Temperature, Precipitation and Emission Statistics

Jericho Timbol

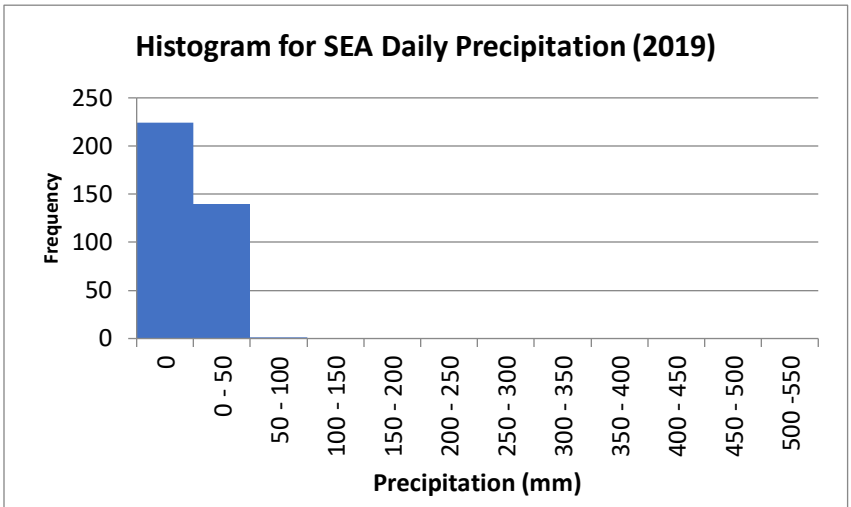
07/10/23

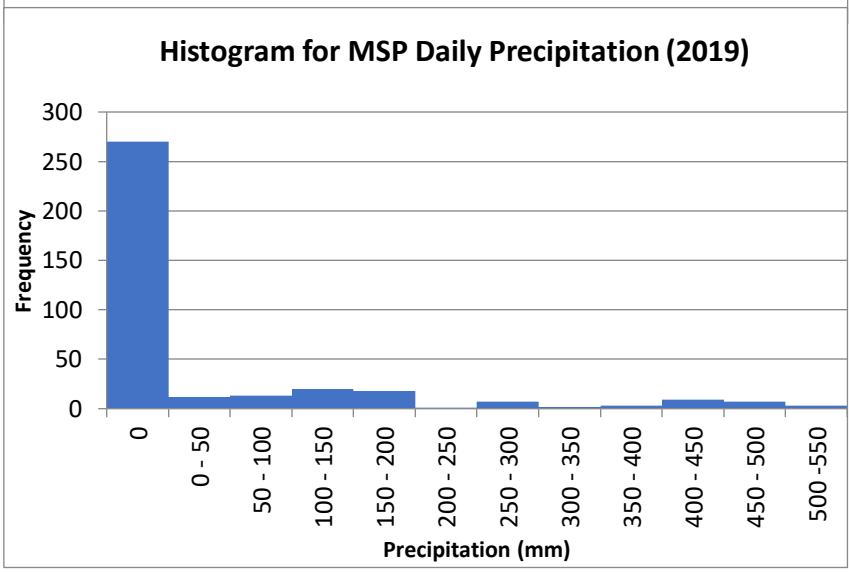
Questions

a) We have calculated the descriptive statistics of our data set and visualized our data set with histograms and box and whisker plots. Now please compare weather data between Seattle and Minneapolis based on your analysis. What can you say about the precipitation and temperature patterns for the two stations?

- Compare precipitation patterns by describing its “central tendency” and “dispersion/spread” between the two stations.

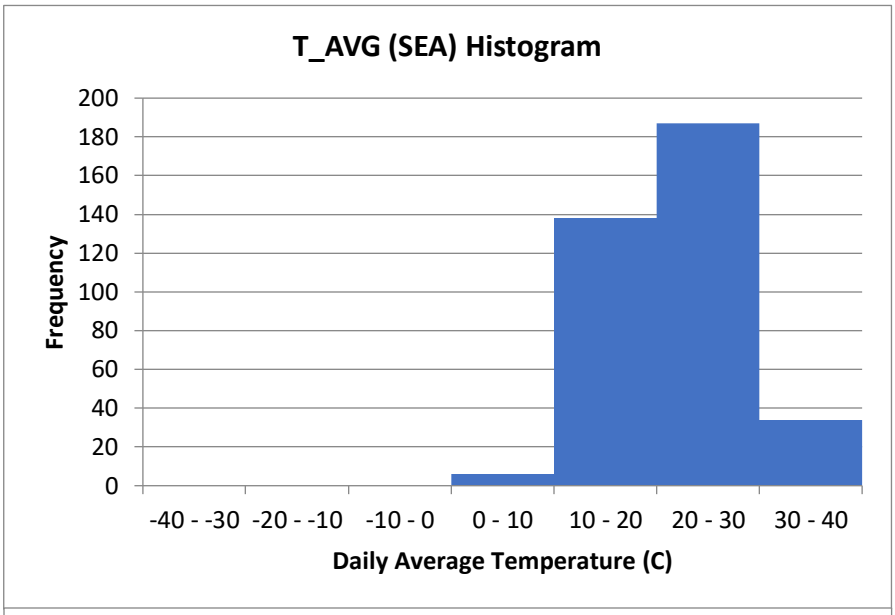
In terms of similarity, both of these visuals show that both Seattle and Minneapolis have a right skew. We can see this in the grouping of data based on our precipitation bins, with many of the data points for both SEA and MSP having more data points and values in the lower (left) end of the visual. This means that both of them strongly tend to have precipitation values closer to 0 in terms of mean, medians and modes. MSP has a wider spread with a much higher range. MSP has values for every bin whilst Seattle has values for the first 3.

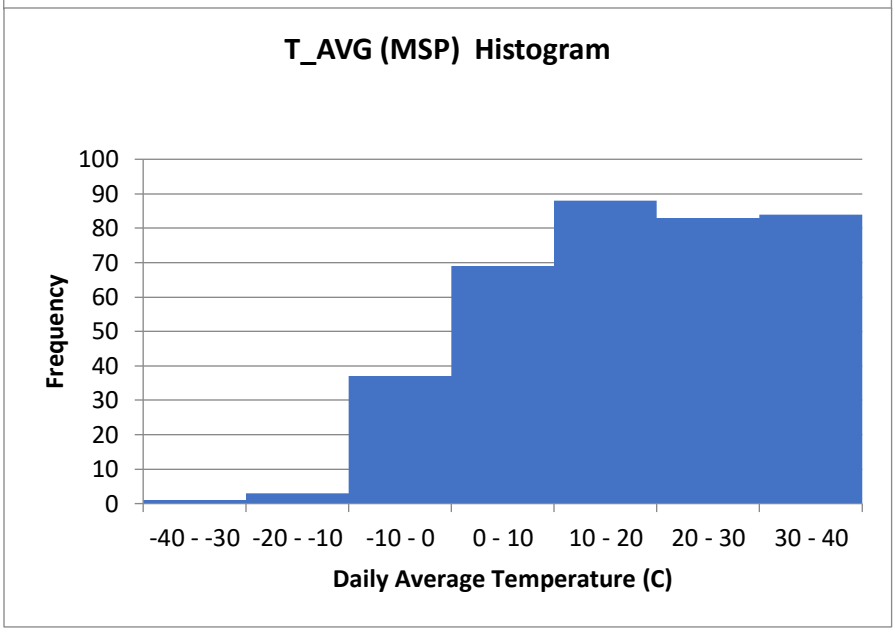




- Compare temperature patterns by describing its “central tendency” and “dispersion/spread” between the two locations.

In terms of similarity, both of these visuals show that both Seattle and Minneapolis have a left skew. We can see this in the grouping of data based on our temperature bins, with many of the data points for both SEA and MSP having more data points and values in the higher (right) end of the visual. This means that both of them strongly tend to have temperatures values closer to the high end of the histogram. There is a strong tendency for both time series to contain data between the 10-20 through 30 – 40 range. MSP has a greater distribution of points with the minimum being lower as well as a bigger left tail for the low-end values. MSP seems to be more stable in temperatures with gradual changes and minimal extremities. This is in contrast with the static cold nature of Seattle with few extreme values in the high end and little low values on the low end.





Application – Present Descriptive Statistics for a Data Set

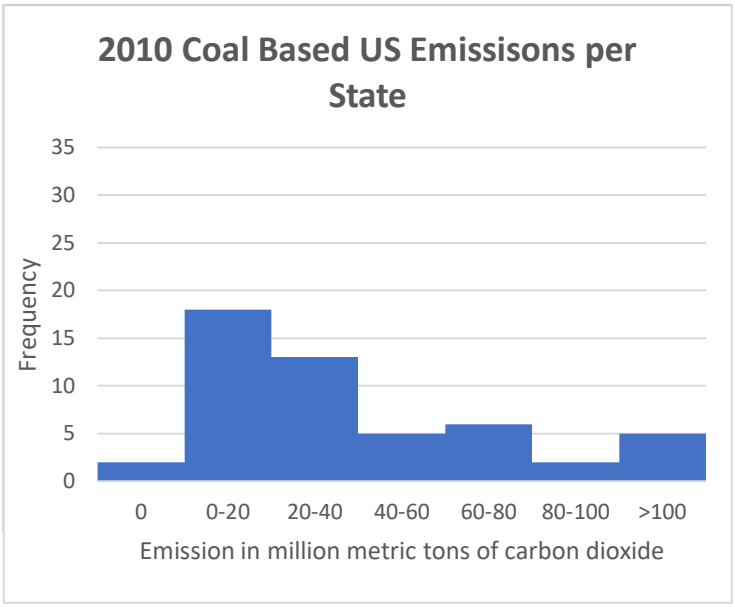
Data set: Coal energy-related carbon dioxide emissions:
https://www.eia.gov/environment/emissions/state/

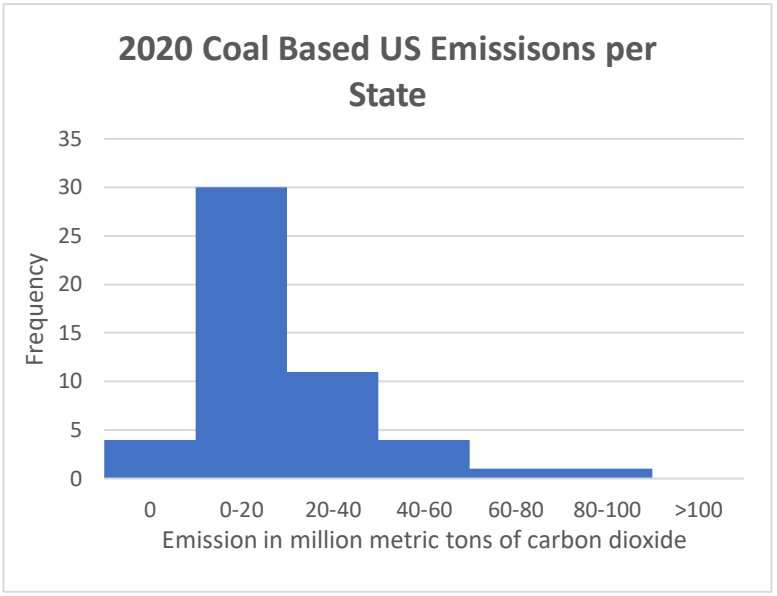
This dataset was intriguing to me in the sense that I was looking to explore more data in an environmental context. That being said, emissions were an area I found to be rather comparable in regards to different series. While this data set includes state data, the numeric values are able to be summarized through statistics and distributions and values can be compared between the years. It is interesting to see the improvements we made as a country as well in relation to coal-based fuel sources.

b) A table of descriptive statistics **(5 points)**

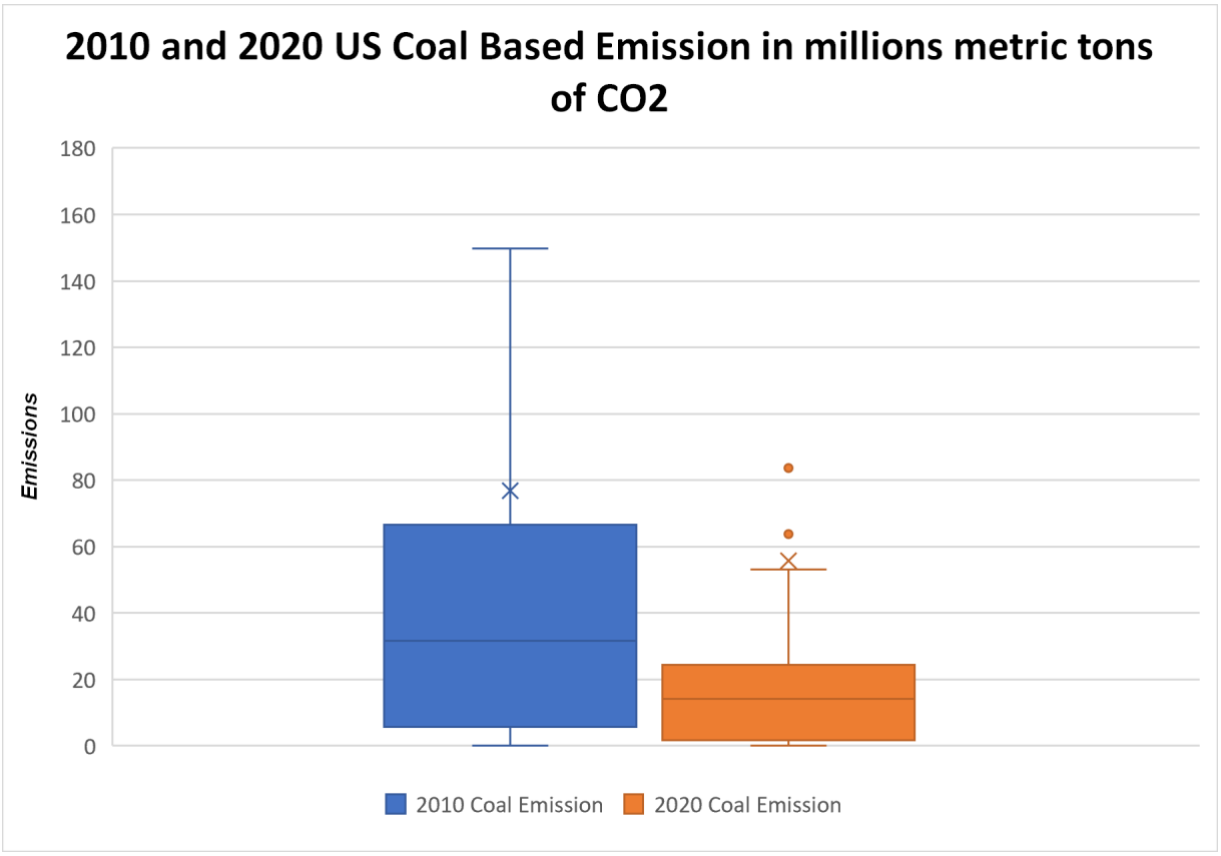
2010 Emissions		*2020 Emissions*	
Summary Stats		**Summary Stats**	
Mean	38.948	Mean	17.21421
Standard Error	5.499427	Standard Error	2.689996
Median	30.11138	Median	13.30958
Mode	0	Mode	0
Standard Deviation	39.27376	Standard Deviation	19.21042
Sample Variance	1542.429	Sample Variance	369.0401
Kurtosis	0.901191	Kurtosis	2.006835
Skewness	1.22713	Skewness	1.455869
Range	149.7659	Range	83.5955
Minimum	0	Minimum	0
Maximum	149.7659	Maximum	83.5955
Sum	1986.348	Sum	877.9245
Count	51	Count	51

c) A histogram for each data series.





d) A box and whisker plot with two data series



e) A summary paragraph describing the two data series – how they compare in terms of central tendency and dispersion/spread. Can you hypothesize a reason to explain the differences?

When looking at the distribution of the data, we see a wider spread in central tendency in 2010. There is a normal distribution in the data set with a slight right skew. The range in 2010 is far greater showing 5 values in the highest bin of over 100. Compared to 2020, this year has no values in the highest bins. Similar to 2010, the 2020 emissions distribution shows a more prominent right skew. There are much more values lying in the lower end of coal-based emissions. While both datasets contain the most values in bins 0-20, and 20-40, the tendency for higher values in the distribution for 2010 make for a year with far higher emissions. To put it into perspective the 2010 mean and median emissions were 38.95 and 30.11. This is much higher than the means and medians in 2020 at 17.21 and 13.31. I speculate that this trend to lower **Coal** based CO2 emissions is attributed to our shift towards different forms of energy and fuel. Coal is long documented to produce harmful byproducts; we have actively been shifting away from these forms of energy in America as we explore a more sustainable future. Efficiency and improvement in technologies using coal-based fuels could be attributed to the lower emissions over the years.