DANA4800 – Summer 2023

Data Screening/Cleaning Assignment 1

Download the four (4) datasets file to analyze. This data aims to explore and understand the data itself. You want to screen/clean the whole dataset at once to look for issues of data in terms of data entry and participants' responses.

This assignment has a total of **50 points**. Deadline of submission is **Friday, Week 3 at 5:00 PM (PDT time zone).**

Data Description

Purpose

The datasets provide the profiling of recent PM2.5 concentration levels in Ulaanbaatar city. The purpose is to elucidate the patterns of air pollution in the city in relation to their health implications, in accordance with the U.S. Environmental Protection.

Data is stratified from historical measurements of $PM_{2.5}$ concentration levels, acquired from fixed air quality monitoring instruments in the U.S. Embassy in Ulaanbaatar, Mongolian. Air pollution includes period throughout 2018 to 2021.

In order to conduct any inferential analysis, there is a need to screen/clean the data for any entry errors, missing data, and distributions of data. You want to merge and check the data for the following:

1. Merge the monthly data into a master dataset and categorize on types of variables (categorical and numerical) – provide an explanation the categories that you outlined.
2. Accuracy
    a. Check the data for out of range scores. Include the codes and its outputs to show any out range.
    b. If necessary, fix the out of range scores.
        i. Describe how you fixed them.
        ii. Include a R/Python codes and its outputs showing that you fixed the accuracy issues.
    c. Other accuracy issues that you might detect
3. Missing data
    a. Include a R/Python output that shows that there is not missing data.
    b. What type of missing data do you appear to have?
    c. If necessary, "fix" the missing data (remember there are several options).
        i. Describe what you did to the missing data.
        ii. Include a R/Python output showing that you fixed the missing data (you may repeat a box you had earlier).

Note: on this section, you will walk through univariate screening/cleaning options because it helps to practice.

4. Outliers:
    i. Use scatterplots to detect outliers for each continuous variable.
    ii. How many outliers did you have for each continuous variable?
    iii. Explain the rationale of the outliers that you identified in (ii)?
5. Univariate Normality
    a. Include histograms of the continuous variables.
    b. Identify the shape of histograms?

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**
Question 1	Merge datasets & Categorize variables	16
Question 2	Accuracy check	4
Question 3	Missing data	4
Question 4	Outliers	4
Question 5	Univariate Normality	8
Summary	Report on the master data and summarize issues that were identified in the 5 questions	14
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 data screening in Word/PDF
2. R codes
3. The dataset after cleaning.