In [1]:
```
# libraries
import numpy as np
import pandas as pd
import altair as alt
```

PSTAT 100 Project plan

Group information

Group members:

Kathy Wu, Nathan Lai, Tymee Wang, Yuchen Fang

Contributions:

Kathy Wu: Tidied and visulized the dataset, Planned work

Nathan Lai: Data Description, Planned work

Tymee Wang: Background Information, Planned work

Yuchen Fang: Initial exploration, Planned work

---

0. Background

ESG is an abbreviation for Environmental, Social, and Governance, which is a combination of three categories of non-financial factors that are increasingly applied by investors as part of their analysis process to evaluate material risks and growth opportunities nowadays. However, in order to better align with the global goals, the World Bank Group rearranges it in a new data framework which further classifies 17 key sustainability themes based on the original environmental, social, and governance categories. The World Bank Group believes that these themes are crucial for financial sector representatives to consider when assessing the contribution of investments or policies to sustainable development.

Our project will mainly focus on analyzing the reported ESG data from the year 2000 to 2020. We would like to see which region or continent is the most sustainable based on the assessment and whether there is any correlation among the three categories. Also, any events that are not included in the assessing framework but would influence the whole sustainability result is also the question that we would like to pay attention to.

For our data here, we keep the division of the three parts: Environmental, Social, and Governance. The Environmental part encompasses key themes that focus on the economic performance given a country's natural resource endowment, management and supplementation, and also accounts for other factors such as food security for stable long-term economic growth. For the Social part, it indicates how good a country's performance is on its efficacy in meeting the basic needs of its population and reducing poverty, management of social and equity issues and investment in human capital and productivity. For the Governance part, it evaluates a country's sustainability by its institutional capacity to support long-term development, including political, financial and legal aspects.

The motivation of collecting this data is to study how large the gap among countries all over the world would have on their development and sustainability in these three aspects and what factors may contribute to such a situation. The things that we can potentially learn is what changes the country with lower sustainability can make to improve their current status and the overall developing trend of the world as a whole.

---

1. Data description

Basic information

General description:
In order to shift financial flows so that they are better aligned with global goals, the World Bank Group (WBG) is working to provide financial markets with improved data and analytics that shed light on countries' sustainability performance.

This dataset provides information on sustainability themes spanning environmental, social, and governance categories. Along with new information and tools, the World Bank can develop research on the correlation between countries' sustainability performance and the risk and return profiles of relevant investments.

Source:
Environment, Social and Governance Data, The World Bank is classified as Public under the Access to Information Classification Policy.

This dataset is licensed under Creative Commons Attribution 4.0.

Collection methods:
Our data is census data, most of the data values in topic Governance and Social are obtained from surveys, and most of the data in topic Environment is collected by using scientific equipment.

Sampling design and scope of inference:
Sampling frame: All countries reporting environment, social and governance data.

Sampling mechanism: Census

Scope of inference: None

Data semantics and structure

Units and observations:
State the observational units.

Variable descriptions:

Name	Variable Description	Topic	Type	Units of measurement
fore_area	*Forest area*	Environment	Numeric	% of land area
fore_dep	Adjusted savings: *net forest depletion*	Environment	Numeric	% of GNI
natu_res_dep	Adjusted savings: *natural resources depletion*	Environment	Numeric	% of GNI
pop_denst	*Population density*	Environment	Numeric	people per sq. km of land area
rate_labor	Ratio of female to male *labor force participation rate*	Governance	Numeric	% (modeled ILO estimate)
gdp_grow	*GDP growth*	Governance	Numeric	annual %
unemp_rate	*Unemployment*, total	Social	Numeric	% of total labor force (modeled ILO estimate)
life_exp	*Life expectancy* at birth, total	Social	Numeric	years
acce_electr	Access to *electricity*	Social	Numeric	% of population
mortal_rate	*Mortality* rate, under-5	Social	Numeric	per 1,000 live births
acce_fuel_tech	Access to *clean fuels* and *technologies* for cooking	Social	Numeric	% of population
pop_65	Population *ages 65 and above*	Social	Numeric	% of total population
ferti_rate	*Fertility rate*, total	Social	Numeric	births per woman

In [2]:
```
show = pd.read_csv('test.csv').drop(columns = 'Unnamed: 0')
show.head()
```

Out[2]:

	Region	Country Name	Year	Access to clean fuels and technologies for cooking (% of population)	Access to electricity (% of population)	Forest area (% of land area)	GDP growth (annual %)	Life expectancy at birth, total (years)	Population density (people per sq. km of land area)	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)	Unemployment, total (% of total labor force) (modeled ILO estimate)
0	Africa	Congo, Dem. Rep.	2000	1.0	6.700000	63.474118	-6.910927	50.041	20.778470	96.881158	2.904
1	Africa	Congo, Dem. Rep.	2001	1.2	7.314364	63.177257	-2.100173	50.667	21.361917	96.724567	2.888
2	Africa	Congo, Dem. Rep.	2002	1.4	7.915845	62.880395	2.947765	51.385	21.998487	96.617267	2.871
3	Africa	Congo, Dem. Rep.	2003	1.6	8.512090	62.583534	5.577822	52.144	22.683921	96.539211	2.860
4	Africa	Congo, Dem. Rep.	2004	1.9	9.105449	62.286672	6.738374	52.917	23.408777	96.488953	2.853

2. Initial explorations

Basic properties of the dataset

Dimensions:
There are 378 rows and 8 columns of variables in the dataset after cleaning.

Missing values:
Since the original data is collected by census, it is possible that some values are missing by chance. Hence, we only select variables from non-missing ones and rank their importance from each category.

Variable summaries:
The dataset consists of 8 numeric variables which are divided into 3 categories: Environment, Governance, and Social.

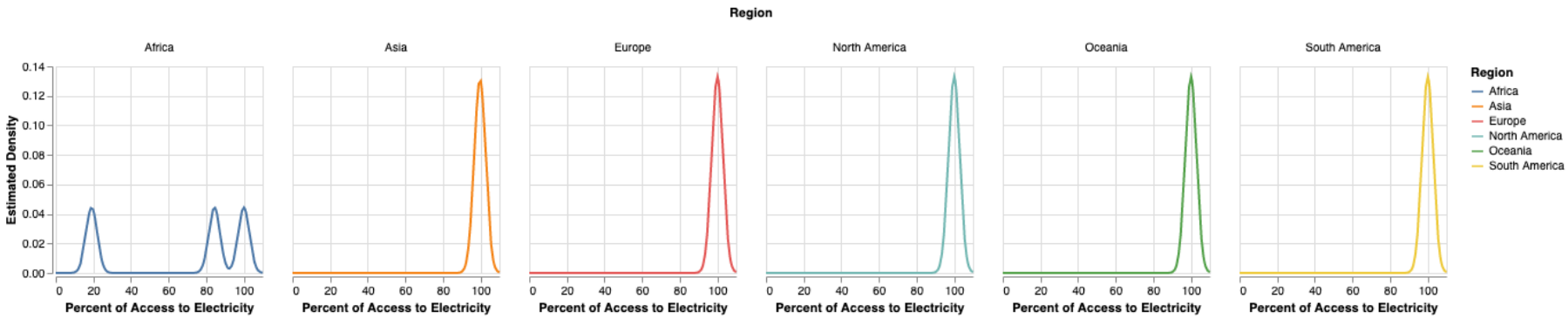
We select 18 countries from 6 continents and 8 most representative variables of them from 2000 to 2020 and no missing values after cleaning.

Under the environment part, we have `Population density` and `Forest area`. Governance variables include `GDP growth (annual %)` and `Ratio of female to male labor force participation rate (%)`. Lastly, social variables consist of `Access to clean fuels and technologies for cooking`, `Life expectancy at birth, total (years)`, `Unemployment, total (% of total labor force)`, and `Access to electricity (% of population)`.

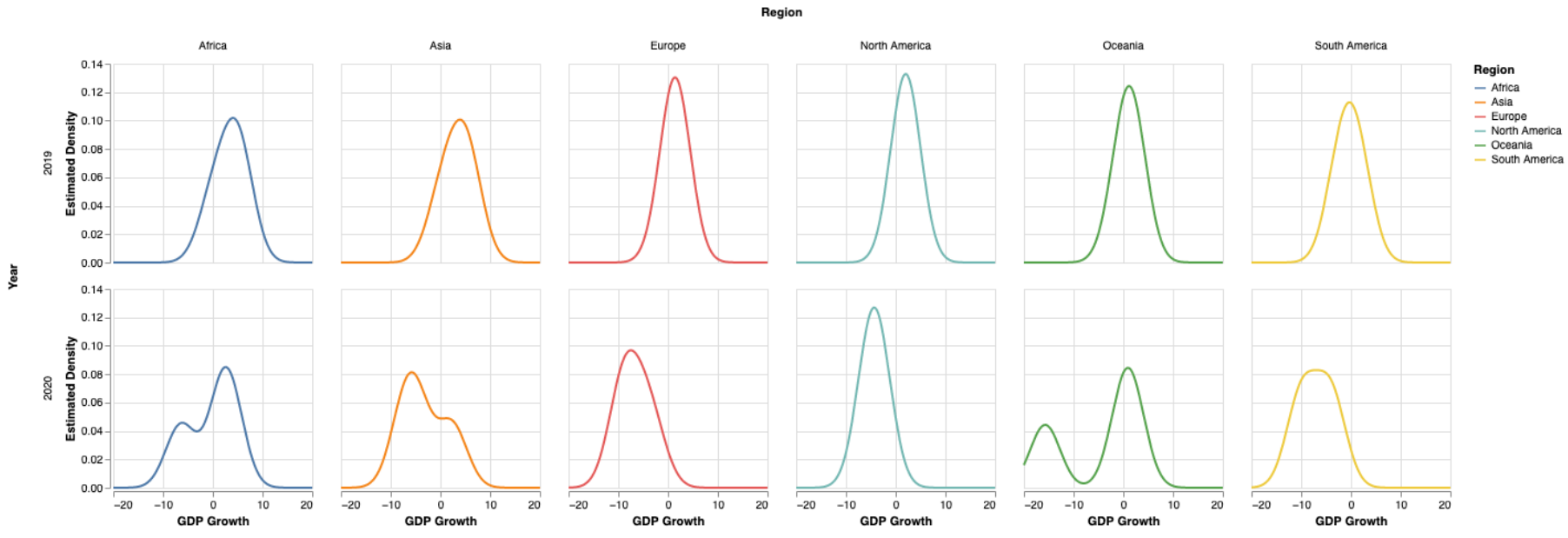
Thus, there are 168 different values of variables for each country.

Exploratory analysis

The following graphs are the plots of distribution of accessibility to electricity in different regions in year 2020. We can see that except Africa region, other regions all have a relatively high accessibility to electricity.



The follow graphs are the plots of the distribution of growth of GDP in 2019 and 2020. We choose these two years because COVID-19 happened in 2020, and thus we can see that in 2020, the GDP growth shift slightly to negative in almost all the regions.



3. Planned work

Questions

1. *Which region is more sustainable and how COVID-19 affects the GDP growth?*
2. *Will the environment and social motivate governance?*

Proposed approaches

1. *We plan to plot a density plot showing the change during the pandemic and see how the curve differs.*
2. *Using heatmap to see the correlation between environmental, social, and governance variables. Or probably using PCA.*