AN7914 Week 6 Python

March 14, 2023

1 Week 6 Python

1.1 Setting Indexes

Let's look at simple example. Let's first create at toy dataset

[83]:
```
import pandas as pd

people = {
    "first": ["Sakib", 'Jane', 'John',],
    "last": ["Anwar", 'Doe', 'Doe'],
    "email": ["SakibAnwar@winchester.ac.uk", 'JaneDoe@email.com',␣
 ↪'JohnDoe@email.com'],
    "age":[100,24,32],
    "degree":['Economics','Economics','Management'],
    "role":['Programmer','Analyst','HR'],
}
df = pd.DataFrame(people)
df
```

[83]:

	first	last	email	age	degree	role
0	Sakib	Anwar	SakibAnwar@winchester.ac.uk	100	Economics	Programmer
1	Jane	Doe	JaneDoe@email.com	24	Economics	Analyst
2	John	Doe	JohnDoe@email.com	32	Management	HR

We can look at the email column.

[84]:
```
df['email']
```

[84]:
```
0    SakibAnwar@winchester.ac.uk
1              JaneDoe@email.com
2              JohnDoe@email.com
Name: email, dtype: object
```

The current index in the dataframe:

[85]:
```
df.index
```

[85]:
```
RangeIndex(start=0, stop=3, step=1)
```

If we want to change the index. Maybe we want the index to the email addresses.

[86]:
```
df.set_index('email')
```

[86]:
```
                            first   last  age      degree        role
email
SakibAnwar@winchester.ac.uk  Sakib  Anwar  100   Economics  Programmer
JaneDoe@email.com             Jane    Doe   24   Economics     Analyst
JohnDoe@email.com             John    Doe   32  Management          HR
```

But we have not actually changed anything in the dataframe.

[87]:
```
df
```

[87]:
```
   first   last                         email  age      degree        role
0  Sakib  Anwar  SakibAnwar@winchester.ac.uk  100   Economics  Programmer
1   Jane    Doe            JaneDoe@email.com   24   Economics     Analyst
2   John    Doe            JohnDoe@email.com   32  Management          HR
```

To change we can do the following:

[88]:
```
df.set_index('email',inplace=True)
df
```

[88]:
```
                            first   last  age      degree        role
email
SakibAnwar@winchester.ac.uk  Sakib  Anwar  100   Economics  Programmer
JaneDoe@email.com             Jane    Doe   24   Economics     Analyst
JohnDoe@email.com             John    Doe   32  Management          HR
```

Now let's look a the indexes.

[89]:
```
df.index
```

[89]:
```
Index(['SakibAnwar@winchester.ac.uk', 'JaneDoe@email.com',
       'JohnDoe@email.com'],
      dtype='object', name='email')
```

Now we can find more information for the person holding the email address `Sakib.Anwar@winchester.ac.uk`

[90]:
```
df.loc['SakibAnwar@winchester.ac.uk']
```

[90]:
```
first          Sakib
last           Anwar
age              100
degree     Economics
role      Programmer
Name: SakibAnwar@winchester.ac.uk, dtype: object
```

If we want to go back to the default index. You can just reset index.

[91]:
```
df.reset_index(inplace=True)
df
```

[91]:

	email	first	last	age	degree	role
0	SakibAnwar@winchester.ac.uk	Sakib	Anwar	100	Economics	Programmer
1	JaneDoe@email.com	Jane	Doe	24	Economics	Analyst
2	JohnDoe@email.com	John	Doe	32	Management	HR

Why do you think this might be useful? Let's look a a different dataframe. We will read in two csv files.

[70]:
```
df_stckov=pd.read_csv('stack-overflow-developer-survey-2022/
↪survey_results_public.csv')
```

[71]:
```
df_schema=pd.read_csv('stack-overflow-developer-survey-2022/
↪survey_results_schema.csv')
```

[72]:
```
df_stckov.head(10)
```

[72]:

	ResponseId	MainBranch	Employment	RemoteWork
0	1	None of these	NaN	NaN
1	2	I am a developer by profession	Employed, full-time	Fully remote
2	3	I am not primarily a developer, but I write co…	Employed, full-time	
3	4	I am a developer by profession	Employed, full-time	
4	5	I am a developer by profession	Employed, full-time	
5	6	I am not primarily a developer, but I write co…	Student, full-time	
6	7	I code primarily as a hobby	Student, part-time	
7	8	I am a developer by profession	Not employed, but looking for work	
8	9	I am a developer by profession	Employed, full-time	
9	10	I am a developer by profession	Independent contractor, freelancer, or self-em…	

```
2  Hybrid (some remote, some in-person)
3                            Fully remote
4  Hybrid (some remote, some in-person)
5                                    NaN
6                                    NaN
7                                    NaN
8  Hybrid (some remote, some in-person)
9                            Fully remote

                                    CodingActivities  \
0                                                NaN
1           Hobby;Contribute to open-source projects
2                                              Hobby
3                         I don't code outside of work
4                                              Hobby
5                                                NaN
6                                                NaN
7                                                NaN
8                         I don't code outside of work
9  Hobby;Contribute to open-source projects;Boots…

                                              EdLevel  \
0                                                NaN
1                                                NaN
2    Master's degree (M.A., M.S., M.Eng., MBA, etc.)
3       Bachelor's degree (B.A., B.S., B.Eng., etc.)
4       Bachelor's degree (B.A., B.S., B.Eng., etc.)
5    Master's degree (M.A., M.S., M.Eng., MBA, etc.)
6  Secondary school (e.g. American high school, G…
7  Some college/university study without earning …
8    Master's degree (M.A., M.S., M.Eng., MBA, etc.)
9  Some college/university study without earning …

                                            LearnCode  \
0                                                NaN
1                                                NaN
2  Books / Physical media;Friend or family member…
3  Books / Physical media;School (i.e., Universit…
4  Other online resources (e.g., videos, blogs, f…
5  Books / Physical media;School (i.e., Universit…
6  Other online resources (e.g., videos, blogs, f…
7                    Online Courses or Certification
8                On the job training;Coding Bootcamp
9  Books / Physical media;Other online resources …

                                    LearnCodeOnline LearnCodeCoursesCert  \
0                                               NaN                   NaN
```

```
1                                                  NaN                  NaN
2  Technical documentation;Blogs;Programming Game…                  NaN
3                                                  NaN                  NaN
4  Technical documentation;Blogs;Stack Overflow;O…                  NaN
5                                                  NaN                  NaN
6           Stack Overflow;Video-based Online Courses                  NaN
7                                                  NaN       Coursera;Udemy
8                                                  NaN                  NaN
9  Technical documentation;Blogs;Written Tutorial…                  NaN

  YearsCode  …        TimeSearching           TimeAnswering     Onboarding  \
0       NaN  …                  NaN                     NaN            NaN
1       NaN  …                  NaN                     NaN            NaN
2        14  …                  NaN                     NaN            NaN
3        20  …                  NaN                     NaN            NaN
4         8  …                  NaN                     NaN            NaN
5        15  …                  NaN                     NaN            NaN
6         3  …                  NaN                     NaN            NaN
7         1  …                  NaN                     NaN            NaN
8         6  …  15-30 minutes a day  Over 120 minutes a day  Somewhat long
9        37  …                  NaN                     NaN            NaN

                                    ProfessionalTech TrueFalse_1 TrueFalse_2  \
0                                                NaN         NaN         NaN
1                                                NaN         NaN         NaN
2                                                NaN         NaN         NaN
3                                                NaN         NaN         NaN
4                                                NaN         NaN         NaN
5                                                NaN         NaN         NaN
6                                                NaN         NaN         NaN
7                                                NaN         NaN         NaN
8  Innersource initiative;DevOps function;Microse…         Yes         Yes
9                                                NaN         NaN         NaN

  TrueFalse_3           SurveyLength                  SurveyEase  \
0         NaN                    NaN                         NaN
1         NaN              Too long                   Difficult
2         NaN  Appropriate in length  Neither easy nor difficult
3         NaN  Appropriate in length                        Easy
4         NaN              Too long                        Easy
5         NaN  Appropriate in length                        Easy
6         NaN  Appropriate in length                        Easy
7         NaN  Appropriate in length                        Easy
8         Yes  Appropriate in length                        Easy
9         NaN  Appropriate in length                        Easy

  ConvertedCompYearly
```

```
0                NaN
1                NaN
2            40205.0
3           215232.0
4                NaN
5                NaN
6                NaN
7                NaN
8            49056.0
9                NaN

[10 rows x 79 columns]
```

[73]:
```
df_schema.head(10)
```

[73]:
```
      qid             qname  \
0   QID16               S0
1   QID12          MetaInfo
2    QID1               S1
3    QID2        MainBranch
4  QID296        Employment
5  QID308        RemoteWork
6  QID297  CodingActivities
7  QID190               S2
8   QID25           EdLevel
9  QID276         LearnCode

                                            question force_resp  type selector
0  <div><span style="font-size:19px;"><strong>Hel…      False    DB       TB
1                                Browser Meta Info      False  Meta  Browser
2  <span style="font-size:22px; font-family: aria…      False    DB       TB
3  Which of the following options best describes …       True    MC     SAVR
4  Which of the following best describes your cur…      False    MC     MAVR
5  Which best describes your current work situation?      False    MC     SAVR
6  Which of the following best describes the code…      False    MC     MAVR
7  <span style="font-size:22px; font-family: aria…      False    DB       TB
8  Which of the following best describes the high…      False    MC     SAVR
9  How did you learn to code? Select all that apply.      False    MC     MAVR
```

Looking at the stack-overflow survey data we see there is a `'ResponseId'` column. It is the id variable in the dataset. So setting the index to this id would be useful.

[74]:
```
df_stckov.set_index('ResponseId',inplace=True)
df_stckov
```

[74]:
```
                                                MainBranch  \
ResponseId
```

```
1                                          None of these
2                           I am a developer by profession
3          I am not primarily a developer, but I write co…
4                           I am a developer by profession
5                           I am a developer by profession
…                                                      …
73264                       I am a developer by profession
73265                       I am a developer by profession
73266      I am not primarily a developer, but I write co…
73267                       I am a developer by profession
73268      I used to be a developer by profession, but no…

                                                 Employment  \
ResponseId
1                                                       NaN
2                                       Employed, full-time
3                                       Employed, full-time
4                                       Employed, full-time
5                                       Employed, full-time
…                                                       …
73264                                   Employed, full-time
73265                                   Employed, full-time
73266                                   Employed, full-time
73267                                   Employed, full-time
73268      Independent contractor, freelancer, or self-em…

                                    RemoteWork  \
ResponseId
1                                          NaN
2                                Fully remote
3          Hybrid (some remote, some in-person)
4                                Fully remote
5          Hybrid (some remote, some in-person)
…                                          …
73264                            Fully remote
73265                           Full in-person
73266      Hybrid (some remote, some in-person)
73267      Hybrid (some remote, some in-person)
73268                            Fully remote

                                          CodingActivities  \
ResponseId
1                                                     NaN
2                  Hobby;Contribute to open-source projects
3                                                   Hobby
4                           I don’t code outside of work
5                                                   Hobby
```

```
…                                                       …
73264                                Freelance/contract work
73265                                                 Hobby
73266                          Hobby;School or academic work
73267                                                 Hobby
73268       Hobby;Contribute to open-source projects;Boots…

                                                     EdLevel  \
ResponseId
1                                                        NaN
2                                                        NaN
3           Master's degree (M.A., M.S., M.Eng., MBA, etc.)
4              Bachelor's degree (B.A., B.S., B.Eng., etc.)
5              Bachelor's degree (B.A., B.S., B.Eng., etc.)
…                                                        …
73264          Bachelor's degree (B.A., B.S., B.Eng., etc.)
73265       Master's degree (M.A., M.S., M.Eng., MBA, etc.)
73266          Bachelor's degree (B.A., B.S., B.Eng., etc.)
73267          Bachelor's degree (B.A., B.S., B.Eng., etc.)
73268          Bachelor's degree (B.A., B.S., B.Eng., etc.)

                                                    LearnCode  \
ResponseId
1                                                        NaN
2                                                        NaN
3           Books / Physical media;Friend or family member…
4           Books / Physical media;School (i.e., Universit…
5           Other online resources (e.g., videos, blogs, f…
…                                                        …
73264       Books / Physical media;Other online resources …
73265       Other online resources (e.g., videos, blogs, f…
73266       Books / Physical media;Other online resources …
73267              Books / Physical media;On the job training
73268       Books / Physical media;Friend or family member…

                                               LearnCodeOnline  \
ResponseId
1                                                        NaN
2                                                        NaN
3           Technical documentation;Blogs;Programming Game…
4                                                        NaN
5           Technical documentation;Blogs;Stack Overflow;O…
…                                                        …
73264       Technical documentation;Blogs;Written Tutorial…
73265       Technical documentation;Blogs;Written Tutorial…
73266       Technical documentation;Programming Games;Stac…
73267                                                    NaN
```

```
73268       Technical documentation;Blogs;Programming Game…

                        LearnCodeCoursesCert YearsCode YearsCodePro  …  \
ResponseId                                                           …
1                                        NaN       NaN          NaN  …
2                                        NaN       NaN          NaN  …
3                                        NaN        14            5  …
4                                        NaN        20           17  …
5                                        NaN         8            3  …
…                                          …       …          …  …
73264                                 Udemy         8            5  …
73265                 Coursera;Udemy;Udacity         6            5  …
73266       Udemy;Codecademy;Pluralsight;edX        42           33  …
73267                                    NaN        50           31  …
73268                    Udemy;Pluralsight        16            5  …

                  TimeSearching               TimeAnswering  Onboarding  \
ResponseId
1                           NaN                         NaN         NaN
2                           NaN                         NaN         NaN
3                           NaN                         NaN         NaN
4                           NaN                         NaN         NaN
5                           NaN                         NaN         NaN
…                             …                         …         …
73264       30-60 minutes a day  Less than 15 minutes a day  Just right
73265       15-30 minutes a day        60-120 minutes a day   Very long
73266       30-60 minutes a day        60-120 minutes a day  Just right
73267                       NaN                         NaN         NaN
73268                       NaN                         NaN         NaN

                                           ProfessionalTech TrueFalse_1  \
ResponseId
1                                                        NaN         NaN
2                                                        NaN         NaN
3                                                        NaN         NaN
4                                                        NaN         NaN
5                                                        NaN         NaN
…                                                          …         …
73264       DevOps function;Microservices;Developer portal…         Yes
73265                                          None of these          No
73266                                          None of these          No
73267                                                    NaN         NaN
73268                                                    NaN         NaN

           TrueFalse_2  TrueFalse_3           SurveyLength  \
ResponseId
1                  NaN          NaN                    NaN
```

```
2                  NaN          NaN                Too long
3                  NaN          NaN  Appropriate in length
4                  NaN          NaN  Appropriate in length
5                  NaN          NaN                Too long
…                  …            …                  …
73264              Yes          Yes                Too long
73265              Yes          Yes                Too long
73266               No           No  Appropriate in length
73267              NaN          NaN  Appropriate in length
73268              NaN          NaN  Appropriate in length

                              SurveyEase ConvertedCompYearly
ResponseId
1                                    NaN                NaN
2                              Difficult                NaN
3           Neither easy nor difficult             40205.0
4                                   Easy            215232.0
5                                   Easy                NaN
…                                   …                  …
73264                               Easy                NaN
73265                               Easy                NaN
73266                               Easy                NaN
73267                               Easy                NaN
73268                               Easy                NaN

[73268 rows x 78 columns]
```

OR, We could just tell pandas to use a particular index when we read the data.

[75]:
```
df_stckov=pd.read_csv('stack-overflow-developer-survey-2022/
 ↪survey_results_public.csv',index_col='ResponseId')
```

We also use a different index for the schema datataframe.

[92]:
```
df_schema=pd.read_csv('stack-overflow-developer-survey-2022/
 ↪survey_results_schema.csv',index_col='qname')
df_schema
```

[92]:
```
                qid                                              question  \
qname
S0            QID16  <div><span style="font-size:19px;"><strong>Hel…
MetaInfo      QID12                                  Browser Meta Info
S1             QID1  <span style="font-size:22px; font-family: aria…
MainBranch     QID2  Which of the following options best describes …
Employment   QID296  Which of the following best describes your cur…
…              …                                                  …
Frequency_2  QID290  Interacting with people outside of your immedi…
Frequency_3  QID290  Encountering knowledge silos (where one indivi…
```

```
TrueFalse_1  QID294  Are you involved in supporting new hires durin…
TrueFalse_2  QID294  Do you use learning resources provided by your…
TrueFalse_3  QID294  Does your employer give you time to learn new …
```

qname	force_resp	type	selector
S0	False	DB	TB
MetaInfo	False	Meta	Browser
S1	False	DB	TB
MainBranch	True	MC	SAVR
Employment	False	MC	MAVR
…	…	…	…
Frequency_2	NaN	MC	MAVR
Frequency_3	NaN	MC	MAVR
TrueFalse_1	NaN	MC	MAVR
TrueFalse_2	NaN	MC	MAVR
TrueFalse_3	NaN	MC	MAVR

```
[79 rows x 5 columns]
```

Let's use these new indices. Let's say we want to look at Respondent 3.

[77]:
```
df_stckov.iloc[3]
```

[77]:
```
MainBranch                          I am a developer by profession
Employment                                    Employed, full-time
RemoteWork                                           Fully remote
CodingActivities                       I don't code outside of work
EdLevel                Bachelor's degree (B.A., B.S., B.Eng., etc.)
                                                  …
TrueFalse_2                                                   NaN
TrueFalse_3                                                   NaN
SurveyLength                                 Appropriate in length
SurveyEase                                                   Easy
ConvertedCompYearly                                      215232.0
Name: 4, Length: 78, dtype: object
```

Maybe we are not sure what some of these columns in survey data mean. Let's say we don't know what `'MainBranch'` mean. We can use the schema to find out. But before we do that let's sort index.

[78]:
```
df_schema.sort_index(inplace=True)
df_schema
```

[78]:

qname	qid	\
Accessibility	QID124	
Age	QID127	

```
Blockchain             QID305
BuyNewTool             QID279
CodingActivities       QID297
…                        …
VersionControlSystem   QID283
Webframe               QID264
WorkExp                QID288
YearsCode               QID32
YearsCodePro            QID34

                                                                question  \
qname
Accessibility          Which of the following describe you, if any? P…
Age                                                  What is your age?
Blockchain             How favorable are you about blockchain, crypto…
BuyNewTool             When buying a new tool or software, how do you…
CodingActivities       Which of the following best describes the code…
…                                                                    …
VersionControlSystem   What are the primary <b>version control system…
Webframe               Which <b>web frameworks and web technologies</…
WorkExp                How many years of working experience do you have?
YearsCode              Including any education, how many years have y…
YearsCodePro           NOT including education, how many years have y…

                      force_resp    type selector
qname
Accessibility              False      MC     MAVR
Age                        False      MC     MAVR
Blockchain                 False      MC     SAVR
BuyNewTool                 False      MC     MAVR
CodingActivities           False      MC     MAVR
…                            …      …      …
VersionControlSystem       False      MC     MAVR
Webframe                   False  Matrix   Likert
WorkExp                    False  Slider  HSLIDER
YearsCode                  False      MC       DL
YearsCodePro               False      MC       DL

[79 rows x 5 columns]
```

[79]:
```
df_schema.loc['MainBranch']
```

[79]:
```
qid                                                        QID2
question      Which of the following options best describes …
force_resp                                                 True
type                                                         MC
selector                                                   SAVR
```

```
Name: MainBranch, dtype: object
```

Still not clear! Let's only look the `'question'`.

[80]:
```
df_schema.loc['MainBranch','question']
```

[80]:
```
'Which of the following options best describes you today? Here, by "developer"
we mean "someone who writes code." <b>*</b>'
```

1.2 Filtering Data

Let's look at the original dataframe we created with email adresses, age, etc

[81]:
```
df
```

[81]:

	email	first	last	age	degree	role
0	SakibAnwar@winchester.ac.uk	Sakib	Anwar	100	Economics	Programmer
1	JaneDoe@email.com	Jane	Doe	24	Economics	Analyst
2	JohnDoe@email.com	John	Doe	32	Management	HR

Now here if we want to look the information for individuals with the last name 'Doe'. We can *filter* the dataframe.

[82]:
```
df['last']=='Doe'
```

[82]:
```
0    False
1     True
2     True
Name: last, dtype: bool
```

Now this returns a `Series` Object.

[96]:
```
filt1= (df['last']=='Doe')
```

[97]:
```
df[filt1]
```

[97]:

	email	first	last	age	degree	role
1	JaneDoe@email.com	Jane	Doe	24	Economics	Analyst
2	JohnDoe@email.com	John	Doe	32	Management	HR

[98]:
```
filt2= (df['last']=='Doe') & (df['age']>25)
df[filt2]
```

[98]:

	email	first	last	age	degree	role
2	JohnDoe@email.com	John	Doe	32	Management	HR

[ ]:
```

```