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 [1]:
```
import pandas as pd
```

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

In [3]:
```
dataset_url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/I
```

Load the data available at dataset_url into a dataframe.

In [4]:
```
# your code goes here
df = pd.read_csv(dataset_url)
```

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.

In [7]:
```
# your code goes here
df.head(5)
```

Out[7]:

	Respondent	MainBranch	Hobbyist	OpenSourcer	OpenSource	Employment	Country
0	4	I am a developer by profession	No	Never	The quality of OSS and closed source software ...	Employed full-time	United States
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
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
3	16	I am a developer by profession	Yes	Never	The quality of OSS and closed source software ...	Employed full-time	United Kingdom
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

5 rows × 85 columns

Find out the number of rows and columns

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

In [22]:
```
print(df.shape)
```

```
(11552, 85)
```

Print the number of rows in the dataset.

In [23]:
```
print(df.shape[0])
print(len(df))
```

```
11552
11552
```

Print the number of columns in the dataset.

In [24]:
```
print(len(df.columns))
print(df.shape[1])
```

```
85
85
```

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 [33]:
```
# Set display option to show all data types without truncation
pd.set_option('display.max_columns', None)  # Show all columns without truncation
pd.set_option('display.max_rows', None)     # Show all rows without truncation

# your code goes here
print("Datatype of all columns:")
print(df.dtypes)
```

```
Datatype of all columns:
Respondent                 int64
MainBranch                object
Hobbyist                  object
OpenSourcer               object
OpenSource                object
Employment                object
Country                   object
Student                   object
EdLevel                   object
UndergradMajor            object
EduOther                  object
OrgSize                   object
DevType                   object
YearsCode                 object
Age1stCode                object
YearsCodePro              object
CareerSat                 object
JobSat                    object
MgrIdiot                  object
MgrMoney                  object
MgrWant                   object
JobSeek                   object
LastHireDate              object
LastInt                   object
FizzBuzz                  object
JobFactors                object
ResumeUpdate              object
CurrencySymbol            object
CurrencyDesc              object
CompTotal                float64
CompFreq                  object
ConvertedComp            float64
WorkWeekHrs              float64
WorkPlan                  object
WorkChallenge             object
WorkRemote                object
WorkLoc                   object
ImpSyn                    object
CodeRev                   object
CodeRevHrs               float64
UnitTests                 object
PurchaseHow               object
PurchaseWhat              object
LanguageWorkedWith        object
LanguageDesireNextYear    object
DatabaseWorkedWith        object
DatabaseDesireNextYear    object
PlatformWorkedWith        object
PlatformDesireNextYear    object
WebFrameWorkedWith        object
WebFrameDesireNextYear    object
MiscTechWorkedWith        object
MiscTechDesireNextYear    object
DevEnviron                object
OpSys                     object
```

```
Containers              object
BlockchainOrg           object
BlockchainIs            object
BetterLife              object
ITperson                object
OffOn                   object
SocialMedia             object
Extraversion            object
ScreenName              object
SOVisit1st              object
SOVisitFreq             object
SOVisitTo               object
SOFindAnswer            object
SOTimeSaved             object
SOHowMuchTime           object
SOAccount               object
SOPartFreq              object
SOJobs                  object
EntTeams                object
SOComm                  object
WelcomeChange           object
SONewContent            object
Age                    float64
Gender                  object
Trans                   object
Sexuality               object
Ethnicity               object
Dependents              object
SurveyLength            object
SurveyEase              object
dtype: object
```

Print the mean age of the survey participants.

In [41]:
```
# your code goes here
print(df['Age'].mean())
```

```
30.77239449133718
```

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

In [44]:
```
# your code goes here
print(len(df['Country'].unique()))
```

```
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