**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 [4]:

*# your code goes here*

df\_survey **=** 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 [5]:

*# your code goes here*

df\_survey**.**head()

Out[5]:

|  | **Respondent** | **MainBranch** | **Hobbyist** | **OpenSourcer** | **OpenSource** | **Employment** | **Country** | **Student** | **EdLevel** | **UndergradMajor** | **...** | **WelcomeChange** | **SONewContent** | **Age** | **Gender** | **Trans** | **Sexuality** | **Ethnicity** | **Dependents** | **SurveyLength** | **SurveyEase** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 4 | I am a developer by profession | No | Never | The quality of OSS and closed source software ... | Employed full-time | United States | No | Bachelor’s degree (BA, BS, B.Eng., etc.) | Computer science, computer engineering, or sof... | ... | Just as welcome now as I felt last year | Tech articles written by other developers;Indu... | 22.0 | Man | No | Straight / Heterosexual | White or of European descent | No | Appropriate in length | Easy |
| **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 | Some college/university study without earning ... | Computer science, computer engineering, or sof... | ... | Just as welcome now as I felt last year | NaN | 23.0 | Man | No | Bisexual | White or of European descent | No | Appropriate in length | Neither easy nor difficult |
| **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 | Master’s degree (MA, MS, M.Eng., MBA, etc.) | Computer science, computer engineering, or sof... | ... | Somewhat more welcome now than last year | Tech articles written by other developers;Cour... | 28.0 | Man | No | Straight / Heterosexual | White or of European descent | Yes | Appropriate in length | Easy |
| **3** | 16 | I am a developer by profession | Yes | Never | The quality of OSS and closed source software ... | Employed full-time | United Kingdom | No | Master’s degree (MA, MS, M.Eng., MBA, etc.) | NaN | ... | Just as welcome now as I felt last year | Tech articles written by other developers;Indu... | 26.0 | Man | No | Straight / Heterosexual | White or of European descent | No | Appropriate in length | Neither easy nor difficult |
| **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 | Bachelor’s degree (BA, BS, B.Eng., etc.) | Computer science, computer engineering, or sof... | ... | Just as welcome now as I felt last year | Tech articles written by other developers;Indu... | 29.0 | Man | No | Straight / Heterosexual | Hispanic or Latino/Latina;Multiracial | No | Appropriate in length | Easy |

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.

Print the number of rows in the dataset.

In [14]:

*# your code goes here*

print('The number of rows in the dataframe is ' **+** str(df\_survey**.**shape[0]) **+**'.')

The number of rows in the dataframe is 11552.

---------------------------------------------------------------------------

NameError Traceback (most recent call last)

<ipython-input-14-94ead6c633d8> in <module>

**1** # your code goes here

**2** print('The number of rows in the dataframe is ' + str(df\_survey.shape[0]) +'.')

----> 3 rows(df\_survey)

NameError: name 'rows' is not defined

Print the number of columns in the dataset.

In [15]:

*# your code goes here*

print('The number of columns in the dataframe is ' **+** str(df\_survey**.**shape[1]) **+**'.')

The number of columns in the dataframe is 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 [36]:

*# your code goes here*

*# option 1 (doesn't show all columns):*

*#df\_survey.dtypes*

*# option 2 (prints more info than needed):*

*#df\_survey.info(verbose=True)*

*# option 3 (works, but wordy)*

**with** pd**.**option\_context('display.max\_rows', **None**, 'display.max\_columns', **None**):

print(df\_survey**.**dtypes)

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 [42]:

*# your code goes here*

print('The mean age of the survey participants is ', df\_survey['Age']**.**mean()**.**round(1))

The mean age of the survey participants is {} 30.8

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

In [54]:

*# your code goes here*

print('There are', df\_survey['Country']**.**nunique(), 'unique countries in the survey:')

print('\nCountry Respondants')

**with** pd**.**option\_context('display.max\_rows', **None**, 'display.max\_columns', **None**):

print(df\_survey['Country']**.**value\_counts())

There are 135 unique countries in the survey:

Country Respondants

United States 3173

India 911

United Kingdom 841

Germany 715

Canada 442

France 339

Brazil 328

Australia 287

Netherlands 259

Spain 257

Russian Federation 211

Poland 205

Italy 188

Sweden 162

Switzerland 151

Ukraine 111

Israel 104

South Africa 104

Turkey 98

Mexico 98

Romania 95

Pakistan 93

Austria 89

Belgium 89

Denmark 85

Norway 84

Iran 83

Bulgaria 77

Greece 76

New Zealand 76

Argentina 73

Finland 72

Czech Republic 72

China 69

Portugal 69

Ireland 67

Hungary 64

Bangladesh 58

Serbia 53

Japan 50

Colombia 50

Nigeria 44

Sri Lanka 43

Malaysia 42

Egypt 39

Indonesia 36

Philippines 36

Slovenia 35

Lithuania 35

Hong Kong (S.A.R.) 34

Singapore 33

Croatia 31

Slovakia 28

Estonia 27

Belarus 24

Taiwan 23

Chile 21

Kenya 20

Latvia 19

Thailand 19

United Arab Emirates 19

Other Country (Not Listed Above) 18

Uruguay 16

Dominican Republic 16

South Korea 15

Costa Rica 15

Nepal 15

Guatemala 13

Armenia 12

Peru 12

Albania 12

Lebanon 12

Saudi Arabia 12

Viet Nam 12

Ecuador 11

Luxembourg 11

Bosnia and Herzegovina 10

Republic of Moldova 10

Tunisia 9

Georgia 9

El Salvador 8

Malta 8

Morocco 8

Jordan 7

The former Yugoslav Republic of Macedonia 7

Cyprus 7

Venezuela, Bolivarian Republic of... 7

Ghana 7

Paraguay 6

Uganda 6

Iceland 6

Honduras 5

Zimbabwe 5

Azerbaijan 5

Côte d'Ivoire 5

Kyrgyzstan 4

Cameroon 4

Bolivia 4

Cuba 4

Ethiopia 4

Nicaragua 3

Myanmar 3

Swaziland 3

Algeria 3

Bahrain 3

Syrian Arab Republic 3

Congo, Republic of the... 2

United Republic of Tanzania 2

Kuwait 2

Montenegro 2

Mauritius 2

Panama 2

Afghanistan 2

Senegal 2

Mongolia 2

Burundi 2

Sudan 2

Uzbekistan 2

Monaco 1

Togo 1

Somalia 1

Cambodia 1

Timor-Leste 1

Republic of Korea 1

Cape Verde 1

Jamaica 1

Mozambique 1

Rwanda 1

Liechtenstein 1

Yemen 1

Libyan Arab Jamahiriya 1

Iraq 1

Qatar 1

Turkmenistan 1

Brunei Darussalam 1

Name: Country, dtype: int64

**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 |