Capstone Project Data Wrangling Process

Jake Schroeder

**Step 1: Get Demographic Data**

As my mentor Andrew and I discussed during our first call, finding yearly demographic data was going to be an adventure. The typical demographic data comes in ten-year increments (census data). Given the young nature of Google, and the limited historical data provided by Google Trends (2004 as the earliest year of data), this would not do.

As Andrew recommended, there was more data selection available at the city-level for major cities. However, this brings in other variables that could skew the data early on, and make the creation of a “straw man” solution difficult. For example, what if changes in keyword trends were attributable to regional changes, such as a large migration of residents from a particular section of the country?

With a national scope, these variables would largely be nullified since each region’s growth and decline are mixed together into a larger, higher-level picture of the nation as a whole. At this high-level position, an overarching, simplistic model could be created more easily. This “straw man” model could then be improved to scale locally as the availability of data allows.

Therefore, I use annual demographic projections from 2008 and 2012 from the US Government. While 2012 offers four more years of up-to-date information; the 2008 model has more overlap with the Google Trends historical data. Specifically, the 2008 model starts from 2000, while the 2012 model starts from 2012.

Statistically speaking, the 2008 model is more likely to produce actionable results given the better quality of historical data. That said, I’ve retained the 2012 data in case an opportunity arises where I can combine the datasets to get both the extended historical data of the 2008 dataset and the modernity of the 2012 dataset.

**Step 2: Connect to Google Trends**

To begin, I sought out a Google Trends API that I could use to fetch the data more efficiently. Through a stubborn research process, I came across the “pytrends” library, which is an unofficial API structured for easy use by python developers. It is based on a similar project written in JavaScript, and reliably pulls historical data for up to 5 keywords at a time.

After downloading the library using pip install, I looked through some example documentation to setup the first pull. As the documentation promised, the results came in an easy-to-comprehend DataFrame format, indexed by date (in weekly intervals) with each column representing a keyword’s popularity.

**Step 3: Cleaning the Data**

My first step in cleaning the data was to drop unnecessary columns from the demographics information. These columns were RACE, SEX and HISP, representing the race, gender and Hispanic ethnicity of the population being recorded and forecast.

The demographic dataset included additional rows below the general forecast for each race, gender and Hispanic ethnicity combination. These rows were removed from the file to make for faster processing and easier data wrangling.

Therefore, this cleaning left me only the rows and columns needed for a general analysis. I made a note in my script to consider the unused data in the future if modeling or data analysis calls for it.

From a Google Trends perspective, the data was relatively easy to clean, as it contained its data in an easy-to-manipulate format. I simply group the data by year and use the mean aggregation function to get the yearly average. When inspecting the data beforehand, I did not notice large outliers, therefore giving me confidence to use the mean, rather than any other measure of center.

**Step 4: Combining the Data**

To get around the five-keyword limitation of the Google Trends API, I loop over lists of keywords and join the resulting DataFrames together. To ensure the index of the Google Trends and demographics data line up, I set the index of each to the year attribute of its DateTime value.

I then use a simple DataFrame join to get the data together into a master DataFrame, df.

**Step 5: Dealing with Null Values**

Null values are kept, since there are no projections for Google Trends (data stops in 2018) and the demographic data starts four years before Google Trends begins recording.

As for Google Trends itself, I used the Keyword Planner tool to generate 20 relevant, sought-after keywords in the banking industry. These keywords did not have brand-related terms in them - this was done to preserve data integrity. I then tested these keywords to ensure there was Google Trends data for each.

Some returned very low data values, and others returned all zeros. To deal with this, I shortened up the keywords to make them more generalized. After confirming quality data for each of the 20 keywords, I verified that the data was cleaned and ready for analysis.

*Examples of keyword modifications include:*

“how to open a foreign bank account” => “bank account”

“bonds vs savings account” => “savings account”

“1 year savings plan” => “savings plan”

“direct deposit savings” => “direct deposit”