Mid-Term Project

Ambrose M Malagon Cordero

Bellevue University

DSC 540 – Data Preparation

Friday January 19, 2020

**Food Access Research Atlas – Data Set.**

The data set chosen for this Mid-Term project was the Food Access Research Atlas [1]. The Excel file contains three sheets:

* Read Me: A description of the file contents.
* Variable Look-Up: Translation table of the field names in the Data Table with the long name and the description of the field.
* Food Access Research Atlas: The Data Table: Consists of 147 variables with 72,864.

A first glance of the file shows a neat file with no visible anomalies such as missing values. There were no challenges importing the data into the data dictionary, although I did struggle with a bit with identifying outliers, missing data and data types steps; I was following the steps detailed in *Data Wrangling in Python* [2] and the results did not look right at first: It was only giving counts of one. I managed to figure it out in the end, so it should all look correct.

Each of the fields contained a unique identifier, which makes me work under the assumption that there should be no duplicate records. I went ahead and followed the steps to de-duplicate the data to see if the assumption was incorrect, however the result was as expected: 72,864 unique values.

With Fuzzy String Matching, I went with the notion that someone wanted to see if there was find a State and County name combination, with the understanding that they might not have all details or the incorrect spelling. After testing out several threshold levels, above 60 percent worked as the best. Any higher, resulted in no returned values.

**Conclusions: Thoughts and Next Steps.**

The resulting data import has left me with questions. In particular the high number of observations that have a 0 value: Are these true values or do these represent missing values? I will need to do some further research on this. Once I have an answer, then I can proceed to do some Exploratory Data Analysis to see what kinds of insights this data yields.

**References:**

[1] Alana Rone. Food Access Research Atlas. United States Department of Agriculture – Economic Research Service. Website. Last Update October 31, 2019. Retrieved January 16, 2020 from <https://www.ers.usda.gov/data-products/food-access-research-atlas/>.

[2] Jacqueline Kazil, Katharine Jarmul. Data Wrangling with Python. O’Reily. Chapter 7. Pages 169-178. Retrieved January 16, 2020