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.

In [2]:
```
import pandas as pd
```

Load the dataset into a dataframe.

In [21]:
```
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/LargeData/m2_surve
```

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.

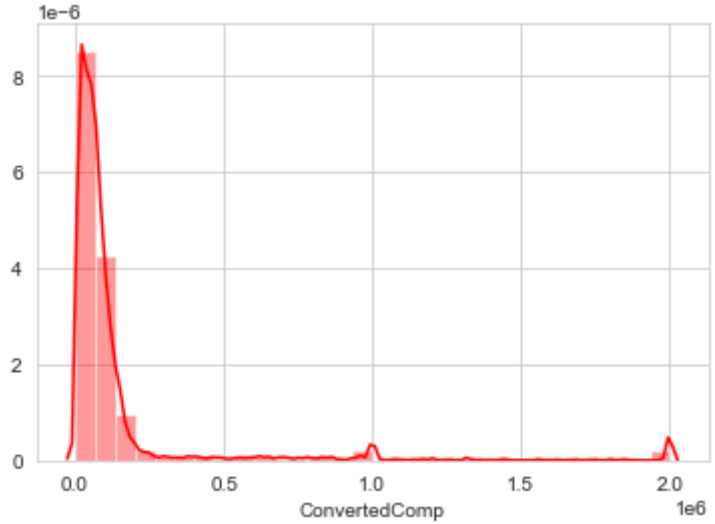
This assumes 12 working months and 50 working weeks.

Plot the distribution curve for the column `ConvertedComp`.

In [4]:
```
import seaborn as sns
%matplotlib inline
import matplotlib.pyplot as plt
```

In [5]:
```
sns.set_style('whitegrid')
sns.distplot(df['ConvertedComp'], kde = True, color ='red', bins = 30)
```

Out[5]:
```
<AxesSubplot:xlabel='ConvertedComp'>
```

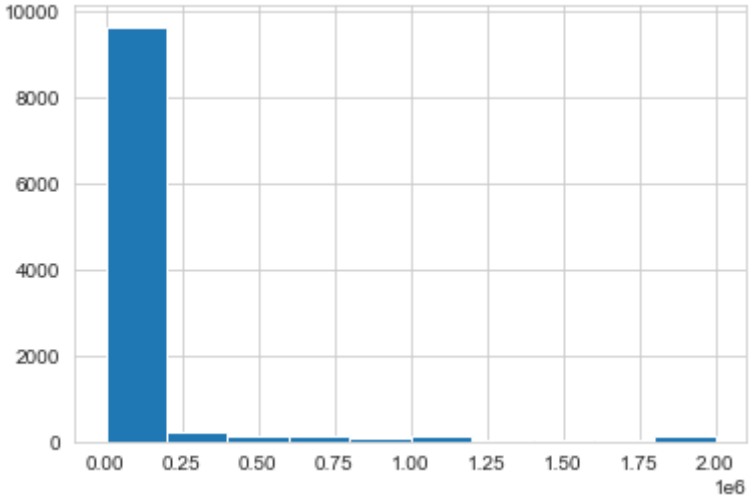


Plot the histogram for the column `ConvertedComp`.

In [6]:
```
plt.hist(df['ConvertedComp'])
```

Out[6]:
```
(array([9659.,  238.,  115.,  125.,   99.,  131.,   34.,   15.,   15.,
         151.]),
```

```
array([      0.,  200000.,  400000.,  600000.,  800000., 1000000.,
       1200000., 1400000., 1600000., 1800000., 2000000.]),
<BarContainer object of 10 artists>)
```



What is the median of the column `ConvertedComp` ?

In [7]:
```
df["ConvertedComp"].median()
```

Out[7]: 57745.0

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

In [8]:
```
df["Gender"].value_counts().Man
```

Out[8]: 10480

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

In [9]:
```
df.loc[df['Gender'] == 'Woman', 'ConvertedComp'].median()
```

Out[9]: 57708.0

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

Double click here for hint.

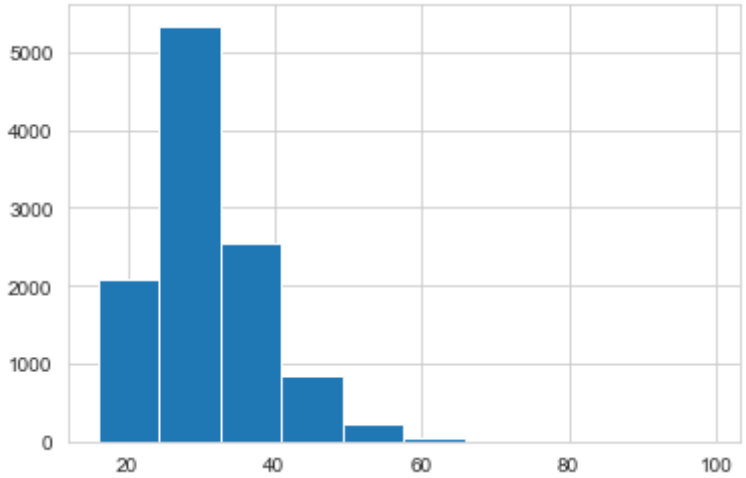
In [10]:
```
df["Age"].describe()
```

Out[10]:
```
count    11111.000000
mean        30.778895
std          7.393686
min         16.000000
25%         25.000000
50%         29.000000
75%         35.000000
max         99.000000
Name: Age, dtype: float64
```

Plot a histogram of the column `Age`.

In [11]:
```
plt.hist(df['Age'])
```

Out[11]:
```
(array([2.094e+03, 5.337e+03, 2.557e+03, 8.420e+02, 2.250e+02, 4.900e+01,
        6.000e+00, 0.000e+00, 0.000e+00, 1.000e+00]),
 array([16. , 24.3, 32.6, 40.9, 49.2, 57.5, 65.8, 74.1, 82.4, 90.7, 99. ]),
 <BarContainer object of 10 artists>)
```



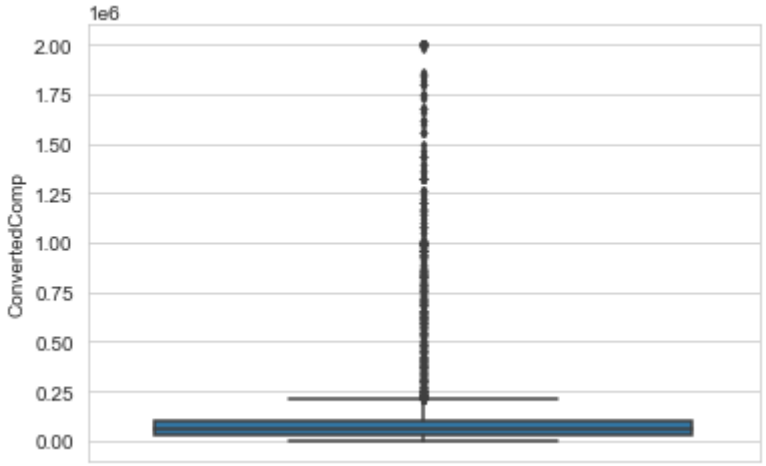
Outliers

Finding outliers

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

In [12]:
```
sns.boxplot(x=df["ConvertedComp"],data=df,orient="v",width=0.8,fliersize=2.9)
```

Out[12]:
```
<AxesSubplot:ylabel='ConvertedComp'>
```



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

In [13]:
```
df['ConvertedComp'].dropna(axis=0, inplace=True)
Q1,Q3 = df['ConvertedComp'].quantile(.25),df['ConvertedComp'].quantile(.75)
IQR = Q3 - Q1
print('The Inter Quartile Range for ConvertedComp: ', IQR)
```

```
The Inter Quartile Range for ConvertedComp:  73132.0
```

Find out the upper and lower bounds.

In [14]:
```
upper = Q3+(IQR*1.5)
lower = Q1-(IQR*1.5)

print('Upper bound: ', upper)
print('Lower bound: ', lower)
```

```
Upper bound:  209698.0
Lower bound:  -82830.0
```

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

In [15]:
```
outliers = df[((df['ConvertedComp']<lower )| (df['ConvertedComp']>upper))]
outliers["ConvertedComp"].count()
```

Out[15]:
```
879
```

In [16]:
```
df['ConvertedComp']
```

Out[16]:
```
0         61000.0
1         95179.0
2         90000.0
3        455352.0
4         65277.0
           ...
11393    130000.0
11394     19880.0
11395    105000.0
11396     80371.0
11397         NaN
Name: ConvertedComp, Length: 11398, dtype: float64
```

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

In [17]:
```
new_df=df[((df['ConvertedComp']>lower )& (df['ConvertedComp']<upper))]

new_df.describe()
```

Out[17]:

	Respondent	CompTotal	ConvertedComp	WorkWeekHrs	CodeRevHrs	Age
count	9703.000000	9.703000e+03	9703.000000	9664.000000	7612.000000	9493.000000
mean	12501.007317	7.241139e+05	59883.208389	41.864782	4.737455	30.695860
std	7235.627217	7.186806e+06	43394.336755	24.613489	4.420472	7.346625
min	4.000000	0.000000e+00	0.000000	3.000000	0.000000	16.000000
25%	6237.000000	2.000000e+04	24060.000000	40.000000	2.000000	25.000000
50%	12571.000000	6.300000e+04	52704.000000	40.000000	4.000000	29.000000
75%	18787.500000	1.150000e+05	85574.500000	42.000000	5.000000	34.000000
max	25141.000000	3.900000e+08	209356.000000	1012.000000	99.000000	99.000000

In [18]:
```
new_df.median()
```

Out[18]:
```
Respondent       12571.0
CompTotal        63000.0
ConvertedComp    52704.0
WorkWeekHrs         40.0
CodeRevHrs           4.0
Age                 29.0
dtype: float64
```

Correlation

Finding correlation

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

In [30]:
```
new_df.corr()['Age']
```

Out[30]:
```
Respondent       0.002180
CompTotal        0.006337
ConvertedComp    0.401821
WorkWeekHrs      0.032032
CodeRevHrs      -0.012878
Age              1.000000
Name: Age, dtype: float64
```

Quiz:

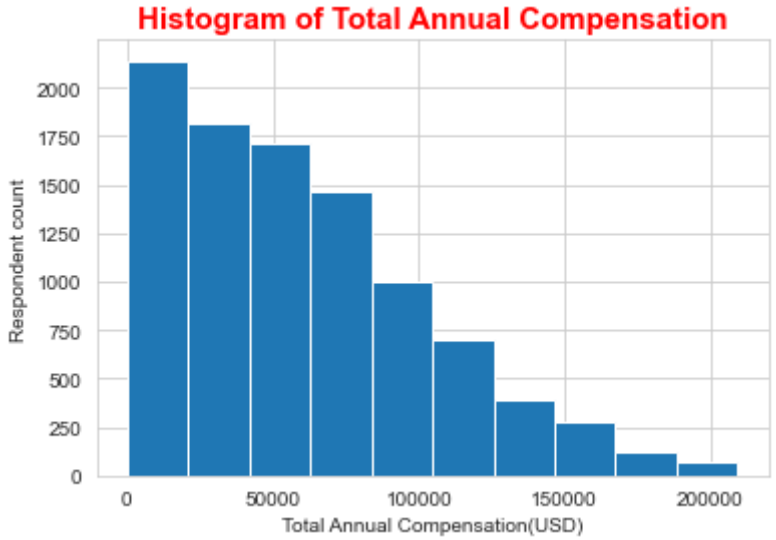
In [20]:
```
df["Age"].median()
```

Out[20]:
```
29.0
```

In [29]:
```
plt.hist(new_df['ConvertedComp'])
plt.ylabel('Respondent count')
plt.xlabel('Total Annual Compensation(USD)')

plt.title('Histogram of Total Annual Compensation',
          fontweight ="bold",color="RED",size=15)
```

Out[29]:
```
Text(0.5, 1.0, 'Histogram of Total Annual Compensation')
```



In [31]:
```
new_df.loc[df['Gender'] == 'Woman', 'ConvertedComp'].median()
```

Out[31]: 54956.0

In [32]:
```
new_df.loc[df['Gender'] == 'Man', 'ConvertedComp'].median()
```

Out[32]: 52339.0

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