Analytics In Practice (MIS 64038)
Assignment I – Data Preparation

Total Points: 50

Exercise 1:

The Open Data 500 is the first comprehensive study of U.S. companies that use open government data to generate new business and develop new products and services. Open Data is free, public data that can be used to launch commercial and nonprofit ventures, conduct research, make data-driven decisions, and solve complex problems [https://www.opendata500.com/]

In this assignment you will be analyzing two datasets – "US_agency.csv" and US_companies.csv". You can download the data from the website:

https://www.opendata500.com/

Or from blackboard, on which I already have downloaded.

Your task is to explore the dataset and answer the following questions:

1. Are there any missing columns?
2. Are there any missing column names or errors in the column names? If so, name those columns.
3. Are there any values in the columns missing?
4. How is data organized in each column? Is it properly organized?
5. Is data in the proper shape for further analysis? If not, why? Explain.
6. How will you fix this dataset? Describe the methods you will use to fix this dataset for further analysis? It can be missing values, NAs, etc. (OPTIONAL: Uploading clean dataset)
7. How are the two datasets linked to each other? Is there a common "primary key" to connect the two datasets?

Total Points: 35

Exercise 2:

JSON (JavaScript Object Notation) is a most commonly used data format today and as a data scientist, you must know how to access JSON data sets. JSON is easy for machines to parse and generate. "It is based on a subset of the JavaScript Programming Language Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent [JSON.ORG]."

For this case study, you will parse JSON file, which has city traffic details. "Average Daily Traffic (ADT)" counts are analogous to a census count of vehicles on city streets. These counts provide a close approximation to the actual number of vehicles passing through a given location on an average weekday. Since it is not possible to count every vehicle on every city street, sample counts are taken along larger streets to get an estimate of traffic on half-mile or one-mile street segments. ADT counts are used by city planners, transportation engineers, real-estate developers, marketers and many others for myriad planning and operational purposes. Data Owner: Transportation. Time Period: 2006. Frequency: A citywide count is

taken approximately every 10 years. A limited number of traffic counts will be taken and added to the list periodically [https://catalog.data.gov/]".

Your task is to process the JSON file and answer the following questions:

1. How many variables are in the dataset?
2. Name all the variables?
3. What is the total traffic of vehicles on 100th street to 115th street?
4. What is the total traffic of vehicles on geolocations, (41.651861, -87.54501) and (41.66836, -87.620176)

Instructions:

1. Use "R" or "python" programming language and Data wrangling techniques you have learnt in your other courses.
2. Provide your answers by copying the output (screenshot).
3. List all the libraries you have tried (and used) to solve this problem
4. Summarize your learning in one paragraph (400 words maximum) at the beginning of the report

Total Points: 15