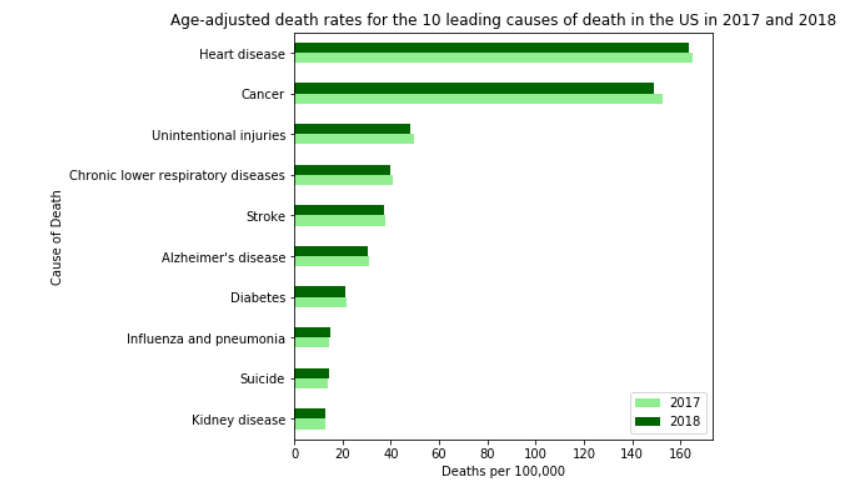
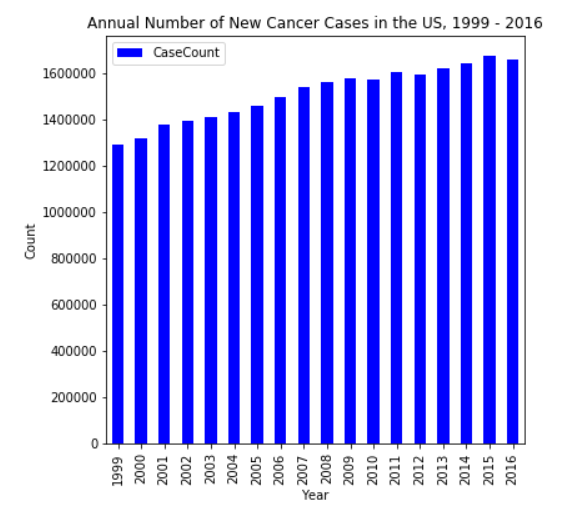
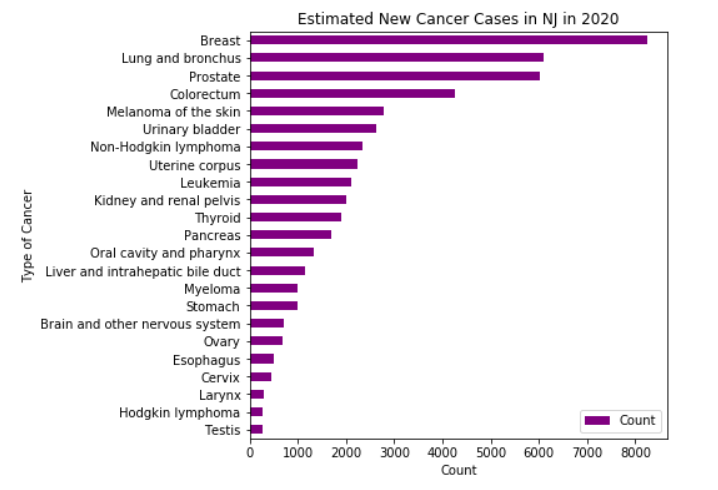
**Introduction**

**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.





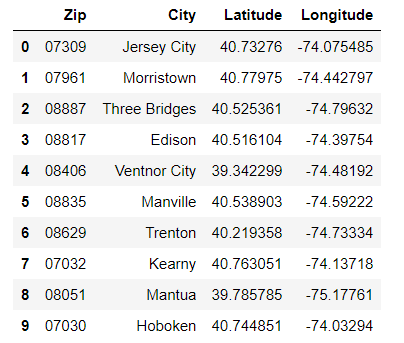
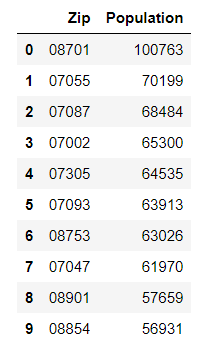
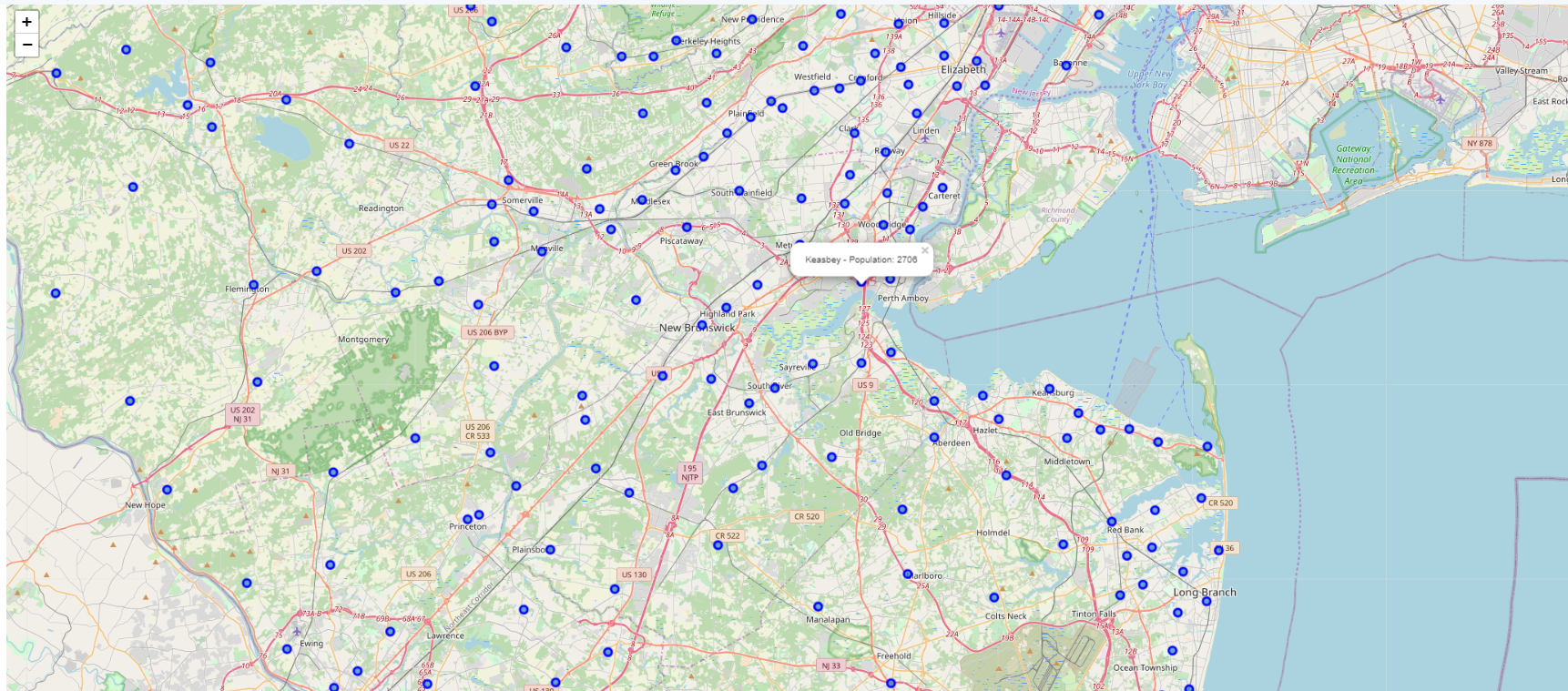
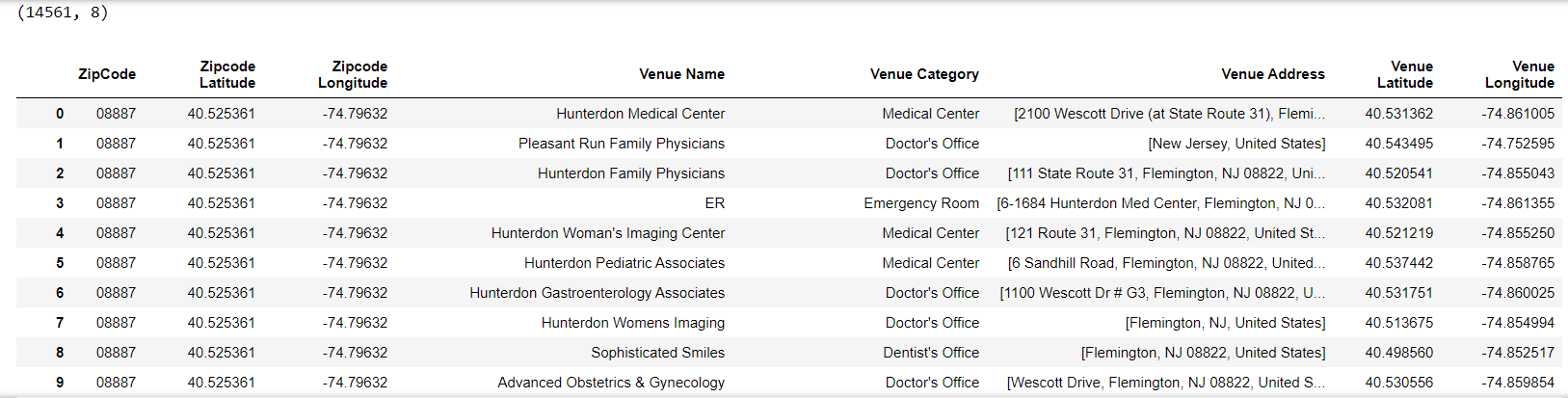
**Problem**Cancer Hospitals USA(a fictitious group) is a group of premier cancer hospitals located in several locations in the US, including in New York City (NYC). They have a large percentage of patients who live in New Jersey (NJ). It is very difficult for the patients and their families to travel 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.



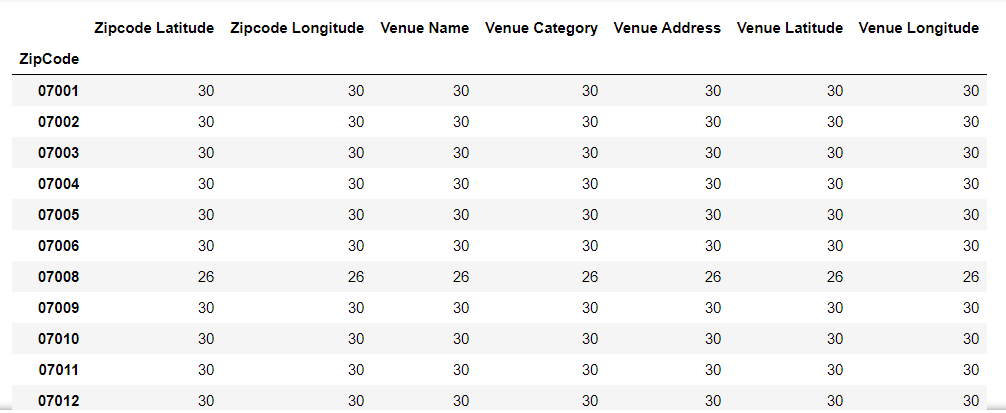
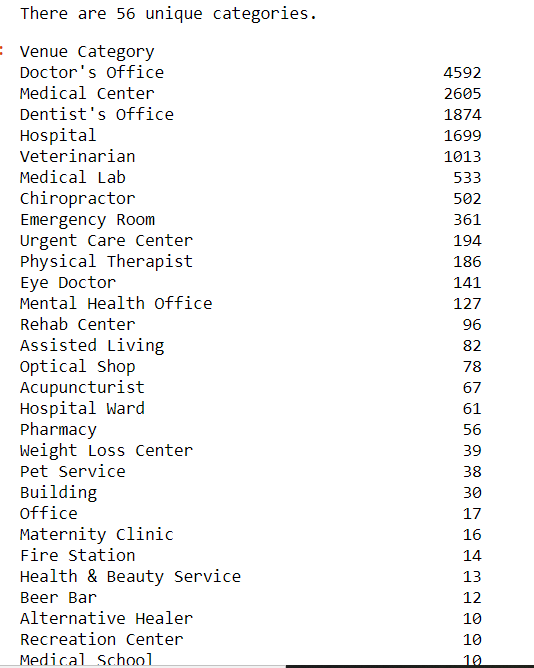
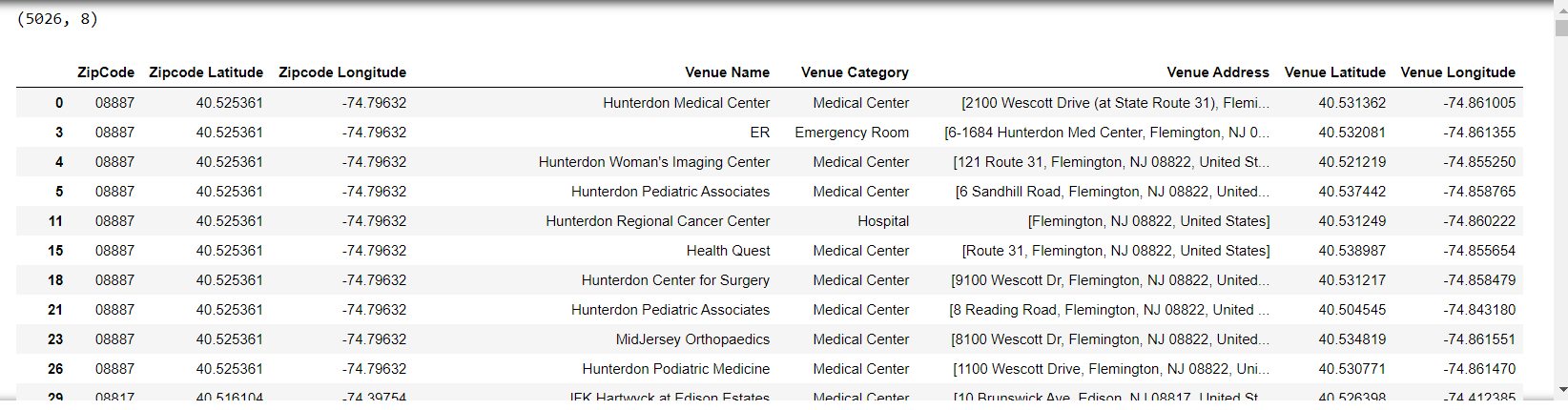
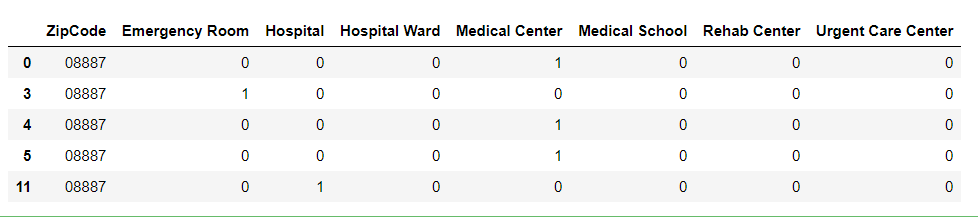
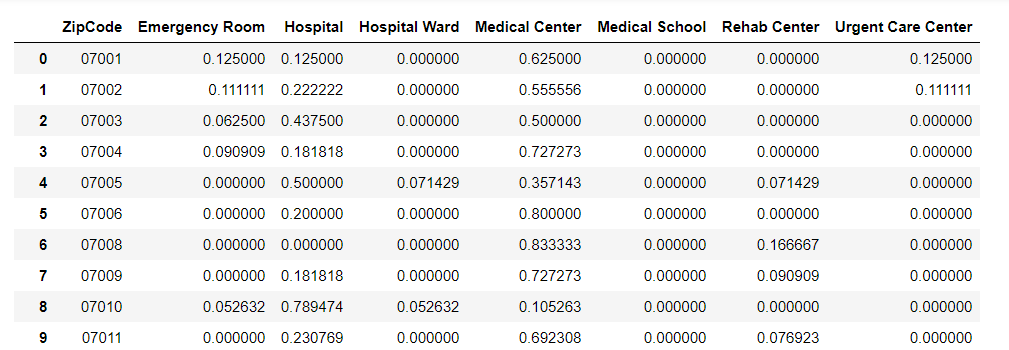
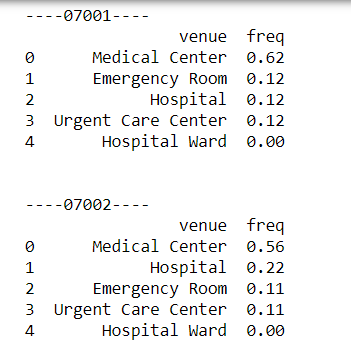
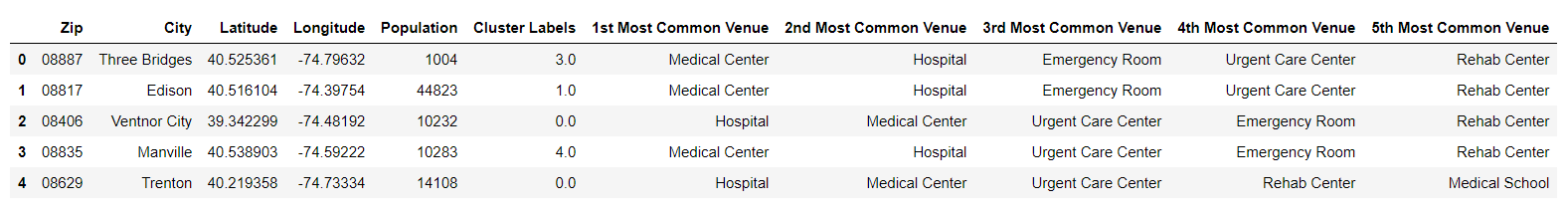
**Data Acquisition and Cleaning**

* **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/](https://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  
  
* Perform web scraping, clean, load and display the NJ population data  
  
* Merge the 2 dataframes and display the merged data  
  
* 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  
  

**Analysis of Results and Machine Learning**

* List the count of venues for each neighborhood  
  
* Count and list number of unique categories in all the venues combined  
  
* 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.  
  
* Then analyze each ZipCode using one hot encoding  
  
* Group rows by ZipCode and by taking the mean of the frequency of occurrence of each category  
  
* Use K-Means clustering to further analyze the data. Set the number of clusters to 5.
* Find the top 5 venues in each ZipCode  
  
* Create a dataframe with the venues in descending order
* 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.