Victor

Kaitlyn

Miranda

Matthew

IST 707

**Introduction**

**Analysis and Models**

**About the Data / Exploring the Data**

**Clustering**

For the Food Desert data set, the machine learning algorithm known as ‘clustering’ was chosen to find interesting or valuable outcomes within the data set. This method is viewed as an ‘unsupervised’ algorithm due to the lack of a training aspect in which a target variable is already determined. In the field, data miners/scientists/engineers/etc. utilize unsupervised clustering to explore data that may not be well researched. It can often lead to the assistance of developing future supervised machine learning algorithms in which known groups can be assigned.

Within the scope of this project, possible limitations occur in terms of the size of the data set. After removing ‘NA’ values, the data set is left with ~70,000+ rows. It was determined to simply remove NAs since an abundance of data is still available for use within various algorithms. From here, scaling numeric variables is necessary to ‘level the playing field’ and ensure all numeric variables are comparable (i.e., all numeric variables contain a standard deviation of 1 and an overall mean of 0).

Following these basic pre-processing steps – assuming other pre-processing actions were taken (as explained in an earlier section) – the ‘kmeans,’ ‘hclust,’ and ‘agnes’ functions within R were utilized (as well as others) to develop these algorithms and visualize the results. This covers both the K-Means (i.e., ‘kmeans’) and Hierarchical (i.e., ‘hclust,’ ‘agnes’) methods.

**Results**

**Conclusions**