In [1]:
```
# libraries
import pandas as pd
import seaborn as sns
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.graph_objects as go
import altair as alt
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

PSTAT 100 Project plan

Group information

Group members: Jen Rink, Roshan Mehta, Kasturi Sharma, and Jake Jensema.

Contributions:

1. Jen Rink worked on tidying the data, contributed to the data description, proposed exploration approaches, and compiled group work.
2. Roshan Mehta worked on tidying the data and contributed to the data description; discussing the data collection, structure, sampling design and also filling in data for the year 2020 that our dataset was missing.
3. Kasturi Sharma worked on the background portion, introducing the topic of our project and generally describing the data and presenting our goals and questions.
4. Jake Jensema worked on the initial explorations; producing variable summaries, correlation heatmaps, and explored the missing data and how to deal with it.

---

0. Background

We chose to do our project on World Development Indicators and how they compare to one another. The World Development Indicators are the primary collection of development indicators in the World Bank. It complies data from around the globe and has the most current and accurate information.

We chose a few indicators from a collection of many to see how they correlate with one another. The indicators we chose were government expenditures on education, CO2 emissions, forest area, and GDP per capita. The government expenditures on education indicator is expressed as a percentage of the GDP. It includes local, regional, and central governments. The CO2 emissions indicator showcases the amount of CO2 in kilotons produced in the area. The forest area indicator showcases the how many square kilometers of forest area are found within the specific country, excluding agricultural areas, urban parks, and gardens. Lastly, the indicator GDP per capita is the gross domestic produce per person. It measures the economic output of a country per person.

The motivation for collecting data such as this is to find out what sort of effect, if any, does government expenditure have on these various factors. As college students, we have been in school for the majority of our lives. Many of us will start working or already do work and pay taxes. It's important that we are aware of how much of our money is spent on our educations. On top of that, it is important we realize how these expenses affect the other aspects of our lives. Also, the topic of the cost of education and student loans have been widely discussed within the United States and internationally. With this data, can we say if education can have an impact on things such as CO2 emissions or GDP? In turn, can we say that there is any correlation between CO2 emissions and forest area? What would be some of the possible explanations as to why we see these trends?

---

1. Data description

Basic information

General description: The data consists of different variables (Population, GDP, GDP Expenditures for education, CO2 emissions, and forest area) from 2012 to 2020, collected for 217 countries/territories around the world.

Source: The data used in this course project comes from 'The World Bank,' specifically the World Development Indicators, which can be found here (https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators). The WDI is comprised of numerous different topics which were collected by different entities such as the UNESCO Institute for Statistics, or even the countires themselves, and for various reasons, but are all compiled here and give the most accurate picture of global estimates.

Collection methods: The data was "compiled from officially-recognized international sources" like Climate Watch, and UNESCO. They use different collection methods based on the variable being collected. The environmental related variables were found using scientific equipment, while others used census surveys, or transactional tracking.

Sampling design and scope of inference: The statistical population includes all countries and territories around the globe existing between 2012 and 2020. The sampling frame is all countries reporting economic output and environmental data from 2012 to 2020. The sample is 217 countries and is equal to the frame. So, the frame completely overlaps the population. The sampling mechanism: the sample is a census of the frame. This is also population census data and there is no scope of inference.

Data semantics and structure

Units and observations: The Countries are our observational units.
The variables: Government expenditure was measured by taking the % spent of total GDP the country had, CO2 emissions by kilotons of emissions produced by each country, Forest area by the square kilometers of forested area each country had, and GDP by the gross domestic product per capita/person.

Variable descriptions:

Name	Variable description	Type	Units of measurement
Country Name	Name of each Country	Categorical	N/A
Year	Years range from 2012-2020	Categorical	Year
Government expenditure on education, total (% of GDP)	The total each country spends on education	Numeric	% of GDP
CO2 emissions (kt)	The total CO2 emissions each country exerts	Numeric	Kilotons
Forest area (sq. km)	The total forested area a country has	Numeric	Square Kilometers
GDP per capita	The gross domestic product of a country	Numeric	Per Capita
Population	The total estimated population of a country	Numeric	Individuals

Example rows:

In [2]:
```
# load tidied data and print rows
data = pd.read_csv('data-classproject/WDI_data_full.csv', encoding = 'latin1')
df = data.rename(columns = {'Country Name': 'Country',
                            'Government expenditure on education, total (% of GDP)': 'GEOC',
                            'CO2 emissions (kt)': 'CO2_emissions',
                            'Forest area (sq. km)': 'Forest_area'})
df.head()
```

Out[2]:

	Country	Year	GEOC	CO2_emissions	Forest_area	GDP per capita	Population
0	Afghanistan	2012	3.32	8080.00	12084.40	663.14	30466479
1	Albania	2012	2.93	4360.00	7849.17	4247.63	2900401
2	Algeria	2012	7.64	134929.99	19332.00	5610.73	37260563
3	American Samoa	2012	NaN	NaN	173.70	11920.06	53691
4	Andorra	2012	NaN	490.00	160.00	44904.58	71013

---

2. Initial explorations

We can see some NaN values already just from the first 5 rows above so we should look at how many values are gonna be missing in our dataset.

In [3]:
```
((df.isna().sum()) / (df.count()) * 100)
```

Out[3]:
```
Country           0.000000
Year              0.000000
GEOC             25.112108
CO2_emissions    27.814136
Forest_area       1.877934
GDP per capita    3.827751
Population        0.000000
dtype: float64
```

We can see that for 2 variables there is high percentage of missing data which we will need to deal with.
First we can try and look to see if there is a pattern within our missing data or if it is missing completely at random.

First for **CO2 Emissions**:

In [4]:
```
df.loc[df['CO2_emissions'].isnull()].drop(columns = ['Year', 'CO2_emissions']).mean()
```

Out[4]:
```
GEOC              4.663414e+00
Forest_area       1.044448e+05
GDP per capita    3.083606e+04
Population        1.871200e+07
dtype: float64
```

In [5]:
```
pd.set_option('display.float_format', str)
df.loc[df['CO2_emissions'].notnull()].drop(columns = ['Year', 'CO2_emissions']).mean()
```

Out[5]:
```
GEOC              4.548666158536585
Forest_area       213120.3979646597
GDP per capita   14522.145223184541
Population        38759267.01374345
dtype: float64
```

Second for **GEOC (Government expenditure on education)**:

In [6]:
```
df.loc[df['GEOC'].isnull()].drop(columns = ['Year', 'GEOC']).mean()
```

Out[6]:
```
CO2_emissions     56421.94500000001
Forest_area       65792.28973262032
GDP per capita   22135.275975975976
Population       16088708.339285715
dtype: float64
```

In [7]:
```
df.loc[df['GEOC'].notnull()].drop(columns = ['Year', 'GEOC']).mean()
```

Out[7]:
```
CO2_emissions    191434.74846036587
Forest_area       221432.7114257939
GDP per capita     16889.1455878553
Population        38994226.94106342
dtype: float64
```

When looking at the comparison between the NaN values and when they are not NaN we do not see a clear trend in any direction. Some values are higher when a variable is null while in other times the value is lower when a variable is NaN. This can lead us to believe that the values are MAR (Missing at random). Meaning dropping the NaN values will not bias the data heavily.

In [8]:
```
df = df.dropna()
df.head()
```

Out[8]:

	Country	Year	GEOC	CO2_emissions	Forest_area	GDP per capita	Population
0	Afghanistan	2012	3.32	8080.0	12084.4	663.14	30466479
1	Albania	2012	2.93	4360.0	7849.17	4247.63	2900401
2	Algeria	2012	7.64	134929.99	19332.0	5610.73	37260563
5	Angola	2012	3.08	23870.0	710478.76	4962.55	25188292
6	Antigua and Barbuda	2012	2.49	700.0	86.48	13686.48	87674

Data summary

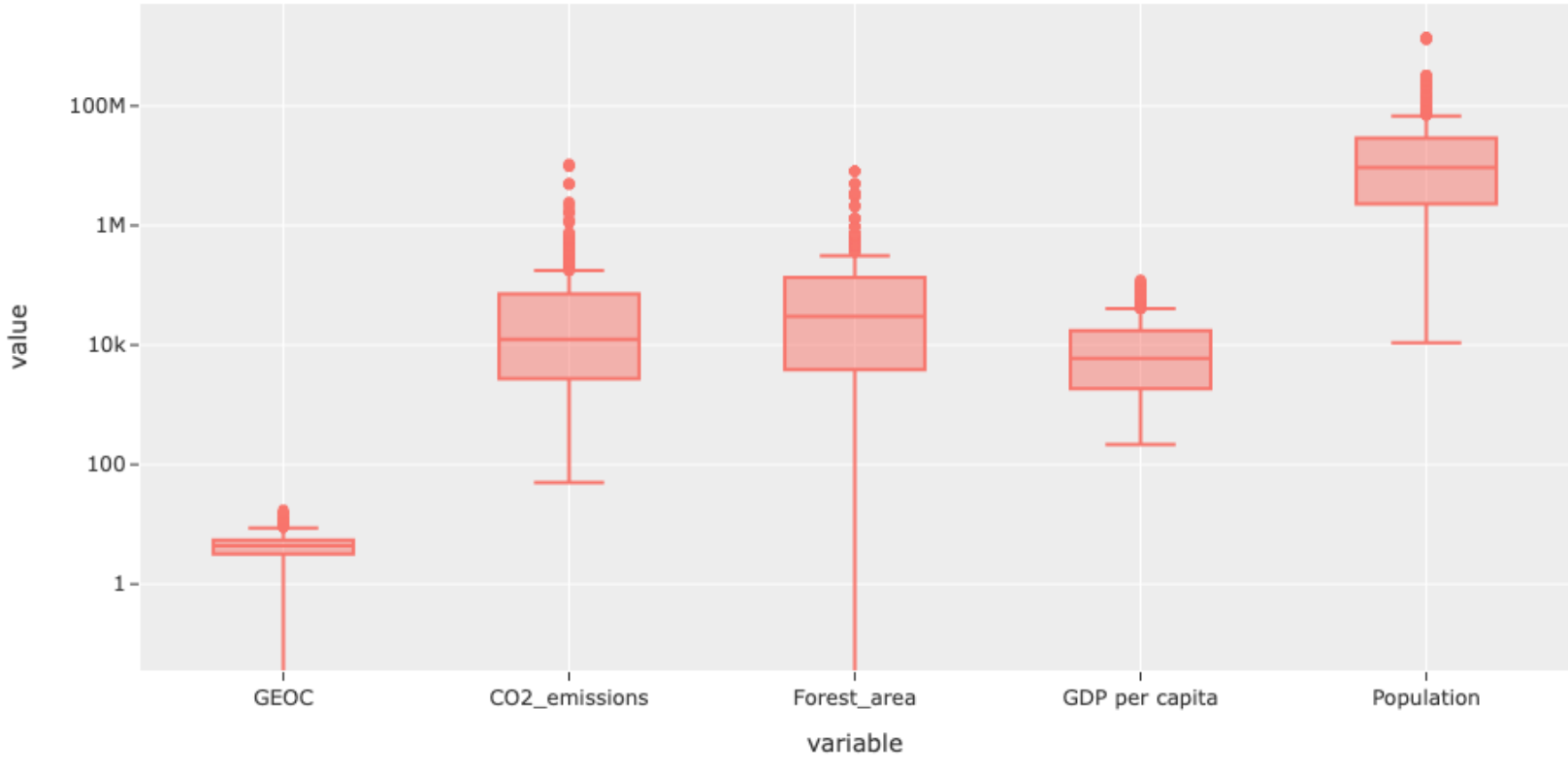
In [9]:
```
df.describe()
```

Out[9]:

	Year	GEOC	CO2_emissions	Forest_area	GDP per capita	Population
count	1308.0	1308.0	1308.0	1308.0	1308.0	1308.0
mean	2015.5366972477063	4.5579892966360855	191670.9785626911	235255.8676376147	14511.83857798165	41281363.1559633
std	2.2774699939082246	2.0462013271784163	900164.8856545219	845993.0031725528	19943.455839430702	149617932.78422123
min	2012.0	0.0	50.0	0.0	216.97	10940.0
25%	2014.0	3.2175000000000002	2720.0	3865.0	1880.2975000000001	2314197.25
50%	2016.0	4.39	12405.0	30256.5	5967.675	9376902.0
75%	2018.0	5.46	71675.0	133863.95	17505.555	28842106.75
max	2019.0	17.63	10707219.73	8153116.0	123678.7	1407745000.0

Now we will use plotly to make box plot to visualize these statistics.

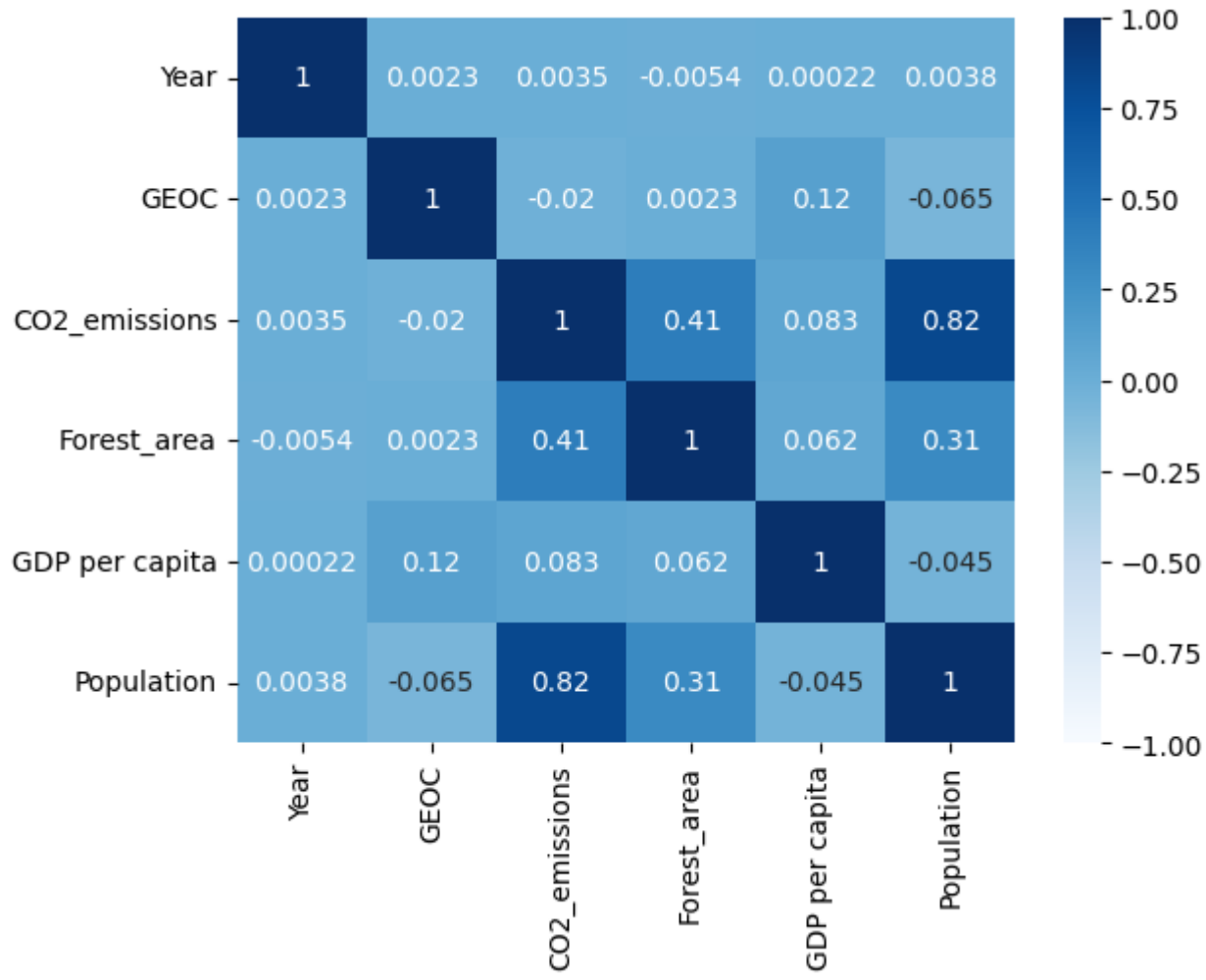
In [ ]:
```
px.box(df.drop(['Country', 'Year'], axis = 1), log_y = True, template = 'ggplot2')
```



One of our biggest analysis goals for this project is to look at the correlation between these variables and discover patterns.
One exploratory question could be: Does higher population mean more CO2 emissions?

So we could create a correlation heatmap using seaborn to visualize the significance of each correlation between the variables.

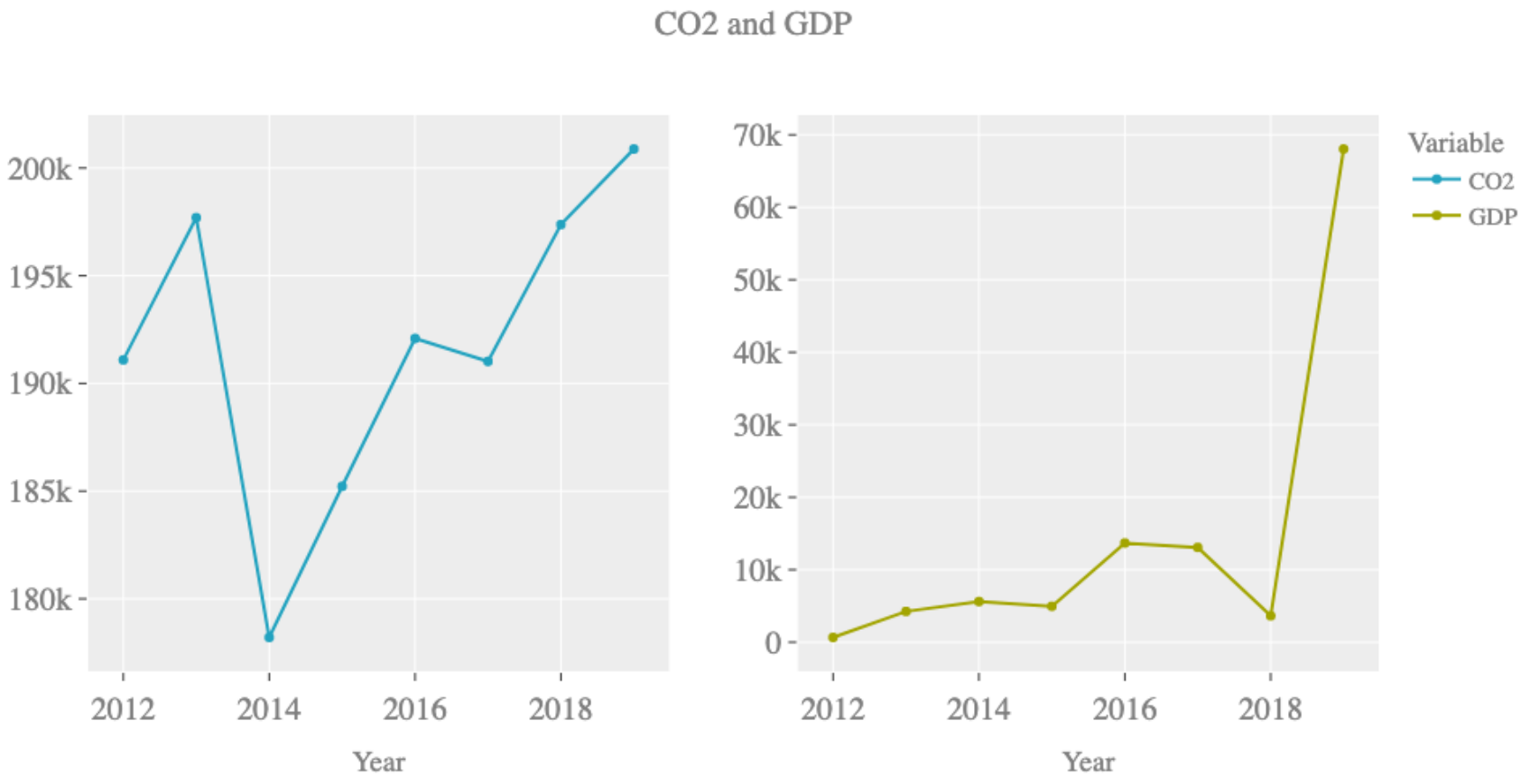
In [10]:
```
correlation_plot = sns.heatmap(df.corr(), vmin = -1, cmap="Blues", annot=True)
```



Intially we can see that there is hardly any type of negative correlation, meaning that for all of our variables as they get larger all of the other variables get larger as well.
Now let's further explore the correlation between CO2 emissions and GDP per capita with a line plot.

In [ ]:
```
plot1_data = df.groupby('Year').CO2_emissions.mean().reset_index()
plot2_data = df.groupby('Year')['GDP per capita'].mean().reset_index()

trace1 = go.Scatter(
    x=plot1_data['Year'],
    y=plot1_data['CO2_emissions'],
    name='CO2',
    marker=dict(
        color='rgb(34,163,192)'
               )
)
trace2 = go.Scatter(
    x=plot2_data['Year'],
    y=df['GDP per capita'],
```

```
    name='GDP',
    yaxis='y2'

)

fig = make_subplots(rows=1, cols=2)
fig.add_trace(trace1, row=1, col=1)
fig.add_trace(trace2, row=1, col=2)

fig.update_layout(legend_title="Variable",
    title = 'CO2 and GDP',
    template = 'ggplot2',
    font=dict(
        family="Roboto",
        size=15,
        color="Grey"
    )
)
fig.update_xaxes(title = 'Year', tickfont_size=20)
fig.update_yaxes(tickfont_size=20)
fig.show()
```



---

3. Planned work

Questions

1. Does a country's GDP increase/decrease as it spends more on education based on the size of the population of the country?
2. How has CO2 Emissions and Forest Area changed since the year 2000, and is there any correlation between them?

Proposed approaches

1. Fit Multiple Linear Regression Model with GDP as the outcome variable and Education Expenditure and Population as the predictors. Compare GDP and Expenditure and Population individually to discover how they interact with each other by themselves, then explore the MLR model.
2. Create Line Graph showing CO2 emissions on the Y-axis and Forest Area on the X axis and produce lines for each year. Create Scatterplots for each year to compare the changes in CO2 emissions and Forest Area. Create correlation matrix to find correlations between variables.