Assignment 9: Time Series Graphs

80 Points scaled to 20 Points

Introduction

In this assignment, you will work with time series data of monthly average temperatures in Rio de Janeiro, Brazil and COVID-19 cases aggregated to the state-level.

The temperature data ("station_rio.csv") were obtained from Kaggle: https://www.kaggle.com/datasets/volpatto/temperature-timeseries-for-some-brazilian-cities.

The COVID-19 data were also obtained from Kaggle: https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-usa.

Objectives

- *Clean data to prepare for graphing*
- *Generate and interpret time series graphs*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and graphs embedded. Files can be rendered to HTML webpages if your instructor requires this. Graph prompts should be stated within Markdown cells.*

Tasks and Graphs

This assignment can be conducted using either Python (matplotlib, seaborn, and pandas) or R (ggplot2), whichever you prefer or whichever you instructor requires.

Rio de Janeiro Temperature Graphs

The temperature data represent monthly average temperature from January 1973 to October 2019. Some monthly data points are missing. Also, the table is in the wrong shape to perform the analyses.

Task 1: Reshape the data so that the year, month, and temperature measurements are presented in three columns. In other words, reshape the data so that each month is not represented in a separate column. (5 Points)

Task 2: Remove any rows that have missing data. The missing code in the datasets is 999.9. (5 Points)

Task 3. Create a new column to recode the months into seasons. DEC, JAN, and FEB should be recoded to “SUMMER”; MAR, APR, and MAY, should be coded to “FALL”; JUN, JUL, and AUG should be coded to “WINTER”; and SEP, OCT, and NOV should be coded to “SPRING”. (5 Points).

Task 4: Aggregate the data to obtain yearly median temperatures from the monthly data. (5 Points)

Task 5: Aggregate the data to obtain seasonal medians for each year from the monthly data. (5 Points)

Task 6: Generate a grouped box plot to summarize the monthly data. Make sure the months are in calendar as opposed to alphabetical order. (5 Points).

Task 7: Generate a grouped box plot to summarize the seasonal data. Make sure the seasons are in calendar order as opposed to alphabetical order. (5 Points)

Task 8: Create a line graph or time series of yearly median temperature. (5 Points)

Task 9: Create a line graph or time series of seasonal median temperatures where each line represents a different time series for each season. (5 Points)

Task 10: Summarize your results in a paragraph. Discuss both the seasonal central tendency and seasonal variability in the data. Do these data suggest a warming trend in the city over the decades included? (5 Points)

COVID-19 by State

The COVID-19 data consist of cumulative counts of cases (“cases”) on a daily basis from January 21, 2020 to December 5, 2020. Each row consists of a date (“date”), the county name (“county”), the state in which the county occurs (“state”), the county FIPS code (“fips”), the number of cumulative cases (“cases”), and the number of cumulative deaths (“deaths”).

Task 11: Aggregate the county-level data to obtain the total cumulative number of cases in the state per day. (5 Points)

Task 12: Aggregate the county-level data to obtain the total cumulative cases for the entire United states per day. (5 Points)

Task 13: From the data aggregated by state per day, extract out only the records for New York, North Dakota, and Washington. (5 Points)

Task 14: Create a time series graph of cumulative cases for the entire United States over the provided time period. (5 Points)

Task 15: Create a time series graph of cumulative cases for the three selected states with three separate lines to differentiate the states. (5 Points)

Task 16: Write a paragraph to compare the three states that were graphed in regards to when COVID-19 became prominent and how fast the number of cases grew. (5 Points)