GEOG/EOS 230

Tutorial 2: Basic stats and plotting

In this tutorial we have the following objectives:

- Review the metadata.
- Review the reading data with the Python library, Pandas.
- Practice cleaning data.
- Practice plotting the data.

BirdLife Australia, Birdata

For this assignment you will be using a bird occurrence and nesting observations dataset from Australia. This dataset can be used to draw bird distribution maps and generate bird lists for any part of the country. The original dataset is over 11GB, too large for our purposes, so we have trimmed it down to a 600MB (and 1GB if you are comfortable with the size and the additional processing time) file and placed it on Brightspace for use in the assignment. Additional information on the source dataset can be found here: https://www.gbif.org/dataset/4bf1cca8-832c-4891-9e17-7e7a65b7cc81

*** Please note that Birdlife data is available under a non-commercial license and consultants should contact Birdlife directly to arrange access to full resolution records for commercial purposes.*

Plotly Open Source Graphing Library for Python

Plotly is a powerful and versatile graphing library for Python that allows you to create interactive graphs and visualizations. You can create interactive plots that allow users to zoom, pan, and hover over data points to see more details. This feature is different from the other Python libraries like Pandas and NumPy. Since Plotly integrates well with libraries like Pandas and NumPy, it makes for an easy fit to visualize data from dataframes. Plotly supports a variety of chart types, including line plots, scatter plots, bar charts, histograms, heatmaps, and offers extensive customization options for styling and formatting your plots. See Brightspace for links to Plotly Bar and Scatter plots.

Awesome Plotly with code series (Part 1): Alternatives to bar charts - *Jose Parreño Garcia*

This article provides a great introduction to the concept of how others view our data results/visualizations and how important it is to consider the target audience and message we are trying to communicate with our data analysis. The article shares the authors view on "how to transform Plotly's powerful features into sleek, professional-grade charts that meet data journalism standards.". Please read this article, we also encourage you to follow the example code and use it as your reference for the rationales to the visualization choices on the assignment plots.

Submission

Submit a PDF file with your plots and answers to the questions plus a python notebook .ipynb file with your code. There is no sentence limitation for each question, but please write neatly and clearly. Please change the colour of your answer to add visual separation from the question.

To submit (55 marks total)

Check metadata and reading the data (10 marks)

1) Download the "occurrence_trimmed2.tsv" from the Brightspace. Open the file with notepad++ and describe the file format .tsv. How is the data separated? (5 marks).

2) Read in the data into the Google Colab and use the Pandas library to read the data with the argument (header=None, low_memory=False). For the separator, try the types of separators listed below. Find the right one. Show the result. How many columns are there and how do we find the names? (5 marks) ** hint check the metadata.

List of Separators in Pandas read_csv().	
Comma (,)	The default separator for standard CSV files.
Tab (\t)	Used for TSV (Tab-Separated Values) files.
Semicolon (;)	Sometimes used in certain regions or specific data formats.
Space ( ):	Used for space-separated data.
Pipe (|)	Used for pipe-separated data

Cleaning the Data (10 marks)

3) Can you plot the data without any changes? What cleaning/processing may be needed for this dataset? Discuss briefly (5 marks).

4) Download the "meta.xml" file from the Brightspace. Select columns with data you believe would be interesting to analyze/plot. Subset your full dataframe into a new dataframe with only the columns of interest you selected. If using dates, set as datetime object and set it as the index. Show the top few records (5 marks).

Plot the data with Plotly (35 marks)

5) Use a Plotly Express bar chart to plot your data and show the result. Do you see any patters? Are the data distributed well? Does the plot clearly communicate the data? Explain (10 marks).

6) For the "tourism trap" plot in the reading: Why did the author choose the final plot layout to communicate the data? What considerations did they make in building the plot? Briefly explain. (5 marks).

7) Using the process discussed in the reading improve the Plotly Express default plot you created in question 5. Describe your final plot, including the message you are trying to communicate with your data (10 marks). Explain the improvements you made and the reason(s) why (10 marks).

Submission

Submit a Python ipynb file containing your code results, comments, and markdown. Submit a pdf file with the answers to the questions above including your plots, you do not need to include your code in the pdf.

- Ensure your code is well-commented and follows good coding practices.
- Use meaningful variable and function names.