After the search for data sets, the next step is data cleansing and exploration. We used variety of tools to do this like Excel, Pandas, Jupyter Notebooks, Matplotlib, and other Python libraries. One of the key methods to learn about data is to visually inspect it. Excel is a good way to start depending on the size of the data set and our data sets were small enough to work well in Excel. We sorted and filtered the data to get an idea of the distribution of records by year and country. Pandas and Jupyter Notebook were used to clean up and prepare the data for preliminary charting. Once the data was cleaned and saved into clean csv files, we again loaded the data using Pandas into Jupyter notebooks, and started creating charts like lag plots, histograms, density plots, line plots, and scatter plots. We also used a Pandas correlation function, corr() to show what correlation exists between our temperature and CO2 data. All this is done to understand and be able to effectively chart data for presentation to our audience. The main plot on this page is a scatter plot of CO2 data with a regression line fit to the data. What we are looking for are direction of the line, the distribution of data points on each side of the line, how close to the line and outliers which are data points way off the line. In this case, for the USA, we have a positive upwards trending regression line that follows our data increasing in CO2 metric tons as the years move forward thru time. This is what we expected so that is a good plot that tells us we are on the right track and should continue our exploration.

<br>

<br>

Clicking on the smaller plot on this page will bring up a page to compare the various exploratory plots we used in this part of the process. We are showing only a small fraction of the charts we completed, and the rest of the charts are available in our Jupyter Notebooks.

This USA 1878 - 2012 temperature lag plot shows the data has structure which is excellent because it is temperature and time related so it shouldn't be randomly distributed on plot.

This France 1850 - 2012 temperature lag plot is very similar to the USA one. It has a few more years of data points so it has a little bit better structure but again the important thing is that the data should be relatively structured.

This USA 1878 - 2012 histogram has frequency up the y axis and CO2 metric tons across the x axis, and it shows the distribution of CO2 amounts. No surprises here. The data has more frequencies of lower CO2 amounts which makes sense in that the amount of CO2 output increases over time with the increase use of fossil fuels.

This France 1850 - 2012 histogram is similar to the USA one, and again no surprises it follows the same pattern of distribution and frequency as the USA chart with explainable differences due to size of population.

This image is from a Jupyter notebook showing the corr() function with inputs of USA temperature and CO2 data. The numbers to look at are where the columns and rows meet for the data points you want to compare. The data points are temperature (F) and CO2, and we have a value of 0.623459. Any correlation value greater than 0.5 or -0.5 is a good correlation number.

This image is from a Jupyter notebook showing the corr() function with inputs of France temperature and CO2 data. The data points are temperature (F) and CO2, and we have a value of 0.532010. This shows a positive correlation value greater than 0.5, so the two sets of numbers correlate in a positive direction as we hope and expect.