Lab 5: Visualizing Coronavirus Data

Name: Sal Figueroa

2025-04-28

Contents

Github Repository *Holds all related files*

github: Figgs0bit-Lab4 (https://github.com/Figgs0bit/CSIT165-Lab5.git)

Required data sets *This lab represents data downloaded on 04/xx/2025*

Human Proteins Data Set ()

1. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Global Confirmations.

This data set is operated by the John Hopkins University Center for Systems Science and Engineering (JHU CSSE). Data set includes a daily time series CSV summary table confirmed cases of COVID-19. Lat and Long refer to coordinate references for the data field. Date fields are stored in MM/DD/YYYY format.

2019 Novel Coronavirus COVID-19 (2019-nCoV) Global Confirmations

2. Human Proteins Data Set

This data set is a tibble created from parsing Homo sapiens protein fasta files curated by the Genome Reference Consortium as part of the Human Genome Project. The original protein fasta file can be found in NCBI, here. Data consists of two columns, Gene and Protein.Sequence. Gene represents every gene product, or protein, made by humans. Protein.Sequence represents the primary amino acid structure of its correspondent gene. Each amino acid is represented as a single capital letter and the sequence of letters is unique to each gene.

Instructions

Before beginning your objectives in your final document, please state which day you downloaded the data sets on for analysis. The objectives for this lab will cumulatively cover many subjects discussed in this course and will also contain an objective for manipulating strings.

The surgeon general for the United States recently created a new data science initiative, CSIT-165, that uses data science to characterize pandemic diseases. CSIT-165 disseminates data driven analyses to state governors. You are a data scientist for CSIT-165 and it is up to you and you alone to manipulate and visualize COVID-19 data for disease control.

Objective 1

Create a scatterplot for counts per day of the the top five confirmed countries. For this objective, please use dplyr and tidyr to manipulate data and ggplot2 to create the visualization. Scatter plot must have specified colors, a non-standard theme for display, and custom a customized titles, axis labels, and legend labels.

Objective 2

Understanding how COVID-19 enters human cells requires that we have a better understanding of human proteins. It has been shown that COVID-19 is able to enter cells by binding to Angiotensin-Converting Enzyme 2 receptors in the heart, lungs, and intestines. Angiotensin-Converting Enzyme 2 is used by the body to regulate blood pressure and inflammation.

1. Show how many different isoforms of Angiotensin-Converting Enzyme 2 humans make using str_detect and regular expressions. Use the pattern "Angiotensin-Converting Enzyme 2 isoform" with regular expressions to include all variations irrespective of first letter capitalization.

2. Show the amino acid sequence between the 27th and 63rd amino acid sequence for each isoform using str_sub.

3. Print a statement using cat with paste or sprintf of the first thirty amino acids for each isoform with a new line after each isoform and sequence listing. cat is necessary to use in combination with paste or sprintf to output concatenation with a new line.