Data Visualization Lab

Estimated time needed: **45 to 60** minutes

In this assignment you will be focusing on the visualization of data.

The data set will be presented to you in the form of a RDBMS.

You will have to use SQL queries to extract the data.

Objectives

In this lab you will perform the following:

- Visualize the distribution of data.
- Visualize the relationship between two features.
- Visualize composition of data.
- Visualize comparison of data.

Demo: How to work with database

Download database file.

Connect to the database.

```
!wget https://cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/Capstone_edX/
Module%204/master.db

--2024-06-23 19:34:40--  https://cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/Capstone_edX/
Module%204/master.db
Resolving cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud
(cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud)...
169.63.118.104, 169.63.118.104
Connecting to cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud (cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud)|169.63.118.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8691712 (8.3M) [binary/octet-stream]
Saving to: ‘master.db.4’

master.db.4         100%[===================>]   8.29M  --.-KB/s    in
0.1s

2024-06-23 19:34:40 (63.5 MB/s) - ‘master.db.4’ saved
```

```
[8691712/8691712]
```

```
import sqlite3
import pandas as pd
conn = sqlite3.connect('master.db')
cur = conn.cursor()
QUERY = "SELECT * FROM master"
# the read_sql_query runs the sql query and returns the data as a
dataframe
df = pd.read_sql_query(QUERY,conn)
print(df.head())
```

```
   Respondent                      MainBranch Hobbyist  Age Age1stCode
\
0           1  I am a developer by profession      Yes   22         14

1           2  I am a developer by profession      Yes   39         14

2           3  I am a developer by profession       No   32         14

3           4  I am a developer by profession      Yes   53         10

4           5  I am a developer by profession      Yes   25         12


  CompFreq  CompTotal  ConvertedComp         Country
CurrencyDesc  \
0   Yearly      25000          32315  United Kingdom        Pound
sterling
1  Monthly       4900          63564         Belgium         European
Euro
2   Yearly     130000         130000   United States  United States
dollar
3   Yearly      58000          74970  United Kingdom        Pound
sterling
4   Yearly     550000         594539          France         European
Euro

   ...                         SOVisitFreq                   SurveyEase  \
0  ...           Multiple times per day                         Easy
1  ...            Daily or almost daily  Neither easy nor difficult
2  ...  A few times per month or weekly                         Easy
3  ...             A few times per week  Neither easy nor difficult
4  ...             A few times per week                         Easy

            SurveyLength Trans  \
0  Appropriate in length    No
1  Appropriate in length    No
2  Appropriate in length    No
3  Appropriate in length    No
```

```
4              Too short    No

                                    UndergradMajor  \
0                          Mathematics or statistics
1  Computer science, computer engineering, or sof...
2  Computer science, computer engineering, or sof...
3  A natural science (such as biology, chemistry,...
4  Computer science, computer engineering, or sof...

                              WelcomeChange WorkWeekHrs YearsCode  \
0  Somewhat more welcome now than last year        36.0         8
1   Just as welcome now as I felt last year        40.0        20
2  Somewhat less welcome now than last year        37.0        16
3   Just as welcome now as I felt last year        40.0        43
4   Just as welcome now as I felt last year        40.0        13

  YearsCodePro NormalizedAnnualCompensation
0            4                       25000
1           14                       58800
2           10                      130000
3           28                       58000
4            3                      550000

[5 rows x 49 columns]
```

Import pandas module.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import plotly.express as px
```

Demo: How to run an sql query

```
# print how many rows are there in the table named 'master'
QUERY = """
SELECT COUNT(*)
FROM master
"""

# the read_sql_query runs the sql query and returns the data as a
dataframe
df1 = pd.read_sql_query(QUERY,conn)
df1.head()
```

```
   COUNT(*)
0      4387
```

Demo: How to list all tables

```
# print all the tables names in the database
QUERY = """
SELECT name as Table_Name FROM
sqlite_master WHERE
type = 'table'
"""
# the read_sql_query runs the sql query and returns the data as a
dataframe
pd.read_sql_query(QUERY,conn)
```

```
  Table_Name
0     MASTER
```

Demo: How to run a group by query

Hands-on Lab

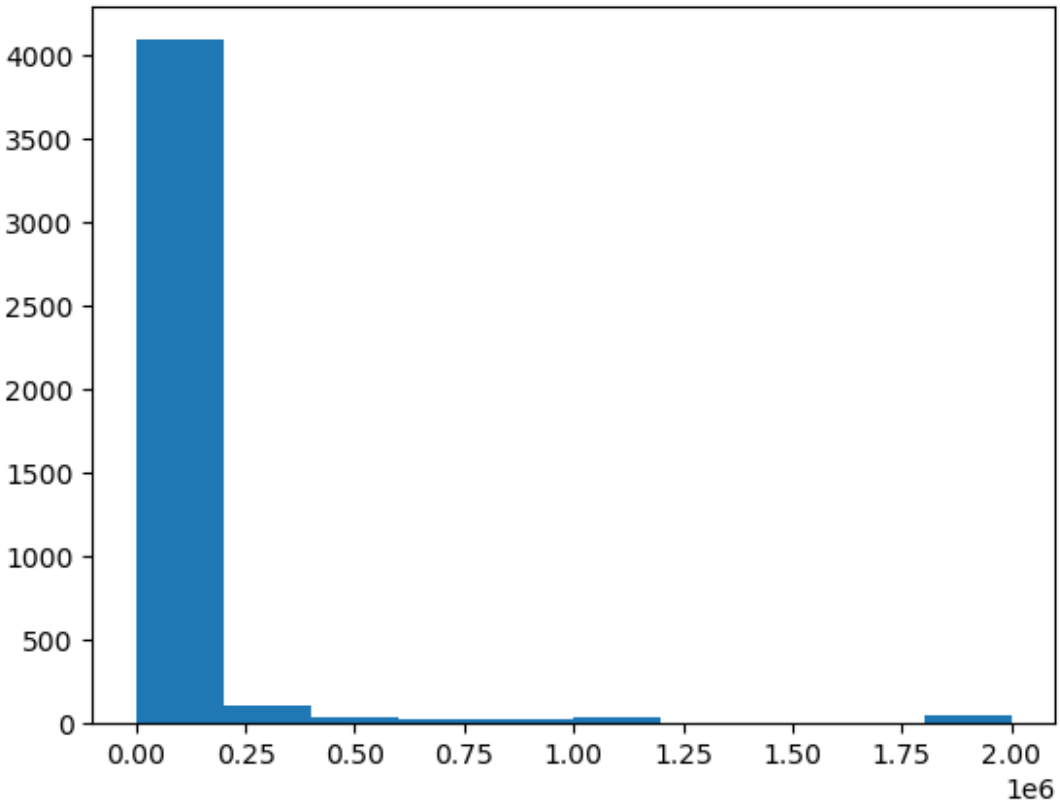
Visualizing distribution of data

Histograms

Plot a histogram of `ConvertedComp.`

```
# your code goes here
plt.hist(df['ConvertedComp'])
```

```
(array([4.094e+03, 1.060e+02, 3.600e+01, 2.900e+01, 2.200e+01,
3.900e+01,
        3.000e+00, 6.000e+00, 4.000e+00, 4.800e+01]),
 array([      0.,  200000.,  400000.,  600000.,  800000., 1000000.,
        1200000., 1400000., 1600000., 1800000., 2000000.]),
 <BarContainer object of 10 artists>)
```

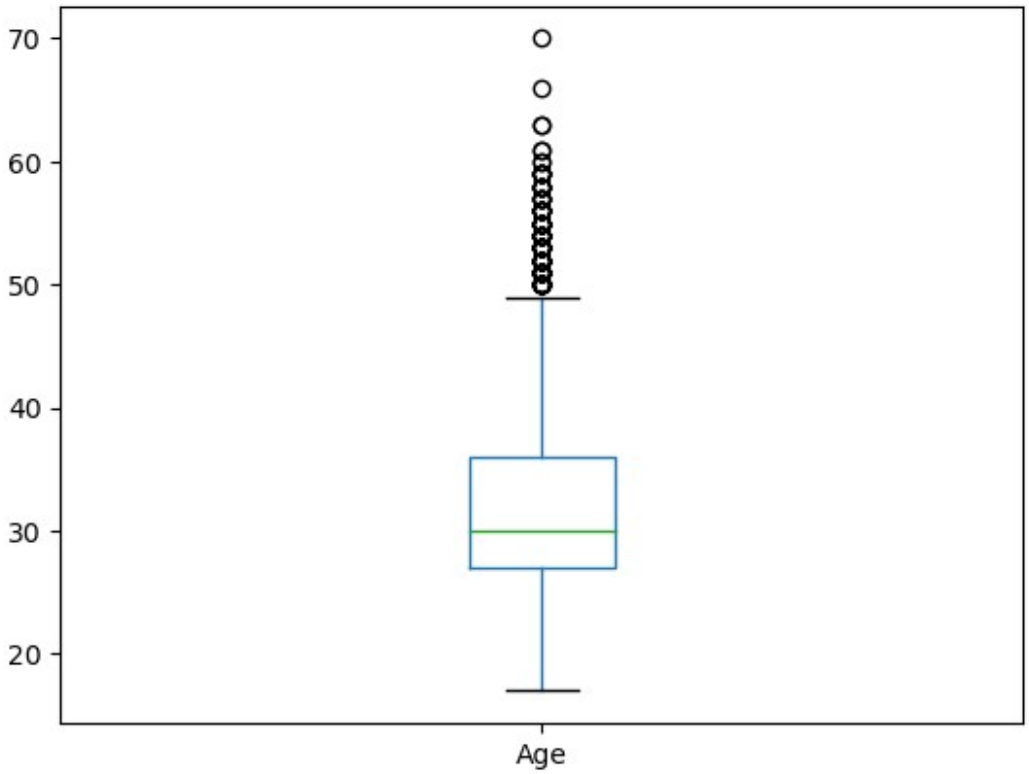


Box Plots

Plot a box plot of `Age.`

```
# your code goes here
df.boxplot(column =['Age'],grid=False)
```

```
<AxesSubplot:>
```



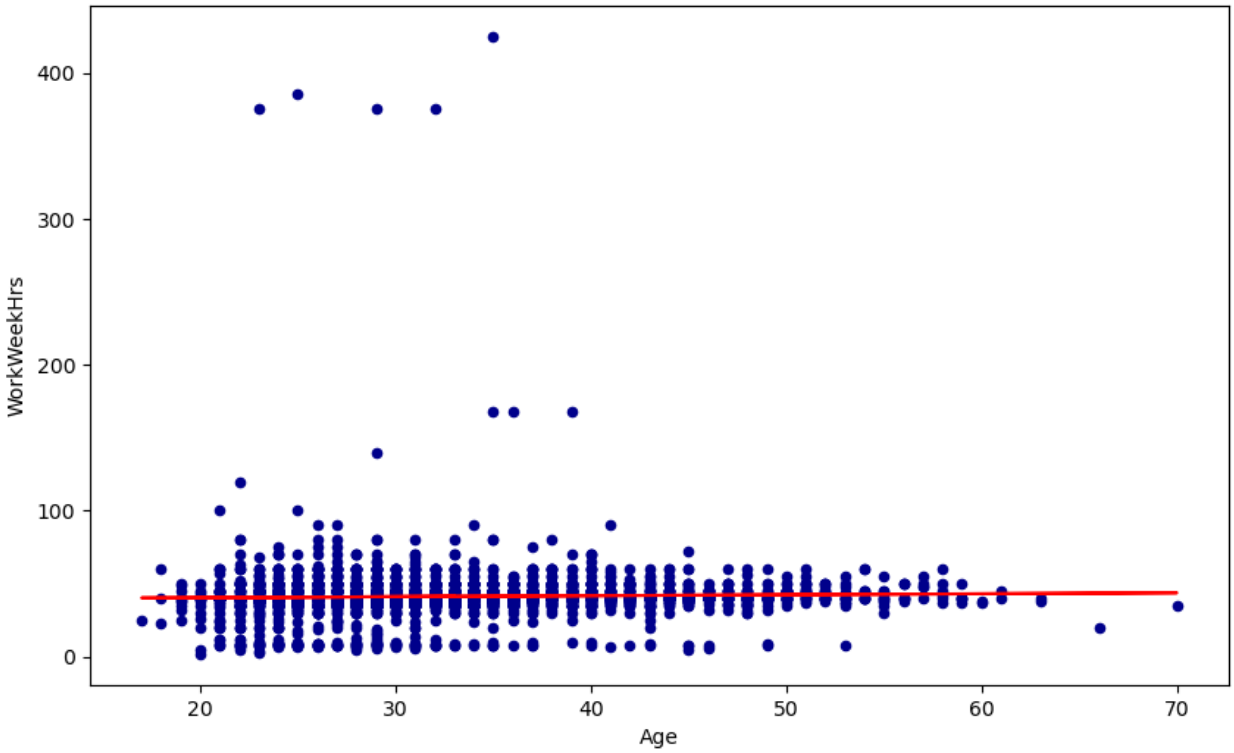
Visualizing relationships in data

Scatter Plots

Create a scatter plot of `Age` and `WorkWeekHrs.`

```
# your code goes here
df.plot(kind='scatter', x='Age', y='WorkWeekHrs', figsize=(10, 6),
color='darkblue')
# plot line of best fit
x = df['Age']      # year on x-axis
y = df['WorkWeekHrs']     # total on y-axis
fit = np.polyfit(x, y, deg=1)
fit

plt.plot(x, fit[0] * x + fit[1], color='red') # recall that x is the
Years
plt.annotate('y={0:.0f} x + {1:.0f}'.format(fit[0], fit[1]), xy=(2000,
150000))
plt.show()
```



Bubble Plots

Create a bubble plot of `WorkWeekHrs` and `Age`, use `Age` column as bubble size.

Hint: Use plotly.express to create a bubble chart

```
# your code goes here
fig = px.scatter(df, x="WorkWeekHrs", y="Age", size="Age", log_x=True,
size_max=10)
fig.show()
```

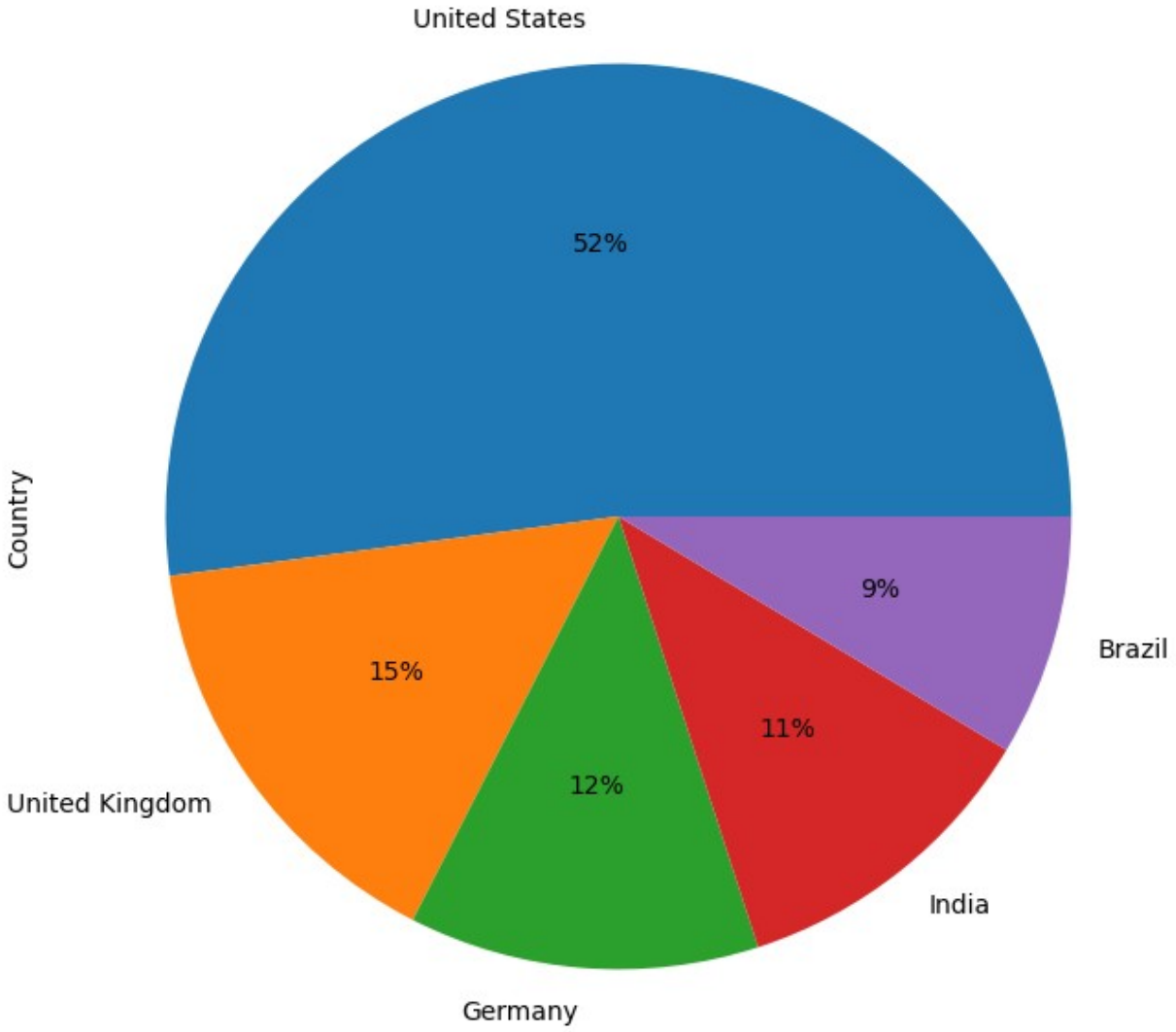
Visualizing composition of data

Pie Charts

Create a pie chart of the top 5 Country that respondents filled the survey . Display percentages of each database on the pie chart.

```
# your code goes here
df_pie = df['Country'].value_counts()
df_pie=df_pie.head(5)
df_pie.plot(kind='pie', figsize=(8,8), autopct='%1.0f%%')
```

```
<AxesSubplot:ylabel='Country'>
```



Visualizing comparison of data

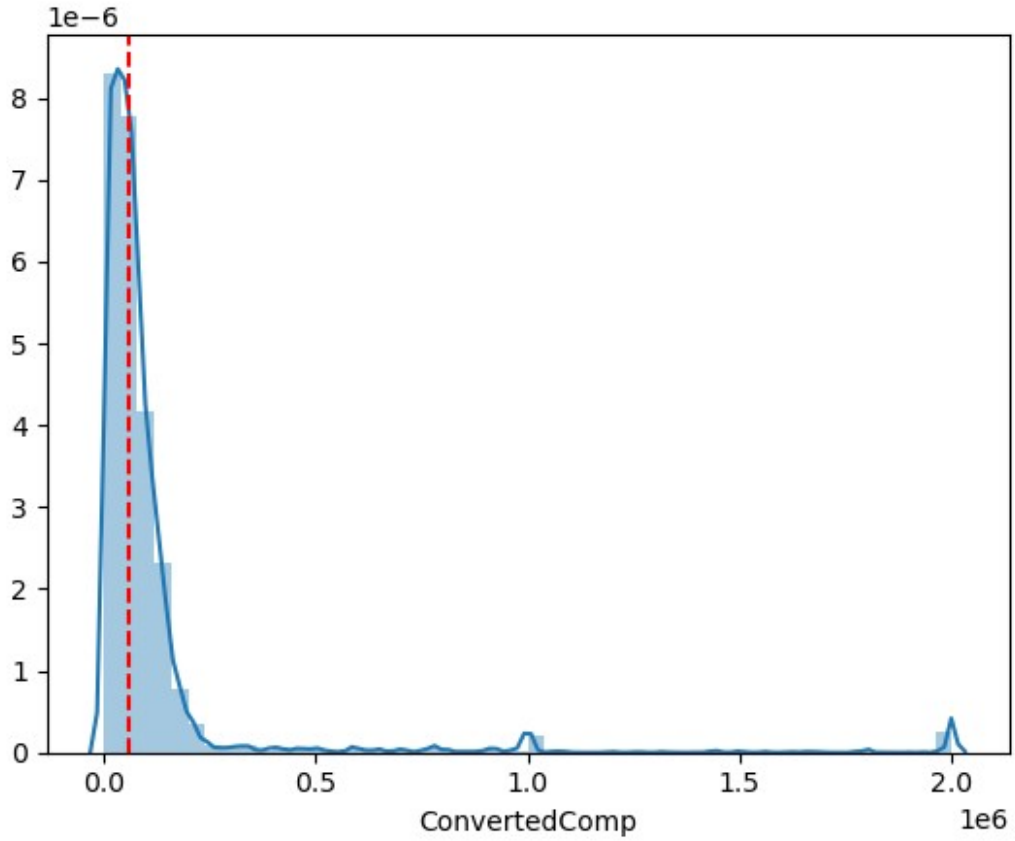
Line Chart

Draw distribution plot for ConverteComp and plot the median

Hint: Use seaborn library for distribution plot

```
# your code goes here
median=df['ConvertedComp'].median()
sns.distplot(df["ConvertedComp"])
plt.axvline(median, color='r', linestyle='--')
```

```
<matplotlib.lines.Line2D at 0x7fbe80310c50>
```

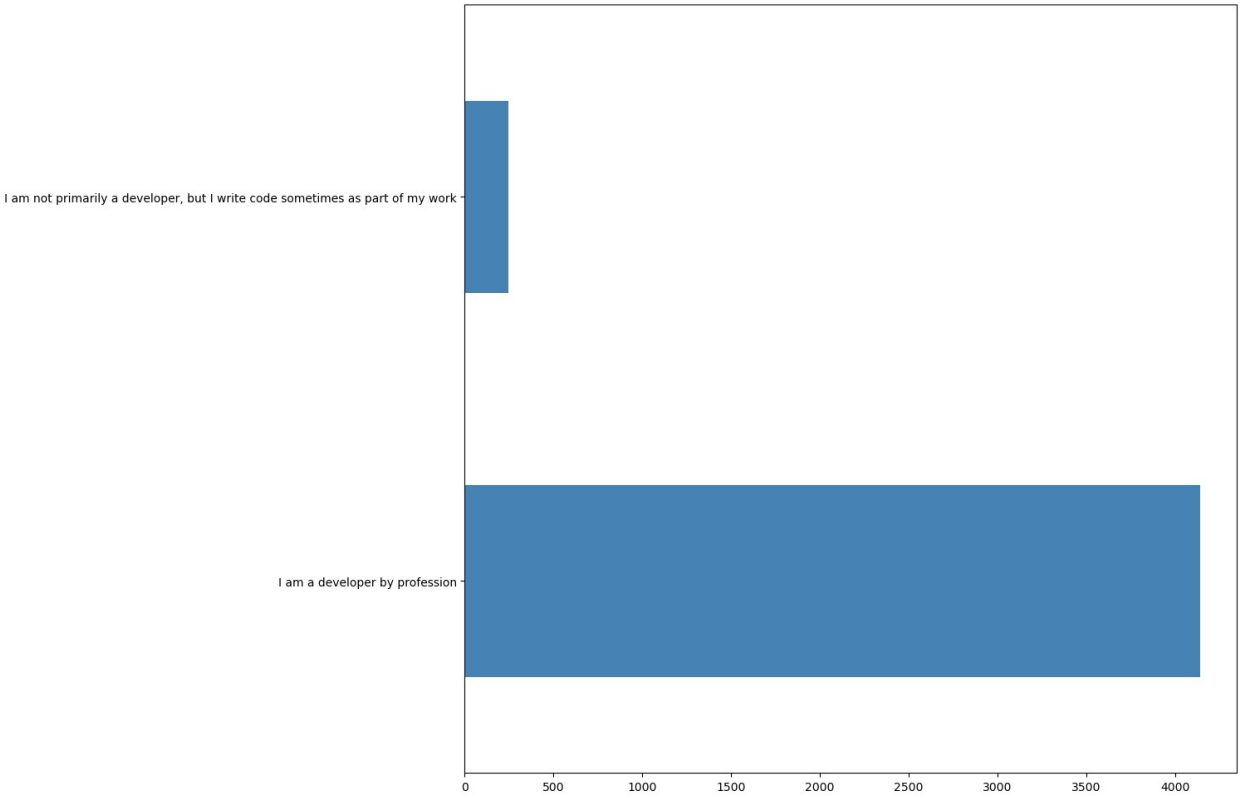


Bar Chart

Create a horizontal bar chart using column `MainBranch.`

```
# your code goes here
df_mbranch = df['MainBranch'].value_counts()
df_mbranch
df_mbranch.plot(kind='barh', figsize=(12, 12), color='steelblue')
```

```
<AxesSubplot:>
```



Close the database connection.

```
conn.close()
```