"Know your enemy and know yourself, and you can fight a hundred battles without disaster." - Sun Tzu. While Sun Tzu was referring to military strategy, this quote is relevant to cancer analysis as well. By understanding the characteristics and patterns of cancer incidence and mortality in different regions, we can better equip ourselves to fight this disease and ultimately improve cancer outcomes.

The problem addressed in the executive summary is the need to understand the prevalence of cancer in different regions and the factors that influence its incidence and mortality rates. This information is critical for public health services to take appropriate action, allocate resources effectively, and provide better cancer care. Additionally, there is a need to develop accurate regression models that can predict cancer incidence and death rates, which can assist in making informed decisions about cancer prevention and treatment.

Overview

The executive summary is organized into five sections. Firstly, it provides a detailed analysis of cancer incidence and mortality rates based on location, including state-wise, region-wise, and county-wise data. This analysis offers valuable insights into the prevalence of cancer in different areas, enabling public health services to take necessary action at higher severity regions. Secondly, the summary explores the relationship between income level and cancer rates, taking into account factors such as poverty and income that can affect cancer incidence and mortality rates. The third section examines the correlation between various factors and cancer incidence and death rates to identify which factors are most strongly associated with cancer. Furthermore, the summary presents the key findings of the analysis and recommendations, which can be used to guide public health policies, resource allocation, and cancer care delivery. The report concludes with the presentation of multiple regression models aimed at predicting cancer incidence and death rates, followed by the selection of the best model.

---

1. Location-wise investigation

State-wise Analysis:

By understanding the variation in cancer rates across different states, healthcare providers and policymakers operating in the state level can implement targeted cancer prevention and control measures. For instance, states with high rates of cancer can focus on increasing cancer screening and early detection programs. The state-wise summary statistics of incidence rate is presented below:

Average incidence rate	512.74	502.13	498.29	494.75	486.14	483.83
Standard Deviation	40.38	22.02	26.35	37.89	25.27	34.89
State	KY	DE	NY	NJ	NH	LA

Table 1: Top 5 State wise Average incidence rate

According to Table 1, the state of Kentucky has the highest cancer incidence rate compared to all other states in the US. Following Kentucky, states such as Delaware, New York, New Jersey, New Hampshire, and Louisiana also have a high prevalence of cancer. However, it should be noted that the standard deviation for Kentucky is high, which suggests that the incidence rate of cancer has more variance (widely dispersed) in Kentucky when compared to other states.

Region-wise Analysis:

Region-wise analysis helps to identify the prevalence of cancer in specific regions and sub-regions, allowing for targeted supervision to reduce cancer risk and improve outcomes.

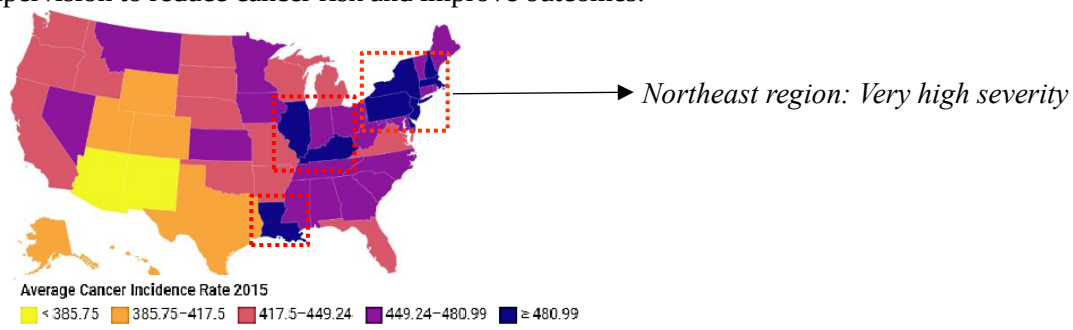

Figure 1: State wide cancer incidence per 100K people

The Northeast region of the US had a significantly higher incidence of cancer in 2015, with a staggering average rate of more than 480 cases per 100K people, compared to the Western and Central states which had a lower incidence of cancer. Interestingly, four out of the six states with the highest cancer rates (from table 1 - DE, NY, NJ, NH) belong to the Northeast region, indicating an intersection between the state-wise and region-wise analysis.

County-wise Analysis:

Summary statistics – Florida County		
Statistics	*Cases*	*County, state*
Minimum cases per 100K	336.6	Hamilton County, FL
Maximum cases per 100K	1206.9	Union County, FL
Range	1005.6	-
Mean	448.28	-

Table 2: County wise summary statistics

Based on Table 2, Union County in Florida had the highest incidence of cancer with 1206.9 cases per 100,000 people, while Hamilton County in Florida reported the lowest number of cases at 336.6. This substantial difference between the two counties highlights the significant variation in cancer incidence rates within the same state. To reduce the burden of cancer, counties in a particular state with high incidence rates should take more proactive measures to prevent and control the disease. In contrast, counties with lower incidence rates may require fewer interventions to reduce the incidence and mortality rates of cancer.

---

2. Incident rate and death rate analysis for four income levels

For this analysis, the US population is classified into four groups based on the median income level of individuals, which are presented as follows:

Income level	**Range** (*median income - $*)	**Statistics**
Very High	52,512 – 125,635	Q3 - Max
High	45,201 – 52,512	Median – Q3
Low	38,873 – 45,201	Q1 - Median
Very Low	22,640 – 38,873	Min – Q1

Table 3: 4-level indicator variable for median income

According to figure 2-a below, individuals whose incomes fall between the high and low-income brackets (approximately ranging between 38K – 52K) are at a higher risk of developing cancer compared to those with very high and very low-income levels, indicating that the middle class is more susceptible to cancer. People with lower and medium income levels may have higher exposure to environmental risk factors, such as pollution and toxic substances, which can increase their risk of developing cancer. Additionally, individuals with lower to medium income may have less access to healthy food options and may engage in unhealthy behaviours, such as smoking and physical inactivity, which can also increase their risk of developing cancer.

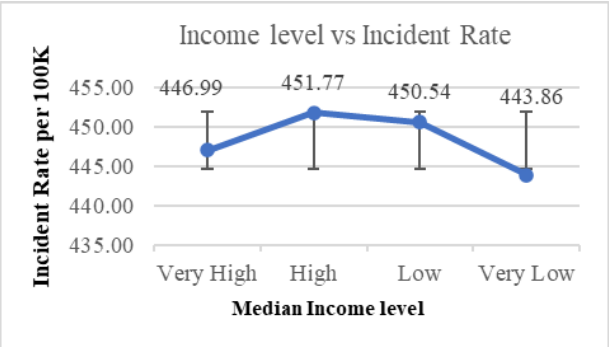

Figure 2-a: Income level vs Incident rate

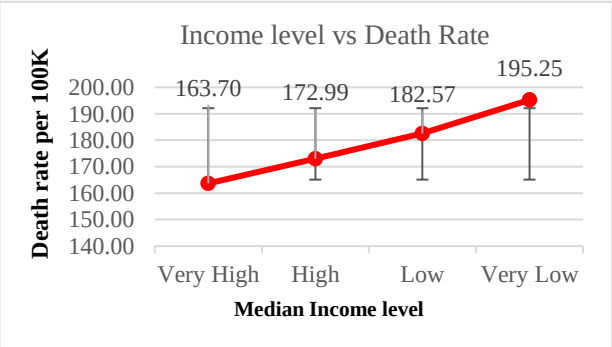

Figure 2-b: Income level vs Death rate

Similarly, figure 2-b clearly shows that there is a high death rate when the income level is very low, while there is a low death rate when the income level is very high. This finding suggests that individuals with lower incomes may not be able to afford adequate treatment for cancer, leading to higher death rates. This trend is because people with lower income levels often have limited access to healthcare services and may not be able to afford the cost of cancer treatment. Therefore, they may be more likely to be diagnosed with cancer at a later stage when the disease is more advanced and difficult to treat, leading to higher mortality rates. This analysis is extensively presented in the graphs above.

---

3. Relationship between various factors vs cancer incidence and death rates

Poverty versus location analysis:

Poverty vs. location analysis is crucial in cancer analysis because poverty is known to be a significant risk factor for cancer. People living in impoverished areas often have limited access to healthcare services, healthy food options, and lifestyle choices that can reduce the risk of cancer. Therefore, examining the correlation between poverty and cancer incidence rates can help identify areas with high cancer rates that are also economically disadvantaged.

State	Average number of people below poverty line	Average of Poverty Estimate in %	Death rate
DC	114790.00	100.00%	182.3
CA	109705.46	95.57%	158.0964912
AZ	79725.93	69.45%	149.0866667
-	-	-	-

States with higher poverty

Table 5: Income versus location analysis

Table 4 clearly illustrates that Washington DC has the highest number of individuals living below the poverty line with a higher death rate. This indicates a high relation between poverty and death rate. This trend is not only experienced in the US but also in many parts of the world.

Similar trend in India

The trend of significant relationship between poverty and cancer death rates is heavily experienced in highly poverty prone country like India. According to the National Statistical Office (NSO) report, around 271 million people in India live below the poverty line and the age-standardized death rate due to cancer in India in 2015 was 96.1 per 100,000 population. This means that out of every 100,000 people in India, 96.1 died due to cancer in 2015. There is a significant relationship between poverty and cancer death rates, and there are several reasons for this. One of the main factors is limited access to healthcare, including cancer screening, diagnosis, and treatment. People living in poverty may not have the resources to access healthcare, which can result in delayed diagnosis and treatment, reducing the chances of survival. Additionally, lack of health insurance can make it difficult for those living in poverty to afford cancer treatment.

Another factor is exposure to environmental toxins and carcinogens, which people living in poverty are more likely to experience. This can increase the risk of developing cancer. Poor diet is also a significant issue for those living in poverty, as they may not have access to healthy and nutritious food, which can increase the risk of certain types of cancer.

Income versus location analysis:

State	Average of Medium Income in terms of %	Average of Medium income ($)	Standard deviation of median Income ($)
NJ	100.00%	72135.28	16879.39
CT	97.62%	70415.37	9587.42
DC	97.03%	69992	NA
MD	93.26%	67269.91	20369.98
MA	90.56%	65328.64	13166.9
AK	88.01%	63487.77	11859.40

Table 5: Income versus location analysis

Table 5 shows that New Jersey has the highest median household income among all states in the country. Using New Jersey as a reference point with 100%, it can be observed that Connecticut, Washington DC, Maryland, and Massachusetts follow closely in terms of income levels. Unlike the poverty percentage, the income percentage does not drop significantly, indicating that the income levels among the states are relatively close to the national average income of $50,528, with a lower standard deviation. From the region-wise analysis we could recall, even though NJ has the highest average income it was one among the states in high severity regions which shows that income does not affect cancer prevalence. This finding highlights the complex relationship between income and cancer outcomes and underscores the importance of examining other factors that may contribute to cancer risk, such as lifestyle behaviours, environmental exposures, and access to healthcare.

4. Correlation analysis to see which factors are most highly correlated with cancer incidence/death rate

	Poverty Estimate	2015 Population Estimate	Average Annual Count	Incident Rate	Five Year Trend	Death Rate	Average Deaths Per Year
Poverty Estimate	1						
2015 Population Estimate	0.968736422	1					
Average Annual Count	0.888021164	0.926632598	1				
Incident Rate	0.013413481	0.023980934	0.071422337	1			
Five Year Trend	-0.032924816	-0.038489501	-0.00726907	0.16875537	1		
Death Rate	-0.084419833	-0.123186203	-0.144071898	0.447938904	0.05564483	1	
Average Deaths Per Year	0.943291214	0.976883759	0.939701957	0.060646862	-0.039382624	-0.092432865	1
Medium Income	0.116401162	0.238723356	0.269429877	-0.002948919	-0.052230137	-0.430815269	0.224363883

Figure 3: Correlation analysis to see which factors are most highly correlated with cancer incidence/death rate

- Positive correlation
- Negative correlation

The correlation figure reveals several noteworthy observations. Firstly, there exists a significant positive correlation between cancer incident rate and death rate, suggesting that regions with a higher incidence of cancer also experience higher death rates. Additionally, the data suggests that individuals with lower median incomes are more likely to experience higher death rates when affected by cancer, as indicated by the strong negative correlation between death rate and median income.

Furthermore, the predictor variables of population estimate and poverty estimate are highly correlated, indicating a positive relationship between population and poverty. Additionally, the average annual cancer count demonstrates positive correlations with both poverty estimate and population estimate, suggesting that a higher number of cancer cases tend to occur in areas with higher poverty and higher population.

Key Findings:

- Higher number of cancer cases tend to occur in areas with higher poverty and population.
- The Northeast region of the US have a significantly higher incidence of cancer.
- There is a high death rate when the income level is very low, while there is a low death rate when the income level is very high. This finding suggests that individuals with lower incomes may not be able to afford adequate treatment for cancer, leading to higher death rates.
- The income levels among the states are relatively close to the national average.
- Regions with a higher incidence of cancer also experience higher death rates.

Recommendations:

Based on the key findings, the following recommendations are proposed:

1. **Develop targeted interventions**: Given that higher poverty and population are associated with a higher number of cancer cases, it is important to develop targeted interventions to address these issues. These interventions may include increasing access to healthcare services, implementing cancer prevention and screening programs, and improving health education in underserved communities.
2. **Invest in cancer research**: The Northeast region of the US has a significantly higher incidence of cancer. Therefore, it is recommended that the government and private sector invest more in cancer research in this region to identify the underlying causes and develop effective prevention and treatment strategies.
3. **Increase access to affordable cancer treatment:** The high death rate among individuals with lower incomes suggests that they may not be able to afford adequate treatment for cancer. Therefore, policymakers and healthcare providers should explore ways to increase access to affordable cancer treatment, such as through expanding Medicaid coverage or providing subsidies for cancer medications.
4. **Promote healthy lifestyle behaviours:** While income levels among the states are relatively close to the national average, lifestyle behaviours such as smoking and poor diet can contribute to cancer risk. Therefore, promoting healthy lifestyle behaviours through public health campaigns and education programs can help reduce cancer incidence and mortality rates.
5. **Implement comprehensive cancer control programs**: Given that regions with higher incidence of cancer also experience higher death rates, it is recommended that comprehensive cancer control programs be implemented to address all aspects of cancer prevention, early detection, treatment, and survivorship.

---------- *End of exploratory data analysis section* ----------

5. Regression Analysis

A regression model to predict the incidence rate.

Predicting cancer incidence rates using regression with various factors is prominent because it allows us to understand the relationships between different risk factors and cancer incidence rates. This information can be used to predict the likelihood of cancer occurrence in different populations and to develop prevention and early detection strategies that are tailored to the specific needs of those populations. Four distinct models were created based on different factors that can affect cancer incidence rates. Each model utilizes a unique set of predictor variables and innovative features. To determine the best model, the suitability of the predictor variables as well as the proportion of variability explained by each model will be considered. Below are the descriptions of the four models developed in the project to predict the incidence rate.

Variable	Model 1	Model 2	Model 3	Model 4
Constant	201.227* (~0)	165.1220* (~0)	218.5590* (~0)	213.977* (~0)
Poverty Estimate	-0.00003633* (~0)	0.00045617* (~0)	-0.00003383	-0.000413424* (~0)
Population Estimate	-0.00006937 (0.607)	-0.00023293* (~0)	0.00001098	0.00016475* (~0)
Death Rate	1.10612340* (0.000092)	1.23663275* (0.02)	1.09307898* (~0)	1.100147* (~0)
Average Annual Count	-	0.14740775* (~0)	-	-
Average Deaths Per Year	0.05507653* (~0)	-0.29539087* (~0)	-	-
Medium Income	0.00099402* (~0)	0.00086576* (~0)	-	0.00086101* (~0)
Five Year trend (numerical)	-	-	2.18399064* (~0)	2.1872988* (~0)
Five Year trend (low categorical)	-	-3.41290284 (0.1572)	-	-
Five Year Trend (Stable categorical)	-	5.89774524 (0.0286)	-	-
Recent Trend	-	Rising 32.435 stable 6.6271 falling 21.906 rising 3.98897 stable 7.0735 *None significant*	-	-
Region North East	-	-	35.18261456* (~0)	34.8060* (~0)
Region South	-	-	-5.89282972* (~0)	-5.495474 (0.007)
Region West	-	-	-10.28535226* (~0)	-9.337701* (0.0012)
Medium Income * Population Estimate	-	-		0.0000164* (~0)
R^2	0.248	0.2825	0.3222	0.3267
Adjusted R^2	0.2467	0.2789	0.3203	0.3245

Note: The asterisk notifies the significance of the particular variable towards the model output.

Model Selection: Based on careful examination, **Model-4 has been selected as the most appropriate for predicting the incidence rate** not only due to its higher R^2 value compared to the other models but also due to the appropriateness of the set of predictor variables. Due to higher R^2, it can account for a greater proportion of variability in the incidence rate. Model-4 has novel features like feature generation, interaction effect which contributes heavily in the performance. Additionally, the difference between R^2 and adjusted R^2 values is minimal in this model, indicating that the predictor variables are well-suited for the task at hand. Perhaps most importantly, all of the predictor variables in Model-4 are statistically significant, meaning that they make a meaningful contribution to the prediction of the incidence rate. *(Note: All assumptions were checked for model-4 – please check R script.)*

A regression model to predict the death rate.

Regression analysis can also provide a better understanding of the relationship between cancer affecting factors and mortality rates, allowing for the development of more effective treatments and interventions for individuals at higher risk of cancer mortality. Ultimately, predicting cancer death rates can help healthcare providers and policymakers identify areas that may require additional resources or interventions to improve cancer outcomes. Three unique models were developed and compared. Certain cancer affecting factors were found to be unsuitable for predicting the death rate and thus were excluded. Each model was then developed using a distinct set of predictor variables and incorporating important features. The following are descriptions of the three models that were developed in predicting the death rate.

Variable	Model 1	Model 2	Model 3
Constant	128.411	122.272044694* (~0)	122.85500* (~0)
Poverty Estimate	-0.00009 (~0)	-0.000071629*	0.00008956520 (0.06)
Population Estimate	-0.00006937* (0.607)	0.000009657* (0.0613)	-0.0000555161728* (~0)
Incidence Rate	0.21938986342* (0.000092)	0.214339976* (~0)	0.2158515343803* (~0)
Average Deaths Per Year	0.00847604407* (~0)	-	-
Medium Income	-0.00102288015* (~0)	-0.000852164 (0.0175)	-0.00086847* (~0)
Five Year trend (numerical)	-	-0.208496045* (0.03)	-0.2147719446026* (~0)
Region North East	-	-6.954555779* (~0)	-6.8518996917422* (~0)
Region South	-	6.634572285* (~0)	6.4427183440371* (0.007)
Region West	-	-11.802234270* (~0)	-12.102829281664* (0.0012)
Medium Income * Population Estimate	-	-	0.0000000006659* (0.000002)
R^2	0.389	0.4586	0.4619
Adjusted R^2	0.388	0.4571	0.4602

Note: The asterisk notifies the significance of the particular variable towards the model output and the value inside the parenthesis corresponds to the p-value.

Model Selection: Based on a solid examination, Model-3 has been determined as the most suitable option for predicting the death rate, as it demonstrates a higher R^2 value compared to the other models. This signifies that Model-3 is able to explain a larger portion of the variability in the death rate, thereby making it a more reliable predictor. Additionally, the trend of increasing R^2 values from Model-1 to Model-3 suggests a progressive improvement in the models' ability to explain the variability of the death rate. This model can be used to evaluate the effectiveness of healthcare programs and policies by comparing predicted and actual death rates. Also, it can help healthcare professionals and policymakers to identify the factors that contribute to high death rates and develop appropriate interventions to reduce the death rate. (*Note: All assumptions were checked for model-3 – please check R script.)*

---------- *End of regression analysis section* ----------