Programming for Bioinformatics | BIOL 7200

Week 9 Exercise

October 17, 2023

Instructions for submission

- Run a Linux or Mac terminal on your computer
- You may want to create a directory to work in (e.g., "~/biol7200/class9/ex9")
- Download the data file, week9_data.tar.gz, into your working directory
- Write your script(s) to be compatible with Python version 3.11
- Name your script file(s): "gtusername_questionnumber.py"
- If you wish to write and use local modules, submit your script and associated modules in a tar.gz file. Make sure everything works by copying the tarball somewhere and unzipping it and ensuring everything still runs. Your tarball and compressed directory should have a name of the format "gtusername_questionnumber.tar.gz" and "gtusername_questionnumber/", respectively.
- Name your other submission file "gtusername_questionnumber.(jpg|png)"
- Upload your script files and other submission file on canvas.

Grading Rubric

This assignment will be graded out of 100.

You must only submit a solution to one question this week. Choose your favourite out of the options. Don't submit solutions to more than one, but of course you are encouraged to solve all of them if you would like to!

This assignment is intended to assess two things:

1. Your ability to use Python to read, process, and filter data into whichever format you need.
2. To design and impliment an effective visualization using Python and matplotlib.

You do not need to identify and use appropriate statistical tests to support your analysis of the provided data. You only need to produce a visualization. You can use statistical tests if you wish, though.

There is no right answer as to which is the best way to visualize these data. Feel free to experiment and submit something creative. As long as you follow the below submission specifications and produce a sensible visualization, you will get the points.

We will discuss different visualizations (what was more or less effective about each) and review different approaches to creating them during next Thursday's class.

Submission specifications

1. You must perform all data manipulation, analysis, and plotting using Python.
2. You may write your analysis to be specific to the provided files. Your script does not need to be generalizable to another input file. However, do not hard-code the path of the input file. The file path should be provided as a command line argument.

3. You may use the following packages in your solution (Note, only the core library and matplotlib are required. You need only use the others if you would like to):
    - Any Python core library package
    - matplotlib
    - numpy
    - pandas
    - scipy
4. Your submission should consist of two files:
    1. Your Python script (.py) or a tar.gz file containing your script and any local modules that are used by the script.
    2. An image file of your plot in .jpg or .png format
5. You do not need to plot every datapoint. If you think that the message conveyed by your visualization is made clearer by only showing a summary statistic such as the median value, then you should do so.

Assignment options

If you need any clarification about the datasets provided, use the discussion page for this week to ask for clarification.

Each assignment option includes a description of the dataset and a prompt about what you should generate a plot to address.

1. Influenza hospitalization rates

Dataset description

The tarball included on canvas contains a file "Flu/FluSurveillance_Data.csv". This file contains data describing the rate of hospitalizations due to Influenza infections between 2009 and 2023. The file was downloaded from the CDC.

The Influenza hospitalization data include hospitalization rates (per 100,000) people broken down by several age and race categories as well as sex. "Overall" is used to indicate a summary of the rate of all catagories. i.e., "Overall" in the "SEX CATEGORY" column represents the overall rate for males and females of the specified age range and race.

Both a weekly rate as well as a cumulative rate are included. The weekly rate represents the proportion of hospitalizations in a given week, while the cumulative rate represents the sum of hospitalizations up to that point within a given flu season (i.e., the sum of weekly rates for those categories of patients up to that point in the current season).

Data are reported per week, with week 1 being the first week of January and week 52 being the last week of December. A flu season in the US starts in the Fall and ends in the Spring, so each flu season is represented with data from two years and with week numbers starting around 35 and ending around 15. i.e., week 1 is not the first week of the flu season. Week 1 is the first week in January, but is in the middle of the flu season.

Assignment prompt

Did the COVID-19 pandemic, which began in 2020, impact Influenza hospitalizations?

2. Global average temperature anomaly

Dataset description

The tarball included on canvas contains a file "global_temperature/temp_anomalies.csv". This file contains data describing global average temperature anomolies for each month of the years 1850 to 2023. The data were downloaded from the NOAA. A temperate anomaly is defined on the NOAA website as follows.

> The term temperature anomaly means a departure from a reference value or long-term average. A positive anomaly indicates that the observed temperature was warmer than the reference value, while a negative anomaly indicates that the observed temperature was cooler than the reference value.

Within the file, it is stated that the anomalies reported here are relative to the average temperature between 1901-2000.

Each temperature anomaly is reported as a difference from the average temperature as measured in degrees celcius. The period of time during which each reading was recorded is reported as a 6 digit number of the format YYYYMM. i.e., the first 4 digits are the year and the 5th and 6th digit are the month. e.g., 185001 is January of 1850.

Assignment prompt

Have global temperatures changed between 1850 and now?

3. SARS-CoV-2 spike protein mutations

Dataset description

The tarball included on canvas contains a file "covid_spike/VOI_RepresentativeSpike_Translated.fasta". This file is a multifasta of aligned sequences of the SARS-CoV-2 spike protein. Each sequence is a variant of the spike protein from a different lineage of the SARS-CoV-2 virus. The data were downloaded from the GISAID

In each sequence, gaps in the alignment are represented with "-" characters. All other characters represent the amino acid at that position in the alignment.

The spike protein of SARS-CoV-2 is an important protein involved in viral pathogenesis. In addition, it is a common target of antibodies. Therefore, mutation in the spike protein can have impacts on viral fitness and immune evasion. Not all mutations will impact these traits. However, many mutations do impact one or both of these traits. Sometimes, by comparing sequences of the same gene or protein from different isolates, we can infer which parts of a gene or protein are significant to the function. For example, positions that are under strong diversifying selection often have more mutations.

Assignment prompt

Which amino acid positions in the spike protein are the most mutated?