8.5 Statistical Analysis for Capstone 1

by Chantel Clark

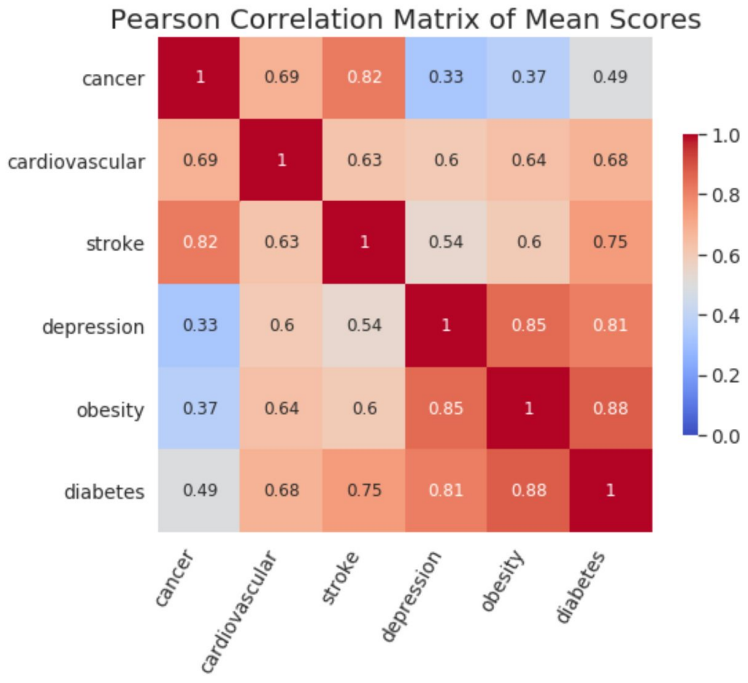
Descriptive statistics were printed to understand the distribution of the data. The describe method printed out the count, mean, standard deviation, minimum, maximum, and quantiles of the search interest scores.

```
[ ]  healthSearchData.describe().T.head()
```

	count	mean	std	min	25%	50%	75%	max
2004+cancer	210.0	43.904762	7.618944	27.0	40.0	43.0	47.0	100.0
2004+cardiovascular	210.0	7.433333	7.909647	0.0	5.0	6.0	9.0	100.0
2004+stroke	210.0	17.642857	8.135284	0.0	14.0	16.0	18.0	100.0
2004+depression	210.0	45.623810	13.715720	0.0	37.0	44.0	51.0	100.0
2004+rehab	210.0	18.890476	10.157723	0.0	15.0	17.0	21.0	100.0

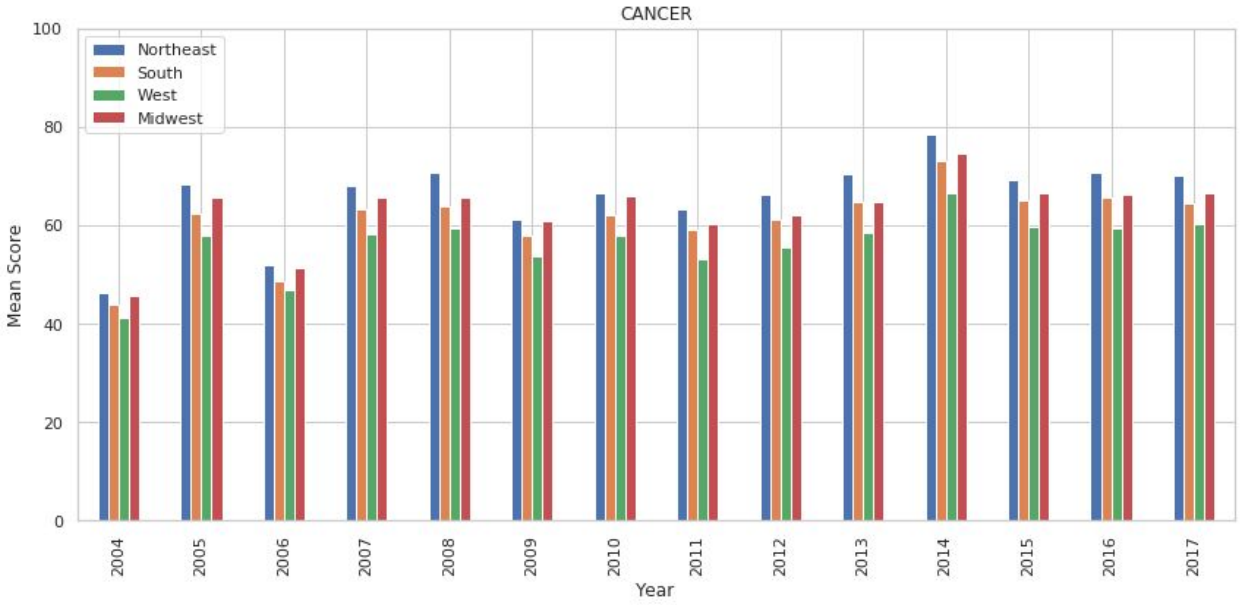
From the print out, I found that the data has a range of 0-100. On Google Search API the 'interest score' is proportional to the fraction of all searches, so a score of 100 represents very high interest, and a score of 50 is half as popular as the previous search. A score of 0 means that there was not enough data. Because the zero scores were at most 3.2% of the total data and did not represent a low interest score, they were omitted from the analysis, before calculating mean interest scores.

To find whether or not queries were correlated with each other, a correlation matrix for the means of each health query was created.



The strongest correlation among the health queries was 0.88 between obesity and diabetes, which makes sense because people who are obese are at a high rist of getting Type 2 diabetes. The next strongest correlation was 0.85 between obesity and depression. Studies have also shown that people with obesity are at a higher risk of being depressed than people without obesity. The intertwined nature of obesity and depression is still being studied, medical professionals are unsure if the relationship might be bidirectional, meaning that depression could potentially cause obesity in some cases (https://www.medicalnewstoday.com/articles/323668).

Does the Northeast have a significantly higher interest than the West in cancer?



To determine whether or not there was a significant difference, a student t-test was used to compare the means. The null hypothesis was that the Northeast has a mean interest score less than or equal to the mean interest score in the West for cancer. The alternative hypothesis was that the Northeast has a significantly higher interest score than the West.

A student t-test was performed to compare the means of the Northeast and West interest scores. There were 322 samples with a variance of about 85.9 from the Northeast, and 686 samples with a variance of about 85.3 from the West. Because the sample sizes were large and variances were close, equal variances were assumed. The resulting t-statistic was 15.3, with p-value 6.22e-48.

Variances were not equal within each year, therefore a Welch's t-test was performed to test the same hypothesis for each year. A p-value less than 0.05 means that there is less than a 5% chance that given the condition of the null hypothesis that the Northeast has a mean interest score less than or equal to the mean interest score in the West for cancer, there is less than a 5% chance that we would obtain the means that are in our data given a random sample. All p-values were less than 0.05, therefore we can reject the null hypothesis and conclude that the Northeast mean interest scores are significantly higher than the West mean interest score for every year in this dataset.

Welch's t-test results

```
2004
Northeast: n = 23 var = 25.516068052930056
West: n = 49 var = 40.231570179092046
t-stat = 3.6226944024476295 p-value = 0.00032805604204169225
2005
Northeast: n = 23 var = 21.890359168241968
West: n = 49 var = 49.812578092461465
t-stat = 7.245338218446185 p-value = 4.2874076845945e-10
2006
Northeast: n = 23 var = 24.722117202268425
West: n = 49 var = 114.6938775510204
t-stat = 2.598001919945983 p-value = 0.005711927674284372
2007
Northeast: n = 23 var = 46.389413988657836
West: n = 49 var = 135.54935443565188
t-stat = 4.914913235709949 p-value = 3.0817234734110875e-06
2008
Northeast: n = 23 var = 40.124763705103966
West: n = 49 var = 62.3990004164931
t-stat = 6.467188791119975 p-value = 1.702536168288156e-08
2009
Northeast: n = 23 var = 30.36672967863894
West: n = 49 var = 41.87755102040817
t-stat = 4.932798898111672 p-value = 4.769924088847964e-06
2010
Northeast: n = 23 var = 31.77693761814744
West: n = 49 var = 56.92378175760099
t-stat = 5.437145022489872 p-value = 6.203305667354999e-07
2011
Northeast: n = 23 var = 14.73345935727788
West: n = 49 var = 23.018742190753873
t-stat = 9.573873978655275 p-value = 2.092368245722694e-13
2012
Northeast: n = 23 var = 17.561436672967865
West: n = 49 var = 20.389837567680136
t-stat = 9.552618804065018 p-value = 9.11210555371442e-13
2013
Northeast: n = 23 var = 17.863894139886575
West: n = 49 var = 21.695960016659733
t-stat = 10.363377496508333 p-value = 5.421474347542823e-14
2014
Northeast: n = 23 var = 20.15122873345936
```

West: n = 49 var = 49.91836734693879
t-stat = 8.709133755918957 p-value = 1.026384341202711e-12
2015
Northeast: n = 23 var = 18.952741020793948
West: n = 49 var = 26.38150770512287
t-stat = 7.89684993492182 p-value = 1.2424891598511318e-10
2016
Northeast: n = 23 var = 5.387523629489602
West: n = 49 var = 28.41232819658473
t-stat = 12.36723444075989 p-value = 1.455804560109636e-19
2017
Northeast: n = 23 var = 11.08506616257089
West: n = 49 var = 27.959183673469393
t-stat = 9.22487029496813 p-value = 1.2463713540173798e-13