DANA4800 – SUMMER 2023

Descriptive Analysis Assignment 2

DEADLINE: FRIDAY – WEEK 5 – 5:00PM (VANCOUVER TIME).

Download file of datasets to analyze from this site: https://www.airnow.gov/international/us-embassies-and-consulates/.

The main aim of this exercise is to explore and identify patterns.

Datasets: the datasets provide the profiling of recent PM2.5 concentration levels in cities over a number of years. The purpose is to detect patterns of air pollution in the cities in relation to their health implications, in accordance with the U.S. Environmental Protection guidelines.

The cities you work on are assigned in the spreadsheet.

Data is stratified from historical measurements of $PM_{2.5}$ concentration levels, acquired from various fixed air quality monitoring instruments in the U.S. Embassy. Air pollution includes the analysis of periods including months, days and hours within the above timeframe.

Your job is to clean and format the datasets by following the below steps:

- Drop all invalid values in the NowCastConc., AQI, and Raw Conc.
- Select values of air pollution levels are stored every 6 hours
- Combine all years into a master dataset
- The sample of clean data is provided (Hanoi 2019 – every 6 hours)

Using monthly data, generate:

1. Boxplots to compare air pollution levels on a yearly basis
2. Scatterplots to compare trends of air pollution levels on yearly basis
3. Histograms to compare the distribution of air pollution levels on yearly basis

What do these graphs tell you about the data and their structure?

Using systematic sampling method, select:

4. 4 records per month.

and conduct:

5. AQI Category comparison among all years.
6. Air quality (AQI) correlation among all years.

What do these tests tell you about the pattern of air quality in the cities?

Write up an analysis of what you find in this data, including all the information you answered above. This write up should include the following for credit:

Questions	**Contents**	**Points allocations**
Cleaning and Screening tasks *(Dataset and variable descriptions and issues and methods relating to data cleaning/screening)*		8
Descriptive analyses *(Graphs and descriptions of graphs)*		16
Interpretation & Discussion *(Meaningful interpretation and identification of patterns. Pros and cons of descriptive analysis techniques discussed)*		26
Total		**50**

Format of the report

a. A4 paper standard
b. 1.5-line space
c. Times New Roman 12 point
d. Normal margin

Submit the following files:

1. The report of your analysis in Word/PDF
2. Python or R codes
3. The dataset after cleaning.