Applying Data Science Methods to Cancer Research

Kaitlyn Torres

Charles J. Howard

December 6, 2023

Table of Contents

Introduction

Cancer, as a relentless and seemingly unavoidable adversary, stands as one of the most formidable challenges in modern medicine and the human experience. While families continue to grieve the losses of their loved ones, the incidence of cancer continues its upward trajectory. With an optimistic view, as the rise of cancer continues, so has our understanding of the disease and our ability to treat it. One reason for our increase in understanding, is the rise of data availability, and global efforts to research and unravel the disease. Data science plays a crucial role in cancer research by leveraging data analysis and computational techniques to better understand, diagnose, treat, and prevent the further escalation of cancer incidences.

This project employs data science and data visualization techniques to collect and analyze incidence of different types of cancer in the United States, focusing primarily on New York and Florida. Based on our analysis, we were able to discover several trends from deviations in expected rates of cancer from population rates, and possible correlations in communities and their environmental factors.

Background

Introduction to Data Science

Data Science is a field that allows us to propose and support hypotheses by organizing and studying data. By performing statistical analysis on said data, we can extrapolate insight from it to find evidence that supports or challenges a hypothesis. The process begins with

understanding a problem, and deciding on a goal for the analysis. Collecting accurate data and identifying accurate sources of data is an important step, as accurate data allows us to draw true conclusions on a topic. This data is then prepared, by cleaning and transforming it into analyzable data. It is used to build predictive models, selecting the best algorithms to use, tuning parameters, and finally validating the model.

Data Science in Healthcare

Data Science is used in many different fields and industries, such as education, the natural sciences, finance, retail, marketing, and other technology disciplines. Its daily importance is increasing everyday. Data Science allows us to make better decisions by informing us of the pros and cons of our current decisions. This quality makes its use in healthcare especially important. In Healthcare, Data Science gives us better diagnostics, allows us to detect possible issues earlier on, and lets us visualize, analyze, and improve demographic health through medical heatmaps. Using Data Science, scientists were able to successfully identify gene-specific signatures of epilepsy in children. The US Department of Defense's Defense Innovation Unit launched an initiative for utilizing AI to detect early signs of cancer in medical images. A study published in Cancer Epidemiology, Biomarkers & Prevention used a predictive analytics model to tell which patients were at a higher than normal risk of pancreatic cancer, improving prevention and early screening efforts (Mack).

Specifically, Data Science plays a huge role in advancing cancer research. Despite its fairly recent introduction to the field of oncology, data science has been used to predict the occurrence of cancer, as well as recommend treatment options for those with a certain type of cancer. A study in Korea used deep learning techniques to develop a chemotherapy

recommendation model for patients with colorectal cancer. Instead of giving a generic recommendation for those with colorectal cancer as a whole, data science was used to recommend personal treatment based on the type of cancer, how much it has advanced, and other key personal characteristics (Cronin, Patrick). Motivated by the potential that data science offers to advance cancer research, we aimed to demonstrate its capability in uncovering the reasons behind high cancer rates in a community.

Data Science in Cancer Research

Earlier studies have demonstrated the role of data science in advancing cancer research. Two studies in particular served as inspiration for us to conduct this analysis. Sivaram et al. focused on the increase in cancer incidence and mortality rates in low and middle income countries. Data science was used to extract information from a credible database, GLOBOCAN, and visually show the data in order to highlight trends and anomalies. What stood out prominently was the heatmaps that were created to show the top cancer per country, one for female and one for male. Each color represented the different type of cancer that was most prevalent in a given country. Rather than concentrating on the most prevalent global cancer types, we chose to shift our attention exclusively to the United States and refine our focus to the states that stood out.

The next research paper that we took a keen interest in focused more on how Machine Learning and Artificial Intelligence can progress cancer research through the capability to forecast early indicators of cancer. Zhang et al. concentrated on proving how different types of artificial techniques can estimate the likelihood of a certain type of cancer in an individual. Like the previous research paper, the GLOBOCAN database was used to grab valid data. After the

data was grabbed and cleaned up, it was once again visualized, this time being shown in a pie chart. The research paper continued on to explain how machine learning algorithms are now being used to analyze the unstructured data in an attempt to determine the probability of a patient contracting a specific type of cancer or other illnesses.

Due to our short time span of four months, we did not have the time, money, and ability to use broader machine learning and artificial intelligence techniques. Instead, our goal is to showcase the potential of data science in unveiling patterns associated with a variety of cancer types in specified locations, to enhance our overall understanding of the factors behind the elevated prevalence of certain cancers in particular regions. Rather than concentrating on the most prevalent global cancer types, we chose to shift our attention exclusively to the United States and refine our focus to the states that stood out, focusing on data from the States of Florida and New York and their respective counties.

Methodology

The methodology for this research outlines the advanced data science techniques and tools employed to enhance the efficiency and effectiveness of cancer research, from data collection and processing to in-depth analysis and visualization. To start off, the selection of a programming language was necessary in order to begin data extraction. The language Python was chosen, as it contains a variety of libraries for data analysis, visualization, and machine-learning techniques. Our Python code was run in Jupyter Notebook, which is an open source web application that allows code to be edited, shared, and run in realtime. As Jupyter allows you to run code line by line, it is highly effective in conducting data analysis.

United States

Moving on to data extraction, the Playwright API as well as the BeautifulSoup package were both used to retrieve text from most of our utilized sources. Playwright is an application programming interface (API) that launches a browser instance of a desired webpage.(Playwright Library) From there, the BeautifulSoup Package was used to parse through the HTML of the webpage and grab the data we wanted to analyze. For the United States, the data was always extracted from a HTML table. BeautifulSoup allowed us to find the table we wanted using a class identifier, then appending each row of the table to a list. Once finished, the list was transformed into a dataframe.

Our first source that was used was "CA: A Cancer Journal for Clinicians". By using playwright and BeautifulSoup, table 2 was grabbed which shows the estimated new cases of certain types of cancer per state for 2023. Once we converted the data to a dataframe, the data had to be cleaned up before further analysis. This included removing string characters and converting the data type to integer. Using the same technique as before, we grabbed a table from Ballotpedia to show the United States census for 2020. By grabbing the 2020 data, our objective was to assess the two dataframes to determine whether the increase in cancer rates in 2023 was attributable to population growth or other influencing factors.

Now that the two dataframes for United States data were cleaned up, we proceeded to visualize the data in a way that clearly shows any outliers. Plotly is an open source python package that is used to create a variety of interactive graphs including bar graphs, scatter plots, and pie charts ("Getting Started with Plotly."). Since we had data on all states in the United States, we were able to use plotly and create a choropleth map, which is a map of colored polygons used to represent spatial variations of a quantity (in this case number of cancer cases

and population). By looking up each state's FIP code and putting it into a list, as well as using the plotly express choropleth feature with the dataframe of our choice, we were able to create an interactive United States Map. Two of these types of maps were created; One for population and one for number of cancer cases. Each state will have a color depending on its value. If the state appears darker, then the number value is low. For example, since Wyoming has a low population, the color is purple and since California has a high population, the color is a yellow-ish color. We compared the two graphs and used a state's color change to choose points of focus.

New York

Due to the two maps, New York and Florida were the states that were chosen to take a deeper dive into. Starting with New York, we grabbed data from the Cancer Incidence and Mortality by County and Sex, 2016-2020 pdf (Department of Health). Since this was a pdf rather than a webpage, playwright and beautiful soup couldn't be used to extract the data in the table. Instead the python package tabula was used, as its main function is to read tables in a pdf. As the pdf had 62 pages for 62 counties, tabula puts the data for each table as an element for a list. Once this is done, each element in the list gets merged together to create one big dataframe with all of the counties and their data. Population for NY between 2016 and 2020 was grabbed right after from the New York State Department of Health website , using the same technique to grab the data as with the United States (playwright, beautifulsoup). Both data frames were cleaned up like the United States data, removing any unnecessary string characters and converting values to their appropriate data types. From there, two choropleth maps were created to compare population and cancer data. Since plotly express only recognizes country and state

fips rather than county, plotly's figure factory feature was used to create the choropleth maps instead.

Once we identified specific areas of focus, we started using the python package 'pandas' for data manipulation. The pandas package is an open-source, easy to use package that provides data structures and functionalities for handling numerical tables and time series. When analyzing dataframes, pandas is often the most preferred package to use. Given the NY dataframe, we were curious as to what type of cancer was most prevalent in each individual county. To determine this, the pandas idxmax method was used. The idxmax method grabs the index of the row with the greatest value. For this case, since we want to grab the column name where the highest value in the row is (type of cancer) instead of the index, we set the axis equal to 1 (dataframe.idxmax(axis=1)). All of this was then put into a new column, called Area Site. We also aimed to determine the frequency of a specific cancer being the most widespread in a count so we called the groupby.count() function on that column, which groups the data if it is the same and gets the count of each value.

Due to lack of outliers, we also decided to see which county had the greatest number of each type of cancer. To do this, the old dataframe was transposed. This means the row elements were changed into column elements and the column elements into row elements. Once the index for this new dataframe was set to the different types of cancer, we iterated through each row using the pandas function .iterrows(), and grabbed the largest value per row. The index as well as the column name that contained the largest value were appended to a list and eventually created into a new dataframe. As a result, one column was the type of cancer and the next column contained the county name that had the highest number of cases. Upon examining the

new dataframe, noticeable outliers were observed, prompting further investigation into the reasons behind specific occurrences.

Florida

Similarly to how the New York Data was extracted, tabula was used to extract the number of Florida cases between 2016 and 2020 from the Florida Cancer Data System website. On the contrary, the Population data was grabbed a bit differently. The excel spreadsheet of County Population Totals from 2010-2019 was grabbed from the United States Census Bureau. The excel sheet was cleaned up in excel, making a new column which was the total of the populations from 2016-2019. Using the pandas function .read_excel, the excel spreadsheet was able to be brought into our Jupyter notebook. Again, both the cancer data and population data were created into two choropleth maps.

Again, we were curious as to what type of cancer is most prevalent in each county. By using the idxmax(axis=1) function as well as the groupby.count() function like we did with New York, we were able to get a count of how many times each certain cancer was most prevalent. Having identified outliers on this occasion, we were once more able to conduct a thorough investigation into the reasons behind their presence.

Analysis

United States

To start our analysis, we focused on the United States first and then narrowed down our search to those states that stood out. For the United States, we first started by grabbing its population data as well as cancer rate data per state by type of cancer both in the timeframe from 2016-2020. As an overview, the United States had approximately 1,958,330 cases of cancer from any site, with California leading the way with 192,770 of those cases. The most prevalent type of cancer within these 5 years was female breast cancer with 297,790 cases, followed shortly behind by Prostate cancer with 288,300. The two least prevalent types of cancer were Uterine cervix and Uterine corpus with about 65,000 a piece. The population data showed as expected, having California, Texas, Florida, and New York being the top 4 most populated states. We decided to create two heatmaps: One showing the sum of cancer from all sites per state and the other showing each state's population. The lighter the color is for that state, the higher their total number is whether that's for cancer or population.

fig 1.1- Cancer per State USA

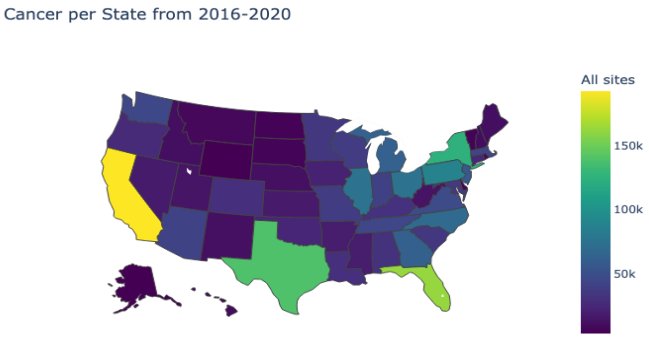
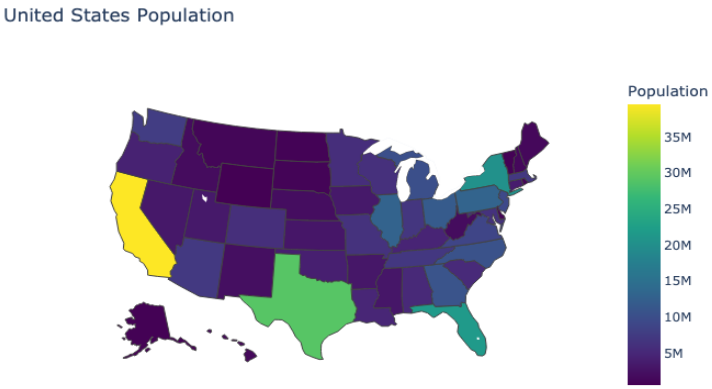


fig 1.2- United States Population



With these two graphs, we were able to spot some outliers to focus on. It's important to note that our project's primary goal is not to discover a cure for cancer; rather, it aims to

investigate the factors contributing to the high incidence of a specific type of cancer in particular regions. The first state we chose to focus on is New York, as it's our place of residence. The second state we chose to take a deeper dive into is Florida. As you can see, California's color stood consistent in population and cancer incidences, as it was yellow on each chart. Florida's, on the other hand, did not. Florida's population color was in the darker green area while its cancer rate was almost yellow,indicating that Florida has a higher than expected incidence of cancer. Given this data, we were able to ask the question as to why Florida's population is lower than Texas but its cancer rate is about 20,000 cases higher.

New York

New York has 62 counties, with an average population per year of 19.66 million from 2016-2020. As expected, the five boroughs, Suffolk, and Nassau county are all in the top 10 for population and cancer rates, with Brooklyn leading both races. In the five year time span, Brooklyn had about 12,000 cases of cancer. The most prevalent type of cancer in all counties was Lung and Bronchus cancer. During the entire five-year period, the state accumulated a total of 13,944 cases. The least prevalent type of cancer in New York was Hodgkin Lymphoma with 619 cases. Just like we did for the United States, two heatmaps were created in an attempt to spot any outliers between the two. The first map shows the total population while the second shows the number of cancer cases in New York from 2016-2020.

fig 2.1 - Count of Cancer NY

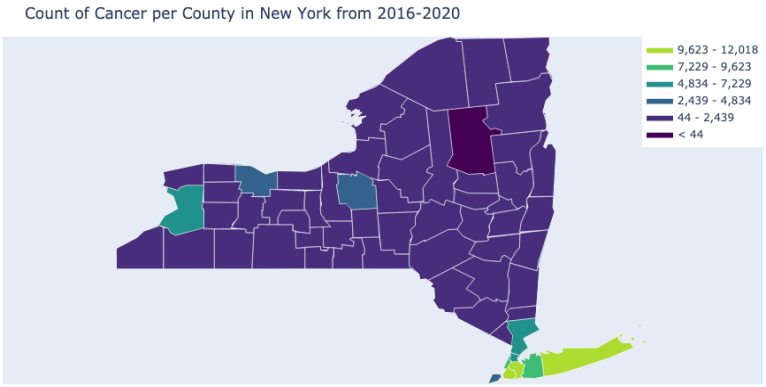
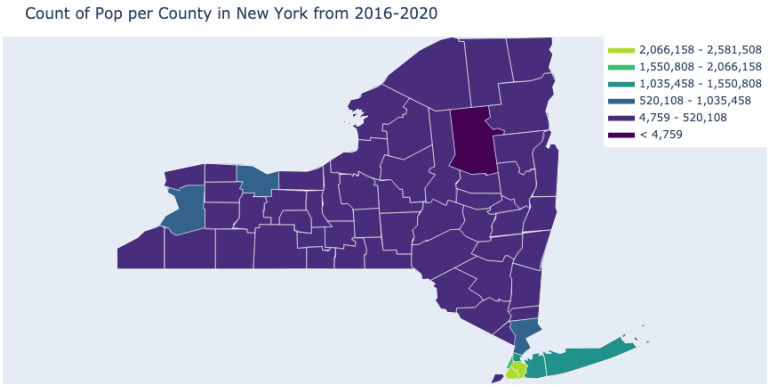


fig 2.2- Population of NY



With these two heatmaps, there are some evident outliers that warrant a closer look. As expected, most of upstate New York is purple for cancer, which means low cancer rates. This is because the population upstate isn't as big as it is in the five boroughs. The county that stood out the most was Suffolk County. As shown in **fig 2.2**, Long Island's population is in the 1.0-1.5 million population range, meaning its color is a darker green.When you look at **fig 2.1,** Long Island is in the top 3 for cancer rates, especially Suffolk county having a color of bright green on the map.

As further evaluation continues, the main focus was to understand why Suffolk County's cancer rate is so high, when their population isn't as high as some other counties. To do so, we grabbed each type of cancer and found the county that had the greatest number of cases for the certain type. Initially, we predicted that Brooklyn and Queens would lead the way, as they were the two most populated counties. What we found was that Brooklyn had the most cancer cases for 12 different types of cancer, followed by Suffolk with 4 and then Queens with 2. When we took a deeper dive into these 4 types of cancer, we saw that it was bladder, lung, melanoma, and leukemia.

Melanoma is typically caused by exposure to UV radiation. Anyone can get melanoma but it is more common in fair-skinned people (National Library of Medicine). According to Census.gov, Suffolk county's population is 82.8% White. The high number of Melanoma cases in Suffolk County may be attributed to this factor, along with its abundance of over 20 beaches.

Several factors, including smoking and exposure to certain chemicals, can contribute to the development of bladder cancer, leukemia, and lung cancer. In 2018, Centereach, Farmingville, and Selden had increased incidences of Bladder, lung cancer, and leukemia than the rest of the state. According to Citizens Campaign, these towns in Eastern Suffolk County "which have a combined population of about 65,000 — had a leukemia rate 51% higher than elsewhere in the state, a thyroid cancer rate 46% higher, a lung cancer rate 40% higher and a urinary bladder cancer rate 30% higher" (Winters, Kristina). Considering the potential link between these cancers and chemical exposure, we chose to explore whether there are any chemicals more commonly used in Suffolk than in any other part of the state. After further research it was discovered that Suffolk County utilizes the most pesticides annually among all the counties in New York. For example, in 2021, approximately 6.5 million pounds of pesticides were used in Suffolk county alone when most counties use less than 100,000 pounds (Gangemi, Christopher). The drastic amount of chemicals used here may cause or elevate the risk of developing these types of cancers.

Florida

Florida has 67 counties, with an average population per year of 21.21 million from 2016-2020. With the majority of the cancer rates being down south, Miami-Dade triumphs with about 66,000 cases. Neighboring counties Broward and Palm Beach follow with about 45,000 a

piece. The cancer that is most prevalent in Florida from 2016-2020 is Breast Cancer with 88,433 cases followed shortly after by Bronchus Cancer with about 86,000 cases. Cervix cancer is the least prevalent with about 4.5 thousand. Again, to visualize our analysis, a heatmap was created to show the rates of cancer per county in Florida, as well as population between 2016 and 2020.

fig 3.1 - Count of Cancer FL

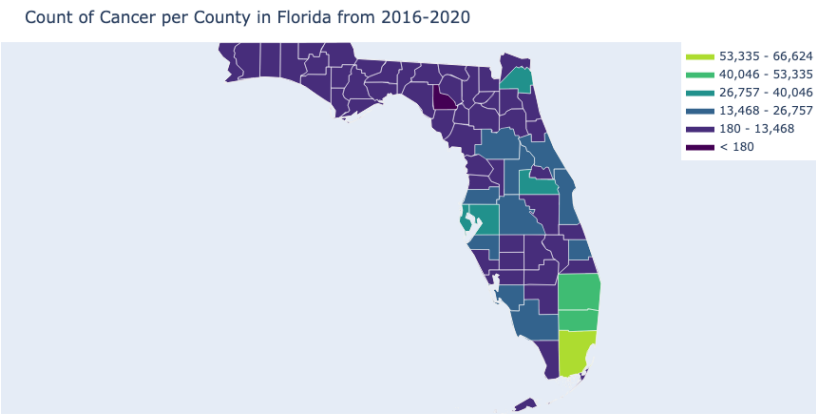
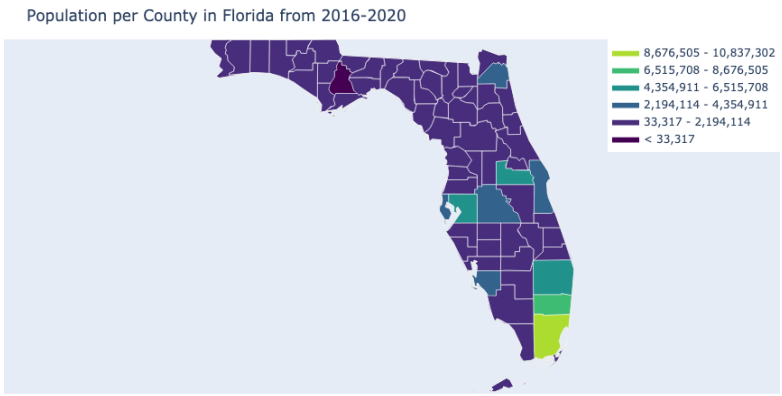


fig 3.2 - Population in FL



As you can see here, the majority of the counties with a greater number of cancer cases as well as the higher population reside in southern Florida. Unlike the United States and New York heatmap, there weren't any outliers that stood out initially. Due to this, we decided to dig deeper into the underlying data. Firstly, we grabbed each cancer site and got the count of how many counties that specific cancer is the most prevalent in. Results showed that Bronchus was most prevalent in 50/67 counties, Breast was most prevalent in 15/67, as well as Melanoma and Prostate being the most prevalent in one county each.

Having acquired the data, our next objective is to comprehend the reasons behind its occurrence in certain areas. Starting with Melanoma, the number one risk factor is UV radiation. In Florida, the region with the highest UV index is the southwest region, whose UV index

typically reaches an index of 12. According to the United States Environmental Protection Agency, an UV index of 11 or higher is considered “extreme” (Naples Center for Functional Medicine). The county that had Melanoma as its most prevalent cancer is Glades County. Due to Glades being located in the southwest region of Florida and considering that this area typically experiences a notably high UV index, the increase in melanoma cases there is understandable. A lot of the other counties in this area have a higher incidence of Bronchus cancer than Melanoma. Glades County is the third smallest county in Florida, both in population and cancer cases. Given its small population, this county has fewer smokers compared to its neighboring counties, which as a result lowers the incidence of bronchus cancer.

As of right now, it is not clear as to what exactly causes Prostate Cancer. The only known fact regarding prostate cancer is its exclusive occurrence in men. The county where this type of cancer is identified as the most common cancer is Union County. The number of Prostate cancer cases is 189 from 2016-2020 followed shortly after by Bronchus with 183. The only reason as to why Prostate is higher in comparison to other counties is because of its male to female ratio. While the other counties with a predominantly male population have a 100 to 97 male to female ratio, Union county has an estimated 183 to 100 male to female ratio. Given that this cancer exclusively affects males, the notable surge in the male population could account for the high prevalence of this cancer in this region.

Most counties in Florida experience bronchus cancer as the most commonly occurring type of cancer. For this type of cancer, the number one risk factor is smoking. Florida has a 14.5% smoking rate, with Putnam county having a 25% smoking rate on its own. Not only does the high smoking rate contribute to the high number of lung and bronchus but also does the lack of screening. According to the American Lung Association’s “State of Lung Cancer Report”,

“Florida ranks among the states with the lowest lung cancer screening rates” (American Lung Association). While the national rate of screening for those at high risk of lung and bronchus cancer is 4.5, in Florida this rate drops significantly at 2.4%. In conclusion, the most widespread occurrence of Bronchus cancer across the 50 counties is a direct outcome of both smoking and a lack of proper cancer screening.

Like Prostate cancer, Breast cancer can occur due to an increase in age and gender. Whereas Prostate cancer can’t occur in females, there are rare occurrences where males may have breast cancer. In 15 out of 67 counties, breast cancer stood out as the most widespread form of cancer. In 14 out of 15 counties, the explanation for the increase in breast cancer cases was clear-cut given the higher ratio of females to males in these areas. The one county, Sumter, stands out due to its higher male population, prompting curiosity about the elevated breast cancer rate in the area. As further research was done, we discovered that Sumter County is ranked best in the state for smoking with the lowest recorded rate at 12%. As a result of this, the incidence of bronchus cancer in this county is notably reduced, making breast cancer the predominant type by default.

Conclusion

Data science introduces a new realm of opportunities to the field of oncology. The integration of data science techniques for identifying outliers in datasets has proven to be a transformative tool in advancing cancer research. By leveraging sophisticated algorithms and analytical methods, data scientists can pinpoint anomalies in diverse datasets related to cancer, shedding light on potential outliers that may signify unique patterns or factors. This significantly

enhances the accuracy of cancer research, enabling researchers and healthcare professionals to focus on specific areas, discover new indicators, and ultimately deepen our understanding of the complex mechanisms underlying various types of cancer. As data science progresses, its crucial role in uncovering anomalies is expected to persist as a fundamental aspect in the ongoing efforts to improve cancer diagnosis, treatments, and overall outcomes.

Bibliography

- *American Lung Association. "State of Lung Cancer: Florida." State Data | Florida | American Lung Association, www.lung.org/research/state-of-lung-cancer/states/florida#:~:text=Screening%20for%20High%20Risk%3A,the%20national%20rate%20of%204.5%25. Accessed 28 Nov. 2023.*
- *Bureau, US Census. "County Population Totals: 2010-2019." Census.Gov, 29 Mar. 2023, www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.*
- *Cancer Incidence and Mortality by County, New York State*, www.health.ny.gov/statistics/cancer/registry/pdf/volume1.pdf. Accessed 29 Nov. 2023.
- Cronin, Patrick. "How Data Science Is Helping Oncologists Battle Cancer." *Health IT Answers*, 27 Apr. 2023, www.healthitanswers.net/how-data-science-is-helping-oncologists-battle-cancer/.
- "Department of Health." *Average Annual Population of Counties, New York State, 2016-2020*, www.health.ny.gov/statistics/cancer/registry/appendix/countypop.htm. Accessed 28 Nov. 2023.
- Gangemi, Christopher. "Dubious Distinction for Suffolk; It Leads the State in Pesticide Use." *Dubious Distinction for Suffolk; It Leads the State in Pesticide Use | The East Hampton Star*, www.easthamptonstar.com/government/202346/suffolk-leads-state-pesticide-use. Accessed 28 Nov. 2023.
- "Getting Started with Plotly." Plotly.com, plotly.com/python/getting-started/.

- “Florida Statewide Cancer Registry.” *The Florida Cancer Data System - Publications*, fcds.med.miami.edu/inc/publications.shtml. Accessed 28 Nov. 2023.
- National Library of Medicine. (n.d.). What increases your risk of melanoma? - informedhealth.org - NCBI bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK321118/
- Sheridan, Terry. “Elevated Cancer Rates in 3 Suffolk Towns Prompt Study.” *WSHU*, 19 Oct. 2021, www.wshu.org/news/2018-07-19/elevated-cancer-rates-in-3-suffolk-towns-prompt-study.
- Siegel, Rebecca L, Kimberly D Miller, Nikita Sandeep Wegel, et al. *Cancer Statistics, 2023 - Siegel - Wiley Online Library*, Acs Journals, acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21763. Accessed 29 Nov. 2023.
- Sivaram, Sudha, et al. “Building Capacity for Global Cancer Research: Existing Opportunities and Future Directions.” *Journal of Cancer Education : The Official Journal of the American Association for Cancer Education*, U.S. National Library of Medicine, July 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8285681/.
- Mack, Jonathan. “How Data Science Is Reshaping Healthcare.” *University of San Diego Online Degrees*, University of San Diego , 25 July 2022, onlinedegrees.sandiego.edu/data-science-health-care/.
- Naples Center for Functional Medicine .*Surviving Florida’s never-ending summer sun*. Naples Center for Functional Medicine. (2023, October 23). https://naplescfm.com/uncategorized/surviving-floridas-never-ending-summer-sun/#:~:text=Southwest%20Florida%20often%20reaches%20an,skin%20is%20to%20stay%20indoors.

- "Playwright Library." *Playwright*, playwright.dev/docs/api/class-playwright. Accessed 1 Dec. 2023.
- "United States Census, 2020." *Ballotpedia*, ballotpedia.org/United_States_census,_2020. Accessed 28 Nov. 2023.
- "UV Index Overview | US EPA - U.S. Environmental Protection Agency." *United States Environmental Protection Agency*, www.epa.gov/enviro/uv-index-overview. Accessed 29 Nov. 2023.
- Winter, Kristina. "High Cancer Rates in 3 Suffolk Communities Discussed." *Citizens Campaign for the Environment*, Citizens Campaign for the Environment, 28 Mar. 2020, www.citizenscampaign.org/whats-new-at-cce/2019/11/13/high-cancer-rates-in-3-suffolk-communities-discussed.
- Zhang, Bo, et al. "Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach." *Journal of Multidisciplinary Healthcare*, U.S. National Library of Medicine, 26 June 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC10312208/.