y2vlr4job

April 11, 2023

1 Data Munging

[1]:

```
#import pandas library as pd
import pandas as pd
```

[2]:

```
#Load the data
df1 = pd.read_csv('C:\\Users\\AKSHARA RAJ␣
↪MAKWANA\\Desktop\\hinge\\interviews-data_akshara1412-main\\interviews-data_akshara1412-main
↪csv')
df1
```

[2]:

```
       first_name   last_name         dob  company_id last_active  score  \
0          Robert  Mclaughlin  1967/03/26           3  2018/08/25     57
1        Brittany      Norris  1972/09/06          12  2018/03/29     73
2          Sharon     Nichols  1971/04/19           7  2018/04/11     92
3     Christopher        Ware  1977/05/25          11  2018/07/20     74
4           Kevin       Scott  1981/12/15           8  2018/11/20     42
…             …         …         …         …         …     …
8461      Michael         Kim  1981/12/02           0  2018/03/18     92
8462         John       White  1994/10/02          13  2018/09/28     56
8463         Eric        Cook  1965/10/10           6  2018/10/08     43
8464   Alexandria       Smith  1982/07/27           6  2018/10/22     66
8465        James    Anderson  1968/01/12           7  2018/07/07     92

      member_since state
0             2013    OR
1             1986    MD
2             1985    WY
3             2003    PA
4             1994    MN
…              …   …
8461          2000    NM
8462          2016    FL
8463          1998    MN
8464          1983    NV
8465          2014    SC
```

```
[8466 rows x 8 columns]
```

[3]:
```
# Join the first name and last name columns into a name column

df1['name'] = df1['first_name'] + ' ' + df1['last_name']
df1
```

[3]:
```
       first_name   last_name         dob  company_id last_active  score  \
0          Robert  Mclaughlin  1967/03/26           3  2018/08/25     57
1        Brittany      Norris  1972/09/06          12  2018/03/29     73
2          Sharon     Nichols  1971/04/19           7  2018/04/11     92
3     Christopher        Ware  1977/05/25          11  2018/07/20     74
4           Kevin       Scott  1981/12/15           8  2018/11/20     42
…             …           …           …         …           …     …
8461      Michael         Kim  1981/12/02           0  2018/03/18     92
8462         John       White  1994/10/02          13  2018/09/28     56
8463         Eric        Cook  1965/10/10           6  2018/10/08     43
8464   Alexandria       Smith  1982/07/27           6  2018/10/22     66
8465        James    Anderson  1968/01/12           7  2018/07/07     92

      member_since state               name
0             2013    OR  Robert Mclaughlin
1             1986    MD    Brittany Norris
2             1985    WY     Sharon Nichols
3             2003    PA   Christopher Ware
4             1994    MN        Kevin Scott
…              …   …                  …
8461          2000    NM        Michael Kim
8462          2016    FL         John White
8463          1998    MN          Eric Cook
8464          1983    NV   Alexandria Smith
8465          2014    SC     James Anderson

[8466 rows x 9 columns]
```

[4]:
```
# datatype
print(df1['dob'].dtype)
```

```
object
```

[5]:
```
#converting dob & last_active into datetime datatype
df1['dob'] = pd.to_datetime(df1['dob'])
df1['last_active'] = pd.to_datetime(df1['last_active'])
```

[6]:
```
#converting dob('%m/%d/%Y) format into '%Y-%m-%d'
df1['dob'] = pd.to_datetime(df1['dob'], format='%m/%d/%Y').dt.
 ↪strftime('%Y-%m-%d')
```

```
df1['last_active'] = pd.to_datetime(df1['last_active'], format='%m/%d/%Y').dt.
 ↪strftime('%Y-%m-%d')
df1
```

```
[6]:       first_name   last_name         dob  company_id last_active  score  \
     0          Robert  Mclaughlin  1967-03-26           3  2018-08-25     57
     1        Brittany      Norris  1972-09-06          12  2018-03-29     73
     2          Sharon     Nichols  1971-04-19           7  2018-04-11     92
     3     Christopher        Ware  1977-05-25          11  2018-07-20     74
     4           Kevin      Scott  1981-12-15           8  2018-11-20     42
     …             …          …           …          …          …      …
     8461      Michael         Kim  1981-12-02           0  2018-03-18     92
     8462         John       White  1994-10-02          13  2018-09-28     56
     8463         Eric        Cook  1965-10-10           6  2018-10-08     43
     8464   Alexandria       Smith  1982-07-27           6  2018-10-22     66
     8465        James    Anderson  1968-01-12           7  2018-07-07     92

           member_since state                name
     0             2013    OR   Robert Mclaughlin
     1             1986    MD     Brittany Norris
     2             1985    WY      Sharon Nichols
     3             2003    PA   Christopher Ware
     4             1994    MN         Kevin Scott
     …                …     …                   …
     8461          2000    NM         Michael Kim
     8462          2016    FL          John White
     8463          1998    MN           Eric Cook
     8464          1983    NV   Alexandria Smith
     8465          2014    SC      James Anderson

     [8466 rows x 9 columns]
```

```
[7]: #Load the tsv data file
     #df = pd.read_csv('filename.tsv', delimiter='\t')
     df2 = pd.read_csv('C:\\Users\\AKSHARA RAJ␣
      ↪MAKWANA\\Desktop\\hinge\\interviews-data_akshara1412-main\\interviews-data_akshara1412-main
      ↪tsv', delimiter='\t')
     df2
```

```
[7]:                    name date_of_birth  company_id last_active  score  \
     0       Mikayla Brennan    11/02/1966           2  07/04/2018     84
     1         Thomas Holmes    11/29/1962           1  05/15/2018     92
     2           Corey Jones    12/20/1964           7  08/25/2018     47
     3          Laura Howard    04/26/1989           8  04/15/2018     76
     4     Daniel Mclaughlin    06/19/1966          13  05/10/2018     56
     …                     …             …           …          …      …
     7647         John Lopez    02/19/1985           5  07/31/2018     95
```

```
7648       Janice Perez    03/28/1968           4  11/25/2018     88
7649      Deborah Walls    11/22/1993          15  08/04/2018     87
7650  Michael Schneider    06/26/1997           5  08/22/2018     44
7651      Bradley Horne    07/08/1972           6  01/08/2019     92

      joined_league        us_state
0              1989        Illinois
1              1972       Wisconsin
2              2007      New Mexico
3              1976      New Jersey
4              1986    Rhode Island
…                …              …
7647           1975        Virginia
7648           1994         Vermont
7649           1994  South Carolina
7650           2007         Arizona
7651           1980            Ohio

[7652 rows x 7 columns]
```

[8]:

```
#datatype
print(df2['date_of_birth'].dtype)
```

```
object
```

[9]:

```
##converting date_of_birth & last_active into datetime datatype

df2['date_of_birth'] = pd.to_datetime(df2['date_of_birth'])
df2['last_active'] = pd.to_datetime(df2['last_active'])
```

[10]:

```
#converting date_of_birth('%m/%d/%Y) format into '%Y-%m-%d'

df2['date_of_birth'] = pd.to_datetime(df2['date_of_birth'], format='%m/%d/%Y').
 ↪dt.strftime('%Y-%m-%d')
df2['last_active'] = pd.to_datetime(df2['last_active'], format='%m/%d/%Y').dt.
 ↪strftime('%Y-%m-%d')
df2
```

[10]:

```
                   name date_of_birth  company_id last_active  score  \
0      Mikayla Brennan    1966-11-02           2  2018-07-04     84
1        Thomas Holmes    1962-11-29           1  2018-05-15     92
2          Corey Jones    1964-12-20           7  2018-08-25     47
3         Laura Howard    1989-04-26           8  2018-04-15     76
4     Daniel Mclaughlin    1966-06-19          13  2018-05-10     56
…                  …            …          …          …    …
7647        John Lopez    1985-02-19           5  2018-07-31     95
7648      Janice Perez    1968-03-28           4  2018-11-25     88
```

```
7649      Deborah Walls    1993-11-22          15  2018-08-04     87
7650  Michael Schneider    1997-06-26           5  2018-08-22     44
7651      Bradley Horne    1972-07-08           6  2019-01-08     92

      joined_league        us_state
0              1989        Illinois
1              1972       Wisconsin
2              2007      New Mexico
3              1976      New Jersey
4              1986    Rhode Island
…             …             …
7647           1975        Virginia
7648           1994         Vermont
7649           1994  South Carolina
7650           2007         Arizona
7651           1980            Ohio

[7652 rows x 7 columns]
```

[11]:

```
# Convert the state names in data file to two character abbreviations

state_codes = {
    'Alabama': 'AL',
    'Alaska': 'AK',
    'Arizona': 'AZ',
    'Arkansas': 'AR',
    'California': 'CA',
    'Colorado': 'CO',
    'Connecticut': 'CT',
    'Delaware': 'DE',
    'Florida': 'FL',
    'Georgia': 'GA',
    'Hawaii': 'HI',
    'Idaho': 'ID',
    'Illinois': 'IL',
    'Indiana': 'IN',
    'Iowa': 'IA',
    'Kansas': 'KS',
    'Kentucky': 'KY',
    'Louisiana': 'LA',
    'Maine': 'ME',
    'Maryland': 'MD',
    'Massachusetts': 'MA',
    'Michigan': 'MI',
    'Minnesota': 'MN',
    'Mississippi': 'MS',
    'Missouri': 'MO',
    'Montana': 'MT',
    'Nebraska': 'NE',
    'Nevada': 'NV',
    'New Hampshire': 'NH',
    'New Jersey': 'NJ',
    'New Mexico': 'NM',
    'New York': 'NY',
    'North Carolina': 'NC',
    'North Dakota': 'ND',
    'Ohio': 'OH',
    'Oklahoma': 'OK',
    'Oregon': 'OR',
    'Pennsylvania': 'PA',
    'Rhode Island': 'RI',
    'South Carolina': 'SC',
    'South Dakota': 'SD',
    'Tennessee': 'TN',
    'Texas': 'TX',
    'Utah': 'UT',
    'Vermont': 'VT',
    'Virginia': 'VA',
    'Washington': 'WA',
    'West Virginia': 'WV',
    'Wisconsin': 'WI',
    'Wyoming': 'WY'
}
```

[12]:

```
#using pandas map function

df2['us_state'] = df2['us_state'].map(state_codes)
```

[13]:

```
df2
```

[13]:

```
                   name date_of_birth  company_id last_active  score  \
0       Mikayla Brennan    1966-11-02           2  2018-07-04     84
1         Thomas Holmes    1962-11-29           1  2018-05-15     92
2           Corey Jones    1964-12-20           7  2018-08-25     47
3          Laura Howard    1989-04-26           8  2018-04-15     76
4     Daniel Mclaughlin    1966-06-19          13  2018-05-10     56
…                 …           …         …         …    …
7647         John Lopez    1985-02-19           5  2018-07-31     95
7648       Janice Perez    1968-03-28           4  2018-11-25     88
7649      Deborah Walls    1993-11-22          15  2018-08-04     87
7650  Michael Schneider    1997-06-26           5  2018-08-22     44
7651       Bradley Horne    1972-07-08           6  2019-01-08     92

      joined_league us_state
```

```
0              1989       IL
1              1972       WI
2              2007       NM
3              1976       NJ
4              1986       RI
…             …      …
7647           1975       VA
7648           1994       VT
7649           1994       SC
7650           2007       AZ
7651           1980       OH

[7652 rows x 7 columns]
```

[14]:
```
df1['state'] = df1['state'].map(state_codes)
df1
```

[14]:
```
       first_name   last_name         dob  company_id last_active  score  \
0          Robert  Mclaughlin  1967-03-26           3  2018-08-25     57
1        Brittany      Norris  1972-09-06          12  2018-03-29     73
2          Sharon     Nichols  1971-04-19           7  2018-04-11     92
3     Christopher        Ware  1977-05-25          11  2018-07-20     74
4           Kevin       Scott  1981-12-15           8  2018-11-20     42
…           …         …         …         …         …    …
8461      Michael         Kim  1981-12-02           0  2018-03-18     92
8462         John       White  1994-10-02          13  2018-09-28     56
8463         Eric        Cook  1965-10-10           6  2018-10-08     43
8464   Alexandria       Smith  1982-07-27           6  2018-10-22     66
8465        James    Anderson  1968-01-12           7  2018-07-07     92

      member_since state               name
0             2013   NaN  Robert Mclaughlin
1             1986   NaN    Brittany Norris
2             1985   NaN     Sharon Nichols
3             2003   NaN   Christopher Ware
4             1994   NaN        Kevin Scott
…            …   …               …
8461          2000   NaN        Michael Kim
8462          2016   NaN         John White
8463          1998   NaN          Eric Cook
8464          1983   NaN   Alexandria Smith
8465          2014   NaN     James Anderson

[8466 rows x 9 columns]
```

[15]:
```
#df1.rename(columns={'full_name': 'Name', 'dob': 'DOB', 'state': 'State',␣
 ↪'company_id': 'Company ID'},
                  #  inplace=True)
```

[16]:

```
# Merge the unity and us_softball data frames on the company_id column

o1 = pd.merge(df1, df2, on='company_id',how='outer')
o1
```

[16]:

```
        first_name   last_name         dob  company_id last_active_x  score_x  \
0           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
1           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
2           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
3           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
4           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
…            …         …         …         …           …      …
3084333     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084334     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084335     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084336     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084337     Sherry      Kelley  1967-01-28          10    2018-10-20       30

         member_since state             name_x             name_y  \
0                2013   NaN  Robert Mclaughlin       Brian Oliver
1                2013   NaN  Robert Mclaughlin     Denise Webster
2                2013   NaN  Robert Mclaughlin       Gordon Hines
3                2013   NaN  Robert Mclaughlin        Jerry Wells
4                2013   NaN  Robert Mclaughlin      Robert Nelson
…               …   …                …                …
3084333          2018   NaN      Sherry Kelley       Renee Potter
3084334          2018   NaN      Sherry Kelley  Wendy Jackson DDS
3084335          2018   NaN      Sherry Kelley      Darryl Garcia
3084336          2018   NaN      Sherry Kelley     Jesus Williams
3084337          2018   NaN      Sherry Kelley      Denise Chavez

        date_of_birth last_active_y  score_y  joined_league us_state
0          1992-03-21    2019-01-07       68           1991       ME
1          1973-03-28    2018-07-17       75           1992       SD
2          1962-07-08    2018-03-02       46           2011       MS
3          1997-06-18    2018-04-13       41           2008       TN
4          1983-03-08    2018-06-03       96           1985       AZ
…               …           …      …            …      …
3084333    1971-04-21    2018-06-08       50           1992       ME
3084334    1988-11-05    2019-01-03       87           2002       WV
3084335    1967-04-25    2018-10-17       72           2011       AZ
3084336    1984-02-11    2018-09-23       88           2007       UT
3084337    1973-05-10    2018-08-28       97           2015       GA

[3084338 rows x 15 columns]
```

2 Use companies.csv to replace company_id with the company name.

```
[17]: #Load the data file of companies
      df3 = pd.read_csv('C:\\Users\\AKSHARA RAJ␣
       ↪MAKWANA\\Desktop\\hinge\\interviews-data_akshara1412-main\\interviews-data_akshara1412-main
       ↪csv')
      df3
```

```
[17]:     id                              name
      0    0              Williams-Stephenson
      1    1         Brown, Vasquez and Sanchez
      2    2                      Keller Group
      3    3                      Mcdonald Inc
      4    4                         Bruce Inc
      5    5        Kelley, Gilbert and Jackson
      6    6                    Taylor-Alvarez
      7    7   Alvarez, Schaefer and Robertson
      8    8         Smith, Torres and Matthews
      9    9         Martin, Mcknight and Clark
      10  10                    Scott and Sons
      11  11                     Rivera-Morrow
      12  12                        Hunter Ltd
      13  13        Mullen, Huffman and Vasquez
      14  14    Jackson, Carlson and Contreras
      15  15                     Pearson Group
      16  16                      Parker Group
      17  17                Peterson and Sons
      18  18           Hopkins, Barnes and Ward
      19  19                        Rivera Ltd
```

```
[18]: df3.rename(columns={'id': 'company_id'},
                         inplace=True)
```

```
[19]: o2 = pd.merge(o1, df3, on='company_id',how='outer')
      o2
```

```
[19]:         first_name   last_name         dob  company_id last_active_x  score_x  \
      0           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
      1           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
      2           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
      3           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
      4           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
      …              …          …          …          …          …          …
      3084333     Sherry      Kelley  1967-01-28          10    2018-10-20       30
      3084334     Sherry      Kelley  1967-01-28          10    2018-10-20       30
      3084335     Sherry      Kelley  1967-01-28          10    2018-10-20       30
```

```
3084336     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084337     Sherry      Kelley  1967-01-28          10    2018-10-20       30

         member_since state            name_x             name_y  \
0                2013   NaN  Robert Mclaughlin       Brian Oliver
1                2013   NaN  Robert Mclaughlin     Denise Webster
2                2013   NaN  Robert Mclaughlin       Gordon Hines
3                2013   NaN  Robert Mclaughlin        Jerry Wells
4                2013   NaN  Robert Mclaughlin      Robert Nelson
…               …   …                …                …
3084333          2018   NaN      Sherry Kelley       Renee Potter
3084334          2018   NaN      Sherry Kelley  Wendy Jackson DDS
3084335          2018   NaN      Sherry Kelley      Darryl Garcia
3084336          2018   NaN      Sherry Kelley     Jesus Williams
3084337          2018   NaN      Sherry Kelley      Denise Chavez

        date_of_birth last_active_y  score_y  joined_league us_state  \
0          1992-03-21    2019-01-07       68           1991       ME
1          1973-03-28    2018-07-17       75           1992       SD
2          1962-07-08    2018-03-02       46           2011       MS
3          1997-06-18    2018-04-13       41           2008       TN
4          1983-03-08    2018-06-03       96           1985       AZ
…               …           …      …            …      …
3084333    1971-04-21    2018-06-08       50           1992       ME
3084334    1988-11-05    2019-01-03       87           2002       WV
3084335    1967-04-25    2018-10-17       72           2011       AZ
3084336    1984-02-11    2018-09-23       88           2007       UT
3084337    1973-05-10    2018-08-28       97           2015       GA

                   name
0          Mcdonald Inc
1          Mcdonald Inc
2          Mcdonald Inc
3          Mcdonald Inc
4          Mcdonald Inc
…                 …
3084333  Scott and Sons
3084334  Scott and Sons
3084335  Scott and Sons
3084336  Scott and Sons
3084337  Scott and Sons

[3084338 rows x 16 columns]
```

```
[21]: # Drop the company_id column from the merged data frame
      #o2.drop('first_name','last_name','company_id','state', axis=1, inplace=True)
```

```
C:\Users\AKSHARA RAJ MAKWANA\AppData\Local\Temp\ipykernel_2164\2734389604.py:2:
FutureWarning: In a future version of pandas all arguments of DataFrame.drop
except for the argument 'labels' will be keyword-only.
  o2.drop('first_name','last_name','company_id','state', axis=1, inplace=True)
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Input In [21], in <cell line: 2>()
      1 # Drop the company_id column from the merged data frame
----> 2␣
  ↪o2.drop('first_name','last_name','company_id','state', axis=1, inplace=True)

File ~\Documents\A\lib\site-packages\pandas\util\_decorators.py:311, in␣
  ↪deprecate_nonkeyword_arguments.<locals>.decorate.<locals>.wrapper(*args,␣
  ↪**kwargs)
    305 if len(args) > num_allow_args:
    306     warnings.warn(
    307         msg.format(arguments=arguments),
    308         FutureWarning,
    309         stacklevel=stacklevel,
    310     )
--> 311 return func(*args, **kwargs)

TypeError: drop() got multiple values for argument 'axis'
```

[24]:

```
# Identify bad records using boolean indexing
bad_records = o2[o2['dob'] < '2020-01-01']
bad_records
```

[24]:

```
        first_name   last_name         dob  company_id last_active_x  score_x  \
0           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
1           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
2           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
3           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
4           Robert  Mclaughlin  1967-03-26           3    2018-08-25       57
…              …          …          …          …           …         …
3084333     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084334     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084335     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084336     Sherry      Kelley  1967-01-28          10    2018-10-20       30
3084337     Sherry      Kelley  1967-01-28          10    2018-10-20       30

         member_since state             name_x          name_y  \
0                2013   NaN  Robert Mclaughlin    Brian Oliver
1                2013   NaN  Robert Mclaughlin  Denise Webster
2                2013   NaN  Robert Mclaughlin    Gordon Hines
3                2013   NaN  Robert Mclaughlin     Jerry Wells
```

```
4                2013   NaN  Robert Mclaughlin      Robert Nelson
…               …   …                …                …
3084333          2018   NaN      Sherry Kelley       Renee Potter
3084334          2018   NaN      Sherry Kelley  Wendy Jackson DDS
3084335          2018   NaN      Sherry Kelley      Darryl Garcia
3084336          2018   NaN      Sherry Kelley     Jesus Williams
3084337          2018   NaN      Sherry Kelley      Denise Chavez

        date_of_birth last_active_y  score_y  joined_league us_state  \
0          1992-03-21    2019-01-07       68           1991       ME
1          1973-03-28    2018-07-17       75           1992       SD
2          1962-07-08    2018-03-02       46           2011       MS
3          1997-06-18    2018-04-13       41           2008       TN
4          1983-03-08    2018-06-03       96           1985       AZ
…               …           …      …            …      …
3084333    1971-04-21    2018-06-08       50           1992       ME
3084334    1988-11-05    2019-01-03       87           2002       WV
3084335    1967-04-25    2018-10-17       72           2011       AZ
3084336    1984-02-11    2018-09-23       88           2007       UT
3084337    1973-05-10    2018-08-28       97           2015       GA

                   name
0          Mcdonald Inc
1          Mcdonald Inc
2          Mcdonald Inc
3          Mcdonald Inc
4          Mcdonald Inc
…                 …
3084333  Scott and Sons
3084334  Scott and Sons
3084335  Scott and Sons
3084336  Scott and Sons
3084337  Scott and Sons

[3084338 rows x 16 columns]
```

[25]:

```python
# Write bad records to a separate file
bad_records.to_csv('bad_records.csv', index=False)
```

[27]:

```python
# Drop bad records from the main data frame
o2 = o2[o2['dob'] >= '2022-01-01']
```