NAME: Devika Pillai

MGIS PROJECT 1

Cancer is a complex disease and is the second leading cause of death in the country, with approximately 1.8 million new cases and 606,520 deaths projected to occur in 2022. The American Cancer Society (ACS) is committed to reducing the burden of cancer through research, education, advocacy, and service. In order to achieve this goal, it is essential to understand the factors that contribute to cancer incidence and death rates in the US. This report aims to provide an analysis of the factors related to cancer incidence and death in the US using county-level data. The dataset used, includes information on demographic, socioeconomic, behavioral, and environmental factors that may influence cancer outcomes. By exploring these factors, I hope to identify regions and partners for targeted cancer interventions across the US. The data has been collected for 47 different States and for more clarity I added an additional feature of region in the dataset which classified the state into northern, southern, northwestern etc. The main features of the dataset to predict the onset of cancer or the death due to cancer is the number of people below poverty line, median household income, and the estimated population in the country in the year of 2015.

State	County Name	Region
Florida	Union County	Southeastern
Virginia	Williamsburg city	Southeastern
Kentucky	Bracken County	Southeastern
West Virginia	Clay County	Southeastern
Oregon	Sherman County	Western

Table 1: Sates and counties most prone to cancer

Table 1 lists the names of the top 5 states, their counties, and the region they lie in, which are the most prone to cancer in the descending order. As it is shown, the top 5 states prone to cancer are Florida, Virginia, Kentucky, West Virginia, and Oregon and the top 5 counties are Union County, Williamsburg city, Bracken County, Clay County and Sherman County. The top 4 states lie in the southeastern region of the United States and the 5th most prone state, Oregon, lies in the western region. This shows that the southeastern region is the most prone to cancer. The state of Florida is the most vulnerable to cancer and the Oregon state is the least. The Union County in Florida is specifically the most susceptible to cancer and the Sherman County is the least susceptible to cancer.

State	PovertyEst	medIncome	Name	popEst2015	incidenceRate	deathRate	Region	income_level
<chr>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<fct>
AL	7204	54366	Autauga County	55347	475	178.	Southeastern	Very High
AL	25696	49626	Baldwin County	203709	455.	174.	Southeastern	High
AL	5943	34971	Barbour County	26489	478.	193.	Southeastern	Very Low
AL	3666	39546	Bibb County	22583	495.	212.	Southeastern	Low
AL	10000	45567	Blount County	57673	430.	175.	Southeastern	High
AL	3179	26580	Bullock County	10696	489.	176.	Southeastern	Very Low

Fig.1: New dataset with 4 – level indicator variable

Now, I create a 4-level indicator variable for Median Income so it can help in identifying income-based disparities that may exist in the population. The figure 1 shows the dataset with the new variable. This can be particularly relevant when analyzing data on cancer incidence and death rates and we can understand how different income groups are distributed across the population and assess whether there are any disparities in health outcomes based on these categories. Additionally,

it can simplify the data analysis process by making it easier to identify patterns and trends in the data. The quartiles divide the median variable into 4 parts, and I used this to then create the 4-level indicator variable.

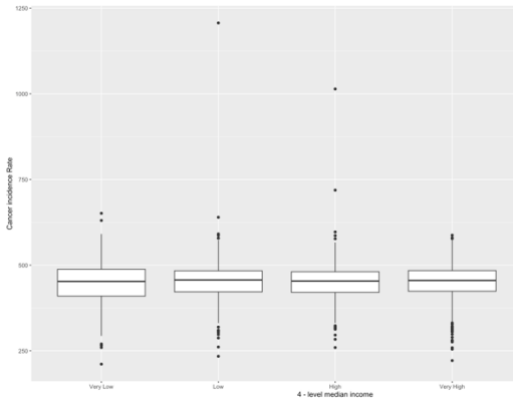

Fig. 2(a)Boxplot showing median income vs cancer incidence rate

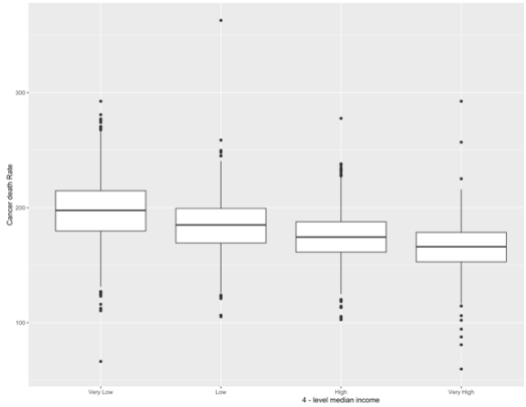

Fig. 2(b) Boxplot showing median income vs cancer death rate

Figure 2 (a) and (b) contain boxplots which are used to show the distribution of the incidence rates and the death rates across the income levels respectively. In figure 2(a) the median of all 4 income levels is approximately the same. There may be a correlation between income level and certain lifestyle factors that increase the risk of cancer, such as smoking or alcohol consumption. Individuals with very low median income may have a higher prevalence of these risk factors, which could contribute to the higher incidence rates observed in this group. While it is true that anyone can get cancer regardless of their income level, individuals with higher incomes may have access to better healthcare and other resources that could help with cancer prevention, early detection, and treatment. This could contribute to the lower incidence rates observed in the low and high income groups. The presence of outliers in the very high income group suggests that even among this relatively affluent group, there may be individuals who are at a higher risk for cancer due to factors such as genetics or environmental exposures.

Figure 3(b) shows the medians of all 4 income levels are different. The median is the highest for very low income and keeps decreasing till very high income which shows there is a negative correlation between income and cancer incidence rate. This means that as income increases, cancer incidence rate decreases. Looking at the ranges of the boxplots, cancer death rates are the highest for individuals with very low median income and lowest for individuals with very high income. The data is skewed towards the lower income groups, as the boxes are shorter for the lower income groups and there are more outliers in these groups. The variation in cancer death rate decreases as income increases. This is indicated by the decreasing size of the boxes as income increases. There may be socioeconomic factors that contribute to differences in cancer death rates between income groups. For example, individuals with higher incomes may have access to better healthcare and healthier lifestyles, which may lower their risk of death by cancer.

To look at how the factors affecting cancer vary by location, I have displayed them on the map of the United States. Figure 3 shows the average median household income in each state on the map of United States. It shows that on average the highest median income is in the state of New Jersey, in the Northeastern region and the lowest is in the state of Mississippi, in the Southeastern region. The highest average population is in the state of California, in the Western region and the lowest

is in the state of South Dakota, in the Midwestern region. On average the highest number of individuals below poverty line is in the District of Columbia, in the Mid-Atlantic region and the lowest is in the state of North Dakota, in the Midwestern region. While this information can provide insights into these factors, it does not directly provide information on cancer incidence or mortality rates. Therefore, we cannot make any specific inferences regarding cancer incidence and mortality rates based solely on the information provided. However, it is possible to make some general hypotheses, for example, regions with higher median household incomes may have better access to healthcare and may be able to afford healthier lifestyles, which could lead to lower cancer incidence and mortality rates. On the other hand, regions with higher rates of poverty and lower median household incomes may have less access to healthcare and may be more likely to engage in unhealthy behaviors, which could potentially lead to higher cancer incidence and mortality rates.

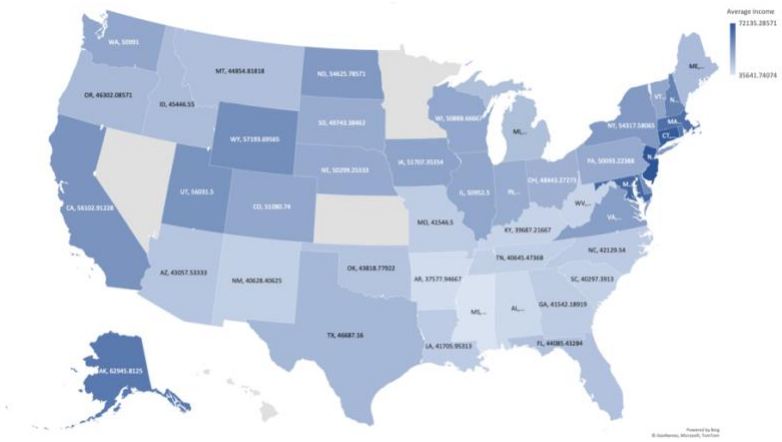

Fig. 3 Average median household income in each state

	PovertyEst	medIncome	popEst2015	incidenceRate	avgAnnCount	fiveYearTrend	deathRate	avgDeathsPerYear
PovertyEst	1.000000000	0.12287327	0.96868580	0.007146165	0.93782969	-0.02690429	-0.10144592	0.94264500
medIncome	0.122873273	1.00000000	0.24736641	-0.010514987	0.25936846	-0.06319818	-0.44263500	0.23429207
popEst2015	0.968685796	0.24736641	1.00000000	0.017248618	0.98073666	-0.03252009	-0.14304070	0.97645183
incidenceRate	0.007146165	-0.01051499	0.01724862	1.000000000	0.05929436	0.18463262	0.45688947	0.05447092
avgAnnCount	0.937829688	0.25936846	0.98073666	0.059294358	1.00000000	-0.03002280	-0.13159395	0.99699028
fiveYearTrend	-0.026904294	-0.06319818	-0.03252009	0.184632615	-0.03002280	1.00000000	0.07822178	-0.03208696
deathRate	-0.101445915	-0.44263500	-0.14304070	0.456889465	-0.13159395	0.07822178	1.00000000	-0.11290888
avgDeathsPerYear	0.942644996	0.23429207	0.97645183	0.054470922	0.99699028	-0.03208696	-0.11290888	1.00000000

Fig. 4 Correlation matrix for all numerical variables

The correlation matrix shown in figure 4 provides important information about the relationships between variables in the dataset and the insights can help make decisions related to resource allocation and interventions to reduce cancer incidence and mortality rates in specific areas. The variables representing cancer incidence rate and death rate of cancer are moderately correlated with each other, suggesting that there may be a relationship between the two variables, but it may not be strong enough to make a significant impact. The variable representing poverty is highly correlated with population estimates, and the average number of cancer incidences and deaths per year. This suggests that higher levels of poverty may be associated with higher rates of cancer in a given county. The population estimates are also highly correlated with the average number of cancer incidences and deaths per year, indicating that larger populations may be associated with higher rates of cancer. Finally, the average number of cancer incidences is highly correlated with the average number of cancer deaths per year, suggesting that counties with higher rates of cancer incidences may also have higher rates of cancer-related deaths.

	Model 1	Model 2	Model 3	Model 4
Intercept	462.95398*	462.1048*	-136.9482	-110.73300
PovertyEst	-0.00011	0.000036	-0.00002	-0.00015
medIncome	-0.00009	-0.00010	-0.0012*	-0.0011*
popEst2015	0.000024	0.00006*	0.000064*	0.000010
RegionMid-Atlantic	24.636590	37.8633	35.4939	14.6627
RegionMidwestern	-8.6998614	-8.08263	-11.57909	-7.6547
RegionNortheastern	27.325670*	27.2395*	24.1202*	23.79939*
RegionNorthwestern	-16.65846	-15.6935	-18.9419	-13.71572
RegionSoutheastern	3.209247	3.63281	2.18675	2.84644
RegionSouthern	-45.128876*	-44.4959*	-47.32649*	-44.37879*
RegionSouthwestern	-86.136660*	-86.05220*	-86.9281*	-86.5216*
RegionWestern	-37.892426*	-36.6264*	-39.68342*	-38.24066*
log(medIncome)	NA	NA	61.2999*	54.30132*
log(PovertyEst)	NA	NA	NA	4.9484101*
PovertyEst:medIncome	NA	-0.00000*	-0.00000	0.000000
se	50.5	50.47	50.41	50.19
Multiple R^2	0.159	0.1603	0.1626	0.1699
Adjusted R^2	0.1556	0.1567	0.1586	0.1657
F statistic	47.62(0.00)	44.07(0.00)	41.35(0.00)	40.47(0.00)

Table 2: Regression analysis for the incidence rate

Table 2 shows the results of four different regression models for the incidence rate of cancer. A regression model is a type of statistical analysis that helps us understand how different variables are related to each other. The table has four different models, each with a different set of independent variables, which are used to predict the dependent variable, cancer incidence and death rates. Since the p – value of all models is less than α, the models are significant. The values of the Intercept in the first two models are positive, but in the third and fourth models it is negative. Model 1 shows that as the number of people below the poverty line in the county increases, the cancer incidence rates decrease. In model 2, as the number of people below the poverty line in the county increases, the cancer incidence rates increase. MedIncome has a negative coefficient in all models, which means that as median household income in the county increases, the cancer incidence rates decrease. The magnitude of the coefficient is larger in models 3 and 4, indicating that income has a stronger effect on cancer rates in those models. In all 4 models, the Northeast region have higher cancer incidence rates than the other parts. In conclusion, the results of the regression analysis indicate that income levels have a significant effect on cancer incidence rates in counties across the United States. Additionally, the region in which the county is located also affects cancer rates. Model 4 has the highest Multiple R^2 value of 0.1699, which indicates that 16.99% of the variation in the response variable (cancer incidence rate) can be explained by the independent variables included in the model. Model 4 also has the lowest F statistic value and a low p-value, indicating that the model is statistically significant and that the independent variables included in the model are jointly significant in explaining the variation in the response variable and it includes all the independent variables that were found to be significant in the other models. This suggests that Model 4 is a more comprehensive and accurate representation of the relationship between the independent variables and the response variable. Therefore, we can consider Model 4 to be the best model for this regression analysis.

	Model 1	Model 2	Model 3	Model 4
Intercept	137.5145*	137.59477*	892.27469*	897.65742*
PovertyEst	-0.00007*	-0.000176*	-0.00010*	-0.000131*
medIncome	-0.0008*	-0.00084*	0.00066*	0.000687*
popEst2015	0.000009	-0.000014	-0.00002*	-0.00003*
RegionMid-Atlantic	5.43309	-3.49291	-0.63139	-5.04690
RegionMidwestern	-12.1990*	-12.60514*	-8.16342	-7.3408
RegionNortheastern	-18.80358*	-18.77463*	-14.92993*	-14.9534*
RegionNorthwestern	-23.1992*	-23.83077*	-19.67616*	-18.58730*
RegionSoutheastern	-5.18412	-5.47253	-3.65942	-3.5133
RegionSouthern	-6.95803	-7.33601	-3.61031	-3.065272
RegionSouthwestern	-25.40803*	-25.37348*	-23.97296*	-24.04941*
RegionWestern	-24.01398*	-24.8256*	-20.84097*	-20.60535*
incidenceRate	0.20811*	0.20917*	0.21259*	0.21071*
log(medIncome)	NA	NA	-77.3869*	-78.7772*
log(PovertyEst)	NA	NA	NA	1.06479*
PovertyEst:medIncome	NA	0.000000*	0.00000*	0.000000*
se	20.08	20.04	19.77	19.75
Multiple R^2	0.4542	0.4566	0.4715	0.4729
Adjusted R^2	0.4518	0.4541	0.4688	0.47
F statistic	192.1	179	176.4	165.5

Table 3: Regression analysis for the death rate

Table 3 contains the results of a statistical analysis that examines the relationship between various factors and the death rate due to cancer in different counties. The four different models (Model 1 to Model 4) indicate how different factors may be related to the death rate. From Model 1 to Model 4, additional factors are introduced into the analysis. In Model 3, the logarithm of median income is added as a predictor variable, and in Model 4, the logarithm of poverty estimate is added. The p-value for the Intercept in all models is less than 0.05, meaning that the intercept term is significant, indicating the baseline death rate in the counties. In all 4 models, the Northeast and Northwestern regions have lower cancer death rates than the other parts. The counties with higher poverty, and higher incidence rates tend to have higher death rates. There is no evidence of a relationship between population size and death rate in model 1 and 2 but there is in model 3 and 4. After adding more variables, the Multiple R2 and Adjusted R2 increase in Model 3 and Model 4. This means that these models better explain the variability in the data than Model 1 and Model 2. However, Model 4 is the best model because it has the highest Adjusted R2 value, a lower F statistic and a low p-value, indicating that it is more parsimonious and has fewer predictors that explain the same amount of variation. In addition, the logarithm of poverty estimate and the interaction term between poverty estimate and median income are significant, which suggests that the relationship between poverty estimate and death rate depends on the median income.