy
                      ...
variety       White Blend
winery            Nicosia
Name: 0, Length: 14, dtype: object
```

```
data.iloc[:3, 1]
```

```
0       Italy
1    Portugal
2          US
Name: country, dtype: object
```

```
data.iloc[-5:] #selecting last 5 rows, plus all columns
```

	Unnamed: 0	country	description	designation
129966	129966	Germany	Notes of honeysuckle and cantaloupe sweeten th...	Brauneberger Juffer-S
129967	129967	US	Citation is given as much as a decade of bottl...	NaN
129968	129968	France	Well-drained gravel soil gives this wine its c...	Kritt
129969	129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN
129970	129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée

```
data.loc[:, ['taster_name', 'variety', 'winery']]
```

	taster_name	variety	winery
0	Kerin O'Keefe	White Blend	Nicosia
1	Roger Voss	Portuguese Red	Quinta dos Avidagos
...	...	...	...
129969	Roger Voss	Pinot Gris	Domaine Marcel Deiss
129970	Roger Voss	Gewürztraminer	Domaine Schoffit

Manipulating the index

```
data.set_index('title') #now first column is title
```

title	Unnamed: 0	country	de
Nicosia 2013 Vulkà Bianco (Etna)	0	Italy	Ar
Quinta dos Avidagos 2011 Avidagos Red (Douro)	1	Portugal	Th
...	...	...	...
Domaine Marcel Deiss 2012 Pinot Gris (Alsace)	129969	France	A

title	Unnamed: 0	country	des
Domaine Schoffit 2012 Lieu-dit Harth Cuvée Caroline Gewurztraminer (Alsace)	129970	France	Bi

```
# conditional selection
# selects data with US in columns names for countries
data.loc[data.country == 'US']
```

	Unnamed: 0	country	description	designation
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest
...	...	...	...	...
129952	129952	US	This Zinfandel from the eastern section of Nap...	NaN
129967	129967	US	Citation is given as much as a decade of bottl...	NaN

```
# selecting particular rows
indices = [1, 2, 3, 5, 8]
sample_rows = data.loc[indices]
print(sample_rows)
```

```
   Unnamed: 0   country                                        description  \
1           1  Portugal  This is ripe and fruity, a wine that is smooth...
2           2        US  Tart and snappy, the flavors of lime flesh and...
3           3        US  Pineapple rind, lemon pith and orange blossom ...
5           5     Spain  Blackberry and raspberry aromas show a typical...
8           8   Germany  Savory dried thyme notes accent sunnier flavor...

            designation  points  price        province             region_1  \
1             Avidagos      50   15.0           Douro                  NaN
2                  NaN      50   14.0          Oregon    Willamette Valley
3  Reserve Late Harvest      50   13.0        Michigan  Lake Michigan Shore
5          Ars In Vitro      50   15.0  Northern Spain              Navarra
8                Shine      50   12.0     Rheinhessen                  NaN

           region_2         taster_name taster_twitter_handle  \
1               NaN          Roger Voss           @vossroger
2  Willamette Valley        Paul Gregutt          @paulgwine
3               NaN  Alexander Peartree                  NaN
```

```
5                NaN   Michael Schachner           @wineschach
8                NaN  Anna Lee C. Iijima                   NaN

                                                 title             variety  \
1      Quinta dos Avidagos 2011 Avidagos Red (Douro)      Portuguese Red
2      Rainstorm 2013 Pinot Gris (Willamette Valley)          Pinot Gris
3  St. Julian 2013 Reserve Late Harvest Riesling ...            Riesling
5  Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...  Tempranillo-Merlot
8  Heinz Eifel 2013 Shine Gewürztraminer (Rheinhe...      Gewürztraminer

                winery
1  Quinta dos Avidagos
2            Rainstorm
3           St. Julian
5               Tandem
8          Heinz Eifel
```

```
# selecting costly wines from US
data.loc[(data.country == 'US') & (data.price >= 75)]
```

	Unnamed: 0	country	description	designation
60	60	US	Syrupy and dense, this wine is jammy in plum a...	Estate
73	73	US	Juicy plum, raspberry and pencil lead lead the...	Bella Vetta Vineyard
...	...	...	...	...
129919	129919	US	This ripe, rich, almost decadently thick wine ...	Reserve
129967	129967	US	Citation is given as much as a decade of bottl...	NaN

```
# wines from Australia and New Zealand
data.loc[
    (data.country.isin(['Australia', 'New Zealand']))
]
```

	Unnamed: 0	country	description	designation
77	77	Australia	This medium-bodied Chardonnay features aromas ...	Made With Org
83	83	Australia	Pale copper in hue, this wine exudes passion f...	Jester Sangioves
...	...	...	...	...
129956	129956	New Zealand	The blend is 44% Merlot, 33% Cabernet Sauvigno...	Gimblett Grave
129958	129958	New Zealand	This blend of Cabernet Sauvignon-Merlot and Ca...	Irongate

```
# selecting rows and columns
columns = ['price', 'region_1', 'region_2']
rows = [1, 10, 100]
df = data.loc[rows, columns]
print(df)
```

```
     price      region_1      region_2
1     15.0           NaN           NaN
10    19.0   Napa Valley          Napa
100   18.0  Finger Lakes  Finger Lakes
```

```
# selecting notnull values
data.loc[data.price.notnull()]
```

	Unnamed: 0	country	description	designation
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN
...	...	...	...	...
129969	129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN
129970	129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée Car

Assigning data

```
data['points'] = 50
print(data['points'])
```

```
0         50
1         50
          ..
129969    50
129970    50
Name: points, Length: 129971, dtype: int64
```