Name:	Chaitya Arun Dobariya
UID:	2021600017
Course:	Advanced Data Visualization

Experiment 2

Aim:	Create advanced charts using Tableau / Power BI / R / Python / Plotly or Chart or D3.js to be performed on the dataset - Socio-economic data Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter, Line, Area, Waterfall, Donut, Treemap, Funnel Write observations from each chart Practice dataset: World Socio-Economic dataset and Power BI file

1. Dataset

You can find the dataset here.

Description

The study by Acemoglu and Johnson demonstrated the relationship between increased life expectancy and improvement in economic growth (GDP per capita), controlling for country-fixed effects. However, further analysis is necessary to determine how the allocation of a country's wealth through certain investments in healthcare, education, environmental management, and some socioeconomic factors have an overall effect in determining average life expectancy.

Main sources of data - World Bank Open Data & Our World in Data

Fields

Country Name: 174 countries - list (String)
Country Code: 3-letter code (String)
Region: region of the world country is located in (String)
IncomeGroup: country's income class (String)

Year: 2000-2019 (both included) (IntegeUnemployment refers to the % sharer)
Life Expectancy World Bank: data (Float)
Prevelance of Undernourishment: Prevalence of undernourishment is the percentage of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are required to maintain a normally active and healthy life (Float)
CO2: Carbon dioxide emissions (Integer)
Health Expenditure %: Level of current health expenditure (Float)
Education Expenditure %: General government expenditure on education (Float)
Unemployment: Unemployment refers to the % share (Float)
Corruption: the extent to which the executive can be held accountable for its use of funds and for the results of its actions by the electorate and by the legislature and judiciary, and the extent to which public employees within the executive are required to account for administrative decisions, use of resources, and results obtained. (Float)
Sanitation: The percentage of people using improved sanitation facilities (Float)
Injuries: (Float)
Communicable: (Float)
NonCommunicable: (Float)

Importing the libraries and data:

```
!pip install squarify
```

```
Collecting squarify
  Downloading squarify-0.4.4-py3-none-any.whl.metadata (600 bytes)
Downloading squarify-0.4.4-py3-none-any.whl (4.1 kB)
Installing collected packages: squarify
Successfully installed squarify-0.4.4
```

```
[ ] import pandas as pd
    import seaborn as sns
    import squarify
    import matplotlib.pyplot as plt
    import matplotlib.ticker as ticker
```

```
[ ] df = pd.read_csv('data.csv')
    # df = df.dropna()
```

```
df.head()
```

	Country Name	Country Code	Region	IncomeGroup	Year	Life Expectancy World Bank	Prevelance of Undernourishment	CO2	Health Expenditure %	Education Expenditure %	Unemployment	Corruption	Sanitation		Injuries	Communicable	NonCommunicable
0	Afghanistan	AFG	South Asia	Low income	2001	56.308	47.8	730.0	NaN	NaN	10.809000	NaN	NaN	2	2179727.10	9689193.70	5795426.38
1	Angola	AGO	Sub-Saharan Africa	Lower middle income	2001	47.059	67.5	15960.0	4.483516	NaN	4.004000	NaN	NaN	1	1392080.71	11190210.53	2663516.34
2	Albania	ALB	Europe & Central Asia	Upper middle income	2001	74.288	4.9	3230.0	7.139524	3.4587	18.575001	NaN	40.520895		117081.67	140894.78	532324.75
3	Andorra	AND	Europe & Central Asia	High income	2001	NaN	NaN	520.0	5.865939	NaN	NaN	NaN	21.788660		1697.99	695.56	13636.64
4	United Arab Emirates	ARE	Middle East & North Africa	High income	2001	74.544	2.8	97200.0	2.484370	NaN	2.493000	NaN	NaN		144678.14	65271.91	481740.70

data.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Country N	Country C	Region	IncomeGr	Year	Life Expec	Prevelanc	CO2	Health Ex	Education	Unemploy	Corruptio	Sanitation	Injuries	Communi	NonCommunicable	
2	Afghanist	AFG	South Asi	Low incon	2001	56.308	47.8	730			10.809			2179727	9689194	5795426	
3	Angola	AGO	Sub-Saha	Lower mic	2001	47.059	67.5	15960	4.483516		4.004			1392081	11190211	2663516	
4	Albania	ALB	Europe & (	Upper mic	2001	74.288	4.9	3230	7.139524	3.4587	18.575		40.5209	117081.7	140894.8	532324.8	
5	Andorra	AND	Europe & (	High incor	2001			520	5.865939				21.78866	1697.99	695.56	13636.64	
6	United Ara	ARE	Middle Ea	High incor	2001	74.544	2.8	97200	2.48437		2.493			144678.1	65271.91	481740.7	
7	Argentina	ARG	Latin Ame	Upper mic	2001	73.755	3	125260	8.371798	4.83374	17.32		48.054	1397676	1507069	8070910	
8	Armenia	ARM	Europe & (	Upper mic	2001	71.8	26.1	3600	4.645627	2.46944	10.912		46.3519	103371.8	122238.1	767916.2	
9	American	ASM	East Asia	Upper mic	2001									1683.98	2933.98	10752.13	
10	Antigua a	ATG	Latin Ame	High incor	2001	74.171		350	5.435876					2201.12	3279.72	14289.69	

2. Data Preprocessing

```
df_india = df[df['Country Name'] == 'India']
# df_india = df_india.dropna()
df_india.head()
```

	Country Name	Country Code	Region	IncomeGroup	Year	Life Expectancy World Bank	Prevelance of Undernourishment	CO2	Health Expenditure %	Education Expenditure %	Unemployment	Corruption	Sanitation	Inj
72	India	IND	South Asia	Lower middle income	2001	62.907	18.4	9.535400e+05	4.262781	NaN	5.576	NaN	7.847825	54362
246	India	IND	South Asia	Lower middle income	2002	63.304	20.1	9.854500e+05	4.240167	NaN	5.530	NaN	9.168575	52183
420	India	IND	South Asia	Lower middle income	2003	63.699	21.5	1.011770e+06	4.008480	3.61341	5.643	NaN	11.045160	50948
594	India	IND	South Asia	Lower middle income	2004	64.095	22.1	1.085670e+06	3.957392	3.35254	5.629	NaN	13.377481	52419
768	India	IND	South Asia	Lower middle income	2005	64.500	21.6	1.136470e+06	3.791162	3.18875	5.613	3.5	15.321118	53131

Selecting the ‘India’ Region

```
df_brazil = df[df['Country Name'] == 'Brazil']
df_brazil.head()
```

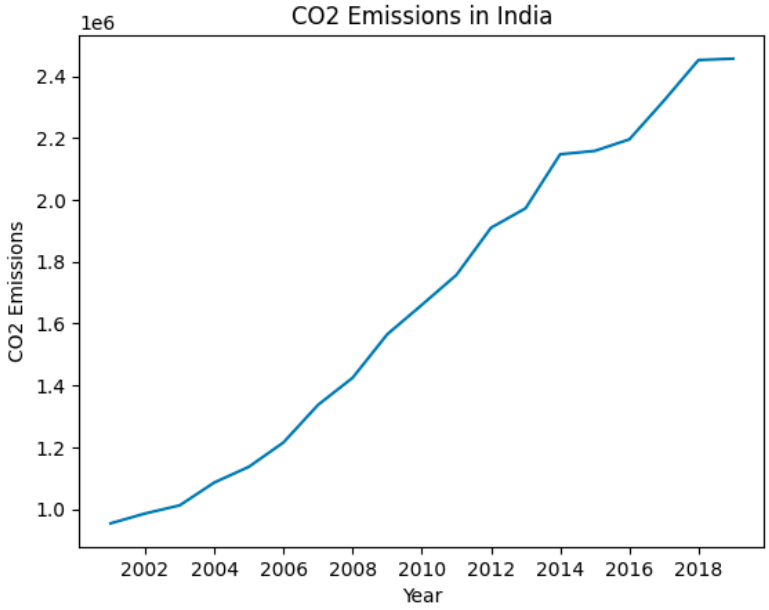
	Country Name	Country Code	Region	IncomeGroup	Year	Life Expectancy World Bank	Prevelance of Undernourishment	CO2	Health Expenditure %	Education Expenditure %	Unemployment	Corruption	Sanitation	
24	Brazil	BRA	Latin America & Caribbean	Upper middle income	2001	70.462	10.7	319380.000000	8.549606	3.84468	9.61	NaN	36.167624	9
198	Brazil	BRA	Latin America & Caribbean	Upper middle income	2002	70.813	9.3	317760.009766	8.696857	3.75037	9.37	NaN	36.585898	9
372	Brazil	BRA	Latin America & Caribbean	Upper middle income	2003	71.170	7.9	310809.997559	8.188999	NaN	9.99	NaN	37.002145	9
546	Brazil	BRA	Latin America & Caribbean	Upper middle income	2004	71.531	6.8	328519.989014	8.124920	3.97448	9.11	NaN	37.416236	9
720	Brazil	BRA	Latin America & Caribbean	Upper middle income	2005	71.896	6.5	331690.002441	8.035410	4.47908	9.57	NaN	37.828215	9

Selecting the ‘Brazil’ Region

3. Charts & Plots

3.1 Line Chart:

```
# Line Chart
sns.lineplot(data=df_india, x='Year', y='CO2')
plt.title('CO2 Emissions in India')
plt.xlabel('Year')
plt.ylabel('CO2 Emissions')
plt.gca().xaxis.set_major_locator(ticker.MaxNLocator(integer=True))
plt.show()
```

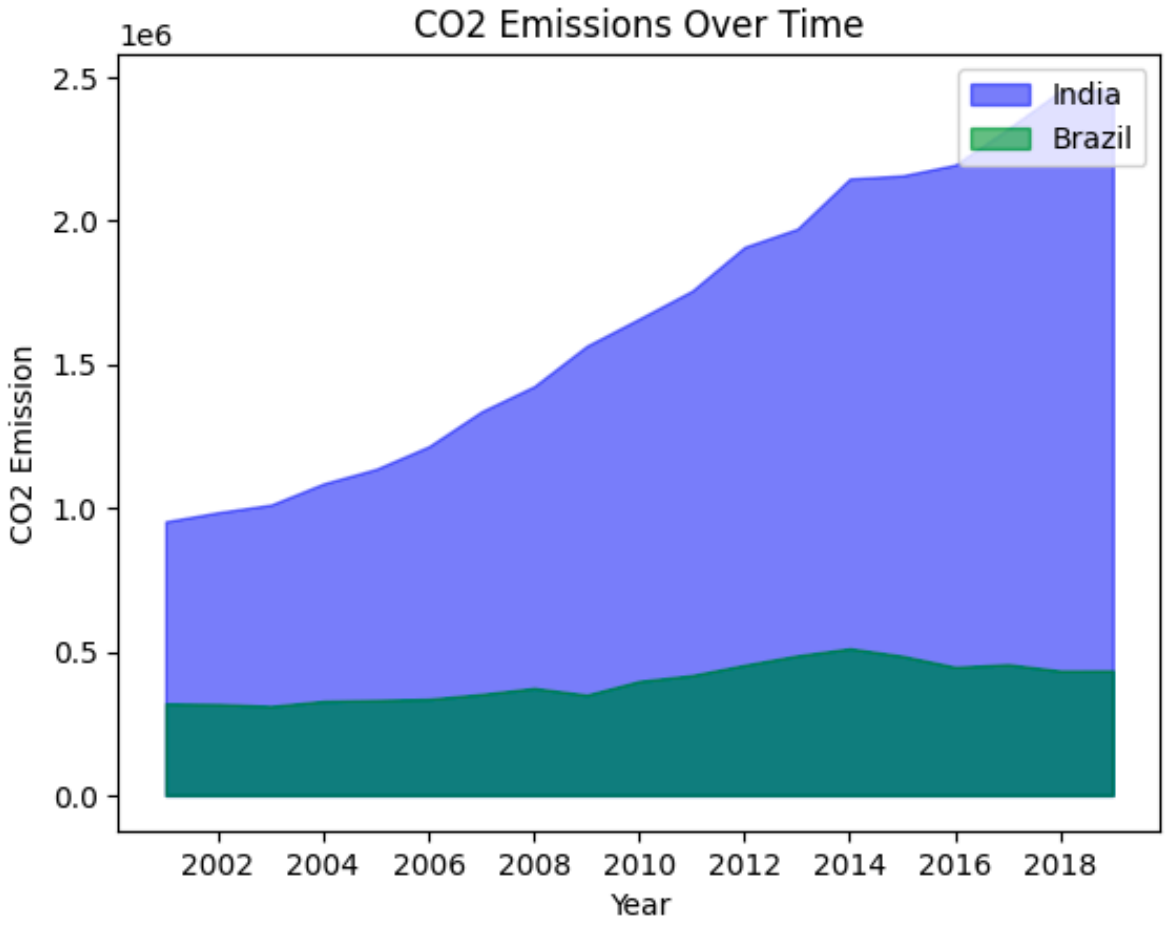


Observation:

The line chart shows the increase in CO2 emissions in India from 2001 to 2020.

3.2 Area Chart

```
# Area Chart
# plt.stackplot(df_india['Year'], df_india['Life Expectancy World Bank'], df_india['CO2'], labels=['Life Expectancy World Bank', 'CO2'])


plt.fill_between(df_india['Year'], df_india['CO2'], color='blue', alpha=0.5, label='India')
plt.fill_between(df_brazil['Year'], df_brazil['CO2'], color='green', alpha=0.5, label='Brazil')
plt.title('CO2 Emissions Over Time')
plt.xlabel('Year')
plt.ylabel('CO2 Emission')
plt.gca().xaxis.set_major_locator(ticker.MaxNLocator(integer=True))
plt.legend()
plt.show()
```

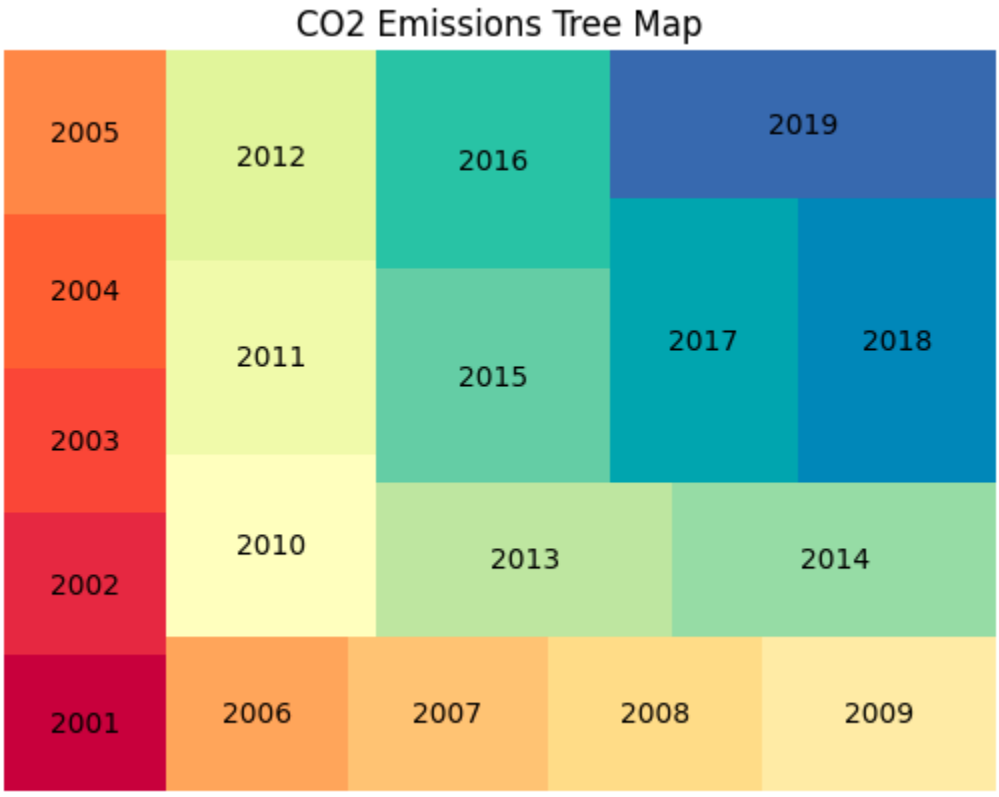


Observation:

This area chart shows the increase in CO2 emissions comparing India and Brazil from 2001 to 2020.

3.3 Treemap

```
# Tree Map
squarify.plot(sizes=df_india['CO2'], label=df_india['Year'], color=sns.color_palette("Spectral", len(df_india['CO2'])))
plt.title('CO2 Emissions Tree Map')
plt.axis('off')
plt.show()
```

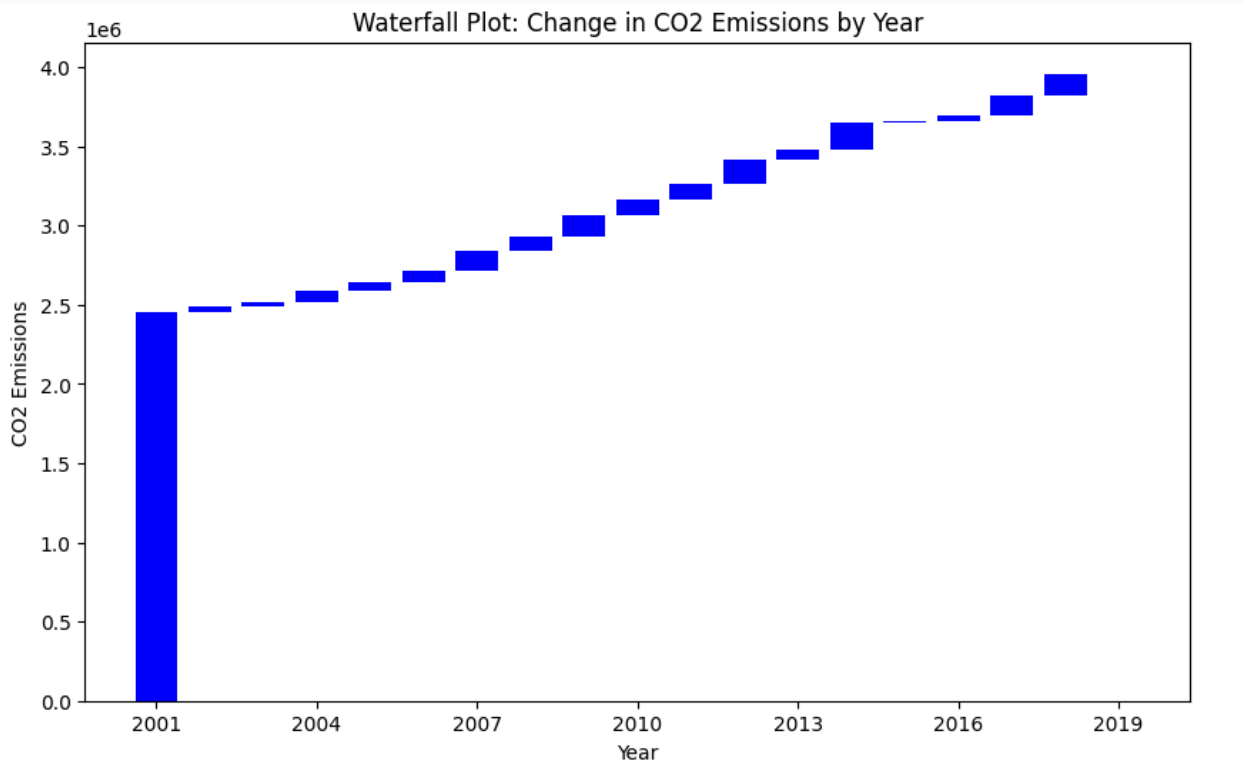


Observation:

This treemap shows the yearly CO2 emissions.

3.4 Waterfall Chart

```
# Waterfall Chart

df_grouped = df_india.groupby('Year')['CO2'].sum().reset_index()

df_grouped['CO2_Diff'] = df_grouped['CO2'].diff().fillna(df_grouped['CO2'].max())

fig, ax = plt.subplots(figsize=(10, 6))

cumulative = 0
for i, (income_group, co2_diff) in enumerate(zip(df_grouped['Year'], df_grouped['CO2_Diff'])):
    ax.bar(income_group, co2_diff, bottom=cumulative, color='blue' if co2_diff >= 0 else 'red')
    cumulative += co2_diff

plt.title('Waterfall Plot: Change in CO2 Emissions by Year')
plt.ylabel('CO2 Emissions')
plt.xlabel('Year')
plt.show()
```

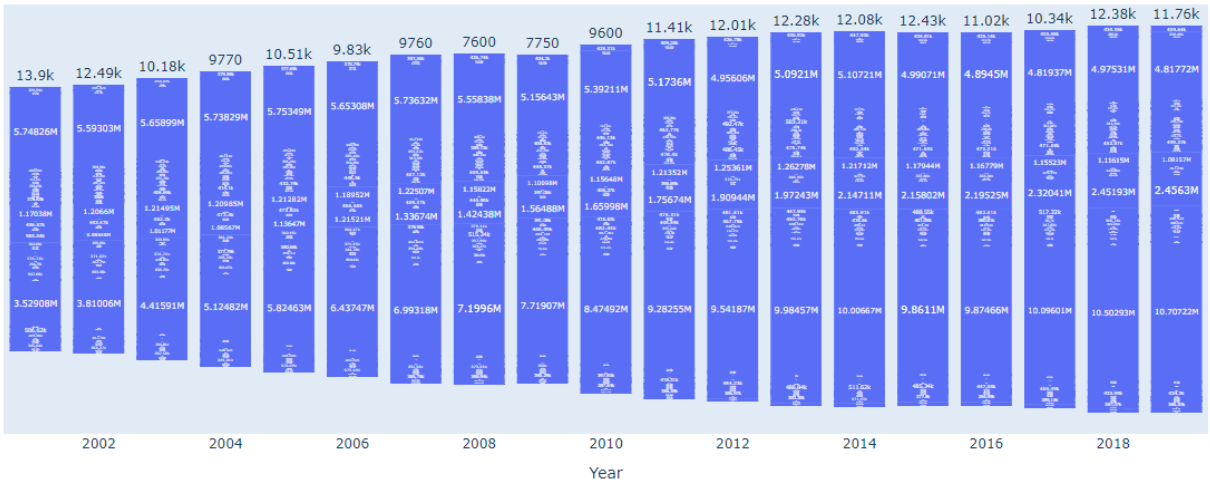


Observation:

This waterfall chart shows the increase in CO2 emissions year on year with a maximum increase in 2001 for India.

3.5 Funnel Chart

```
# Funnel Chart

import plotly.express as px

fig = px.funnel(df, x='Year', y='CO2')
fig.show()
```

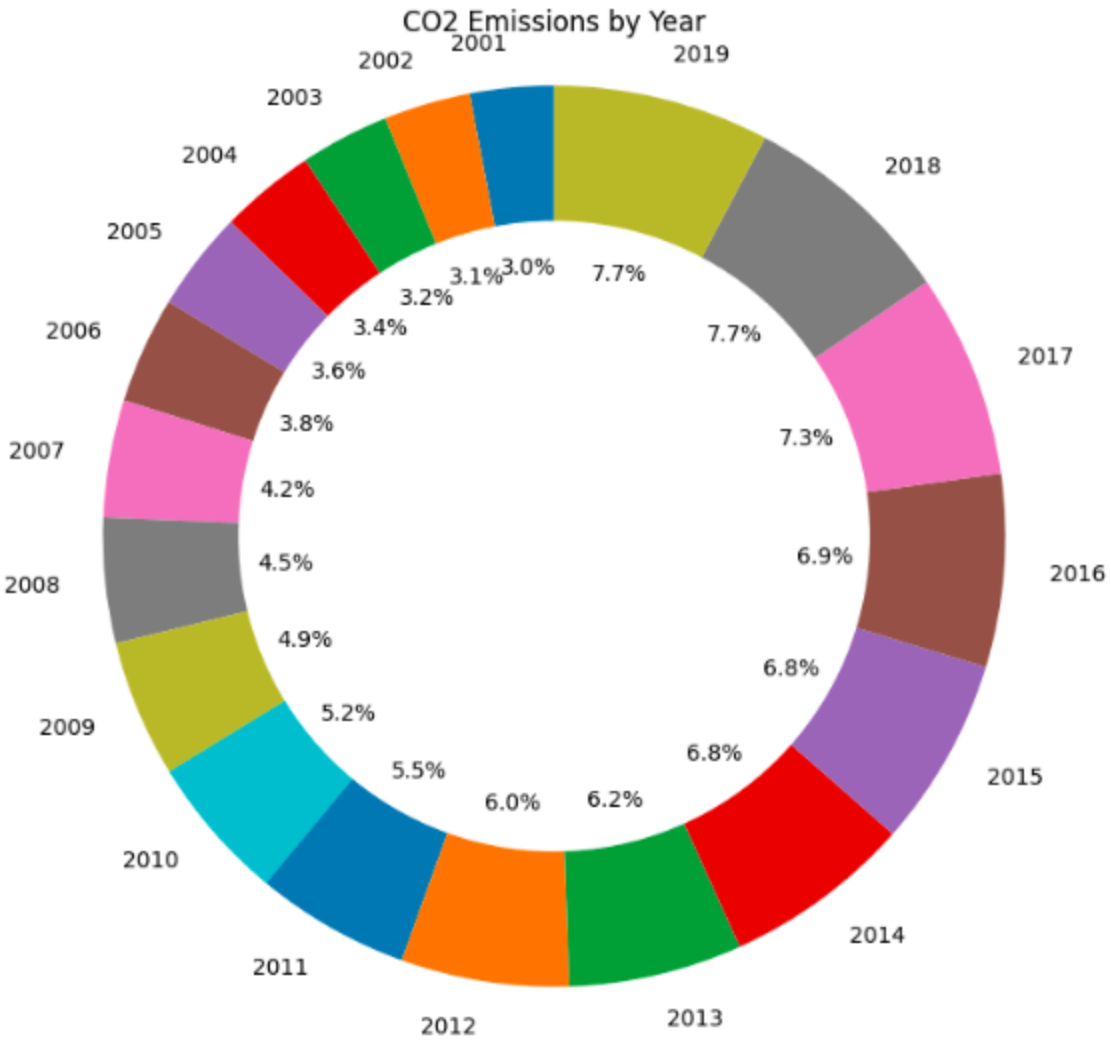


Observation:

This funnel chart shows the increase in CO2 emissions year on year.

3.6 Donut Chart

```
# Donut Chart

fig, ax = plt.subplots(figsize=(8, 8))

ax.pie(df_grouped['CO2'], labels=df_grouped['Year'], autopct='%1.1f%%', startangle=90)

centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig.gca().add_artist(centre_circle)

ax.axis('equal')
plt.title('CO2 Emissions by Year')

plt.show()
```

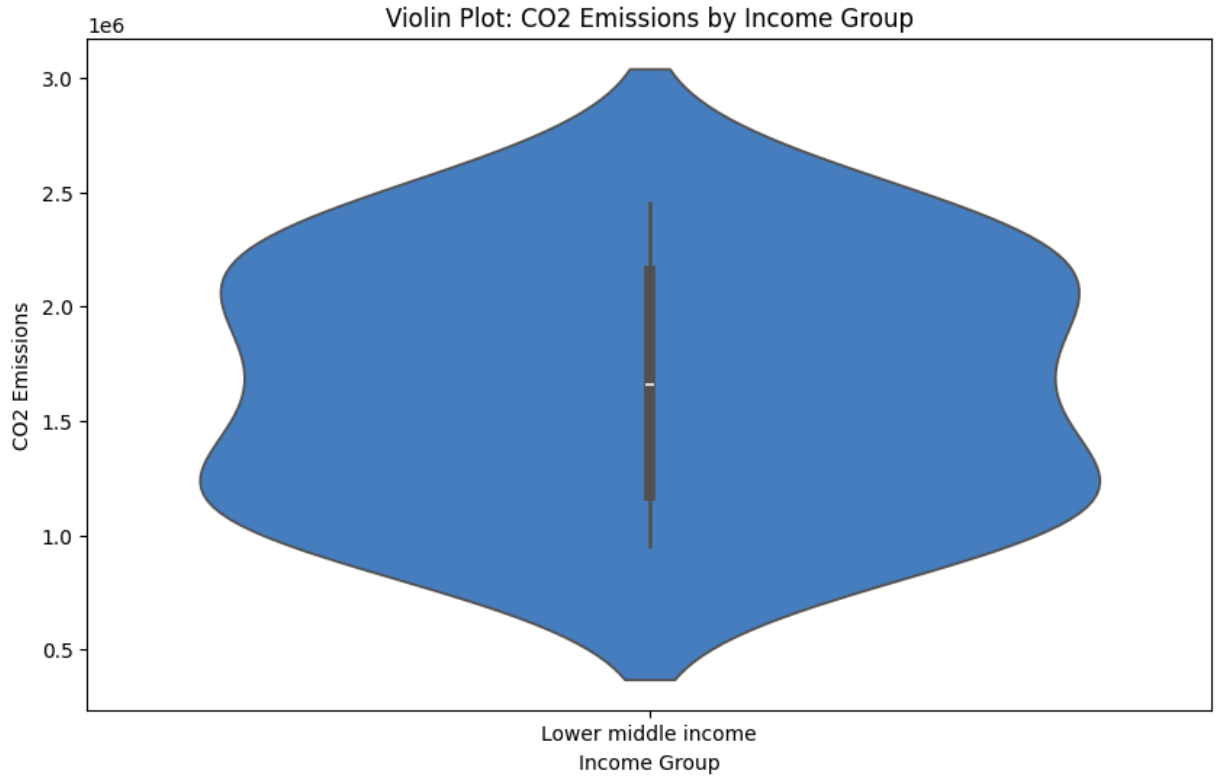


Observation:

This donut chart shows the distribution of CO2 emissions on yearly basis.

3.7 Violin Plot

```
# Violin Plot

plt.figure(figsize=(10, 6))
sns.violinplot(x='IncomeGroup', y='CO2', data=df_india, palette="muted")

plt.title('Violin Plot: CO2 Emissions by Income Group')
plt.xlabel('Income Group')
plt.ylabel('CO2 Emissions')

plt.show()
```



Conclusion

I worked on the analysis of the socio-economic dataset through various advanced data visualization techniques, which has provided valuable insights into the trends and distributions of key indicators like CO2 emissions, life expectancy, and socio-economic factors across different countries and regions, with a specific focus on India and Brazil. Got to know about various plots like Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter, Line, Area, Waterfall, Donut, Treemap and Funnel.