|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | |  | |  | |  | |
| Population & Search  Understanding the relationship between population demographic trends, and the trends of our digital behavior. |
| Jake Schroeder  Springboard Intermediate Data Science  2018 |

Table of Contents

I. Understanding the Problem 1

II. Identifying the Client 2

III. Describing the Data 3

IV. Other Data Sources 4

V. Initial Findings 5

## I. Understanding the Problem

## Population demographic changes have been upon us for decades. Our digital data footprint spans 14 years. How can we draw understanding from such a brief relationship?

### This research aims to illustrate the relational impact of demographic impact upon Google search data.

### Specifically, given a set of twenty highly popular words from a specific industry, what correlations, if any, can we draw? For example, if the older demographic tends to be interested in using online banking (since limited mobility makes traditional banking a literal pain in the rear), do we observe a statistically significant correlation between this age demographic and the related keywords?

To take things further, we aim to forecast this relationship alongside the forecast population demographic data from the United States of America Federal Government. Such a forecast would enable proactive strategic planning by digital marketers, especially those in the field of content creation and optimization. By creating an effective forecast model, we can more accurately align content marketing to its various personas (demographically-related audiences). Such alignment would also generate a more favorable long-term return on investment for the parties involved.

## II. Identifying the Client

## Content marketing can take years to develop. Investors need to have some way of understanding the long-term outlook for content before investing.

The target client for this project is any company seriously considering an investment in digital marketing; namely content. Through the use of this project and its analytics, the target client can better understand which keywords, and their corresponding topics, share the strongest relationship with their target audience.

In content marketing, the target audience of a product is broken down to *personas* - subsets of the target audience whose need is situational in nature. These personas often carry distinct demographical attributes. By using this analysis, a content team would be able to create more relevant content for its personas based on its understanding of its personas age.

This analysis serves the client - companies looking to invest in an enhanced digital presence - by providing their content writers the validation to pursue topics that their core audience (and its subset of personas) will actually care about.

## III. Describing the Data

## This analysis focuses on an annualized dataset containing both Google Trends and population forecast data.

The Google Trends data was retrieved using the unofficial Google Trends python library, pytrends. The data came in biweekly increments, and was aggregated to a yearly form so that it could be compared alongside the population data. Population data was sourced from the census bureau in the form of population forecasts, since the output was yearly and considered the population demographics in the years beyond. This future-oriented nature of the dataset lends it well to work such as forecasting and projection.

Using pandas, the data was wrangled in such a manner than all unnecessary columns were dropped (notably, race, ethnicity and gender). With the data aggregated yearly, each row (year) from 2004 to 2018 shares both population and Google Trends data for the selected keywords. This allows ample space for further data manipulation and statistical inference.

## IV. Other Data Sources

## The current datasets may not meet the needs of statistically rigorous machine learning and forecasting. Therefore, an alternative dataset is provided, below.

The National Cancer Institute, in cooperation with the US Census Bureau, grants access to the annually released population estimates at the county level. At the time of this writing, the estimates span the years 1969-2015. This data is available in 19 age groups (5-year spans) or by single year age groups.

By aggregating the county level data into a statewide dataset, we would have yearly population data for over 40 years across all US States. We could use this data to apply relationship statistics across a range of scenarios, thus granting enhanced data integrity. Simply put, this added data could help answer the question, “Does the age demographic and keyword relationship remain as strong from one state to another?”

Below are the links to the downloads and documentation for parsing the population data:

<https://seer.cancer.gov/popdata.thru.2015/download.html#single>

<https://seer.cancer.gov/popdata.thru.2015/popdic.html>

## V. Initial Findings

## The strong trends shown by the population data are mirrored in statistically sound Pearson Correlation Coefficients. The results urge for further investigation and discovery.

Through a two-sample permutation t-test, the analysis produced high confidence in the repeatability of the strong Pearson Correlation Coefficients initially gathered from the data. Thus, we rejected the null hypothesis that such strong coefficients were not valid or reproducible.

The permutation test involved randomly sampling pairs of data, such as:

(x = Population of Age Group, y = Popularity of Keyword)

This was accomplished through random sampling of the index that the two datasets share (2004-2018). This entire process, including the production of a Pearson Correlation Coefficient, was repeated 10,000 times per keyword, per age group. The large amount of iteration and calculation (a staggering 20,000,000 times total) consumed such computing power that the computation nearly took an hour to complete.

Each age group and keyword permutation test yielded a mean Pearson Correlation Coefficient and a p-value. The p-value was calculated based on the number of permutation test replicates with a stronger correlation coefficient, divided by the total number of permutation test replicates (10,000).

Across the board, p-values were remarkably strong. The vast majority was above .5, suggesting that strong correlations are the norm between these two sets of data. The statistical significance of this finding lends itself well to the practical significance of solving the initial problem stated in this report; by exploring the relationship between these datasets in such a thorough manner, we can confidently confirm the existence of a meaningful relationship worth further exploration by more advanced means (such as machine learning).