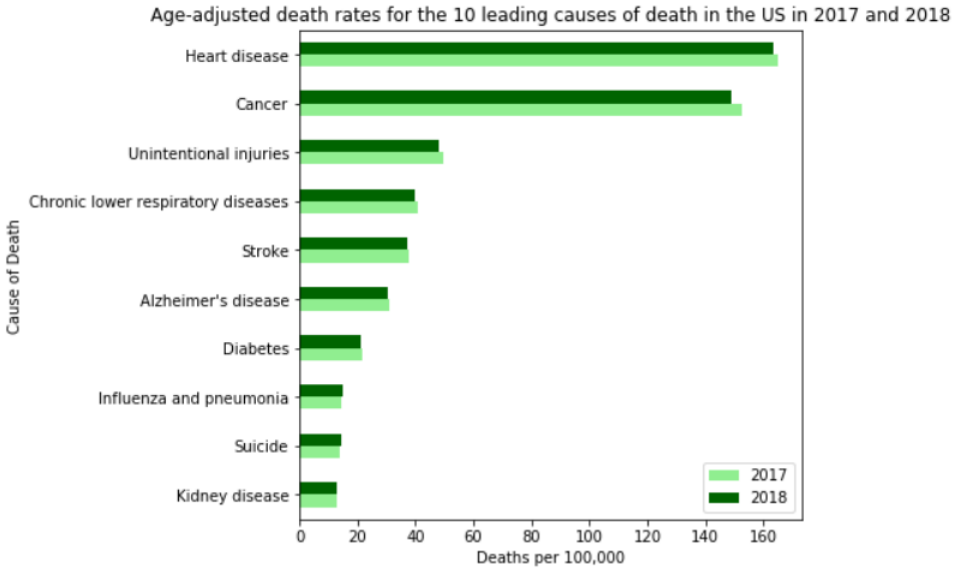
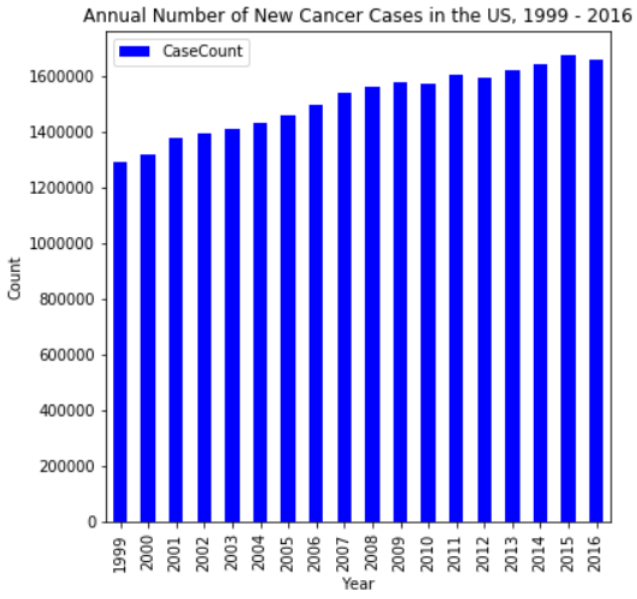
Introduction

Background

Cancer is the second leading cause of death in the Unites States, second only to heart disease. Even though the overall death rate from cancer has been decreasing due to a multitude of reasons including improved treatment options, increased awareness and lifestyle changes among the population; the number of cancer cases year over year has been steadily rising as evidenced by the charts below.

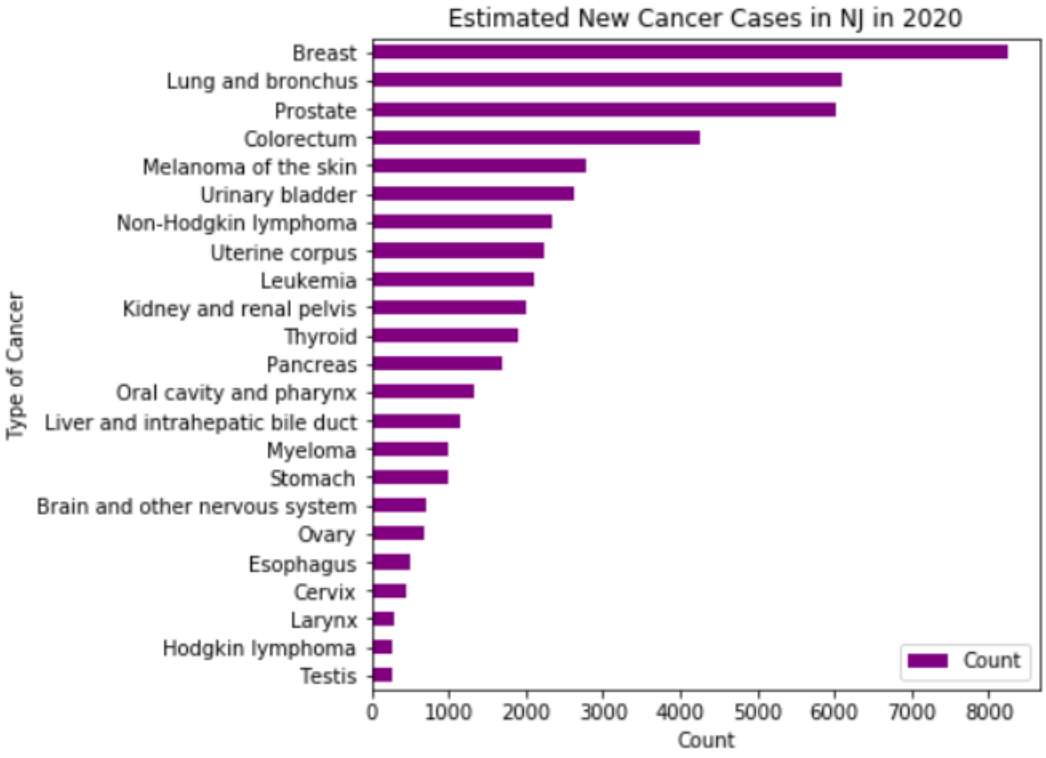




Problem

Cancer Hospitals USA (a fictitious group) is a group of premier cancer hospitals located in several locations in the US, including in New York City (NYC). They have a large percentage of patients who live in New Jersey (NJ). It is very difficult for the patients and their families to travel to NYC regularly – whether it be for consultation or outpatient treatment. Additionally, if the patient needs surgery then the caregiver often needs to stay in NYC for a few days. Obviously, this imposes a huge physical and financial burden on the families.

In 2020, an estimated 53,340 new cases of cancer are expected to occur in NJ with the highest incidence for breast, lung and prostate cancers (see chart below). Considering this large number as well as the inconvenience posed to NJ residents when travelling to their NYC hospital, Cancer hospitals USA intends to open a brand-new state-of-the-art hospital in NJ. The management and Board of Directors would like data on the existing hospitals in NJ based on zip code so that they can pinpoint the location that would be most beneficial to the community as well as to the hospital.



Data Acquisition and Cleaning

- **Data Sources and Acknowledgements**
    - New Jersey zip code and latitude/longitude data was obtained from https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/.
      This data is free and is licensed under Creative Commons Attribution Share-Alike (cc-by-sa) http://creativecommons.org/licenses/by-sa/4.0/
    - Population data for zip codes in New Jersey was obtained from https://www.newjersey-demographics.com/zip_codes_by_population
    - Leading causes of death in the US – CDC website at https://www.cdc.gov/nchs/products/databriefs/db355.htm
    - US Cancer Statistics Year Over Year Trend – CDC website at https://gis.cdc.gov/Cancer/USCS/DataViz.html
    - New Jersey 2020 Estimates – Cancer Statistics Center of the American Cancer Society website at https://cancerstatisticscenter.cancer.org/#!/state/New%20Jersey
- **Data Cleaning**

  The following steps were performed to clean and prepare the data
    - Death Rates for the 10 Leading Causes of Death in the US: The report on the CDC website was in PDF format and the chart was displayed as an image – hence a downloadable Excel file was not available, nor was it possible to do web scraping. The data was manually entered into a spreadsheet and then loaded into a Pandas dataframe to display the horizontal bar chart in the Introduction section.
    - Annual Number of New Cancer Cases in the US from 1999 - 2016: A CSV file was downloaded from the CDC website, nonessential columns were removed and data was formatted appropriately. The CSV file was then loaded into a Pandas dataframe and a vertical bar chart was used to display the data.
    - Estimated New Cancer Cases in NJ in 2020: A CSV file was downloaded from the Cancer Statistics Center website. Only data for the state of NJ was retained, all other data was deleted. Non-required columns were deleted and any rows that were missing counts of no. of cases were removed as well. The data was then loaded into a Pandas dataframe and visualized via a horizontal bar chart.
    - New Jersey Zip Code and Latitude/Longitude Data: A CSV file was downloaded from the OpenDataSoft website and loaded into a Pandas dataframe. Unneeded columns were dropped. No further data cleaning was necessary.
    - Population Data for Zip Codes in NJ: Web scraping methodology was employed to obtain data from https://www.newjersey-demographics.com/zip_codes_by_population. After loading into a Pandas dataframe, 1 column was dropped and another was renamed to match the column name in the latitude/longitude dataframe. There were 6 rows that contained multiple zip codes with an aggregated population value – these were dropped since there was no way to identify and separate the values for the individual zip codes involved.
    - Finally, the 2 dataframes containing latitude/longitude and population data for each zip code in NJ were merged based on the common field zip code.

Methodology

- Clean, load and display the NJ Latitude/Longitude data

	Zip	City	Latitude	Longitude
0	07309	Jersey City	40.73276	-74.075485
1	07961	Morristown	40.77975	-74.442797
2	08887	Three Bridges	40.525361	-74.79632
3	08817	Edison	40.516104	-74.39754
4	08406	Ventnor City	39.342299	-74.48192
5	08835	Manville	40.538903	-74.59222
6	08629	Trenton	40.219358	-74.73334
7	07032	Kearny	40.763051	-74.13718
8	08051	Mantua	39.785785	-75.17761
9	07030	Hoboken	40.744851	-74.03294

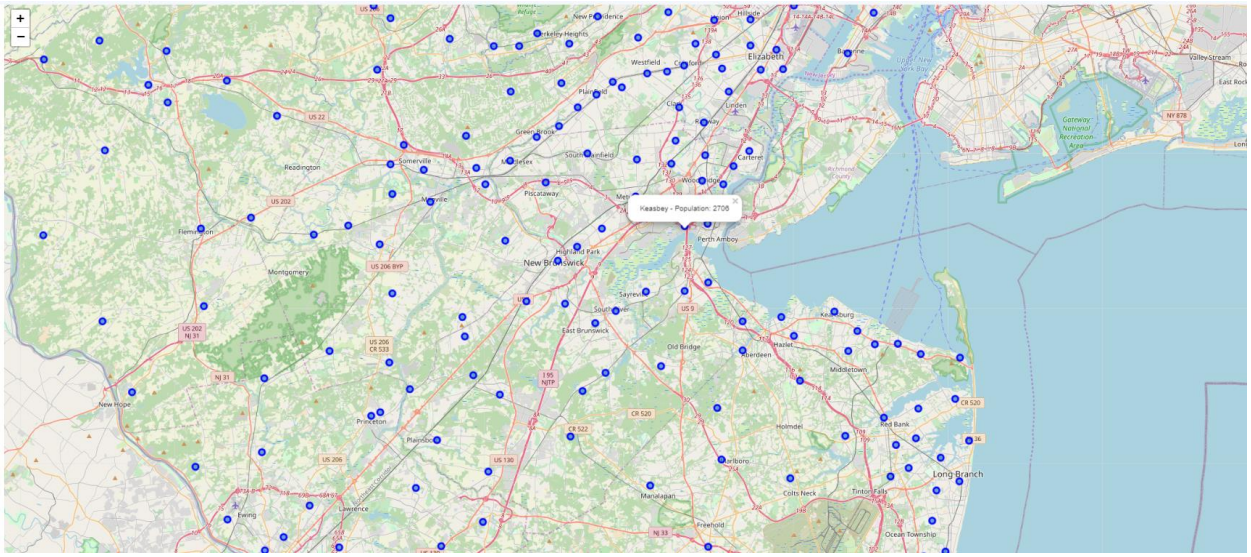
- Perform web scraping, clean, load and display the NJ population data

	Zip	Population
0	08701	100763
1	07055	70199
2	07087	68484
3	07002	65300
4	07305	64535
5	07093	63913
6	08753	63026
7	07047	61970
8	08901	57659
9	08854	56931

- Merge the 2 dataframes and display the merged data

	Zip	City	Latitude	Longitude	Population
0	08887	Three Bridges	40.525361	-74.79632	1004
1	08817	Edison	40.516104	-74.39754	44823
2	08406	Ventnor City	39.342299	-74.48192	10232
3	08835	Manville	40.538903	-74.59222	10283
4	08629	Trenton	40.219358	-74.73334	14108
5	07032	Kearny	40.763051	-74.13718	41523
6	08051	Mantua	39.785785	-75.17761	10409
7	07030	Hoboken	40.744851	-74.03294	53211
8	07939	Lyons	40.566553	-74.599801	212
9	08075	Riverside	40.029361	-74.9541	28731
10	07666	Teaneck	40.890964	-74.01115	40543

- Draw a Folium map to show different zip codes based on latitude and longitude values and display markers showing the city and population



- Search for medical centers in NJ using the FourSquare API
- Extract the different categories of Medical Centers. Then clean the json, structure it into a pandas dataframe, filter out unnecessary columns and display it

```
(14561, 8)
```

	ZipCode	Zipcode Latitude	Zipcode Longitude	Venue Name	Venue Category	Venue Address	Venue Latitude	Venue Longitude
0	08887	40.525361	-74.79632	Hunterdon Medical Center	Medical Center	[2100 Wescott Drive (at State Route 31), Flemi...	40.531362	-74.861005
1	08887	40.525361	-74.79632	Pleasant Run Family Physicians	Doctor's Office	[New Jersey, United States]	40.543495	-74.752595
2	08887	40.525361	-74.79632	Hunterdon Family Physicians	Doctor's Office	[111 State Route 31, Flemington, NJ 08822, Uni...	40.520541	-74.855043
3	08887	40.525361	-74.79632	ER	Emergency Room	[6-1684 Hunterdon Med Center, Flemington, NJ 0...	40.532081	-74.861355
4	08887	40.525361	-74.79632	Hunterdon Woman's Imaging Center	Medical Center	[121 Route 31, Flemington, NJ 08822, United St...	40.521219	-74.855250
5	08887	40.525361	-74.79632	Hunterdon Pediatric Associates	Medical Center	[6 Sandhill Road, Flemington, NJ 08822, United...	40.537442	-74.858765
6	08887	40.525361	-74.79632	Hunterdon Gastroenterology Associates	Doctor's Office	[1100 Wescott Dr # G3, Flemington, NJ 08822, U...	40.531751	-74.860025
7	08887	40.525361	-74.79632	Hunterdon Womens Imaging	Doctor's Office	[Flemington, NJ, United States]	40.513675	-74.854994
8	08887	40.525361	-74.79632	Sophisticated Smiles	Dentist's Office	[Flemington, NJ 08822, United States]	40.498560	-74.852517
9	08887	40.525361	-74.79632	Advanced Obstetrics & Gynecology	Doctor's Office	[Wescott Drive, Flemington, NJ 08822, United S...	40.530556	-74.859854

Analysis of Results and Machine Learning

- List the count of venues for each neighborhood

ZipCode	Zipcode Latitude	Zipcode Longitude	Venue Name	Venue Category	Venue Address	Venue Latitude	Venue Longitude
07001	30	30	30	30	30	30	30
07002	30	30	30	30	30	30	30
07003	30	30	30	30	30	30	30
07004	30	30	30	30	30	30	30
07005	30	30	30	30	30	30	30
07006	30	30	30	30	30	30	30
07008	26	26	26	26	26	26	26
07009	30	30	30	30	30	30	30
07010	30	30	30	30	30	30	30
07011	30	30	30	30	30	30	30
07012	30	30	30	30	30	30	30

- Count and list number of unique categories in all the venues combined

```
There are 56 unique categories.
```

```
: Venue Category
Doctor's Office                 4592
Medical Center                  2605
Dentist's Office                1874
Hospital                        1699
Veterinarian                    1013
Medical Lab                      533
Chiropractor                     502
Emergency Room                   361
Urgent Care Center               194
Physical Therapist               186
Eye Doctor                       141
Mental Health Office             127
Rehab Center                      96
Assisted Living                   82
Optical Shop                      78
Acupuncturist                     67
Hospital Ward                     61
Pharmacy                          56
Weight Loss Center                39
Pet Service                       38
Building                          30
Office                            17
Maternity Clinic                  16
Fire Station                      14
Health & Beauty Service           13
Beer Bar                          12
Alternative Healer                10
Recreation Center                 10
Medical School                    10
```

- This data contains multiple venues mislabeled under the Medical Center category so it will need to be cleaned up. We also want to remove categories such as Doctor's Office, Dentist's Office etc. which are

correctly categorized but not relevant to our analysis.

```
(5026, 8)
```

	ZipCode	Zipcode Latitude	Zipcode Longitude	Venue Name	Venue Category	Venue Address	Venue Latitude	Venue Longitude
0	08887	40.525361	-74.79632	Hunterdon Medical Center	Medical Center	[2100 Wescott Drive (at State Route 31), Flemi...	40.531362	-74.861005
3	08887	40.525361	-74.79632	ER	Emergency Room	[6-1684 Hunterdon Med Center, Flemington, NJ 0...	40.532081	-74.861355
4	08887	40.525361	-74.79632	Hunterdon Woman's Imaging Center	Medical Center	[121 Route 31, Flemington, NJ 08822, United St...	40.521219	-74.855250
5	08887	40.525361	-74.79632	Hunterdon Pediatric Associates	Medical Center	[6 Sandhill Road, Flemington, NJ 08822, United...	40.537442	-74.858765
11	08887	40.525361	-74.79632	Hunterdon Regional Cancer Center	Hospital	[Flemington, NJ 08822, United States]	40.531249	-74.860222
15	08887	40.525361	-74.79632	Health Quest	Medical Center	[Route 31, Flemington, NJ 08822, United States]	40.538987	-74.855654
18	08887	40.525361	-74.79632	Hunterdon Center for Surgery	Medical Center	[9100 Wescott Dr, Flemington, NJ 08822, United...	40.531217	-74.858479
21	08887	40.525361	-74.79632	Hunterdon Pediatric Associates	Medical Center	[8 Reading Road, Flemington, NJ 08822, United ...	40.504545	-74.843180
23	08887	40.525361	-74.79632	MidJersey Orthopaedics	Medical Center	[8100 Wescott Dr, Flemington, NJ 08822, United...	40.534819	-74.861551
26	08887	40.525361	-74.79632	Hunterdon Podiatric Medicine	Medical Center	[1100 Wescott Drive, Flemington, NJ 08822, Uni...	40.530771	-74.861470
29	08817	40.516104	-74.39754	JFK Hartwyck at Edison Estates	Medical Center	[10 Brunswick Ave, Edison, NJ 08817, United St...	40.526398	-74.412385

- Then analyze each ZipCode using one hot encoding

	ZipCode	Emergency Room	Hospital	Hospital Ward	Medical Center	Medical School	Rehab Center	Urgent Care Center
0	08887	0	0	0	1	0	0	0
3	08887	1	0	0	0	0	0	0
4	08887	0	0	0	1	0	0	0
5	08887	0	0	0	1	0	0	0
11	08887	0	1	0	0	0	0	0

- Group rows by ZipCode and by taking the mean of the frequency of occurrence of each category

	ZipCode	Emergency Room	Hospital	Hospital Ward	Medical Center	Medical School	Rehab Center	Urgent Care Center
0	07001	0.125000	0.125000	0.000000	0.625000	0.000000	0.000000	0.125000
1	07002	0.111111	0.222222	0.000000	0.555556	0.000000	0.000000	0.111111
2	07003	0.062500	0.437500	0.000000	0.500000	0.000000	0.000000	0.000000
3	07004	0.090909	0.181818	0.000000	0.727273	0.000000	0.000000	0.000000
4	07005	0.000000	0.500000	0.071429	0.357143	0.000000	0.071429	0.000000
5	07006	0.000000	0.200000	0.000000	0.800000	0.000000	0.000000	0.000000
6	07008	0.000000	0.000000	0.000000	0.833333	0.000000	0.166667	0.000000
7	07009	0.000000	0.181818	0.000000	0.727273	0.000000	0.090909	0.000000
8	07010	0.052632	0.789474	0.052632	0.105263	0.000000	0.000000	0.000000
9	07011	0.000000	0.230769	0.000000	0.692308	0.000000	0.076923	0.000000

- Use K-Means clustering to further analyze the data. Set the number of clusters to 5.
- Find the top 5 venues in each ZipCode

```
----07001----
              venue  freq
0      Medical Center  0.62
1      Emergency Room  0.12
2            Hospital  0.12
3  Urgent Care Center  0.12
4       Hospital Ward  0.00


----07002----
              venue  freq
0      Medical Center  0.56
1            Hospital  0.22
2      Emergency Room  0.11
3  Urgent Care Center  0.11
4       Hospital Ward  0.00
```

- Create a dataframe with the venues in descending order

	Zip	City	Latitude	Longitude	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	08887	Three Bridges	40.525361	-74.79632	1004	3.0	Medical Center	Hospital	Emergency Room	Urgent Care Center	Rehab Center
1	08817	Edison	40.516104	-74.39754	44823	1.0	Medical Center	Hospital	Emergency Room	Urgent Care Center	Rehab Center
2	08406	Ventnor City	39.342299	-74.48192	10232	0.0	Hospital	Medical Center	Urgent Care Center	Emergency Room	Rehab Center
3	08835	Manville	40.538903	-74.59222	10283	4.0	Medical Center	Hospital	Urgent Care Center	Emergency Room	Rehab Center
4	08629	Trenton	40.219358	-74.73334	14108	0.0	Hospital	Medical Center	Urgent Care Center	Rehab Center	Medical School

- Examine each cluster to see the most common venues

Discussion

Cluster # 4 has the highest number of zip codes with hospitals within 5000 meters radius as the most common venue, whereas in clusters # 3 and # 5 there is a preponderance of medical centers. Cluster # 2 has a mix of medical centers and emergency rooms and cluster # 1 has both medical centers and urgent care centers as the most common venue.

One caveat to keep in mind when interpreting the results is that some hospitals might have been counter multiple times since a radius of 5000 meters was used. However, anything less than that would not make practical sense.

Conclusion

Analysis of the distribution of hospitals in NJ shows definitively that there are certain areas that are medically underserved and would benefit from building of a new hospital. This data can be used along with other factors such as real estate cost, easy access from major highways and rail stations etc. to pinpoint the ideal location for the new hospital that Cancer Hospitals USA will be opening.