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, "ml_survey_data.csv")
file_name="ml_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.

Yes							
1		2			I am	a developer by	profession
No							
2		3			I	code primarily	as a hobby
Yes							
3 Yes		4			I am	a developer by	profession
4		5	I us	sed to be	a developer	by profession	, but no
Yes							
64456 Yes		64858					NaN
64457		64867					NaN
Yes							
64458 Yes		64898					NaN
64459		64925					NaN
Yes		65110					
64460		65112					NaN
Yes							
C		Age1st	Code	CompFreq	CompTotal	ConvertedComp	
Country 0	\ NaN		13	Monthly	NaN	NaN	
Germany	IVAIN		13	rioritity	INGIN	Ivaiv	
1	NaN		19	NaN	NaN	NaN	United
Kingdom 2	NaN		15	NaN	NaN	NaN	Russian
Z Federat:			13	IVAIN	IValv	NaN	Kusstall
	25.0		18	NaN	NaN	NaN	
Albania	23.0		10	Nan	Nan	IVAIV	
	31.0		16	NaN	NaN	NaN	United
States							
64456	NaN		16	NaN	NaN	NaN	United
States					A. A.		
64457 Morocco	NaN		NaN	NaN	NaN	NaN	
64458	NaN		NaN	NaN	NaN	NaN	
Viet Nar	n						
64459	NaN		NaN	NaN	NaN	NaN	
Poland 64460	NaN		NaN	NaN	NaN	NaN	
Spain	IVAIV		IVAIN	IVAIN	INGIN	ivalv	
	Cuu	rrencyD	ASC			SurveyEase	
SurveyLe			CSC			JulveyLase	
,	3	•					

```
0
        European Euro ... Neither easy nor difficult Appropriate in
length
1
       Pound sterling ...
                                                     NaN
NaN
                  NaN ... Neither easy nor difficult Appropriate in
length
         Albanian lek ...
                                                     NaN
3
NaN
                  NaN ...
                                                    Easy
Too short
. . .
64456
                                                     NaN
                  NaN ...
NaN
64457
                  NaN ...
                                                     NaN
NaN
                  NaN ...
                                                     NaN
64458
NaN
                                                     NaN
64459
                  NaN ...
NaN
64460
                                                     NaN
                  NaN
NaN
      Trans
                                                 UndergradMajor \
             Computer science, computer engineering, or sof...
0
         No
1
        NaN
             Computer science, computer engineering, or sof...
2
        NaN
3
         No
             Computer science, computer engineering, or sof...
4
             Computer science, computer engineering, or sof...
         No
. . .
64456
        NaN
             Computer science, computer engineering, or sof...
64457
        NaN
                                                             NaN
64458
        NaN
                                                             NaN
64459
        NaN
64460
        NaN
             Computer science, computer engineering, or sof...
            WebframeDesireNextYear
WebframeWorkedWith \
                      ASP.NET Core
                                                       ASP.NET:ASP.NET
Core
                                NaN
1
NaN
2
                                NaN
NaN
                                NaN
3
NaN
              Django; Ruby on Rails
4
                                                               Ruby on
Rails
. . .
```

```
64456
                                 NaN
NaN
64457
                                 NaN
NaN
64458
                                 NaN
NaN
64459 Angular; Angular.js; React.js
NaN
                ASP.NET Core; jQuery Angular; Angular.js; ASP.NET
64460
Core; jQuery
                                    WelcomeChange WorkWeekHrs YearsCode
/
0
        Just as welcome now as I felt last year
                                                                       36
                                                           50.0
       Somewhat more welcome now than last year
                                                            NaN
                                                                        7
       Somewhat more welcome now than last year
                                                                        4
2
                                                            NaN
3
       Somewhat less welcome now than last year
                                                           40.0
                                                                        7
        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
2
                     NaN
3
                       4
4
                       8
       Less than 1 year
64456
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
MainBranch
                         299
Hobbyist
                          45
                       19015
Age
Age1stCode
                        6561
                       22182
WebframeWorkedWith
WelcomeChange
                       11778
WorkWeekHrs
                       23310
YearsCode
                        6777
YearsCodePro
                       18112
Length: 61, dtype: int64
# your code goes here
df.isna()
       Respondent MainBranch Hobbyist Age Age1stCode
CompFreq
            False
                         False
                                   False
                                            True
                                                       False
                                                                  False
            False
                         False
                                   False
                                           True
                                                       False
                                                                   True
2
            False
                         False
                                   False
                                           True
                                                       False
                                                                   True
                                                       False
                                                                   True
3
            False
                         False
                                   False False
            False
                         False
                                   False False
                                                       False
                                                                   True
64456
            False
                          True
                                   False
                                            True
                                                       False
                                                                   True
64457
            False
                          True
                                   False
                                            True
                                                        True
                                                                   True
64458
            False
                                                                   True
                          True
                                   False
                                           True
                                                        True
64459
            False
                                   False
                                                        True
                                                                   True
                          True
                                            True
```

64460	False	T	rue	False	True	True	True
	CompTotal	Converted	dComp	Country	CurrencyDe	esc	
Survey	•	Converte	acomp	country	currencybe	.50	
0	True		True	False	Fal	.se	
False				_			
1	True		True	False	Fal	.se	
True	Truc		True	F21.00	т.		
2 False	True		True	False	11	rue	
3	True		True	False	Fal	.se	
True							
4	True		True	False	Tr	ue	
False							
						• • • • • •	
64456	True		True	False	Tr	ue	
True	True		TTUE	racse		ue	
64457	True		True	False	Tr	ue	
True							
64458	True		True	False	Tr	ue	
True	_		_		_		
64459	True		True	False	I r	ue	
True 64460	True		True	False	Tr	ue	
True	1140		TTUC	racse		uc III	
0 1 2 3 4	SurveyLengt Fals Tru Fals Tru Fals	e False e True e True e False	Unde	rgradMajo False False True False False	e e e	eDesireNe	extYear \ False True True True False
64456	 <u>T</u> ru	e True		False	e		True
64457	Tru			True			True
64458 64459	Tru Tru			True True			True False
64460	Tru			False			False
0.100					-		. 4 . 5 .
	WebframeWor	kedWith	Welco	meChange	WorkWeekHr	s Years	Code
YearsC	odePro					_	
0 Falso		False		False	Fals	se l	- alse
False 1		True		False	Tru	ا ما	alse
False		TTUE		1 4 1 3 6	110		u cse
2		True		False	Tru	ie l	alse
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 8046)	(61 columnal			
[64461 rows >	C OT 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
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 \
                 1
                                        I am a developer by profession
Yes
                 2
                                        I am a developer by profession
1
No
                                           I code primarily as a hobby
2
                 3
Yes
                 4
                                        I am a developer by profession
3
Yes
                    I used to be a developer by profession, but no...
4
Yes
. . .
64456
            64858
                                                                     NaN
Yes
64457
            64867
                                                                     NaN
Yes
64458
            64898
                                                                     NaN
Yes
64459
            64925
                                                                     NaN
Yes
64460
                                                                     NaN
            65112
Yes
        Age Age1stCode CompFreq CompTotal ConvertedComp
Country
       29.0
                     13 Monthly
0
                                         NaN
                                                         NaN
Germany
1
       29.0
                     19
                             NaN
                                         NaN
                                                         NaN
                                                                  United
Kingdom
       29.0
                     15
                             NaN
                                         NaN
                                                         NaN
                                                              Russian
Federation
3
       25.0
                     18
                             NaN
                                         NaN
                                                         NaN
Albania
       31.0
                     16
                             NaN
                                         NaN
                                                         NaN
                                                                    United
States
                                         . . .
                                                         . . .
. . .
64456 29.0
                                                                    United
                     16
                             NaN
                                         NaN
                                                         NaN
States
64457 29.0
                                         NaN
                                                         NaN
                    NaN
                             NaN
Morocco
64458 29.0
                    NaN
                                         NaN
                                                         NaN
                             NaN
Viet Nam
```

64459	29.0	NaN	l Na	aN	N	laN		Nā	aN	
Poland 64460	29.0	NaN	l Na	a N	N	laN		Na	aN.	
Spain	23.0	Nuiv	140	4114	•	· ·		140	aiv	
	Curi	rencyDesc				(SurveyEa	se		
Survey		\				•	5u. 10, Lu			
0		pean Euro	Ne	either	easy	/ nor	difficu	lt	Appropriate i	n
length 1	Pound	sterling					N	aN		
NaN	Tourid	Jeer cang	• • •					an.		
2		NaN	Ne	either	easy	nor nor	difficu	lt	Appropriate i	n
length 3	۸lh	anian lek					N	aN		
NaN	ALD	alitali tek					IV	alv		
4		NaN					Ea	sy		
Too she	ort									
64456		NaN					N	aN		
NaN		Itali	• • •					ai.		
64457		NaN					N	aN		
NaN		NI - NI						- NI		
64458 NaN		NaN					N	aN		
64459		NaN					N	aN		
NaN										
64460		NaN					N	aN		
NaN										
	Trans						Unde	rara	adMajor \	
0	No	Computer	science,	compu	uter	engi				
1	NaN	Computer	science,	compu	uter	engi	neering,	or		
2	NaN No	Computer	ccience	COMPI	ıtar	anaiı	neering	or	NaN	
4	No	Computer								
		·		•		J	•			
64456	NaN	Computer	science,	compu	ıter	engi	neering,	or		
64457 64458	NaN NaN								NaN NaN	
64459	NaN								NaN	
64460	NaN	Computer	science,	compu	uter	engi	neering,	or	-	
	1	WebframeDe	ci roMovt	Voor						
Webfra			STICNEX	. i eai						
0		· · · · · · · · · · · · · · · · · · ·	ASP.NET	Core				AS	SP.NET;ASP.NET	
Core										
1 NaN				NaN						
NaN 2				NaN						
_				ITOIT						

NaN								
3			Na	ıN				
NaN 4	Di	ango;Ruby o	n Dail	c				Ruby on
4 Rails	ינט	aligo, Ruby C	JII Nait	. 5				Ruby on
64456			Na	ıΝ				
NaN								
64457			Na	ıN				
NaN 64458			Na	M				
NaN			IVa	IIN				
64459	Angular;A	ngular.js;F	React.j	S				
NaN	Δ.	CD NET Com					1 '- ACD N	T
64460		SP.NET Core	e;jQuer	'y Ar	ngu La	r;Angu	ılar.js;ASP.N	NE I
Core;j	query							
				Weld	comeCl	nange	WorkWeekHrs	YearsCode
\								
0	Just as v	welcome now	v as I	felt	last	year	50.0	36
1	Compubat i	mana valaar		+ h - n	1	\\00 B	NoN	7
1	Somewhat i	more welcon	ne now	Liidii	last	year	NaN	7
2	Somewhat	more welcom	ne now	than	last	year	NaN	4
3	Somewhat	less welcom	ne now	than	last	year	40.0	7
4	Just as v	welcome now	v as I	felt	last	year	NaN	15
64456						NaN	NaN	10
64457						NaN	NaN	NaN
04437						IVAIN	Ivaiv	Ivaiv
64458						NaN	NaN	NaN
64459						NaN	NaN	NaN
64460						NaN	NaN	NaN
0	Years	CodePro						
0 1 2 3 4		27 4						
2		NaN						
3		4						
4		8						

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

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
                   I am a developer by profession
               10
                                                        Yes
                                                             22.0
14
32
               33
                   I am a developer by profession
                                                        Yes
                                                             39.0
14
                                                         No 32.0
41
               42
                   I am a developer by profession
14
46
               47
                   I am a developer by profession
                                                             53.0
                                                        Yes
10
68
               69
                   I am a developer by profession
                                                             25.0
                                                        Yes
12
. . .
                  I am a developer by profession
                                                        Yes 32.0
61636
            62886
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
                                                             33.0
                                                         No
```

```
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
. . .
                  102700.0
                                  102700.0
                                              United States
61636
        Yearly
61654
        Yearly
                   95000.0
                                   95000.0
                                              United States
                                            United Kingdom
        Yearly
61993
                   65000.0
                                   84019.0
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
                                                              . . .
       United States dollar
61636
                                    Appropriate in length
                                                               No
       United States dollar
61654
                                                  Too lona
                                                               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...
       A natural science (such as biology, chemistry,...
46
68
       Computer science, computer engineering, or sof...
       Information systems, information technology, o...
61636
61654
       Computer science, computer engineering, or sof...
       Computer science, computer engineering, or sof...
61993
63141
       Computer science, computer engineering, or sof...
       Computer science, computer engineering, or sof...
63517
                               WebframeDesireNextYear
9
                                         Flask; iQuery
32
               Express; Gatsby; React.js; Ruby on Rails
41
                                  ASP.NET Core; Spring
46
                                         Flask; Spring
                                         Django; Flask
68
. . .
61636
                                               Angular
61654
                                     Express; React. js
```

61993 63141 63517	Angular;Angular.js;Expre Angular;ASP.NET;ASP.NET Core;React.js;Vue. Django;Fla	js	
	WebframeWorkedWith	1	
9	Flask;jQuery		
32	Angular; Angular.js; Django; Express; React.js		
41 46	ASP.NET;Flask;React.js;Spring		
68	Flask;Spring Django;Flask		
61636	Angular;Angular.js;ASP.NET Core		
61654	Express; Laravel; React. js; Vue. js		
61993 63141	Angular;Angular.js;Express ASP.NET;ASP.NET Core;jQuery		
63517	Django;Flask;jQuery;Ruby on Rails		
			6 1
\	welcomeChange w	lorkWeekHrs Year	scode
9	Somewhat more welcome now than last year	36.0	8
22	Total and the second se	40.0	20
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 32	4 25000.0 14 58800.0		
41	10 130000.0		
46	28 58000.0		
68	3 550000.0		
61636	5 102700.0		
61654	7 95000.0		

61993 63141	7 17 12	65000.0 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 "CompFreg".

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
                I am a developer by profession
                                                         39.0
            33
                                                     Yes
14
            42 I am a developer by profession
41
                                                      No 32.0
14
46
            47
                I am a developer by profession
                                                     Yes 53.0
10
68
                I am a developer by profession
            69
                                                     Yes 25.0
12
   CompFreq CompTotal ConvertedComp
                                              Country
CurrencyDesc
     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
     Yearlv
68
              550000.0
                             594539.0
                                                France
                                                               European
Euro
                  SurveyLength Trans \
9
         Appropriate in length
                                  No
         Appropriate in length
32
                                  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...
    Computer science, computer engineering, or sof...
41
   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
                                    Diango; Flask
                                 WelcomeChange WorkWeekHrs YearsCode \
    Somewhat more welcome now than last year
                                                       36.0
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
     Just as welcome now as I felt last year
                                                                    13
68
                                                       40.0
   YearsCodePro NormalizedAnnualCompensation
9
              4
                                       25000.0
32
             14
                                       58800.0
41
              10
                                      130000.0
46
              28
                                       58000.0
              3
                                      550000.0
68
[5 rows x 62 columns]
df['NormalizedAnnualCompensation'].median()
104000.0
```

Authors

Ramesh Sannareddy

Other Contributors

Rav Ahuja

Change Log

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

Copyright © 2020 IBM Corporation. This notebook and its source code are released under the terms of the MIT License.