**Notes and next Steps from Peleke 11/14**

Regarding e.g., countries in Africa w/ confusing and potentially overlapping category information, try to recombine into a single category (e.g., simply "Sub-Saharan Africa"), or omit the data points for this particular project

\*Baseline Exploratory Steps\*

```# Step 1: Read CSV

df = pd.read\_csv('co2\_emissions.csv')

# Step 2: Inspect

## Direct overview of which columns you actually have, and roughly what the data looks like

df.head()

## Describe

df.describe()

## Create a scatter matrix

df.corr()

## Create a series of plots

### Boxplot: df['co2\_emissions'].plot(kind='box')

df.box()

### Start w/ Histogram: df['co2\_emissions'].hist()

df.hist()

```

(edited)

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html

`Carbon\_Dioxide\_Kilotons` (edited)

\*Baseline Analytical Steps\*

1. \*Step 1\*: Create a list of hypothesis/questions to answer.

Hypothesis: "CO2 emissions are a factor in global warming."

Thus, we need to:

1. Search for correlations between C02 emission and temperature data

2. Search for corrrelations between C02 emissions and \_proxies\_ for temperature data (e.g., humidity, etc.)

To answer \*1\*, try looking at aggregate \_annual\_ C02 emissions and \_annual\_ temperature data. N.b., I would \_omit\_ regional data at this stage. To do this, we'll need a column w/ total c02 emissions by year; and also, a column w/ "temperature data" (e.g., a column tracking some sort of indicator for temperature over the years)

\*Regarding Emissions vs Temp.\*

\*Line Plot\* of both temperatures and emissions over time

\*Correlation coefficient\*: In your c02 emissions data, try grouping by year, and then summing the C02 emissions column. Then, find your annual mean temperature data, and create a new data frame w/ these two columns. Next, look for correlations using `df.corr`.

\*Lag Plot\*. Lag plots reveal whether data is autocorrelated (i.e., cyclical/periodic). In this case, I'd create a lag plot of your temperature data (i.e., any of your indicator columns.

If your lag plot suggests autocorrelation, the next step is to do a \*linear regression\* on the lagged data. (edited)

]

http://www.itl.nist.gov/div898/handbook/eda/section3/lagplot.htm

https://docs.google.com/forms/d/e/1FAIpQLSdb4ejjbqoqKO-Q4k7zeO\_xwykwB0dxYLWYm1mX5Ik45MzEeg/viewform

Here's the Pandas documentation on lag plots. They're \_very\_ easy to create (just two lines of code, essentially), so they're powerful tools to be aware of: https://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-lag