cognitiveclass.ai logo

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 [96]:
import pandas as pd
%matplotlib inline
import plotly.express as px
import plotly.figure_factory as ff
import numpy as np
import seaborn as sns
```

Load the dataset into a dataframe.

```
In [2]:
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/LargeData/m2_survey_data.csv")
```

Distribution

Determine how the data is distributed

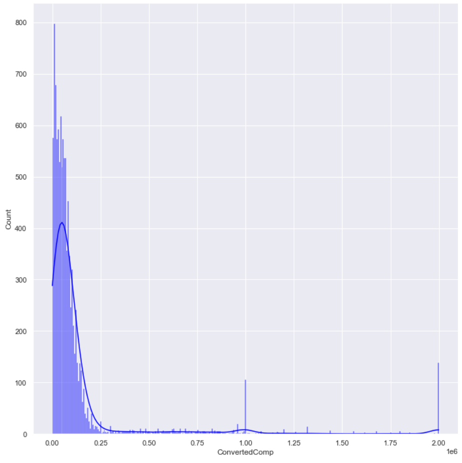
The column `ConvertedComp` contains Salary converted to annual USD salaries using the exchange rate on 2019-02-01.

This assumes 12 working months and 50 working weeks.

Plot the distribution curve for the column `ConvertedComp` .

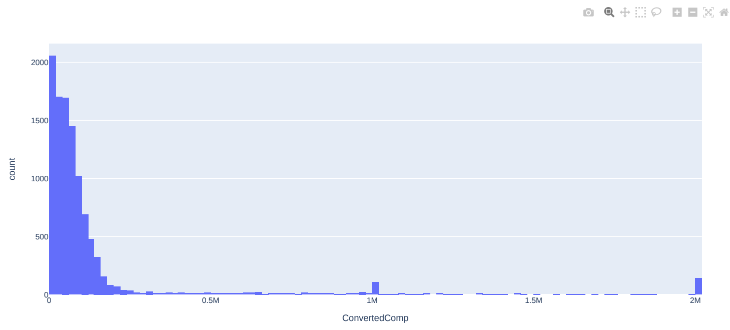
```
In [21]:
sns.set_theme()
sns.displot(df['ConvertedComp'], kde = True, color='blue', height = 10)
```

```
Out[21]: <seaborn.axisgrid.FacetGrid at 0x21b40fbabe0>
```



Plot the histogram for the column `ConvertedComp` .

```
In [4]:
# your code goes here
fig = px.histogram(df, x='ConvertedComp')
fig.show()
```



What is the median of the column `ConvertedComp` ?

```
In [5]:
# your code goes here
df.ConvertedComp.median()
```

```
Out[5]: 57745.0
```

```
In [6]:
df.Age.median()
```

```
Out[6]: 29.0
```

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

```
In [7]:
# your code goes here
df.Gender.value_counts()
```

```
Out[7]:
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 [8]:
# your code goes here
df.loc[df['Gender'] == 'Woman', ['ConvertedComp']].median()
```

```
Out[8]:
ConvertedComp    57708.0
dtype: float64
```

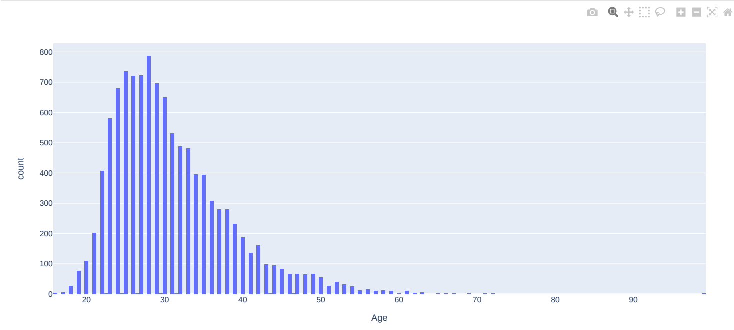
Give the five number summary for the column `Age` ?

```
In [9]:
# your code goes here
df['Age'].describe()
```

```
Out[9]:
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 [10]:
# your code goes here
histAge = px.histogram(df, x="Age")
histAge.show()
```

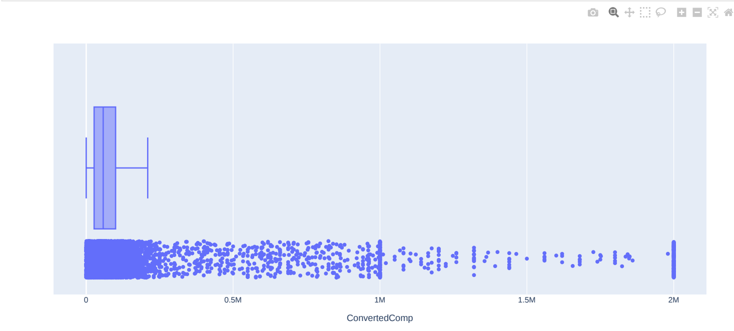


Outliers

Finding outliers

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

```
In [11]:
# your code goes here
#sns.boxplot(x=df['ConvertedComp'])
boxplot = px.box(df, x='ConvertedComp', points='all')
boxplot.show()
```



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

```
In [12]:
# your code goes here
df['ConvertedComp'].dropna(axis=0,inplace=True)
Q1_ConvertedComp = df['ConvertedComp'].quantile(0.25)
print("Primer cuartil: ", Q1_ConvertedComp)
Q3_ConvertedComp = df['ConvertedComp'].quantile(0.75)
print("Tercer cuartil: ", Q3_ConvertedComp)
IQR_ConvertedComp = Q3_ConvertedComp - Q1_ConvertedComp
print("Rango Intercuartil: ", IQR_ConvertedComp)
median_ConvertedComp= df['ConvertedComp'].median()
print('Mediana: ', median_ConvertedComp)
max_ConvertedComp = df["ConvertedComp"].max()
print("El valor maximo es: ", max_ConvertedComp)
min_ConvertedComp = df["ConvertedComp"].min()
print("el valor minimo es: ",min_ConvertedComp)
```

```
Primer cuartil:  26868.0
Tercer cuartil:  100000.0
Rango Intercuartil:  73132.0
Mediana:  57745.0
El valor maximo es:  2000000.0
el valor minimo es:  0.0
```

Find out the upper and lower bounds.

```
In [13]:
#limites superiores e inferiores de la caja.
q_inf=Q1_ConvertedComp-1.5*(IQR_ConvertedComp)
q_sup=Q3_ConvertedComp+1.5*(IQR_ConvertedComp)
```

```
print("limite inferior: ",q_inf)
print("limite superior: ",q_sup)
```

```
limite inferior:  -82830.0
limite superior:  209698.0
```

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

In [133…

```
#Conteo de valores atipicos.
identify_outliers_i = df["ConvertedComp"] < q_inf
identify_outliers_s = df["ConvertedComp"] > q_sup
identify_outliers = identify_outliers_i | identify_outliers_s
identify_outliers.value_counts()
```

Out[133…

```
False    10519
True       879
Name: ConvertedComp, dtype: int64
```

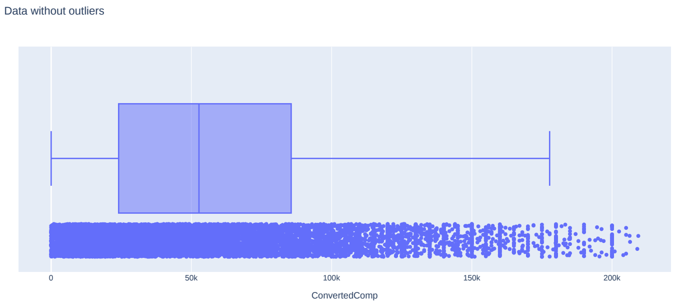
Create a new dataframe by removing the outliers from the `ConvertedComp` column.

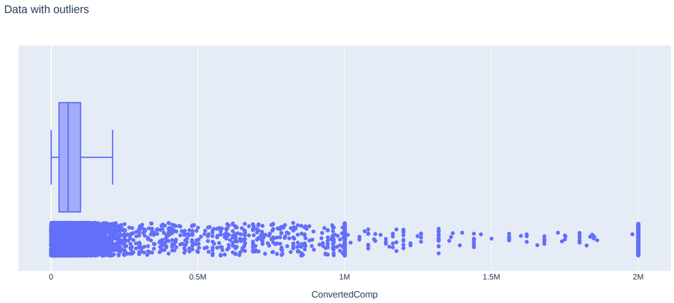
In [110…

```
df1 = (df[(df['ConvertedComp'] > q_inf) & (df['ConvertedComp'] < q_sup)])
```

In [111…

```
df1box = px.box(df1, x='ConvertedComp', points='all', title= 'Data without outliers')
dfbox = px.box(df, x='ConvertedComp', points='all', title='Data with outliers')
df1box.show()
dfbox.show()
```





Correlation

Finding correlation

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

In [143…

```
# your code goes
df.corr()['Age']
```

Out[143…

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

In [ ]:

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

Copyright © 2020 IBM Corporation. This notebook and its source code are released under the terms of the MIT License.