# SI_618_Day_02_Pandas_I_Inclass

September 7, 2023

## 1 SI 618: Data Manipulation and Analysis

### 1.1 02 - Introduction to pandas

#### 1.1.1 Dr. Chris Teplovs, School of Information, University of Michigan



Version 2023.09.06.1.CT

### 1.2 Objectives:

- Know how to manipulate Series and DataFrame
- Draw a random sample of data
- Select subset of data using boolean masking
- Compute descriptive and summary statistics
- Sort a DataFrame by index or column
- Group data and calculate aggregate statistics
- Make basic plots (scatter plot, histogram, bar chart, etc)

### 1.3 Submission Instructions:

Please turn in this Jupyter notebook file (both .ipynb and .html formats) on Canvas before the end of class.

#### 1.3.1 IMPORTANT: Replace ? in the following code with your uniqname.

[ ]:
```
MY_UNIQNAME = 'yanlunar'
```

### 1.4 NumPy

Let's set up a couple of plain old python lists

[ ]:
```
names = ['Alphonso','Beata','Cal','Din','Ella']
scores = [3,5,4,4,5]
```

#### 1.4.1 Q1: Write code to iterate through the two lists to produce the following output:

```
Alphonso has a score of 3.
Beata has a score of 5.
```

```
Cal has a score of 4.
Din has a score of 4.
Ella has a score of 5.
```

Do not import any additional packages (yet).

[ ]:
```
# insert your code here
for i in range(5):
    print(f"{names[i]} has a score of {scores[i]}.")
```

```
Alphonso has a score of 3.
Beata has a score of 5.
Cal has a score of 4.
Din has a score of 4.
Ella has a score of 5.
```

## 1.5 NumPy

[ ]:
```
import numpy as np
```

[ ]:
```
ar_names = np.array(names)
ar_names
```

[ ]:
```
array(['Alphonso', 'Beata', 'Cal', 'Din', 'Ella'], dtype='<U8')
```

### 1.5.1 Q2: Create ar_scores that contains an array of the scores from above:

[ ]:
```
ar_scores = np.array(scores)
```

[ ]:
```
ar_scores
```

[ ]:
```
array([3, 5, 4, 4, 5])
```

Now, let's say we wanted to modify the scores by multiplying each one by 1.25.

### 1.5.2 Q3: Write some code that would do that using plain old python (i.e. without using pandas, numpy, etc.)

[ ]:
```
# insert your code here
new_scores = list( )
for i in scores:
    new_scores.append(i*1.25)
new_scores
```

[ ]:
```
[3.75, 6.25, 5.0, 5.0, 6.25]
```

## 1.6 ufuncs

We can use ufuncs to multiply each score by 1.25:

[ ]:
```
modified_scores = ar_scores * 1.25
modified_scores
```

[ ]:
```
array([3.75, 6.25, 5.  , 5.  , 6.25])
```

[ ]:
```
modified_scores
```

[ ]:
```
array([3.75, 6.25, 5.  , 5.  , 6.25])
```

#### 1.6.1 Q4: write code to create a new array called sqrt_scores that contains the square roots of each of the original scores

[ ]:
```
# insert your code here
# %%timeit
sqrt_scores=np.sqrt(ar_scores)
sqrt_scores
```

[ ]:
```
array([1.73205081, 2.23606798, 2.        , 2.        , 2.23606798])
```

[ ]:
```
%%timeit
ar_scores**0.5
```

```
1.31 µs ± 56.3 ns per loop (mean ± std. dev. of 7 runs, 1,000,000 loops each)
```

### 1.7 pd.Series

[ ]:
```
import pandas as pd
```

[ ]:
```
s_names = pd.Series(names)
```

[ ]:
```
s_names
```

[ ]:
```
0    Alphonso
1       Beata
2         Cal
3         Din
4        Ella
dtype: object
```

[ ]:
```
s_scores = pd.Series(scores)
s_scores
```

[ ]:
```
0    3
1    5
2    4
3    4
4    5
dtype: int64
```

[ ]:
```
names # just to remind ourselves what names looks like
```

[ ]:
```
['Alphonso', 'Beata', 'Cal', 'Din', 'Ella']
```

[ ]:
```
s_scores = pd.Series(scores,index=names)
s_scores
```

[ ]:
```
Alphonso    3
Beata       5
Cal         4
Din         4
Ella        5
dtype: int64
```

### 1.8 pd.DataFrame

[ ]:
```
df = pd.DataFrame({"name": names, "score": scores})
```

[ ]:
```
df
```

[ ]:
```
       name  score
0  Alphonso      3
1     Beata      5
2       Cal      4
3       Din      4
4      Ella      5
```

[ ]:
```
specializations = ['DS', 'UX', 'UX', 'DS', 'DS']
```

[ ]:
```
df['specialization'] = specializations
df
```

[ ]:
```
       name  score specialization
0  Alphonso      3             DS
1     Beata      5             UX
2       Cal      4             UX
3       Din      4             DS
4      Ella      5             DS
```

Let's say we wanted to set the "name" column to be the index:

[ ]:
```
df.set_index("name")
```

[ ]:
```
           score specialization
name
Alphonso       3             DS
Beata          5             UX
Cal            4             UX
```

```
Din           4             DS
Ella          5             DS
```

```
[ ]: df_indexed_by_name = df.set_index("name")
```

```
[ ]: df_indexed_by_name
```

```
[ ]:          score specialization
name
Alphonso      3             DS
Beata         5             UX
Cal           4             UX
Din           4             DS
Ella          5             DS
```

```
[ ]: df.set_index("name",inplace = True) # equivalent to df = df.set_index("name")
```

```
[ ]: df
```

```
[ ]:          score specialization
name
Alphonso      3             DS
Beata         5             UX
Cal           4             UX
Din           4             DS
Ella          5             DS
```

```
[ ]: df.describe().T
```

```
[ ]:        count  mean      std  min  25%  50%  75%  max
score     5.0   4.2  0.83666  3.0  4.0  4.0  5.0  5.0
```

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

```
[ ]:        name  score specialization
0  Alphonso      3             DS
1     Beata      5             UX
2       Cal      4             UX
3       Din      4             DS
4      Ella      5             DS
```

## 2 Part 1 (as a group): Mental Health Disorders In the Tech Workplace

From https://www.kaggle.com/osmi/mental-health-in-tech-survey

## 2.1 Data Description

This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.

## 2.2 Metadata

| Field | Description |
|---|---|
| **Timestamp** | |
| **Age** | |
| **Gender** | |
| **Country** | |
| **state** | If you live in the United States, which state or territory do you live in? |
| **self_employed** | Are you self-employed? |
| **family_history** | Do you have a family history of mental illness? |
| **treatment** | Have you sought treatment for a mental health condition? |
| **work_interfere** | If you have a mental health condition, do you feel that it interferes with your work? |
| **no_employees** | How many employees does your company or organization have? |
| **remote_work** | Do you work remotely (outside of an office) at least 50% of the time? |
| **tech_company** | Is your employer primarily a tech company/organization? |
| **benefits** | Does your employer provide mental health benefits? |
| **care_options** | Do you know the options for mental health care your employer provides? |
| **wellness_program** | Has your employer ever discussed mental health as part of an employee wellness program? |
| **seek_help** | Does your employer provide resources to learn more about mental health issues and how to seek help? |
| **anonymity** | Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources? |
| **leave** | How easy is it for you to take medical leave for a mental health condition? |
| **mental_health_consequence** | Do you think that discussing a mental health issue with your employer would have negative consequences? |
| **phys_health_consequence** | Do you think that discussing a physical health issue with your employer would have negative consequences? |
| **coworkers** | Would you be willing to discuss a mental health issue with your coworkers? |

| Field | Description |
|---|---|
| **supervisor** | Would you be willing to discuss a mental health issue with your direct supervisor(s)? |
| **mental_health_interview** | Would you bring up a mental health issue with a potential employer in an interview? |
| **phys_health_interview** | Would you bring up a physical health issue with a potential employer in an interview? |
| **mental_vs_physical** | Do you feel that your employer takes mental health as seriously as physical health? |
| **obs_consequence** | Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace? |
| **comments** | Any additional notes or comments |

Let's load the usual libraries and also ask for plots to be rendered inside the notebook:

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

Then read the CSV file into a DataFrame:

[ ]:
```
df = pd.read_csv("https://raw.githubusercontent.com/umsi-data-science/data/main/
↪survey.csv")
```

It's common to look at the resulting DataFrame using .head()

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

[ ]:
```
             Timestamp  Age  Gender         Country state self_employed  \
0  2014-08-27 11:29:31   37  Female   United States    IL           NaN
1  2014-08-27 11:29:37   44       M   United States    IN           NaN
2  2014-08-27 11:29:44   32    Male          Canada   NaN           NaN
3  2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
4  2014-08-27 11:30:22   31    Male   United States    TX           NaN

  family_history treatment work_interfere    no_employees  …  \
0             No       Yes          Often            6-25  …
1             No        No         Rarely  More than 1000  …
2             No        No         Rarely            6-25  …
3            Yes       Yes          Often          26-100  …
4             No        No          Never         100-500  …

               leave mental_health_consequence phys_health_consequence  \
0       Somewhat easy                       No                       No
1          Don't know                    Maybe                       No
2  Somewhat difficult                       No                       No
3  Somewhat difficult                      Yes                      Yes
```

```
4          Don't know                          No                          No

      coworkers supervisor mental_health_interview phys_health_interview  \
0  Some of them        Yes                       No                  Maybe
1            No         No                       No                     No
2           Yes        Yes                      Yes                    Yes
3  Some of them         No                    Maybe                  Maybe
4  Some of them        Yes                      Yes                    Yes

  mental_vs_physical obs_consequence comments
0                Yes              No      NaN
1         Don't know              No      NaN
2                 No              No      NaN
3                 No             Yes      NaN
4         Don't know              No      NaN

[5 rows x 27 columns]
```

If you want to look at a random sample, you can use .sample()

[ ]:
```
df.sample(5)
```

[ ]:
```
                Timestamp  Age  Gender        Country state self_employed  \
261  2014-08-27 13:52:05   27    Male         Canada   NaN            No
227  2014-08-27 13:19:40   34    Male      Australia   NaN           Yes
90   2014-08-27 12:12:47   31    Male  United States    NY            No
724  2014-08-28 10:18:34   33    Male  United States    WY            No
479  2014-08-27 16:17:05   30  female  United States    NY            No

    family_history treatment work_interfere no_employees  …  \
261             No        No      Sometimes         6-25  …
227             No        No      Sometimes          1-5  …
90              No        No          Never     500-1000  …
724            Yes       Yes      Sometimes          1-5  …
479             No        No            NaN      100-500  …

                 leave mental_health_consequence phys_health_consequence  \
261      Very difficult                      Yes                     Yes
227  Somewhat difficult                      Yes                     Yes
90        Somewhat easy                       No                      No
724          Don't know                       No                      No
479          Don't know                    Maybe                      No

        coworkers    supervisor mental_health_interview phys_health_interview  \
261  Some of them  Some of them                      No                  Maybe
227            No            No                      No                  Maybe
90   Some of them           Yes                      No                     No
724  Some of them           Yes                   Maybe                  Maybe
```

```
479  Some of them          Yes                       No                    Maybe

    mental_vs_physical obs_consequence comments
261                 No              No      NaN
227                Yes              No      NaN
90                 Yes              No      NaN
724         Don't know              No      NaN
479         Don't know              No      NaN

[5 rows x 27 columns]
```

Finally, you can get some basic information about the size and shape of the DataFrame:

[ ]:
```
print("The number of rows of the dataset is: ", len(df))
print("The number of columns of the dataset is: ", len(df.columns))
print("The shape of the dataset is: ", df.shape)
```

```
The number of rows of the dataset is:  1259
The number of columns of the dataset is:  27
The shape of the dataset is:  (1259, 27)
```

You can list the columns:

[ ]:
```
df.columns
```

[ ]:
```
Index(['Timestamp', 'Age', 'Gender', 'Country', 'state', 'self_employed',
       'family_history', 'treatment', 'work_interfere', 'no_employees',
       'remote_work', 'tech_company', 'benefits', 'care_options',
       'wellness_program', 'seek_help', 'anonymity', 'leave',
       'mental_health_consequence', 'phys_health_consequence', 'coworkers',
       'supervisor', 'mental_health_interview', 'phys_health_interview',
       'mental_vs_physical', 'obs_consequence', 'comments'],
      dtype='object')
```

And you can extract one or more columns. The following pair of commands do exactly the same thing:

[ ]:
```
print(df['Country'])
```

```
0         United States
1         United States
2                Canada
3        United Kingdom
4         United States
              ...
1254     United Kingdom
1255      United States
1256      United States
1257      United States
```

```
1258     United States
Name: Country, Length: 1259, dtype: object
```

[ ]:
```
country_state = df[['Country', 'state']]
country_state.head()
```

[ ]:
```
          Country state
0   United States    IL
1   United States    IN
2          Canada   NaN
3  United Kingdom   NaN
4   United States    TX
```

### 2.3 Extracting rows

[ ]:
```
df.iloc[0]
```

[ ]:
```
Timestamp                    2014-08-27 11:29:31
Age                                          37
Gender                                   Female
Country                           United States
state                                        IL
self_employed                               NaN
family_history                               No
treatment                                   Yes
work_interfere                            Often
no_employees                               6-25
remote_work                                  No
tech_company                                Yes
benefits                                    Yes
care_options                          Not sure
wellness_program                             No
seek_help                                   Yes
anonymity                                   Yes
leave                             Somewhat easy
mental_health_consequence                    No
phys_health_consequence                      No
coworkers                         Some of them
supervisor                                  Yes
mental_health_interview                      No
phys_health_interview                     Maybe
mental_vs_physical                          Yes
obs_consequence                              No
comments                                    NaN
Name: 0, dtype: object
```

[ ]:
```
df.loc[0]
```

[ ]:
```
Timestamp                    2014-08-27 11:29:31
Age                                          37
Gender                                   Female
Country                           United States
state                                        IL
self_employed                               NaN
family_history                               No
treatment                                   Yes
work_interfere                            Often
no_employees                               6-25
remote_work                                  No
tech_company                                Yes
benefits                                    Yes
care_options                           Not sure
wellness_program                             No
seek_help                                   Yes
anonymity                                   Yes
leave                             Somewhat easy
mental_health_consequence                    No
phys_health_consequence                      No
coworkers                          Some of them
supervisor                                  Yes
mental_health_interview                      No
phys_health_interview                     Maybe
mental_vs_physical                          Yes
obs_consequence                              No
comments                                    NaN
Name: 0, dtype: object
```

[ ]:
```
df.head(1)
```

[ ]:
```
             Timestamp  Age  Gender        Country state self_employed  \
0  2014-08-27 11:29:31   37  Female  United States    IL           NaN

  family_history treatment work_interfere no_employees  …          leave  \
0             No       Yes          Often         6-25  …  Somewhat easy

  mental_health_consequence phys_health_consequence     coworkers supervisor  \
0                        No                      No  Some of them        Yes

  mental_health_interview phys_health_interview mental_vs_physical  \
0                      No                Maybe               Yes

  obs_consequence comments
0              No      NaN

[1 rows x 27 columns]
```

[ ]:
```
df_gender = df.set_index('Gender')
```

[ ]:
```
df_gender.head()
```

[ ]:
```
                  Timestamp  Age         Country state self_employed  \
Gender
Female  2014-08-27 11:29:31   37   United States    IL           NaN
M       2014-08-27 11:29:37   44   United States    IN           NaN
Male    2014-08-27 11:29:44   32          Canada   NaN           NaN
Male    2014-08-27 11:29:46   31  United Kingdom   NaN           NaN
Male    2014-08-27 11:30:22   31   United States    TX           NaN

        family_history treatment work_interfere    no_employees remote_work  \
Gender
Female             No       Yes          Often            6-25          No
M                  No        No         Rarely  More than 1000          No
Male               No        No         Rarely            6-25          No
Male              Yes       Yes          Often          26-100          No
Male               No        No          Never         100-500         Yes

         …               leave mental_health_consequence  \
Gender   …
Female   …       Somewhat easy                         No
M        …          Don't know                      Maybe
Male     …  Somewhat difficult                         No
Male     …  Somewhat difficult                        Yes
Male     …          Don't know                         No

        phys_health_consequence     coworkers supervisor  \
Gender
Female                      No  Some of them        Yes
M                           No            No         No
Male                        No           Yes        Yes
Male                       Yes  Some of them         No
Male                        No  Some of them        Yes

        mental_health_interview phys_health_interview mental_vs_physical  \
Gender
Female                      No                Maybe               Yes
M                           No                   No        Don't know
Male                       Yes                  Yes                No
Male                     Maybe                Maybe                No
Male                       Yes                  Yes        Don't know

        obs_consequence comments
Gender
Female              No      NaN
```

```
M                    No      NaN
Male                 No      NaN
Male                Yes      NaN
Male                 No      NaN

[5 rows x 26 columns]
```

[ ]:
```
# df_gender.loc[219] #  will throw an exception
# df_gender.iloc[219]
```

[ ]:
```
# df.iloc['Gender'] # will throw an exception
```

[ ]:
```
import traceback
try:
    df.iloc['Gender'] # generates error
except TypeError as e:
    print(traceback.format_exc())
```

```
Traceback (most recent call last):
  File "/var/folders/v2/8xbh4n71287gwfzxz784538c0000gn/T/ipykernel_15315/2787970
466.py", line 3, in <module>
    df.iloc['Gender'] # generates error
  File "/Users/luyan/Documents/UM/23Fall/si618/.venv/lib/python3.10/site-
packages/pandas/core/indexing.py", line 1103, in __getitem__
    return self._getitem_axis(maybe_callable, axis=axis)
  File "/Users/luyan/Documents/UM/23Fall/si618/.venv/lib/python3.10/site-
packages/pandas/core/indexing.py", line 1653, in _getitem_axis
    raise TypeError("Cannot index by location index with a non-integer key")
TypeError: Cannot index by location index with a non-integer key
```

[ ]:
```
df.iloc[0]
```

[ ]:
```
Timestamp                   2014-08-27 11:29:31
Age                                         37
Gender                                  Female
Country                          United States
state                                       IL
self_employed                              NaN
family_history                              No
treatment                                  Yes
work_interfere                           Often
no_employees                              6-25
remote_work                                 No
tech_company                               Yes
benefits                                   Yes
care_options                           Not sure
wellness_program                            No
```

```
seek_help                                  Yes
anonymity                                  Yes
leave                               Somewhat easy
mental_health_consequence                     No
phys_health_consequence                       No
coworkers                           Some of them
supervisor                                 Yes
mental_health_interview                       No
phys_health_interview                      Maybe
mental_vs_physical                         Yes
obs_consequence                               No
comments                                     NaN
Name: 0, dtype: object
```

### 2.4 Sorting

You can use either sort_values() or sort_index():

[ ]:
```
df_sorted = df.sort_values('Age')
df_sorted.tail(10)
```

[ ]:
```
                Timestamp          Age  Gender        Country state  \
466   2014-08-27 16:06:46           57       M  United States    CA
471   2014-08-27 16:13:40           58    Male  United States    CA
1236  2015-02-24 10:32:32           60    Male  United States    CA
297   2014-08-27 14:18:41           60    male  United States    CA
952   2014-08-29 01:20:32           61    male   South Africa   NaN
520   2014-08-27 17:12:01           62       M  United States    CA
560   2014-08-27 19:17:07           65    Male  United States    FL
1182  2014-10-02 21:25:16           72  Female  United States    IN
364   2014-08-27 15:05:21          329    Male  United States    OH
390   2014-08-27 15:24:47  99999999999     All       Zimbabwe   NaN

     self_employed family_history treatment work_interfere    no_employees  \
466             No            Yes       Yes         Rarely  More than 1000
471             No             No       Yes         Rarely  More than 1000
1236            No             No       Yes          Often  More than 1000
297             No             No        No            NaN  More than 1000
952            Yes             No       Yes      Sometimes             1-5
520             No             No        No          Never  More than 1000
560            Yes             No        No            NaN            6-25
1182            No            Yes       Yes          Never        500-1000
364             No             No       Yes          Often            6-25
390            Yes            Yes       Yes          Often             1-5

      …           leave mental_health_consequence phys_health_consequence  \
466   …      Don't know                     Maybe                       No
```

```
471   …   Somewhat easy                    Maybe                      No
1236  …   Somewhat easy                    Maybe                   Maybe
297   …      Don't know                      No                      No
952   …  Very difficult                     Yes                   Maybe
520   …      Don't know                    Maybe                      No
560   …       Very easy                    Maybe                      No
1182  …   Somewhat easy                    Maybe                   Maybe
364   …      Don't know                    Maybe                      No
390   …  Very difficult                     Yes                      Yes

         coworkers supervisor mental_health_interview phys_health_interview  \
466   Some of them        Yes                      No                   Maybe
471   Some of them        Yes                      No                     Yes
1236  Some of them         No                      No                   Maybe
297   Some of them        Yes                   Maybe                   Maybe
952   Some of them        Yes                      No                   Maybe
520   Some of them        Yes                   Maybe                   Maybe
560   Some of them         No                      No                      No
1182  Some of them        Yes                      No                      No
364   Some of them         No                      No                      No
390             No         No                     Yes                      No

     mental_vs_physical obs_consequence comments
466                  No              No      NaN
471                 Yes              No      NaN
1236         Don't know              No      NaN
297                 Yes              No      NaN
952                  No             Yes      NaN
520                 Yes              No      NaN
560                 Yes              No      NaN
1182         Don't know             Yes      NaN
364                  No              No      NaN
390                  No             Yes      NaN

[10 rows x 27 columns]
```

### 2.5 Filtering using Boolean Masking

[ ]:
```
df.Age
```

[ ]:
```
0       37
1       44
2       32
3       31
4       31
        ..
1254    26
```

```
1255    32
1256    34
1257    46
1258    25
Name: Age, Length: 1259, dtype: int64
```

[ ]:
```
df['Age'] > 40
```

[ ]:
```
0       False
1        True
2       False
3       False
4       False
        …
1254    False
1255    False
1256    False
1257     True
1258    False
Name: Age, Length: 1259, dtype: bool
```

[ ]:
```
df[df['Age'] > 0]
```

[ ]:
```
                 Timestamp  Age  Gender         Country state self_employed  \
0      2014-08-27 11:29:31   37  Female   United States    IL           NaN
1      2014-08-27 11:29:37   44       M   United States    IN           NaN
2      2014-08-27 11:29:44   32    Male          Canada   NaN           NaN
3      2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
4      2014-08-27 11:30:22   31    Male   United States    TX           NaN
…                      …  …     …             …   …           …
1254   2015-09-12 11:17:21   26    male  United Kingdom   NaN            No
1255   2015-09-26 01:07:35   32    Male   United States    IL            No
1256   2015-11-07 12:36:58   34    male   United States    CA            No
1257   2015-11-30 21:25:06   46       f   United States    NC            No
1258   2016-02-01 23:04:31   25    Male   United States    IL            No

     family_history treatment work_interfere    no_employees  …  \
0                No       Yes          Often            6-25  …
1                No        No         Rarely  More than 1000  …
2                No        No         Rarely            6-25  …
3               Yes       Yes          Often          26-100  …
4                No        No          Never         100-500  …
…               …         …            …              …  …
1254             No       Yes            NaN          26-100  …
1255            Yes       Yes          Often          26-100  …
1256            Yes       Yes      Sometimes  More than 1000  …
1257             No        No            NaN         100-500  …
```

```
1258            Yes       Yes      Sometimes          26-100  …

                  leave mental_health_consequence phys_health_consequence  \
0          Somewhat easy                        No                       No
1             Don't know                     Maybe                       No
2     Somewhat difficult                        No                       No
3     Somewhat difficult                       Yes                      Yes
4             Don't know                        No                       No
…                    …                         …                        …
1254       Somewhat easy                        No                       No
1255  Somewhat difficult                        No                       No
1256  Somewhat difficult                       Yes                      Yes
1257          Don't know                       Yes                       No
1258          Don't know                     Maybe                       No

         coworkers    supervisor mental_health_interview  \
0     Some of them           Yes                       No
1               No            No                       No
2              Yes           Yes                      Yes
3     Some of them            No                    Maybe
4     Some of them           Yes                      Yes
…              …             …                        …
1254  Some of them  Some of them                       No
1255  Some of them           Yes                       No
1256            No            No                       No
1257            No            No                       No
1258  Some of them            No                       No

     phys_health_interview mental_vs_physical obs_consequence comments
0                    Maybe                Yes              No      NaN
1                       No         Don't know              No      NaN
2                      Yes                 No              No      NaN
3                    Maybe                 No             Yes      NaN
4                      Yes         Don't know              No      NaN
…                      …                  …             …       …
1254                    No         Don't know              No      NaN
1255                    No                Yes              No      NaN
1256                    No                 No              No      NaN
1257                    No                 No              No      NaN
1258                    No         Don't know              No      NaN

[1256 rows x 27 columns]
```

[ ]:
```
df['Age'] > 40
```

[ ]:
```
0       False
1        True
```

```
2       False
3       False
4       False
        …
1254    False
1255    False
1256    False
1257     True
1258    False
Name: Age, Length: 1259, dtype: bool
```

[ ]:
```
df[df['Age'] > 40]
```

[ ]:
```
                Timestamp  Age  Gender        Country state self_employed  \
1     2014-08-27 11:29:37   44       M  United States    IN           NaN
8     2014-08-27 11:32:39   42  Female  United States    IL           NaN
12    2014-08-27 11:33:23   42  female  United States    CA           NaN
18    2014-08-27 11:34:53   46    male  United States    MD           Yes
22    2014-08-27 11:35:48   46    Male  United States    MA            No
…                     …  …     …            …   …           …
1222  2015-02-21 11:48:52   41  female    Netherlands   NaN           Yes
1236  2015-02-24 10:32:32   60    Male  United States    CA            No
1243  2015-05-05 14:22:18   43       f  United States    FL            No
1248  2015-06-25 12:24:31   41  Female  United States    WA            No
1257  2015-11-30 21:25:06   46       f  United States    NC            No

     family_history treatment work_interfere    no_employees  …  \
1                No        No         Rarely  More than 1000  …
8               Yes       Yes      Sometimes         100-500  …
12              Yes       Yes      Sometimes          26-100  …
18              Yes        No      Sometimes             1-5  …
22               No       Yes          Often          26-100  …
…             …       …            …             …  …
1222             No       Yes         Rarely             1-5  …
1236             No       Yes          Often  More than 1000  …
1243            Yes       Yes         Rarely  More than 1000  …
1248            Yes       Yes      Sometimes          26-100  …
1257             No        No            NaN         100-500  …

                  leave mental_health_consequence phys_health_consequence  \
1             Don't know                    Maybe                      No
8         Very difficult                    Maybe                      No
12    Somewhat difficult                      Yes                     Yes
18             Very easy                       No                      No
22             Don't know                    Maybe                      No
…                     …                       …                       …
1222       Somewhat easy                       No                      No
```

```
1236       Somewhat easy                      Maybe                     Maybe
1243          Don't know                        No                        No
1248          Don't know                       Yes                     Maybe
1257          Don't know                       Yes                        No

         coworkers supervisor mental_health_interview phys_health_interview  \
1               No         No                       No                      No
8              Yes        Yes                       No                   Maybe
12             Yes        Yes                    Maybe                   Maybe
18             Yes        Yes                       No                     Yes
22    Some of them        Yes                       No                   Maybe
…              …          …                         …                       …
1222           Yes        Yes                      Yes                     Yes
1236  Some of them         No                       No                   Maybe
1243  Some of them        Yes                       No                      No
1248            No         No                       No                      No
1257            No         No                       No                      No

     mental_vs_physical obs_consequence  \
1            Don't know              No
8                    No              No
12                   No             Yes
18                  Yes             Yes
22                   No              No
…                   …               …
1222                Yes              No
1236         Don't know              No
1243         Don't know              No
1248         Don't know              No
1257                 No              No

                                                comments
1                                                    NaN
8                                                    NaN
12                                                   NaN
18                                                   NaN
22                                                   NaN
…                                                    …
1222  The data will be skewed for self-employed peop…
1236                                                 NaN
1243                                                 NaN
1248                                                 NaN
1257                                                 NaN

[150 rows x 27 columns]
```

### 2.5.1 Example: Find people who reported a family history of mental health conditions.

Solution:

```
[ ]: df[df.family_history == 'Yes']
```

```
[ ]:                Timestamp  Age  Gender         Country state self_employed  \
3     2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
5     2014-08-27 11:31:22   33    Male   United States    TN           NaN
6     2014-08-27 11:31:50   35  Female   United States    MI           NaN
8     2014-08-27 11:32:39   42  Female   United States    IL           NaN
12    2014-08-27 11:33:23   42  female   United States    CA           NaN
…                     …  …     …              …   …           …
1252  2015-08-20 16:52:09   29    male   United States    NC            No
1253  2015-08-25 19:59:38   36    Male   United States    UT            No
1255  2015-09-26 01:07:35   32    Male   United States    IL            No
1256  2015-11-07 12:36:58   34    male   United States    CA            No
1258  2016-02-01 23:04:31   25    Male   United States    IL            No

     family_history treatment work_interfere    no_employees  …  \
3               Yes       Yes          Often          26-100  …
5               Yes        No      Sometimes            6-25  …
6               Yes       Yes      Sometimes             1-5  …
8               Yes       Yes      Sometimes         100-500  …
12              Yes       Yes      Sometimes          26-100  …
…               …         …              …              …  …
1252            Yes       Yes      Sometimes         100-500  …
1253            Yes        No         Rarely  More than 1000  …
1255            Yes       Yes          Often          26-100  …
1256            Yes       Yes      Sometimes  More than 1000  …
1258            Yes       Yes      Sometimes          26-100  …

                   leave mental_health_consequence phys_health_consequence  \
3     Somewhat difficult                       Yes                     Yes
5             Don't know                        No                      No
6     Somewhat difficult                     Maybe                   Maybe
8         Very difficult                     Maybe                      No
12    Somewhat difficult                       Yes                     Yes
…                     …                       …                       …
1252          Don't know                       Yes                      No
1253       Somewhat easy                     Maybe                   Maybe
1255  Somewhat difficult                        No                      No
1256  Somewhat difficult                       Yes                     Yes
1258          Don't know                     Maybe                      No

          coworkers    supervisor mental_health_interview  \
3     Some of them            No                    Maybe
```

```
5              Yes           Yes                      No
6     Some of them            No                      No
8              Yes           Yes                      No
12             Yes           Yes                   Maybe
…              …             …                        …
1252  Some of them            No                      No
1253  Some of them  Some of them                      No
1255  Some of them           Yes                      No
1256            No            No                      No
1258  Some of them            No                      No

     phys_health_interview mental_vs_physical obs_consequence comments
3                    Maybe                 No             Yes      NaN
5                    Maybe        Don't know              No      NaN
6                       No        Don't know              No      NaN
8                    Maybe                 No              No      NaN
12                   Maybe                 No             Yes      NaN
…                    …                  …              …       …
1252                 Maybe                 No              No      NaN
1253                    No        Don't know              No      NaN
1255                    No               Yes              No      NaN
1256                    No                No              No      NaN
1258                    No        Don't know              No      NaN

[492 rows x 27 columns]
```

You can use a simple expression like `df[df['family_history'] == 'Yes']` or you can make more complex boolean expressions using parentheses:

[ ]:
```
df_filtered = df[(df['family_history'] != 'No') & (df['treatment'] == 'Yes')]
df_filtered.head()
```

[ ]:
```
              Timestamp  Age  Gender         Country state self_employed  \
3   2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
6   2014-08-27 11:31:50   35  Female   United States    MI           NaN
8   2014-08-27 11:32:39   42  Female   United States    IL           NaN
12  2014-08-27 11:33:23   42  female   United States    CA           NaN
15  2014-08-27 11:34:00   29  female   United States    IL           NaN

   family_history treatment work_interfere no_employees  …  \
3             Yes       Yes          Often       26-100  …
6             Yes       Yes      Sometimes          1-5  …
8             Yes       Yes      Sometimes      100-500  …
12            Yes       Yes      Sometimes       26-100  …
15            Yes       Yes         Rarely       26-100  …

                 leave mental_health_consequence phys_health_consequence  \
3   Somewhat difficult                       Yes                      Yes
```

```
6   Somewhat difficult                    Maybe                    Maybe
8       Very difficult                    Maybe                       No
12  Somewhat difficult                      Yes                      Yes
15       Somewhat easy                       No                       No

       coworkers    supervisor mental_health_interview phys_health_interview  \
3   Some of them            No                   Maybe                   Maybe
6   Some of them            No                      No                      No
8            Yes           Yes                      No                   Maybe
12           Yes           Yes                   Maybe                   Maybe
15           Yes  Some of them                   Maybe                   Maybe

   mental_vs_physical obs_consequence  \
3                  No             Yes
6          Don't know              No
8                  No              No
12                 No             Yes
15         Don't know              No

                                              comments
3                                                  NaN
6                                                  NaN
8                                                  NaN
12                                                 NaN
15  I have chronic low-level neurological issues t…

[5 rows x 27 columns]
```

[ ]:
```
df.coworkers.value_counts()
```

[ ]:
```
coworkers
Some of them    774
No              260
Yes             225
Name: count, dtype: int64
```

### 2.5.2 Q5: How many people are willing to discuss a mental health issue with their supervisor or their coworkers?

[ ]:
```
df['supervisor'].value_counts()
```

[ ]:
```
supervisor
Yes             516
No              393
Some of them    350
Name: count, dtype: int64
```

[ ]:
```
# insert your code here
df[(df['supervisor'] == 'Yes') | (df['coworkers'] == 'Yes')]
```

[ ]:

```
                Timestamp  Age  Gender        Country state self_employed  \
0     2014-08-27 11:29:31   37  Female  United States    IL           NaN
2     2014-08-27 11:29:44   32    Male         Canada   NaN           NaN
4     2014-08-27 11:30:22   31    Male  United States    TX           NaN
5     2014-08-27 11:31:22   33    Male  United States    TN           NaN
8     2014-08-27 11:32:39   42  Female  United States    IL           NaN
…                     …  …     …            …   …           …
1245  2015-05-06 10:14:50   22    Male      Australia   NaN            No
1249  2015-07-22 18:57:54   30       M  United States    CA            No
1250  2015-07-27 23:25:34   30    Male  United States    CA           Yes
1251  2015-08-17 09:38:35   36    Male   South Africa   NaN            No
1255  2015-09-26 01:07:35   32    Male  United States    IL            No

     family_history treatment work_interfere no_employees  …  \
0                No       Yes          Often         6-25  …
2                No        No         Rarely         6-25  …
4                No        No          Never      100-500  …
5               Yes        No      Sometimes         6-25  …
8               Yes       Yes      Sometimes      100-500  …
…               …         …            …          …  …
1245            Yes       Yes          Often      100-500  …
1249            Yes       Yes      Sometimes       26-100  …
1250            Yes       Yes          Often       26-100  …
1251            Yes       Yes          Often      100-500  …
1255            Yes       Yes          Often       26-100  …

                   leave mental_health_consequence phys_health_consequence  \
0          Somewhat easy                        No                       No
2     Somewhat difficult                        No                       No
4             Don't know                        No                       No
5             Don't know                        No                       No
8         Very difficult                     Maybe                       No
…                     …                        …                        …
1245          Don't know                     Maybe                    Maybe
1249           Very easy                        No                       No
1250          Don't know                        No                       No
1251       Somewhat easy                        No                       No
1255  Somewhat difficult                        No                       No

         coworkers supervisor mental_health_interview phys_health_interview  \
0     Some of them        Yes                      No                 Maybe
2              Yes        Yes                     Yes                   Yes
4     Some of them        Yes                     Yes                   Yes
5              Yes        Yes                      No                 Maybe
```

```
8              Yes        Yes                       No                  Maybe
…              …          …                         …                   …
1245            No        Yes                       No                  Maybe
1249           Yes        Yes                    Maybe                  Maybe
1250  Some of them        Yes                    Maybe                  Maybe
1251  Some of them        Yes                       No                    Yes
1255  Some of them        Yes                       No                     No

     mental_vs_physical obs_consequence  \
0                   Yes              No
2                    No              No
4            Don't know              No
5            Don't know              No
8                    No              No
…                   …              …
1245         Don't know             Yes
1249                Yes              No
1250                Yes              No
1251                Yes              No
1255                Yes              No

                                                comments
0                                                    NaN
2                                                    NaN
4                                                    NaN
5                                                    NaN
8                                                    NaN
…                                                    …
1245  In australia all organisations of a certain si…
1249                                  Bipolar disorder
1250                                                 NaN
1251                                                 NaN
1255                                                 NaN

[546 rows x 27 columns]
```

### 2.5.3 Q6: Make a new DataFrame `df_millenials` with only millennials (born between 1976 and 1996). Make appropriate assumptions when constructing your filter.

[ ]:
```
df[(df.Timestamp.str[:4].astype(int) - df.Age >= 1976)]
```

[ ]:
```
             Timestamp  Age  Gender         Country state self_employed  \
0  2014-08-27 11:29:31   37  Female   United States    IL           NaN
2  2014-08-27 11:29:44   32    Male          Canada   NaN           NaN
3  2014-08-27 11:29:46   31    Male  United Kingdom   NaN           NaN
4  2014-08-27 11:30:22   31    Male   United States    TX           NaN
5  2014-08-27 11:31:22   33    Male   United States    TN           NaN
```

```
…                    …  …     …              …   …           …
1253  2015-08-25 19:59:38   36    Male   United States    UT            No
1254  2015-09-12 11:17:21   26    male  United Kingdom   NaN            No
1255  2015-09-26 01:07:35   32    Male   United States    IL            No
1256  2015-11-07 12:36:58   34    male   United States    CA            No
1258  2016-02-01 23:04:31   25    Male   United States    IL            No

     family_history treatment work_interfere    no_employees  …  \
0                No       Yes          Often            6-25  …
2                No        No         Rarely            6-25  …
3               Yes       Yes          Often          26-100  …
4                No        No          Never         100-500  …
5               Yes        No      Sometimes            6-25  …
…               …       …             …              …  …
1253            Yes        No         Rarely  More than 1000  …
1254             No       Yes            NaN          26-100  …
1255            Yes       Yes          Often          26-100  …
1256            Yes       Yes      Sometimes  More than 1000  …
1258            Yes       Yes      Sometimes          26-100  …

                  leave mental_health_consequence phys_health_consequence  \
0          Somewhat easy                        No                       No
2     Somewhat difficult                        No                       No
3     Somewhat difficult                       Yes                      Yes
4             Don't know                        No                       No
5             Don't know                        No                       No
…                    …                       …                        …
1253       Somewhat easy                     Maybe                    Maybe
1254       Somewhat easy                        No                       No
1255  Somewhat difficult                        No                       No
1256  Somewhat difficult                       Yes                      Yes
1258          Don't know                     Maybe                       No

         coworkers    supervisor mental_health_interview  \
0     Some of them           Yes                       No
2              Yes           Yes                      Yes
3     Some of them            No                    Maybe
4     Some of them           Yes                      Yes
5              Yes           Yes                       No
…              …             …                        …
1253  Some of them  Some of them                       No
1254  Some of them  Some of them                       No
1255  Some of them           Yes                       No
1256            No            No                       No
1258  Some of them            No                       No

     phys_health_interview mental_vs_physical obs_consequence comments
```

```
0                    Maybe                 Yes              No      NaN
2                      Yes                  No              No      NaN
3                    Maybe                  No             Yes      NaN
4                      Yes         Don't know              No      NaN
5                    Maybe         Don't know              No      NaN
…                      …                  …       …       …
1253                    No         Don't know              No      NaN
1254                    No         Don't know              No      NaN
1255                    No                Yes              No      NaN
1256                    No                 No              No      NaN
1258                    No         Don't know              No      NaN

[1045 rows x 27 columns]
```

[ ]:
```
df_millenials = df[df['Age'] >= 25]
```

**NOTE: We will still use df for the following analysis**

### 2.6 Descriptive and Summary Statistics

Example: What is the mean age of the survey sample?

Solution:

[ ]:
```
df['Age'].mean()
```

[ ]:
```
79428148.31135821
```

#### 2.6.1 Does that look right? What should we do?

[ ]:
```
df.sort_values('Age').tail(10)
```

[ ]:
```
                Timestamp          Age  Gender        Country state  \
466   2014-08-27 16:06:46           57       M  United States    CA
471   2014-08-27 16:13:40           58    Male  United States    CA
1236  2015-02-24 10:32:32           60    Male  United States    CA
297   2014-08-27 14:18:41           60    male  United States    CA
952   2014-08-29 01:20:32           61    male   South Africa   NaN
520   2014-08-27 17:12:01           62       M  United States    CA
560   2014-08-27 19:17:07           65    Male  United States    FL
1182  2014-10-02 21:25:16           72  Female  United States    IN
364   2014-08-27 15:05:21          329    Male  United States    OH
390   2014-08-27 15:24:47  99999999999     All       Zimbabwe   NaN

     self_employed family_history treatment work_interfere    no_employees  \
466             No            Yes       Yes         Rarely  More than 1000
471             No             No       Yes         Rarely  More than 1000
1236            No             No       Yes          Often  More than 1000
297             No             No        No            NaN  More than 1000
```

```
952            Yes             No       Yes      Sometimes             1-5
520             No             No        No          Never  More than 1000
560            Yes             No        No            NaN            6-25
1182            No            Yes       Yes          Never        500-1000
364             No             No       Yes          Often            6-25
390            Yes            Yes       Yes          Often             1-5

      …           leave mental_health_consequence phys_health_consequence  \
466   …      Don't know                     Maybe                      No
471   …   Somewhat easy                     Maybe                      No
1236  …   Somewhat easy                     Maybe                   Maybe
297   …      Don't know                        No                      No
952   …  Very difficult                       Yes                   Maybe
520   …      Don't know                     Maybe                      No
560   …       Very easy                     Maybe                      No
1182  …   Somewhat easy                     Maybe                   Maybe
364   …      Don't know                     Maybe                      No
390   …  Very difficult                       Yes                      Yes

         coworkers supervisor mental_health_interview phys_health_interview  \
466   Some of them        Yes                      No                   Maybe
471   Some of them        Yes                      No                     Yes
1236  Some of them         No                      No                   Maybe
297   Some of them        Yes                   Maybe                   Maybe
952   Some of them        Yes                      No                   Maybe
520   Some of them        Yes                   Maybe                   Maybe
560   Some of them         No                      No                      No
1182  Some of them        Yes                      No                      No
364   Some of them         No                      No                      No
390             No         No                     Yes                      No

     mental_vs_physical obs_consequence comments
466                  No              No      NaN
471                 Yes              No      NaN
1236         Don't know              No      NaN
297                 Yes              No      NaN
952                  No             Yes      NaN
520                 Yes              No      NaN
560                 Yes              No      NaN
1182         Don't know             Yes      NaN
364                  No              No      NaN
390                  No             Yes      NaN

[10 rows x 27 columns]
```

#### 2.6.2 Q7: What is the *median* age of the survey sample?

```
[ ]: # insert your code here
     df['Age'].median()
```

```
[ ]: 31.0
```

#### 2.6.3 Q8: Write one line of code to compute basic statistics (mean, standard deviation, min, 25% percentile, etc) about Age

Hint: see the readings

```
[ ]: # insert your code here
     df['Age'].describe()
     # df['Age'].std()
```

```
[ ]: count    1.259000e+03
     mean     7.942815e+07
     std      2.818299e+09
     min     -1.726000e+03
     25%      2.700000e+01
     50%      3.100000e+01
     75%      3.600000e+01
     max      1.000000e+11
     Name: Age, dtype: float64
```

### 2.7 Unique Values, Counts, Membership

Example: Write one line of code to check unique values of Gender

Solution:

```
[ ]: df.coworkers.unique()
```

```
[ ]: array(['Some of them', 'No', 'Yes'], dtype=object)
```

```
[ ]: df.Gender.unique()
```

```
[ ]: array(['Female', 'M', 'Male', 'male', 'female', 'm', 'Male-ish', 'maile',
            'Trans-female', 'Cis Female', 'F', 'something kinda male?',
            'Cis Male', 'Woman', 'f', 'Mal', 'Male (CIS)', 'queer/she/they',
            'non-binary', 'Femake', 'woman', 'Make', 'Nah', 'All', 'Enby',
            'fluid', 'Genderqueer', 'Female ', 'Androgyne', 'Agender',
            'cis-female/femme', 'Guy (-ish) ^_^', 'male leaning androgynous',
            'Male ', 'Man', 'Trans woman', 'msle', 'Neuter', 'Female (trans)',
            'queer', 'Female (cis)', 'Mail', 'cis male', 'A little about you',
            'Malr', 'p', 'femail', 'Cis Man',
            'ostensibly male, unsure what that really means'], dtype=object)
```

[ ]:
```
df.Gender.value_counts()
```

[ ]:
```
Gender
Male                                615
male                                206
Female                              121
M                                   116
female                               62
F                                    38
m                                    34
f                                    15
Make                                  4
Male                                  3
Woman                                 3
Cis Male                              2
Man                                   2
Female (trans)                        2
Female                                2
Trans woman                           1
msle                                  1
male leaning androgynous              1
Neuter                                1
cis male                              1
queer                                 1
Female (cis)                          1
Mail                                  1
cis-female/femme                      1
A little about you                    1
Malr                                  1
p                                     1
femail                                1
Cis Man                               1
Guy (-ish) ^_^                        1
Enby                                  1
Agender                               1
Androgyne                             1
Male-ish                              1
maile                                 1
Trans-female                          1
Cis Female                            1
something kinda male?                 1
Mal                                   1
Male (CIS)                            1
queer/she/they                        1
non-binary                            1
Femake                                1
woman                                 1
```

```
Nah                                                    1
All                                                    1
fluid                                                  1
Genderqueer                                            1
ostensibly male, unsure what that really means         1
Name: count, dtype: int64
```

Example: Write one line of code to count the occurrences of the countries and show the top 5 countries.

Solution:

[ ]:
```
df.Country.value_counts().head(7)
```

[ ]:
```
Country
United States     751
United Kingdom    185
Canada             72
Germany            45
Ireland            27
Netherlands        27
Australia          21
Name: count, dtype: int64
```

Are you sure that's correct?

### 2.7.1 Q9: Find the unique categories of no_employees. What is the frequency of each category?

[ ]:
```
# insert your code here
df.no_employees.value_counts()
```

[ ]:
```
no_employees
6-25              290
26-100            289
More than 1000    282
100-500           176
1-5               162
500-1000           60
Name: count, dtype: int64
```

### 2.7.2 Q10: Among the people from United States, how many repondents were there from each state?

[ ]:
```
# insert your code here
df.state.value_counts().sort_index()
```

[ ]:
```
state
AL      8
```

```
AZ      7
CA    138
CO      9
CT      4
DC      4
FL     15
GA     12
IA      4
ID      1
IL     29
IN     27
KS      3
KY      5
LA      1
MA     20
MD      8
ME      1
MI     22
MN     21
MO     12
MS      1
NC     14
NE      2
NH      3
NJ      6
NM      2
NV      3
NY     57
OH     30
OK      6
OR     29
PA     29
RI      1
SC      5
SD      3
TN     45
TX     44
UT     11
VA     14
VT      3
WA     70
WI     12
WV      1
WY      2
Name: count, dtype: int64
```

## 2.8 Basic Plots

Example: Investigate the proportion (%) of people receiving health benefits from their employers.

Solution:

[ ]:
```
df.benefits
```

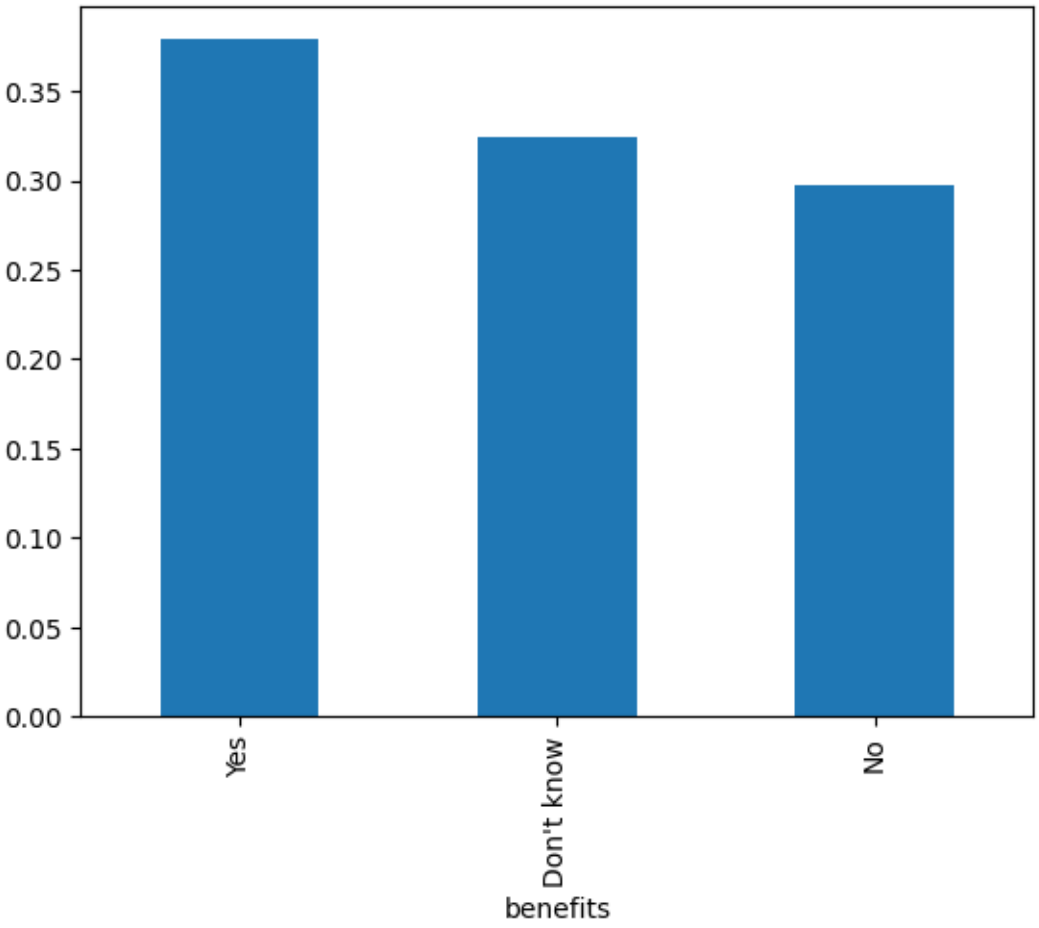
[ ]:
```
0              Yes
1       Don't know
2               No
3               No
4              Yes
           …
1254            No
1255           Yes
1256           Yes
1257            No
1258           Yes
Name: benefits, Length: 1259, dtype: object
```

[ ]:
```
df.benefits.value_counts(normalize=True)
```

[ ]:
```
benefits
Yes           0.378872
Don't know    0.324067
No            0.297061
Name: proportion, dtype: float64
```

[ ]:
```
df.benefits.value_counts(normalize=True).plot.bar()
```

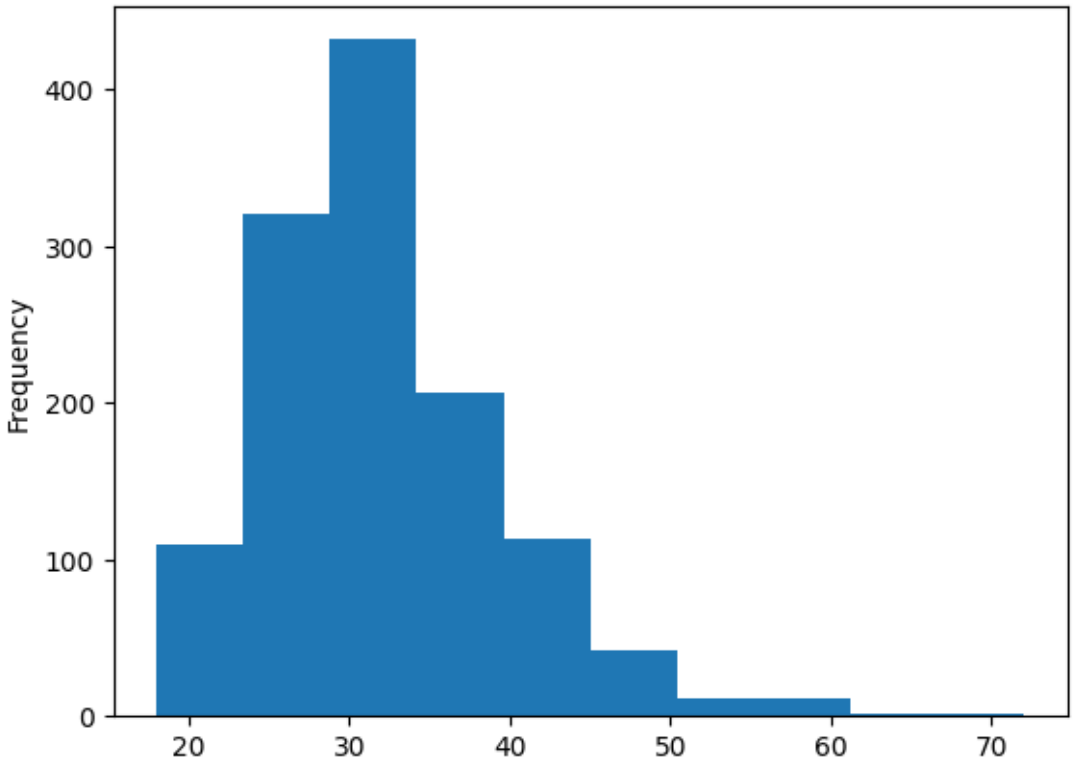
[ ]:
```
<Axes: xlabel='benefits'>
```




Example: Create a histogram of the distribution of Age values:

```
[ ]: df[(df.Age < 100) & (df.Age > 15)].Age.plot.hist()
```

```
[ ]: <Axes: ylabel='Frequency'>
```

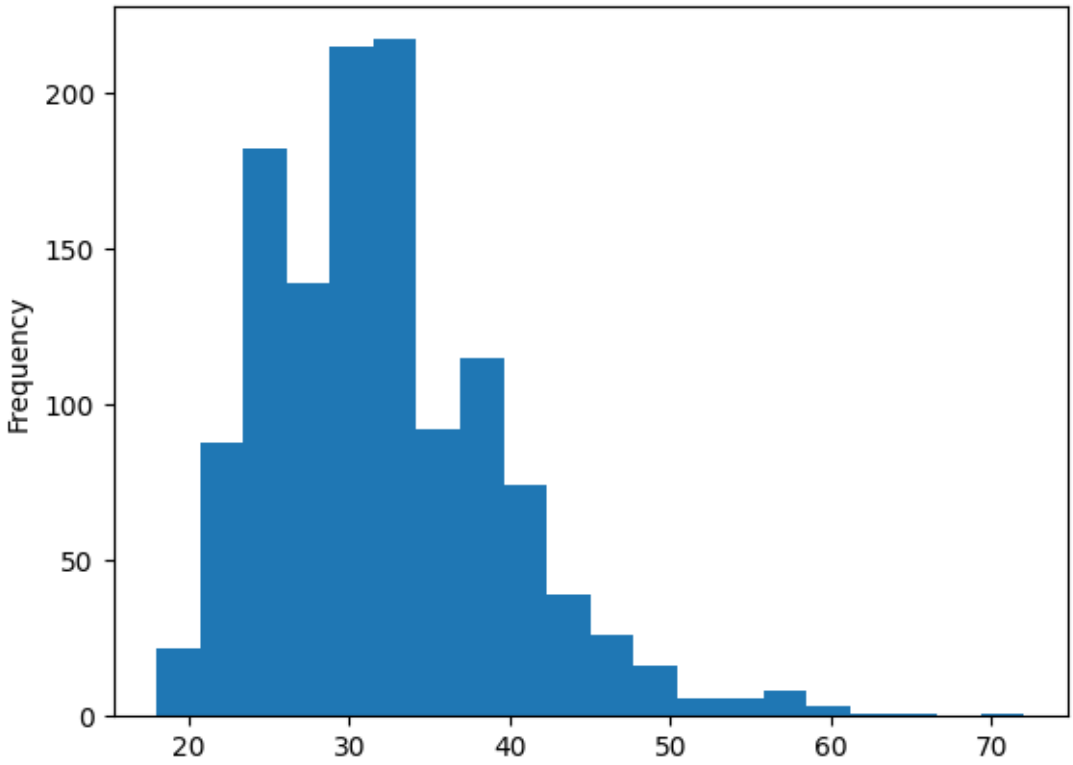



### 2.8.1 Q11: Experiment with the number of bins in the histogram of the Age distribution. Is there a "best" value?

Hint: use the bins= option to plot()

[ ]:
```
# insert your code here
df[(df.Age < 100) & (df.Age > 15)].Age.plot.hist(bins=20)
```

[ ]: <Axes: ylabel='Frequency'>




### 2.9 Aggregation

Example: Find the number of participants from each state.

Solution:

[ ]:
```
df.state.value_counts()
```

[ ]:
```
state
CA    138
WA     70
NY     57
TN     45
TX     44
OH     30
IL     29
OR     29
PA     29
IN     27
MI     22
MN     21
MA     20
FL     15
```

```
NC     14
VA     14
WI     12
GA     12
MO     12
UT     11
CO      9
MD      8
AL      8
AZ      7
OK      6
NJ      6
KY      5
SC      5
IA      4
CT      4
DC      4
NV      3
VT      3
SD      3
KS      3
NH      3
WY      2
NM      2
NE      2
WV      1
ID      1
MS      1
RI      1
LA      1
ME      1
Name: count, dtype: int64
```

[ ]:
```
df.groupby('state').size()
```

[ ]:
```
state
AL      8
AZ      7
CA    138
CO      9
CT      4
DC      4
FL     15
GA     12
IA      4
ID      1
IL     29
```

```
IN     27
KS      3
KY      5
LA      1
MA     20
MD      8
ME      1
MI     22
MN     21
MO     12
MS      1
NC     14
NE      2
NH      3
NJ      6
NM      2
NV      3
NY     57
OH     30
OK      6
OR     29
PA     29
RI      1
SC      5
SD      3
TN     45
TX     44
UT     11
VA     14
VT      3
WA     70
WI     12
WV      1
WY      2
dtype: int64
```

### 2.9.1 Q12: Find the median age of people for each state.

[ ]:
```
# insert your code here
df.groupby('state')['Age'].median()
```

[ ]:
```
state
AL    34.0
AZ    33.0
CA    31.0
CO    31.0
CT    37.5
```

```
DC    37.5
FL    34.0
GA    30.0
IA    40.0
ID    55.0
IL    30.0
IN    34.0
KS    39.0
KY    24.0
LA    35.0
MA    32.0
MD    29.0
ME    40.0
MI    34.0
MN    30.0
MO    33.5
MS    33.0
NC    32.0
NE    26.0
NH    34.0
NJ    32.0
NM    29.5
NV    28.0
NY    29.0
OH    31.0
OK    26.5
OR    32.0
PA    31.0
RI    23.0
SC    30.0
SD    34.0
TN    33.0
TX    31.5
UT    28.0
VA    40.5
VT    34.0
WA    32.5
WI    33.0
WV    23.0
WY    41.5
Name: Age, dtype: float64
```

# 3 Part 2 (on your own): Exploration of Movie Titles and Movie Cast

### 3.1 Time to load some data:

[ ]:
```
titles = pd.read_csv('https://github.com/umsi-data-science/data/raw/main/titles.
↪csv', index_col=None)
```

[ ]:
```
titles.head()
```

[ ]:
```
                    title  year
0           The Rising Son  1990
1  The Thousand Plane Raid  1969
2         Crucea de piatra  1993
3                  Country  2000
4              Gaiking II  2011
```

The titles DataFrame contains a list of movie titles and release year

[ ]:
```
cast = pd.read_csv('https://github.com/umsi-data-science/data/raw/main/cast.
↪zip', index_col=None)
```

The `cast` DataFrame contains the following columns

**title** = name of movie

**year** = year of movie

**name** = name of actor/actress

**type** = actor or actress

**character** = character name

**n** = number in the credits (NaN when not available)

[ ]:
```
titles.head()
```

[ ]:
```
                    title  year
0           The Rising Son  1990
1  The Thousand Plane Raid  1969
2         Crucea de piatra  1993
3                  Country  2000
4              Gaiking II  2011
```

[ ]:
```
cast.sample(5)
```

[ ]:
```
                         title  year              name     type  \
677634        The Grave Bandits  2012  Jack Love Falsis    actor
664960          American Tigers  1996       Joe Estevez    actor
731369   Money Isn't Everything  1918  J. Morris Foster    actor
457243                   Father  1990   Richard Cordner    actor
2996634           Julie Johnson  2001  Denise M. Kelone  actress

                 character     n
```

```
677634                Jack   NaN
664960        General Clay   NaN
731369   Henry P. Rockwell   3.0
457243      Police Officer  21.0
2996634   Friendly Student  18.0
```

### 3.1.1 Q13: How many entries are there in the titles table?

[ ]:
```
# insert your code here
len(titles)
```

[ ]:
```
232330
```

### 3.1.2 Q14: What are the two earliest movies?

[ ]:
```
# insert your code here
titles.sort_values("year").head(2)
```

[ ]:
```
                    title  year
177757          Miss Jerry  1894
215272  The Startled Lover  1898
```

### 3.1.3 Q15: How many movies have the title "Hamlet"?

[ ]:
```
# insert your code here
titles[titles['title'] == 'Hamlet']
```

[ ]:
```
         title  year
6009    Hamlet  1948
45350   Hamlet  1990
46721   Hamlet  1910
92146   Hamlet  1976
94355   Hamlet  1987
94554   Hamlet  2000
98554   Hamlet  1921
102919  Hamlet  2011
122704  Hamlet  1969
128388  Hamlet  1954
166410  Hamlet  1913
168965  Hamlet  2015
176069  Hamlet  2009
182970  Hamlet  1964
201291  Hamlet  1996
209878  Hamlet  1911
221479  Hamlet  1973
223568  Hamlet  2014
```

### 3.1.4 Q16: List all of the "Treasure Island" movies from earliest to most recent.

[ ]:
```
# insert your code here
titles[titles['title'] == 'Treasure Island'].sort_values('year')
```

[ ]:
```
                 title  year
206027  Treasure Island  1918
51287   Treasure Island  1920
191050  Treasure Island  1934
96934   Treasure Island  1950
89534   Treasure Island  1972
111343  Treasure Island  1973
205397  Treasure Island  1985
179354  Treasure Island  1999
```

### 3.1.5 Q17: What are the ten most common movie names of all time?

[ ]:
```
titles.title.value_counts().head(10)
```

[ ]:
```
title
Hamlet                  18
Carmen                  16
Macbeth                 15
The Outsider            12
Maya                    12
Temptation              11
Othello                 11
The Three Musketeers    11
Freedom                 11
Kismet                  11
Name: count, dtype: int64
```

[ ]:
```
# insert your code here
titles.groupby('title').size().sort_values(ascending=False).head(10)
```

[ ]:
```
title
Hamlet                  18
Carmen                  16
Macbeth                 15
The Outsider            12
Maya                    12
Kismet                  11
The Three Musketeers    11
Freedom                 11
Temptation              11
Othello                 11
dtype: int64
```

### 3.1.6 Stretch goals

The following questions are extra material and need not be completed as part of this notebook. We will, however, start next class by considering this material, so it's worth attempting if you have time.

### 3.1.7 EXTRA: Who are the 10 people most often credited as "Herself" in film history?

[ ]:
```
# insert code here
cast[cast['type'] == 'actress'].groupby('name').size().
 ↪sort_values(ascending=False).head(10)
```

[ ]:
```
name
Bess Flowers       819
Aruna Irani        290
Mary Gordon        282
Helen              259
Gertrude Astor     235
Leonor G?mez       227
Minerva Urecal     227
Sukumari           208
Dorothy Vernon     207
Suzanne Ridgway    205
dtype: int64
```

[ ]:
```
cast[cast['type'] == 'actress'].name.value_counts()
```

[ ]:
```
name
Bess Flowers               819
Aruna Irani                290
Mary Gordon                282
Helen                      259
Gertrude Astor             235
                          …
Paulette Kniseley            1
Laura Knirsch                1
Knirke                       1
May Cathala Knipschildt      1
Rosa ? R?gvu                 1
Name: count, Length: 492421, dtype: int64
```

### 3.1.8 EXTRA: What are the 10 most frequent roles that start with the word "Science"?

Hint: read docs on str.startswith()

[ ]:
```
# insert code here
cast_test=cast.dropna(subset=['character'], how='any')
```

```
cast_test[cast_test['character'].str.startswith("Science")].
  ↪groupby('character').size().sort_values(ascending=False).head(10)
```

[ ]:
```
character
Science Teacher         60
Science Student          9
Science Fair Student     9
Science Fair Judge       6
Science Reporter         5
Science Promo Cadet      5
Science Club Member      5
Science Kid              5
Science                  4
Science Officer          3
dtype: int64
```

#### 3.1.9 EXTRA: Comment on the differences in gender ratios for leading vs. supporting roles in the 1950s. Does there appear to be a bias?

Insert your response here.