Survey Dataset Exploration Lab

Estimated time needed: **30** minutes

Objectives

After completing this lab you will be able to:

- Load the dataset that will used thru the capstone project.
- Explore the dataset.
- Get familier with the data types.

Load the dataset

Import the required libraries.

In [6]:
```
import pandas as pd
from bs4 import BeautifulSoup
import requests
```

The dataset is available on the IBM Cloud at the below url.

In [4]:
```
dataset_url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-
dataset_url
```

Out[4]:
```
'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-Skil
lsNetwork/LargeData/m1_survey_data.csv'
```

In [7]:
```
data = requests.get(dataset_url).text
```

In [10]:
```
soup = BeautifulSoup(data)
```

Load the data available at dataset_url into a dataframe.

In [13]:
```
# your code goes here
table = soup.find('table')
for row in table.find_all('tr'):
    col = row.find_all('td')
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In[13], line 3
      1 # your code goes here
      2 table = soup.find('table')
----> 3 for row in table.find('tr'):
      4     col = row.find_all('td')

AttributeError: 'NoneType' object has no attribute 'find'
```

In [16]:
```
df = pd.read_csv('m1_survey_data.csv')
```

In [ ]:
```

```

Explore the data set

It is a good idea to print the top 5 rows of the dataset to get a feel of how the dataset will look.

Display the top 5 rows and columns from your dataset.

Find out the number of rows and columns

Start by exploring the numbers of rows and columns of data in the dataset.

Print the number of rows in the dataset.

In [17]:
```
# your code goes here
df.head()
```

Out[17]:

	Respondent	MainBranch	Hobbyist	OpenSourcer	OpenSource	Employment	Country	Student	
0	4	I am a developer by profession	No	Never	The quality of OSS and closed source software ...	Employed full-time	United States	No	
1	9	I am a developer by profession	Yes	Once a month or more often	The quality of OSS and closed source software ...	Employed full-time	New Zealand	No	c
2	13	I am a developer by profession	Yes	Less than once a month but more than once per ...	OSS is, on average, of HIGHER quality than pro...	Employed full-time	United States	No	(
3	16	I am a developer by profession	Yes	Never	The quality of OSS and closed source software ...	Employed full-time	United Kingdom	No	(
4	17	I am a developer by profession	Yes	Less than once a month but more than once per ...	The quality of OSS and closed source software ...	Employed full-time	Australia	No	

5 rows × 85 columns

Print the number of columns in the dataset.

In [24]:
```
len(df.columns)
```

Out[24]:
```
85
```

In [27]:
```
df.shape[1]
```

Out[27]:
```
85
```

In [18]:
```
# your code goes here
df.shape
```

Out[18]:
```
(11552, 85)
```

In [20]:
```
df.columns
```

Out[20]:
```
Index(['Respondent', 'MainBranch', 'Hobbyist', 'OpenSourcer', 'OpenSource',
       'Employment', 'Country', 'Student', 'EdLevel', 'UndergradMajor',
       'EduOther', 'OrgSize', 'DevType', 'YearsCode', 'Age1stCode',
       'YearsCodePro', 'CareerSat', 'JobSat', 'MgrIdiot', 'MgrMoney',
       'MgrWant', 'JobSeek', 'LastHireDate', 'LastInt', 'FizzBuzz',
       'JobFactors', 'ResumeUpdate', 'CurrencySymbol', 'CurrencyDesc',
       'CompTotal', 'CompFreq', 'ConvertedComp', 'WorkWeekHrs', 'WorkPlan',
       'WorkChallenge', 'WorkRemote', 'WorkLoc', 'ImpSyn', 'CodeRev',
       'CodeRevHrs', 'UnitTests', 'PurchaseHow', 'PurchaseWhat',
       'LanguageWorkedWith', 'LanguageDesireNextYear', 'DatabaseWorkedWith',
       'DatabaseDesireNextYear', 'PlatformWorkedWith',
       'PlatformDesireNextYear', 'WebFrameWorkedWith',
       'WebFrameDesireNextYear', 'MiscTechWorkedWith',
       'MiscTechDesireNextYear', 'DevEnviron', 'OpSys', 'Containers',
       'BlockchainOrg', 'BlockchainIs', 'BetterLife', 'ITperson', 'OffOn',
       'SocialMedia', 'Extraversion', 'ScreenName', 'SOVisit1st',
       'SOVisitFreq', 'SOVisitTo', 'SOFindAnswer', 'SOTimeSaved',
       'SOHowMuchTime', 'SOAccount', 'SOPartFreq', 'SOJobs', 'EntTeams',
       'SOComm', 'WelcomeChange', 'SONewContent', 'Age', 'Gender', 'Trans',
       'Sexuality', 'Ethnicity', 'Dependents', 'SurveyLength', 'SurveyEase'],
      dtype='object')
```

Identify the data types of each column

Explore the dataset and identify the data types of each column.

Print the datatype of all columns.

In [19]:
```
# your code goes here
df.dtypes
```

Out[19]:
```
Respondent       int64
MainBranch      object
Hobbyist        object
OpenSourcer     object
OpenSource      object
                 ...
Sexuality       object
Ethnicity       object
Dependents      object
SurveyLength    object
SurveyEase      object
Length: 85, dtype: object
```

Print the mean age of the survey participants.

In [21]:
```
# your code goes here
df['Age'].mean()
```

Out[21]:
```
30.77239449133718
```

The dataset is the result of a world wide survey. Print how many unique countries are there in the Country column.

In [23]:
```
# your code goes here
df.Country.nunique()
```

Out[23]: 135

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