Stephanie Schouman
Final Project Report
SI 370

Motivation

I'm very interested in the ways data analysis and technology can play a role in people's health, whether that be in hospital settings, wearable fitness devices, or population health analytics. For this project, I investigated datasets that relate data and healthcare and found the Searching for Health (http://www.searching-for-health.com/) interactive data visualization tool. This tool looks at different Google search trends by U.S. county across the last 13 years. I found the way people are using the internet to investigate and become more attuned with their health very interesting and wanted to explore this data further. I focused primarily on differences across geographic locations and over time to see if there were any interesting patterns in the way people are using Google to search for health-related information.

Data Sources and Description

My primary data set is called "Health searches by US Metropolitan Area" from Kaggle. The original dataset had 2 columns describing location:

- dma: location as a county name and state (string)
    - This dataset only had 210 entries for counties compared to the actual 3141 counties in the United States. Some of these entries had multiple cities/counties combined into a larger category.
- geocode: arbitrary geocode location 500-710 (integer)

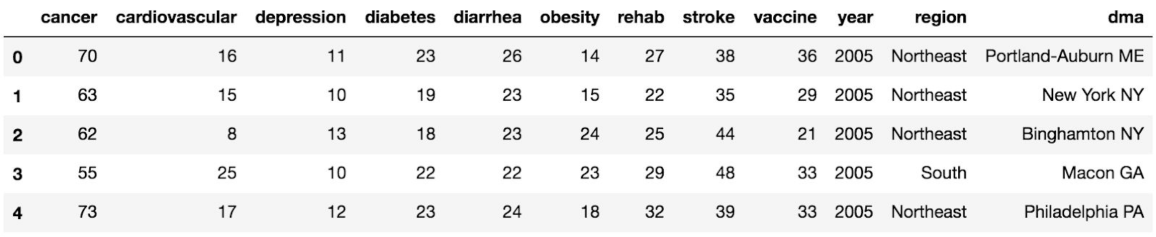
There were also 117 categories describing the google searches, each with a year and the actual search category:

- 2004+cancer: total number of Google searches for "cancer" in thousands (integer)
- ...
- 2017+obesity: total number of Google searches for "obesity" in thousands (integer)

After some cleaning to get these columns separated into search category and year, the final dataframe included the following measurements:

- dma: location as a county name and state (string)
- year: year from 2004 to 2017 (integer)
- region: geographic region (Northeast, Midwest, South, West) (string)
    - Region data from: https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States
- search category: total number of Google searches for category in thousands (integer)
    - Cancer, cardiovascular, depression, diabetes, diarrhea, obesity, rehab, stroke, vaccine
- state: US state listed in the dma column for analysis per state (string)

ex.



	cancer	cardiovascular	depression	diabetes	diarrhea	obesity	rehab	stroke	vaccine	year	region	dma
0	70	16	11	23	26	14	27	38	36	2005	Northeast	Portland-Auburn ME
1	63	15	10	19	23	15	22	35	29	2005	Northeast	New York NY
2	62	8	13	18	23	24	25	44	21	2005	Northeast	Binghamton NY
3	55	25	10	22	22	23	29	48	33	2005	South	Macon GA
4	73	17	12	23	24	18	32	39	33	2005	Northeast	Philadelphia PA

Research Questions and Methods

1. How have health-related Google searches changed over time?

- Time series line graph

2. Are there any expected or unexpected correlations between search categories?

- Bivariate analysis, SPLOM

3. Are there any common clusters within the Google search data? Would principal components analysis find anything interesting?

- K-means clustering, principal components analysis

4. Are the distributions for the searches across each geographic region similar?

- Categorical analysis, Kruskal-Wallis test

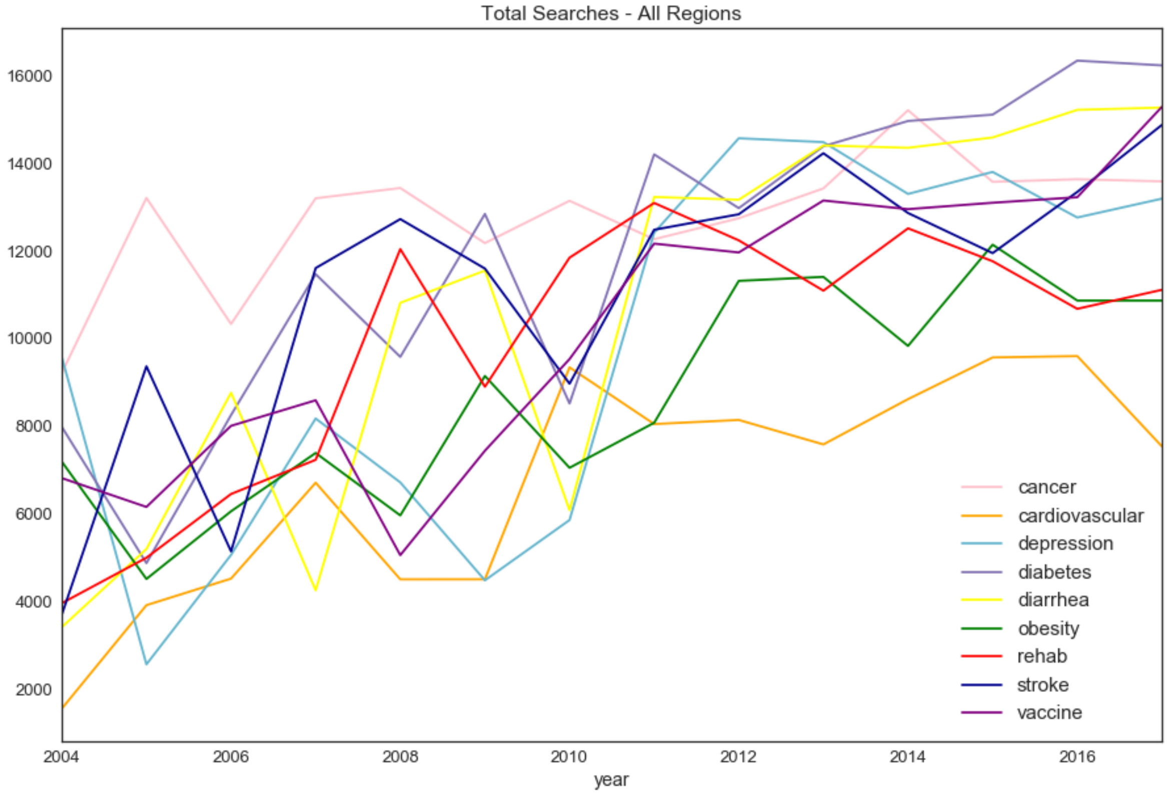
5. Which state has the most total health searches? What about total searches per category?

- Data visualization

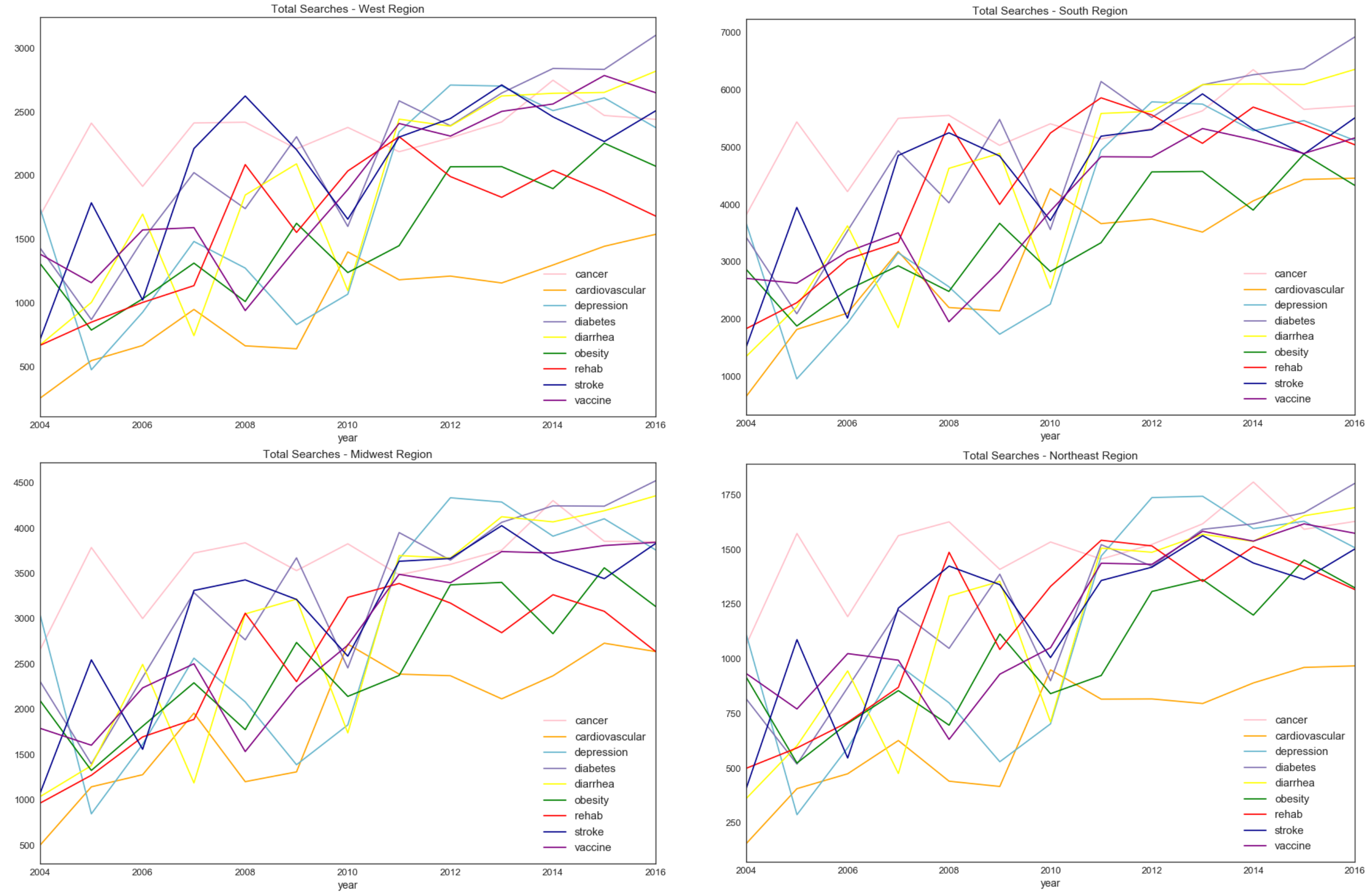
Analysis and Results

Q1: How have health-related Google searches changed over time?

Since the total number of Google searches has increased over the last 13 years with the increased availability and accessibility of technology that connects to the internet, I would expect the number of health-specific Google searches to also have increased. To investigate this change, I decided to look at the totals for each search category for each year and then plot these collapsed totals against years. I wasn't able to do an actual time series plot, as the data only had data for each total year, meaning finding seasonality would be difficult. The following line graph shows the total number of searches per category over the years:



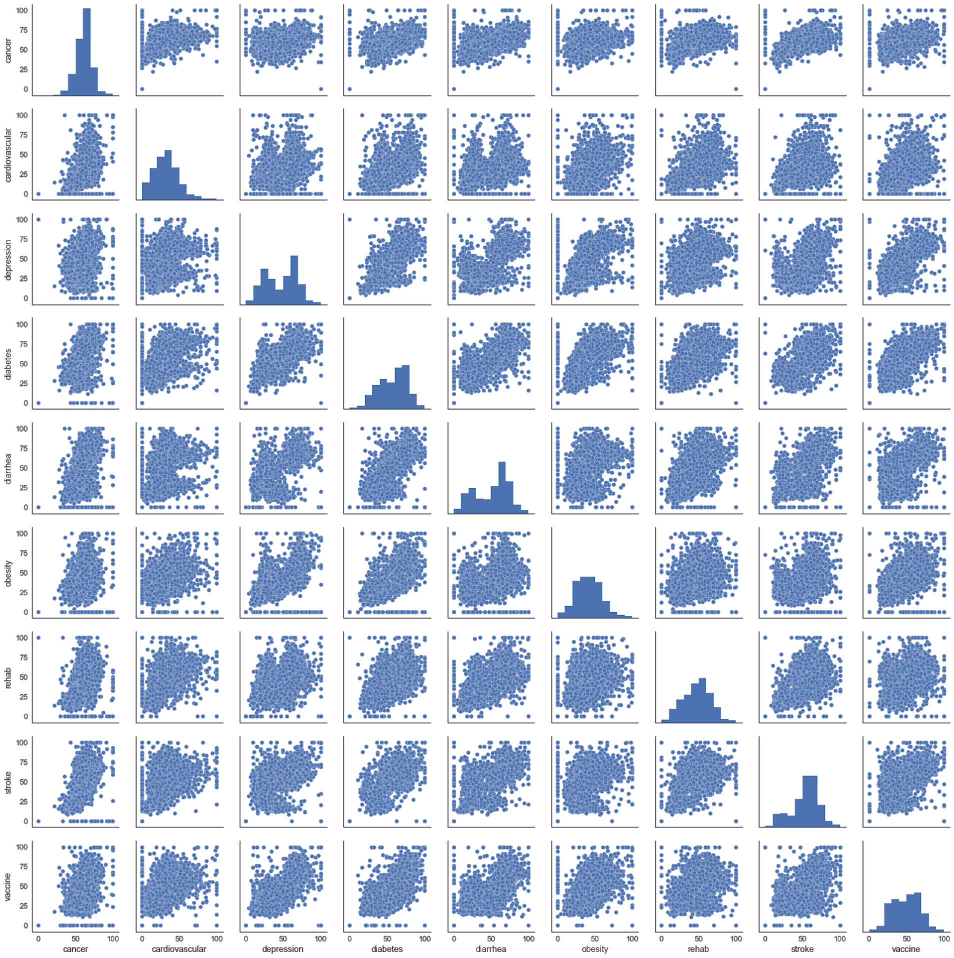
It appears that most health-related searches have increased over time, increasing from between 1000 to 9000 (in thousands), to between 7000 to 16000 (in thousands) on average. I decided to break down these totals across the four geographic regions to see if the same increasing trend was still present for all locations:

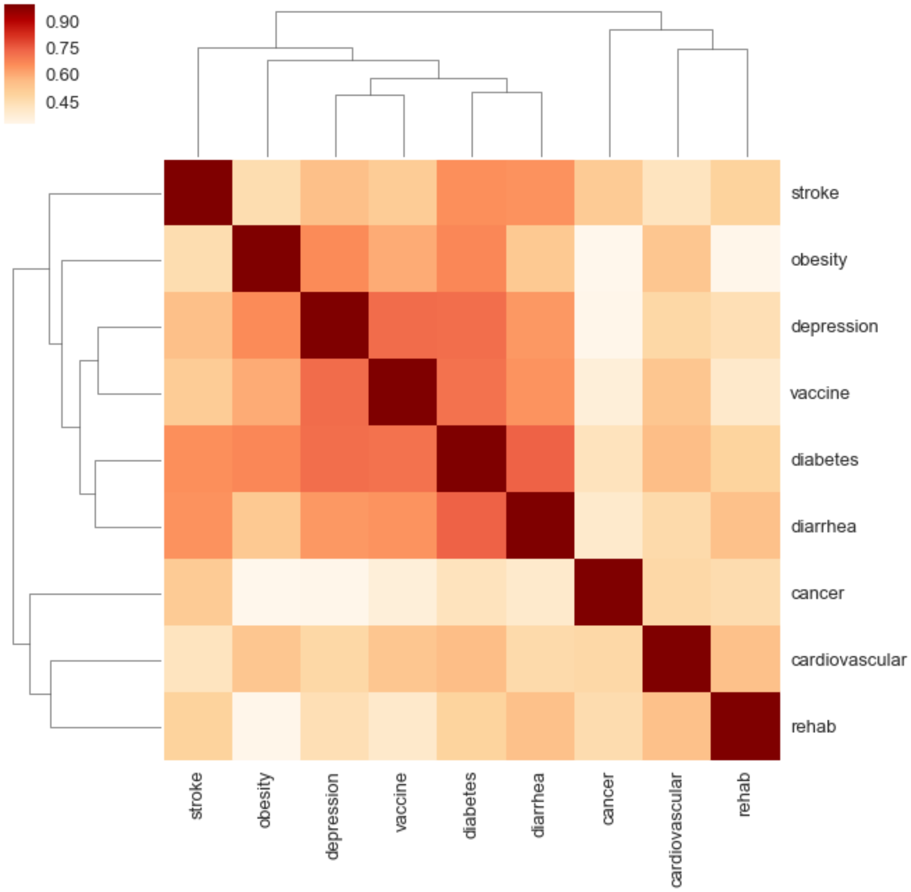


When breaking down into regions, it seems that the majority of search categories are increasing over time. While some categories are increasing at a faster rate, both the totals collapsed over region and each region show an upwards trend. We could assume at least one of the reasons for this increase can be attributed to the increased use of the internet via tablets and smartphones, which could lead to increased Google searches in general, as previously stated. Another interesting find is that the sharpest increase in searches in both the collapsed totals and all regions seems to occur around 2010, which aligns with the signing of the Affordable Care Act. This suggests a pivotal point where individuals are taking more interest and agency in their own healthcare.

Q2. Are there any expected or unexpected correlations between search categories?

To investigate the correlations between search categories, I conducted scatterplot matrix (SPLOM) pairplot, as well as a clustered heatmap to better visualize the correlations. The correlations run are using the Spearman correlation because the data is non-parametric:


cancer
cardiovascular
depression
diabetes
diarrhea
obesity
rehab
stroke
vaccine
0
25
50
75
100
0
50
100
cancer
cardiovascular
depression
diabetes
diarrhea
obesity
rehab
stroke
vaccine


0.90
0.75
0.60
0.45
stroke
obesity
depression
vaccine
diabetes
diarrhea
cancer
cardiovascular
rehab
stroke
obesity
depression
vaccine
diabetes
diarrhea
cancer
cardiovascular
rehab

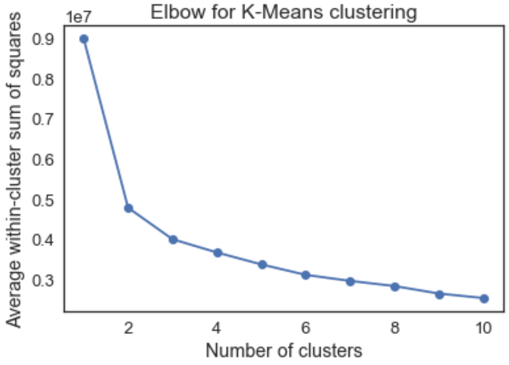
The strongest Spearman's correlations are listed below:

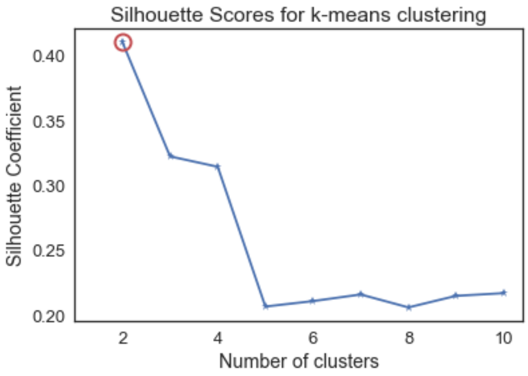
- Diabetes and diarrhea: 0.749477
- Depression and vaccine: 0.729196
- Diabetes and depression: 0.723749
- Diabetes and vaccine: 0.717544

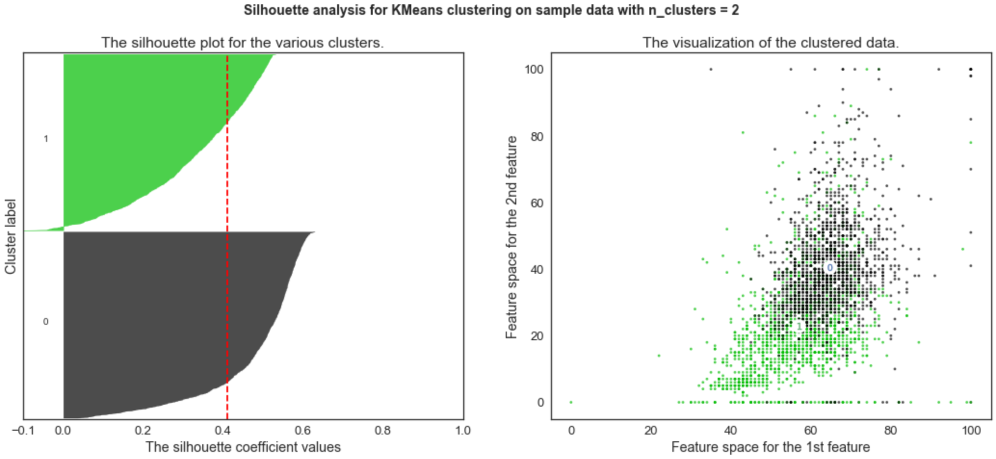
These correlations are not the correlations I expected to see but many of them make sense. The search categories with the most correlations stronger than 0.6 are depression, diabetes and vaccine. These categories often have some sort of effect on other health conditions. For example, depression is a common side effect of diabetes and obesity, and if a patient is being treated via antidepressants, that has implications on how they can be treated for other conditions. In addition, with the anti-vaccination movement, many people are investigating if and how vaccinations are impacting health issues like autism or diabetes.

Q3. Are there any common clusters within the Google search data? Would principal components analysis find anything interesting?

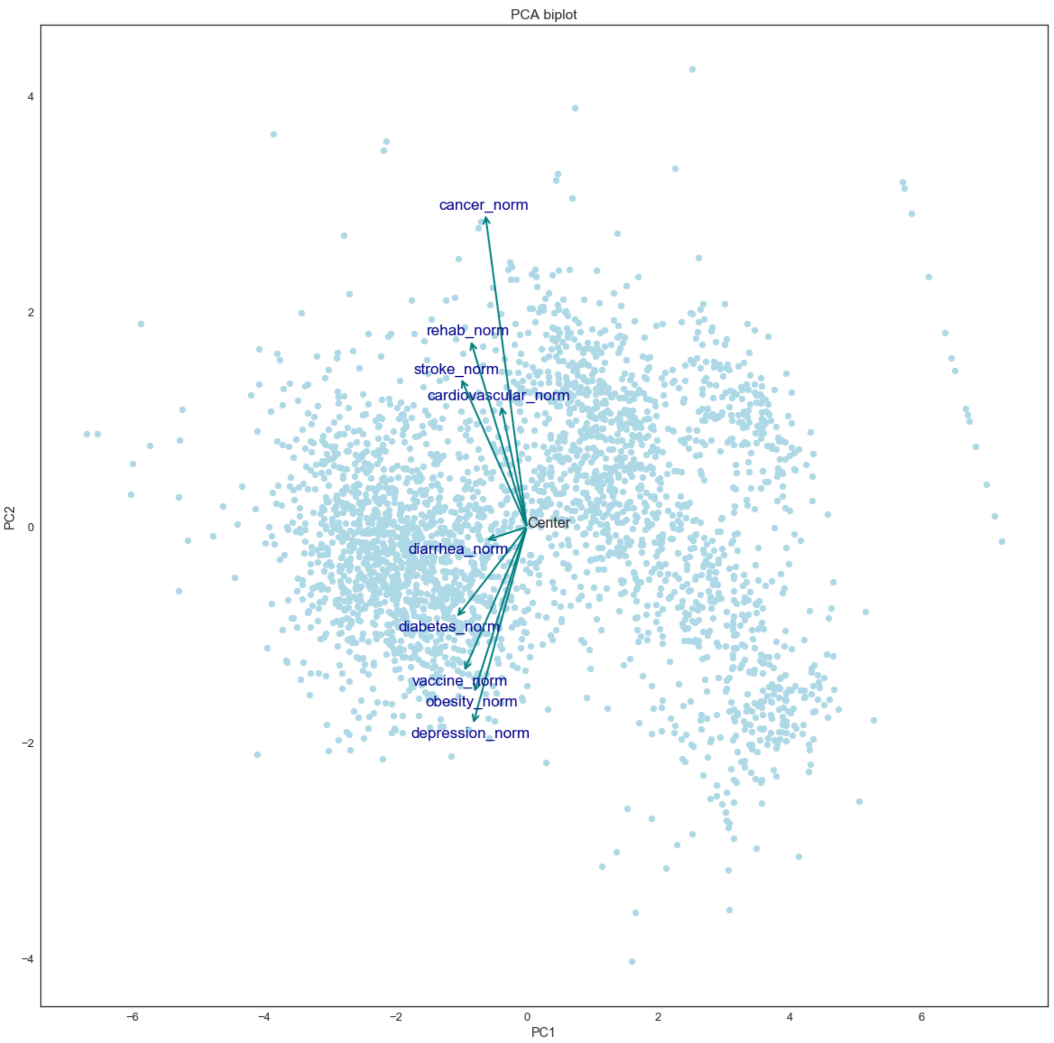
To determine how many clusters to include in my principal components analysis, I used k-means analysis followed by the elbow plot and silhouette scores methods:







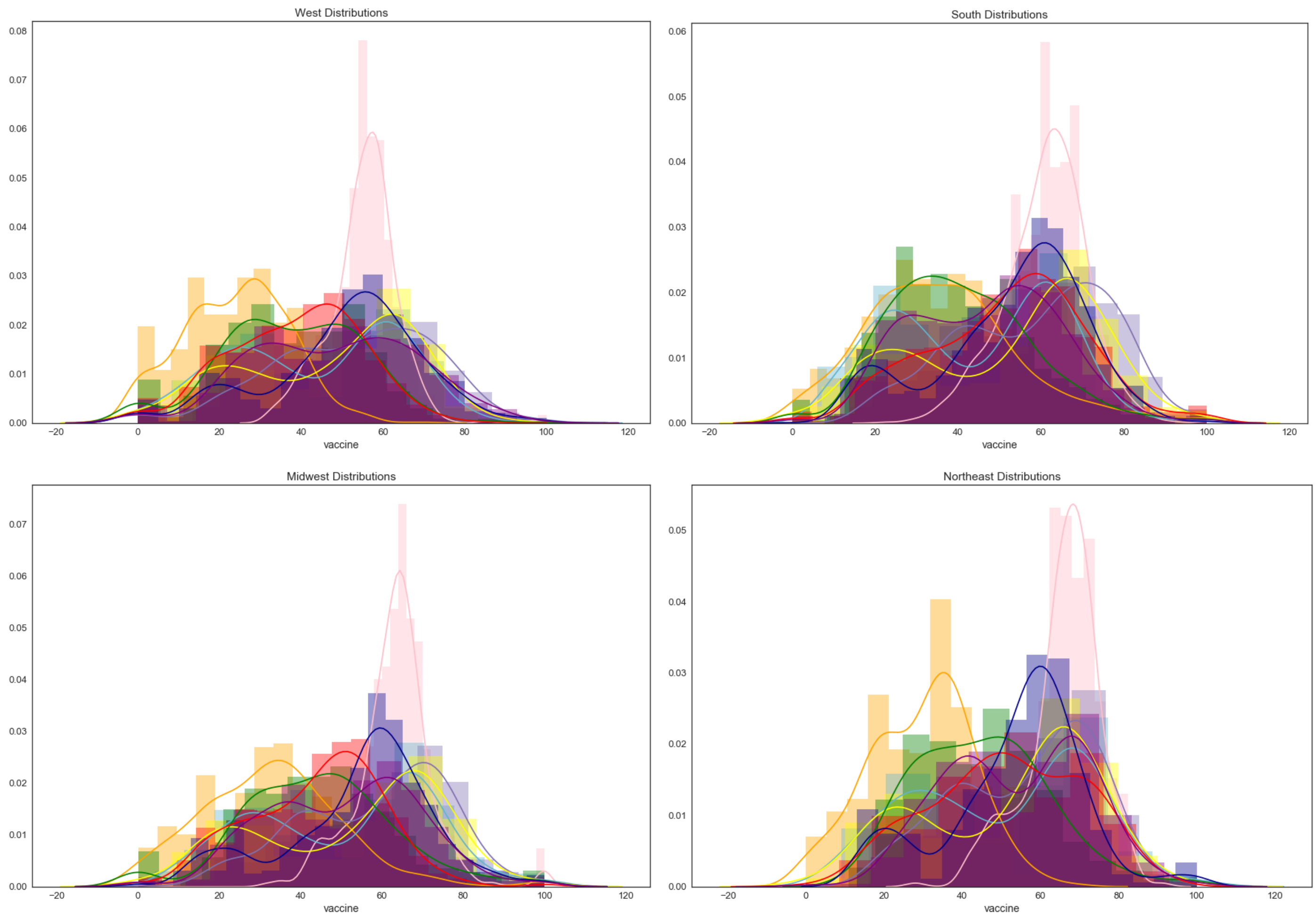
Both methods determined that two clusters would be most efficient for a principal components analysis and the visualization of the clustered data shown above shows a distinct difference in the two main features of the data. For the principal components analysis biplot I normalized the data using z-score normalization. The biplot for principal components analysis with two main features is shown below:



A search with a high PC1 value was most likely a search relating to cancer, rehab, stroke, or cardiovascular issues. These health conditions are more temporary or represent a specific health-related event. On the other hand, a search with a higher PC2 value was probably associated with diabetes, vaccine, obesity, depression, which are all longer-term conditions or events. These longer-term conditions can often have implications for shorter-term conditions, as discussed in the correlations between search categories in research question 2.

Q4. Are the distributions for the searches across each geographic region similar?

For my first steps into investigating the distributions across regions, I looked at histograms for all categories and all years per region:



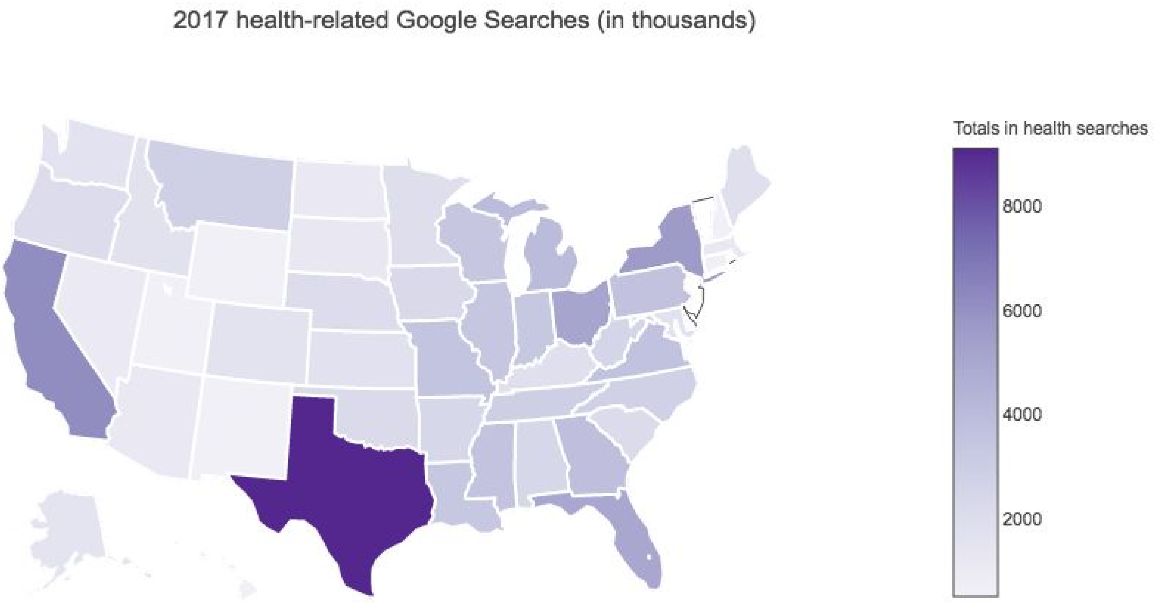
These plots show that the distributions for most search categories look similar across all regions, with a few exceptions. All regions have large frequencies of cancer (pink) and stroke (dark blue) searches. Because the distributions are not normal for all search categories, I used the nonparametric Kruskal-Wallis test to determine if there are significant differences across regions:

```
cancer:  KruskalResult(statistic=385.13220392976967, pvalue=3.6768235189121981e-83)
cardiovascular:  KruskalResult(statistic=189.15151911523492, pvalue=9.3087755348311756e-41)
depression:  KruskalResult(statistic=87.149255607753915, pvalue=8.9691599628896055e-19)
diabetes:  KruskalResult(statistic=39.137865050414327, pvalue=1.6227764464418079e-08)
diarrhea:  KruskalResult(statistic=40.354910255453483, pvalue=8.9602168745664793e-09)
obesity:  KruskalResult(statistic=61.132904534535925, pvalue=3.3664603086020478e-13)
rehab:  KruskalResult(statistic=250.22679939234513, pvalue=5.8448475625123722e-54)
stroke:  KruskalResult(statistic=18.9153136364848, pvalue=0.00028464602323376869)
vaccine:  KruskalResult(statistic=51.54824190169402, pvalue=3.7383589008697845e-11)
```

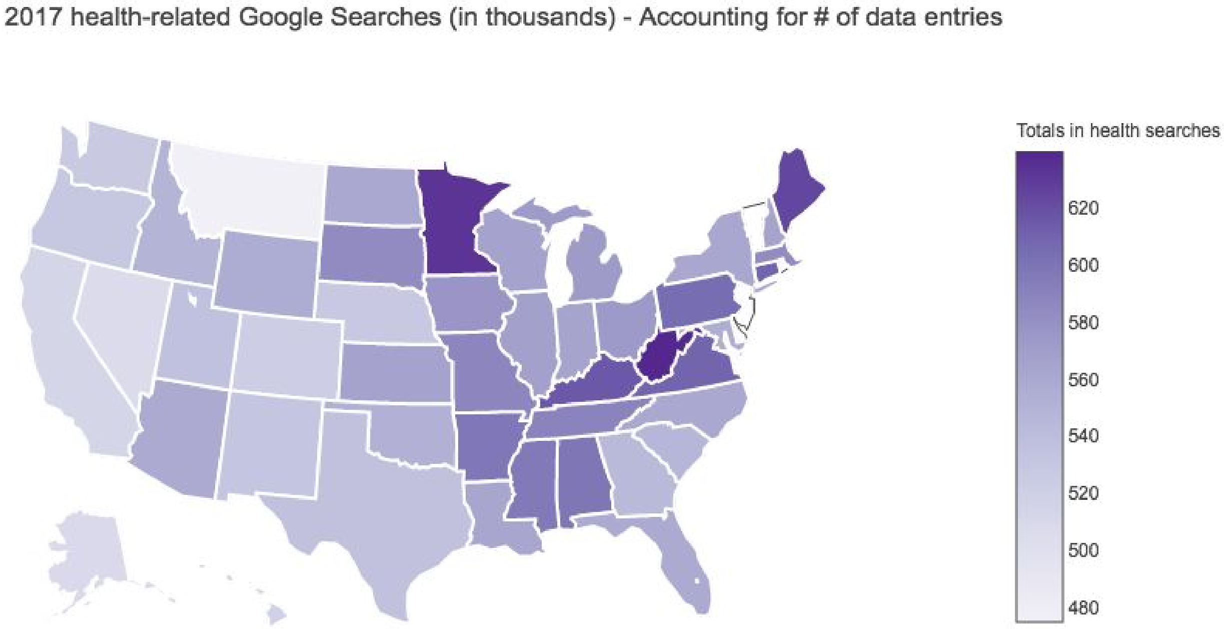
All search categories have p-values much smaller than the 0.05 alpha level, meaning we can reject the null that the distributions for all geographic regions are the same and conclude that the distributions in health related searches is statistically significant. These differences could have a number of causes, but one of the main causes could be in the number of counties that were chosen for this dataset in each region. For example, the northeast data set has 312 entries while the south has 1218 entries. This also shows that the 4 geographic regions don’t divide the U.S. evenly.

Q5. Which state has the most total health searches? What about total searches per category?

To determine which state has the most total health searches, I decided to just use data for 2017, as it’s the most recent data available. After collapsing the data over search category to find the totals, I utilized plotly to visualized the total searches:


2017 health-related Google Searches (in thousands)
Totals in health searches
8000
6000
4000
2000

Here we can see that Texas has the largest number of total health-related searches by far, but I did not controlling for state population. California and Texas have the two largest state populations respectively, which explains why they have the greatest search totals. But, for a more fair analysis, we need to control for population somehow. Because the dataset only has 210 county entries and this isn't necessarily proportional to the number of actual counties or state population in the United States, let's try to normalize the searches by how many county entries for each state:


2017 health-related Google Searches (in thousands) - Accounting for # of data entries
Totals in health searches
620
600
580
560
540
520
500
480

After normalizing for the number of entries in the dataset we can see that the top 3 states by total health-searches are West Virginia, Minnesota and Maine. The analysis that controls for the number of entries in the dataset per state, while having a better spread, still has issues with population size. Because we still don't know how the counties were selected - whether it be by population size, population density, or general popularity - controlling for the number of entries per state is arbitrary. Because of these issues, it’s hard to determine what is causing these states to have more health-related Google searches; which could have interesting implications in the way we understand how people are investigating their own health online.

Future Steps

It’s difficult to determine the exact implications of much of the previous analysis because the dataset was limited to 210 entries and it wasn’t specified why these specific counties were selected. Any future steps should include the full dataset from the Searching for Health application to better determine if the analysis is accurate or meaningful.

One direction I would want to take is to look deeper into the differences between states, regions or other demographics. Again, this would require more data points and could also involve issues of data privacy. But, if we understand how people are currently using Google to become more attuned to their health and the differences between these populations, we can develop better tools for retrieving and understanding this information. Having access to reliable and comprehensive information is key in increasing our awareness about our health. If we find there are significant differences across populations, we should start taking the steps to minimize these disparities to ensure all people have the same chance at maximizing their health.