Programming for Bioinformatics | BIOL 7200

Exercise 7

Choose your favourite dataset out of the options below and write a script that creates a plot. The plot should communicate something about the data, perhaps focusing on analyzing the data in a way that you find interesting. I have included an example prompt for each dataset but you do not have to use those prompts. You can focus your plot on whatever you would like.

You must only submit a solution to one question this week. 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 implement 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.
3. The autograder will put your script and all provided data in the same directory. You should hard-code the path to your chosen dataset as <folder>/<file> (e.g., "Flu/FluSurveillance_Data.csv")
4. You may use the Python core library packages and matplotlib
5. Your submission should consist of two files:
    1. Your Python script (plot.py)
    2. An image file of your plot in .png format (plot.png)
6. Your script should generate the uploaded image using `plt.savefig()`. Don't upload a screenshot of the popup you get from `plt.show()`. The autograder will assess whether the plot your script produces matches the uploaded plot (if you use jitter or other randomization ensure consistent output by setting the seed)
7. 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 a subset of the data or 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 Ed Discussion page to ask for clarification. This week you may ask about and post code publicly. This assignment is intended to be an opportunity to use code creatively to achieve desired visualizations. The goal is for each of you to decide what ***you*** want to plot and to figure out how to do that. As opposed to other weeks when the goal is for you to learn specific code syntax or concepts.

You can bounce ideas off one another, discuss how to achieve an effect or talk about analyzing the data. Just remember the guiding principle that you shouldn't do anything that undermines your learning. Remember that there is no right answer so you should absolutely do your own thing rather than think that anyone elses approach is better. Experimenting with wild plotting ideas can be a valuable part of research - it's a good way to explore your data.

Each assignment option includes a description of the dataset and a prompt about what you might want to generate a plot to address. You can also generate a plot to show something else. This assignment is going to be assessed based on your ability to generate a plot using Python, but the content of the plot can be whatever you like so feel free to experiment.

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 this year. 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 categories. 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?
- Is Influenza incidence different between group X and group Y (where X and Y are any groups you are interested to compare)

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 this year. 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, we might expect to see positions that are under strong diversifying selection have more mutations, while positions under stabilizing selection may have fewer mutations and positions under no selection have an intermediate number (genetic drift rather than natural selection).

Assignment prompt

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

N.B. while this prompt is short, the term "most mutated" is not simple. What could "most mutated" mean? I used a deliberately imprecise term here partly so that you could determine how you want to interpret it, but also to illustrate a case in which what initially might read like quite a plainly clear sentence actually falls apart when you think about what exactly it means. Communicating data as a scientist is two skills in one: visualizing data in clear and compelling forms, and describing what you did and what your data show in clear and precise language. Both of these skills take a lot of work to develop and require a great deal of thought about the nature of your data and the audience to which you are attempting to communicate. It is worth the effort to

develop these skills though. As science is a communal effort (and you are generally assessed based on other people's perception of your work), the ability to communicate your work to others effectively is essential!

How do you assess the amount of mutation at a position? Is it the number of sequences that differ relative to a reference, the number of different amino acids at a position among the sequences, or something else? Do you need to consider evenness in the dataset (i.e., is the dataset 99% one thing and single sequences with variants or are variants roughly equally abundant with no dominant sequence)? A huge amount of thought has been put into questions like quantifying diversity in the field of ecology and the theory developed there for communities can be applied to sequences. If you are interested, this review offers both a detailed description of mathematical approaches and an approachable narrative in between the dense maths. For this assignment you should think about what you would consider a useful way to define "most mutated" and plot that. Perhaps try defining it in different ways and see what each approach indicates about the sequences being analyzed.

4. COVID-19 death and test positivity rates

he tarball included on canvas contains a file "covid_rates/covid_rates.csv". This file contains measures of weekly deaths attributed to COVID-19 infection (see source for data details) and weekly test counts and positivity rates for nucleic acid amplification tests (NAAT). Data were downloaded from the CDC COVID data tracker.

The global focus on COVID-19 means that there exists an abundance and quality of data relating that virus which would be unthinkable for any other - even such common and widespread viruses as influenza. One example of such data is estimates of the number of infections at any given time over several years through extensive surveillance efforts (both per-person testing and community surveillance through sewage monitoring). The provided dataset describes the number of deaths attributed to COVID-19 per week and the number, and positivity rate, of NAATs over the same period.

Surveillance data such as these do have flaws (though I haven't found a detailed description of how these data were generated so I don't mean to suggest the following definitely apply in this case - I just want to prompt consideration of possible issues). For example, when considering test positivity at different points in time you should consider who was being tested (did everyone have equal access to testing?), under what circumstances (were people more likely to be tested if symptomatic?), what kind of test was conducted (should a self-administered antigen test be considered equivalent to a hospital-conducted PCR test?) and whether reporting of test results was skewed based on what that result was (would you be more likely to tell your doctor if you tested negative or positive?). Furthermore, were the answers to those questions the same at all times between 2020 and now and in all geographic regions?

In any effort to gather large quantities of data there is a tension between gathering as much data as possible and gathering data that is as high quality as possible. Effective communication of research based on data such as these should always grapple with questions about what can be concluded from the data given the weaknesses and biases of those data. Whether you should account for the details of your dataset generation in your visualization or in the accompanying description (figure legend, main text, spoken words) depends on the specifics of your situation and should be considered carefully.

Assignment prompt

Does increasing test positivity predict a later increase in weekly deaths? Did the rollout of vaccines starting at the end of 2021 impact that relationship?