PSTAT 100 Project

By Azfal Peermohammed, Raghav Chhabra, Navneet Rajagopal, Isaac Chang

Data Description

We chose the ClimateWatch historical emissions dataset for our project. The data was collected from each country's NDC (Nationally Determined Contributions), which represents a country's commitment to addressing climate change and outlines its specific targets, policies, and measures for mitigating greenhouse gas emissions and adapting to the impacts of climate change. Under the Paris Agreement, every nation participating is required to submit reports every 5 years detailing their green house emission. It consists of observational data of the total greenhouse gas emissions, measured in MtCO2e (Metric Tons of Carbon Dioxide Emitted), of 193 different countries, the European Union, and the World from 1990- 2019. The important variables that we will use in our analysis are the countries and the 30 measurements of greenhouse gas emissions from different years.

It is important to note that some countries have negative emissions; this being Montenegro, Georgia, Fiji, Latvia, Micronesia, Cape Verde, Finland, Bhutan, Solomon Islands, and Romania. This is because these countries remove more carbon than they emit, either due to techniques in place to reduce emissions or due to a high volume of greenery and vegitation.

Question

Understanding how different countries and regions contribute to greenhouse gas emissions provide insights into the sources of global climate change, enabling more targeted and effective mitigation strategies. Additionally, identifying nations with particularly notable emissions profiles can highlight areas for potential policy influence and technological intervention. This analysis ultimately equips us to more effectively combat the climate crisis and work towards global environmental sustainability.

This led us to the question: How do different countries and regions contribute to greenhouse gas emissions around the world in the past thirty years and can we use this data to explain what causes variability in emissions?

Data Analysis

In [1]:
```
%%capture --no-display
!pip install pycountry_convert
import pandas as pd
import numpy as np
import altair as alt
import pycountry_convert as pc
```

Due to the immense amount of countries in the data set and number of years, it was very difficult to make effective plots. Since we are trying to visually represent the conditional distributions of CO2 emissions based on the region and year, we used an external package (pycountry_convert) to bin the countries into continents which allow us to effectively visualize the data.

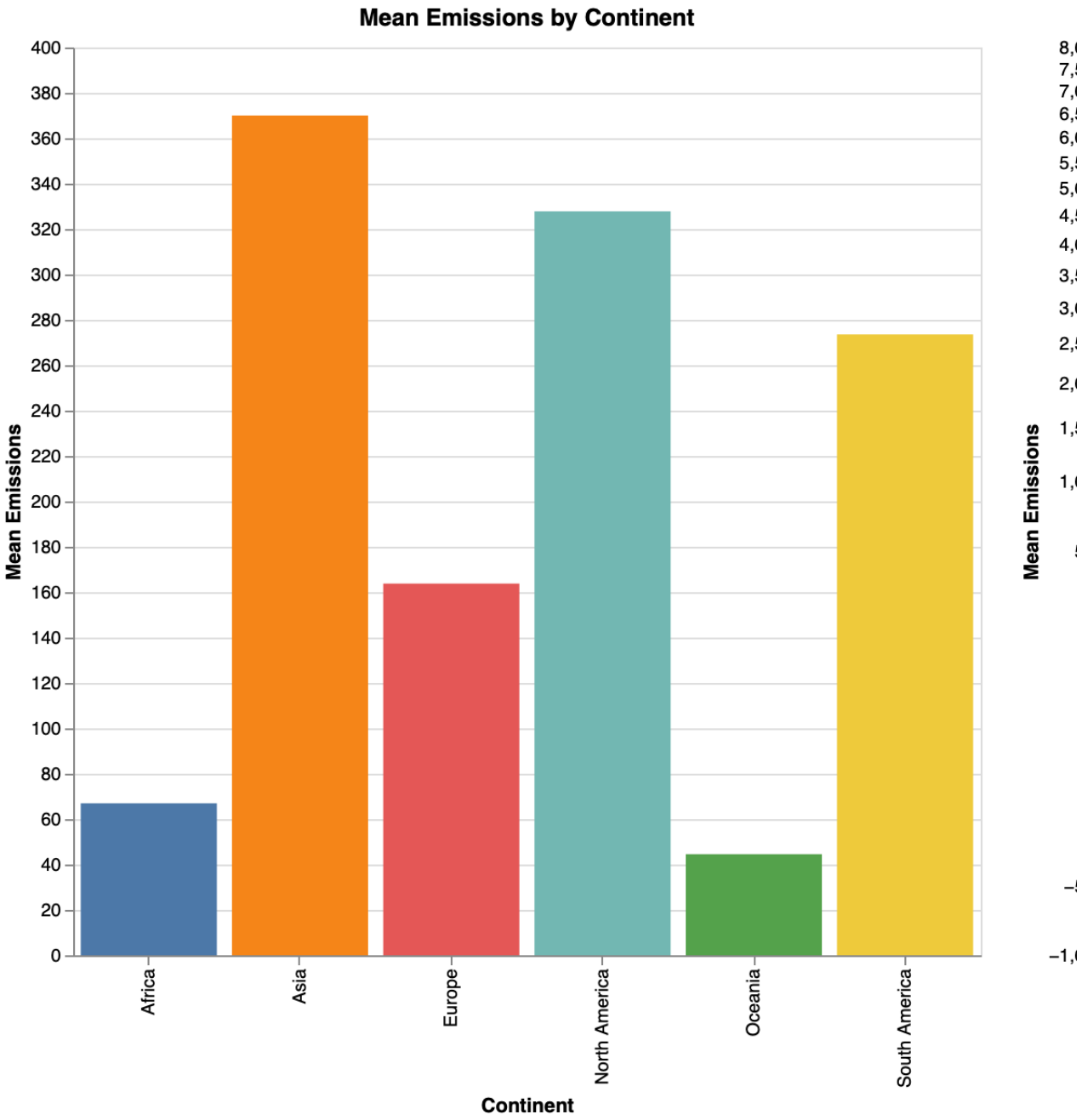
In [2]:
```
data= pd.read_csv('data/historical_emissions/historical_emissions.csv')

# get rid of the columns with no varaince
data_subset = data.drop(columns = ['Data source', 'Sector', 'Gas', 'Unit'])


# loking at univariate data
# let's look at the year
# we can make a verticle boxplot of all of the years
# talk about the before and after of missing data. Add this part in. Justify
data_subset.dropna(inplace = True)


# this function uses an externel library to convert all the country names in
# we are doing this to bin the countries into continents to directly answer
# write about this

def country_to_continent(country_name):
    try:
        country_alpha2 = pc.country_name_to_country_alpha2(country_name)
        country_continent_code = pc.country_alpha2_to_continent_code(country
        country_continent_name = pc.convert_continent_code_to_continent_name
        return country_continent_name
    except:
        return None

# this calls the function
data_subset['Continent'] = data_subset['Country'].apply(country_to_continent

# Filter out rows where the continent is None
# explain that some of the rows were things like World and European Union.
# we take them out since we are not interested in them
df_filtered = data_subset.dropna(subset=['Continent'])
continent_df = df_filtered.groupby('Continent').apply(lambda x:x).drop(colum
```

Our motivation is here to aggrigate across the mean for each continent in order to compare the overall emissions through the time frame. This is represented in the left panel below.

The motivation of the right panel is to show the distributions of the carbon emissions in each continent in a alternate meathod.

In [3]:
```
# we take the mean of each continent and compare them
# this is the plot on the left
# we see that asia, north america, and sout america are a lot higher than th
chart = alt.Chart(continent_df.mean(axis = 1).groupby(level = 0).mean().rese
    x='Continent',
    y=alt.Y('Mean:Q', title='Mean Emissions'),
    color='Continent'
).properties(
    title='Mean Emissions by Continent',
    width = 500,
    height = 500

)

chart


# we plot the distributions using boxplots
# talk about how we are skeptical about the means since they could be skewed
# this is the reason we also create the plot on the right
# the right plot reveals that our intution was correct
# The medians by cintinent are a lot closer
# we notice several outliars that are muc h higher for Asia and North Americ
# also just talk about the graph on the right in general this answers the fi
chart2 = alt.Chart(continent_df.mean(axis = 1).reset_index().drop(columns=['
    x='Continent',
    y=alt.Y('Emissions:Q', title='Mean Emissions', scale = alt.Scale(type =
    color='Continent'
).properties(
    title='Distributions of Emissions by Continent',
    width = 500,
    height = 500

)

alt.hconcat(chart,chart2)
```

Out[3]:



The left panel reveals that Asia, South America, and North America produce more carbon emissions than Europe, Africa, and Oceania in the past 30 years. This graph may be misleading, however, when observing the right panel we actually see the distributions of carbon emissions is much closer than the left panel reveals.

What the right panel indicates is that the IQR of Asia and South America is similar. However, North America looks different, which doesn't allign to what the left panel shows. The left panel suggests the distribution of North America should be similar to Asia and South America, but this is not the case, as seen with the right panel. The right panel shows that each continent has outliers that skew the overall mean emissions and indicates that continents may not be the best representation of overall emissions. This led us to use countries to specify our region of interest.

Asian Countries

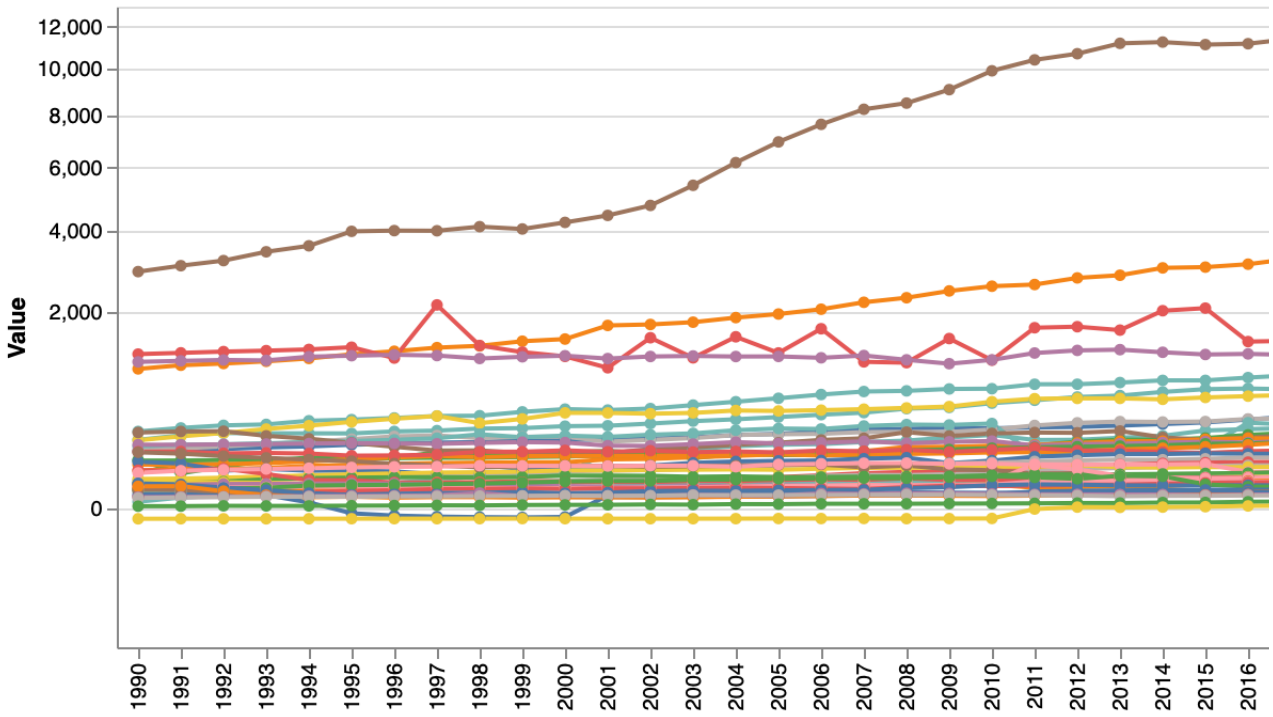
Remember that there are too many countries to successfully plot in Altair, thus we decided the break up the dataset by continent. Below we create a line plot of CO2 emissions colored by country to show the overall trends for all countries in Asia since 1990.

We decided to create tables in order to quanitfy our visual results. The first table we created was the mean of the top 5 emitting countries, sorted in order from highest to lowest. The second table we created was to find the biggest change from 1990 to 2019, also sorted in order from highest to lowest.

Finally we plotted the interesting countries (the ones ranked highest in the list). We are including North Korea since they are a complete dictatorship and we want to see their effects on CO2 emissions out of curiosity.

In [4]:
```
# we will now focus on countries and there differences
# explain that looking at the regions by continent did not explain muc of th
# so now we will look at the indiviual countries
# the below chunk is for Asia

asia_df = df_filtered.loc[df_filtered['Continent'] == 'Asia', :].drop(columr
fig_1 = alt.Chart(asia_df.melt(id_vars='Country', var_name='Year', value_nam
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_1
```

Out[4]:



In [5]:

```
# the above plot is sloppy
# there are too many countries
# we compute coountry means and sort
# we want to rank the following
asia_df['Country means'] = asia_df.iloc[:,1:].mean(axis = 1)
asia_df.sort_values(by = ['Country means'], ascending = False).head()
```

Out[5]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	
1	China	12055.41	11821.66	11385.48	11151.31	11108.86	11228.48	11168.26	10675.66	1
3	India	3363.60	3360.56	3215.07	3076.48	3003.07	2984.52	2804.34	2740.40	
5	Indonesia	1959.71	1692.36	1447.22	1434.46	2067.75	2015.50	1638.39	1702.30	
8	Japan	1134.45	1172.32	1214.59	1229.82	1220.73	1256.16	1298.56	1286.53	
9	Iran	893.78	925.58	912.77	881.05	844.14	844.13	815.31	793.95	

5 rows × 32 columns

In [6]:
```
# find the change of emmissions
# we also comptue the overall change and sort
asia_df["Change"] = asia_df['2019'] - asia_df['1990']
asia_df.sort_values(by = ['Change'], ascending = False).head()
```

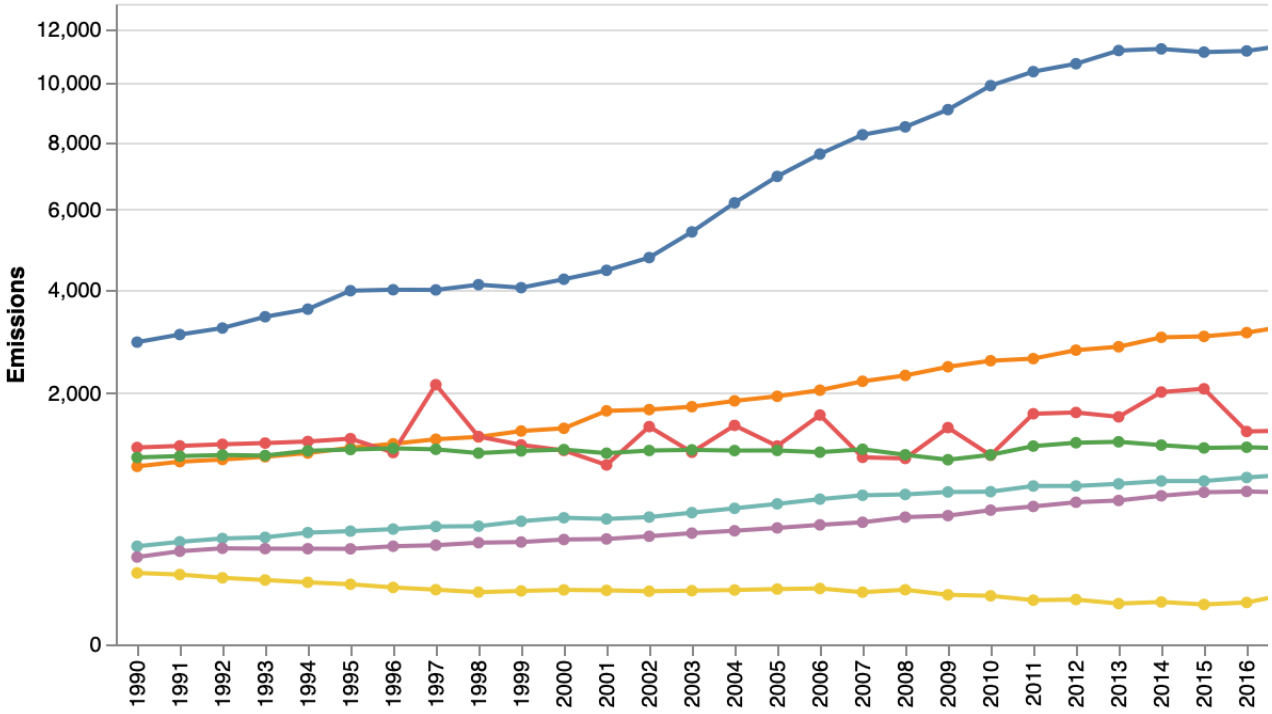
Out[6]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012
1	China	12055.41	11821.66	11385.48	11151.31	11108.86	11228.48	11168.26	10675.66
3	India	3363.60	3360.56	3215.07	3076.48	3003.07	2984.52	2804.34	2740.40
5	Indonesia	1959.71	1692.36	1447.22	1434.46	2067.75	2015.50	1638.39	1702.30
9	Iran	893.78	925.58	912.77	881.05	844.14	844.13	815.31	793.95
11	Saudi Arabia	723.15	715.23	729.31	739.82	731.89	698.29	654.85	638.88

5 rows × 33 columns

In [7]:
```
# now we replot the interesting countris with each other
# We use the above tables to selectt interesting countries
# super interesting ones, China, grew a lot
# india, also grew alot
subset_asia = asia_df.loc[asia_df['Country'].isin(['China', 'India', 'Indone
fig_2 = alt.Chart(subset_asia.melt(id_vars='Country', var_name='Year', value
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt'), title = 'Emissions'
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_2
```

Out[7]:



China has both the greatest mean emissions every year, but also the greatest change since 1990. We also notice, in second, India is present for both mean emissions and change. We notice that most of the countries present on this list are also the fastest changing emission rates among Asian Countries.

After doing some research, we learned that India, China, and Indonesia are all industrial countries, while Iran and Saudi Arabia are petrol states which explains the high emission rates. Interestingly, Japan is present on the list of high mean emissions but not in the top 5 of change since 1990.

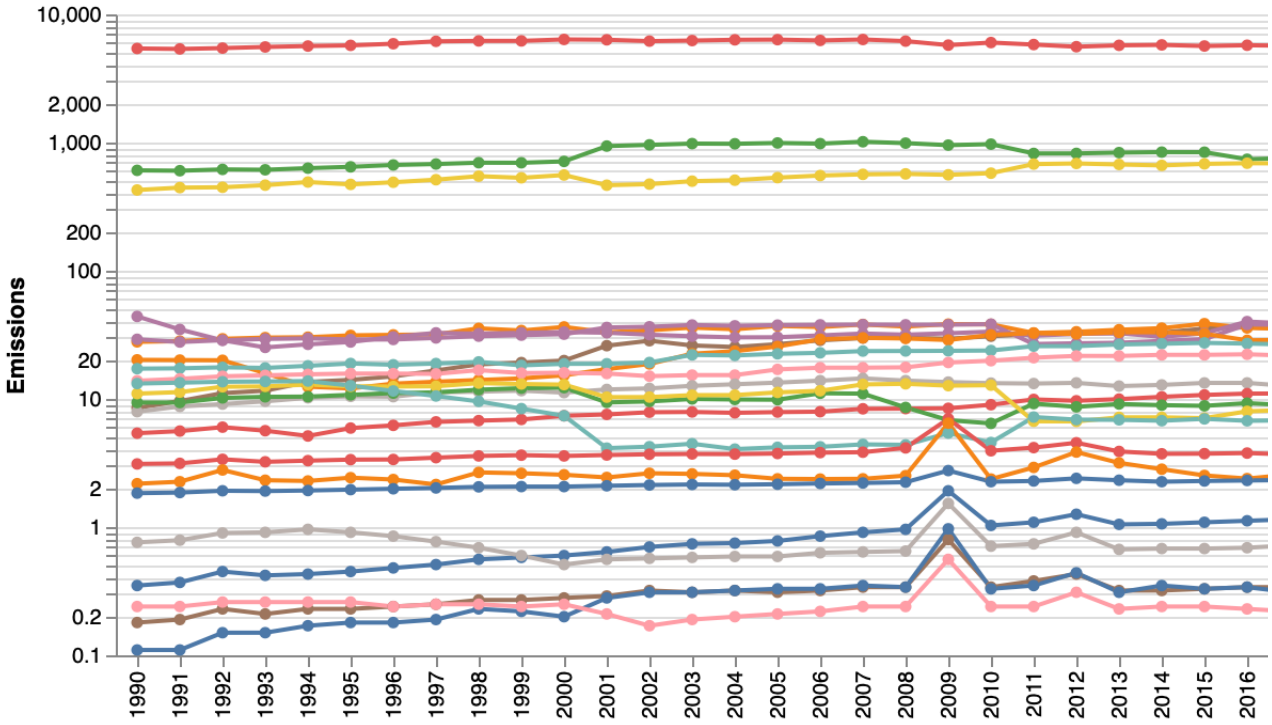
It makes sense that North Korea is low since they are trade limited and lack an industrial sector.

North and Central America

Similar to Asia, we perform a line graph of every countries' mean emissions since 1990 and create a chart of the mean emissions. Unlike Asia, we decided to create a chart of the change from 2008-2009, beacuse we noticed a spike in the CO2 emissions in a lot of countries during this time, we then sorted this change in both the top and bottom 5.

In [8]:
```
# we perform a similar functions to what we did above
# this time though notice that some countries exhibit this trend where there
# be careful about the scale
# just see if you can find a potential reason of why the countries exhibit t

north_df = df_filtered.loc[df_filtered['Continent'] == 'North America', :].c
fig_4 = alt.Chart(north_df.melt(id_vars='Country', var_name='Year', value_na
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'log'), title = 'Emissions')
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_4
```

Out[8]:



In [9]:
```
# looking at the countries with the greatest means
# we do the same above
north_df['Country Means'] = north_df.iloc[:,1:].mean(axis = 1)
north_df.sort_values(by = ['Country Means'], ascending = False).head()
```

Out[9]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	
2	United States	5771.00	5892.37	5689.61	5743.85	5665.21	5779.54	5734.28	5593.25	58
10	Canada	774.29	776.50	757.38	740.67	841.22	842.14	834.47	822.51	82
14	Mexico	670.84	669.63	688.06	689.71	681.94	663.46	674.64	687.03	67
102	Guatemala	38.49	37.05	35.21	35.81	38.64	35.66	34.63	33.39	3
103	Nicaragua	38.41	37.87	37.80	38.05	29.28	28.36	27.42	27.16	2

5 rows × 32 columns

In [10]:
```
# want to see the countries with that big spike in 2009
# notice the recession column this is try to identify wher ethe peaks are oc
north_df['Recession'] = north_df['2009'] - north_df['2008']
north_df.sort_values(by = ['Recession'], ascending = False).head(8)
```

Out[10]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	2011	...	1997	19
162	Bahamas	3.18	3.13	2.57	2.39	2.53	2.83	3.17	3.85	2.92	...	2.15	2
161	Barbados	3.79	3.76	3.69	3.78	3.74	3.73	3.89	4.55	4.17	...	3.48	3
102	Guatemala	38.49	37.05	35.21	35.81	38.64	35.66	34.63	33.39	32.71	...	31.85	31
119	Panama	25.30	21.90	21.74	22.29	22.06	22.07	21.64	21.66	20.96	...	15.69	15
155	Belize	6.85	6.74	6.81	6.75	6.99	6.76	6.87	6.89	7.22	...	10.54	11
171	Antigua and Barbuda	1.22	1.19	1.15	1.12	1.09	1.06	1.05	1.26	1.09	...	0.51	0
176	Saint Lucia	0.74	0.72	0.72	0.69	0.68	0.68	0.67	0.91	0.74	...	0.77	0
104	Cuba	38.19	39.24	38.78	40.10	32.51	30.52	32.51	32.14	31.13	...	32.73	30

8 rows × 33 columns

In [11]:
```
north_df.sort_values(by = ['Recession'], ascending = False).tail()
```

Out[11]:

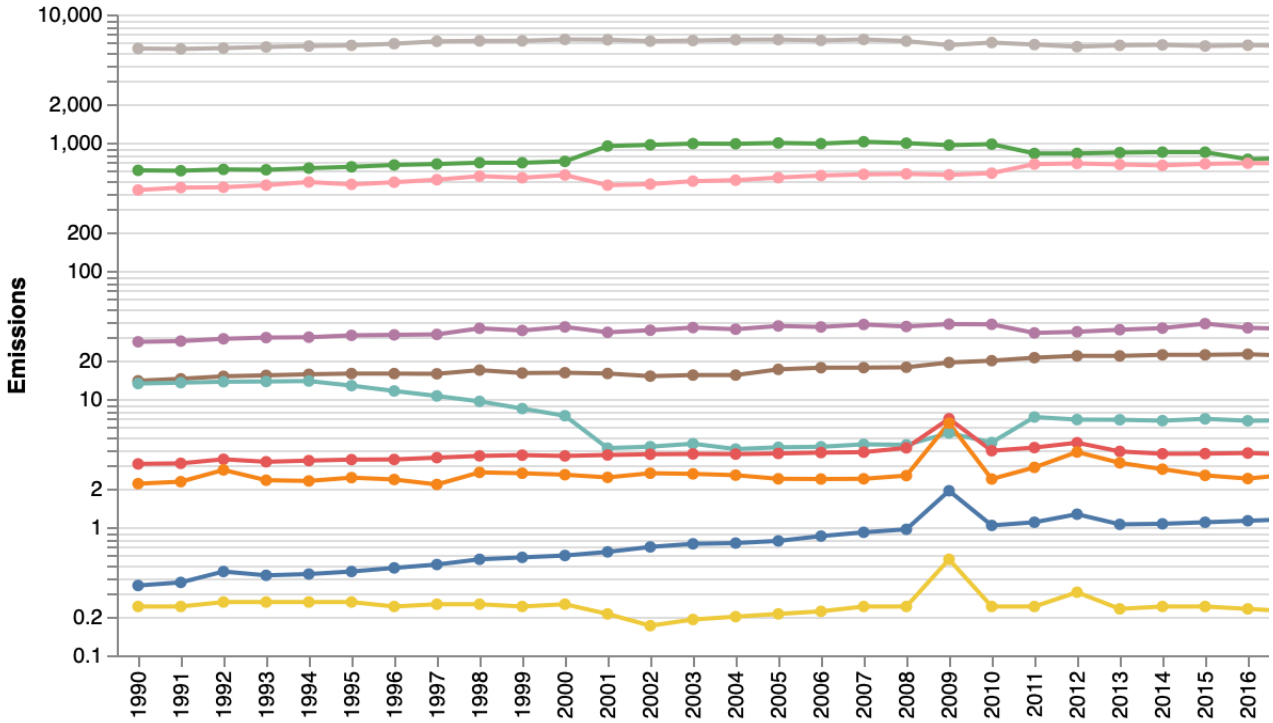
	Country	2019	2018	2017	2016	2015	2014	2013	2012	20
115	Trinidad and Tobago	28.47	28.81	28.87	28.65	32.23	33.14	33.11	32.13	32.
144	Jamaica	10.15	10.34	8.92	9.27	8.83	8.94	9.13	8.67	9
14	Mexico	670.84	669.63	688.06	689.71	681.94	663.46	674.64	687.03	679.
10	Canada	774.29	776.50	757.38	740.67	841.22	842.14	834.47	822.51	821.
2	United States	5771.00	5892.37	5689.61	5743.85	5665.21	5779.54	5734.28	5593.25	5811.

5 rows × 33 columns

In [12]:

```
# This gives you some of the countries talk about the grreatest emmisssions
# Look at some ofthe countries with the peaks
# maybe these epcific countries can provide you a clue about what happened i

subset_north = north_df.loc[north_df['Country'].isin(['Bahamas', 'Barbados',
                                                      'Mexico', 'Antigua and
fig_2 = alt.Chart(subset_north.melt(id_vars='Country', var_name='Year', valu
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'log'), title = 'Emissions')
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_2
```

Out[12]:



We noticed the United States, Canada, and Mexico were the three highest CO2 emitters followed by Guatamala and Nicaragua. This is probably due to the population difference and infrastructure. The United States is an industrial state which is why its so high. Unlike Asia, the countries here do not display as high of a rate of change; furthermore, it is interesting to note the bump in emissions present from 2008 to 2009.

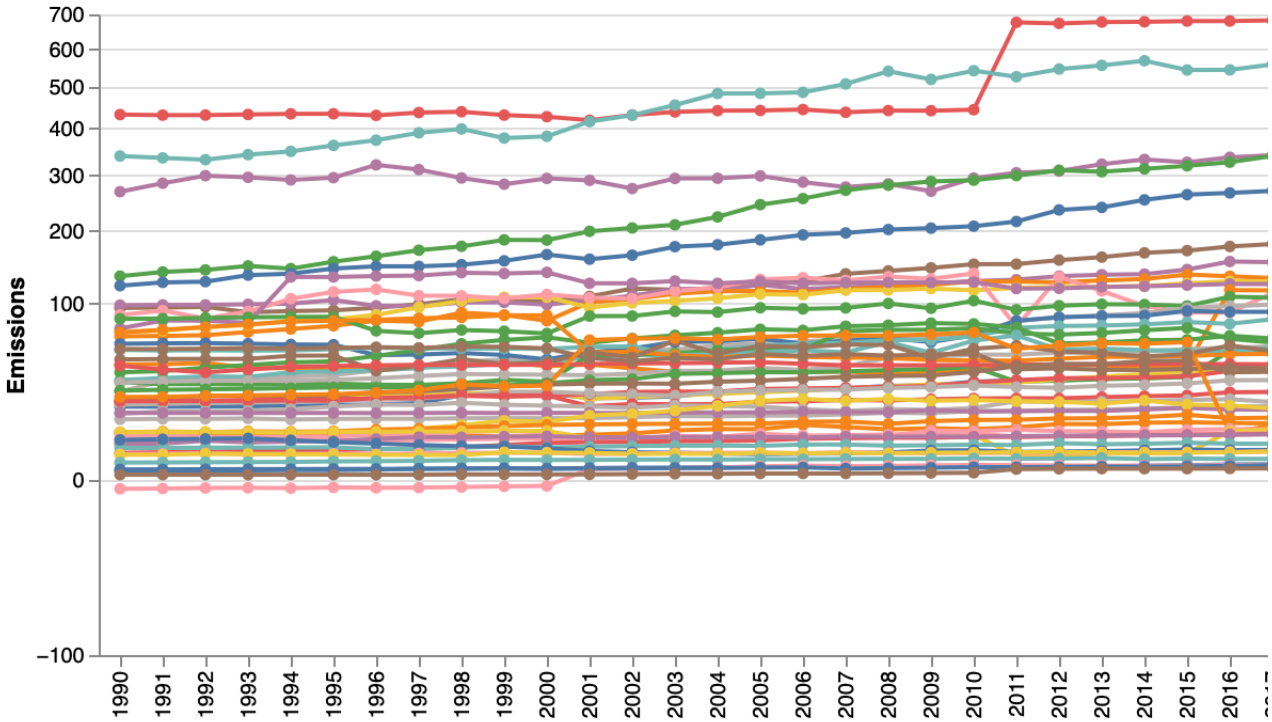
If we look at the graph between 2008 and 2009, the time of the Great Recession, we see many countries have a spike in their emissions. When we analyzed this we found that countries like the Bahamas, Barbados, and Guatemala, among others, had big emission gains over those years. Potentially, as global economic activity contracted, wealthier countries (like the United States and Canada) might have outsourced more industrial production to these smaller nations due to their lower operating costs, thereby increasing local emissions.

African Countries.

Similarly to Asia, we create graphs for the mean emissions per year, create a table, and found the difference. You can go through the tables to see the results, but we are more interested in finding the unique trends that certain countries in this region display.

In [13]:
```
# we do the same process for africa
africa_df = df_filtered.loc[df_filtered['Continent'] == 'Africa', :].drop(cc

fig_4 = alt.Chart(africa_df.melt(id_vars='Country', var_name='Year', value_r
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt'), title = 'Emissions'
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_4
```

Out[13]:



In [14]:

```
# sort by the mean commentate
# grab the top countries and biggest change
africa_df['Mean'] = africa_df.iloc[:,1:].mean(axis = 1)
africa_df['Difference'] = africa_df['2019'] - africa_df['1990']
africa_df.sort_values(by = ['Mean'], ascending=False).head()
```

Out[14]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	2011	...
13	Democratic Republic of the Congo	679.57	676.86	682.01	680.03	679.91	677.77	677.01	673.04	676.09	...
17	South Africa	562.19	556.72	557.46	542.89	542.51	567.33	554.77	544.98	524.74	...
26	Nigeria	354.33	345.70	340.02	335.84	325.43	331.28	321.55	307.78	304.25	...
28	Egypt	351.96	345.01	338.61	325.66	318.29	312.08	306.32	309.38	298.66	...
33	Algeria	282.23	278.23	269.38	265.60	262.52	252.91	239.46	234.99	215.21	...

5 rows × 33 columns

In [15]:
```
# look at the last few
africa_df.sort_values(by = ['Difference'], ascending=False).tail()
```

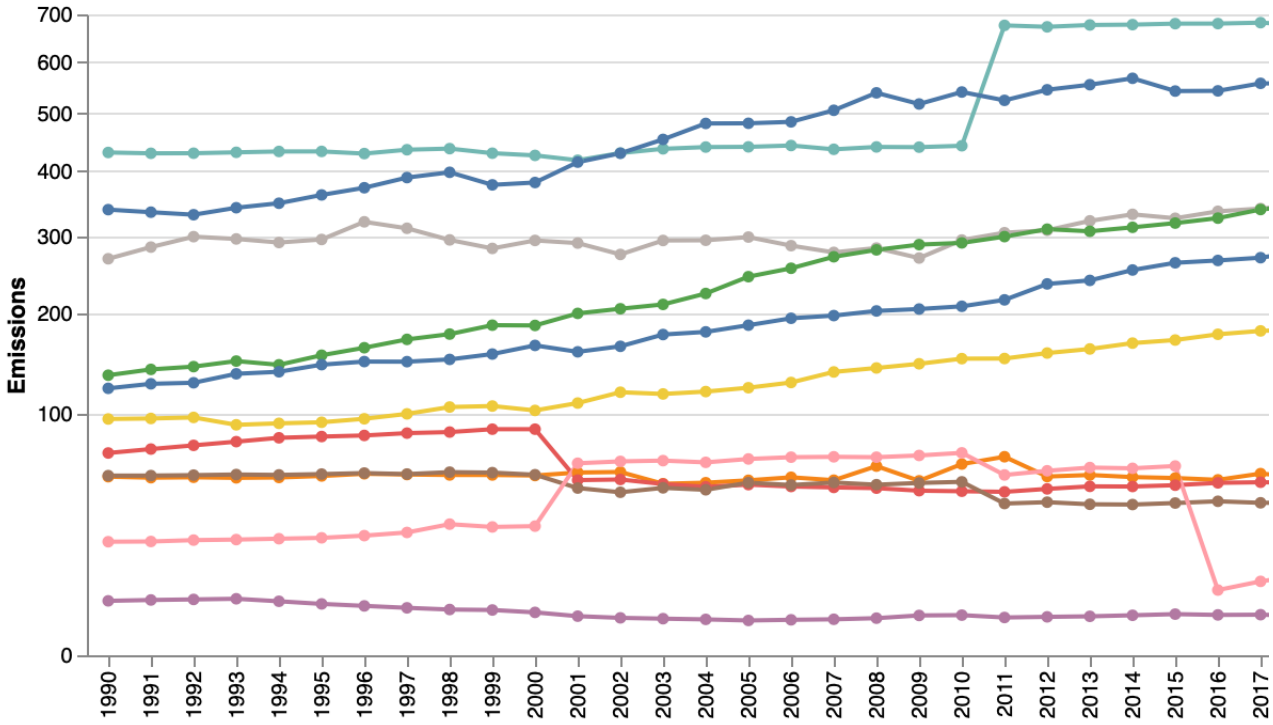
Out[15]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	2011	...	1997
87	Botswana	52.34	54.01	56.22	52.43	53.52	54.09	55.38	54.43	67.14	...	55.66
163	Gambia	2.86	2.66	2.80	2.77	2.88	2.72	2.58	2.51	2.43	...	3.83
140	Ghana	12.75	10.94	9.30	7.24	61.02	59.55	60.02	57.97	55.38	...	25.69
99	Madagascar	40.22	39.67	39.59	40.40	39.48	38.70	38.84	39.93	39.22	...	55.93
88	Côte d'Ivoire	51.51	50.74	51.00	50.65	49.29	48.52	48.61	47.15	45.50	...	84.08

5 rows × 33 columns

In [16]:
```
# look at the top one.
# try to figure out what happened to this ocuntry
# same with Ghana
# just comment whats interesting

subset_africa = africa_df[africa_df['Country'].isin(['Democratic Republic of
                                                     'Gambia', 'Ghana', 'Mac
fig_2 = alt.Chart(subset_africa.melt(id_vars='Country', var_name='Year', val
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt'), title = 'Emissions'
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_2
```

Out[16]:



A lot of the African Countries exibit a trend closer to Asia, with increasing emissions. This could be due to Africa having a lot of third world countries that rely heavily on trade and industry to grow. We see the the Democratic Republic of the Congo as a huge spike in 2010. One potential reason for this is the DRC holds the second bigest rainforest in the world and they have started a deforestation process. Furthermore, the DRC holds a lot of resources like copper, cobalt, and coltan and mining these resources requires a lot of energy rich process.

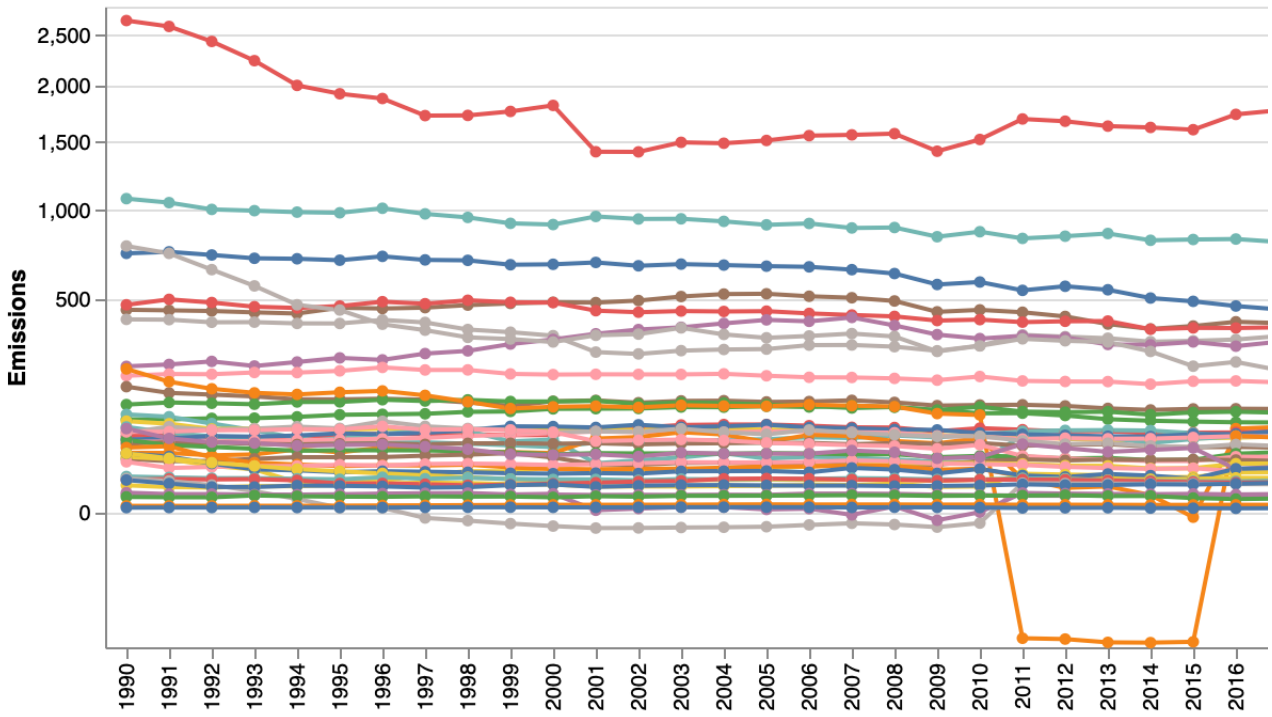
Looking at the graph, we see a drop in emissions from Cote d'Ivoire and Ghana. Cote d'Ivoire actually introduced a plan to tackle the emissions which is how they have dropped significantly from 2000 onward. Among further research from the Clean Air Coalition we found that in 2015, "Ghana committed to unconditionally lower greenhouse gas emissions by 15% relative to business as usual scenario emissions of 73.95 MtCO2e by 2030 and voluntarily pledged an additional 30% emission reduction on the condition that external support is made available to cover full cost of implementing mitigation action." (https://www.ccacoalition.org/en/partners/ghana) This is most likely the reason why Ghana's greenhouse gas emissions dropped rapidly from 2015 to 2016, then became more constant.

Europe, Oceania, and South America

For here the methods are similar to above. The reason we decided to group these countries together is because they all have similar trends separate to the other continents.

In [17]:
```
# comment whatever you want something
# that orange dip seems super interesting

europe_df = df_filtered.loc[df_filtered['Continent'] == 'Europe', :].drop(cc

fig_4 = alt.Chart(europe_df.melt(id_vars='Country', var_name='Year', value_r
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt'), title = 'Emissions'
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_4
```

Out[17]:



In [18]:

```
europe_df['Mean'] = europe_df.iloc[:,1:].mean(axis = 1)
europe_df['Difference'] = europe_df['2015'] - europe_df['2010']
europe_df.sort_values(by = ['2015'], ascending=True).head()
```

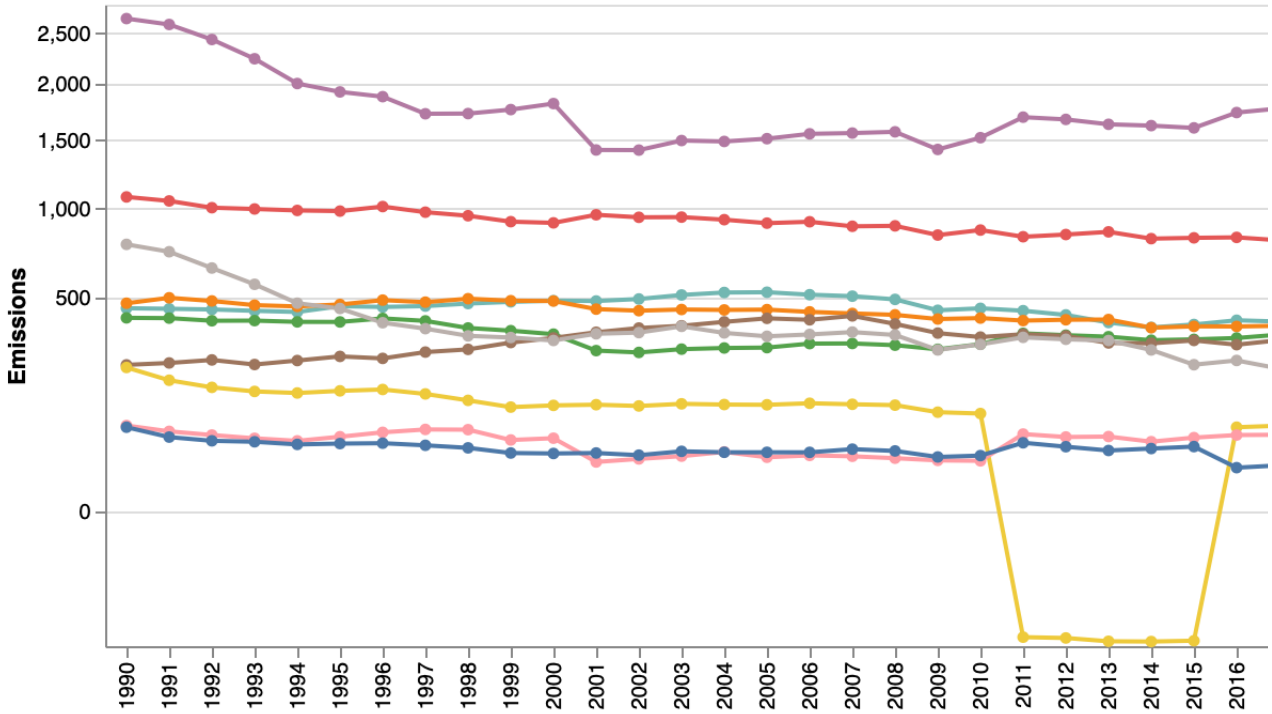
Out[18]:

	Country	2019	2018	2017	2016	2015	2014	2013	2012	2011	...
67	Romania	78.36	79.91	79.25	76.46	-184.27	-186.55	-185.86	-176.58	-174.27	...
82	Finland	58.42	62.43	60.86	63.65	-0.28	3.20	7.24	6.46	12.10	...
188	Liechtenstein	0.16	0.16	0.17	0.16	0.17	0.18	0.21	0.20	0.19	...
178	Andorra	0.63	0.62	0.59	0.59	0.58	0.57	0.59	0.60	0.60	...
169	Malta	2.13	2.04	2.02	1.88	2.18	2.86	2.84	3.17	2.99	...

5 rows × 33 columns

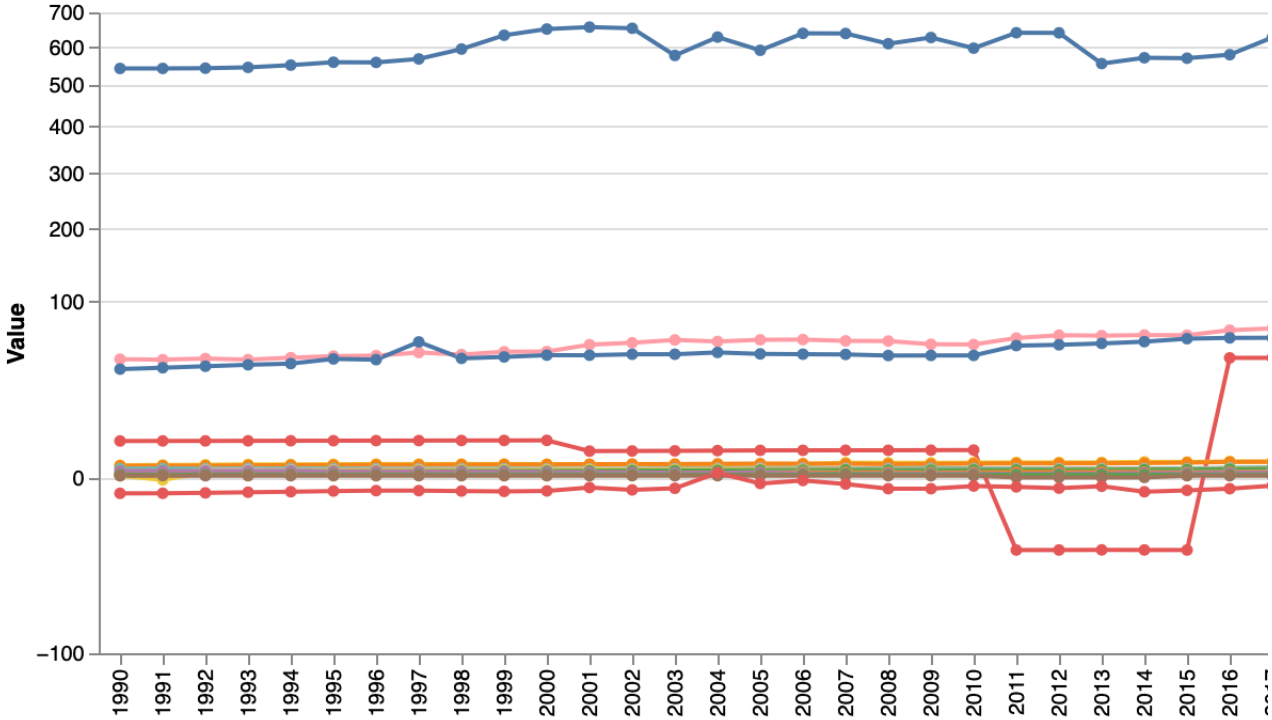
In [19]:
```
# when looking at the graphs below there are several countries that share a
# look up why this could be
subset_europe = europe_df[europe_df['Country'].isin(['Russia', 'Germany', 'F
                                                     'Ukraine', 'Bulgaria',
fig_2 = alt.Chart(subset_europe.melt(id_vars='Country', var_name='Year', val
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt'), title = 'Emissions'
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_2
```

Out[19]:



In [20]:
```
ocean_df = df_filtered.loc[df_filtered['Continent'] == 'Oceania'].drop(colum
fig_4 = alt.Chart(ocean_df.melt(id_vars='Country', var_name='Year', value_na
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'sqrt')),
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_4
```

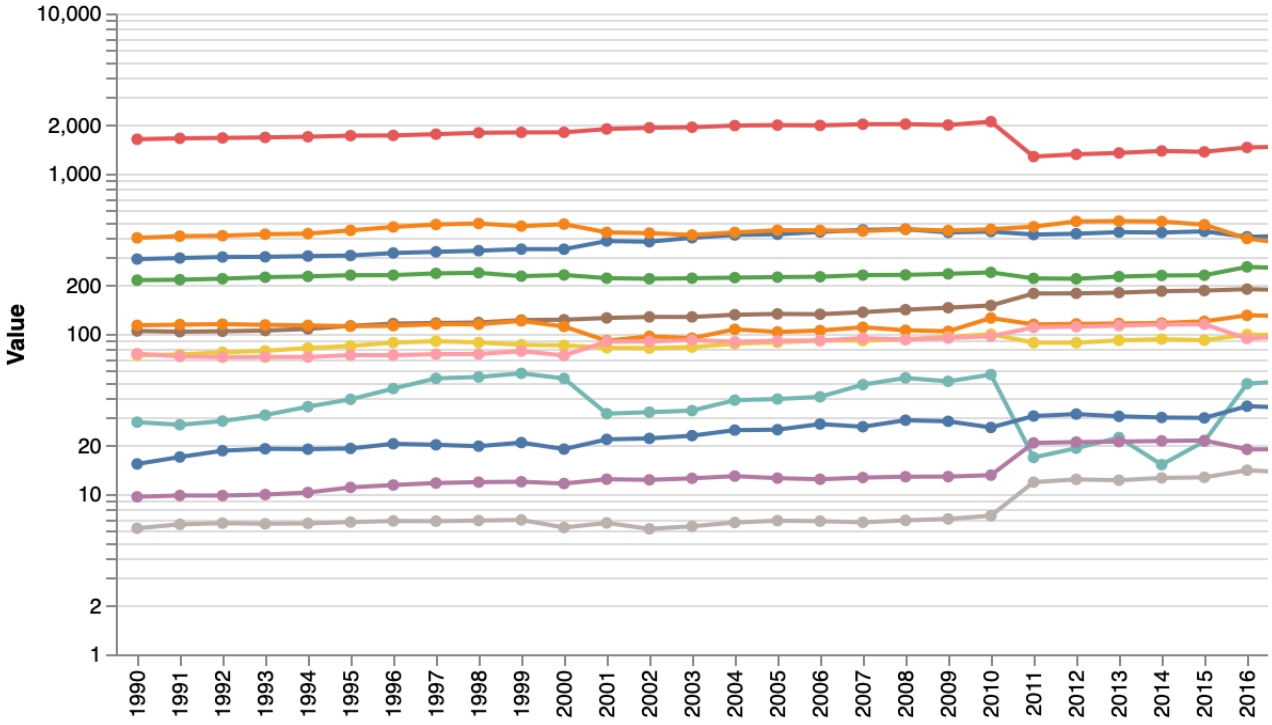
Out[20]:



In [21]:

```
south_df = df_filtered.loc[df_filtered['Continent'] == 'South America', :].c
fig_4 = alt.Chart(south_df.melt(id_vars='Country', var_name='Year', value_na
    x = alt.X('Year', title = None),
    y = alt.Y('Value', scale = alt.Scale(type = 'log')),
    color = alt.Color('Country')
).mark_line(point = True)

# display
fig_4
```

Out[21]:



Looking at all the graphs above, countries like Brazil, Australia, and Russia topped their regional chart. This makes sense since they are the most populated countries in their respective regions. What is interesting to note is Brazil saw a drop in emissions in 2010 potentially due to lower rates of deforestation. Russia meanwhile had a drop overall since 1991 which can be due to the collapse of the Soviet Union. Similar to Brazil, Australia has had a lot of deforestation which could be the reason for its fluctuation in emissions.

So many of these countries like Romania, Fiji and Chile had a drop in 2011 before going back up. Romania has a dip in greenhouse gas emissions from 2011 to 2015, then it rises back up to what it was before. This could be due to a transition in its energy sector that involved a shift towards cleaner and more sustainable energy sources, such as increased deployment of renewable energy projects like wind, solar, and hydroelectric power. This caused energy industry emissions to drop 2005 to 2019, and most of the progress probably happened from 2011 to 2015.

Summary of Findings

What we found is that countries which are industrial or petrol states have higher emissions than other countries in that region. Countries like the United States, India, China, Saudi Arabia, and the Democratic Republic of the Congo are all countries that fall under this category and as seen with the graphs they have much higher emissions compared to countries in their respective region. This is because the countries in Asia and Africa are developing and resort more to using their natural resources for growth.

Also we see that countries with large populations in their respective region tend to have higher emissions. This is seen with the US, Canada, Australia, India, China, and Russia when comparing them to other countries in their region.

There are also a lot of countries with dips in emission, Romania, Chile, and Cote d'Ivoire to name a few. This is due to legislation and policy changes that have pushed a greener environment. These countries have been actively trying to reduce their carbon foot print and we can see the results in the graph comparing them to other nations in that region.

One of the causes for variability in emissions that transcends continents is how public policy influences the CO2 emissions for each country.