CSC 400, Data Analysis #5, Spring 2024 Name _____

Instructions: More information on expectations for the report can be found in the general directions for the data analysis sets. In this document, the specifics for the individual assignment will be discussed. Students are responsible for both the requirements in the general directions, and for the specific directions discussed below.

Topic 1: Outliers and Anomalies

For this analysis, you'll use the dataset posted in the course in the file **400DA5.xlsx**. This dataset contains information on 325 cities (metro-areas) in the United States with information on cost of living, crime, healthcare, transportation and other factors that contribute to overall quality of living. Most of the data is numerical, although some is categorical and may need to be transformed appropriately for further analysis.

Analysis the dataset generally, including each variable, and perform an outlier analysis or other anomaly detection methods (such as clustering methods). Recall that unusual values are considered to be observations that occur less than 5% of the time. In particular, identify any extreme outcomes that occur less than 1% of the time (or meet other extreme criteria such as more than three standard deviations from the mean, more than 3IQR above or below the quartiles, etc.) There are different criteria for being an outlier or an anomaly depending on whether you are considering one variable or multiple variables at a time, so different methods are likely to produce different results.

Identify the cities that stand out the most in your analysis and explain what makes them different from other cities in the dataset (either for good or ill). In your analysis, be sure to explain what types of statistical tests you used to identify the outliers. You should use at least one type of statistical test for individual variables, at least two types for multiple variables (such as regression or clustering). Include any plots you used to identify and analyze the outliers once identified. Do the outliers have anything in common (this could be geographically or other factors).

You may include your code in an appendix or separate file, but the report of approximately 10 pages should focus on the analysis. It should look professionally formatted. Raw code and raw output is frowned upon.