Hands-on Activity 10.1 Data Analysis using Python

Intended Learning Outcome

1. Perform descriptive and correlation analysis to to analyze the dataset.
2. Interpret the results of descriptive and correlation analysis

Resources

- Personal Computer
- Jupyter Notebook
- Internet Connection

Instruction

1. Gather a dataset regarding your identified problem for the ASEAN Data Science Explorer. Make sure that the dataset includes multiple variables.
2. Load the dataset into pandas dataframe.
3. Prepare the data by applying appropriate data preprocessing techniques.
4. Analyze the data using descriptive analysis.
5. Perform correlation analysis.
6. Interpret the results based on the descriptive and correlation analysis.
7. Submit the PDF file.

```
import pandas as pd

df = pd.read_csv('/content/climate_risk_index_1.csv')
df.head()
```

	cartodb_id	the_geom	the_geom_webmercator	country	cri_rank	cri_score	fatalities_per_100k_rank	fatalities_per_1
0	1	NaN	NaN	Saudi Arabia	79	72.50	18	
1	2	NaN	NaN	Romania	61	61.50	112	
2	3	NaN	NaN	Spain	69	66.33	74	
3	4	NaN	NaN	Slovenia	135	124.50	114	
4	5	NaN	NaN	South Sudan	133	117.33	114	

```
df.dtypes
```

```
cartodb_id                   int64
the_geom                   float64
the_geom_webmercator       float64
country                     object
cri_rank                     int64
cri_score                  float64
fatalities_per_100k_rank     int64
fatalities_per_100k_total  float64
fatalities_rank              int64
fatalities_total             int64
losses_per_gdp_rank          int64
losses_per_gdp_total       float64
losses_usdm_ppp_rank         int64
losses_usdm_ppp_total      float64
rw_country_code             object
rw_country_name             object
dtype: object
```

```
#function for deleting columns
def delete_columns(columns):
  df.drop(columns= columns, inplace= True)
```

```
# Delete the column the_geom and the_geom_webmercator because both of these column are empty and delete also rw_conuntry_name because their is already a country column showing all what c
columns = ['the_geom', 'the_geom_webmercator', 'rw_country_name']
delete_columns(columns)
```

```
df
```

	cartodb_id	country	cri_rank	cri_score	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank
0	1	Saudi Arabia	79	72.50	18	0.45	18
1	2	Romania	61	61.50	112	0.01	102
2	3	Spain	69	66.33	74	0.05	47
3	4	Slovenia	135	124.50	114	0.00	114
4	5	South Sudan	133	117.33	114	0.00	114
...	...	...	...	...	...	...	...
177	178	Seychelles	135	124.50	114	0.00	114
178	179	Gambia	135	124.50	114	0.00	114
179	180	Togo	131	114.33	104	0.01	102
180	181	Trinidad and Tobago	135	124.50	114	0.00	114

```
# Check if there is a duplicated values
duplicated = df.duplicated()
print(duplicated.sum())
```

```
0
```

The output is zero that means no need to delete duplicated rows

```
# Get only the asean country in the dataset
asean_country = ['Brunei Darussalam', 'Burma', 'Myanmar', 'Cambodia', 'Indonesia', 'Laos', 'Malaysia', 'Philippines', 'Singapore', 'Thailand', 'Vietnam']

df= df[df['country'].isin(asean_country)]
```

```
df
```

	country	cri_rank	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank	fatalities_total	loss
14	Thailand	53	100	0.02	60	12	
17	Vietnam	29	56	0.10	25	91	
36	Singapore	135	114	0.00	114	0	
49	Indonesia	39	82	0.04	24	104	
62	Cambodia	48	58	0.09	55	14	
102	Malaysia	132	111	0.01	90	2	
133	Brunei Darussalam	135	114	0.00	114	0	
166	Myanmar	6	25	0.33	14	173	
167	Philippines	13	35	0.19	12	196	

```
# The only needed is the fatality and economy loss so delete the other columns that it is not important
columns = ['cartodb_id', 'cri_score', 'cri_rank']
delete_columns(columns)
```

```
df
```

	country	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank	fatalities_total	losses_per_gdp
14	Thailand	100	0.02	60	12	
17	Vietnam	56	0.10	25	91	
36	Singapore	114	0.00	114	0	
49	Indonesia	82	0.04	24	104	
62	Cambodia	58	0.09	55	14	
102	Malaysia	111	0.01	90	2	
133	Brunei Darussalam	114	0.00	114	0	
166	Myanmar	25	0.33	14	173	
167	Philippines	35	0.19	12	196	

```
df.describe()
```

	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank	fatalities_total	losses_per_gdp_rank	losses_per_gdp_total	losses_usdm_ppp_rank	losses_usdm_ppp_total
count	9.000000	9.000000	9.000000	9.000000	9.000000	6.000000	9.000000	9.000000
mean	77.222222	0.086667	56.444444	65.777778	71.222222	0.241267	59.333333	1233.432556
std	34.852467	0.110454	41.207133	78.138943	47.803707	0.125115	57.341085	1485.636204
min	25.000000	0.000000	12.000000	0.000000	22.000000	0.147000	4.000000	0.000000
25%	56.000000	0.010000	24.000000	2.000000	38.000000	0.155450	15.000000	0.276000
50%	82.000000	0.040000	55.000000	14.000000	49.000000	0.208850	26.000000	822.584000
75%	111.000000	0.100000	90.000000	104.000000	132.000000	0.252200	126.000000	1797.737000
max	114.000000	0.330000	114.000000	196.000000	135.000000	0.478600	135.000000	4186.230000

```
df.corr()
```

```
<ipython-input-51-2f6f6606aa2c>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid
  df.corr()
```

	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank	fatalities_total	losses_per_gdp_rank	losses_per_gdp_total	losses_usdm_ppp_rank	losses_u
fatalities_per_100k_rank	1.000000	-0.913195	0.877430	-0.857063	0.810929	-0.529561	0.693736	
fatalities_per_100k_total	-0.913195	1.000000	-0.730990	0.837173	-0.666971	0.829960	-0.545904	
fatalities_rank	0.877430	-0.730990	1.000000	-0.861837	0.904135	-0.306462	0.919048	
fatalities_total	-0.857063	0.837173	-0.861837	1.000000	-0.667229	0.459252	-0.702096	
losses_per_gdp_rank	0.810929	-0.666971	0.904135	-0.667229	1.000000	-0.978814	0.945387	
losses_per_gdp_total	-0.529561	0.829960	-0.306462	0.459252	-0.978814	1.000000	-0.208425	
losses_usdm_ppp_rank	0.693736	-0.545904	0.919048	-0.702096	0.945387	-0.208425	1.000000	
losses_usdm_ppp_total	-0.196839	0.109142	-0.590771	0.433959	-0.591999	-0.122460	-0.786059	

Data Analysis

The summary statistics presented provide insights into many parts of the dataset, and correlation coefficients clarify links between variables. The variables associated to deaths, such as fatalities per 100,000 people and total fatalities, vary significantly among nations, as seen by their standard deviations. Countries' ranks in terms of fatalities per 100,000 and overall fatalities are likewise diverse, indicating varying affects across areas. Furthermore, the correlation coefficients reveal interesting associations between these variables; for example, there is a strong positive correlation between fatalities per 100,000 and total fatalities, indicating that countries with higher fatality rates have higher overall fatality counts. Similarly, indicators associated to economic losses, such as losses per GDP and losses in USD (PPP), show substantial variation, with some nations suffering more severe economic consequences than others. The correlation coefficients shed more light on these linkages, revealing, for example, the substantial positive association between fatality rankings and economic loss rankings, meaning that nations with more deaths also have larger economic losses. Together, these data and correlations highlight the diverse character of the phenomena under investigation, underlining the significance of detailed analysis and focused responses to solve the complex difficulties that various nations confront.