Oasis Infobytes - Internship

Task 2 - Unemployement Analysis with PythonUnemployment Analysis With Python

Unemployment is measured by the unemployment rate which is the number of people who are unemployed as a percentage of the total labour force. We have seen a sharp increase in the unemployment rate during Covid-19, so analyzing the unemployment rate can be a good data science project

In [1]:
```
#import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objects as go

import warnings
warnings.filterwarnings("ignore")
%matplotlib inline

import datetime as dt
import calendar
```

In [2]:
```
df = pd.read_csv('C:/Users/cws/Downloads/Unemployment in India.csv')
df.head()
```

Out[2]:

	Region	Date	Frequency	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)	Area
0	Andhra Pradesh	31-05-2019	Monthly	3.65	11999139.0	43.24	Rural
1	Andhra Pradesh	30-06-2019	Monthly	3.05	11755881.0	42.05	Rural
2	Andhra Pradesh	31-07-2019	Monthly	3.75	12086707.0	43.50	Rural
3	Andhra Pradesh	31-08-2019	Monthly	3.32	12285693.0	43.97	Rural
4	Andhra Pradesh	30-09-2019	Monthly	5.17	12256762.0	44.68	Rural

In [3]:
```
df.shape
```

Out[3]:
```
(768, 7)
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 7 columns):
 #   Column                                    Non-Null Count  Dtype  
---  ------                                    --------------  -----  
 0   Region                                    740 non-null    object 
 1    Date                                     740 non-null    object 
 2    Frequency                                740 non-null    object 
 3    Estimated Unemployment Rate (%)          740 non-null    float64
 4    Estimated Employed                       740 non-null    float64
 5    Estimated Labour Participation Rate (%)  740 non-null    float64
 6   Area                                      740 non-null    object 
dtypes: float64(3), object(4)
memory usage: 42.1+ KB
```

In [5]:
```
df.describe()
```

Out[5]:

	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)
count	740.000000	7.400000e+02	740.000000
mean	11.787946	7.204460e+06	42.630122
std	10.721298	8.087988e+06	8.111094
min	0.000000	4.942000e+04	13.330000
25%	4.657500	1.190404e+06	38.062500
50%	8.350000	4.744178e+06	41.160000
75%	15.887500	1.127549e+07	45.505000
max	76.740000	4.577751e+07	72.570000

In [6]:
```
#missing value in dataset
df.isna().sum()
```

Out[6]:
```
Region                                     28
 Date                                      28
 Frequency                                 28
 Estimated Unemployment Rate (%)           28
 Estimated Employed                        28
 Estimated Labour Participation Rate (%)   28
Area                                       28
dtype: int64
```

In [7]:
```
df.dropna(inplace=True)
```

In [8]:
```
#to Check
df.isna().sum()
```

Out[8]:
```
Region                                     0
 Date                                      0
 Frequency                                 0
 Estimated Unemployment Rate (%)           0
 Estimated Employed                        0
 Estimated Labour Participation Rate (%)   0
Area                                       0
dtype: int64
```

In [9]:
```
df.duplicated().any()
```

Out[9]:
```
False
```

In [10]:
```
#data shape after drop
df.shape
```

Out[10]:
```
(740, 7)
```

In [11]:
```
df.Region.value_counts()
```

Out[11]:
```
Andhra Pradesh      28
Kerala              28
West Bengal         28
Uttar Pradesh       28
Tripura             28
Telangana           28
Tamil Nadu          28
Rajasthan           28
Punjab              28
Odisha              28
Madhya Pradesh      28
Maharashtra         28
Karnataka           28
Jharkhand           28
Himachal Pradesh    28
Haryana             28
Gujarat             28
Delhi               28
Chhattisgarh        28
Bihar               28
Meghalaya           27
Uttarakhand         27
Assam               26
Puducherry          26
Goa                 24
Jammu & Kashmir     21
Sikkim              17
Chandigarh          12
Name: Region, dtype: int64
```

In [12]:
```
df.describe()
```

Out[12]:

	Estimated Unemployment Rate (%)	Estimated Employed	Estimated Labour Participation Rate (%)
count	740.000000	7.400000e+02	740.000000
mean	11.787946	7.204460e+06	42.630122
std	10.721298	8.087988e+06	8.111094
min	0.000000	4.942000e+04	13.330000
25%	4.657500	1.190404e+06	38.062500
50%	8.350000	4.744178e+06	41.160000
75%	15.887500	1.127549e+07	45.505000
max	76.740000	4.577751e+07	72.570000

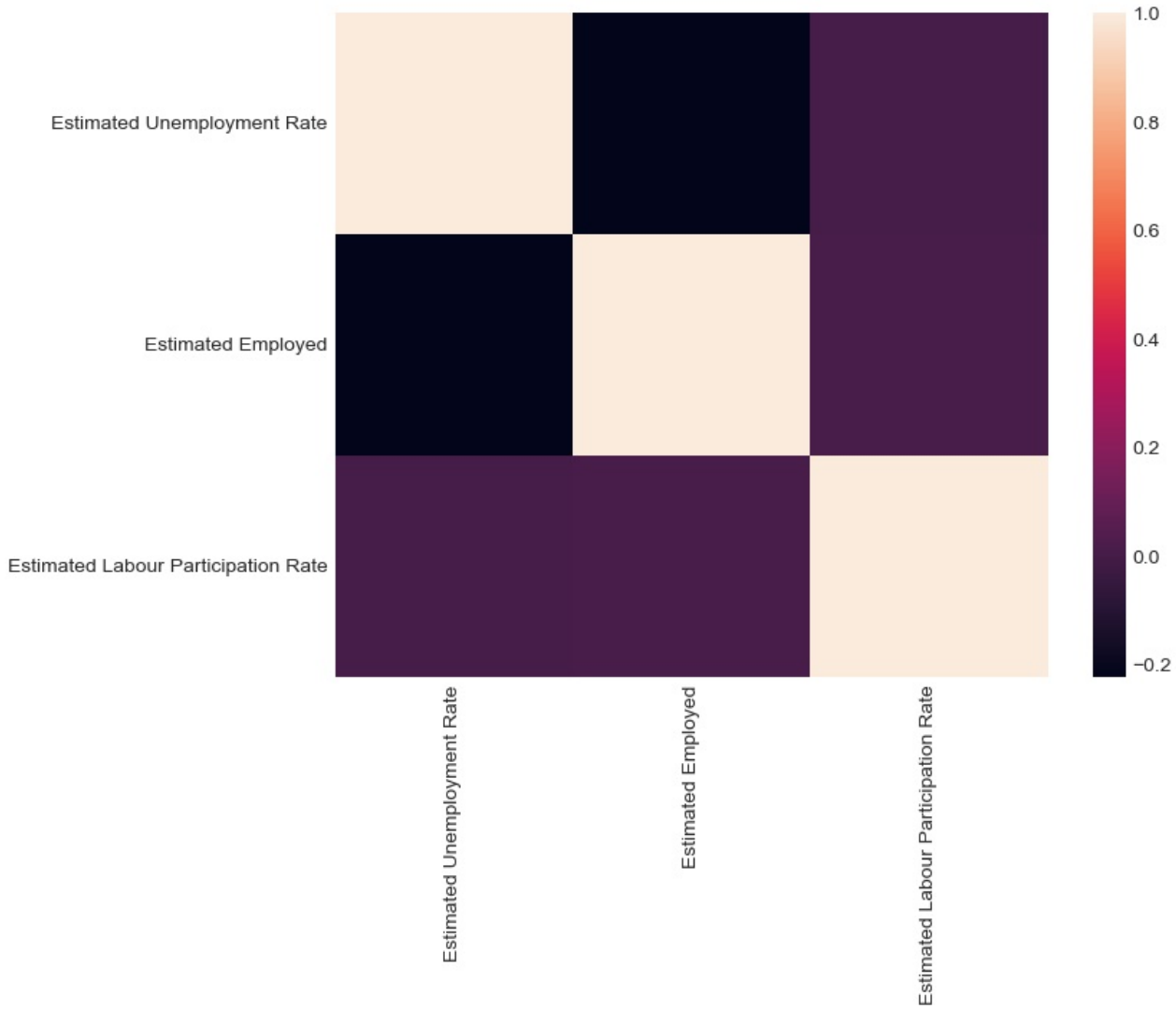
In [13]:
```
##Rename column
df.columns = ['State', 'Date', 'Frequency', 'Estimated Unemployment Rate', 'Estimated Employed',
              'Estimated Labour Participation Rate','Region']
df
```

Out[13]:

	State	Date	Frequency	Estimated Unemployment Rate	Estimated Employed	Estimated Labour Participation Rate	Region
0	Andhra Pradesh	31-05-2019	Monthly	3.65	11999139.0	43.24	Rural
1	Andhra Pradesh	30-06-2019	Monthly	3.05	11755881.0	42.05	Rural
2	Andhra Pradesh	31-07-2019	Monthly	3.75	12086707.0	43.50	Rural
3	Andhra Pradesh	31-08-2019	Monthly	3.32	12285693.0	43.97	Rural
4	Andhra Pradesh	30-09-2019	Monthly	5.17	12256762.0	44.68	Rural
...	...	...	...	...	...	...	...
749	West Bengal	29-02-2020	Monthly	7.55	10871168.0	44.09	Urban
750	West Bengal	31-03-2020	Monthly	6.67	10806105.0	43.34	Urban
751	West Bengal	30-04-2020	Monthly	15.63	9299466.0	41.20	Urban
752	West Bengal	31-05-2020	Monthly	15.22	9240903.0	40.67	Urban
753	West Bengal	30-06-2020	Monthly	9.86	9088931.0	37.57	Urban

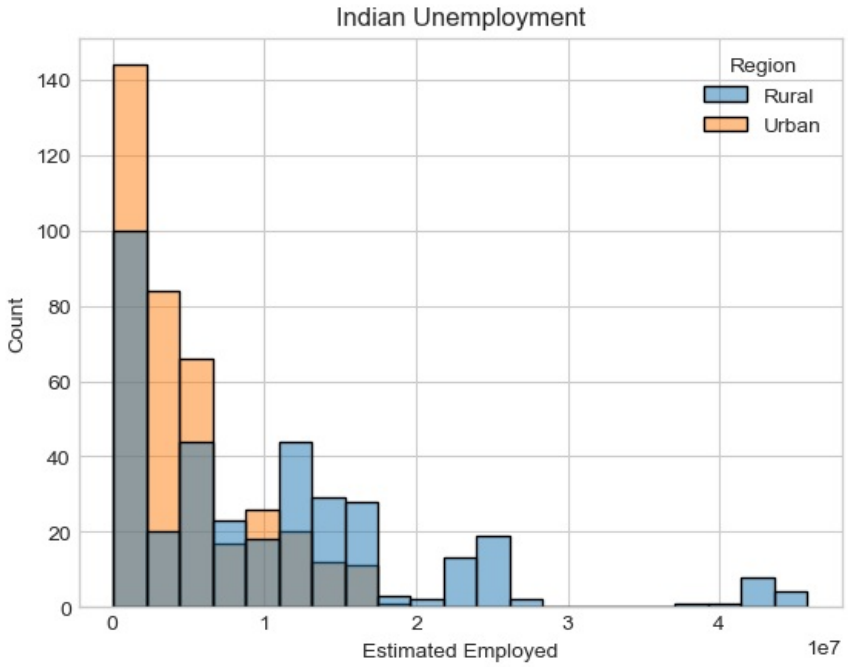
740 rows × 7 columns

In [14]:
```
#Correlation metrics
plt.style.use("seaborn-whitegrid")
plt.figure(figsize=(8,6))
sns.heatmap(df.corr())
plt.show()
```

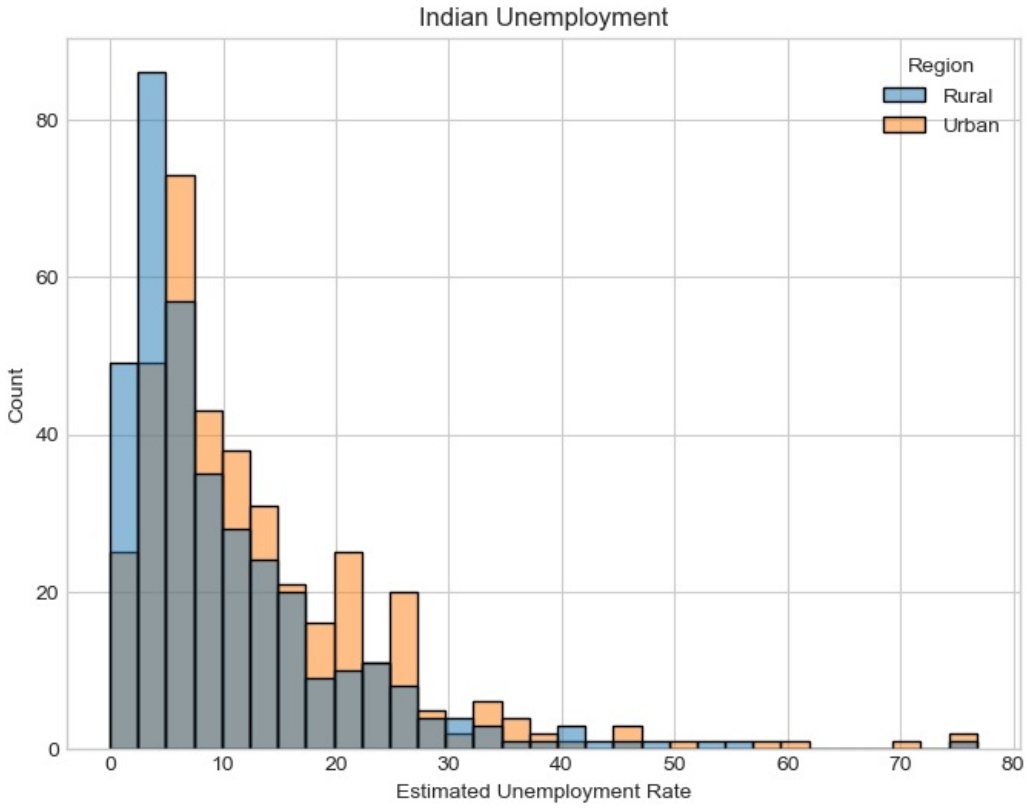


In [15]:
```
#Plot histogram
df.columns = ['State', 'Date', 'Frequency', 'Estimated Unemployment Rate', 'Estimated Employed',
              'Estimated Labour Participation Rate','Region']
plt.title("Indian Unemployment")
```

```
sns.histplot(x="Estimated Employed", hue="Region", data = df)
plt.show()
```



In [16]:
```
#Plot histogram for Estimated Unemployment Rate in different regions
plt.figure(figsize=(8,6))
plt.title("Indian Unemployment")
sns.histplot(x="Estimated Unemployment Rate", hue="Region", data=df)
plt.show()
```



In [19]:
```
import plotly.express as px
```

In [22]:
```
#Dashboard for Estimated Unemployment Rate in Region and State
unemployment = df[['State','Region','Estimated Unemployment Rate']]
figure = px.sunburst(unemployment,path=['Region','State'],
                     values='Estimated Unemployment Rate',
                     width=800, height=800, color_continuous_scale='RDY1Gn',
                     title='Unemployment Rate in India')
figure.show()
```

In [27]:
```
fig = px.box(data_frame=df,x='State',y='Estimated Unemployment Rate',color='State',title='Estimated Unemploymen
fig.update_layout(xaxis={'categoryorder':'total descending'})
fig.show()
```

In [28]:
```
fig = px.box(data_frame=df,x='State',y='Estimated Unemployment Rate',color='Region',title='Estimated Unemployme
fig.update_layout(xaxis={'categoryorder':'total descending'})
fig.show()
```

Thanks