Re: Analysis of Themes in Census Data

This memo contains analysis of census data to identify the themes that exist in the data.

Specifically, a model and analysis that would address the following questions:

1. What are the major ways in which the US distinguishes itself demographically?
2. “I am building some demographic questions into a survey. What are the 5 or 6 most important general demographic questions to ask on that survey?”

Factor analysis is a statistical method used to identify latent variables, or 'factors', in a dataset. By grouping variables based on their associations, it reduces data dimensionality and uncovers underlying patterns or themes. Results of the factor analysis indicate that the themes in which US distinguishes itself demographically are:

* Economic Status and Education
* Household Structure
* Age and Family Composition
* High-Value Home and Ownership
* Racial Demographics

Based on the themes uncovered in the data, the following demographic questions are suggested to be included in the survey:

1. What is your household’s approximate annual income?
2. What is the highest level of education you’ve achieved?
3. How many individuals live in your household?
4. What is the age range of the members?
5. Do you own or rent your place of residence?
6. To which racial or ethnic group do you most identify?

The overall results from factor analysis are shown in Table 1 below. It indicates how strongly each variable is associated with each of the five factors. These associations can vary from -100% to 100%. Numbers that have larger absolute values indicate that the variable is more strongly associated with the factor. An analysis of each factor is provided on the following pages.

**Table 1**

**Detailed Results from Factor Analysis**

Factor 1: Economic Status and Education. Partial output from the summary table as it relates to the first factor is provided in Table 2. The first three variables indicate that the regions scoring high in this factor are overall financially flourished. The next two variables ‘median years..’ and ‘% population..’ show a directly proportional relationship between educational levels and socio-economic status. The OOH home value index and percentile variables depict that wealthier areas are more likely to have higher home values.

**Table 2**



Factor 2: Household Structure and Ownership. reflects a pattern of housing and ownership, dominated by single-unit structures (80% association), signifying neighbourhoods with likely suburban single-family homes. The negative associations with multi-unit structures (-80% for 3+ units, -74% for 10+ units) and renter-occupied housing (-79%) underscore a preference for homeownership (87% association) and family households (80%). This factor captures the essence of traditional, owner-occupied, family-oriented residential areas as shown below in Table 3.

**Table 3**



Factor 3: Age and Family Composition. encompasses demographics related to age and family size depicted in Table 4, with positive associations for households with children (75%) and larger household sizes (71% for 3+ persons), as well as a younger adult population (72% for ages 35-44). The pronounced negative associations with older age groups (-86% and -91%) and median age (-90%) suggest a dynamic of younger family-centric neighborhoods.

**Table 4**



Factor 4: High-Value Home Ownership. characterized by the financial stature of home values within a community, with all associated variables shown in Table 5 below indicating wealthier domiciles: median home values (73%), homes valued over $500k (74%), and homes valued over $200k (79%). This factor delineates areas with a significant presence of high-value real estate.

**Table 5**



Factor 5: Racial Demographics. the racial makeup of neighborhoods, with a marked decrease in the percentage of white residents (-79%) and an increase in Black or African American residents (70%).

**Table 6**



The technical appendix to this document which is attached below contains a discussion of the data preparation and statistical steps that were executed as a part of the analysis.

**Technical Appendix.**

This appendix discusses the technical issues relevant to the factor analysis that are contained in the previous portion of the memo.

The analysis was based on a dataset representing household data about all zip codes in the United States of America, that were collected by the census.

The following data preparation and cleansing steps were executed prior to running the factor analysis:

**Exclusion of Non-Quantitative Variables**: Only quantitative variables were retained for the analysis. Non-quantitative identifiers and categorical variables, such as ADICODE and Zipcode, were excluded to concentrate on variables that provide scalable and measurable data.

**Mean Replacement for Missing Values**: The dataset initially contained 34,297 records spanning 33 variables, with 33,377 of those records being complete. There were 1,650 missing values in total. For example, the “Income Index” variable had 136 missing values, “Persons Pet Household” had 121 etc. To prevent data loss that would have occurred had these missing values been omitted, the following calculation illustrates the preserved data due to mean replacement:

34,297 – 33,377= 920 rows

920 x 33= 30,360 data points

Or 30,360 – 1,650= 28,710 data points retained

Therefore, mean values of the distributions for the respective zip code variables were used to replace the missing data. This approach helped maintain the integrity of the dataset without significantly altering the overall data distribution.

**Standardization of Variables**: Each variable was normalized to a mean of zero and a standard deviation of one. This ensured that variables of widely varying scales contributed on an equal footing to the factor analysis.

All variables were standardized to a mean of zero and a standard deviation of one. This was done so that variables with remarkably different scales would be able to contribute equally to the factor analysis.

Technical decisions relative to the factor analysis itself were:

**Varimax Rotation Application**: Varimax rotation was implemented to maximize the loadings for each variable on its primary factor, thereby enhancing the clarity and interpretability of the factors.

**Threshold for Variable Association**: A threshold of 0.675 was set for the absolute values in the rotated component matrix to identify significant associations between variables and factors. This rigorous threshold was chosen to ensure the retention of only the most pronounced relationships.

**Determining the Number of Factors**: The initial selection of factors was based on eigenvalues greater than one, which indicates factors that account for a significant portion of the variance within the dataset. However, the initial analysis revealed that some factors, such as those representing percentages of the Hispanic population, household size, etc., did not meet the set threshold for associations. As a result, the number of factors was reduced to five to eliminate redundancy and to focus on those that provided a unique variance explanation. A scree plot illustrating this selection process is provided as Figure 1.

**Figure 1**

A graph with a line graph and numbers

Description automatically generated