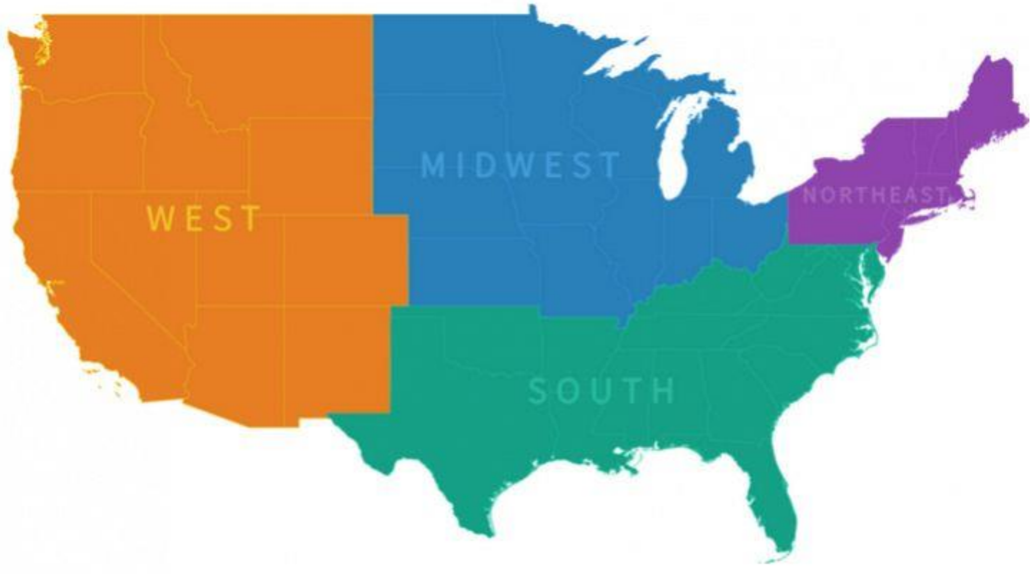
An Analysis of Three of the Leading Causes of Death and their Google Search Results in the US, By Region, Between 2014 and 2018

According to the Centers for Disease Control and Prevention, the leading causes of death as of 2017 in the US are the following:

- Heart disease: 647,457
- Cancer: 599,108
- Accidents (unintentional injuries): 169,936
- Chronic lower respiratory diseases: 160,201
- Stroke (cerebrovascular diseases): 146,383
- Alzheimer's disease: 121,404
- Diabetes: 83,564
- Influenza and pneumonia: 55,672
- Nephritis, nephrotic syndrome, and nephrosis: 50,633
- Intentional self-harm (suicide): 47,173

Source: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Out of these, we will be focusing on Heart Disease, Cancer and Stroke for our data, and will be performing our analysis on a regional basis. We have divided the 50 states (plus the District of Columbia) into 4 regions, as illustrated in the following image:



Source: https://www.worldatlas.com/articles/the-regions-of-the-united-states.html

Our Goal

The main goal of our analysis is to identify trends and patterns in the data, and the significance of these trends. Heart Disease, Cancer and Stroke were chosen as those were the three leading causes of death that are also tied to common google searches of health issues, which allows us to incorporate these searches into our analysis. In order to streamline our report, we are going to try to answer a handful of questions that would yield the most interesting results.

How does each region compare in terms of mortality rates for each cause of death? How about for google searches?

Are there any outliers for any of the causes of death in terms of mortality rates and google searches in a given region or in a given year?

Is there an overall trend in terms of mortality rates for each region over time? Is there a trend for google searches? If so, what is the significance of these trends?

Is there are relationship between overall deaths and searches over time by region?

The Data

All of the data used in this report comes from two sources: The Centers for Disease Control and Prevention (CDC), and an implementation of the Google Search Trends API. Because the data provided by the CDC is limited to 2005 to 2018, we have restricted the scope of our sample data to the years 2014-2018. Here is a sample table with mortality rates for cancer for the year 2018:

YEAR	STATE	RATE
2018	AL	170.4
2018	AK	141.5
2018	AZ	131.9
2018	AR	168.8
2018	CA	135
2018	CO	127.6
2018	CT	134.1
2018	DE	159.4

Source: https://www.cdc.gov/nchs/pressroom/sosmap/cancer_mortality/cancer.htm

These values represent the number of deaths per 100,000 total population, as per the CDC. We used these values for the sake of readability.

Next, we have a sample table with data from google searches for cancer in 2018:

```
+------+--------------------+
|cancer|               State|
+------+--------------------+
|    90|             Alabama|
|    72|              Alaska|
|    86|             Arizona|
|    87|            Arkansas|
|    78|          California|
|    75|            Colorado|
|    96|         Connecticut|
|    91|            Delaware|
|    82|District of Columbia|
```

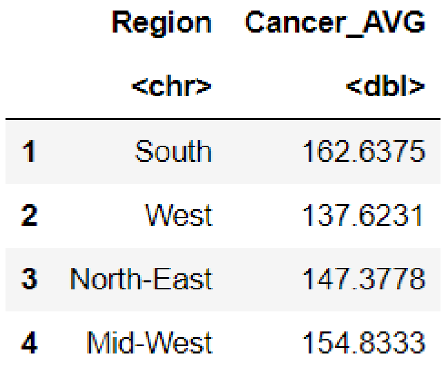
Source: Google Trends API with pytrends

Note that these numbers are not the actual number of google searches for cancer in 2018 but are instead the values reduced to a scale of 100. This is directly done by the Google Search Trends API and improves the readability of the data.

Another thing to note is that the values for search trends and mortality rates are scaled differently, however we are only interested in the relationship between regions and trends over time, so the two categories of values are not going to be compared directly.

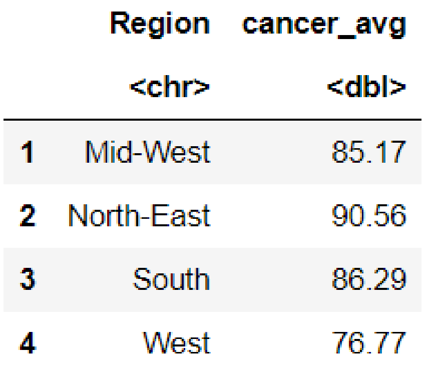
Transformations

The next step in our analysis is to perform transformations of our data. For this, we grouped the data by the 4 regions as we defined them earlier and averaged out the values. Here is a sample table for mortality rates for cancer in 2018:



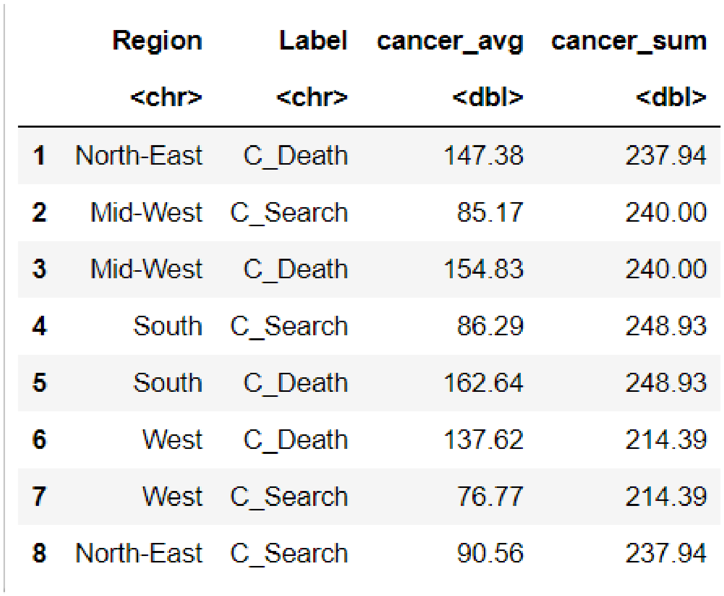
	Region	Cancer_AVG
	<chr>	<dbl>
1	South	162.6375
2	West	137.6231
3	North-East	147.3778
4	Mid-West	154.8333

And we have a sample table for google searches of cancer in 2018 grouped by region:



	Region	cancer_avg
	<chr>	<dbl>
1	Mid-West	85.17
2	North-East	90.56
3	South	86.29
4	West	76.77

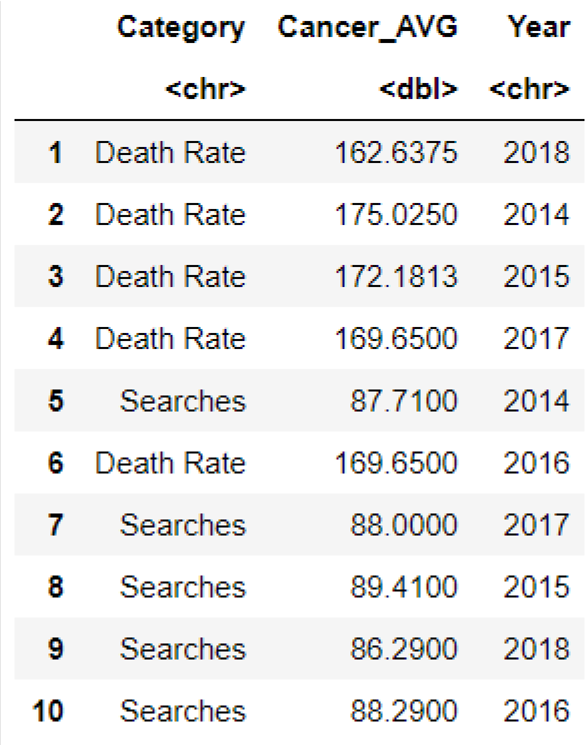
We will do another transformation on our data where we will group both searches and death rates together and categorize them accordingly. This will allow us to visualize both kinds values in on chart.



	Region	Label	cancer_avg	cancer_sum
	<chr>	<chr>	<dbl>	<dbl>
1	North-East	C_Death	147.38	237.94
2	Mid-West	C_Search	85.17	240.00
3	Mid-West	C_Death	154.83	240.00
4	South	C_Search	86.29	248.93
5	South	C_Death	162.64	248.93
6	West	C_Death	137.62	214.39
7	West	C_Search	76.77	214.39
8	North-East	C_Search	90.56	237.94

Note that the cancer_sum field is strictly to make mapping the data to a graph form easier.

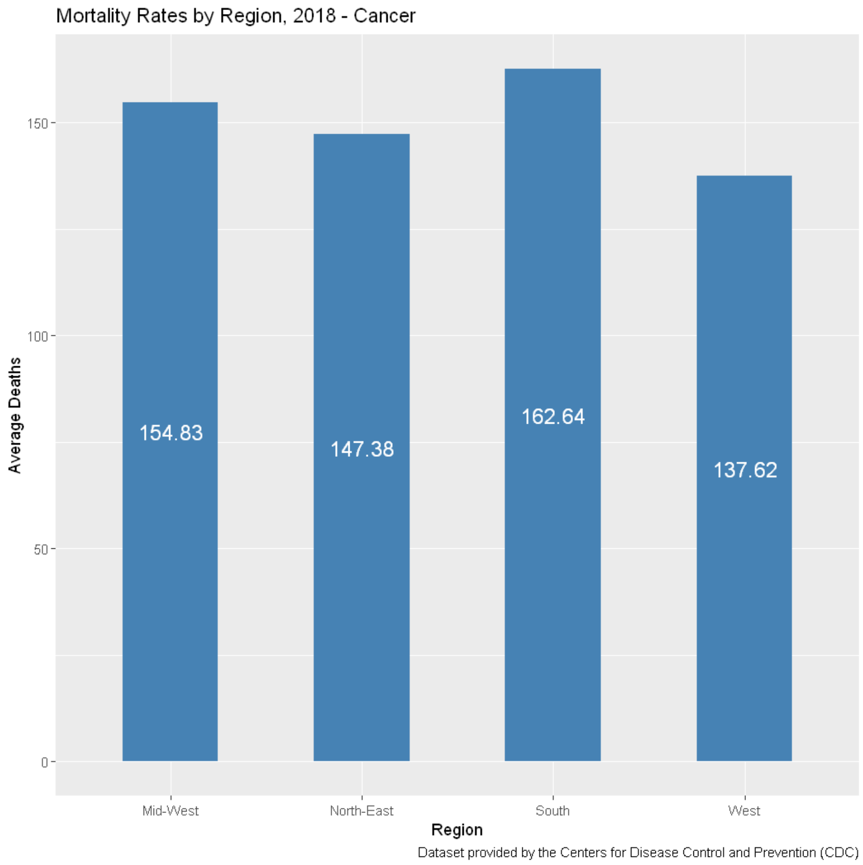
We perform one final transformation on our data, grouping all values for both searches and death rates across all 5 years by region. Here is a sample table the south:



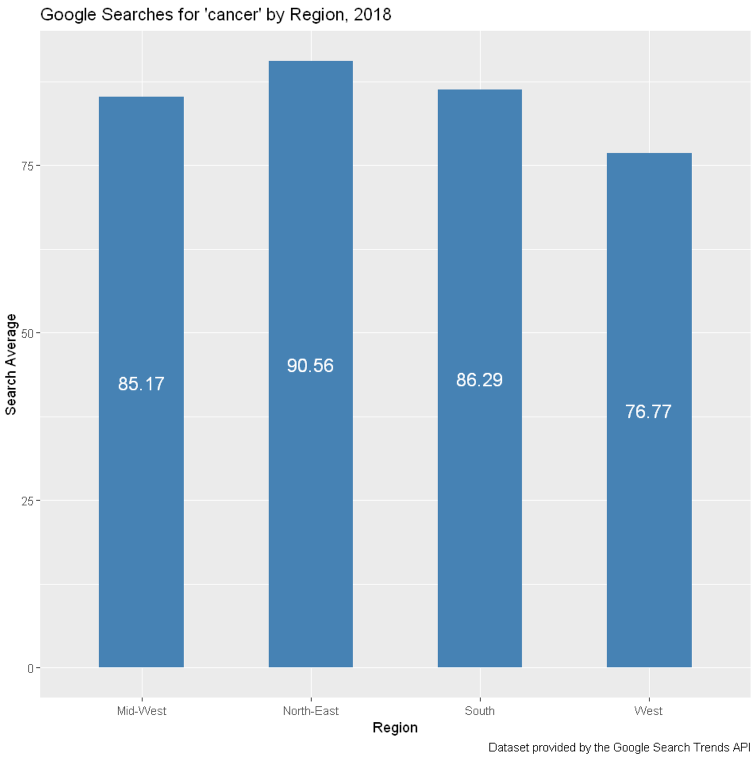
	Category	Cancer_AVG	Year
	<chr>	<dbl>	<chr>
1	Death Rate	162.6375	2018
2	Death Rate	175.0250	2014
3	Death Rate	172.1813	2015
4	Death Rate	169.6500	2017
5	Searches	87.7100	2014
6	Death Rate	169.6500	2016
7	Searches	88.0000	2017
8	Searches	89.4100	2015
9	Searches	86.2900	2018
10	Searches	88.2900	2016

Graphs

Now we can translate our data into graphs so that we can visualize it better and observe the relationship between regions and any possible trends. First, we create a bar chart for each year showing mortality rates and google searches per region:

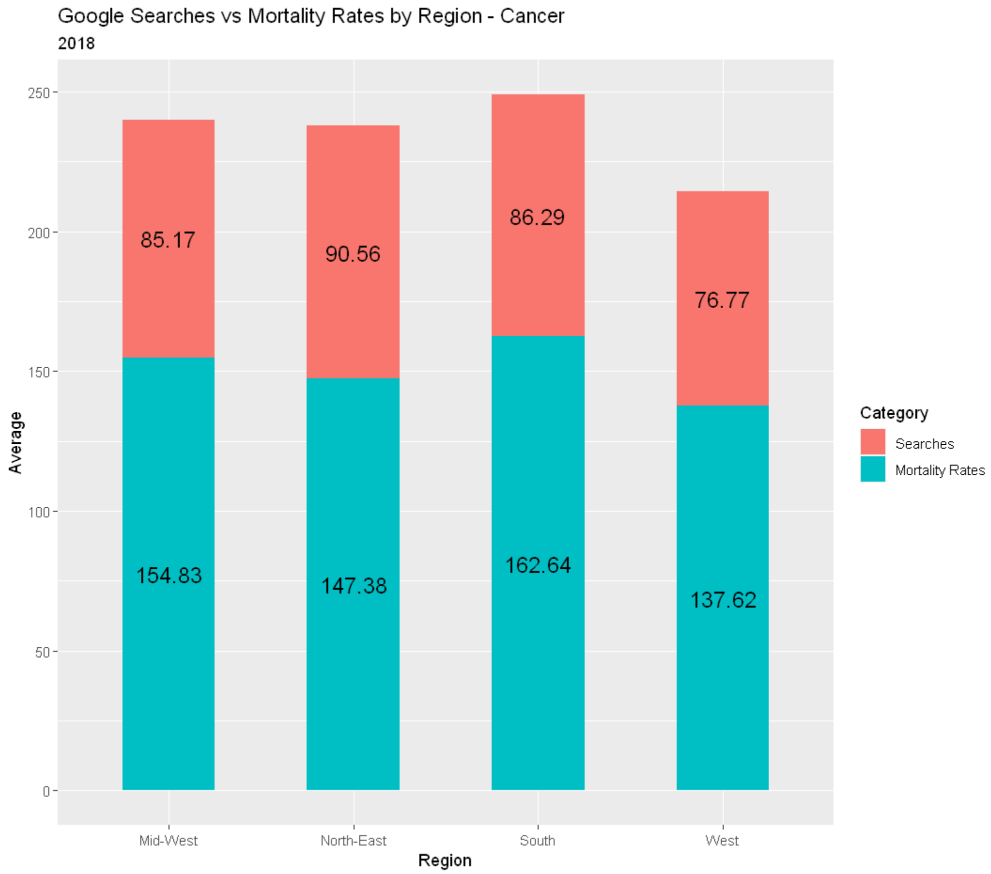


And here is the same chart for searches:



We can already begin to observe certain differences between regions, such as the South having the highest morality rate for cancer in 2018, or the west having both the lowest death rate and number of searches.

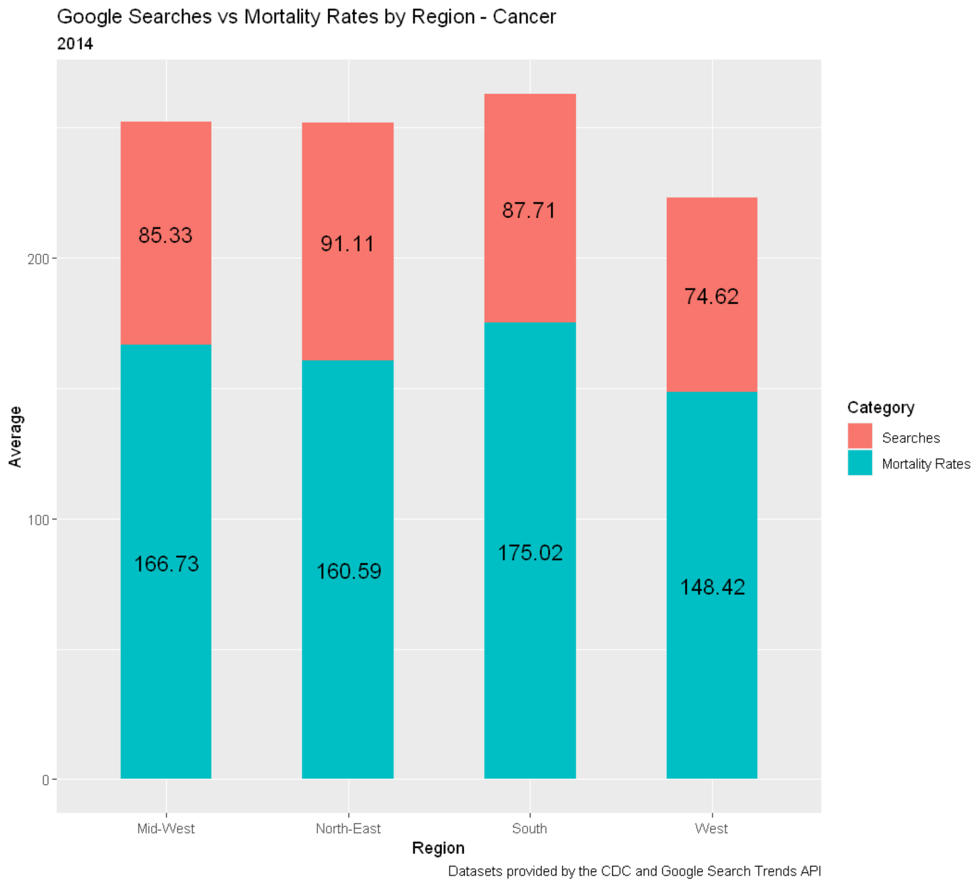
After creating a graph for both death rates and searches for each year, we then group both categories together in one graph. This allows to better compare the proportion of deaths to searches between each region.



Datasets provided by the CDC and Google Search Trends API

Based on this graph we can observe that the proportion of death rates to searches seems to be mostly uniform across all regions. Something of note is that the North-East has the highest number of searches along with the second lowest death rate.

Now let’s take a look at the same graph for 2014, to see if we can notice any patterns.

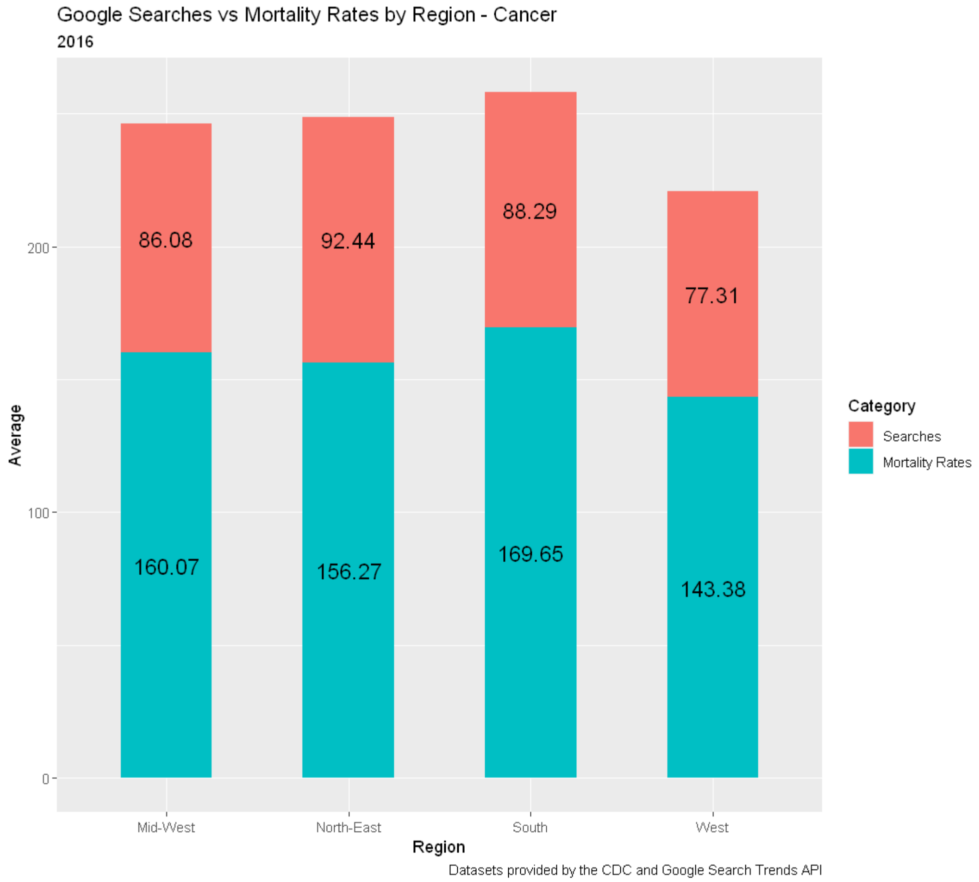

Google Searches vs Mortality Rates by Region - Cancer

2014

Datasets provided by the CDC and Google Search Trends API

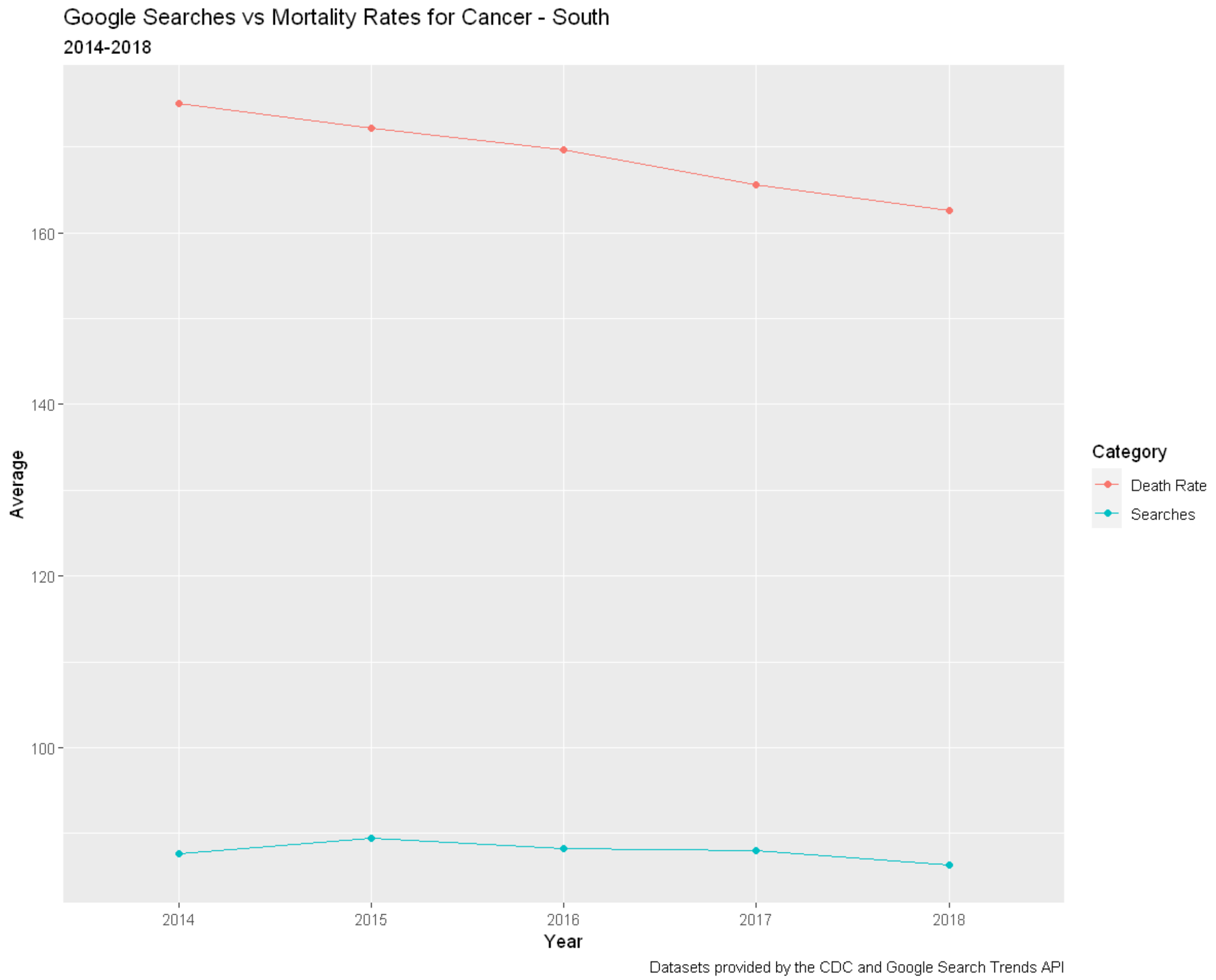
Some patterns are immediately apparent: the North-East still has the highest number of searches, the West has the lowest value in both categories, and the South still has the highest death rate. There are also appears to be both a decrease in death rates between 2014 and 2018 across all regions, while there is a slight decrease in searches in the North-East and South.

We will observe one more graph for 2016, to determine whether these are consistent patterns or not.

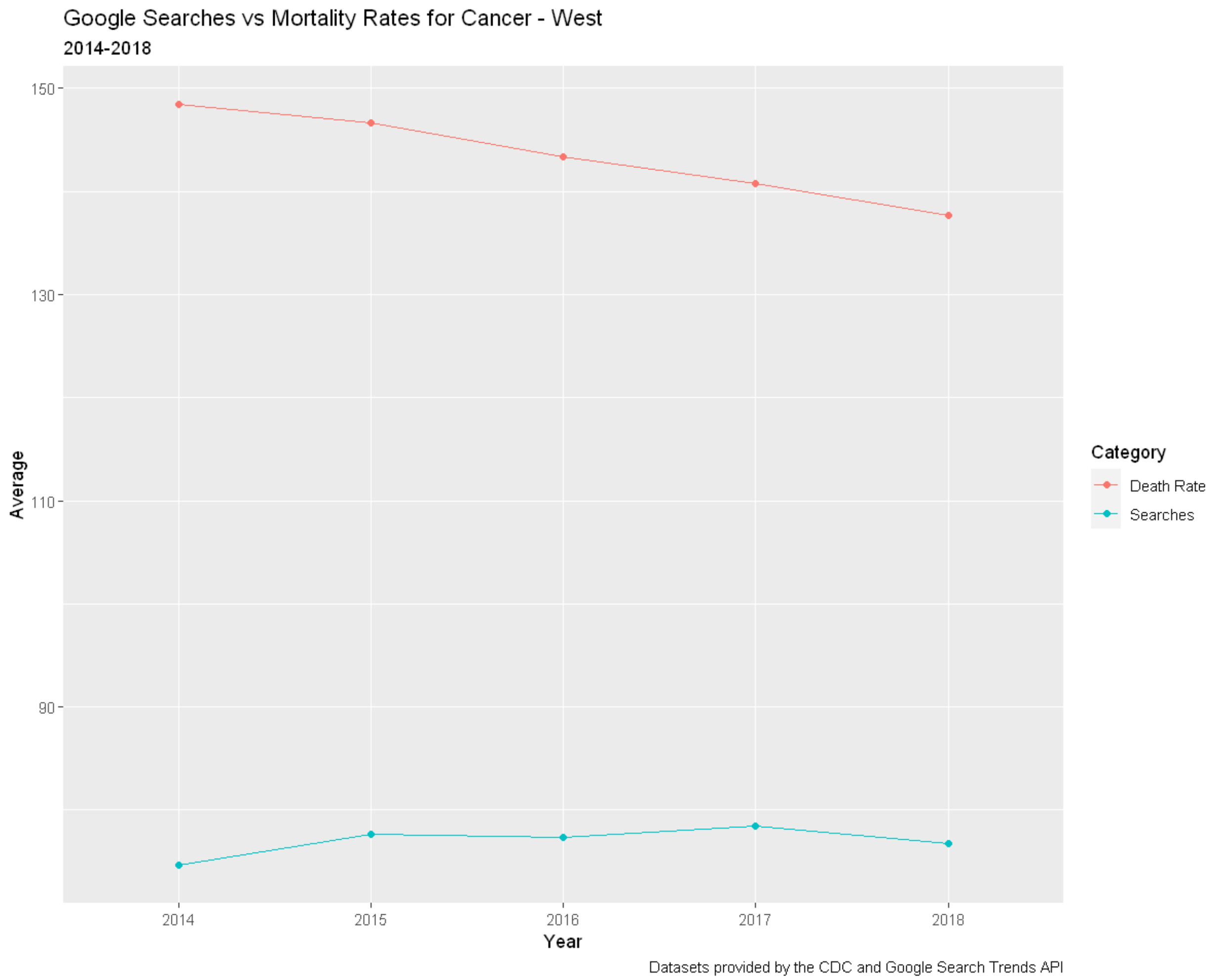


And we notice the same patterns that appeared for the years 2014 and 2018.

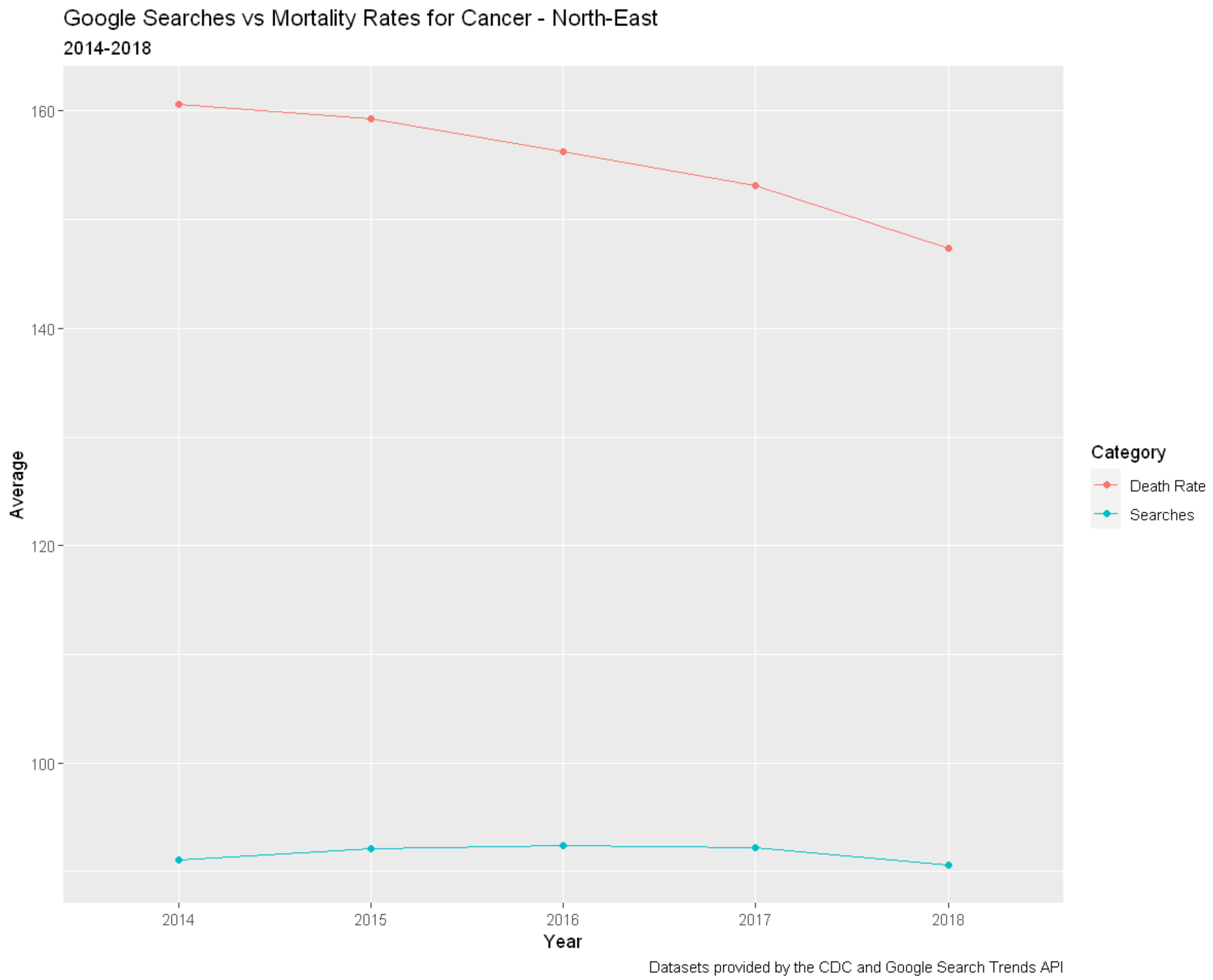
Finally, we will analyze a series of line graphs that show the progress of the death rates and google searches over time on a regional basis.



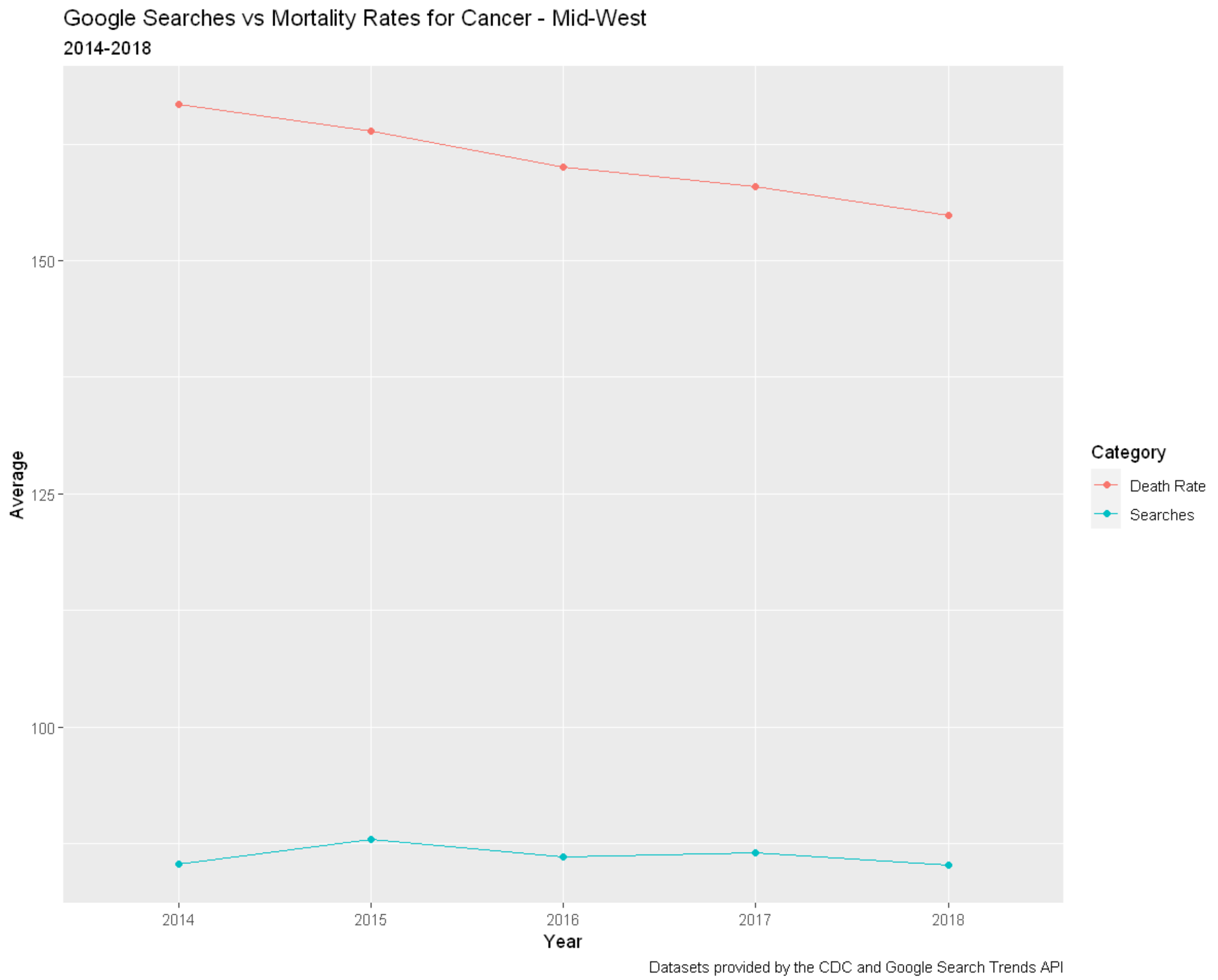
Now we can get a better look at the trends for each category for every region. The most obvious trend is that death rates for cancer have been consistently decreasing between 2014 and 2018. Searches for cancer seem relatively stable, with a slight uptick in 2015. However, it seems that in 2018 searches were lower than in each of the previous 4 years.



We can observe a similar trend in the West, except that searches seem to have increased more than they did in the South.



The same trends seem to appear in the North-East as well, however there seems to be a steeper decrease in death rates than in the other regions.

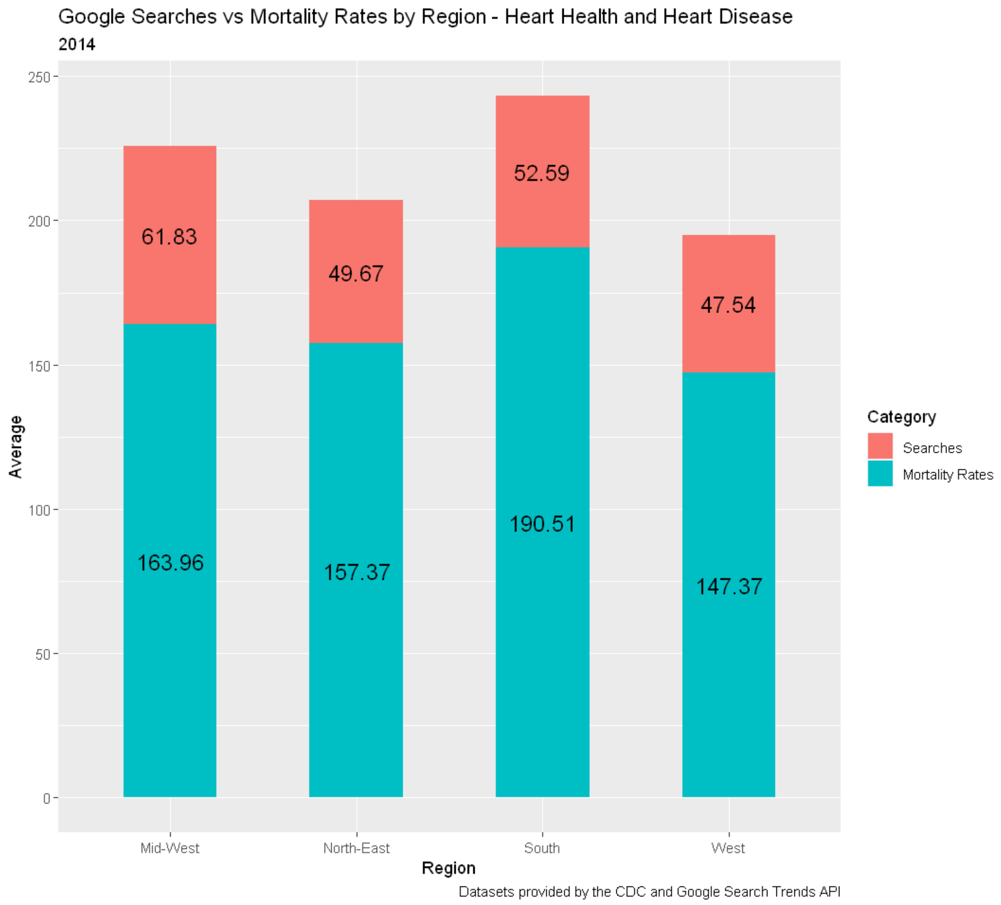


There seems to have been a small spike in searches in the Mid-West in 2015, but afterwards there is a downward trend.

We can already draw some conclusions from the analyses we have performed so far, but first we are going to analyze the other two death causes mentioned, heart disease and stroke, and see how they compare.

Heart Disease

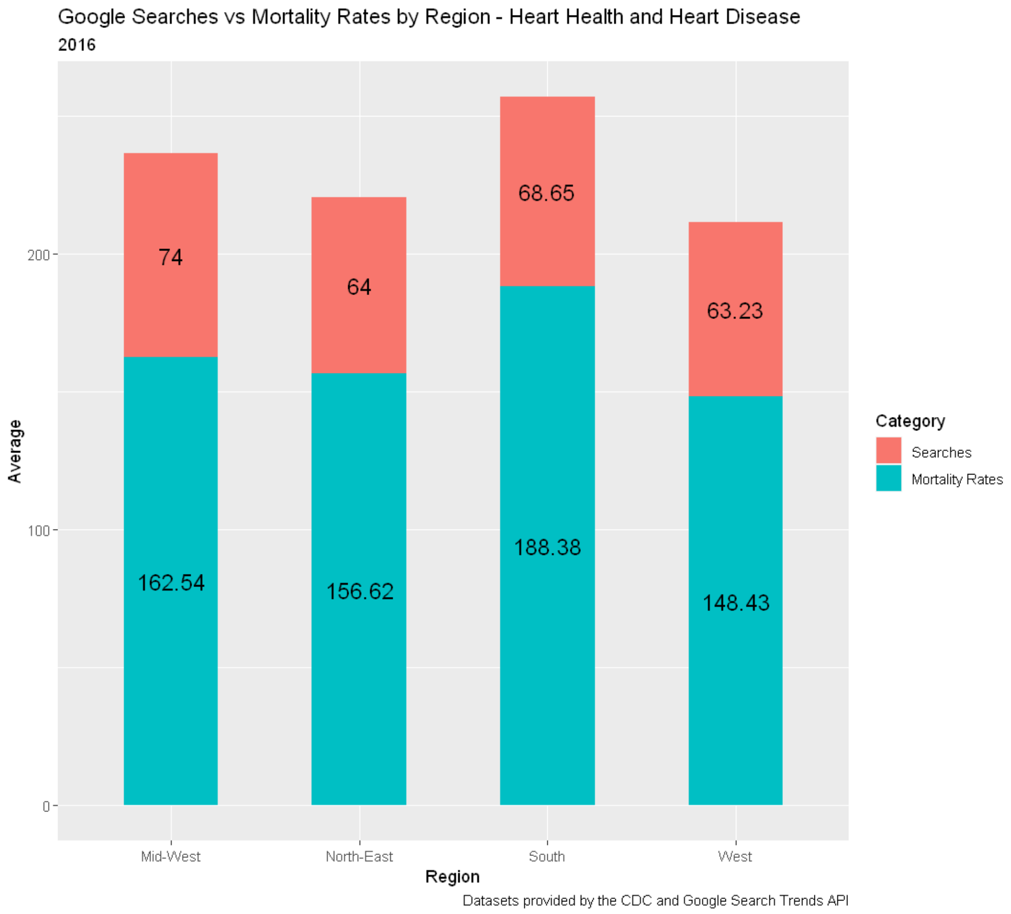
For heart disease, we will be taking a look at the stacked bar charts for 3 separate years. Note that we chose “heart health” as our google search term because it best reflects people’s concern about heart-related health issues.


Google Searches vs Mortality Rates by Region - Heart Health and Heart Disease

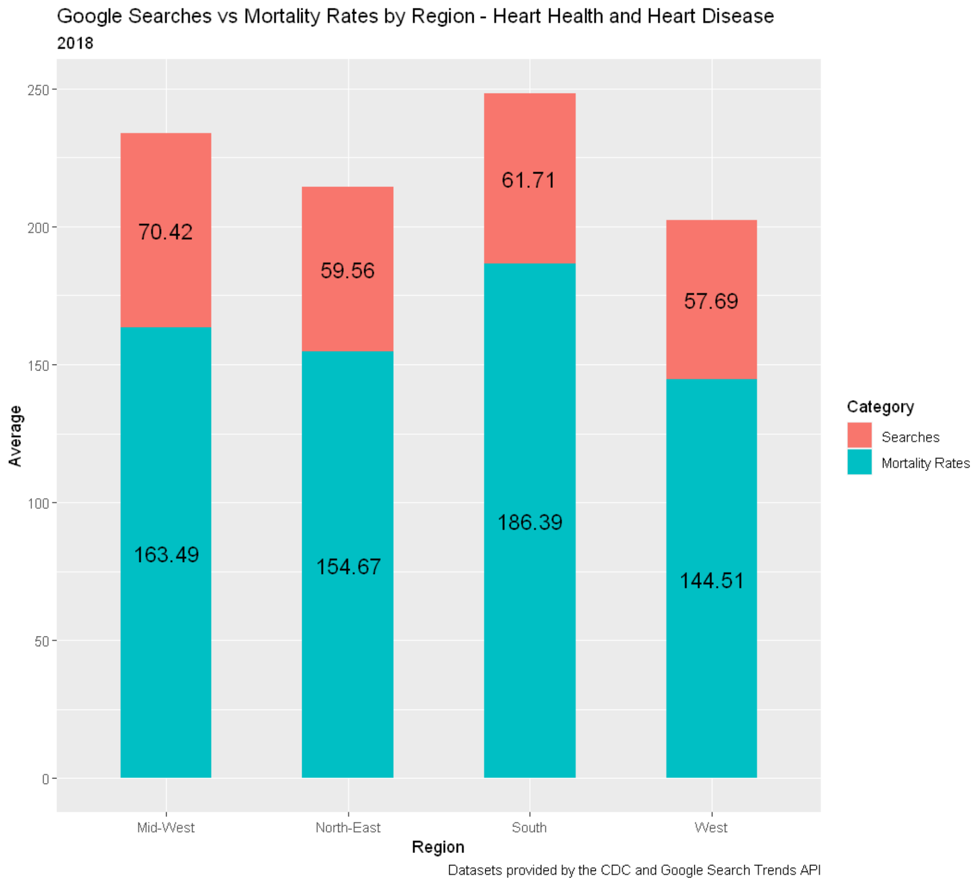
2014

Datasets provided by the CDC and Google Search Trends API

We can already observe some stark contrasts to the charts we analyzed for cancer, most notably the much lower search count. The South seems to have a significantly higher death rate related to heart disease than every other region. However, a pattern we had already observed in the cancer bar charts seems to be present here as well. The West still has both the lowest death rate and search number when compared to the other regions. Also notable is the fact that the Mid-West has the highest amount of searches, which may indicate a higher level of concern for heart related issues in the region.

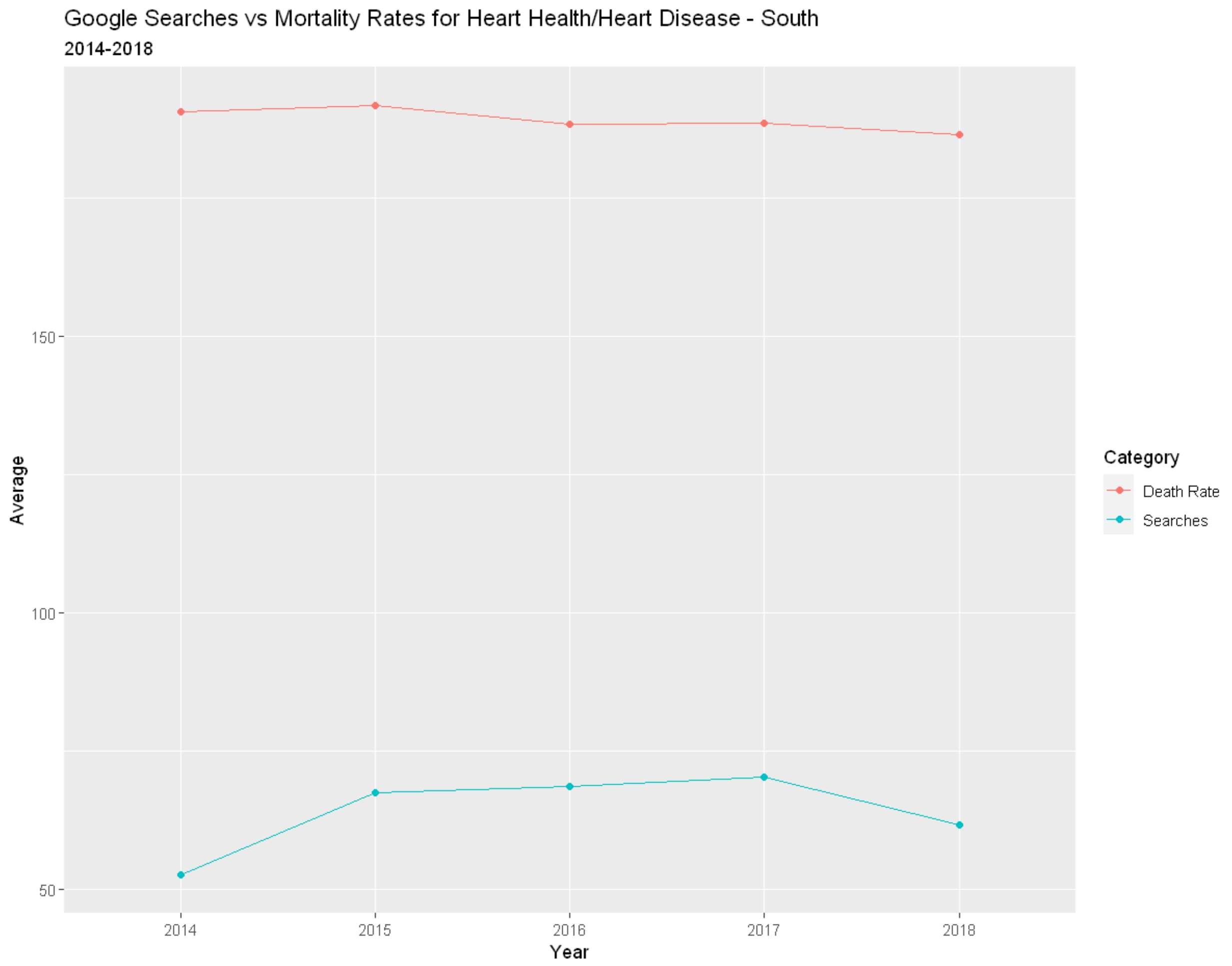


In 2016, we can observe a major increase in searches for heart health across all regions, possibly indicating and increasing interest in the issue. Meanwhile, death rates seem relatively stable, with only minor decreases in 3 regions and a slight increase in the other.



The pattern for the death rates seems to continue for 2018, with the South still having the highest count. However, there appears to be a decrease in searches for all regions.

Now we can observe hoe the data changes over time for each region.

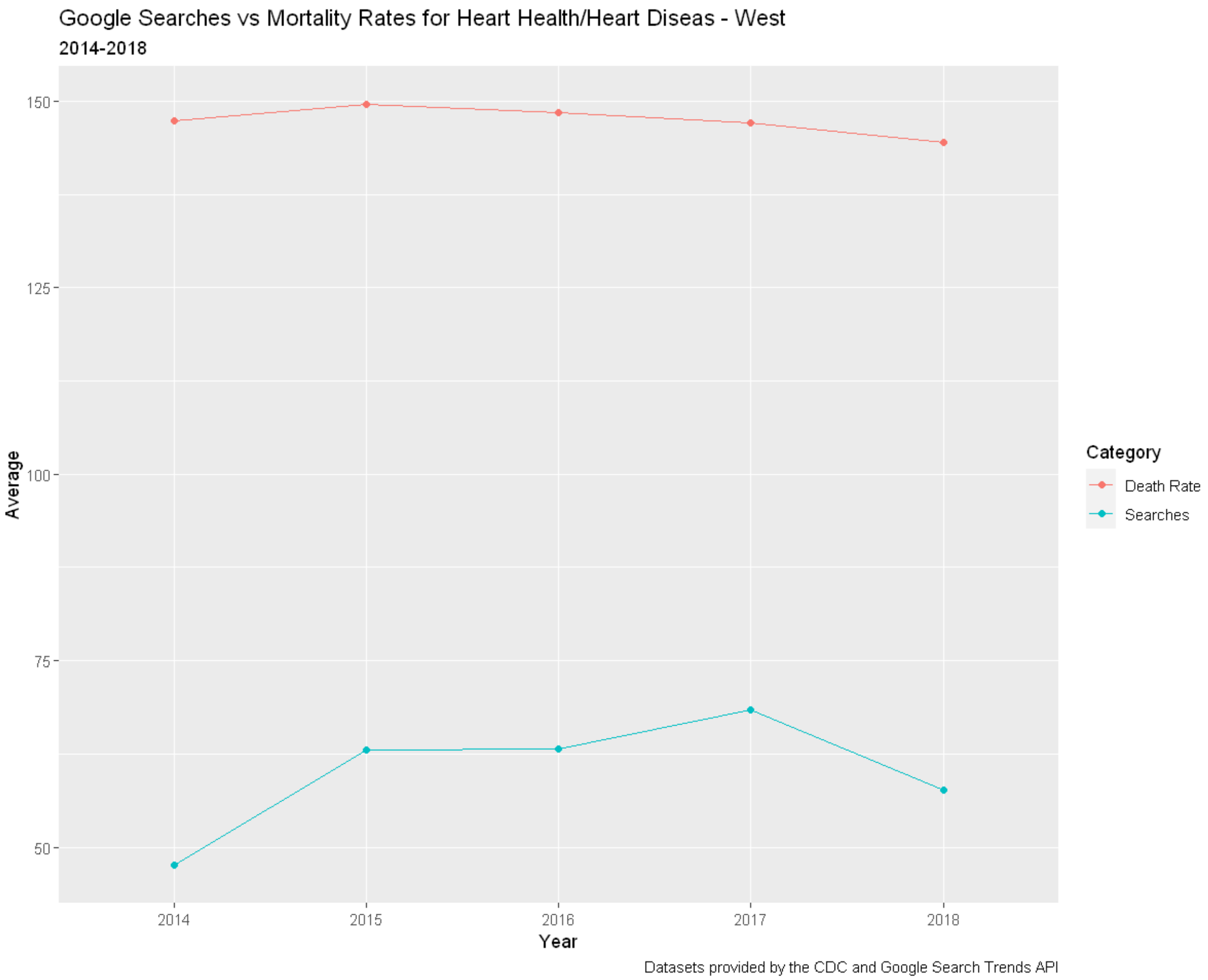


Google Searches vs Mortality Rates for Heart Health/Heart Disease - South

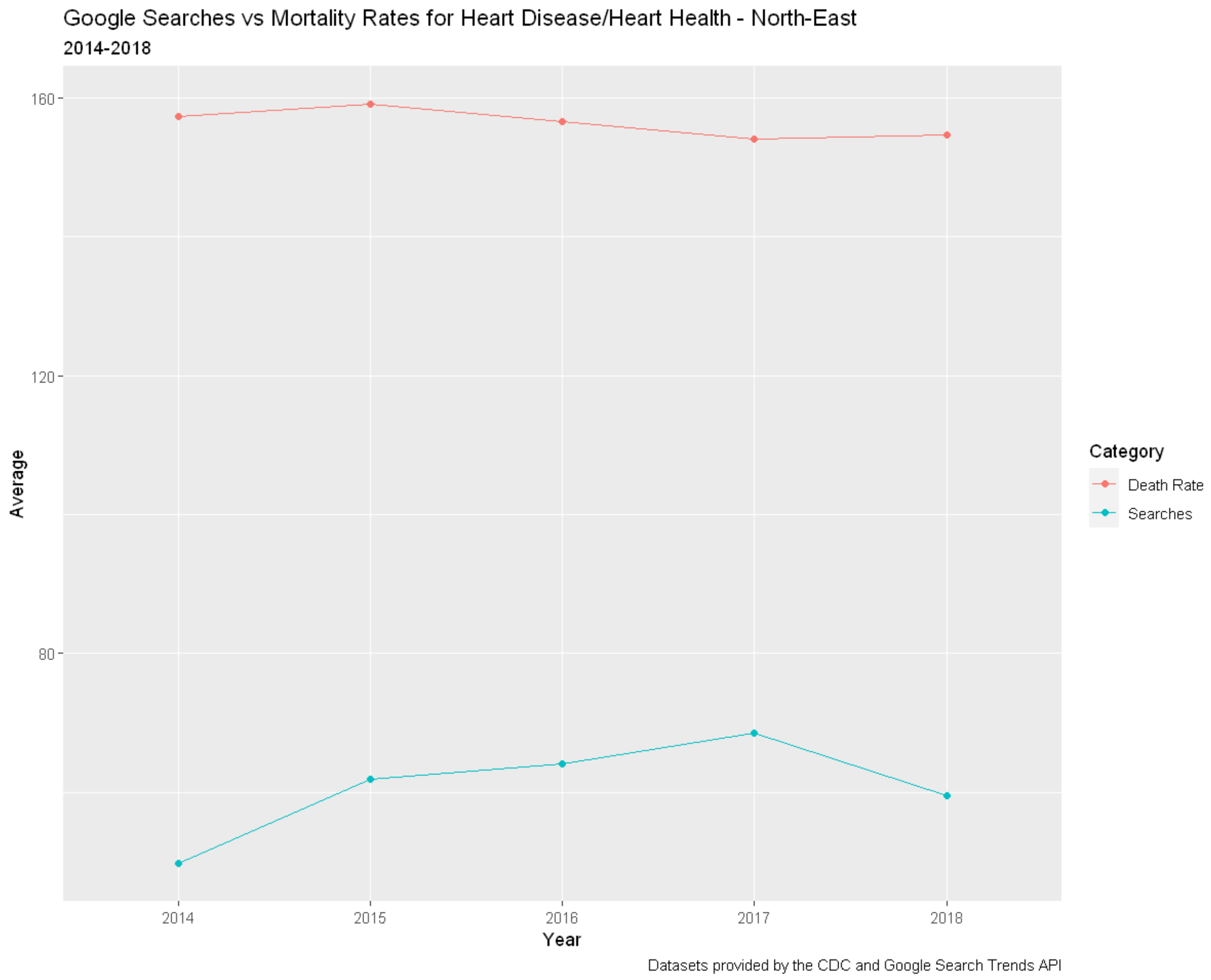
2014-2018

Datasets provided by the CDC and Google Search Trends API

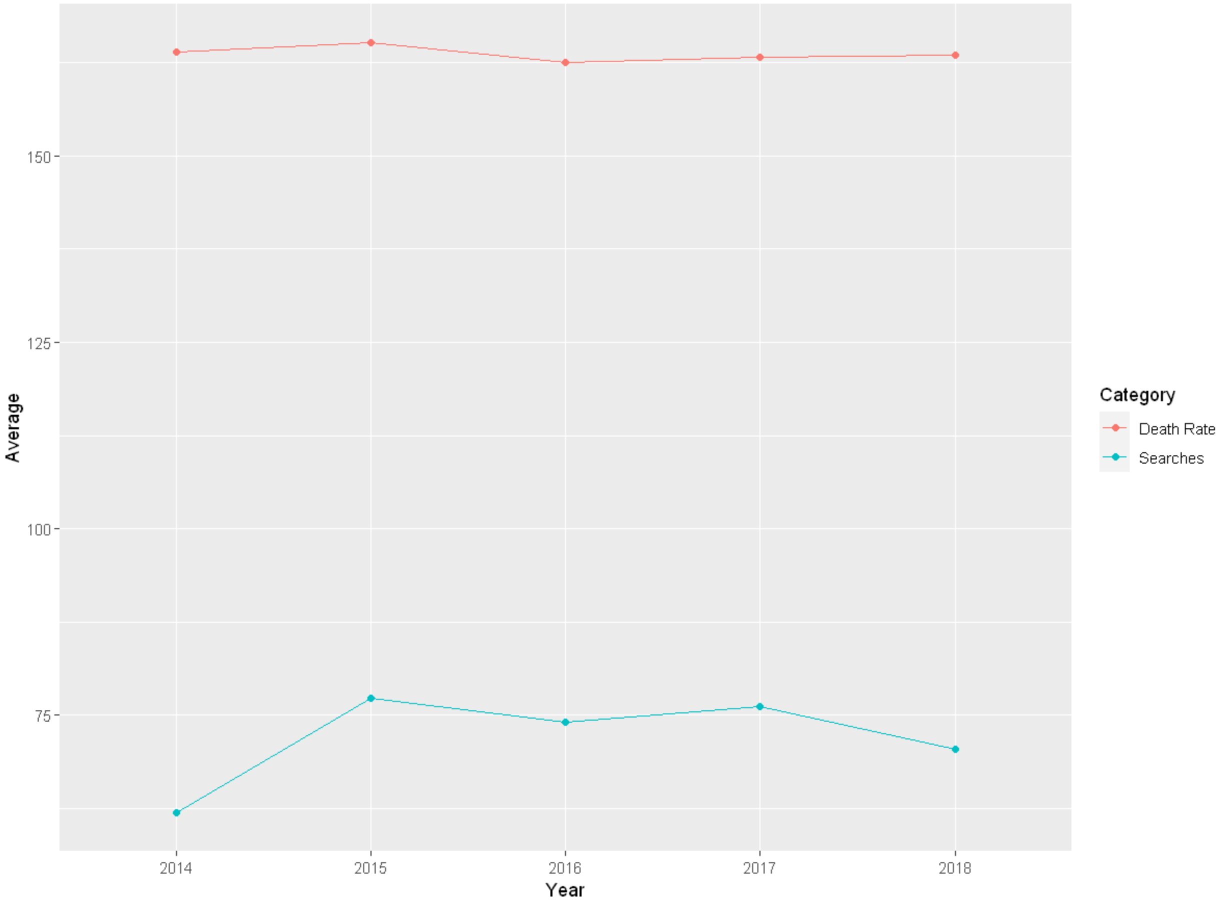
This graph appears to tell a different story than what we saw for cancer death rates and searches. The death rates for heart disease show very little change over time, while searches appear to have spiked between the years 2015 and 2017, only to see a sharp decrease in 2018.



The West seems to maintain the same pattern as shown in the South, except that both the spike and decrease in searches seems to be more pronounced.



The pattern for searches in the North-East seems very similar to that of the West, however there is a slightly more noticeable decrease in death rates over the years.



Google Searches vs Mortality Rates for Heart Disease/Heart Health - Mid-West

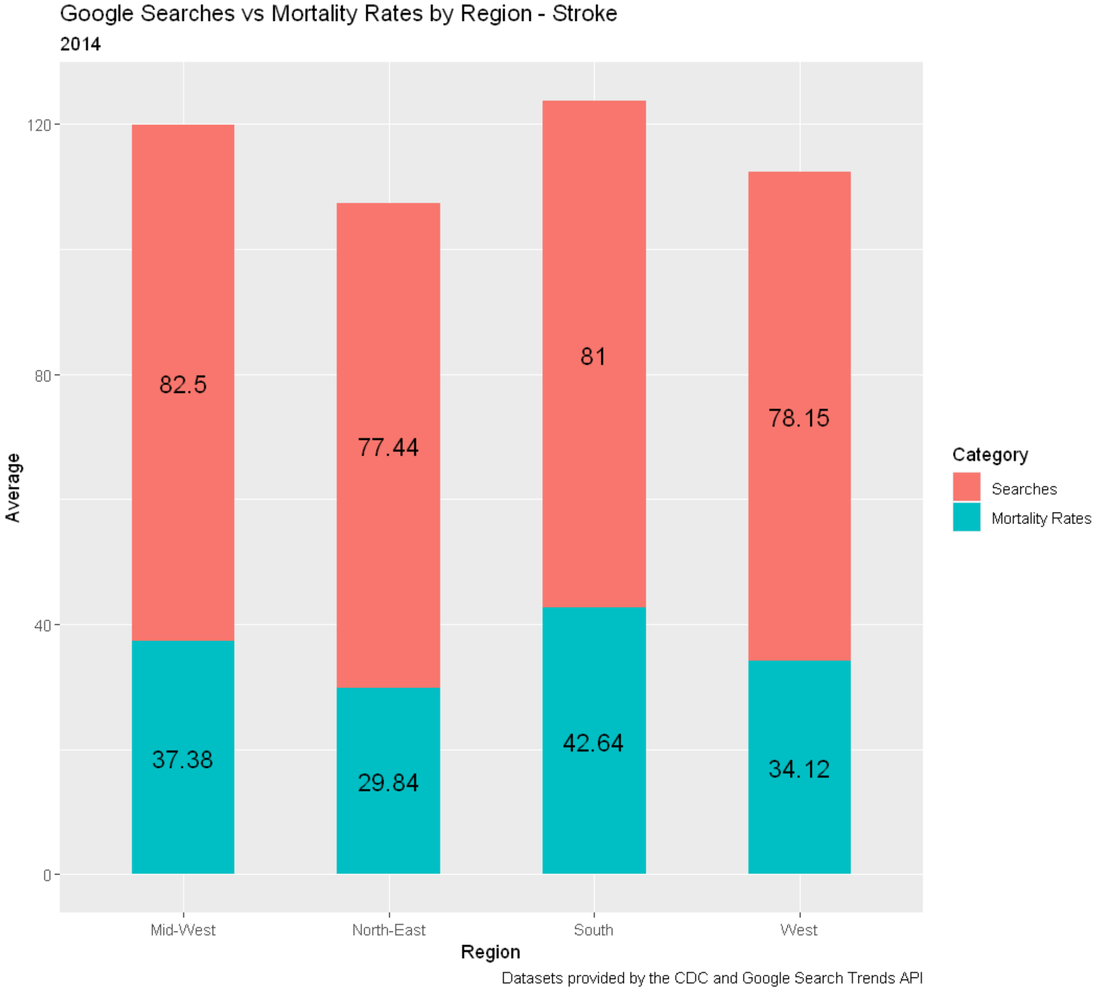
2014-2018

Datasets provided by the CDC and Google Search Trends API

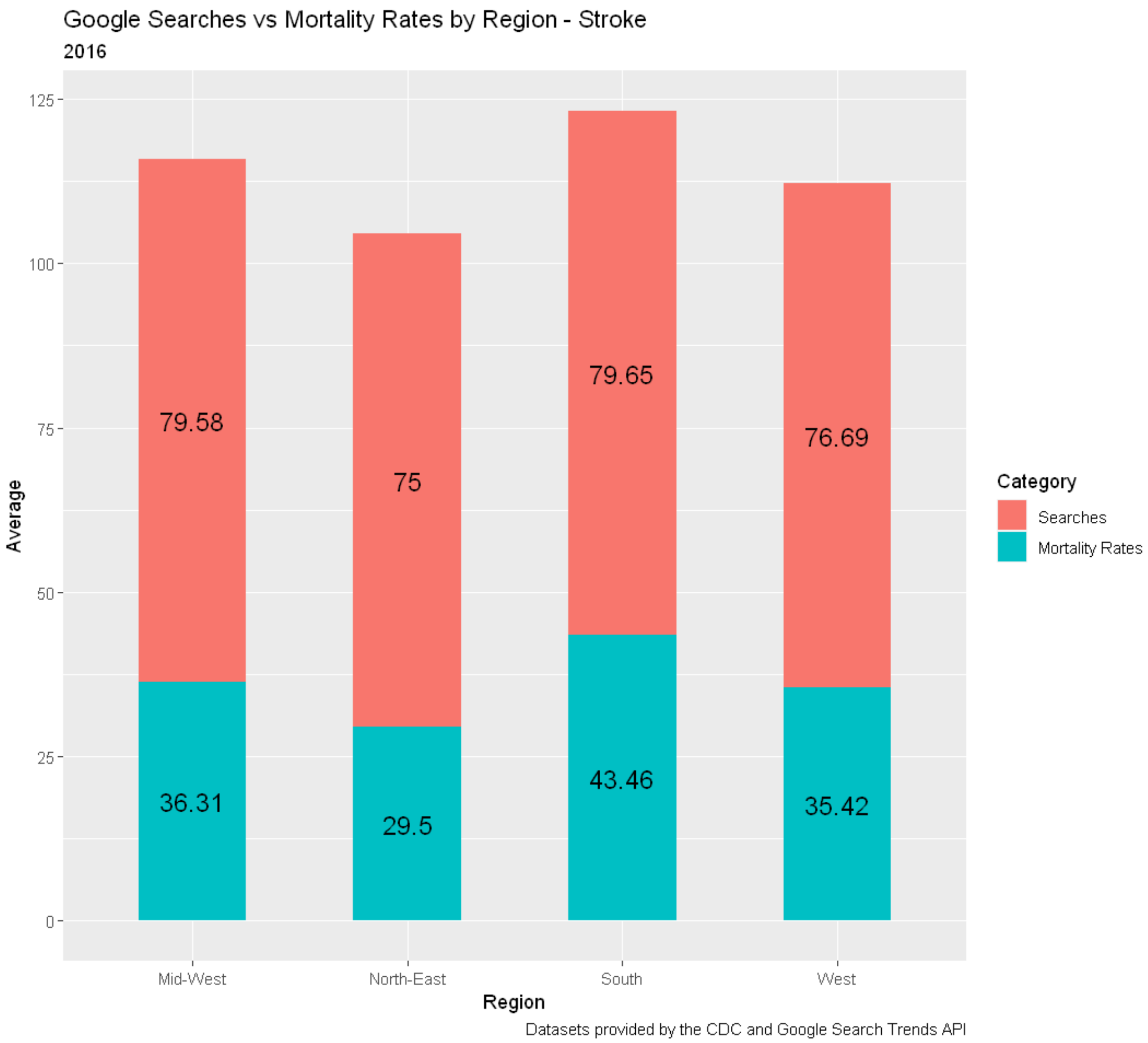
Searches in the Mid-West appear to have spiked back in 2015, and their decline in 2018 is seemingly less pronounced when compared to the rest of the regions.

Lastly, we will take a look at the data for strokes before we make any conclusions about our findings.

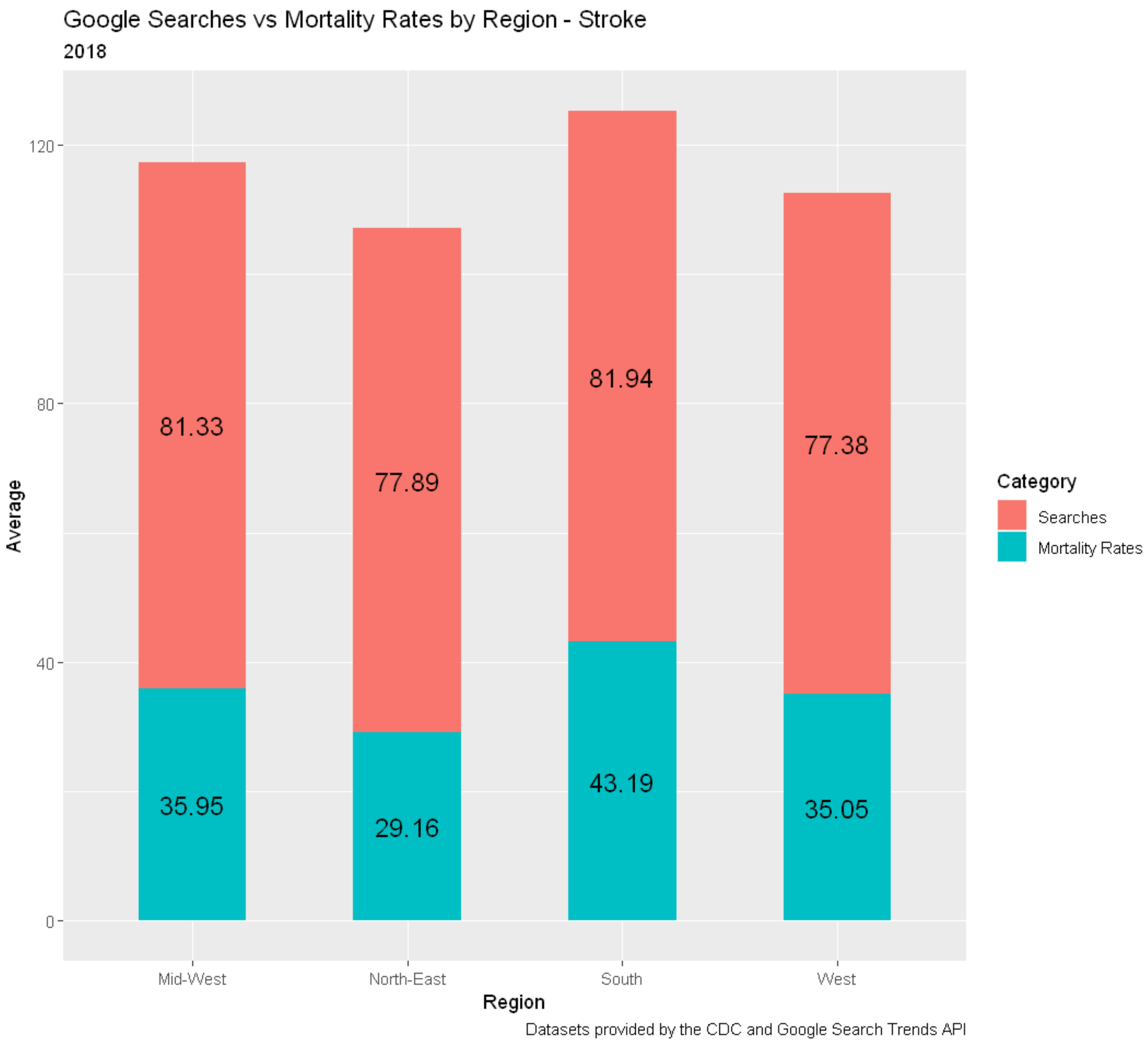
Stroke



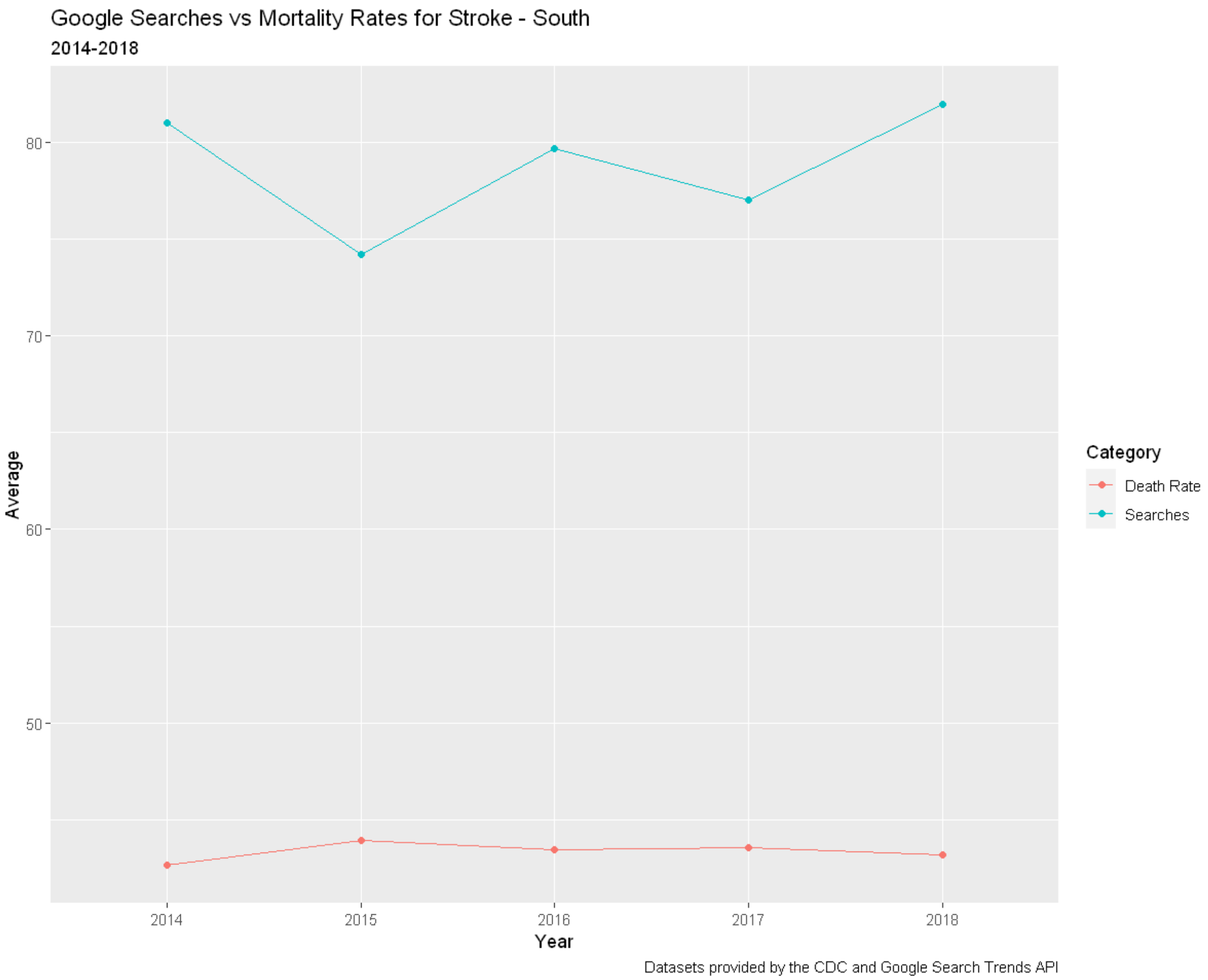
There is a stark contrast in the stacked chart for stroke in comparison to previous charts. This time, the North-East appears to have the lowest mortality rate, while the South is again in the lead. The North-East also has the lowest search count, which, when combined with the lower mortality rates, may indicate that strokes are a less concerning health issue in the region. However, further analysis is need before drawing our conclusions.



For 2016, we can observe the same pattern for the North-East, while the South now has both the highest death rate and highest number of searches, which my indicate a higher level concern on the issue In the region.

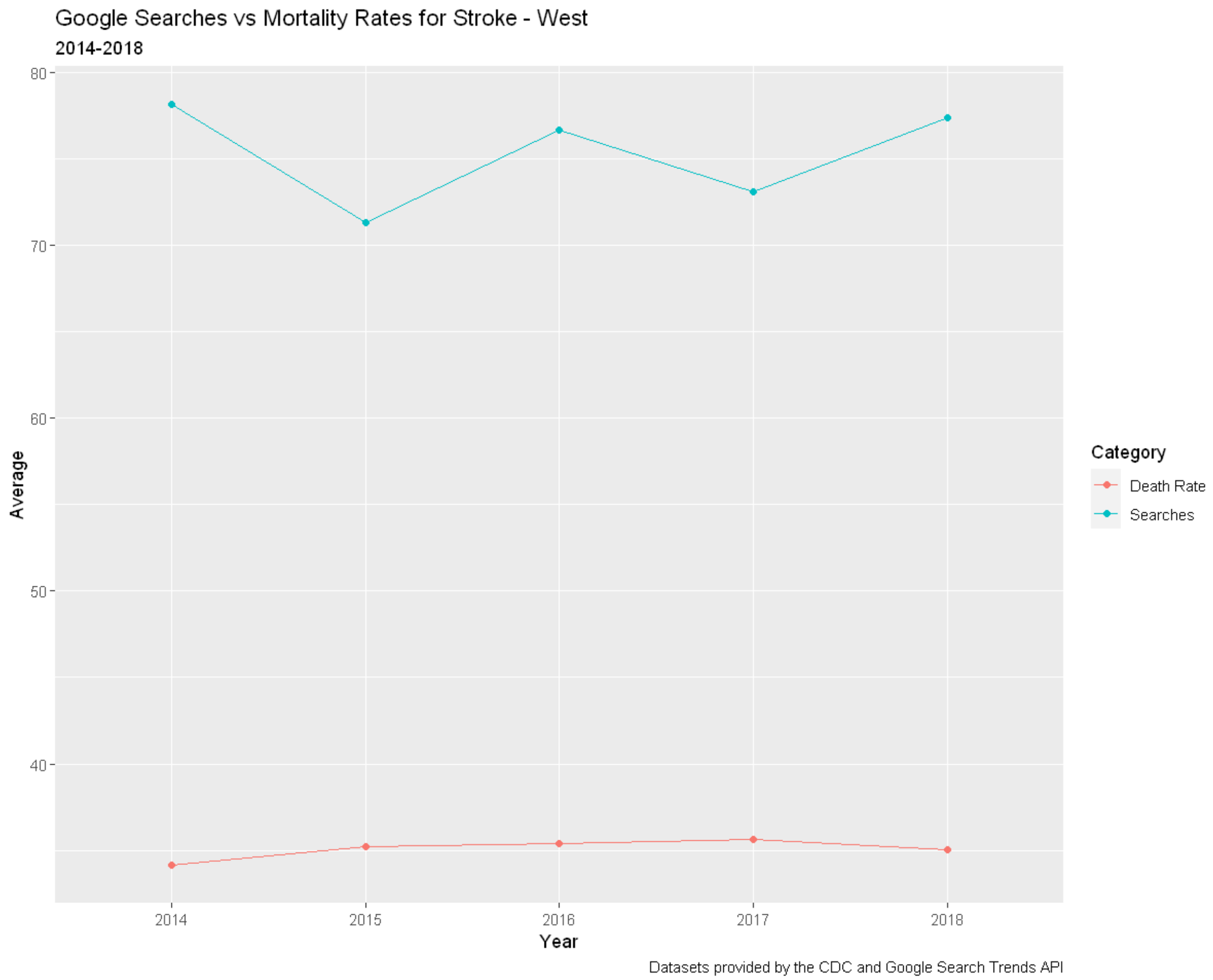


The number of searches show a slight increase from 2016 to 2018, however the mortality rates appear to be stable.

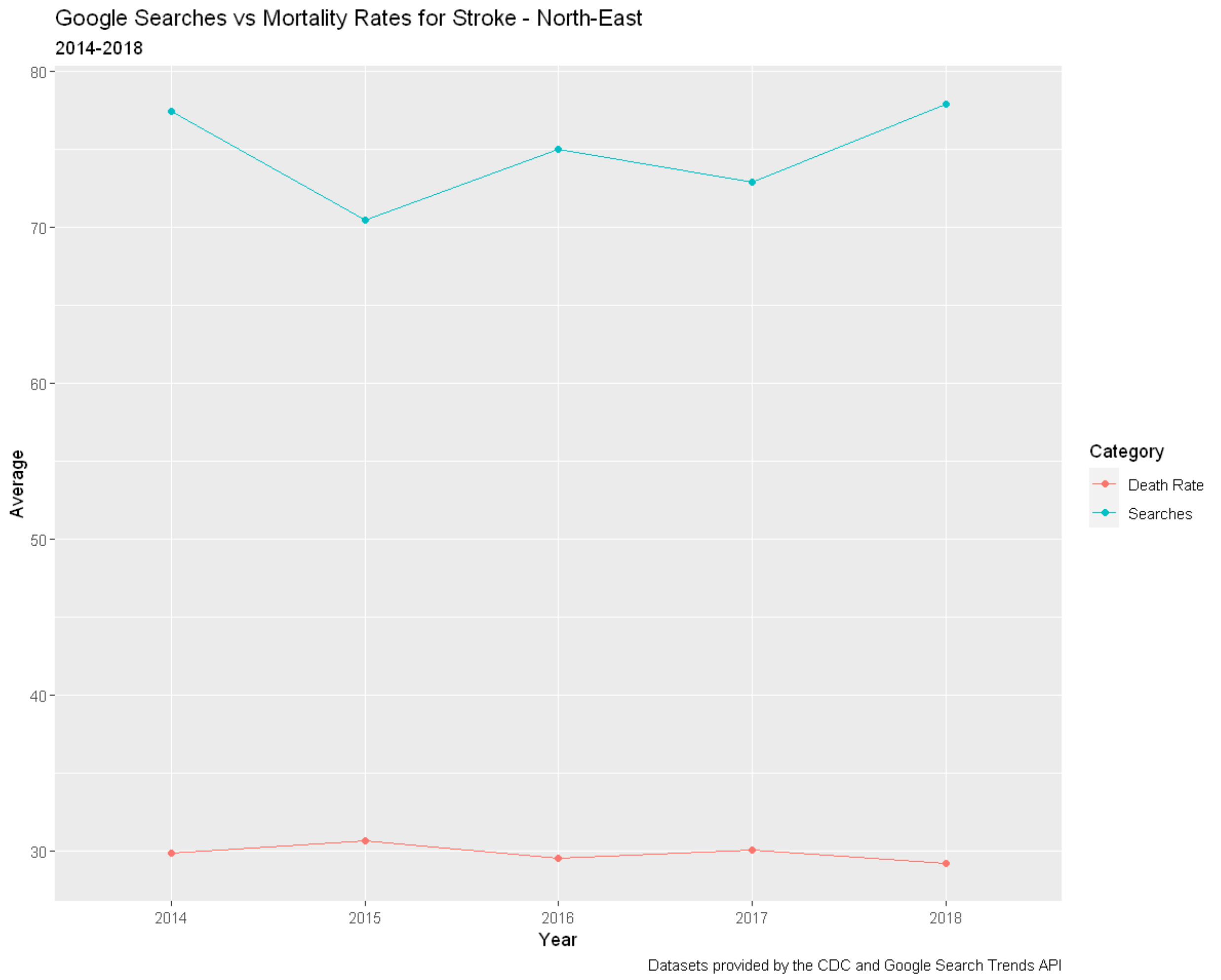


For the South, we can immediately observe a somewhat erratic pattern for the number of searches for stroke across all years, while the death rate is consistently stable with only a slight increase from 2014 to 2015.

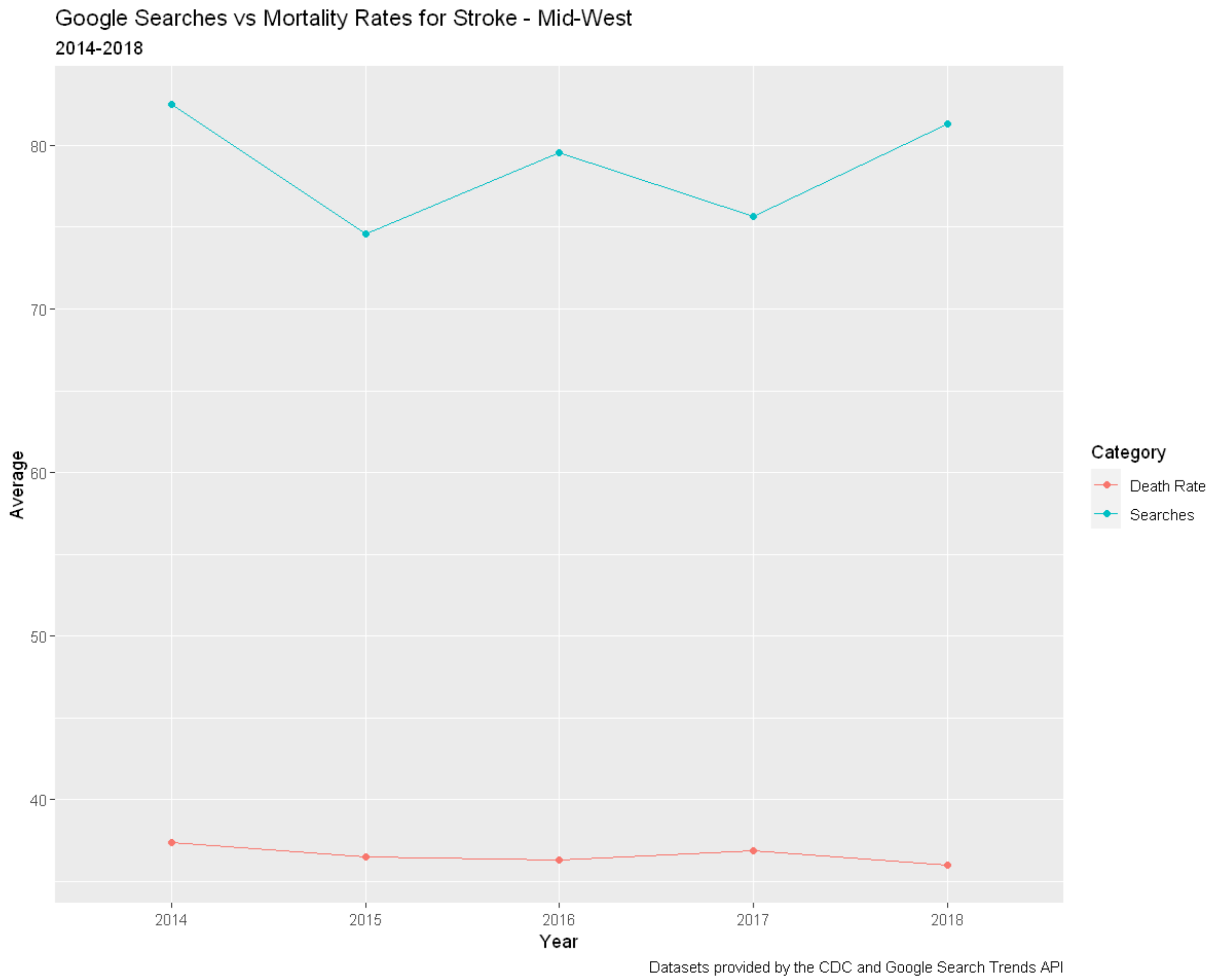
Note that the position of the lines on the graph is of little significance on its own, as we mentioned earlier the values for the death rates and searches are scaled differently, and we are not comparing them directly.



For the West, there is an almost identical pattern in searches when compared to the South, and the mortality rates are also similar, with the main difference being that the increase after 2014 is more gradual.



We can observe a very similar trend in the North-East as far as searches are concerned. The death rate, however, seems to gradually decrease after peaking in 2015.



Finally, the Mid-West shows the same search pattern as the rest of the regions. On the other hand, the death rate appears to be decreasing since 2014, while in every other region there was an increase between that year and 2015.

Now, we are ready to draw our conclusions about each cause of death and their searches for every region.

Conclusions

Out of the three causes of death analyzed, cancer is the only one showing a clear downward trend in all regions, while searches remain relatively consistent between years.

The trend in mortality rates for heart disease is mostly stable, while searches peaked between 2015 and 2017 and show a noticeable decrease in 2018.

Searches for stroke appear to be inconsistent between years across all regions, showing a potentially radical shift in interest in the issue from one year to another. Death rates show a mostly consistent pattern, with some small but important differences between regions.

Regional Differences

As for regional differences, the South was shown to consistently have the highest mortality rates for all three causes of death, possibly indicating an overall lower quality of health in the population. The South was also shown to have either the second highest number of searches (cancer and heart disease), or the consistently highest number for stroke. This may also indicate a higher level of concern for these health issues, and potentially others, when compared to the other regions. It is even possible that this increased level of interest may be an indirect result of these health issues being more commonly responsible for deaths than in other regions, however, the scope of our data is too small to confidently come to this conclusion.

The North-East appears to consistently have the second lowest mortality rate for 3 of the health issues, while having the lowest for stroke. This may indicate a contrast to the south, showing an overall healthier population. Search numbers are the second lowest for heart disease and stroke but are actually consistently the highest for cancer. This may indicate that awareness and/or concern for cancer is generally higher in the region than in the rest of the country, despite the death rates being only higher than in the West. There may be a large number of reasons for this increased interest, such potentially more aggressive awareness campaigns or even a generally more receptive culture and attitude towards health issues involving cancer.

The West consistently had the lowest number of searches and the lowest death rate, except for in the case of stroke where these numbers were slightly higher than in the North-East. There was also a more noticeable downward trend in heart disease than in the other 3 regions, although it is very slight. These factors may indicate an overall healthier population than in the South as well. The low death rate and searches may also support the idea of searches and deaths having a direct relationship, i.e. the less people are dying of a certain cause, the less interest they will show in actively learning about that cause. Nonetheless, more data is necessary to show concrete proof of this relationship.

Last but not least, the Mid-West appears to consistently place second in both death rates and searches. The exception to this is heart disease, where it had the highest number of searches across all years, which may tell us there is a higher interest in heart related health issues in this region in particular. Overall, this region appears to have the second-least healthy population after that of the South, however, one thing of note is that the Mid-West was the only region to show a consistent decrease in death rates for stroke right after 2014.

Some Final Notes

The analyses performed in this report of searches and death rates across all regions serve as a strong basis for the theory that these two elements share a direct relationship between each another, with the only major outlier in the data being that the North-East had the highest number of searches for cancer while also having the second lowest death rate.

Something very important to point out is that there are many other intervening factors when it comes to death rates in a given region, such as population density, the proportion of people in high-risk groups (for example the elderly or those with comorbidities such as obesity or compromised immune systems) to those in low-risk groups in each region, or even a lower availability of data for certain areas compared to others. Therefore, while the data analyzed in this report offers strong support for the conclusions presented about each region, they do not paint a complete picture of the overall health of the respective populations, even in relation to the three health issues that were focused on.

Datasets

All datasets for US mortality rates by state between 2014 and 2018 provided by the Centers for Disease Control and Prevention.

Details about the data and download links to the datasets can be found at:

https://www.cdc.gov/nchs/pressroom/sosmap/cancer_mortality/cancer.htm

https://www.cdc.gov/nchs/pressroom/sosmap/heart_disease_mortality/heart_disease.htm

https://www.cdc.gov/nchs/pressroom/sosmap/stroke_mortality/stroke.htm

Datasets for Google searches provided by the Python implementation of the Google Search Trends API, pytrends.

More details can be found at:

https://pypi.org/project/pytrends/