COOGLE HEALTH SEARCHES IN THE U.S.

Springboard Data Science Career Track

Capstone 1

By Chantel Clark

WHO CAN GOOGLE HEALTH QUERIES HELP?

- Hospitals and healthcare centers
  - Optimize patient education programs by matching patient and community needs
  - Which health queries are trending in real time in your Metropolitan city?
  - Learn when to send out surveys to determine interest in patient education
  - Implement preventative health measures faster if concern is increasing rapidly

- Nonprofit organizations (such as American Heart Association, etc.)
  - Identify locations of target population accurately, and in a timely manner

THE DATASET

- CSV file from Kaggle

- Contains search interest score for 210 cities across the U.S.
  - Years 2004-2017
  - Queries: cancer, cardiovascular, stroke, depression, obesity, diabetes, diarrhea, rehab, vaccine
  - Range of interest score 0-100
    - Proportional to fraction of all searches
    - 100 represents high interest
    - 50 represents a score that is half as popular
    - 0 means not enough data was collected for location

CLEANING UP THE DATA

- Zeros were removed to calculate the mean – because small proportion and did not represent low interest

- Regional information added (as defined by CDC):
  - Northeast
  - South
  - West
  - Midwest

QUESTIONS

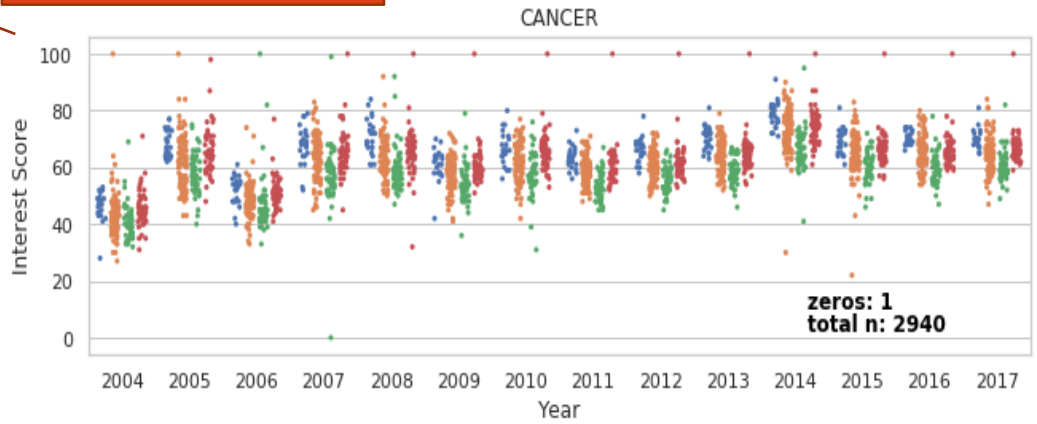
1. Are there any patterns in the regions which contain high outliers?
2. Which health searches are the most and least popular on Google?
3. Are the health searches correlated?
4. Which health queries have the largest change in mean search interest score from year to year, and over the span of 2004-2017?
5. Are there differences in interest score between the four U.S. regions (Northeast, South, West, Midwest)?

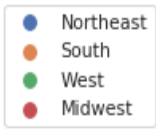
Q1: Are there any patterns in the regions which contain high outliers?

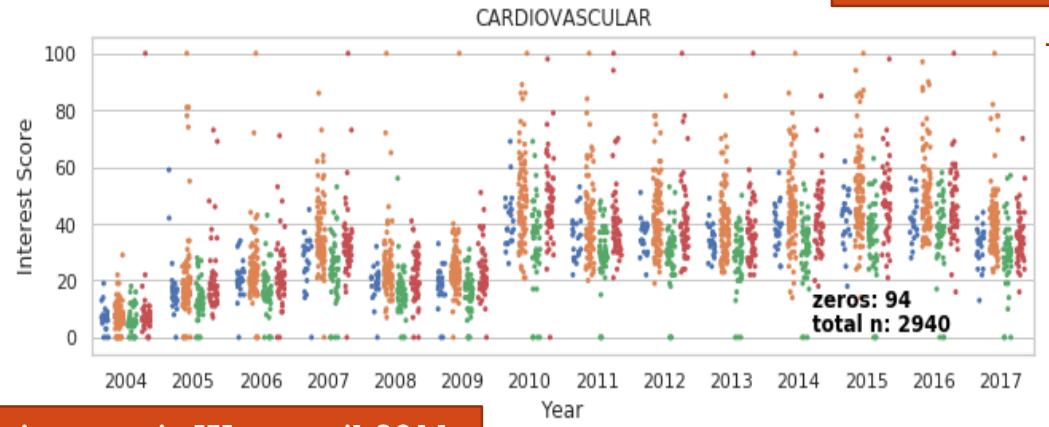
High interest in Midwest

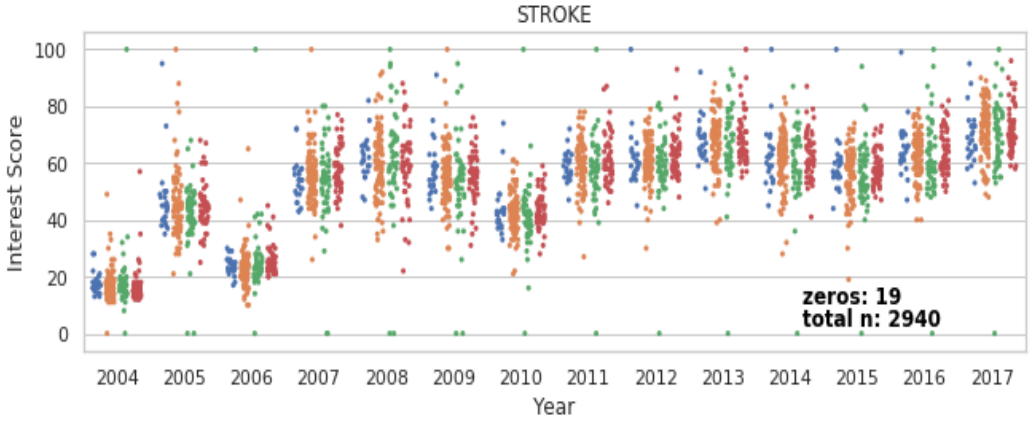
High interest in Midwest and South

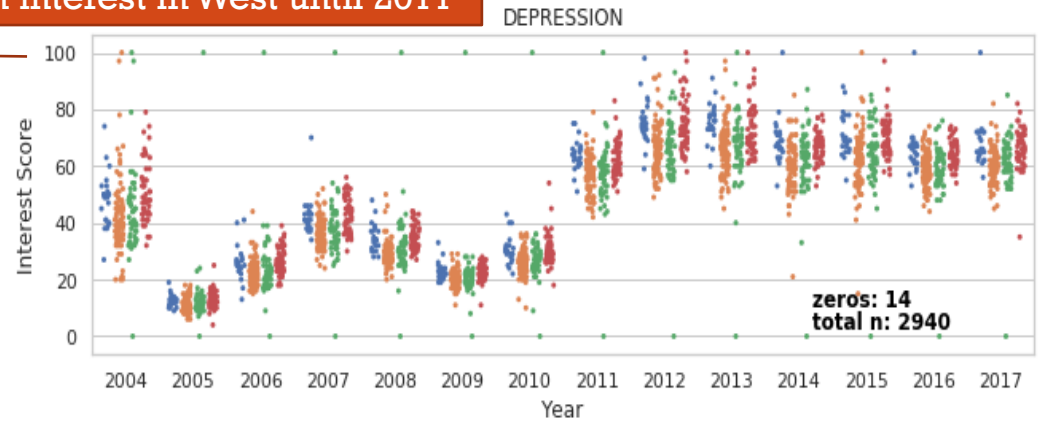
High interest in West until 2011









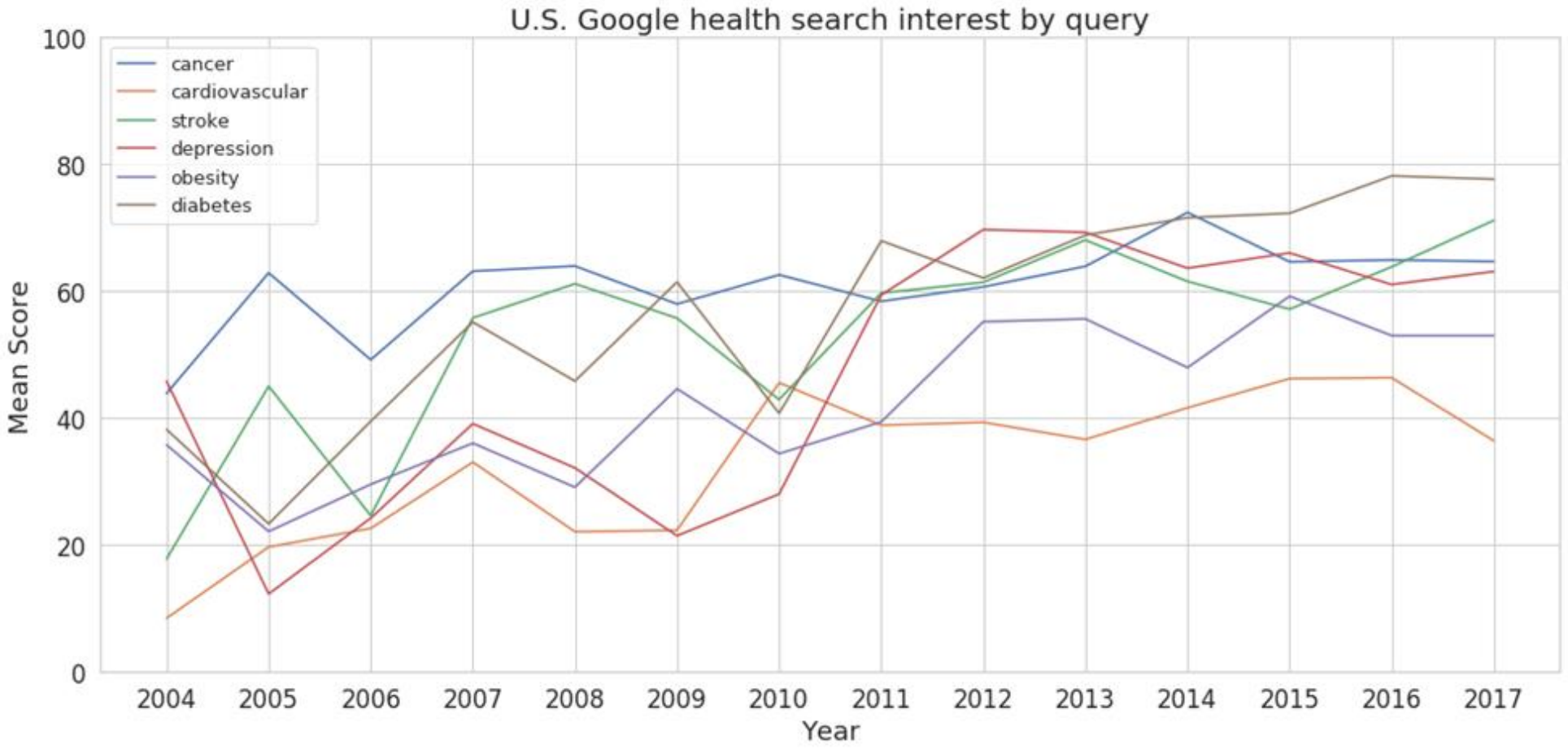


Q1: Are there any patterns in the regions which contain high outliers?

- Cancer queries had an unusually high amount of interest coming from the Midwest from 2007-2017
- Cardiovascular queries had an unusually high amount of interest coming from the Midwest and South from 2004-2017
- Depression queries had an unusually high amount of interest coming from the West from 2004-2011
- Other health queries contained a greater mix of regions

AVERAGE SEARCH INTEREST SCORES

Q2: Which health searches are the most and least popular on Google?



Q2: Which health searches are the most and least popular on Google?

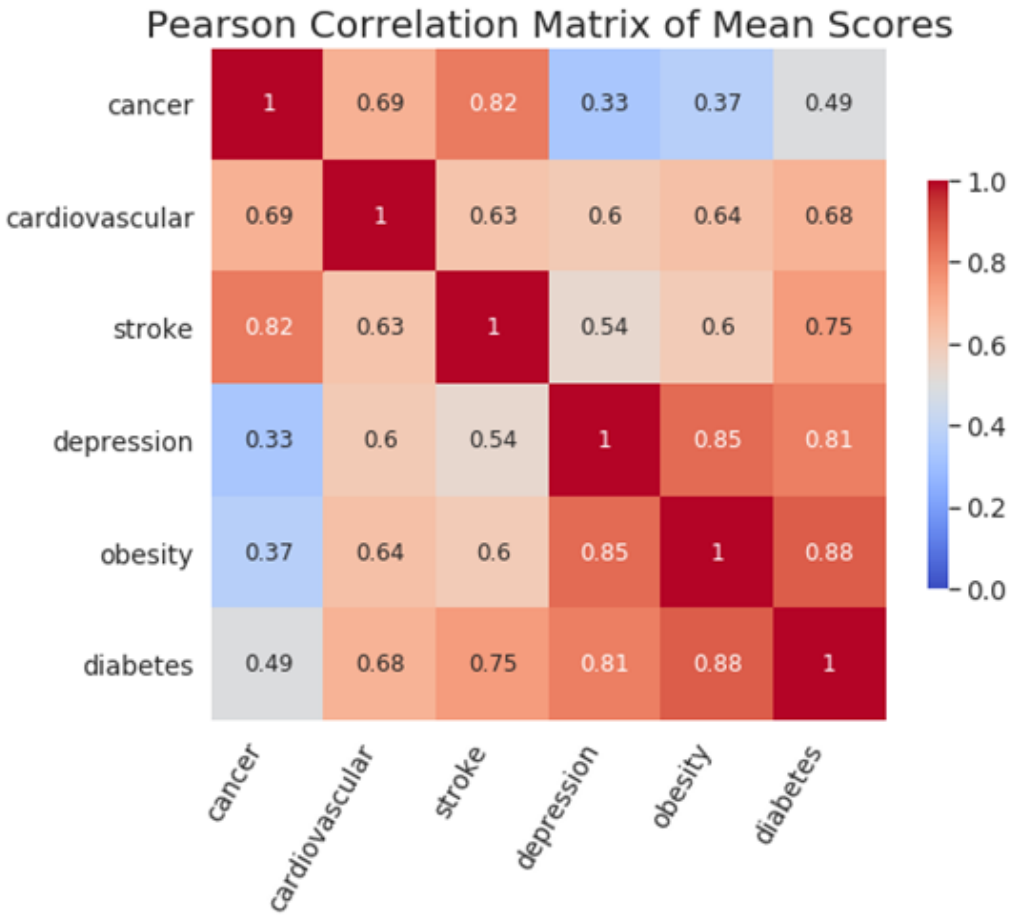
Highest mean scores:

- Cancer from 2005-2008, 2010, 2014
- Diabetes in 2009, 2011, 2013, 2015-2017
- Depression in 2004, 2012-2013

Lowest mean scores:

- Cardiovascular in 2004, 2006-2008, 2011-2017
- Depression in 2005, 2009-2010

Q3: ARE HEALTH SEARCHES CORRELATED?



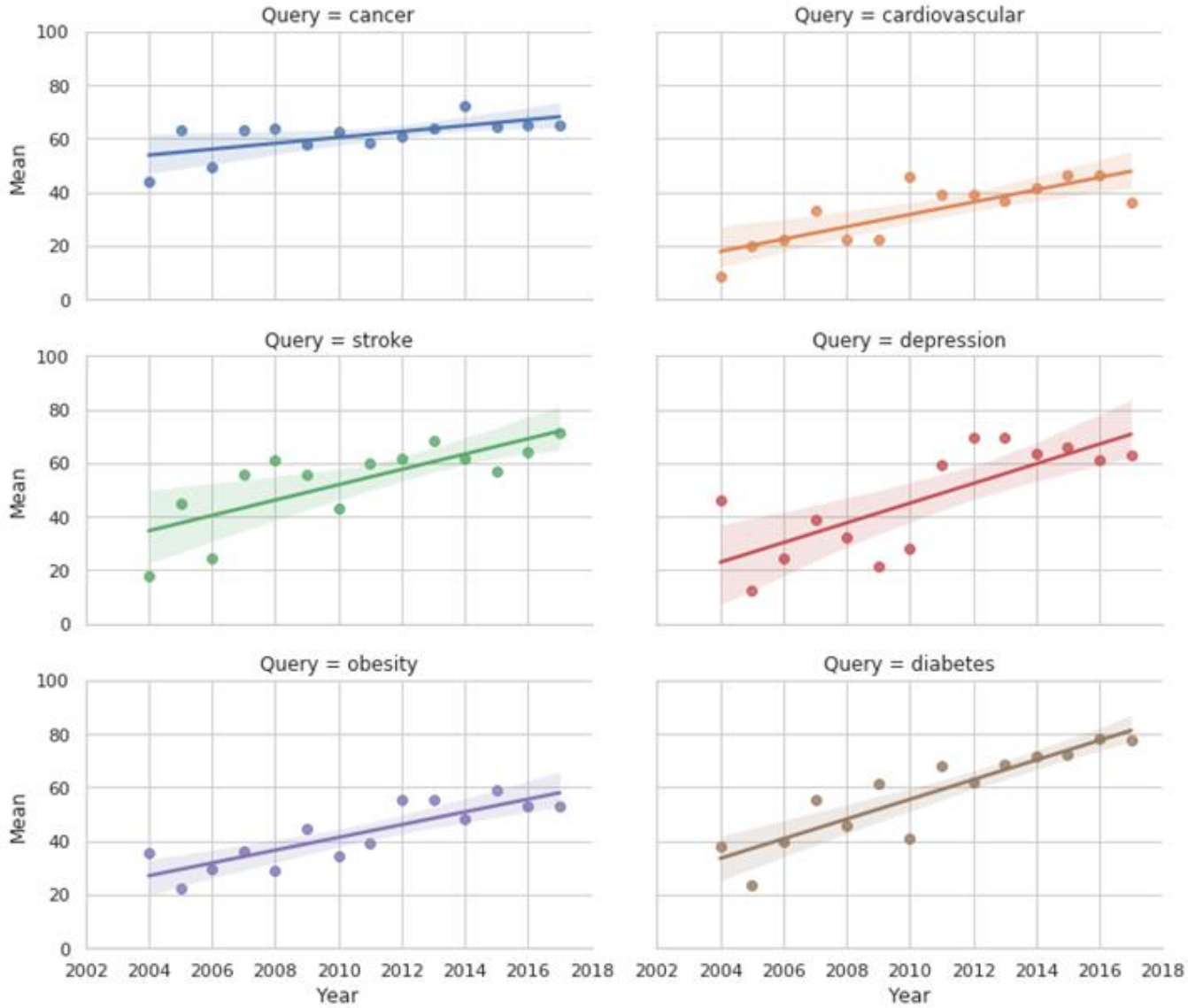
Obesity and diabetes: Strongest correlation 0.88 - people who are obese are at risk of getting Type 2 diabetes (Resnick et al. 2000)

Obesity and depression: 0.85 correlation – obesity and depression often co-occur and risk is bidirectional (Luppino et al. 2010)

CHANGE IN AVERAGE SCORES

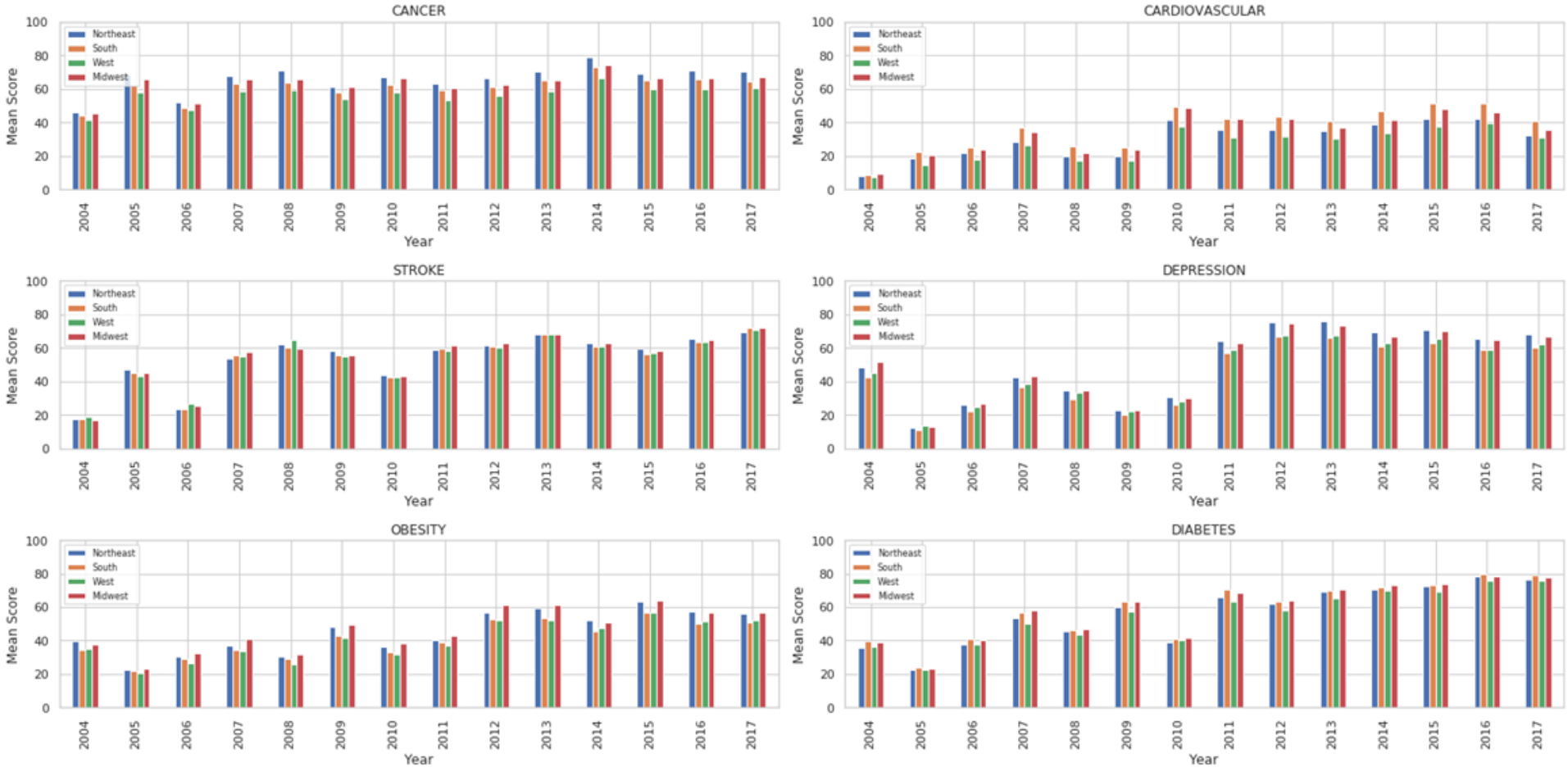
Q4: Which health queries have the largest change in mean search interest score from year to year, and over the span of 2004-2017?

- Stroke and depression scores do NOT appear linear
- Depression large jump from 2010-2011
- Diabetes has the highest rate of increase from 2004-2017



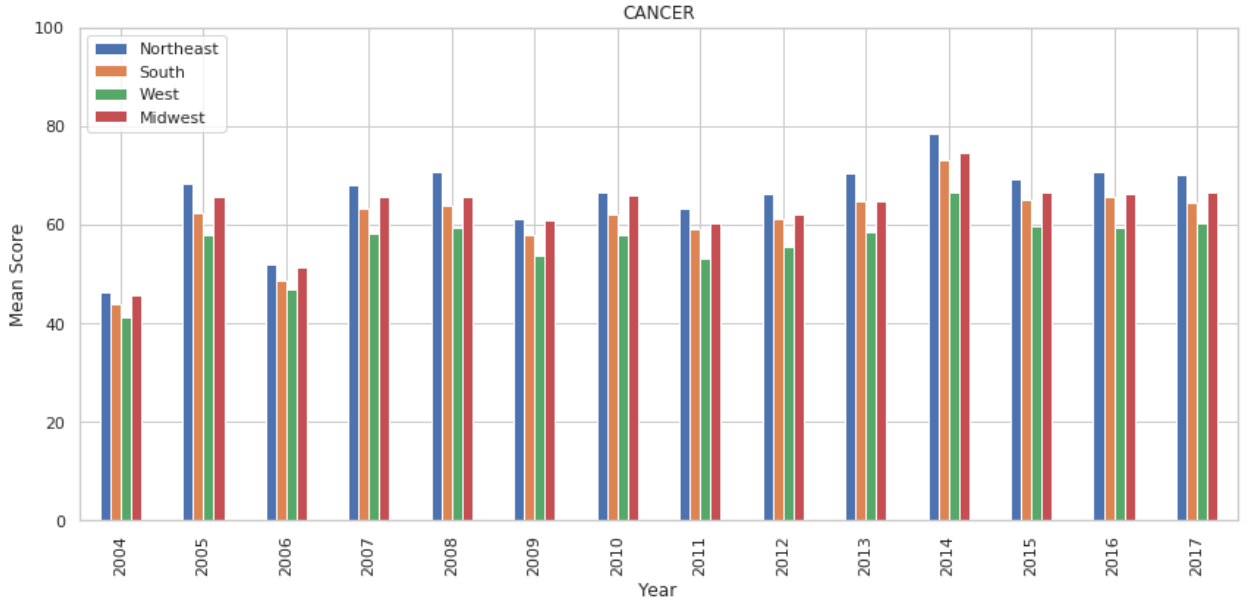
DIFFERENCES BETWEEN REGIONS

Q5: Are there differences in mean interest score between the four U.S. regions?

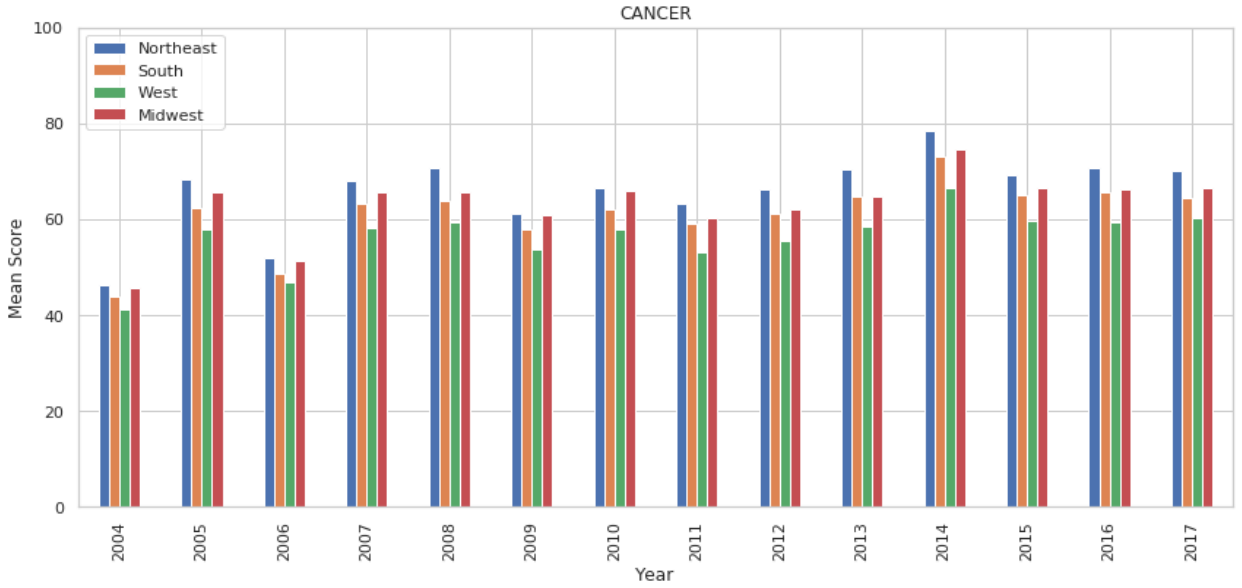


HYPOTHESIS TESTING

- Q5: Does the Northeast have significantly higher interest than the West in cancer?
- One-tailed Student t-test to compare the means
- Null hypothesis: Northeast cancer mean interest score ≤ West cancer mean interest score
- Alternative hypothesis: Northeast cancer mean interest score > West cancer mean interest score



t-statistic =15.3, p-value 6.22e-48; reject null hypothesis
Conclusion: Northeast has a higher mean interest score for cancer than the West



	Northeast_n	Northeast_var	West_n	West_var	t-stat	p-value
2004	23	25.5161	49	40.2316	3.62269	0.000328056
2005	23	21.8904	49	49.8126	7.24534	4.28741e-10
2006	23	24.7221	49	114.694	2.598	0.00571193
2007	23	46.3894	49	135.549	4.91491	3.08172e-06
2008	23	40.1248	49	62.399	6.46719	1.70254e-08
2009	23	30.3667	49	41.8776	4.9328	4.76992e-06
2010	23	31.7769	49	56.9238	5.43715	6.20331e-07
2011	23	14.7335	49	23.0187	9.57387	2.09237e-13
2012	23	17.5614	49	20.3898	9.55262	9.11211e-13
2013	23	17.8639	49	21.696	10.3634	5.42147e-14
2014	23	20.1512	49	49.9184	8.70913	1.02638e-12
2015	23	18.9527	49	26.3815	7.89685	1.24249e-10
2016	23	5.38752	49	28.4123	12.3672	1.4558e-19
2017	23	11.0851	49	27.9592	9.22487	1.24637e-13

All p-values were less than 0.05, therefore reject null hypothesis

Northeast mean interest scores are significantly higher than the West mean interest score for every year

HYPOTHESIS TESTING

- Did the Northeast have significantly higher interest than the West in cancer for every year?
- One-sided Welch's t-test
- Null hypothesis: Northeast cancer mean interest score $\leq$ West cancer mean interest score
- Alternative hypothesis: Northeast cancer mean interest score $>$ West cancer mean interest score

CONCLUSIONS

1. Are there any patterns in the regions which contain high outliers?
   - Yes, the Midwest had unusually high interest scores for cancer and cardiovascular, the South for cardiovascular, and the West for depression
   - Other queries contained a mix
2. Which health searches are the most and least popular on Google?
   - Cancer and diabetes had the most interest over the years 2004-2017
   - Cardiovascular had the lowest interest scores
3. Are the health searches correlated?
   - Obesity searches were correlated with diabetes and depression
4. Which health queries have the largest change in mean search interest score from year to year, and over the span of 2004-2017?
   - Depression had the highest change from 2010 to 2011
   - Diabetes had the larges overall increase in interest from 2004-2017
5. Are there differences in interest score between the four U.S. regions (Northeast, South, West, Midwest)?
   - The Northeast had significantly higher interest scores in cancer than the West

REFERENCES

- Luppino FS, de Wit LM, Bouvy PF, et al. Overweight, Obesity, and Depression: A Systematic Review and Meta-analysis of Longitudinal Studies. *Arch Gen Psychiatry.* 2010;67(3):220–229. doi:10.1001/archgenpsychiatry.2010.2
- Resnick HE, Valsania P, Halter JB*, et al.* Relation of weight gain and weight loss on subsequent diabetes risk in overweight adults. *Journal of Epidemiology & Community Health* 2000;**54:**596-602