M3ExploratoryDataAnalysis-lab

June 24, 2022

1 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.

1.1 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.

1.2 Hands on Lab

Import the pandas module.

[1]:

```
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

Load the dataset into a dataframe.

[2]:

```
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.
 ↪cloud/IBM-DA0321EN-SkillsNetwork/LargeData/m2_survey_data.csv")
```

1.3 Distribution

1.3.1 Determine how the data is distributed

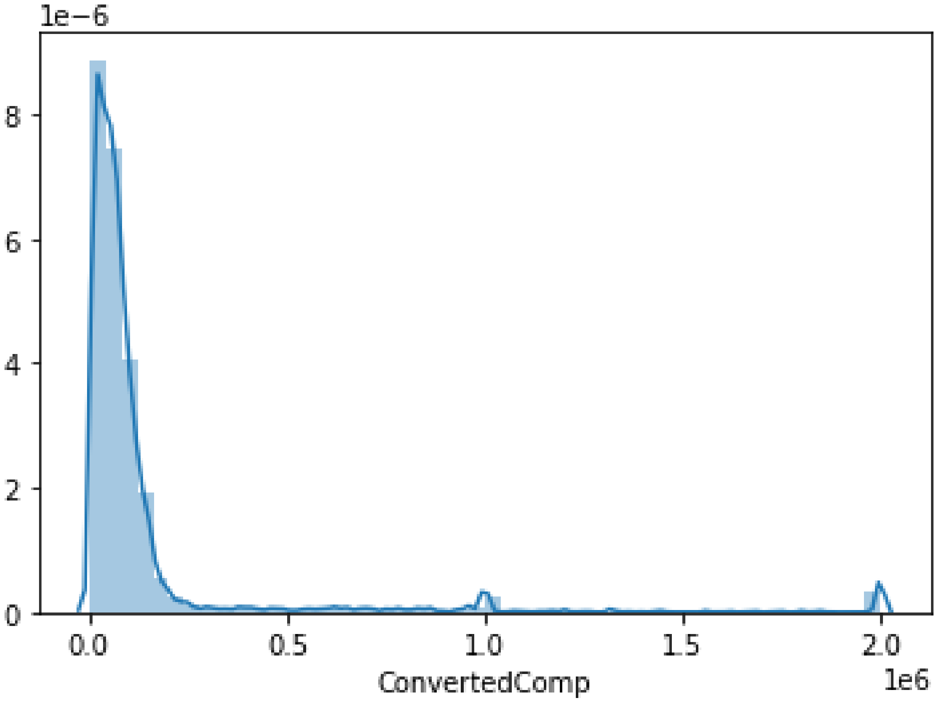
The column `ConvertedComp` contains Salary converted to annual USD salaries using the exchange rate on 2019-02-01.

This assumes 12 working months and 50 working weeks.

Plot the distribution curve for the column `ConvertedComp`.

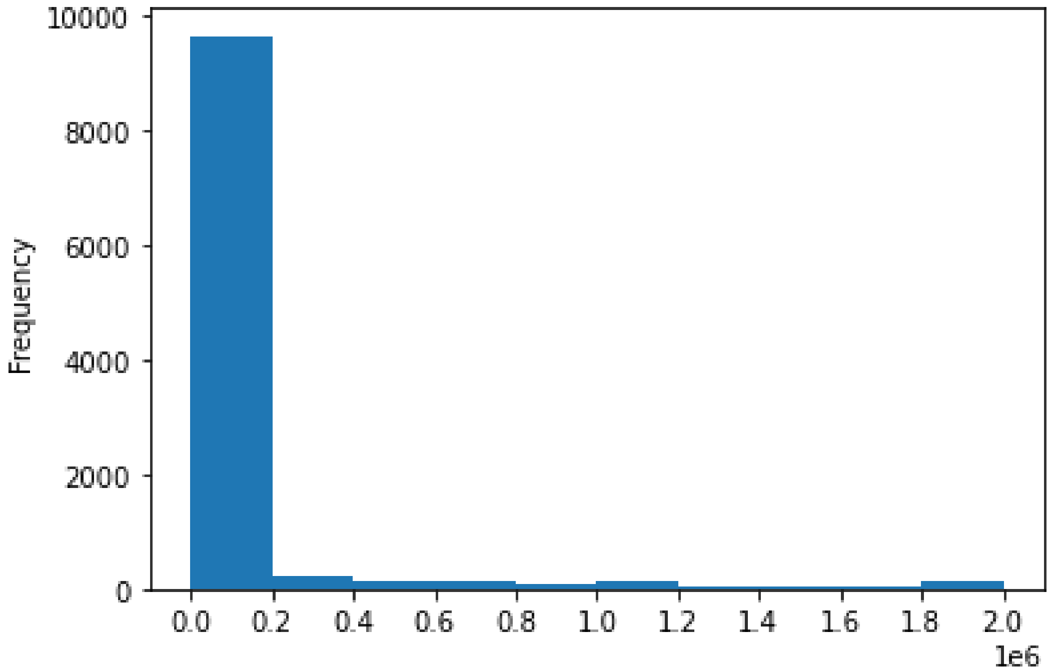
```
[3]: #drop the rows that have a null value in the ConvertedComp column
     df = df.dropna(subset=['ConvertedComp'])
```

```
[4]: # your code goes here
     ax = sns.distplot(df['ConvertedComp'], kde = True)
```



Plot the histogram for the column `ConvertedComp`.

```
[6]: # your code goes here
     count, bin_edges = np.histogram(df['ConvertedComp'])
     df['ConvertedComp'].plot(kind='hist', xticks=bin_edges)
     plt.show()
```



What is the median of the column `ConvertedComp`?

[7]:
```
df['ConvertedComp'].median()
```

[7]:
```
57745.0
```

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

[10]:
```
# your code goes here
df['Gender'].value_counts()
df['Gender'].loc[df['Gender'] == 'Man'].count()
```

[10]:
```
9725
```

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

[11]:
```
# your code goes here
df.loc[df['Gender'] == 'Woman']['ConvertedComp'].median()
```

[11]:
```
57708.0
```

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

Double click here for hint.

[12]:

```
# your code goes here
print('Max: ', df['Age'].max())
print('Min: ', df['Age'].min())
print('Median: ', df['Age'].median())
print('25% quantile: ', df['Age'].quantile(.25))
print('75% quantile: ', df['Age'].quantile(.75))
```

```
Max:  99.0
Min:  16.0
Median:  29.0
25% quantile:  25.0
75% quantile:  35.0
```

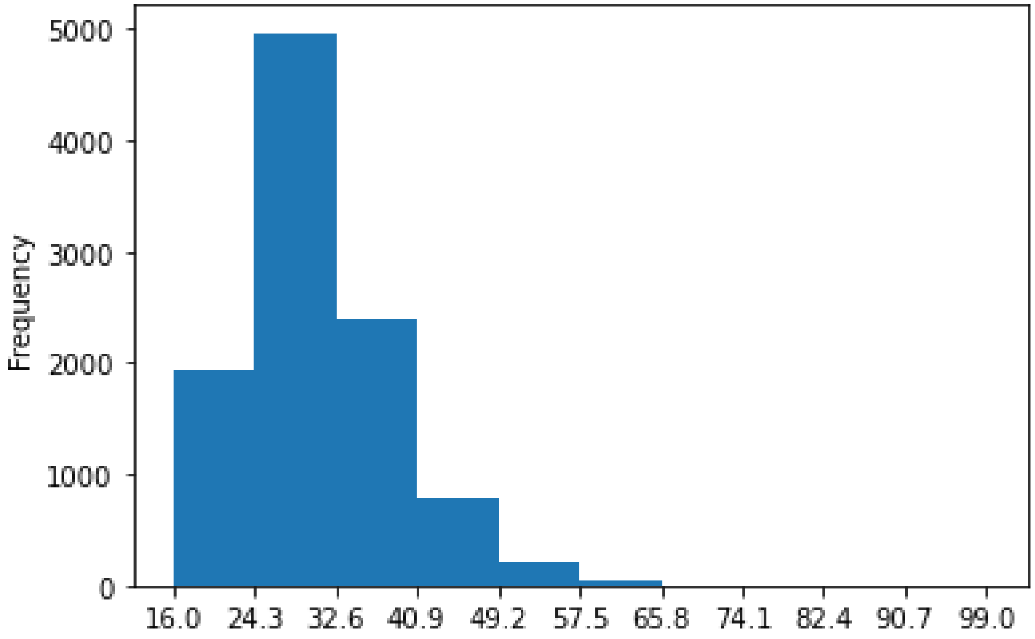
Plot a histogram of the column `Age`.

[13]:

```
df = df.dropna(subset=['Age'])
```

[14]:

```
# your code goes here
count, bin_edges = np.histogram(df['Age'])
df['Age'].plot(kind='hist', xticks=bin_edges)
plt.show()
```

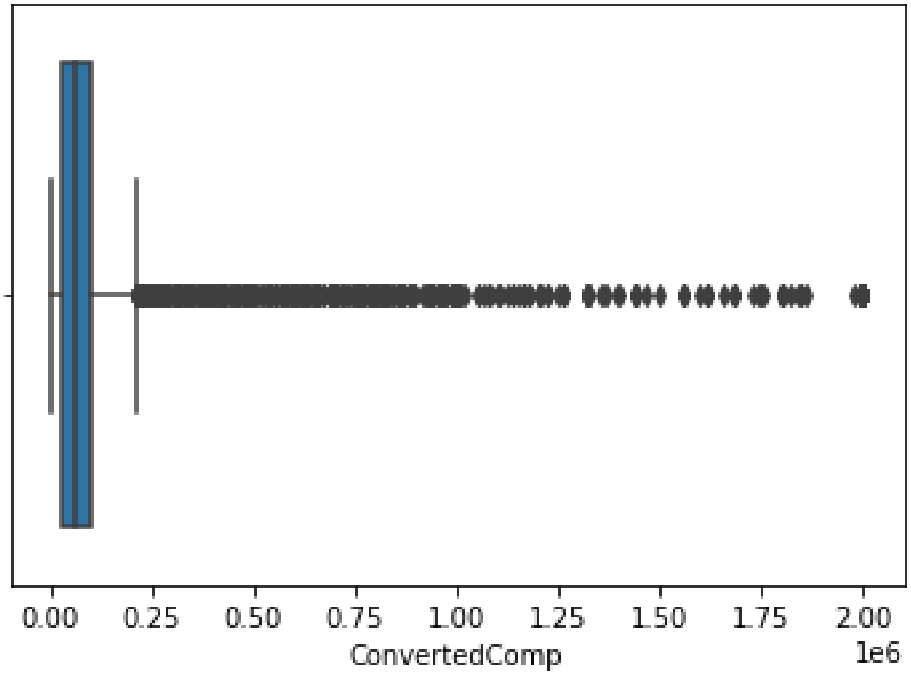


1.4 Outliers

1.4.1 Finding outliers

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

[15]:
```
# your code goes here
ax = sns.boxplot(df['ConvertedComp'])
```



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

[16]:
```
# your code goes here
df['ConvertedComp'].quantile(.75) - df['ConvertedComp'].quantile(.25)
```

[16]:
```
73165.5
```

[17]:
```
from scipy import stats
IQR = stats.iqr(df['ConvertedComp'], interpolation = 'midpoint')
IQR
```

[17]:
```
73155.0
```

Find out the upper and lower bounds.

[18]:
```
# your code goes here
print('Lower bound: ', df['ConvertedComp'].quantile(.25))
lower = df['ConvertedComp'].quantile(.25)
print('Upper bound: ', df['ConvertedComp'].quantile(.75))
upper = df['ConvertedComp'].quantile(.75)
```

```
Lower bound:  26834.5
Upper bound:  100000.0
```

Identify how many outliers are there in the ConvertedComp column.

[59]:

```
# your code goes here
df['ConvertedComp'].loc[(df['ConvertedComp'] > upper) | (df['ConvertedComp'] <␣
 ↪lower)].count()
```

[59]:

```
5081
```

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

[23]:

```
# your code goes here
outliers = df.loc[(df['ConvertedComp'] > upper) | (df['ConvertedComp'] < lower)]
df_new = df.drop(outliers.index)
df_new.shape
```

[23]:

```
(5273, 85)
```

[24]:

```
print('original df shape: ', df.shape)
print('outliers df shape: ', outliers.shape)
print('new df shape: ', df_new.shape)
```

```
original df shape:  (10354, 85)
outliers df shape:  (5081, 85)
new df shape:  (5273, 85)
```

1.5 Correlation

1.5.1 Finding correlation

Find the correlation between Age and all other numerical columns.

[26]:

```
# your code goes here
df_new.corr(method ='pearson')['Age']
```

[26]:

```
Respondent       0.010152
CompTotal        0.024843
ConvertedComp    0.204733
WorkWeekHrs      0.004258
CodeRevHrs       0.055230
Age              1.000000
Name: Age, dtype: float64
```

1.6 Authors

Ramesh Sannareddy

1.6.1 Other Contributors

Rav Ahuja

1.7 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-17	0.1	Ramesh Sannareddy	Created initial version of the lab