Data challenge (ITpS postdoc recruitment)

This document provides instructions for a challenge of data analysis & visualization. Once you finish analyzing and plotting the data to answer the questions below, you'll be asked to give a short slide presentation on March 16th, 2022 (in Portuguese, 5-7 minutes), showing the results you've obtained. If you have questions, please contact us: recruitment@itps.org.br

The dataset (download here)

The file '`metadata_brazil_variants.tsv`' contains the metadata columns below, from more than 92,000 SARS-CoV-2 genomes sequenced from samples collected in Brazil, between March 1st, 2020 and December 31st, 2021:

`identifier` = sample ID;
`date` = date of sample collection;
`pango_lineage` = SARS-CoV-2 lineage, as shown in cov-lineages.org;
`variant` = variant name as designated by the WHO;
`country` = country where the COVID-19 sample was collected;
`state` = state where the COVID-19 sample was collected;
`state_code` = 2-digit state code;
`ibge_code` = state code according to IBGE;
`age` = age of the infected individual;
`sex` = sex of the infected individual;
`lat` = latitude of the state of collection;
`long` = longitude of the state of collection.

Questions

Using your coding and data visualization skills, analyze and plot the data described above to answer the questions below. You can create visualizations with as many plots and subplots as you want, using custom colors schemes, and even creating infographics: just make sure the graphical representations are intelligible, and nearly self-explanatory. Be creative, and choose the most adequate styles of plots to visualize the data. You can use this visual vocabulary to get inspired.

1. Genomic surveillance in Brazil has improved over the past months of COVID-19 pandemic. Using the dataset described above, show us how many daily SARS-CoV-2 samples were sequenced in Brazil, from March 1st, 2020 to December 31st, 2021. How well genome sequencing developed in 2020 and in 2021?

2. Dozens (likely hundreds) of SARS-CoV-2 lineages circulated in Brazil during the current COVID-19 pandemic. Some of these lineages were classified as “variants of concern” (VOC) by the WHO. As in most countries, once a new VOC is introduced in Brazil, it may replace the circulating viruses, becoming dominant. Using the same dataset provided above, how was the distribution of the different variants (Alpha, Beta, Gamma, Delta, and Omicron) in the Brazilian states?

3. The data provided can be visualized in many ways, and this visual vocabulary illustrates alternative visualizations. Using the same dataset, explore your creativity, and provide answers to new questions not yet considered in the previous tasks.

Scripts

At any point between today and March 15th, 2022, upload to your GitHub the scripts you’ve used to analyze and visualize the dataset, and send to recruitment@itps.org.br the URL of the GitHub repository with the scripts. There is no need to provide any documentation or instructions (README) about how to use the scripts: we only want to assess how you approached the questions above using programming languages of your choice.