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 [1]:
```
import pandas as pd
```

Load the dataset into a dataframe.

In [2]:
```
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/
```

Distribution

Determine how the data is distributed

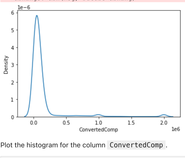
The column `ConvertedComp` contains Salary converted to annual USD salaries using the exchange rate on 2019-02-01.

This assumes 12 working months and 50 working weeks.

Plot the distribution curve for the column `ConvertedComp`.

In [16]:
```
# your code goes here
import seaborn as sns
ax = sns.distplot(df['ConvertedComp'], hist=False)
```

```
/opt/anaconda3/lib/python3.8/site-packages/seaborn/distributions.py:2557: FutureWarnin
g: `distplot` is a deprecated function and will be removed in a future version. Please
adapt your code to use either `displot` (a figure-level function with similar flexibil
ity) or `kdeplot` (an axes-level function for kernel density plots).
  warnings.warn(msg, FutureWarning)
```

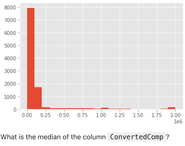


Plot the histogram for the column `ConvertedComp`.

In [28]:
```
# your code goes here
import matplotlib.pyplot as plt

x = df['ConvertedComp']

plt.style.use('ggplot')
plt.hist(x, bins=20)
plt.show()
```



What is the median of the column `ConvertedComp`?

In [30]:
```
# your code goes here
df['ConvertedComp'].median()
```

Out[20]:
```
57745.0
```

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

In [22]:
```
# your code goes here
df['Gender'].value_counts()
```

Out[22]:
```
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**?

In [23]:
```
# your code goes here
df[df['Gender'] == 'Woman']['ConvertedComp'].median()
```

Out[23]:
```
57708.0
```

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

Double click here for hint.

In [25]:
```
# your code goes here
df['Age'].describe()
```

Out[25]:
```
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
```

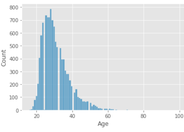
Plot a histogram of the column `Age`.

In [27]:
```
# your code goes here
x = df['Age']

plt.style.use('ggplot')
plt.hist(x, bins=10)
plt.show()
```

In [29]:
```
sns.histplot(df['Age'])
```

Out[29]:
```
<Axes:xlabel='Age', ylabel='Count'>
```

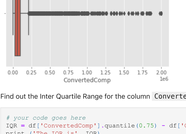


Outliers

Finding outliers

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

In [42]:
```
# your code goes here
ax = sns.boxplot(x=df['ConvertedComp'])
```



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

In [43]:
```
# your code goes here
IQR = df['ConvertedComp'].quantile(0.75) - df['ConvertedComp'].quantile(0.25)
print ('The IQR is', IQR)
```

```
The IQR is 73132.0
```

Find out the upper and lower bounds.

In [67]:
```
# your code goes here
Q1= df['ConvertedComp'].quantile(0.25)
Q3= df['ConvertedComp'].quantile(0.75)

Upper_Bound = Q3 + (IQR * 1.5)
Lower_Bound = Q1 - (IQR * 1.5)

print ('The upper bound is',Upper_Bound)
print ('The lower bound is',Lower_Bound)
```

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

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

In [71]:
```
# your code goes here
len(df[df['ConvertedComp'] > Upper_Bound])
```

Out[71]:
```
879
```

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

In [66]:
```
# your code goes here
df2 = df[df['ConvertedComp'] <= Upper_Bound]
df2['ConvertedComp'].describe()
```

Out[66]:
```
count       9703.000000
mean       59883.208389
std        43394.336755
min            0.000000
25%        24060.000000
50%        52704.000000
75%        85574.500000
max       209356.000000
Name: ConvertedComp, dtype: float64
```

Correlation

Finding correlation

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

In [76]:
```
# your code goes here
df.corr()['Age']
```

Out[76]:
```
Respondent        0.004041
CompTotal        -0.006970
ConvertedComp    -0.105386
WorkWeekHrs       0.036518
CodeRevHrs       -0.020469
Age               1.000000
Name: Age, dtype: float64
```

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