Class Project - Stage 1

Group Information

Group Members:

- Kayla Katakis
- Yihan Cao
- Hung Nguyen

Contributions:

- Yihan - Prepared data description and documentation
- Kayla - Data Tidying and Initial Observations/Analysis
- Hung - Initial Observations/Analysis and Next Steps

Background

This project will cover annual greenhouse gasses (GHG) emissions given the activities and countries that produced these greenhouse gasses in the past five years from 2017 to 2021. GHG emission was and has been one of the most controversial topics worldwide. Although it is inevitable since rapid technological advancements have been taking place on a global scale, the increase in the emission of these gasses have severe consequences to the environment and humans' life quality. Thus, scientists and professionals have been attempting to control and aiming to lower the emission levels while simultaneously maintaining optimal levels of economic and technological activites. In this project, we will utilize and explore the dataset provided by IMF data library which is free for users to access.

Data Description

The orginal website leads us to the GHG data set is World Bank Open Data. Then we navigate to data catalog and find library network where we further to access IMF data library. It is convienent for us to access the GHG data because it is on the dashboard.

There are a lot of variables in the data set. After looking through the table provided by the IMF website, we decided to narrow down our interested variables to the following:

`Country` : The name of the continent or subregion

`Indicator` : General description of GHG

`Industry` : The activities causing the GHG emission

`Gas_Type` : Type of GHG gas

`Unit` : Unit of the measured gas

`F2017` - `F2021` : The measurement of GHG gases in the corresponding year

All observations of the data set are accounts of annual greenhouse gas air emissions, recorded in million metric tons of the carbon dioxide equivalent.

Read in data and Select desired variables

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

In [2]:
```
ghg = pd.read_csv('data/Annual_GHG.csv', encoding = 'latin-1')

ghg = ghg.loc[:,['Country', 'Industry','Gas_Type'\
                 ,'F2017','F2018','F2019','F2020','F2021']]
ghg.head(5)
```

Out[2]:

	Country	Industry	Gas_Type	F2017	F2018	F2019	F2020	F2021
0	Advanced Economies	Agriculture, Forestry and Fishing	Carbon dioxide	193.054238	191.720412	191.165538	187.134711	193.971754
1	Advanced Economies	Agriculture, Forestry and Fishing	Fluorinated gases	0.982652	0.851009	0.816072	0.778334	0.729931
2	Advanced Economies	Agriculture, Forestry and Fishing	Greenhouse gas	1380.725829	1388.771814	1386.321969	1352.443269	1345.263619
3	Advanced Economies	Agriculture, Forestry and Fishing	Methane	618.262461	620.189092	613.713837	611.371924	592.283622
4	Advanced Economies	Agriculture, Forestry and Fishing	Nitrous oxide	568.426479	576.011301	580.626521	553.158300	558.278313

Tidying

Each observation represents a region's emissions of a specific gas type in a given industry from 2017-2021. The data is already tidy, but for clarity's sake we will rename the `Country` column to `Region`.

In [3]:
```
ghg = ghg.rename(columns = {'Country': 'Region', 'Gas_Type': 'Gas Type'})
```

In [4]:
```
ghg.head()
```

Out[4]:

	Region	Industry	Gas Type	F2017	F2018	F2019	F2020	F2021
0	Advanced Economies	Agriculture, Forestry and Fishing	Carbon dioxide	193.054238	191.720412	191.165538	187.134711	193.971754
1	Advanced Economies	Agriculture, Forestry and Fishing	Fluorinated gases	0.982652	0.851009	0.816072	0.778334	0.729931
2	Advanced Economies	Agriculture, Forestry and Fishing	Greenhouse gas	1380.725829	1388.771814	1386.321969	1352.443269	1345.263619
3	Advanced Economies	Agriculture, Forestry and Fishing	Methane	618.262461	620.189092	613.713837	611.371924	592.283622
4	Advanced Economies	Agriculture, Forestry and Fishing	Nitrous oxide	568.426479	576.011301	580.626521	553.158300	558.278313

In this project, we would like to focus on the emissions of greenhouse gasses from each continent/subregion and their industries as well as the global level of greenhouse gas emissions. In the cell below, we filtered the dataset by removing regions that are irrelevant to our topic of interest.

In [5]:
```
ghg = ghg[ghg['Region'] != 'Advanced Economies']
ghg = ghg[ghg['Region'] != 'Emerging and Developing Economies']
ghg = ghg[ghg['Region'] != 'G7']
ghg = ghg[ghg['Region'] != 'G20']
ghg
```

Out[5]:

	Region	Industry	Gas Type	F2017	F2018	F2019	F2020	F2021
50	Africa	Agriculture, Forestry and Fishing	Carbon dioxide	8.598177	8.885628	9.193573	9.519654	9.869859
51	Africa	Agriculture, Forestry and Fishing	Greenhouse gas	801.551149	820.263383	841.909644	859.337790	875.720444
52	Africa	Agriculture, Forestry and Fishing	Methane	533.778860	548.656038	563.256921	574.852827	584.961495
53	Africa	Agriculture, Forestry and Fishing	Nitrous oxide	259.174112	262.721717	269.459149	274.965308	280.889091
54	Africa	Construction	Carbon dioxide	93.987951	95.613220	96.690936	87.940823	93.301037
...	...	...	...	...	...	...	...	...
1134	World	Water supply; sewerage, waste management and r...	Carbon dioxide	212.625849	219.993696	226.031515	227.326438	229.040646
1135	World	Water supply; sewerage, waste management and r...	Fluorinated gases	12.641309	13.518851	13.771323	13.137126	12.133499
1136	World	Water supply; sewerage, waste management and r...	Greenhouse gas	2642.549083	2699.002173	2746.707736	2800.593032	2831.028085
1137	World	Water supply; sewerage, waste management and r...	Methane	2284.968589	2331.721563	2371.492339	2422.875824	2450.209878
1138	World	Water supply; sewerage, waste management and r...	Nitrous oxide	132.313337	133.768063	135.412558	137.253644	139.644062

939 rows × 8 columns

Initial Observations and Explorations

The GHG dataset, as we have tidied, is made up of 939 observations across 8 variables. There is no missing data, which makes our future analysis much easier. The `Region` variable contains the continent, subcontinent, or region in which the gases were measured, and the `Industry` column indicates the industry that caused the emissions for the gas described in the `Gas Type` column. Columns `F2017 - F2021` detail the measurements for that gas that were recorded each year.

There are 5 continents included in the dataset, naming Asia, Europe, Oceania, Americas, and Africa, along with 14 subregions within these continents. The region `World` describes the level of gas emissions globally.

In [6]:
```
ghg.shape
```

Out[6]:
```
(939, 8)
```

In [7]:
```
ghg.Region.value_counts()
```

Out[7]:
```
World                           50
Northern Europe                 50
Asia                            50
Australia and New Zealand       50
Western Europe                  50
Western Asia                    50
Eastern Europe                  50
Europe                          50
Southern Europe                 50
Oceania                         50
Northern America                45
Americas                        45
Latin America and the Caribbean 45
Eastern Asia                    45
Southern Asia                   44
Northern Africa                 43
South-eastern Asia              43
Sub-Saharan Africa              43
Central Asia                    43
Africa                          43
Name: Region, dtype: int64
```

From the table below, it is evident that there is an increase in the level of greenhouse gas emission from 2017 to 2021 within each region and globally. In 2017, the mean greenhouse gas emission was recorded at 9885.44 million metric tons of CO2 equivalent which gradually increased every year to 10192.51 million metric tons of CO2 equivalent in 2021. However, noticeably, there was a slight decerase in greenhouse gas emission in 2020 which was recorded at 9783.68. We would assume that this was caused by the COVID-19 pandemic when social distancing and isolation laws were in place.

In [8]:
```
ghg.groupby(['Region', 'Gas Type']).mean()
```

```
C:\Users\29749\AppData\Local\Temp\ipykernel_16788\159007057.py:1: FutureWarning: The default value of numeric_only in DataFra
meGroupBy.mean is deprecated. In a future version, numeric_only will default to False. Either specify numeric_only or select
only columns which should be valid for the function.
  ghg.groupby(['Region', 'Gas Type']).mean()
```

Out[8]:

Region	Gas Type	F2017	F2018	F2019	F2020	F2021
Africa	**Carbon dioxide**	294.771920	301.757536	305.917024	282.313748	298.213724
	Fluorinated gases	36.848115	41.231802	45.130290	49.100145	53.456377
	Greenhouse gas	624.226028	639.697797	652.554296	628.945105	651.841588
	Methane	252.421211	258.726723	264.583885	262.581504	266.748683
	Nitrous oxide	65.978463	66.843997	68.514300	69.319810	70.842268
...	**...**	...	...	...	...	...
World	**Carbon dioxide**	7349.200712	7505.287384	7522.587904	7152.564764	7523.279165
	Fluorinated gases	186.924313	199.767228	208.482461	218.620274	230.696594
	Greenhouse gas	9885.437494	10082.535706	10118.368100	9783.678030	10192.504839
	Methane	1779.833461	1803.300437	1813.184431	1842.909536	1856.816331
	Nitrous oxide	569.479009	574.180656	574.113305	569.583456	581.712750

100 rows × 5 columns

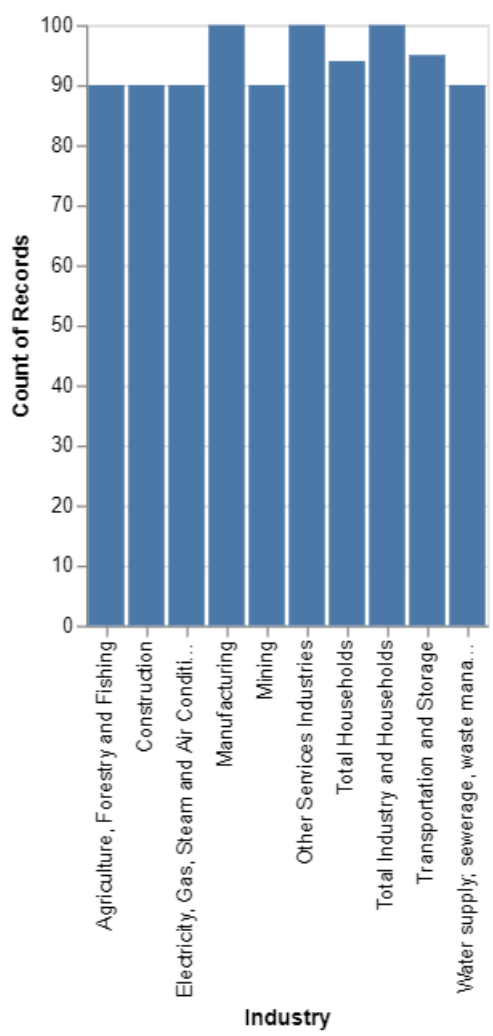
Visualizations

Below, we see that each industry is relatively evenly counted, with Manufacturing, Other Services, and Total Household having slightly more representation

In [9]:
```
alt.Chart(ghg).mark_bar().encode(
    x = alt.X('Industry'),
    y = alt.Y('count()')
)
```

```
C:\Users\29749\AppData\Local\Programs\Python\Python310\lib\site-packages\altair\utils\core.py:317: FutureWarning: iteritems i
s deprecated and will be removed in a future version. Use .items instead.
  for col_name, dtype in df.dtypes.iteritems():
```

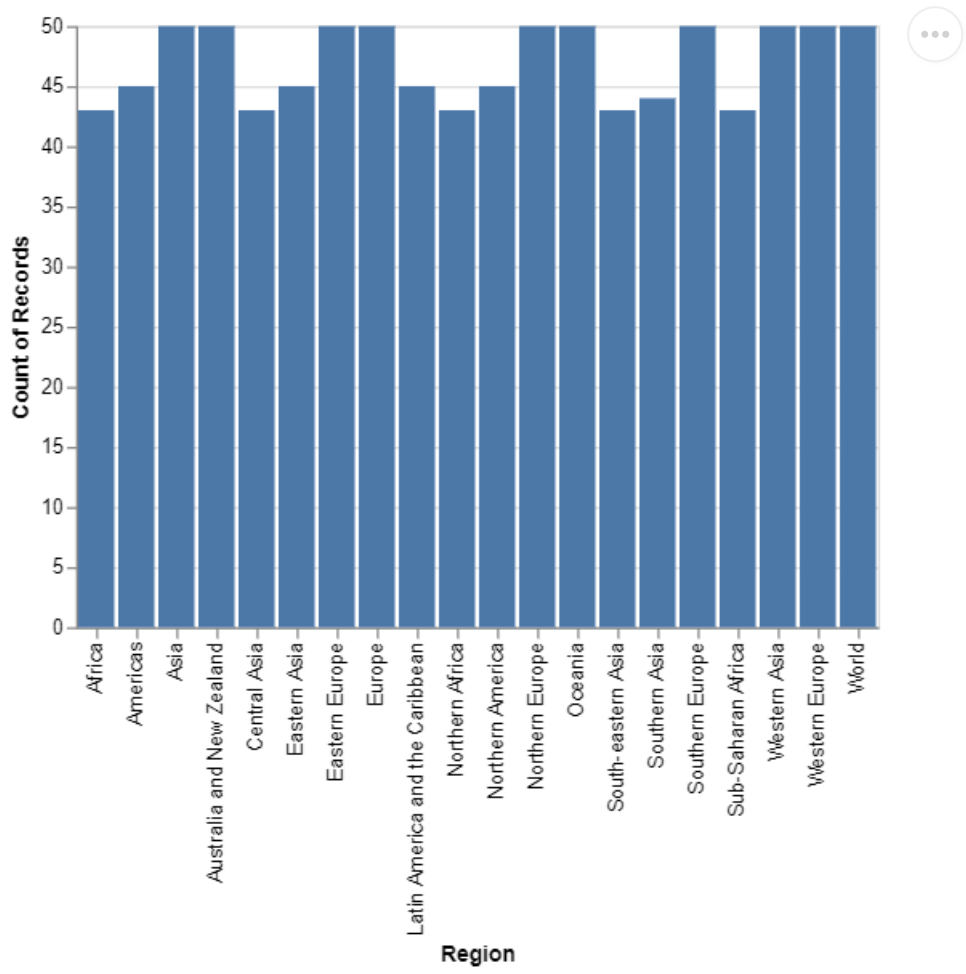
Out[9]:



Below, we can see the different regions and continents in which observations were recorded. The regions are distributed relatively uniformly, with the regions with the most and least observations differing by less than 10 records.

In [10]:
```
alt.Chart(ghg).mark_bar().encode(
    x = alt.X('Region'),
    y= alt.Y('count()'),
)
```

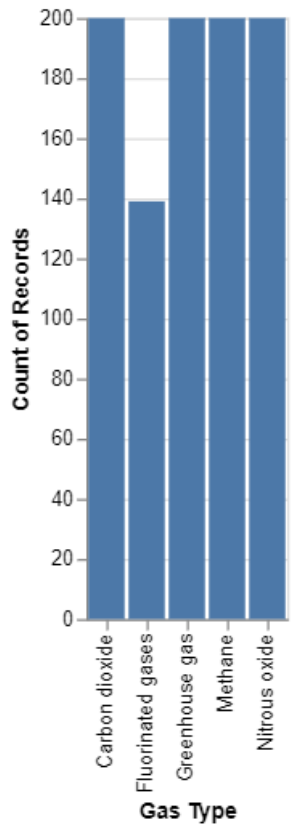
Out[10]:



Below, we can see the different gas types that were measured. Fluorinated gases are slightly less common in this dataset.

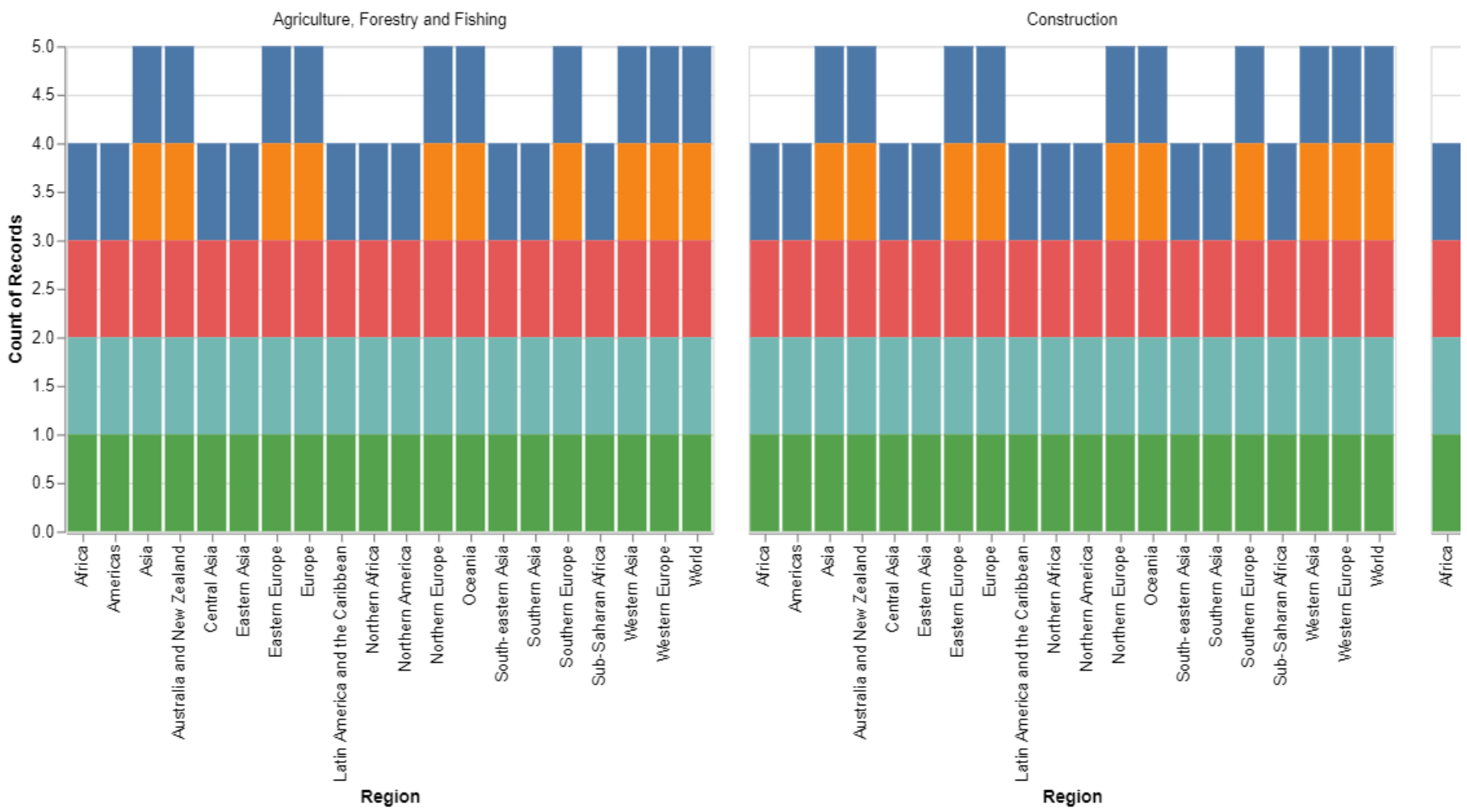
In [11]:
```
alt.Chart(ghg).mark_bar().encode(
    x = alt.X('Gas Type'),
    y = alt.Y('count()')
)
```

Out[11]:



In [12]:
```
alt.Chart(ghg).mark_bar().encode(
    x=alt.X('Region'),
    y=alt.Y('count()'),
    color='Gas Type').facet(
        column = 'Industry')
```

Out[12]:



From this bar chart, we can see that there are fewer records of fluoride gas emission levels. It is also evident that Africa doesn't have as many records of fluoride gas emission levels as other continents.

Next Steps

Now that we have a good idea about what our data looks like and what it means, we can begin to investigate some potnetial research questions.

In this project, there are two main questions we want to answer:

1. What is the relationship between industry and emissions? Which industries produce the most emissions, and which have increased/decreased their emissions over time?
2. How does gas type influence emissions over time? Do certain gases require/result in increased emissions? Do certain regions emit more of any specific gases?

Both of these questions will benefit from linear regression models in order to investigate the relationship between these variables over time. We will also further use visual analysis to look at the relationships between industry and emission as well as region and emission in regards to specific gases.

In [ ]: