CANCER INCIDENCE ANALYSIS REPORT

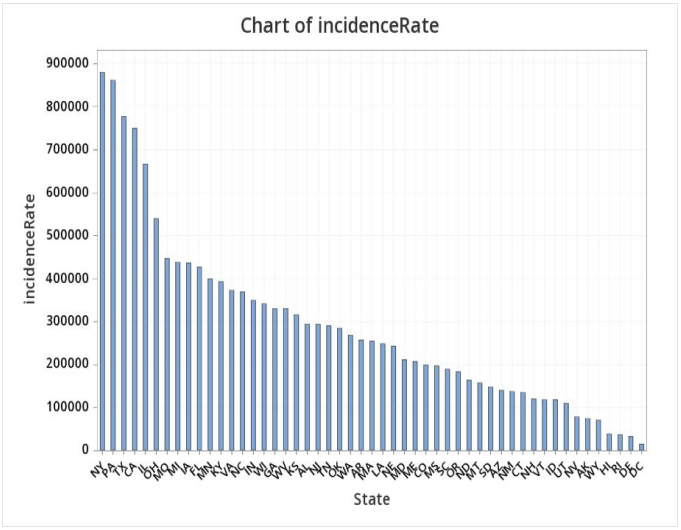
OVERVIEW

This report gives a summary of the analysis done by the American Cancer Society exploring the factors related to cancer diagnosis in the United States. Region, poverty level, income level, population size, and death rates were all statistically compared with cancer incidence level to determine which factors are most significant in determining the likelihood of being diagnosed with cancer. This analysis will be used to identify regions and other associated factors that have the greatest need for cancer interventions. Most importantly, we discuss the importance of using combinations of these factors when interpreting the overall impact of cancer in the United States. Considering that cancer is one of the leading causes of deaths, knowing these factors will assist in improving resources related to public health and intervention efforts for States with high incidence rates. In particular, this information may be used by health officials and policy makers to inform policies and practices that influence cancer outcomes.

STATE VS. CANCER INCIDENCE

Cancer incidence rate is defined as the number of new cancers of a specific type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk, and age-adjusted to a standard population to allow comparisons over time. This can be summarized as: $Incidence\ Rate = (New\ cancers/Population)\ x\ 100,000$, where the numerator of the incidence rate is the number of new cancers, and the denominator is the size of the population. The number of new cancers may include multiple primary cancers occurring in one patient.

In order to have a clear visualization about regions of the country which are most prone to cancer, we organized states from highest rate to lowest rate and created a bar graph. States with incidence rates above 400 are shown in the Chart of IncidenceRate below. Cancer incidence rate is one of the most critical factors to investigate as it identifies the population(s) which are at highest risk of cancer. As can be observed from the graph, the States NY, PA, TX, CA, IL, OH, MO MI, IA, and FL are most prone to cancer. Notably, NY and PA recorded the highest incidence rates,

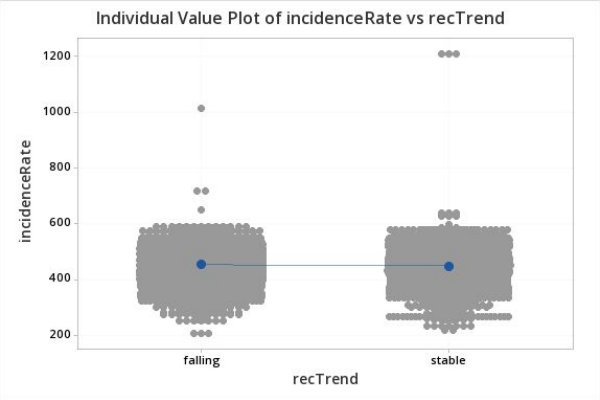


followed by TX, CA, IL and OH. The rest of the States demonstrate a gradual decrease in the cancer incidence rates. However, low incidence rates may not completely imply that the number of cases is small, because not everyone may be able to get tested for cancer. Therefore, based on this graph an urgent intervention in the control of cancer is very necessary in NY, PA, TX, CA, IL and OH.

RECENT TRENDS VS. CANCER INCIDENCE

The data provided illustrate the recent trends in the cancer data. To further analyze the data, we examined the recent trends where numbers of cancer cases are either stable or falling. We observed that there is a statistically significant difference between the falling and stable groups in how they affect the incidence rates. In the falling group, there was evidence of a higher incidence rate than in the stable group.

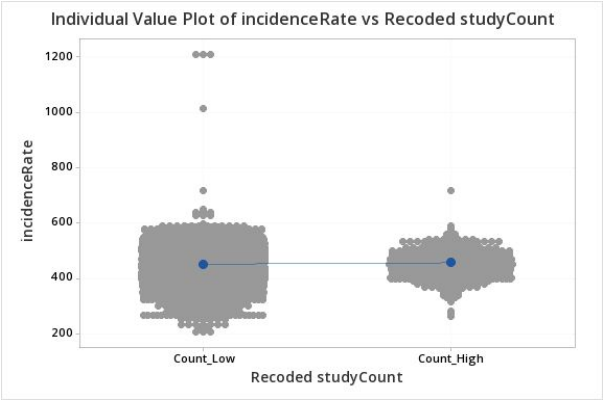
In the figure of the Individual Value Plot of incidenceRate vs recTrend shown to the right, we can see the recent trend groups and their respective incidence rates. The means of the two groups are shown with the blue dots and though they seem to have very similar average incidence rates, our results show that this is enough to be statistically significant and further indicate that the falling group has a higher effect on the average incidence rate than the stable group. *



STUDY COUNT VS. CANCER INCIDENCE

We also investigated the connection between study count, which is the total population being investigated, and cancer incidence rate. We broke our study count data into two groups, the values that fell below the mean were put into the **low** group, and the values that fell above the mean were put into the **high** group. We then hypothesized that the cancer incidence rates would **not** be significantly different between the low and high groups. We tested this assumption with a statistical hypothesis test and found that we had enough evidence to reject the null hypothesis and accept the alternative hypothesis. Therefore, we can conclude that the low study count group **does** in fact report statistically significantly lower cancer incidence rates than the high study count group.

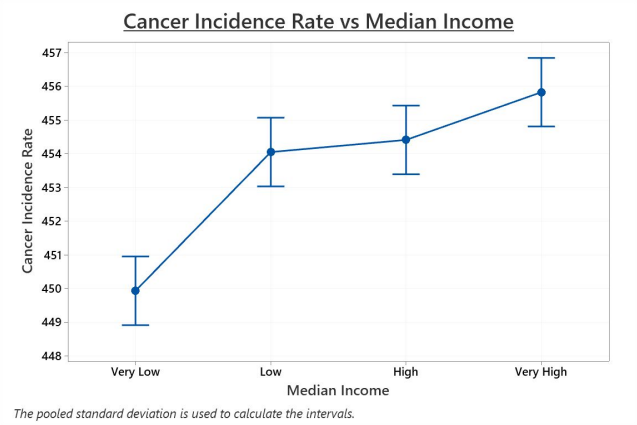
The figure to the right shows the low and high count groups and their respective mean incidence rate data points. We see that although the group means are similar in value, they still are considered significantly different. Therefore, we can observe visually that the low count group has a lower mean cancer incidence, which further confirms our conclusions in the Recent Trend vs Cancer Incidence section. **



MEDIAN INCOME VS. CANCER INCIDENCE

To compare cancer incidence rate to our median income data, we first broke up the median income data into four categorical groups. We named the lowest 25% **very low**, the next 25% **low**, the next 25% **high**, and the top 25% **very high**. Next, we hypothesized that the average cancer rate would be the same for each income group, which allowed us to run an analysis of variance test on the data. This test informs us whether there are any statistically significant differences between the means of any of our groups. The results of the test are displayed below.

The graph displayed here shows the cancer incidence rate determined by median income. Based on this graph, and the numerical results of our analysis, we can conclude that our hypothesis is incorrect and should be rejected. The very low group **is** significantly lower than the other 3 groups, which otherwise do not show any significant statistical difference. Interestingly, the poorest group has the lowest cancer incidence rate, suggesting that those with lower incomes are less likely to get cancer. This conclusion may be false, however, as we are only able to compare income rate with *diagnosed* cancer. Those in the lower income groups may have less access to medical specialists, medical supplies and medications, thus leading to less chances of getting diagnosed even if they do have cancer. Furthermore, in most poor areas there may be lack of proper resources and expertise for instance cancer registry to record, accurate population counting and active follow-up measures. The lack of such resources may limit the efforts of controlling and preventing cancer. Therefore, these conclusions should be tested further with more factors to get more accurate results. ***



CORRELATION ANALYSIS

Correlations

	countyCode	studyCount	PovertyEst	medIncome	popEst2015	incidenceRate
studyCount	-0.013					
PovertyEst	-0.188	0.087				
medIncome	-0.019	0.050	0.111			
popEst2015	-0.189	0.091	0.989	0.204		
incidenceRate	0.019	0.020	-0.097	0.051	-0.083	
avgAnnCount	-0.200	0.094	0.966	0.232	0.983	-0.036
fiveYearTrend	-0.019	-0.007	-0.058	-0.064	-0.062	0.176
deathRate	0.044	-0.027	-0.178	-0.483	-0.213	0.436
avgDeathsPerYear	-0.195	0.096	0.973	0.208	0.988	-0.040

	avgAnnCount	fiveYearTrend	deathRate
studyCount			
PovertyEst			
medIncome			
popEst2015			
incidenceRate			
avgAnnCount			
fiveYearTrend	-0.058		
deathRate	-0.213	0.094	
avgDeathsPerYear	0.993	-0.058	-0.193

The figure to the right is the correlation analysis table which allows us to see the correlation between each numeric variable included in the data set. Correlation values can fall within a range of -1 to 1. Values closer to 0 have a weaker correlation, values closer to -1 have a stronger negative correlation, and values closer to 1 have a stronger positive correlation. We performed the test to establish which variables are most related to each other and which variables are most related to cancer incidence rate. Among the variables that we tested, five had the strongest correlations and these are (1) Population and Poverty, (2) Average Death per Year and Population Estimate, (3) Average Annual Count and Population, (4) Average Annual Count and Poverty, and (5) Average Death per Year and Poverty. The strongest correlation is between population and poverty which showed a correlation value of 0.989. Because this value is so close to 1, it is nearly a perfect positive correlation. Therefore, a positive change in the population variable will also lead to a nearly equal positive change in the poverty variable. This analytical approach can be applied to the other variables in the chart as well to understand how they related to each other. To assess incidenceRate correlation, we looked at the correlation values across the row labeled incidenceRate to assess how each variable correlated with cancer incidence. In this data set, we find that there is relatively low correlation between these variables and the incidenceRate variable.

ACTIONABLE REGRESSION MODEL

Using the data set regarding cancer incidence, we created an actionable regression model through an iterative process using the Minitab software. First, we had to determine which of our factors had the strongest impact on the variability of cancer incidence rate. After some statistical analysis, we found that the most significant factors were death rate, five year trend, poverty percentage, study count, zip code, and average deaths per year. We then created a regression model using these variables as the independent variables, and incidence rate as the dependent variables. The results of this model allowed us to see how each independent variable affects cancer incidence rate, and how strong that effect is.

One challenge for this model is the high potential for multicollinearity, which arises when there are many independent variables in a model. When the independent variables are highly correlated, it is not immediately possible to determine the separate effect of any particular independent variable on the dependent variable. Because many of the variables in our model, such as location data, are highly correlated with each other, we needed to ensure that we minimized the impact that correlation had on the regression model. This is a delicate balancing act between reducing multicollinearity while still having a model that can explain the differences in cancer incidence rate due to our independent variables.

In our final actionable regression model, our variance inflation factor values range from 1.01 to 1.39 showing very little relationship among the variables and representing the lack of multicollinearity in the analysis. Due to these low VIF values, we are able to immediately determine the significance of each independent variable on the response variable. Therefore, this means the death rate, five year trend, poverty percentage, study count, zip code, and average deaths per year have a significant impact on the cancer incidence rate while not having strong effects on the other variables. In fact, this confirms some of the conclusions that we have already made. For instance, we concluded that the level of poverty correlates to the incidence rates. The death rate and average deaths per year, are logically correlated to the response variable. Finally, the study count may have an impact on the incidence rate because as the study count increases, there is a greater chance that the population tested will have cancer thus leading to a high incidence rate.****

OVERALL CONCLUSIONS:

In this project, we were provided with a data set which contained information about different regions in the United States. Based on data manipulation techniques by using Minitab software, we investigated the impact of these factors and created a model to demonstrate how each respective factor related to cancer deaths in the US. In particular, we identified NY, PA, TX, CA, IL, OH, MO MI, IA, and FL as States which are most prone to cancer. This agrees with information in the literature, due to the fact that these places are developed and there is high chance that they have immediate access to expertise, cancer registry, proper medical cancer detection facilities and have updated population databases which makes it easier to obtain good knowledge about factors that relate to cancer incidence rates and draw appropriate conclusions. On recent trends, we examined by means of a plot of incidenceRate and recentTrend (falling, stable) and determined the impact of stable and falling on the incidence rate. With regards to the falling group, there was significant evidence of higher incidence rate than in the stable group. Furthermore, by performing a hypothesis test we note that a lower study count impacts slightly on the incidence cancer rate than a high study count and the converse is correct as well as logical. On the other hand, based on the data we found that the population in the lower income group had low incidence rates. While this may be unexpected, however; it may be correct in the

sense that low income populations usually lack access to proper medical facilities for detection and treatment of cancer. Thus, cancer cases may be present but they are not being properly counted. In order to corroborate our findings, we performed a correlation analysis to find out which factors are correlated with each other and notably those that are most highly correlated with cancer incidence rate. Based on the tested variables, population and poverty were highly correlated with each other while their impact on the cancer incidence rate was found to be low. Moreover, through an iterative process in Minitab, we created an actionable regression model for incidence rate to explain the variability and to identify the most important factors linked to high incidence rates. By using appropriate approaches to identify and address challenges associated with multicollinearity, in order to ascertain that we study only factors which had an impact on the response variable. As a result, we found that the death rate, five year trend, poverty percentage, study count, zip code, and average deaths per year had a significant impact on the incidence rate.

APPENDIX: MINITAB FUNCTIONS PERFORMED

* Manipulated data to a new sheet on Minitab to show only falling and stable values and performed the hypothesis based on those data and the incidence rate. Results from Minitab are shown below.

Method

μ_1: population mean of incidenceRate when recTrend = falling
μ_2: population mean of incidenceRate when recTrend = stable
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 > 0$

T-Value	DF	P-Value
7.00	14541	0.000

** Found mean study count and divide into high and low study count groups using recode function then perform hypothesis test on the new column and incidence rate

Method

μ_1: population mean of incidenceRate when Recoded studyCount = Count_Low
μ_2: population mean of incidenceRate when Recoded studyCount = Count_High
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 < 0$

T-Value	DF	P-Value
-5.77	2844	0.000

*** Created a 4 level indicator variable for medIncome and conducted an ANOVA comparing the four groups. Results from Minitab are displayed below:

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
46.8381	0.22%	0.21%	0.19%

Summary

Lower End	Upper End	Recoded Value	Number of Rows
-1	41800	very low	8134
41800	48544	low	8131
48544	55832	high	8134
55832	1.00000E+11	very high	8152

Source data column medIncome
Recoded data column Recoded medIncome

Each interval includes its lower end.

Means

Recoded medIncome	N	Mean	StDev	95% CI
very low	8134	449.930	57.372	(448.912, 450.948)
low	8131	454.043	43.558	(453.024, 455.061)
high	8134	454.403	40.414	(453.385, 455.421)
very high	8152	455.817	44.198	(454.800, 456.834)

Pooled StDev = 46.8381

**** Attempted fit regression model using all numerical data, identified high VIF values and used an iterative process to remove multicollinearity problems.

Coefficients

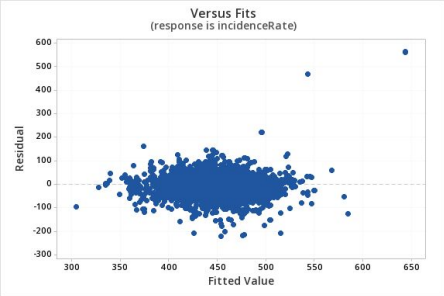
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	340.79	1.72	198.21	0.000	
deathRate	0.9921	0.0101	98.36	0.000	1.39
fiveYearTrend	2.1732	0.0665	32.69	0.000	1.01
povertyPercent	-2.1010	0.0417	-50.44	0.000	1.32
studyCount	0.0674	0.0107	6.32	0.000	1.01
zipCode	-0.000517	0.000008	-64.94	0.000	1.10
avgDeathsPerYear	0.002841	0.000128	22.20	0.000	1.07

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	27177386	4529564	3304.93	0.000
deathRate	1	13259680	13259680	9674.72	0.000
fiveYearTrend	1	1465056	1465056	1068.96	0.000
povertyPercent	1	3486673	3486673	2544.00	0.000
studyCount	1	54789	54789	39.98	0.000
zipCode	1	5780709	5780709	4217.81	0.000
avgDeathsPerYear	1	675697	675697	493.01	0.000
Error	30573	41901790	1371		
Total	30579	69079175			

Attempted to add categorical variables to see if R^2 can be further improved. Recent Trend doesn't worsen VIF scores but only improves R^2 by 1% so it is not included in the actionable model

We also looked at the residuals versus fits result to assess the regression model assumptions and it appears the model assumptions have been met.



See final regression below:

Regression Equation

incidenceRate = 340.79 + 0.9921 deathRate + 2.1732 fiveYearTrend - 2.1010 povertyPercent + 0.0674 studyCount - 0.000517 zipCode + 0.002841 avgDeathsPerYear