Exploratory Data Analysis Lab

Estimated time needed: **30** minutes

In this module you get to work with the cleaned dataset from the previous module.

In this assignment you will perform the task of exploratory data analysis. You will find out the distribution of data, presence of outliers and also determine the correlation between different columns in the dataset.

Objectives

In this lab you will perform the following:

- Identify the distribution of data in the dataset.
- Identify outliers in the dataset.
- Remove outliers from the dataset.
- Identify correlation between features in the dataset.

Hands on Lab

Import the pandas module.

```
%pip install seaborn
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

Load the dataset into a dataframe.

```
df = pd.read_csv('/drive/labs/Capstone_edX/Module
3/m3_survey_data.csv')
df
```

```
       Respondent                      MainBranch Hobbyist  \
0               4  I am a developer by profession       No
1               9  I am a developer by profession      Yes
2              13  I am a developer by profession      Yes
3              16  I am a developer by profession      Yes
4              17  I am a developer by profession      Yes
...           ...                             ...      ...
11393       25136  I am a developer by profession      Yes
```

```
11394       25137  I am a developer by profession      Yes
11395       25138  I am a developer by profession      Yes
11396       25141  I am a developer by profession      Yes
11397       25142  I am a developer by profession      Yes

                                                 OpenSourcer  \
0                                                      Never
1                                 Once a month or more often
2      Less than once a month but more than once per ...
3                                                      Never
4      Less than once a month but more than once per ...
...                                                      ...
11393                                                  Never
11394                                                  Never
11395                               Less than once per year
11396  Less than once a month but more than once per ...
11397  Less than once a month but more than once per ...

                                                  OpenSource
Employment  \
0      The quality of OSS and closed source software ...  Employed
full-time
1      The quality of OSS and closed source software ...  Employed
full-time
2      OSS is, on average, of HIGHER quality than pro...  Employed
full-time
3      The quality of OSS and closed source software ...  Employed
full-time
4      The quality of OSS and closed source software ...  Employed
full-time
...                                                  ...
...
11393  OSS is, on average, of HIGHER quality than pro...  Employed
full-time
11394  The quality of OSS and closed source software ...  Employed
full-time
11395  The quality of OSS and closed source software ...  Employed
full-time
11396  OSS is, on average, of LOWER quality than prop...  Employed
full-time
11397  OSS is, on average, of HIGHER quality than pro...  Employed
full-time

               Country Student  \
0       United States      No
1         New Zealand      No
2       United States      No
3      United Kingdom      No
4           Australia      No
...                ...     ...
```

```
11393   United States      No
11394          Poland      No
11395   United States      No
11396     Switzerland      No
11397  United Kingdom      No

                                                  EdLevel  \
0                Bachelor’s degree (BA, BS, B.Eng., etc.)
1      Some college/university study without earning ...
2            Master’s degree (MA, MS, M.Eng., MBA, etc.)
3            Master’s degree (MA, MS, M.Eng., MBA, etc.)
4               Bachelor’s degree (BA, BS, B.Eng., etc.)
...                                                   ...
11393        Master’s degree (MA, MS, M.Eng., MBA, etc.)
11394        Master’s degree (MA, MS, M.Eng., MBA, etc.)
11395        Master’s degree (MA, MS, M.Eng., MBA, etc.)
11396  Secondary school (e.g. American high school, G...
11397          Other doctoral degree (Ph.D, Ed.D., etc.)

                                           UndergradMajor  ...  \
0      Computer science, computer engineering, or sof...  ...
1      Computer science, computer engineering, or sof...  ...
2      Computer science, computer engineering, or sof...  ...
3                                                    NaN  ...
4      Computer science, computer engineering, or sof...  ...
...                                                  ...  ...
11393  Computer science, computer engineering, or sof...  ...
11394  Computer science, computer engineering, or sof...  ...
11395  Computer science, computer engineering, or sof...  ...
11396                                                NaN  ...
11397  A natural science (ex. biology, chemistry, phy...  ...

                                  WelcomeChange  \
0       Just as welcome now as I felt last year
1       Just as welcome now as I felt last year
2      Somewhat more welcome now than last year
3       Just as welcome now as I felt last year
4       Just as welcome now as I felt last year
...                                        ...
11393   Just as welcome now as I felt last year
11394     A lot more welcome now than last year
11395     A lot more welcome now than last year
11396  Somewhat less welcome now than last year
11397   Just as welcome now as I felt last year

                                          SONewContent   Age Gender
Trans  \
0      Tech articles written by other developers;Indu...  22.0    Man
No
1                                                    NaN  23.0    Man
```

```
No
2      Tech articles written by other developers;Cour...  28.0    Man
No
3      Tech articles written by other developers;Indu...  26.0    Man
No
4      Tech articles written by other developers;Indu...  29.0    Man
No
...                                                  ...   ...    ...
...
11393  Tech articles written by other developers;Cour...  36.0    Man
No
11394  Tech articles written by other developers;Tech...  25.0    Man
No
11395  Tech articles written by other developers;Indu...  34.0    Man
No
11396                                                NaN  25.0    Man
No
11397  Tech articles written by other developers;Tech...  30.0    Man
No

                     Sexuality                                 Ethnicity
\
0      Straight / Heterosexual           White or of European descent

1                     Bisexual           White or of European descent

2      Straight / Heterosexual           White or of European descent

3      Straight / Heterosexual           White or of European descent

4      Straight / Heterosexual  Hispanic or Latino/Latina;Multiracial

...                          ...                                   ...

11393  Straight / Heterosexual           White or of European descent

11394  Straight / Heterosexual           White or of European descent

11395  Straight / Heterosexual           White or of European descent

11396  Straight / Heterosexual           White or of European descent

11397                 Bisexual           White or of European descent


      Dependents           SurveyLength                   SurveyEase
0             No  Appropriate in length                         Easy
1             No  Appropriate in length  Neither easy nor difficult
2            Yes  Appropriate in length                         Easy
3             No  Appropriate in length  Neither easy nor difficult
```

```
4             No  Appropriate in length                          Easy
...          ...                     ...                           ...
11393         No  Appropriate in length                     Difficult
11394         No  Appropriate in length  Neither easy nor difficult
11395        Yes               Too long                          Easy
11396         No  Appropriate in length                          Easy
11397         No  Appropriate in length                          Easy

[11398 rows x 85 columns]
```

Distribution

Determine how the data is distributed

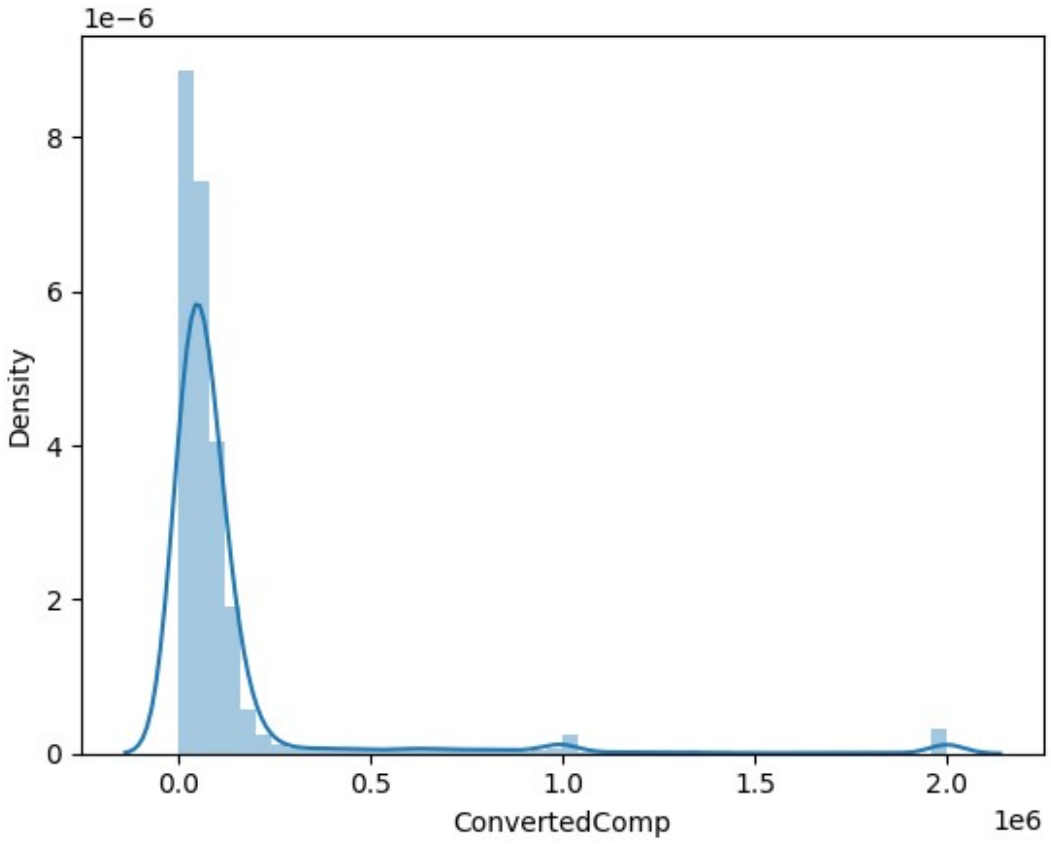
The column `ConvertedComp` contains Salary converted to annual USD salaries using the exchange rate on 2019-02-01.

This assumes 12 working months and 50 working weeks.

Plot the distribution curve for the column `ConvertedComp`.

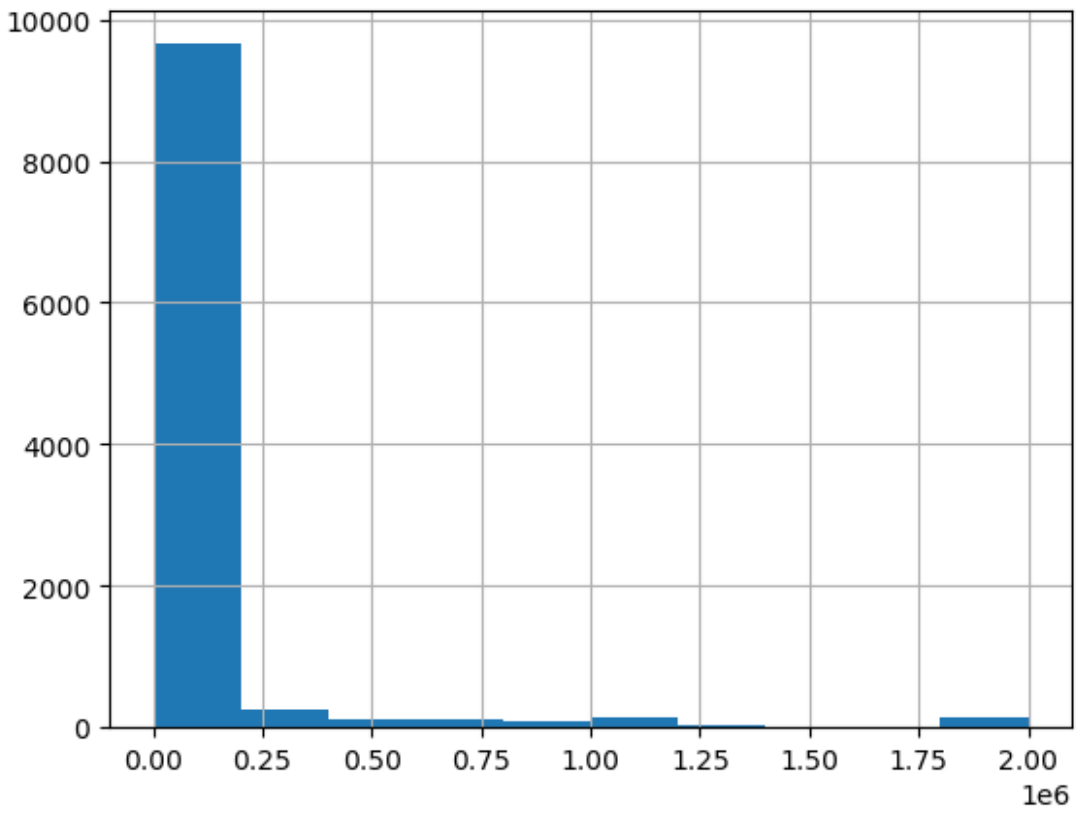
```
# your code goes here
sns.distplot(df["ConvertedComp"])
```

```
<ipython-input-27-6c87e4cbaede>:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn
v0.14.0.

Please adapt your code to use either `displot` (a figure-level
function with
similar flexibility) or `histplot` (an axes-level function for
histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df["ConvertedComp"])

<AxesSubplot:xlabel='ConvertedComp', ylabel='Density'>
```



Plot the histogram for the column `ConvertedComp`.

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



What is the median of the column `ConvertedComp`?

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

```
57745.0
```

How many responders identified themselves only as a **Man**?

```
# your code goes here
df['Gender'].value_counts()
```

```
Man                                                            10480
Woman                                                            731
Non-binary, genderqueer, or gender non-conforming                 63
Man;Non-binary, genderqueer, or gender non-conforming             26
Woman;Non-binary, genderqueer, or gender non-conforming           14
Woman;Man                                                          9
Woman;Man;Non-binary, genderqueer, or gender non-conforming        2
Name: Gender, dtype: int64
```

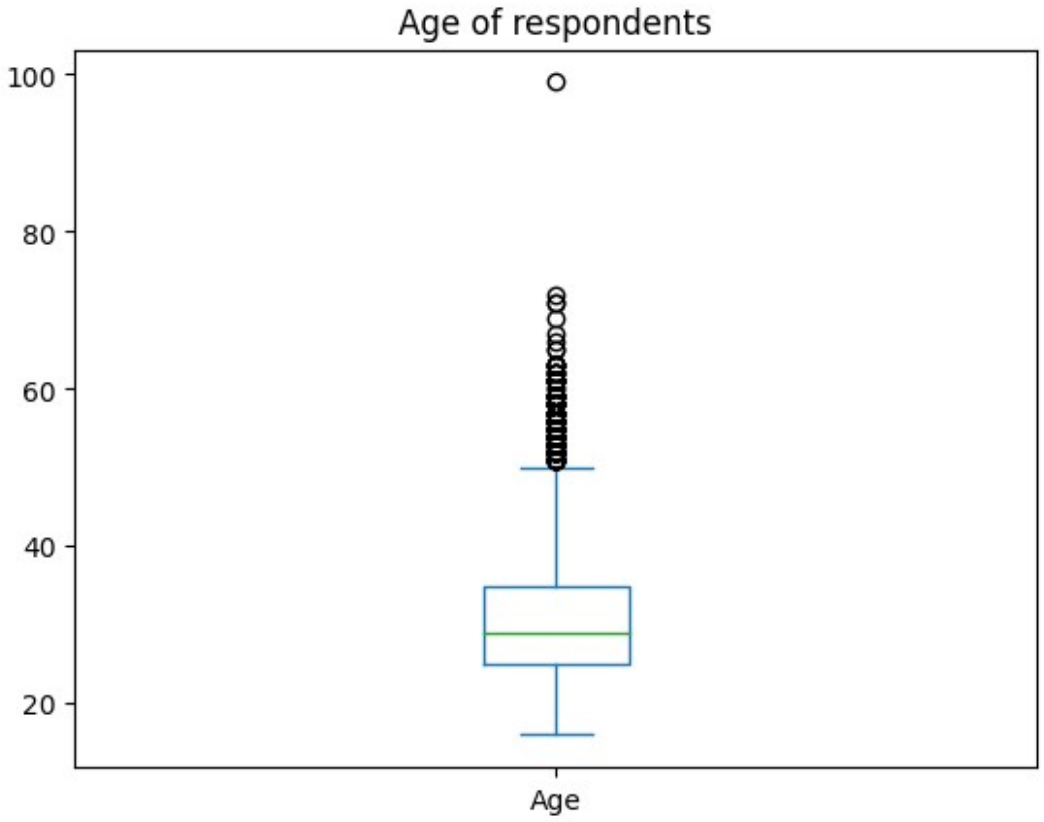
Find out the median ConvertedComp of responders identified themselves only as a **Woman**?

```
# your code goes here
df[['Gender', 'ConvertedComp']]

woman_median = df[df['Gender'] == "Woman"]
woman_median

woman_median['ConvertedComp'].median()
```

```
57708.0
```

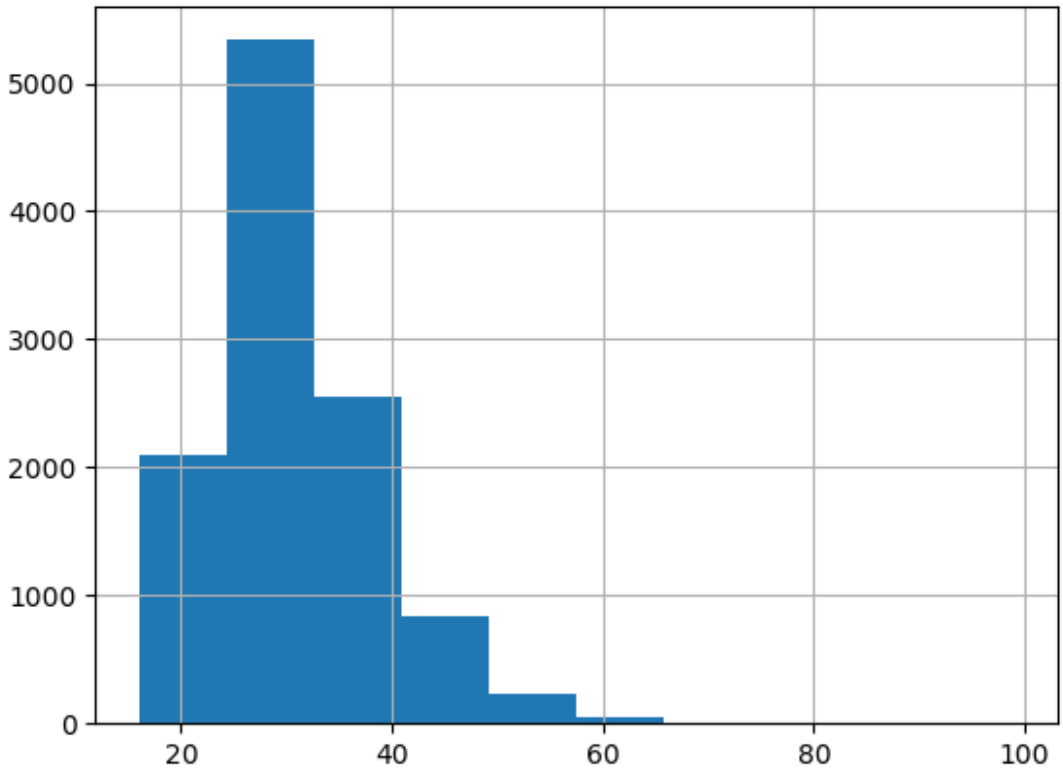
Give the five number summary for the column `Age`?

```
# your code goes here
df['Age'].plot(kind='box', title='Age of respondents')
plt.show()
```



Plot a histogram of the column `Age`.

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

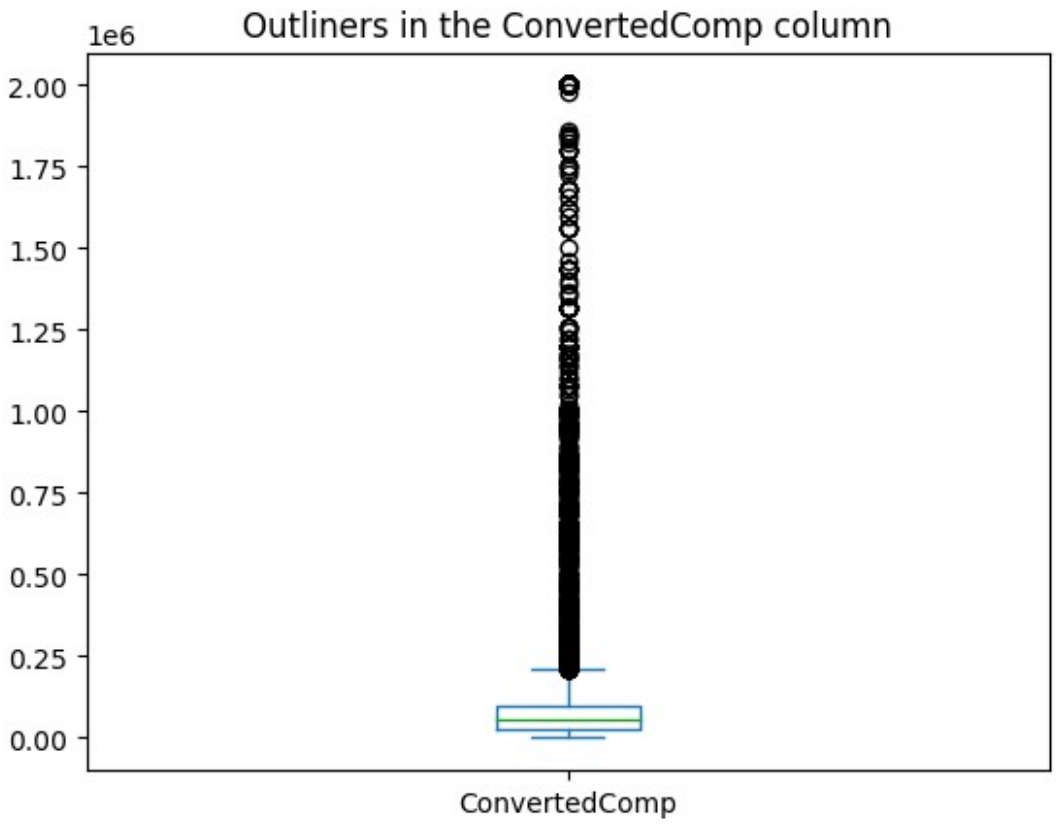


Outliers

Finding outliers

Find out if outliers exist in the column `ConvertedComp` using a box plot?

```
# your code goes here
df['ConvertedComp'].plot(kind='box', title='Outliners in the
ConvertedComp column')
plt.show()
```



Find out the Inter Quartile Range for the column `ConvertedComp`.

```
Q1 = df['ConvertedComp'].quantile(0.25)
Q3 = df['ConvertedComp'].quantile(0.75)
IQR = Q3 - Q1
print("Interquartile Range (IQR):", IQR)
```

```
Interquartile Range (IQR): 73132.0
```

Find out the upper and lower bounds.

```
# your code goes here
Lower = Q1 - 1.5 * IQR
Upper = Q3 + 1.5 * IQR
print("The lower bound is", Lower)
print("The upper bound is", Upper)
```

```
The lower bound is -82830.0
The upper bound is 209698.0
```

Identify how many outliers are there in the `ConvertedComp` column.

```
def find_outliers_IQR(df):
```

```
    q1=df.quantile(0.25)

    q3=df.quantile(0.75)

    IQR=q3-q1

    outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]

    return outliers

# your code goes here
outliers = find_outliers_IQR(df['ConvertedComp'])

print('number of outliers: '+ str(len(outliers)))

print('max outlier value: '+ str(outliers.max()))

print('min outlier value: '+ str(outliers.min()))

outliers
```

```
number of outliers: 879
max outlier value: 2000000.0
min outlier value: 209892.0

3         455352.0
13       1100000.0
45        229016.0
46       2000000.0
60       1000000.0
           ...
11296     840000.0
11303    1000000.0
11350     300000.0
11353     260000.0
11369     701196.0
Name: ConvertedComp, Length: 879, dtype: float64
```

Create a new dataframe by removing the outliers from the `ConvertedComp` column.

```
df_new = df['ConvertedComp'][~((df['ConvertedComp'] < Lower) |
(df['ConvertedComp'] > Upper))]
df_new
```

```
0         61000.0
1         95179.0
2         90000.0
4         65277.0
5         31140.0
           ...
11393    130000.0
```

```
11394     19880.0
11395    105000.0
11396     80371.0
11397         NaN
Name: ConvertedComp, Length: 10519, dtype: float64
```

```
df_new.mean()
```

```
59883.20838915799
```

Correlation

Finding correlation

Find the correlation between `Age` and all other numerical columns.

```
# your code goes here
df[['Respondent', 'YearsCode', 'Age1stCode', 'YearsCodePro',
'CompTotal', 'ConvertedComp', 'WorkWeekHrs', 'CodeRevHrs',
'Age']].corr()
```

```
<ipython-input-41-b9a78f9f5145>:2: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  df[['Respondent', 'YearsCode', 'Age1stCode', 'YearsCodePro',
'CompTotal', 'ConvertedComp', 'WorkWeekHrs', 'CodeRevHrs',
'Age']].corr()
```

```
                Respondent  CompTotal  ConvertedComp  WorkWeekHrs
CodeRevHrs  \
Respondent        1.000000  -0.013490       0.002181    -0.015314
0.004621
CompTotal        -0.013490   1.000000       0.001037     0.003510
0.007063
ConvertedComp     0.002181   0.001037       1.000000     0.021143   -
0.033865
WorkWeekHrs      -0.015314   0.003510       0.021143     1.000000
0.026517
CodeRevHrs        0.004621   0.007063      -0.033865     0.026517
1.000000
Age               0.004041   0.006970       0.105386     0.036518   -
0.020469

                     Age
Respondent      0.004041
CompTotal       0.006970
ConvertedComp   0.105386
WorkWeekHrs     0.036518
```

```
CodeRevHrs    -0.020469
Age            1.000000
```

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