# GEOG/EOS 230

Tutorial 2: Stats and plotting

In this tutorial we have the following objectives:

* Practice using Generative AI.
* Practice retrieving statistics from a dataset.
* Prepare summary plots and tables.
* Start thinking about data distributions.
* Apply linear regression in practice.

**Gemini (Generative AI)**

It has only been four years since the emergence of Chat GPT in 2021, but it has rapidly evolved and is now being used in most IT workplaces - I often hear stories from my friends in IT of how AI is being used. In a few years when you enter the workforce this trend will have evolved even further.

In November 2024, Gemini was integrated into the Google Collaboratory, this is one of the reasons we are including it in our tutorials this year. However, it is not entirely reliable to the point where we can entrust all our coding to it.

Why not? Generative AI does not have the ability (yet) understand the nuances of the intended application of your code, making manual checks necessary. Debugging is a necessity (that's why we covered the basics of coding in Tutorial 1). On the other hand, the benefits of generative AI are significant. It gives us the ability to not have to write long, simple. Or redundant segments of code ourselves.

Using the knowledge from Tutorial 1, we will prompt generative AI to code for us and use it to check code we write.

As generative AI continues to evolve, I predict that the decision on what to do will ultimately be left to human judgment. I hope all of you will study with the thought of expanding your options in coding and continue to develop your skills beyond this course. This is why it's important to study theory and apply it to real data. Let's keep this in mind as we continue our studies.

*Submit a PDF file with your plots and answers to the questions, and include your python notebook .ipynb file with your code.*

# To submit (90 marks total)

## Visual assessment of data distribution (35 marks)

1. Use glob to get the list of weather data, then read the data to Google Colab. You should have 8036 x 30 data. See the table and describe what type of data is this? We explore the metadata in the tutorial 1, as the temperature data will be examined later, what other important information can we get from the table data? Explain (5 marks).
2. Compare the histograms created by the prompt below. Show the two graphs and explain what the difference. Why there is a change? Hint: check the minimum value of the data (5 marks).
   1. Prompt 1: Prepare a histogram of daily values for “Total Precip (mm)”. Use bins=80.
   2. Prompt 2: Prepare a histogram of daily values for “Total Precip (mm)” that are greater than 0. Use bins=80.
3. The first prompt in question 2, likely, gave you code that included “.resample('D').sum()” for the total precipitation. Breakdown this segment and describe: What it is doing to the data? Is this what you asked Gemini to do? What did Gemini add or not include from your prompt? (5 marks)
4. Prepare a bar plot of daily values totaled by year, for “Total Precip (mm)”. Use axis labels: xlabel="Year ", ylabel="Annual Total Precip (mm)". Compared to the question 2: What are the differences between a bar plot and histogram in terms of this “Total Precip (mm)” data. (5 marks)
5. Prepare a histogram of daily values totaled by month, by year, for “Total Precip (mm)”. Use bins=50. What is the difference from question 2? (5 marks)
6. Prepare a bar plot of daily values totaled by month, by year, for “Total Precip (mm)”. Use axis labels: xlabel="Year,month", ylabel="Annual Total Precip (mm)". Show the result graph. This will likely black out your x-label. Why? How can you correct it? Show your solution and explain. (5 marks)
7. Prepare a histogram of daily values totaled by year for “Total Precip (mm)”. Comment on the rough patterns of distributions you see for question 2, 5, 7. (5 marks)

## Mean, standard deviation, correlation (20 marks)

1. Determine the mean and standard deviation of annual sums of “Total Precip (mm)”. Present the numbers in a table as integers. Calculate mean and standard deviation for monthly totals for “Total Precip (mm)”. Plot the two variables on a single bar plot, with the series side-by-side. In what month does Victoria Gonzales station have the greatest average precip? The lowest? Compare to the metadata find (question A.1), does it make sense in terms of weather pattern? (10 marks)
2. Consider the mean and standard deviation for January and November. What observations can you make and what do your observations suggest about how consistent the precipitation total for each month is likely to be? (5 marks)
3. Prepare a correlation matrix for the sums of precipitation data taken by month and by year (i.e. you should have a two-level index and more than two hundred values – if you only have 12 values, one for each month, you have done it incorrectly). Use style format methods in pandas to fix the number of decimal places and use a colour gradient to make the table look like this:

A screenshot of a computer

Description automatically generated

* 1. The matrix will be a 12x12 grid
  2. These resources can help you:
     1. <https://pandas.pydata.org/docs/reference/api/pandas.io.formats.style.Styler.format.html>
     2. <https://pandas.pydata.org/docs/reference/api/pandas.io.formats.style.Styler.background_gradient.html>
     3. Colormaps: <https://matplotlib.org/stable/users/explain/colors/colormaps.html>

1. Which set of months shows the strongest positive correlation in the graph? Why do you think this is the case? (5 marks).
2. You are working with a client who runs a golf course (near the Gonzales weather station) and they are worried about water stress on the grass in the summer, and wants to know if they should set aside funds for higher municipal water use. They ask you, “If July has very little precipitation, can I assume August will also have low precipitation?”. What would you say to your client? (5 marks)

## Scatterplot, linear regression (25 marks)

1. Use plotly express and create a scatterplot with the annualized mean monthly precipitation totals with x=January and y=February. Use proper x and y axis labels. Show the graph. Based on the scatterplot, can you predict February precipitation if you know January precipitation? How accurate would this prediction be? (10 marks).
   1. This website will help:
      1. <https://plotly.com/python/line-and-scatter/>
2. Use plotly express and plot a scatterplot of year on the x-axis and March and September (separate plots) on the y-axis with an “ols” (ordinary least squares) line of best fit through the timeseries. Use the mouseover function and determine what the annualized rate of change in monthly total precipitation is for March and for September (10 marks).
   1. This website will help:
      1. <https://plotly.com/python/linear-fits/>
3. Discuss about the plot created in question 13 in terms of overall comparison, variability, trends over time and outliers (5 marks).

## Reflection (10 marks)

1. What is the debug for python? What are the important things for that? From your experience? (10 marks)