Data Wrangling Lab

Estimated time needed: **45 to 60** minutes

In this assignment you will be performing data wrangling.

Objectives

In this lab you will perform the following:

- Identify duplicate values in the dataset.
- Remove duplicate values from the dataset.
- Identify missing values in the dataset.
- Impute the missing values in the dataset.
- Normalize data in the dataset.

Hands on Lab

Import pandas module.

```
import pandas as pd
print('Done')

Done
```

Load the dataset into a dataframe.

The functions below will download the dataset into your browser:

```
from pyodide.http import pyfetch

async def download(url, filename):
    response = await pyfetch(url)
    if response.status == 200:
        with open(filename, "wb") as f:
            f.write(await response.bytes())
print('Done')

Done

filepath = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/Capstone_edX/Module%201/survey_results_public_2020.csv"
print('Done')
```

```
Done
```

To obtain the dataset, utilize the download() function as defined above:

```
await download(filepath, "m1_survey_data.csv")
file_name="m1_survey_data.csv"
print('Done')
```

```
Done
```

Utilize the Pandas method read_csv() to load the data into a dataframe.

```
df = pd.read_csv(file_name, header=0)
print('Done')
```

```
Done
```

> Note: This version of the lab is working on JupyterLite, which requires the dataset to be downloaded to the interface.While working on the downloaded version of this notebook on their local machines(Jupyter Anaconda), the learners can simply **skip the steps above,** and simply use the URL directly in the `pandas.read_csv()` function. You can uncomment and run the statements in the cell below.

```
#df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/Capstone_edX/
Module%201/survey_results_public_2020.csv")
```

Finding duplicates

In this section you will identify duplicate values in the dataset.

Find how many duplicate rows exist in the dataframe.

```
# your code goes
len(df)-len(df.drop_duplicates())
```

```
0
```

Removing duplicates

Remove the duplicate rows from the dataframe.

```
# your code goes here
df.drop_duplicates(subset=None, keep='first', inplace=False)
```

```
       Respondent                                         MainBranch
Hobbyist  \
0               1                      I am a developer by profession
```

```
Yes
1               2                     I am a developer by profession
No
2               3                        I code primarily as a hobby
Yes
3               4                     I am a developer by profession
Yes
4               5  I used to be a developer by profession, but no...
Yes
...           ...                                                ...
...
64456       64858                                                NaN
Yes
64457       64867                                                NaN
Yes
64458       64898                                                NaN
Yes
64459       64925                                                NaN
Yes
64460       65112                                                NaN
Yes

        Age Age1stCode CompFreq  CompTotal  ConvertedComp
Country  \
0       NaN         13  Monthly        NaN            NaN
Germany
1       NaN         19      NaN        NaN            NaN      United
Kingdom
2       NaN         15      NaN        NaN            NaN  Russian
Federation
3      25.0         18      NaN        NaN            NaN
Albania
4      31.0         16      NaN        NaN            NaN       United
States
...     ...        ...      ...        ...            ...
...
64456   NaN         16      NaN        NaN            NaN       United
States
64457   NaN        NaN      NaN        NaN            NaN
Morocco
64458   NaN        NaN      NaN        NaN            NaN
Viet Nam
64459   NaN        NaN      NaN        NaN            NaN
Poland
64460   NaN        NaN      NaN        NaN            NaN
Spain

         CurrencyDesc  ...                  SurveyEase
SurveyLength  \
0       European Euro  ...  Neither easy nor difficult  Appropriate in
length
1      Pound sterling  ...                         NaN
NaN
2                 NaN  ...  Neither easy nor difficult  Appropriate in
length
3        Albanian lek  ...                         NaN
NaN
4                 NaN  ...                        Easy
Too short
...               ...  ...                         ...
...
64456             NaN  ...                         NaN
NaN
64457             NaN  ...                         NaN
NaN
64458             NaN  ...                         NaN
NaN
64459             NaN  ...                         NaN
NaN
64460             NaN  ...                         NaN
NaN

      Trans                                     UndergradMajor  \
0        No  Computer science, computer engineering, or sof...
1       NaN  Computer science, computer engineering, or sof...
2       NaN                                                NaN
3        No  Computer science, computer engineering, or sof...
4        No  Computer science, computer engineering, or sof...
...     ...                                                ...
64456   NaN  Computer science, computer engineering, or sof...
64457   NaN                                                NaN
64458   NaN                                                NaN
64459   NaN                                                NaN
64460   NaN  Computer science, computer engineering, or sof...

            WebframeDesireNextYear
WebframeWorkedWith  \
0                     ASP.NET Core                    ASP.NET;ASP.NET
Core
1                              NaN
NaN
2                              NaN
NaN
3                              NaN
NaN
4             Django;Ruby on Rails                           Ruby on
Rails
...                            ...
```

```
...
64456                          NaN
NaN
64457                          NaN
NaN
64458                          NaN
NaN
64459  Angular;Angular.js;React.js
NaN
64460          ASP.NET Core;jQuery  Angular;Angular.js;ASP.NET
Core;jQuery

                                  WelcomeChange WorkWeekHrs YearsCode
\
0       Just as welcome now as I felt last year        50.0        36

1      Somewhat more welcome now than last year         NaN         7

2      Somewhat more welcome now than last year         NaN         4

3      Somewhat less welcome now than last year        40.0         7

4       Just as welcome now as I felt last year         NaN        15

...                                         ...         ...       ...

64456                                       NaN         NaN        10

64457                                       NaN         NaN       NaN

64458                                       NaN         NaN       NaN

64459                                       NaN         NaN       NaN

64460                                       NaN         NaN       NaN


           YearsCodePro
0                    27
1                     4
2                   NaN
3                     4
4                     8
...                 ...
64456  Less than 1 year
64457               NaN
64458               NaN
64459               NaN
64460               NaN

[64461 rows x 61 columns]
```

Verify if duplicates were actually dropped.

```
# your code goes here
len(df)-len(df.drop_duplicates())

0
```

Finding Missing values

Find the missing values for all columns.

```
# your code goes here
print(df.isnull().sum())

Respondent                0
MainBranch              299
Hobbyist                 45
Age                   19015
Age1stCode             6561
                      ...
WebframeWorkedWith    22182
WelcomeChange         11778
WorkWeekHrs           23310
YearsCode              6777
YearsCodePro          18112
Length: 61, dtype: int64

# your code goes here
df.isna()

       Respondent  MainBranch  Hobbyist    Age  Age1stCode
CompFreq  \
0           False       False     False   True       False     False

1           False       False     False   True       False      True

2           False       False     False   True       False      True

3           False       False     False  False       False      True

4           False       False     False  False       False      True

...           ...         ...       ...    ...         ...       ...

64456       False        True     False   True       False      True

64457       False        True     False   True        True      True

64458       False        True     False   True        True      True

64459       False        True     False   True        True      True
```

```
64460       False        True     False   True        True      True

       CompTotal  ConvertedComp  Country  CurrencyDesc  ...
SurveyEase  \
0           True          True    False         False  ...
False
1           True          True    False         False  ...
True
2           True          True    False          True  ...
False
3           True          True    False         False  ...
True
4           True          True    False          True  ...
False
...          ...            ...      ...           ...  ...         ..
.
64456       True          True    False          True  ...
True
64457       True          True    False          True  ...
True
64458       True          True    False          True  ...
True
64459       True          True    False          True  ...
True
64460       True          True    False          True  ...
True

       SurveyLength  Trans  UndergradMajor  WebframeDesireNextYear  \
0             False  False           False                   False
1              True   True           False                    True
2             False   True            True                    True
3              True  False           False                    True
4             False  False           False                   False
...             ...    ...             ...                     ...
64456          True   True           False                    True
64457          True   True            True                    True
64458          True   True            True                    True
64459          True   True            True                   False
64460          True   True           False                   False

       WebframeWorkedWith  WelcomeChange  WorkWeekHrs  YearsCode
YearsCodePro
0                   False          False        False      False
False
1                    True          False         True      False
False
2                    True          False         True      False
True
```

```
3                    True           False        False      False
False
4                   False           False         True      False
False
...                   ...             ...          ...        ...
...
64456                True            True         True      False
False
64457                True            True         True       True
True
64458                True            True         True       True
True
64459                True            True         True       True
True
64460               False            True         True       True
True

[64461 rows x 61 columns]
```

Imputing missing values

Find the value counts for the column Age.

```
# your code goes here
df['Age'].value_counts()

25.0    2693
28.0    2412
30.0    2406
26.0    2391
27.0    2338
        ...
34.5       1
14.7       1
97.0       1
3.0        1
14.5       1
Name: Age, Length: 110, dtype: int64
```

Find the median for the column Age

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

29.0
```

Impute the median value to Age column

```
# your code goes here
df['Age'] = df['Age'].fillna(df['Age'].median())

df

       Respondent                                         MainBranch
Hobbyist  \
0               1                     I am a developer by profession
Yes
1               2                     I am a developer by profession
No
2               3                        I code primarily as a hobby
Yes
3               4                     I am a developer by profession
Yes
4               5  I used to be a developer by profession, but no...
Yes
...           ...                                                ...
...
64456       64858                                                NaN
Yes
64457       64867                                                NaN
Yes
64458       64898                                                NaN
Yes
64459       64925                                                NaN
Yes
64460       65112                                                NaN
Yes

        Age Age1stCode CompFreq  CompTotal  ConvertedComp
Country  \
0      29.0         13  Monthly        NaN            NaN
Germany
1      29.0         19      NaN        NaN            NaN      United
Kingdom
2      29.0         15      NaN        NaN            NaN  Russian
Federation
3      25.0         18      NaN        NaN            NaN
Albania
4      31.0         16      NaN        NaN            NaN       United
States
...     ...        ...      ...        ...            ...
...
64456  29.0         16      NaN        NaN            NaN       United
States
64457  29.0        NaN      NaN        NaN            NaN
Morocco
64458  29.0        NaN      NaN        NaN            NaN
Viet Nam
```

```
64459  29.0        NaN      NaN        NaN            NaN
Poland
64460  29.0        NaN      NaN        NaN            NaN
Spain

         CurrencyDesc  ...                    SurveyEase
SurveyLength  \
0       European Euro  ...  Neither easy nor difficult  Appropriate in
length
1      Pound sterling  ...                         NaN
NaN
2                 NaN  ...  Neither easy nor difficult  Appropriate in
length
3        Albanian lek  ...                         NaN
NaN
4                 NaN  ...                        Easy
Too short
...               ...  ...                         ...
...
64456             NaN  ...                         NaN
NaN
64457             NaN  ...                         NaN
NaN
64458             NaN  ...                         NaN
NaN
64459             NaN  ...                         NaN
NaN
64460             NaN  ...                         NaN
NaN

      Trans                                    UndergradMajor  \
0        No  Computer science, computer engineering, or sof...
1       NaN  Computer science, computer engineering, or sof...
2       NaN                                              NaN
3        No  Computer science, computer engineering, or sof...
4        No  Computer science, computer engineering, or sof...
...     ...                                              ...
64456   NaN  Computer science, computer engineering, or sof...
64457   NaN                                              NaN
64458   NaN                                              NaN
64459   NaN                                              NaN
64460   NaN  Computer science, computer engineering, or sof...

            WebframeDesireNextYear
WebframeWorkedWith  \
0                     ASP.NET Core                   ASP.NET;ASP.NET
Core
1                              NaN
NaN
2                              NaN
```

```
NaN
3                              NaN
NaN
4             Django;Ruby on Rails                           Ruby on
Rails
...                            ...
...
64456                          NaN
NaN
64457                          NaN
NaN
64458                          NaN
NaN
64459  Angular;Angular.js;React.js
NaN
64460          ASP.NET Core;jQuery  Angular;Angular.js;ASP.NET
Core;jQuery

                                  WelcomeChange WorkWeekHrs YearsCode
\
0       Just as welcome now as I felt last year        50.0        36

1      Somewhat more welcome now than last year         NaN         7

2      Somewhat more welcome now than last year         NaN         4

3      Somewhat less welcome now than last year        40.0         7

4       Just as welcome now as I felt last year         NaN        15

...                                         ...         ...       ...

64456                                       NaN         NaN        10

64457                                       NaN         NaN       NaN

64458                                       NaN         NaN       NaN

64459                                       NaN         NaN       NaN

64460                                       NaN         NaN       NaN


           YearsCodePro
0                    27
1                     4
2                   NaN
3                     4
4                     8
...                 ...
```

```
64456  Less than 1 year
64457               NaN
64458               NaN
64459               NaN
64460               NaN

[64461 rows x 61 columns]
```

Identify the value that is most frequent (majority) in the Country column.

```
# your code goes here
country = df['Country']
print('Done')

Done

country.mode() #United States the most frequent country

0    United States
Name: Country, dtype: object
```

Drop all the missing values from the dataset

```
# your code goes here
df_NoNaN = df.dropna()
df_NoNaN

       Respondent                      MainBranch Hobbyist   Age
Age1stCode  \
9              10  I am a developer by profession      Yes  22.0
14
32             33  I am a developer by profession      Yes  39.0
14
41             42  I am a developer by profession       No  32.0
14
46             47  I am a developer by profession      Yes  53.0
10
68             69  I am a developer by profession      Yes  25.0
12
...           ...                             ...      ...   ...
...
61636       62886  I am a developer by profession      Yes  32.0
24
61654       62904  I am a developer by profession      Yes  33.0
24
61993       63288  I am a developer by profession       No  31.0
16
63141       64523  I am a developer by profession       No  29.0
15
63517       64938  I am a developer by profession       No  33.0
```

```
13

      CompFreq  CompTotal  ConvertedComp         Country  \
9       Yearly    25000.0        32315.0  United Kingdom   
32     Monthly     4900.0        63564.0         Belgium   
41      Yearly   130000.0       130000.0   United States   
46      Yearly    58000.0        74970.0  United Kingdom   
68      Yearly   550000.0       594539.0          France   
...        ...        ...            ...             ...   
61636   Yearly   102700.0       102700.0   United States   
61654   Yearly    95000.0        95000.0   United States   
61993   Yearly    65000.0        84019.0  United Kingdom   
63141  Monthly     8500.0        23364.0          Brazil   
63517   Yearly    55000.0        59454.0          France   

               CurrencyDesc  ...           SurveyLength Trans  \
9            Pound sterling  ...  Appropriate in length    No   
32            European Euro  ...  Appropriate in length    No   
41     United States dollar  ...  Appropriate in length    No   
46           Pound sterling  ...  Appropriate in length    No   
68            European Euro  ...              Too short    No   
...                     ...  ...                    ...   ...   
61636  United States dollar  ...  Appropriate in length    No   
61654  United States dollar  ...               Too long    No   
61993        Pound sterling  ...  Appropriate in length    No   
63141        Brazilian real  ...  Appropriate in length    No   
63517         European Euro  ...  Appropriate in length    No   

                                          UndergradMajor  \
9                              Mathematics or statistics   
32     Computer science, computer engineering, or sof...   
41     Computer science, computer engineering, or sof...   
46     A natural science (such as biology, chemistry,...   
68     Computer science, computer engineering, or sof...   
...                                                  ...   
61636  Information systems, information technology, o...   
61654  Computer science, computer engineering, or sof...   
61993  Computer science, computer engineering, or sof...   
63141  Computer science, computer engineering, or sof...   
63517  Computer science, computer engineering, or sof...   

                             WebframeDesireNextYear  \
9                                      Flask;jQuery   
32            Express;Gatsby;React.js;Ruby on Rails   
41                              ASP.NET Core;Spring   
46                                     Flask;Spring   
68                                     Django;Flask   
...                                             ...   
61636                                       Angular   
61654                              Express;React.js   
```

```
61993                    Angular;Angular.js;Express
63141  Angular;ASP.NET;ASP.NET Core;React.js;Vue.js
63517                                  Django;Flask

                               WebframeWorkedWith  \
9                                    Flask;jQuery
32     Angular;Angular.js;Django;Express;React.js
41                  ASP.NET;Flask;React.js;Spring
46                                   Flask;Spring
68                                   Django;Flask
...                                           ...
61636             Angular;Angular.js;ASP.NET Core
61654             Express;Laravel;React.js;Vue.js
61993                  Angular;Angular.js;Express
63141                 ASP.NET;ASP.NET Core;jQuery
63517           Django;Flask;jQuery;Ruby on Rails

                                  WelcomeChange WorkWeekHrs YearsCode
\
9      Somewhat more welcome now than last year        36.0         8

32      Just as welcome now as I felt last year        40.0        20

41     Somewhat less welcome now than last year        37.0        16

46      Just as welcome now as I felt last year        40.0        43

68      Just as welcome now as I felt last year        40.0        13

...                                         ...         ...       ...

61636  Somewhat more welcome now than last year        45.0         7

61654   Just as welcome now as I felt last year        50.0         9

61993   Just as welcome now as I felt last year        40.0        14

63141  Somewhat more welcome now than last year        40.0        19

63517   Just as welcome now as I felt last year        40.0        20


      YearsCodePro NormalizedAnnualCompensation
9                4                      25000.0
32              14                      58800.0
41              10                     130000.0
46              28                      58000.0
68               3                     550000.0
...            ...                          ...
61636            5                     102700.0
61654            7                      95000.0
```

```
61993             7                           65000.0
63141            17                          102000.0
63517            12                           55000.0

[4387 rows x 62 columns]
```

Normalizing data

There are two columns in the dataset that talk about compensation.

One is "CompFreq". This column shows how often a developer is paid (Yearly, Monthly, Weekly).

The other is "CompTotal". This column talks about how much the developer is paid per Year, Month, or Week depending upon his/her "CompFreq".

This makes it difficult to compare the total compensation of the developers.

In this section you will create a new column called 'NormalizedAnnualCompensation' which contains the 'Annual Compensation' irrespective of the 'CompFreq'.

Once this column is ready, it makes comparison of salaries easy.

List out the various categories in the column 'CompFreq'

```
# your code goes here
df_NoNaN['CompFreq'].value_counts().unique

<bound method Series.unique of Yearly     2627
Monthly    1689
Weekly       71
Name: CompFreq, dtype: int64>
```

Create a new column named 'NormalizedAnnualCompensation'. Use the hint given below if needed.

```
# your code goes here
def conditions(s):
  if (s['CompFreq'] == 'Yearly'):
    return s['CompTotal']
  elif (s['CompFreq'] == 'Monthly'):
    return (s['CompTotal'] * 12)
  else:
    return (s['CompTotal'] * 52)

df_NoNaN['NormalizedAnnualCompensation'] = df.apply(conditions,
axis=1)
df_NoNaN.head()

<ipython-input-49-422213b9c930>:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df_NoNaN['NormalizedAnnualCompensation'] = df.apply(conditions,
axis=1)

    Respondent                      MainBranch Hobbyist   Age
Age1stCode  \
9           10  I am a developer by profession      Yes  22.0
14
32          33  I am a developer by profession      Yes  39.0
14
41          42  I am a developer by profession       No  32.0
14
46          47  I am a developer by profession      Yes  53.0
10
68          69  I am a developer by profession      Yes  25.0
12

   CompFreq  CompTotal  ConvertedComp         Country
CurrencyDesc  \
9    Yearly    25000.0        32315.0  United Kingdom        Pound
sterling
32  Monthly     4900.0        63564.0         Belgium         European
Euro
41   Yearly   130000.0       130000.0   United States  United States
dollar
46   Yearly    58000.0        74970.0  United Kingdom        Pound
sterling
68   Yearly   550000.0       594539.0          France         European
Euro

    ...             SurveyLength Trans  \
9   ...  Appropriate in length    No
32  ...  Appropriate in length    No
41  ...  Appropriate in length    No
46  ...  Appropriate in length    No
68  ...              Too short    No

                                       UndergradMajor  \
9                            Mathematics or statistics
32  Computer science, computer engineering, or sof...
41  Computer science, computer engineering, or sof...
46  A natural science (such as biology, chemistry,...
68  Computer science, computer engineering, or sof...

                   WebframeDesireNextYear  \
9                            Flask;jQuery
```

```
32  Express;Gatsby;React.js;Ruby on Rails
41                    ASP.NET Core;Spring
46                           Flask;Spring
68                           Django;Flask

                            WebframeWorkedWith  \
9                                 Flask;jQuery
32  Angular;Angular.js;Django;Express;React.js
41               ASP.NET;Flask;React.js;Spring
46                                Flask;Spring
68                                Django;Flask

                               WelcomeChange WorkWeekHrs YearsCode  \
9   Somewhat more welcome now than last year        36.0         8
32   Just as welcome now as I felt last year        40.0        20
41  Somewhat less welcome now than last year        37.0        16
46   Just as welcome now as I felt last year        40.0        43
68   Just as welcome now as I felt last year        40.0        13

   YearsCodePro NormalizedAnnualCompensation
9             4                      25000.0
32           14                      58800.0
41           10                     130000.0
46           28                      58000.0
68            3                     550000.0

[5 rows x 62 columns]
```

```
df['NormalizedAnnualCompensation'].median()
```

```
104000.0
```

Authors

Ramesh Sannareddy

Other Contributors

Rav Ahuja

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-17	0.1	Ramesh Sannareddy	Created initial version of the lab