Big Data Implementation Plan

Mitchell Beckner

1/29/2021

Procured Dataset - Review

Following final import and joining of the US Census, USA Cycling, and NOAA Tables, the resulting dataset is a single 5.39 GB csv file It contains aggregations of 155 variables from 6,495,834 unique census blocks. Census blocks are the smallest level of geography that basic demographic data can be obtained for and are the building blocks for census block groups which generally contain between 600 and 3,000 people. The data collected covers all 50 States plus the District of Columbia.

Data Cleaning and Processing

The 5 year census data tables with the desired information, and the USA Cycling and NOAA data were not available at the block level. To bring this information to the block level of granularity, these tables were expanded from tract level (in the case of census tables) or zip code level (in the case of the USAC table). Using population data from the 10 year P2 table, the population proportion for each tract and zip code was calculated at the block level. For example, block 0001 of a tract might contain 5% of the tract's total population, block 0002 7%, etc. The track and zip code level tables were then expanded to block level by using these population proportions to distribute the less granular data across each tract's or zip code's blocks. Housing data from census table B28011 was expanded in a similar manner based on the proportion of households present. The NOAA data, which existed at the county level was simply joined "as-is" to the main dataset using county fips codes. The assumption was that weather data did not vary significantly at the tract or block level and was sufficiently granular at the existing county level.

Additional data cleaning steps included removing blocks that had no zip code identifiers associated with them and blocks with a population of 0 as these contained no useful information. Several outliers were detected in the *Median Household Income* and *Population* variables. These values were replaced with Nulls (NA). Many blocks contained Null/NA values for the USA Cycling variables. Since it is known that the USAC data is complete, these NA values were not unknown but in fact 0. The data was modified to reflect this fact. The binning of variables in the *Age* and *Education* variables was adjusted to create more uniform divisions in the case of *Age*, and to reduce the granularity of the *Education* categories that were present between nursery school and 11th grade.

The final data preparation step was standardization. This was performed in order to reduce the bias that could be introduced by comparing variables with drastically varying ranges of values. Standardization was accomplished using the z-score method. This step, and all previous cleaning steps were performed in *R*. All R coding produced can be viewed here: Github link. The final cleaned and standardized dataset that will be used for modeling is a 7.2GB csv file with aggregations of 114 variables from 3,761,586 unique census blocks. Using the *ZCTA5* and *block ID* variables, the data will be analyzed at what equates to the Zip+4 level.

Next Steps

The next step in the process will be to actually create the segmentation model. K-means clustering in *R* will be used to accomplish this. K-means clustering is an unsupervised learning process that divides

the observations in a dataset into a predetermined number of groups or clusters based on the similarity between their variables. In order to produce the most appropriate model, several modeling parameters will be varied and the results will be compared. Various distance rules and linkage methods will be tested, and the algorithm will be run many times for each set of parameters in order to reduce the effects of random initial cluster centers. Also, the final numbers of clusters for the model will be varied and the optimum number will be selected using the silhouette method which measures the quality of the clustering. Once a final model has been selected, it may be possible to perform heirarchical clustering on the resulting groups to determine higher level "super groups.'

The final resulting clusters will be named and their characteristics evaluated based on the marketing objectives of the bicycle industry (retail sellers of bicycles and components, bicycle centric organizations such as USA Cycling, etc.) The objective here will be to identify specific areas where marketing efforts are likely to achieve the greatest results. These areas may vary depending on the specific product or service being offered. The final deliverable may be an interface that shows the segmentation groups across a specific local area and/o the ability to select a product or service type and then display the areas best suited for that item.