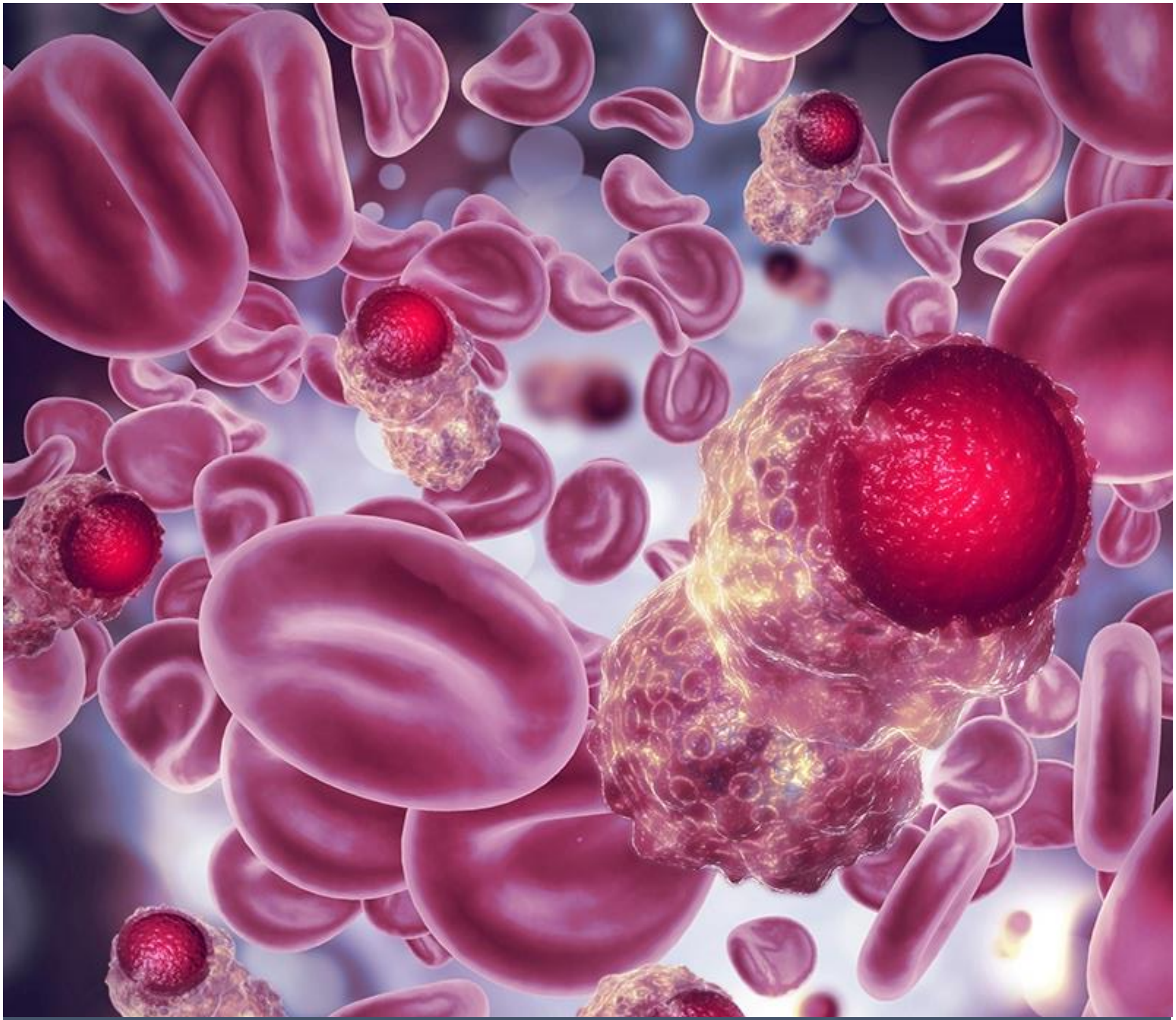


Analysis and Visualization of Cancer factors in the US on Macro and Micro Levels

An Analytical report for RRH to identify key business opportunities

Muhammad Khizar Hayat Tahir MGIS-650 Data Discovery Project

Objective:

To explore, analyze and visualize factors contributing to Cancer Incidence and Death in the United States, and more specifically in the state of New York; and make recommendations to Rochester Regional Health (RRH) on where they can implore business opportunities.

Executive summary:

In this project, we will use a dataset containing information about factors leading to cancer occurrence and also deaths in the US, arranged according to zip codes. Our possible Response variables are Incidence Rate and Death Rate. We will **explore** and **visualize** their relationships with our most important factors to see if we can make conclusions about patterns and create an action plan.

There are 2 broad parts to this study:

1. In the first, we observe relationship between our Response variable(s) and most important factors at a National level.
2. In the second, we will focus specifically at the State of New York which is where RRH can act promptly.

From our earlier research, we determined the following as our most important factors:

- *Recent Trend:* recent trend of incidence of cancer
- *Median Income:* the median household income by county.
- *Population Estimate:* the number of people in a county.
- *Study count:* the number of cancer clinical trials held for all types of cancer by Zip Code.
- *State:* States in the United States.

Problem to solve:

Just looking at the dataset gives no information about trends and patterns Nation-wise or State-wise. Knowing which factors are important, we want to **visualize** their effect on Cancer Incidence and/or Death to see if we can observe any patterns which may be significant for us. Based on visual proof, we can generate insights and create a plan-of-action for RRH.

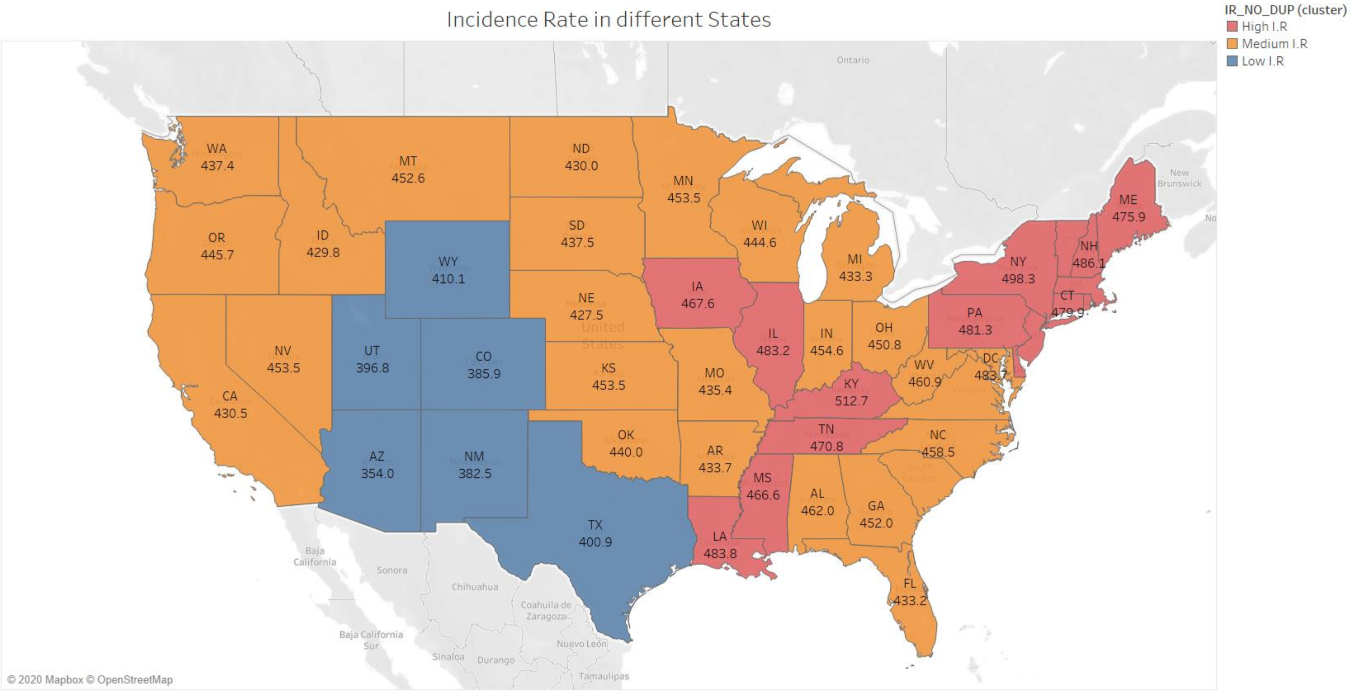
Data cleaning and standardization:

- Some of our factors were Zip Code – wise, others were County – wise. We treated our variables to remove any duplicates.
- We filtered out Alaska from our analysis to remain focused on our region of analysis.
- Some continuous variable like Incidence Rate, Median Income, and StudyCount were recoded as categorical variable to understand impact better. The breakpoints in categories were custom defined to ensure accurate representation and reduce impact of outliers.
- There are about 5 kinds of recent trends. We will consider only Falling and Stable as they are relevant to our research.

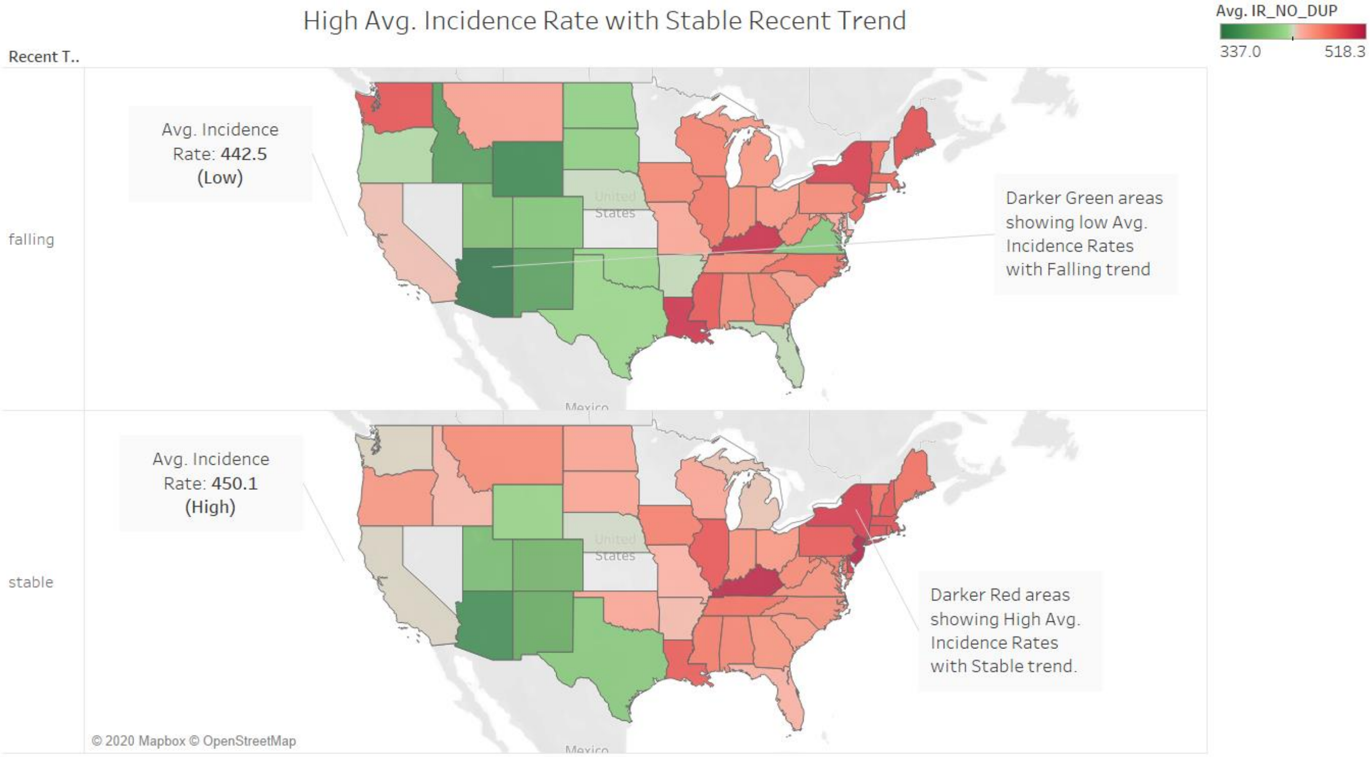
Analysis:

Part 1: Analyzing Cancer Factors across different regions in the US

(a) We first visualize how Average cancer Incidence rate varies by State to know which regions are most prone to cancer. We clustered the Incidence Rate as High, Medium, or Low and visualized it against all the States except for Alaska to remain focused on our region. We observe that High Incidence rate (Red) is only in North East and South East regions. These regions are **most prone** to cancer. Majority of US has Medium Incidence rate (Yellow). South West has the lowest (Blue) average Incidence Rates.


Incidence Rate in different States

(b) Plotting 2 graphs, one for each of our chosen Recent Trend, of Avg. Incidence Rate in each State (depicted by varying color), we observe that for the Recent Trend of "Stable", the Avg. Incidence Rate is higher (more darker red areas) than for the Recent Trend of Falling (more darker green areas).


High Avg. Incidence Rate with Stable Recent Trend

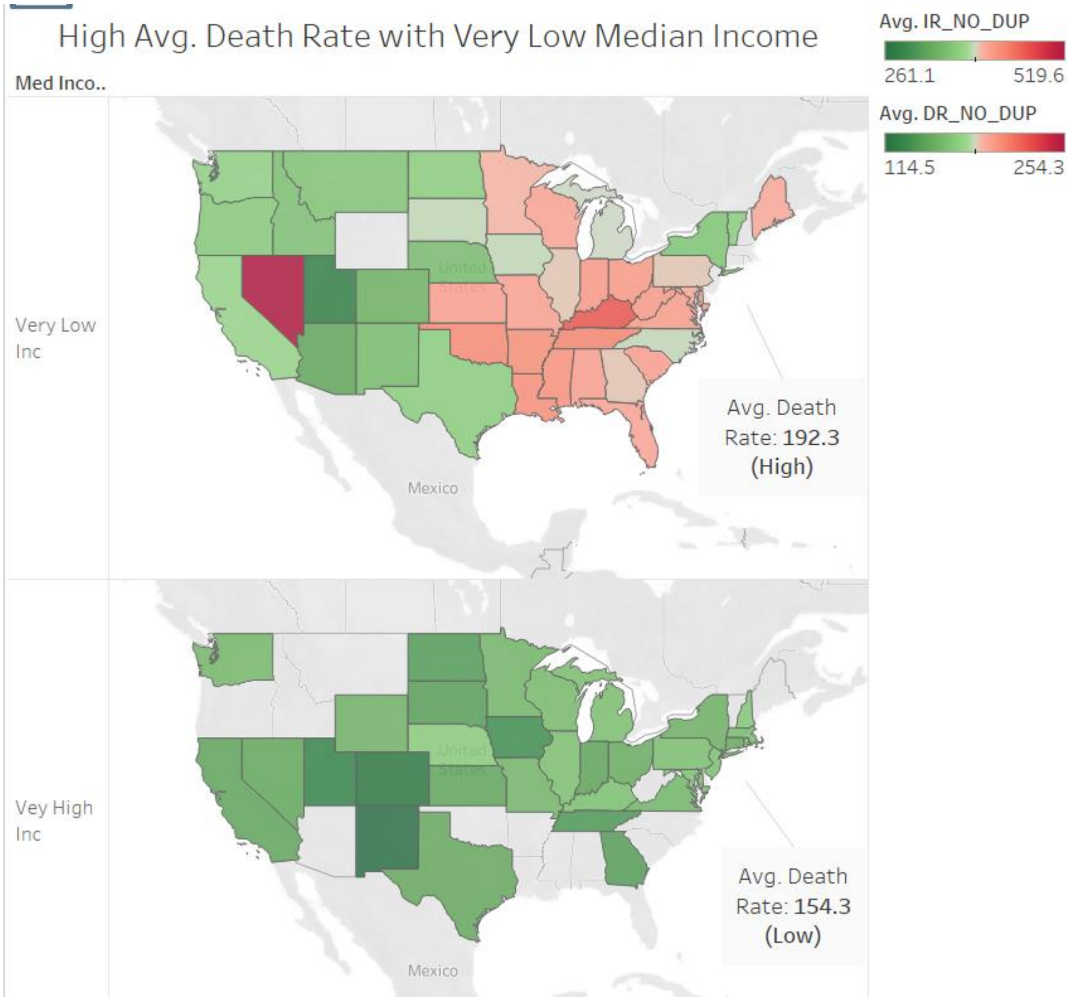
Avg. Incidence Rate: 442.5 (Low)

Darker Green areas showing low Avg. Incidence Rates with Falling trend

Avg. Incidence Rate: 450.1 (High)

Darker Red areas showing High Avg. Incidence Rates with Stable trend.

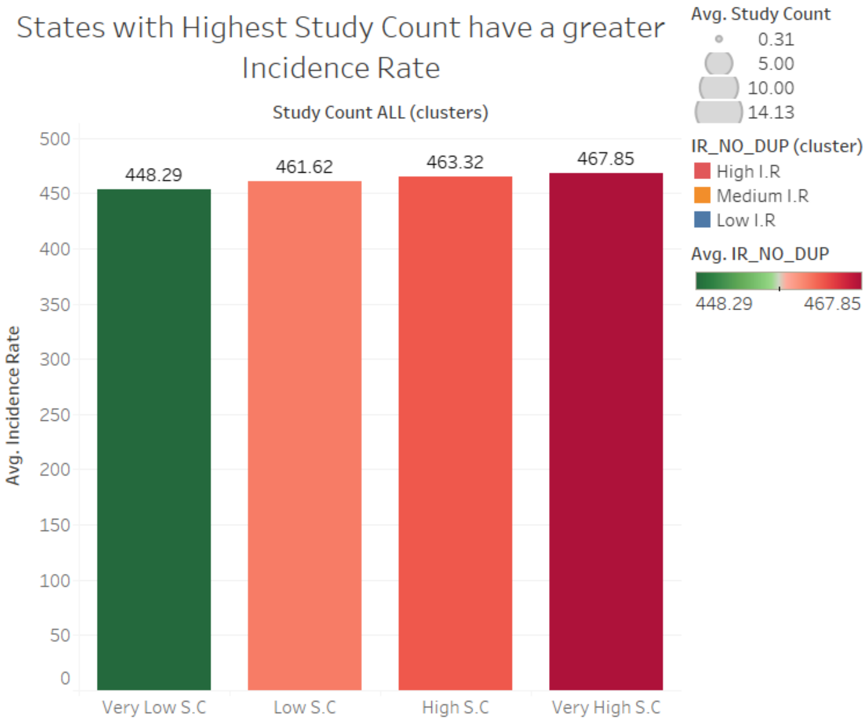
(c) Although a High Median Income positively correlated with a High Avg Incidence Rate, it negatively correlated with Avg. Death Rate. This means that counties with high median incomes were more likely to have cancer but less likely to suffer death from it then counties with low median inc. A plot with median income clustered in 4 categories: Very Low, Low, High, Very High with only extreme categories is shown on the right. Very Low Income graph has many Red regions depicting high death rate. Very High income section, however, has no Red regions depicting Low Death Rates.


High Avg. Death Rate with Very Low Median Income
Med Inco..
Very Low Inc
Vey High Inc
United States
Mexico
Avg. Death Rate: 192.3 (High)
Avg. Death Rate: 154.3 (Low)
Avg. IR_NO_DUP
261.1
519.6
Avg. DR_NO_DUP
114.5
254.3

(d) Neither the Incidence Rate nor the Death Rate showed any correlation with the Population size. Spread of viral diseases are often correlated with population size. Actual causes of cancer are more cellular hence do not show correlation. The factors we are studying can possible "affect" rather than directly cause it.

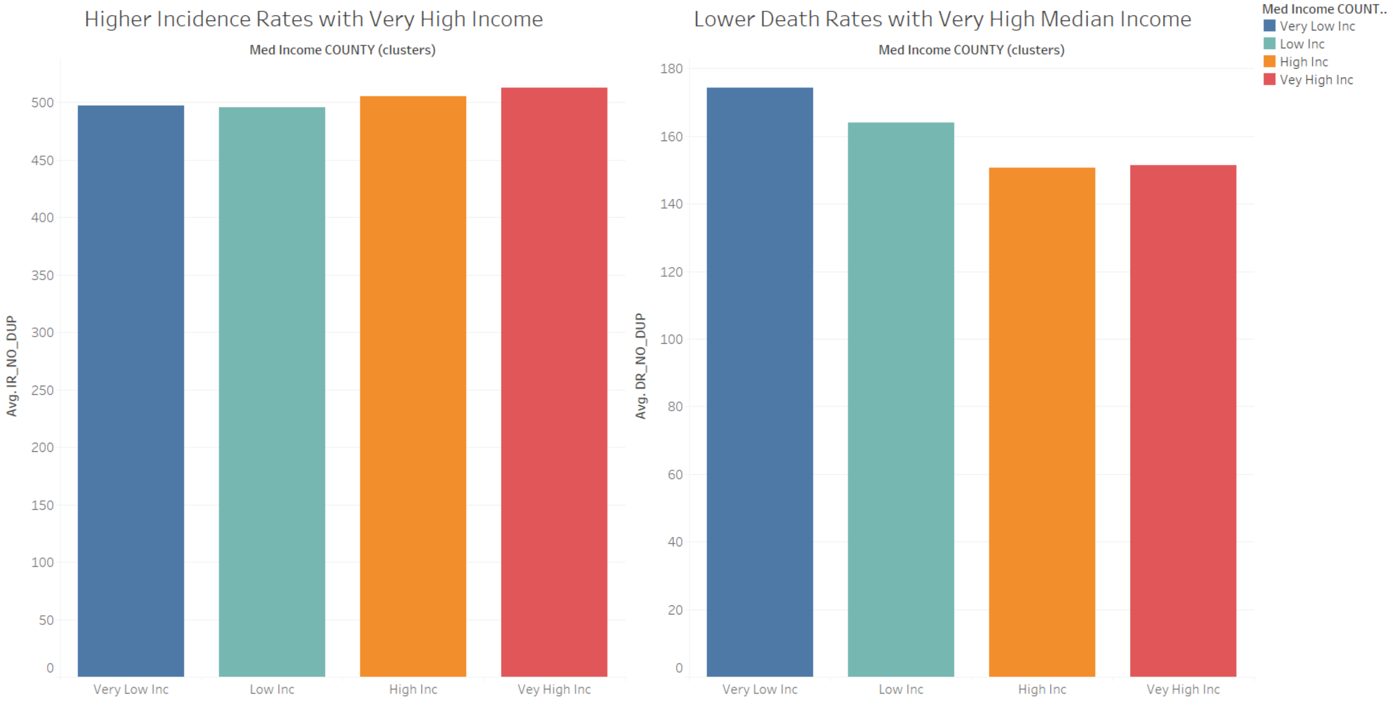
(e) A graph of categorized study count with the Avg Incidence Rate showed a higher StudyCount to be positively correlated with Incidence Rate. States with Very Low StudyCount had a lower Avg Incidence rate (Green); and those with Very High StudyCount had a Higher Avg Incidence Rate (Red).

This is possibly because "generally" the States with a greater cancer incidence rate would be hot places for cancer study and clinical trials.

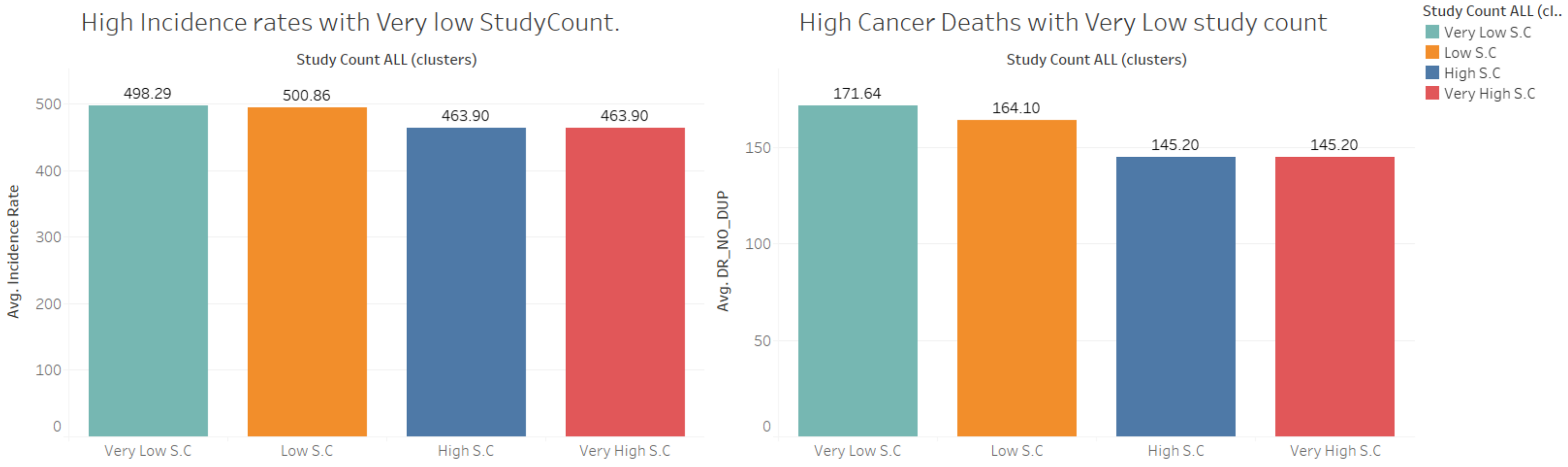


Part 2: Analyzing Cancer Factors specifically in New York State

(a) A graph of Avg. Incidence rate and 4 Median Income categories specifically for New York (see below) shows Positive correlation with Incidence rate but Negative correlation with Death Rate. This is in line with our Nation-wide observation in part 1(c). Again, we conclude that counties with high median income may be more susceptible to cancer occurrence but have a significantly less chances of death compared to low income counties.

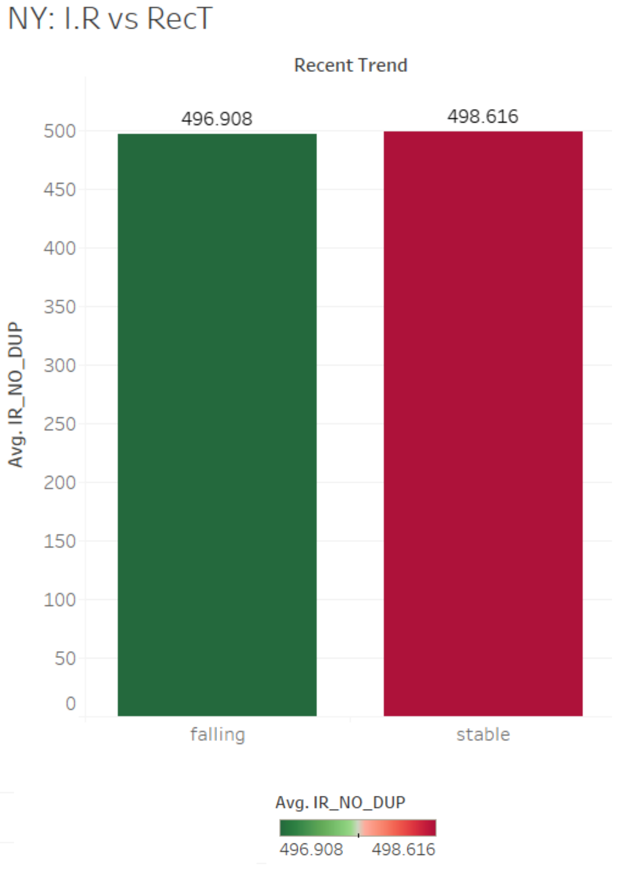


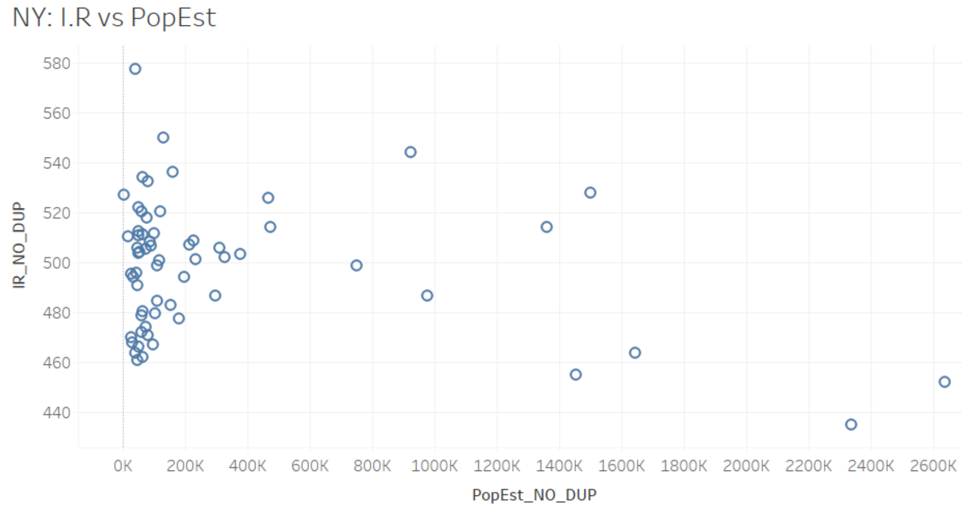
(b) A graph of categorized StudyCount shows a Negative correlation with Incidence and Death rates in New York State (see below). This is counter-intuitive. From the analysis of the whole US (part 1(e)), we identified a positive correlation between Incidence Rate and StudyCount. This may be because, in New York, the counties conducting clinical trials are major counties comprising major cities where incidence and death rate may be lower due to other factors like high income, or greater proportion of younger population. This might be resulting in a low incidence rate in those counties.



(c) Recent Trend seems to marginally impact the incidence rate in New York. From the figure on the right, we can see that the Stable recent trend has a slightly higher (Red) Avg Incidence rate than Falling – like in part 1(b).

(d) Population size does not show clear trends with Incidence rate. However, extremely high Population counties have low Incidence rates (see below). This might be because Counties with extremely high population might have a younger population and rotation of people residing.





We can conclude that the Median Income, Study Count, and RecentTrend are **most important Cancer factors in New York State**; and Population size can be explored more to unearth any hidden relationships.

Insights and Plan-of-action:

- We observed that counties with High Median Income had a greater Avg Incidence rate but lower Avg Death rate. This is because in counties with greater income, more people have access to insurances and health facilities hence more (and timely) detection of cancer. Because of early diagnosis and better healthcare facilities, the death rate in high income counties is less. With this understanding, we can conclude that although our graph shows a low Avg. Incidence rate with Low Median Incomes, it is the lack of detection that causes this rather than an actual low Incidence rate. The fact that Avg Death rate is high with Low Incomes is an evidence that the incidence must be high too – it is just undetected.
  RRH should focus on Counties with Low Median Incomes as undetected cancer cases must be high there. They should set up awareness campaigns and encourage people to utilize insurance and healthcare services if possible. They should also spread awareness about cancers which can be self-assessed like breast cancer.

- In New York State, we noticed that a greater studyCount correlated with lower Avg Incidence rate. A greater studyCount means that more cancer clinical trials are being conducted. We expect that the counties conducting most clinical trials are major counties comprising major cities. These cities have a majority of people who are younger and with high median incomes. These other factors might be contributing to a lower incidence rate in those major counties.
  RRH should keep their smaller centers in smaller counties involved in the clinical trials because those areas would probably have high incidence rate. As our results suggest, less studyCount correlated with high incidence rate. Since RRH has increased its community footprints and has better outreach to communities previously

poorly studied, it should use its smaller centers/clinics in such communities to enroll patients into clinical trials. Doing so will not only help study cancer better but will also help improve clinical outcomes among those patients.

- Since a Stable recent Trend correlated with higher Avg Incidence Rate, RRH can focus more on counties with a Stable trend and invest more in improving preventive health services there while educating those communities on lifestyle medicine.

- We observed that counties with extremely large population had low Avg Incidence rates. Populous counties have greater proportion of younger population because of more job and education opportunities. And, there is a greater flux of people in major counties. We are more likely to detect cancers in stable communities where cohorts live for a longer duration. These two factors might be contributing to lower Incidence rates in highly populous states.
  RRH should collaborate with other healthcare systems across the US to develop better follow up mechanisms and continuity in care. This would ensure that movement of people around would not cause incorrect counts; and that people who are prescribed screening or other follow-up tests are not lost after moving. Secondly, RRH should start awareness campaigns for patients coming into or out of their system for better continuity of care so people with potential malignancies are not lost to follow up.

Post-analysis summary:

In this project, we explored, visualized, and analyzed multiple factors related to cancer; identified the most important ones nation-wide and in New York State; and made recommendations based on our analyses of what RRH can do in response to them. This analysis is not all-encompassing. There are several important factors that were not explored but are known to have correlation and even causation like smoking, obesity, U-V exposure, mean age, etc. These factors can also be studied to enhance our understanding. Also, only one state was focused individually. Other States of varying demographics can also be studied to improve results.

Link for All visualizations:

https://prod-useast-b.online.tableau.com/#/site/mgis6502201/workbooks/11168?:origin=card_share_link