AKRAM ALZAGHIR

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 [2]:

```
dataset_url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/LargeData/m1_survey_data.csv"
```

Load the data available at dataset_url into a dataframe.

In [3]:

```
# 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 [4]:

```
# your code goes here
df.head(3)
```

Out[4]:

	Respondent	MainBranch	Hobbyist	OpenSourcer	OpenSource	Employment	Country	St
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 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.

Print the number of rows in the dataset.

In [7]:

```
# your code goes here
#number_of_rows = len(df)
number_of_rows = len(df.axes[0])
number_of_rows
```

Out[7]:

```
11552
```

Print the number of columns in the dataset.

In [9]:

```
# your code goes here
number_of_columns = len(df.axes[1])
#number_of_columns = len(df.columns)
number_of_columns
```

Out[9]:

```
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 [7]:

```
# your code goes here
df.dtypes
```

Out[7]:

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

In [8]:

```
# your code goes here
df["Age"].dtypes
```

Out[8]:

```
dtype('float64')
```

Print the mean age of the survey participants.

In [9]:

```
# your code goes here
df["Age"].mean()
```

Out[9]:

```
30.77239449133718
```

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

In [10]:

```
# your code goes here
df['Country'].unique()
```

Out[10]:

```
array(['United States', 'New Zealand', 'United Kingdom', 'Australia',
       'Brazil', 'Lithuania', 'Israel', 'South Africa', 'Czech Republic',
       'Spain', 'Germany', 'Serbia', 'India', 'Sweden', 'China', 'France',
       'Netherlands', 'Philippines', 'Ireland', 'Pakistan', 'Austria',
       'Canada', 'Croatia', 'Italy', 'Russian Federation', 'Argentina',
       'Romania', 'Iran', 'Hungary', 'Latvia', 'Hong Kong (S.A.R.)',
       'United Arab Emirates', 'Poland', 'Portugal', 'Bulgaria',
       'Nicaragua', 'Denmark', 'Japan', 'Guatemala', 'Bangladesh',
       'Ukraine', 'Mexico', 'Egypt', 'Switzerland', 'Mauritius',
       'South Korea', 'Slovenia', 'Estonia', 'Norway', 'Singapore',
       'Republic of Moldova', 'Belgium', 'Nigeria', 'Turkey', 'Thailand',
       'Mongolia', 'Chile', 'Malaysia', 'Georgia', 'Luxembourg',
       'Dominican Republic', 'Cape Verde', 'Burundi', 'Finland', 'Greece',
       'Colombia', 'Taiwan', 'Yemen', 'Indonesia', 'Belarus', 'Slovakia',
       'Nepal', 'Kenya', 'Venezuela, Bolivarian Republic of...',
       'Armenia', 'Panama', 'Lebanon', 'Kuwait', 'Algeria',
       "Côte d'Ivoire", 'Bosnia and Herzegovina', 'Brunei Darussalam',
       'Costa Rica', 'Jordan', 'Zimbabwe', 'Ecuador', 'Albania',
       'Azerbaijan', 'Other Country (Not Listed Above)', 'Uzbekistan',
       'The former Yugoslav Republic of Macedonia', 'Sri Lanka', 'Ghana',
       'Paraguay', 'Peru', 'Viet Nam', 'Malta', 'Rwanda', 'El Salvador',
       'Uruguay', 'Tunisia', 'Bolivia', 'Honduras', 'Liechtenstein',
       'Qatar', 'Cameroon', 'Turkmenistan', 'Kyrgyzstan', 'Somalia',
       'Republic of Korea', 'Cuba', 'Montenegro', 'Monaco', 'Cyprus',
       'Uganda', 'Senegal', 'Syrian Arab Republic', 'Morocco', 'Ethiopia',
       'United Republic of Tanzania', 'Iceland', 'Swaziland',
       'Congo, Republic of the...', 'Saudi Arabia', 'Afghanistan',
       'Bahrain', 'Timor-Leste', 'Jamaica', 'Myanmar', 'Sudan',
       'Libyan Arab Jamahiriya', 'Togo', 'Cambodia', 'Mozambique', 'Ira
q'],
      dtype=object)
```

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