project-guidelines

June 17, 2023

[1]:
```
import pandas as pd
import numpy as np
import altair as alt
alt.data_transformers.disable_max_rows()
alt.renderers.enable('mimetype')
```

[1]:
```
RendererRegistry.enable('mimetype')
```

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

# melt the data
cols_to_melt = ['Country', 'Data source', 'Sector', 'Gas', 'Unit']
melted_df = pd.melt(raw_data, id_vars=cols_to_melt, var_name='Year',␣
 ↪value_name='Emissions')

# select the top 5 countries + world
top_countries = ['World', 'China', 'United States', 'India', 'European Union␣
 ↪(27)', 'Indonesia']
tidy_data = melted_df[melted_df['Country'].isin(top_countries)]

# drop the unnecessary columns
tidy_data = tidy_data.drop(['Data source', 'Sector', 'Unit', 'Gas'], axis=1)

tidy_data
```

[2]:

	Country	Year	Emissions
0	World	2019	49758.23
1	China	2019	12055.41
2	United States	2019	5771.00
3	India	2019	3363.60
4	European Union (27)	2019	3149.57
…	…	…	…
5656	China	1990	2891.73
5657	United States	1990	5417.32
5658	India	1990	1002.56
5659	European Union (27)	1990	4187.90

```
5660            Indonesia  1990    1226.82

[180 rows x 3 columns]
```

Data description:

The dataset I chose for my final project is the ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present. This dataset provides information on greenhouse gas emissions by country over several years (1990-2019) and consist of data from various countries and regions, including the world as a whole to represent the global greenhouse gas emissions. The dataset has the following columns/variables: Country, Data source, Sector, Gas, Unit, and the years 1990-2019. The column "Country" indicates the country or region for which the emissions data is reported with the years as columns representing the emissions measurement value. The Data source column has one value: "Climate Watch", the Sector column as well: "Total including LUCF", Gas column with "All GHG", and lastly, the Unit column with the unit of measure: MtCO e. By working with this dataset, we can explore changes in greenhouse gas emissions over the years for different countries as well as analyze or compare emission levels of the countries and observe any patterns or trends in the emissions. This dataset plays a significant role in allowing us to draw deeper insight on how to track the different countries' effect on climate change and potentially reduce their carbon footprint. We will utilize the data analysis and visualization techniques learned throughout this course to gain insight and use this information to understand the global efforts to mitigate climate change.

Question of Interest:

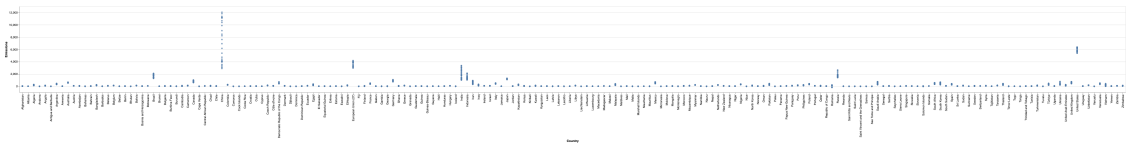
How have greenhouse gas emissions changed over time for the top five emitting countries?

A satisfactory answer might look like visualizations and analysis that clearly shows any patterns or trends in the data to provide insight on how greenhouse gas emissions might change over time for the top five emitting countries.

[31]:
```
# scatter plot aggregated by country
global_plot = melted_df[melted_df["Country"] != "World"]

glob_country = alt.Chart(global_plot).mark_circle().encode(
    x = "Country",
    y = "Emissions"
)

# display
glob_country
```

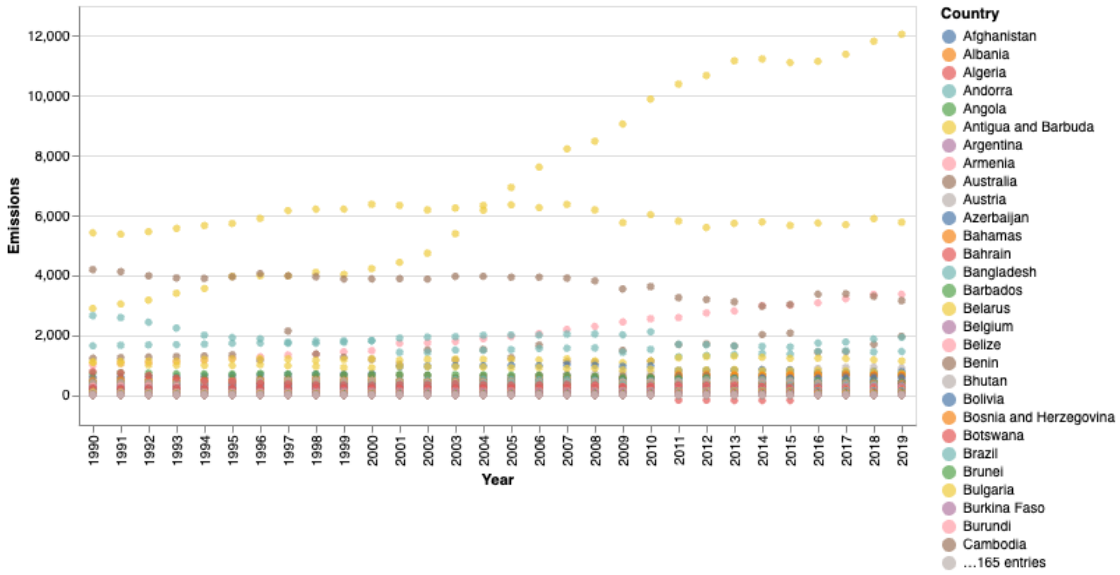
[31]:


[29]:
```
# scatter plot aggregated by year + color coded by country
global_plot = melted_df[melted_df["Country"] != "World"]

glob_country = alt.Chart(global_plot).mark_circle().encode(
    x = "Year",
    y = "Emissions",
    color = 'Country'
)

# display
glob_country
```

[29]:

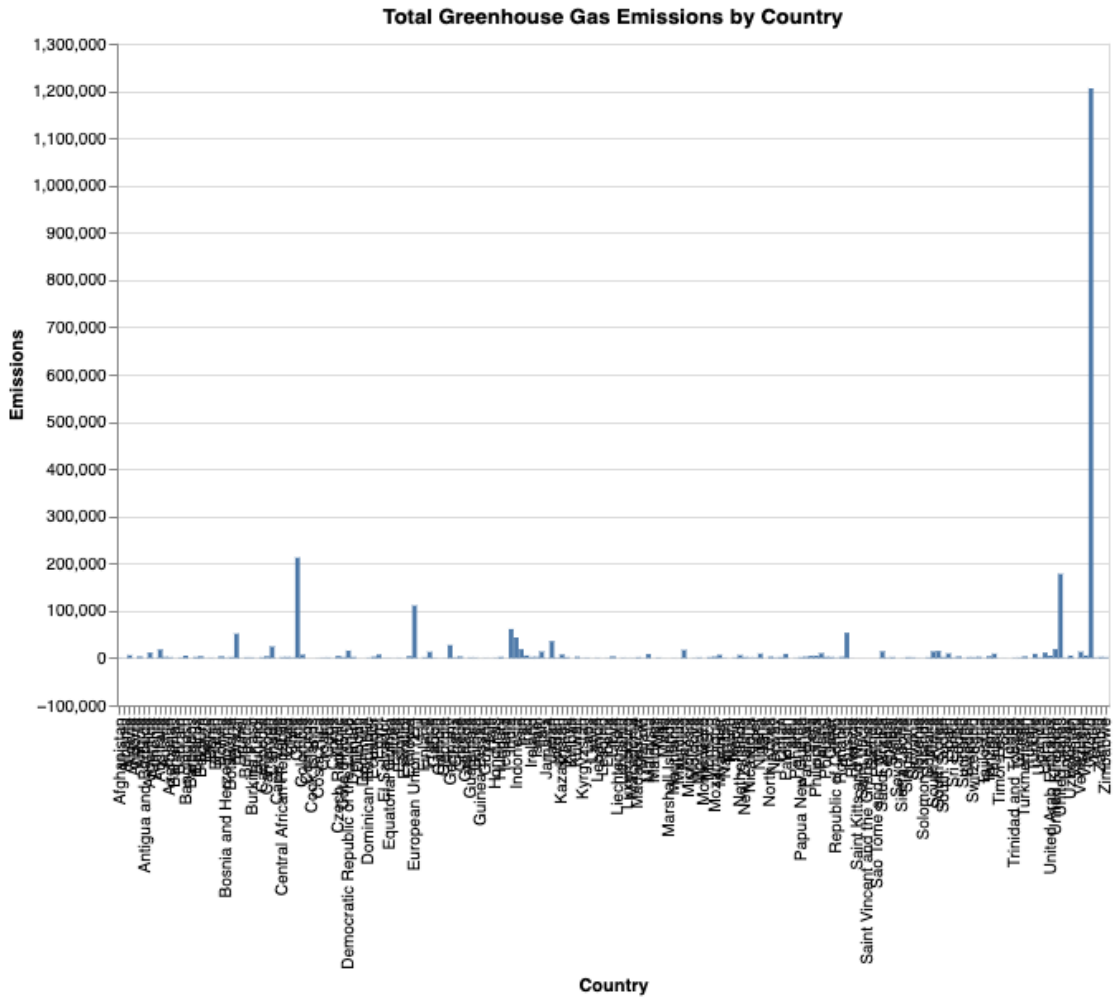


[3]:
```
country_agg = melted_df.groupby('Country').sum().reset_index()

# bar graph plot
bar_country = alt.Chart(country_agg).mark_bar().encode(
    x='Country',
    y='Emissions:Q',
).properties(
    width=600,
    height=400,
    title='Total Greenhouse Gas Emissions by Country'
)

# display
bar_country
```

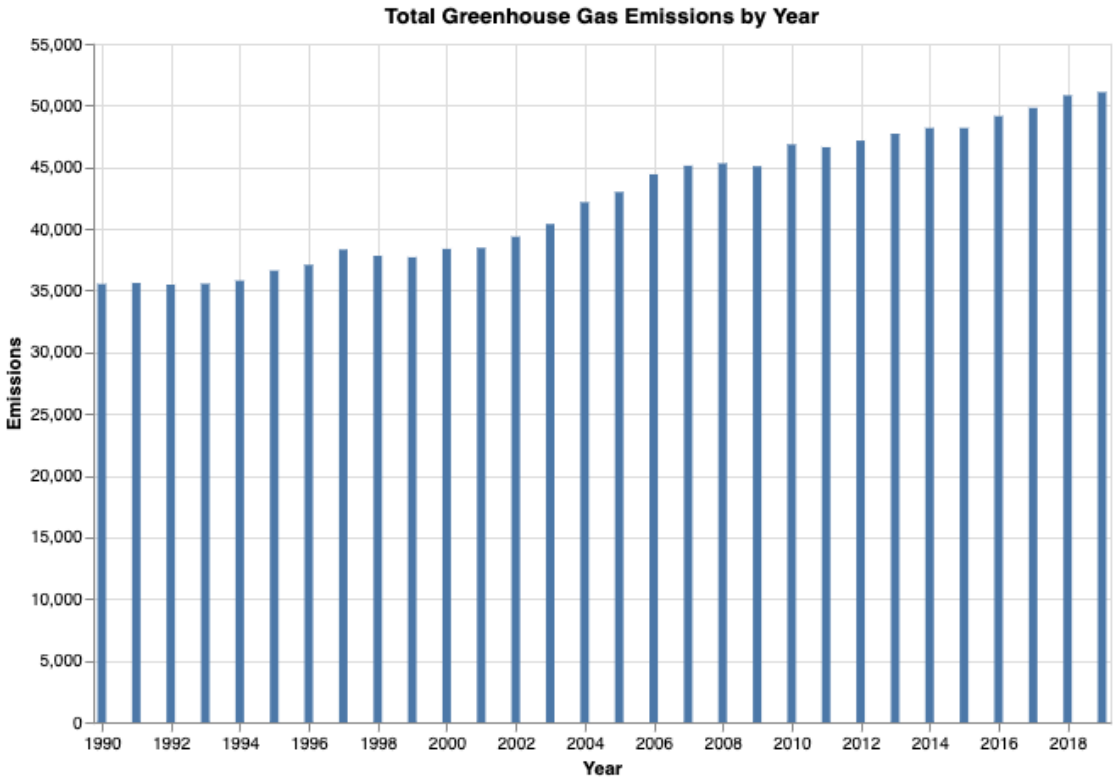
[3]:



[4]:
```
# filter world
year_agg = raw_data[raw_data['Country'] != 'World'].sum().reset_index()
year_agg.columns = ['Year', 'Emissions']

# bar graph plot
bar_year = alt.Chart(year_agg).mark_bar().encode(
    x='Year:T',
    y='Emissions:Q',
).properties(
    width=600,
    height=400,
    title='Total Greenhouse Gas Emissions by Year'
)
#display
bar_year
```

[4]:



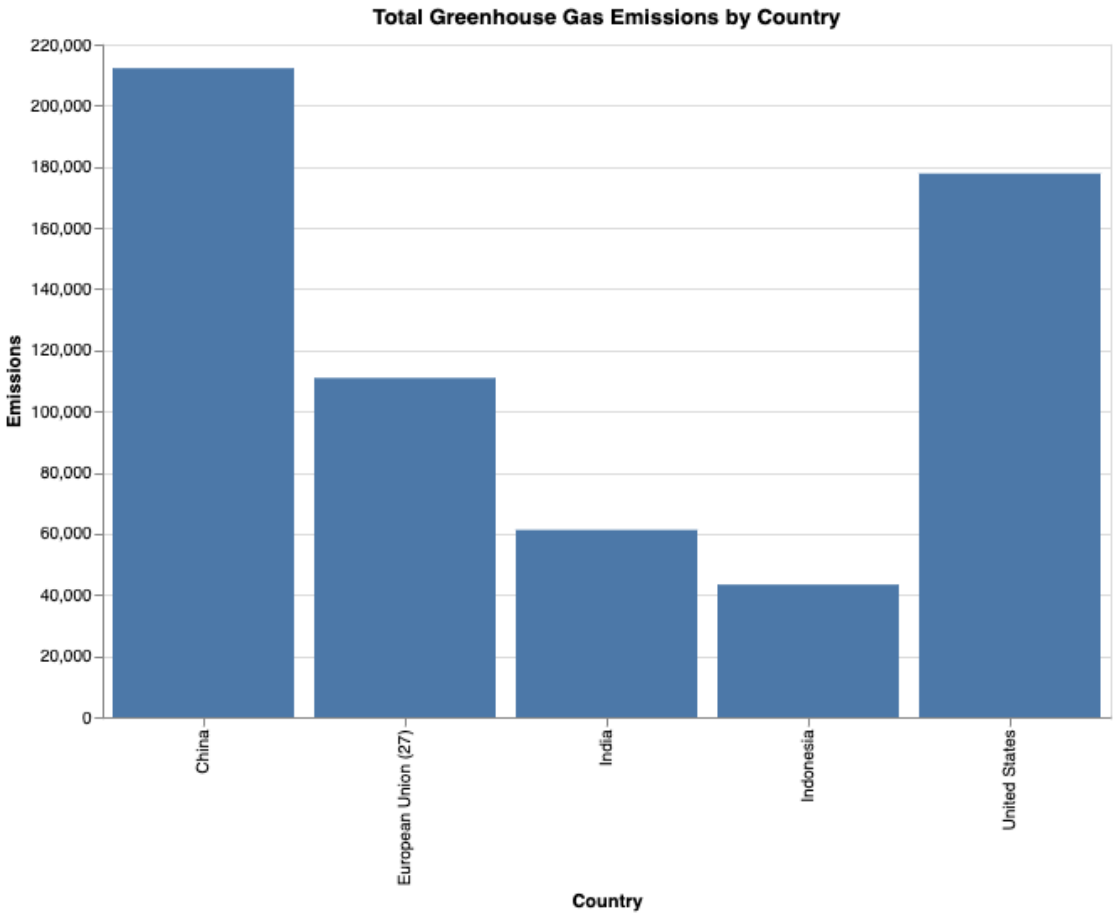
[5]:
```
country_ag = tidy_data.groupby('Country').sum().reset_index()

# filter world data
filtered_data = country_ag[country_ag['Country'] != 'World']

# bar graph aggregated by country
bar_country1 = alt.Chart(filtered_data).mark_bar().encode(
    x='Country:N',
    y='Emissions:Q',
).properties(
    width=600,
    height=400,
    title='Total Greenhouse Gas Emissions by Country'
)

# display
bar_country1
```

[5]:



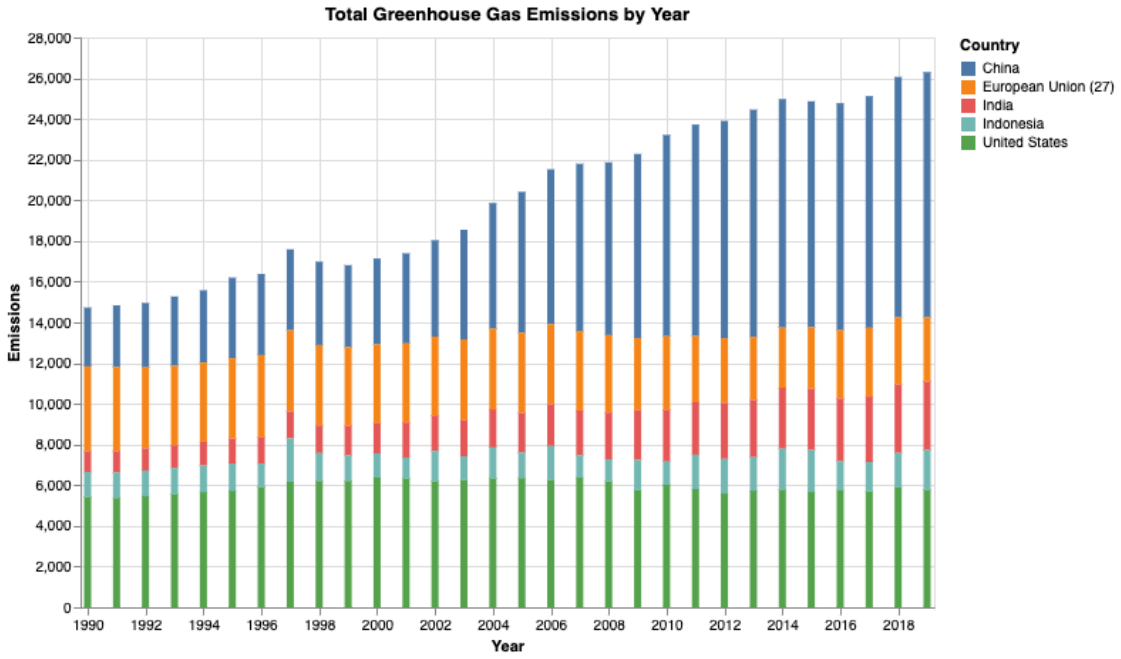
[6]:
```
# aggregating by year and country
country_ag1 = tidy_data.groupby(['Year', 'Country'])['Emissions'].sum().
 ↪reset_index()

# filter world data
filtered_data1 = country_ag1[country_ag1['Country'] != 'World']

# bar graph plot
bar_plot1 = alt.Chart(filtered_data1).mark_bar().encode(
    x='Year:T',
    y='Emissions:Q',
    color='Country:N', # color-coding by country
).properties(
    width=600,
    height=400,
    title='Total Greenhouse Gas Emissions by Year'
)
```

```
# display
bar_plot1
```

[6]:

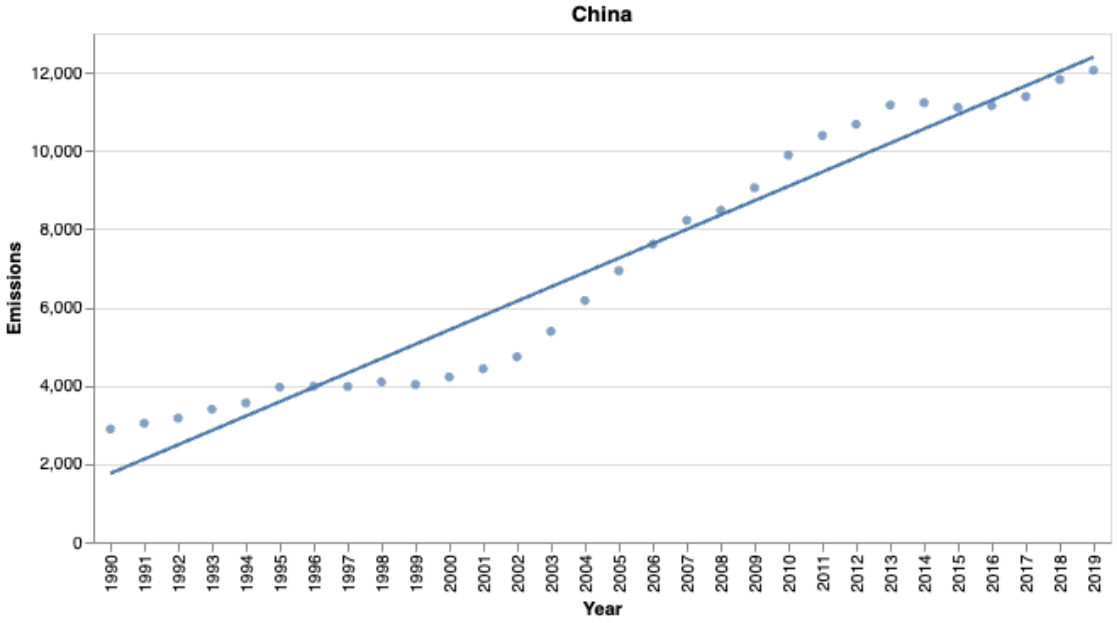


[32]:

```
# number 1: China
# plot and add trendline
china_emission = melted_df[melted_df["Country"] == "China"]
c_scatter = alt.Chart(china_emission).mark_circle().encode(
    x = "Year",
    y = "Emissions"
).properties(
    title = "China"
)

# display
c_scatter + c_scatter.transform_regression("Year", "Emissions").mark_line()
```

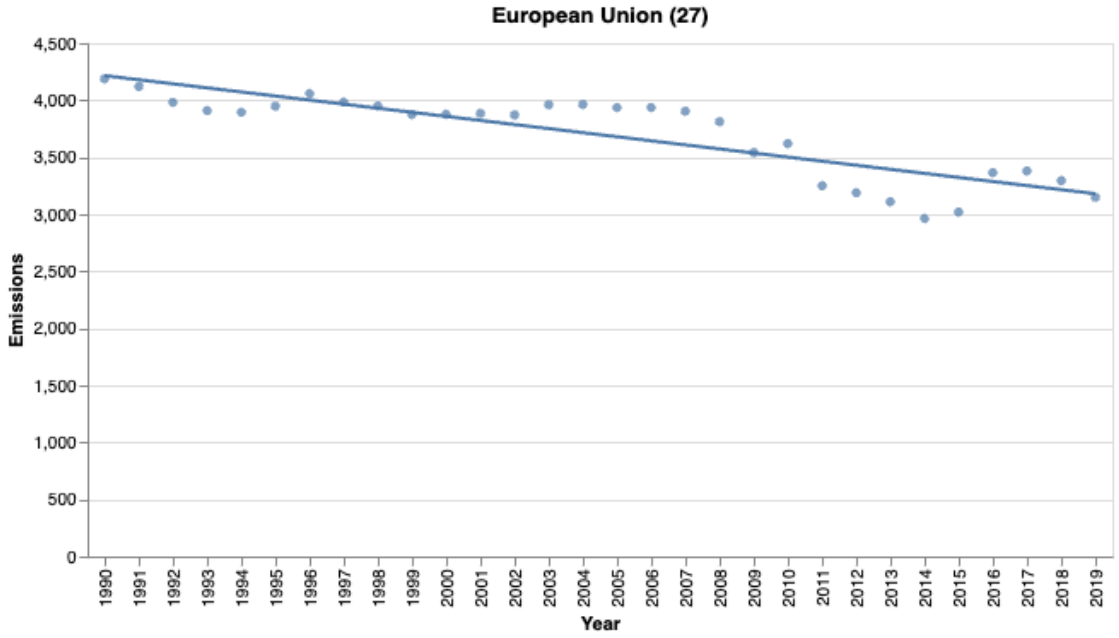
[32]:



[34]:
```
# number 2: European Union
# plot and add trendline
eu_emission = melted_df[melted_df["Country"] == "European Union (27)"]
eu_scatter = alt.Chart(eu_emission).mark_circle().encode(
    x = "Year",
    y = "Emissions"
).properties(
    title = "European Union (27)"
)

# display
eu_scatter + eu_scatter.transform_regression("Year", "Emissions").mark_line()
```

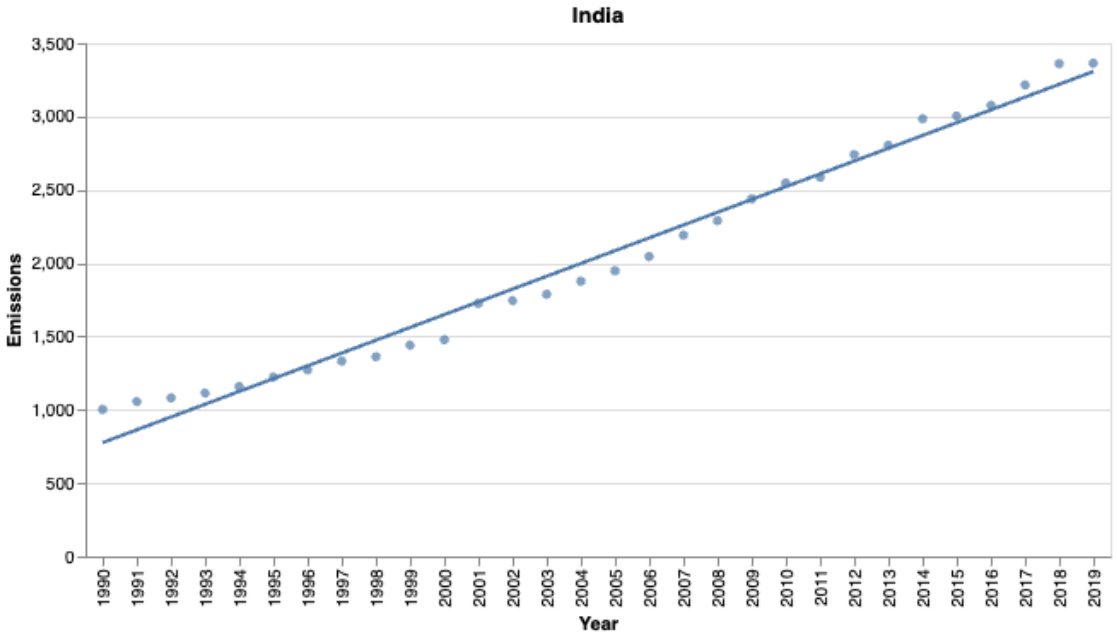
[34]:



[35]:
```
# number 3: India
# plot and add trendline
india_emission = melted_df[melted_df["Country"] == "India"]
in_scatter = alt.Chart(india_emission).mark_circle().encode(
    x = "Year",
    y = "Emissions"
).properties(
    title = "India"
)

# display
in_scatter + in_scatter.transform_regression("Year", "Emissions").mark_line()
```

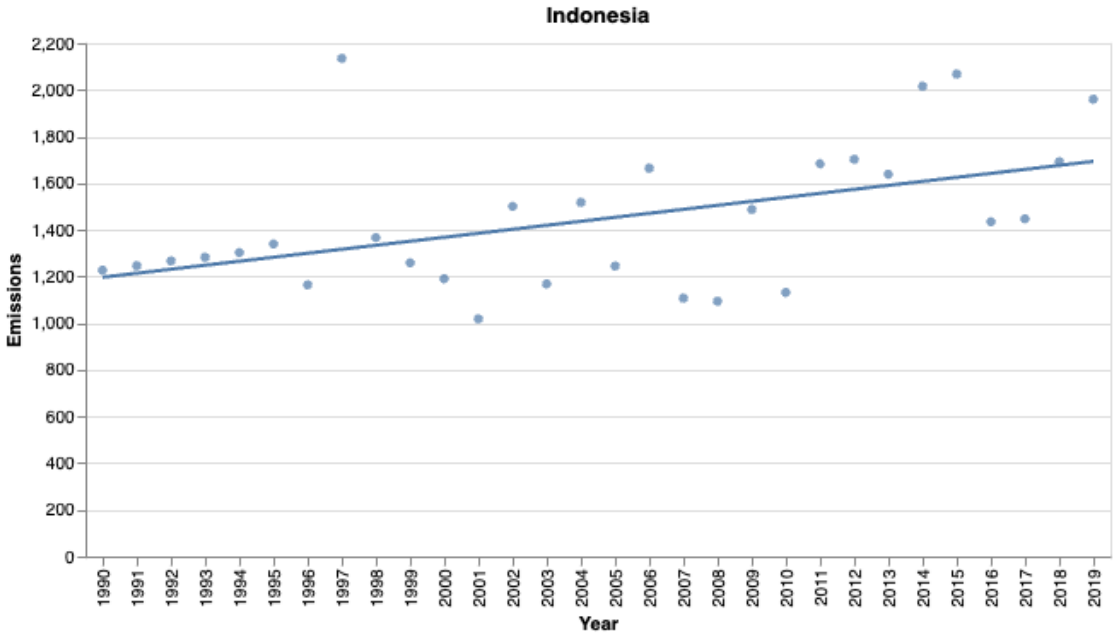
[35]:



[37]:
```
# number 4: Indonesia
# plot and add trendline
indonesia_emission = melted_df[melted_df["Country"] == "Indonesia"]
indo_scatter = alt.Chart(indonesia_emission).mark_circle().encode(
    x = "Year",
    y = "Emissions"
).properties(
    title='Indonesia'
)

# display
indo_scatter + indo_scatter.transform_regression("Year", "Emissions").
 ↪mark_line()
```

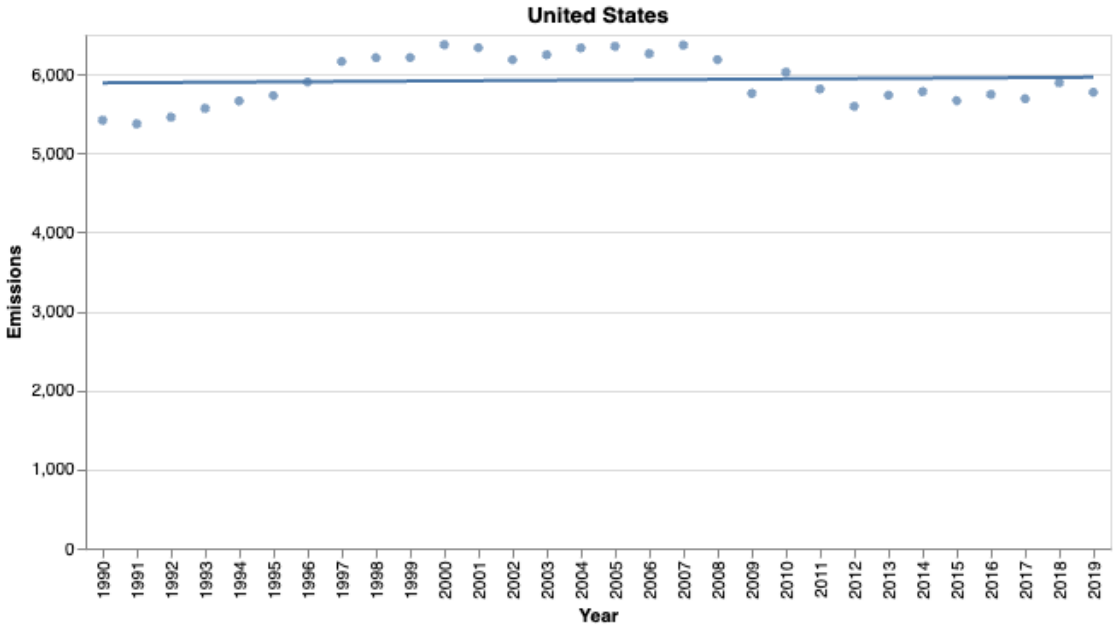
[37]:



[38]:
```
# number 5: United States
# plot and add trendline
us_emission = melted_df[melted_df["Country"] == "United States"]
us_scatter = alt.Chart(us_emission).mark_circle().encode(
    x = "Year",
    y = "Emissions"
).properties(
    title='United States'
)

# display
us_scatter + us_scatter.transform_regression("Year", "Emissions").mark_line()
```

[38]:



[49]:
```
# filter data to only include top 5
top_countries = ['China', 'United States', 'India', 'European Union (27)',␣
 ↪'Indonesia']
top_countries_data = tidy_data[tidy_data['Country'].isin(top_countries)]

# group by country and year
group_data = top_countries_data.groupby(['Country', 'Year']).sum().reset_index()

# calculate emissions for the first year and most recent year for top 5 country
first_year_emissions = group_data.groupby('Country')['Emissions'].first()
most_recent_year_emissions = group_data.groupby('Country')['Emissions'].last()

# calculate the difference
emission_differences = most_recent_year_emissions - first_year_emissions

# create new dataframe
difference_df = pd.DataFrame({
    'Country': emission_differences.index,
    'Emission_Difference': emission_differences.values
})

# sort in descending order
difference_df_sorted = difference_df.sort_values('Emission_Difference',␣
 ↪ascending=False)

# display
```

```
difference_df_sorted
```

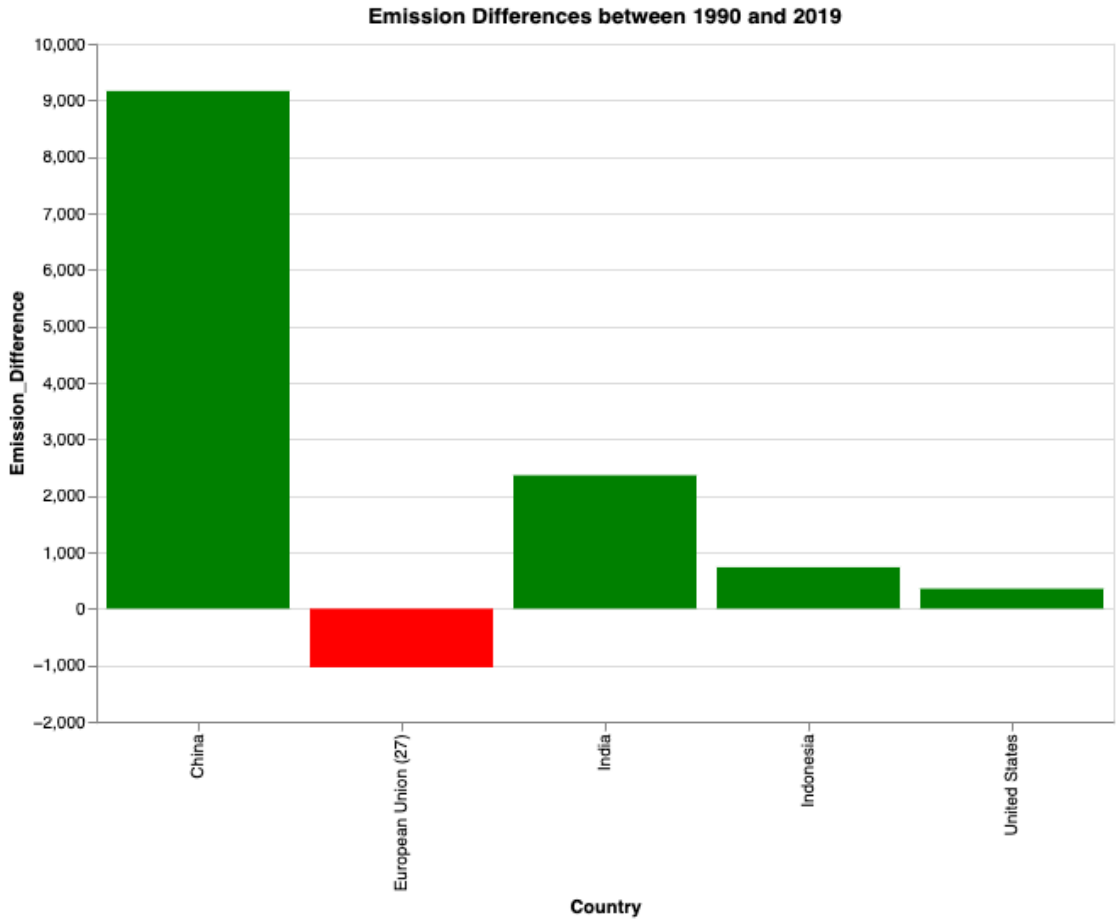
[49]:

	Country	Emission_Difference
0	China	9163.68
2	India	2361.04
3	Indonesia	732.89
4	United States	353.68
1	European Union (27)	-1038.33

[53]:
```
# bar plot
bar_chart = alt.Chart(difference_df_sorted).mark_bar().encode(
    x='Country',
    y='Emission_Difference',
    color=alt.condition(
        alt.datum.Emission_Difference >= 0,
        alt.value('green'),  # positive differences
        alt.value('red')  # negative differences
    )
).properties(
    width=600,
    height=400,
    title='Emission Differences between 1990 and 2019'
)

# display
bar_chart
```

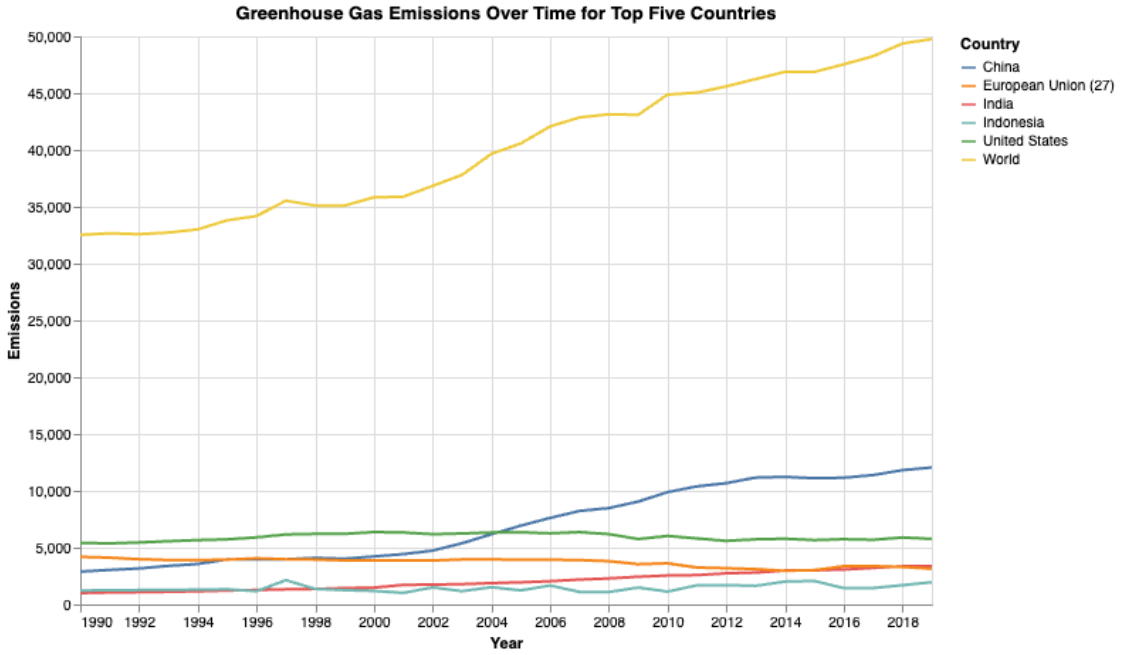
[53]:



[7]:
```
# create a line chart
line_chart = alt.Chart(tidy_data).mark_line().encode(
    x='Year:T',
    y='Emissions:Q',
    color='Country:N',
).properties(
    width=600,
    height=400,
    title='Greenhouse Gas Emissions Over Time for Top Five Countries'
).configure_axis(
    grid=True
).encode(
    color=alt.Color('Country:N', legend=alt.Legend(title='Country'))
)

# display
line_chart
```

[7]:



Data Analysis:

To begin with, I tidied the dataset using the pd.melt() function to reshape it from wide to long format. I then proceeded to create a new dataframe object containing only the top five emitting countries: China, European Union, India, Indonesia, United States, and the world. I proceeded to drop the unnecessary columns, such as Data source, Sector, Gas, and Unit to leave behind just Country, Year, and Emissions columns/variables in the final tidied dataset.

[39]:
```
tidy_data.head()
```

[39]:

	Country	Year	Emissions
0	World	2019	49758.23
1	China	2019	12055.41
2	United States	2019	5771.00
3	India	2019	3363.60
4	European Union (27)	2019	3149.57

In terms of visualization, I created various types of visualization for both the overall data and the tidied dataset containing only the top five emission countries, excluding the world entry.

I started off with creating a scatterplot for the overall data to get a general idea with what we're working with here. Immediately, from the scatterplots, we can see that most countries have a rather small amount of greenhouse gas emissions with China as the leading country in terms of emission. In addition to this, we can see that the global greenhouse gas emissions have a positive trend, consistently increasing over time.

Upon looking at our barplots and scatterplots aggregated by both year and country, it is evident that China is the number one producer of greenhouse gas emissions globally, followed by the United

States, European Union, India, and Indonesia. The rest of the countries seem to be ranked around the same with only a few countries contributing a significant amount of pollution. While these top 5 countries are reasonably the largest greenhouse gas produces with their large population taken into account, it is important to note that we do not take population into account in this project. It is also interesting to note that there are no patterns or sudden changes in trend and the greenhouse gas emissions remain relatively consistent/steadyily increasing throughout the years.

When taking a closer look at the top five emission countries individually, there were interesting trends I noticed. We have China, India, and Indonesia with positive trends, increasing in emissions over time. However, I find it interesting that the European Union has a negative trend, meaning their greenhouse gas emissions have reduced over time. Perhaps this is explainable by various environmental activism laws and organizations that have formed throughout the years, working towards reducting greenhouse gas emissions and their carbon footprints as a whole country. Another thing that caught my eyes was that despite being one of the largest greenhouse gas producers globally, the United States has remained fairly consistent with neither an increasing or decreasing trend over time. Our calculations of differences support these claims as China has the largest difference between 2019 and 1990 with the United States having a meager difference of 353.68 and the European Union having a negative difference value as their overall greenhouse gas emissions have significantly decreased over time.

Summary of findings:

- overall steady increase in trend for the top 5 emission countries in terms of greenhouse gas emissions
- China is the number one greenhouse gas producer; continues to increase in production over time
- United States is a major contributor of greenhouse gas emissions, but has remained relatively consistent in its emissions without any significant increase or decrease
- the European Union has decreased its production of greenhouse gas throughout the years despite being one of the largest producers
- India and Indonesia follows China's trend and continues to increase steadily while being one of the top 5 producing countries
- other countries in the dataset are not significant producers of greenhouse gas

1 Course project guidelines

Your assignment for the course project is to formulate and answer a question of your choosing based on one of the following datasets:

1. ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present
2. World Happiness Report 2023: indices related to happiness and wellbeing by country 2008-present
3. Any dataset from the class assignments or mini projects

A good question is one that you want to answer. It should be a question with contextual meaning, not a purely technical matter. It should be clear enough to answer, but not so specific or narrow that your analysis is a single line of code. It should require you to do some nontrivial exploratory analysis, descriptive analysis, and possibly some statistical modeling. You aren't required to use any specific methods, but it should take a bit of work to answer the question. There may be multiple answers or approaches to contrast based on different ways of interpreting the question or

different ways of analyzing the data. If your question is answerable in under 15 minutes, or your answer only takes a few sentences to explain, the question probably isn't nuanced enough.

1.1 Deliverable

Prepare and submit a jupyter notebook that summarizes your work. Your notebook should contain the following sections/contents:

- **Data description**: write up a short summary of the dataset you chose to work with following the conventions introduced in previous assignments. Cover the sampling if applicable and data semantics, but focus on providing high-level context and not technical details; don't report preprocessing steps or describe tabular layouts, etc.
- **Question of interest**: motivate and formulate your question; explain what a satisfactory answer might look like.
- **Data analysis**: provide a walkthrough with commentary of the steps you took to investigate and answer the question. This section can and should include code cells and text cells, but you should try to focus on presenting the analysis clearly by organizing cells according to the high-level steps in your analysis so that it is easy to skim. For example, if you fit a regression model, include formulating the explanatory variable matrix and response, fitting the model, extracting coefficients, and perhaps even visualization all in one cell; don't separate these into 5-6 substeps.
- **Summary of findings**: answer your question by interpreting the results of your analysis, referring back as appropriate. This can be a short paragraph or a bulleted list.

1.2 Evaluation

Your work will be evaluated on the following criteria:

1. Thoughtfulness: does your question reflect some thoughtful consideration of the dataset and its nuances, or is it more superficial?
2. Thoroughness: is your analysis an end-to-end exploration, or are there a lot of loose ends or unexplained choices?
3. Mistakes or oversights: is your work free from obvious errors or omissions, or are there mistakes and things you've overlooked?
4. Clarity of write-up: is your report well-organized with commented codes and clear writing, or does it require substantial effort to follow?

[ ]: