Socioeconomics and Cancer Mortality Rates in US Counties

Under the supervision of Dr. Ibrahim Mohamed Youssef

Ahmed Kamal, Zeyad Hossam, Abdulrahman Shawqy,
Kareem Salah, Mohamed Ibrahim

Abstract:

The main objective of this study is to analyze and look for the relationship between socioeconomic status and cancer mortality rates in the US from 2010 to 2016. This was done using a comprehensive dataset that was aggregated from several sources, including the American Community Survey [1]. Using this data, we find the correlation coefficient of different parameters and apply a multivariate linear regression model so that this can be used to estimate the number of deaths associated with cancer among 100,000 people each year and assess the effect of different attributes on this number to help reduce it.

Keywords:

Cancer - Mortality - United States - Multivariate Regression

Introduction:

The term "cancer" refers to a set of diseases that are related to the uncontrolled growth and widespread of abnormal cells in a particular bodily region. Cancer is considered a fatal and life-threatening disease considering the significant mortality rates that cancer can produce. It has an impact on how the body functions as a whole and, as it develops, has the potential to damage or invade other body organs. In 2015, a total of 1,633,390 new cancer cases were reported in the United States: 816,937 in men and 816,453 in women. The overall incidence rate was 437.7 per 100,000 persons [2].

Cancer can have several causes, such as consuming tobacco or alcohol, genetic factors, or maybe lifestyle choices such as diet. Our study aims to examine the effect of socioeconomic factors on the incidence and mortality rates of cancer. This encompasses demographic characteristics (e.g., age, race), economic indicators (e.g., income, poverty), educational factors (e.g., education levels), and healthcare-related variables (e.g., health coverage). This was chosen as a factor for the study due to its high effect on cancer incidence and mortality rates. For example, it is believed that people from lower socioeconomic backgrounds often have higher rates of certain cancers; they are more likely to die from cancer compared to those from higher socioeconomic backgrounds due to limited access to medical services such as cancer treatment facilities or the general economic level of each individual. Also, individuals with higher educational levels have better chances to survive the disease since they have the advantage of early detection or association healthcare insurance.

So we are using a dataset that was retrieved from the "data.world" website [1] and mainly collected from many resources, such as the American Community Survey. This data contains several socioeconomic attributes that are hypothesized to affect the incidence and mortality rates of cancer in each county. Examples of these attributes are the different education levels in the age range of 18 to 24, social status, race, age,

income, poverty, health coverage, birth rate, death rate, and finally the city and district as categorical variables.

Methods:

Our journey started with exploring the dataset and understanding the nature of each attribute with the help of several Python packages and libraries:

Package	Usage
Pandas	Easy data manipulation and analysis
numpy	Efficient numerical operations and array manipulation
matplotlib	Visualization and plotting
Seaborn	Statistical graphics
scikit learn	Performing linear regression on the data
scipy	Help in linear regression computation
statsmodels	Statistical modeling and analysis

- Exploring the dataset:

This exploration helped us identify the nature of each attribute and determine what quantitative and categorical variables are. We found out that there are only two categorical variables, which are the city and district from which the data was collected. Through visualization, we also noticed the discrepancy between each county's real and expected death rates.

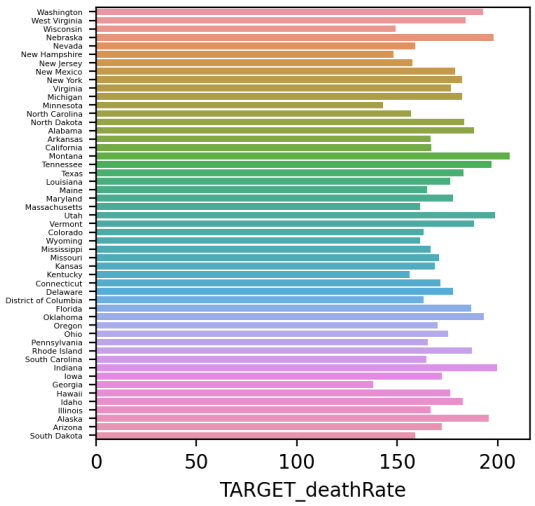


Figure 1, the predicted death rate for each county

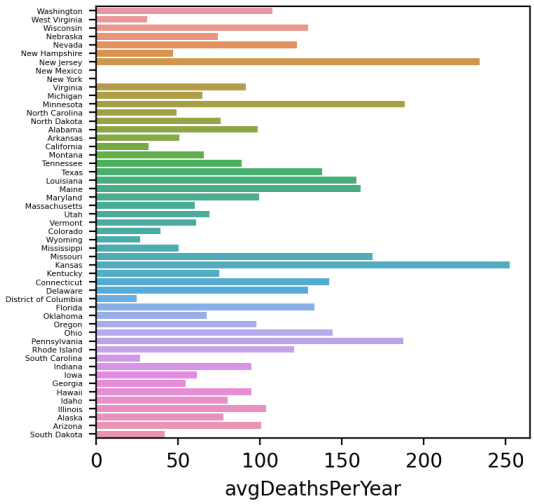


Figure 2, Actual Death Rate for each county

We notice from the two figures that in most counties, the actual death rate is way lower than the predicted one, except for a few counties such as Kansas and New Jersey.

Also, this exploration resulted in the discovery of three attributes that have missing values that need to be handled.

- Data Cleaning:

The attributes that have null values can be listed as follows:

Attribute	No. Null values
The percentage of people between 18 and 24 who got a college degree	2285
The Percentage of county residents with private health coverage alone (no public assistance).	609
The percentage of employees who are 16 years old or more	152

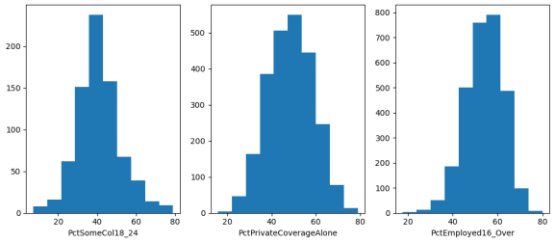


Figure 3, the three attributes with null values

We needed to start handling these null values so they wouldn't affect our study. So the following are the four methods we tried before choosing the best procedure:

- Filling all null values with zero

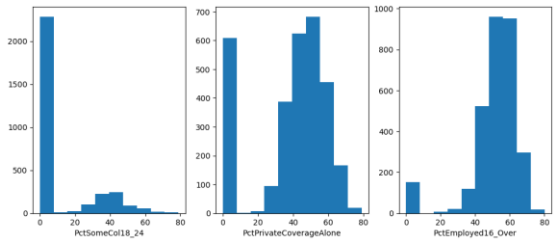


Figure 4, filling null values with zeroes

As shown in Figure 2, we notice that this procedure results in a significant concentration of data points at zero, which distorts the data distribution and skews the statistical properties of each attribute. All this shows a bias in the statistical analysis, so this is an inappropriate way to handle the null values.

- Removing the rows that have null values

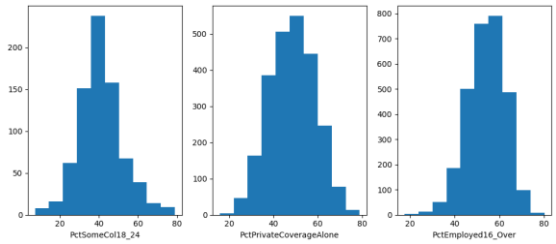


Figure 5, after dropping null values

Figure 3 demonstrates that the graph essentially changed nothing. However, this approach is similarly ineffective for handling missing values. This is because a significant portion of the data points, some of which may contain crucial information, are discarded.

- Filling null values with the mean of the values that are present in their column

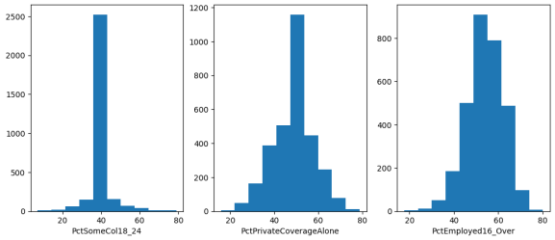


Figure 6, after filling values with each column mean

As shown in Figure 4, we can see that the plot has lost variability, which can lead to a loss of important information in the data and incorrect interpretations of results as this model underestimates the standard error.

- Applying the forward fill method

This method copies the last known non-missing value forward to fill in the missing values. It propagates the value to the next missing value until it reaches the next non-missing value. And this is the most efficient way to deal with the null values, as we don't lose the randomness of the data points and don't need to drop any data or lose information.

We used the function offered by pandas.

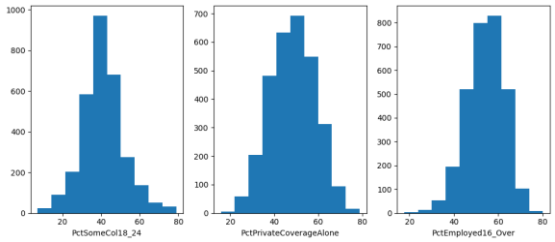
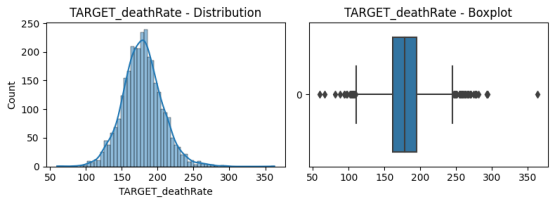
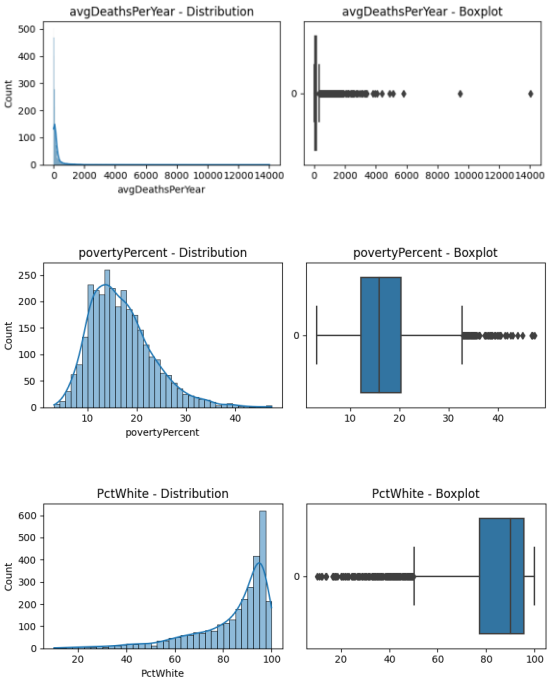


Figure 7, after using the forward fill method

- Handling the outliers:

Visualizing the data helped us see that the data had a lot of outliers that need to be handled. Examples are shown in the following images:



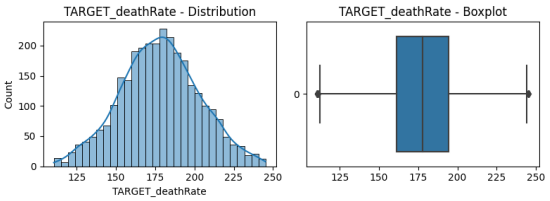


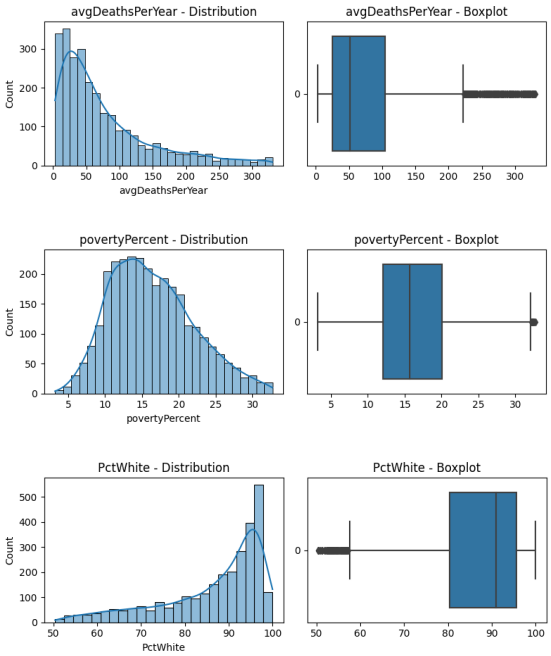
There are two common ways to remove these outliers: the interquartile range and the Z-score. But because most of our data and attributes aren't normally distributed, we followed the procedure of the **interquartile range**. This procedure involves dividing the data into four equal parts, finding the third and first quartiles, and subtracting the first quartile from the third. That will result in a value representing the interquartile range. We applied this operation using the function "**np.percentile()**" from numpy and the "**midpoint**" method.

Finally, we removed all the values in the following range:

$$Q3 + (1.5 * iqr) : Q1 - (1.5 * iqr)$$

What follows are examples of the same attributes shown before after modification:





- Selecting Study Data upon applying the Pearson Correlation Coefficient:

After we had successfully removed the outliers, we went on to choose the attributes that were believed to strongly affect the data.

At first, we calculated the ***Pearson* correlation coefficient** between each attribute of the numeric attributes and the other attributes using *Pandas.* Then, we made a heatmap for these calculations using the *seaborn* python package.

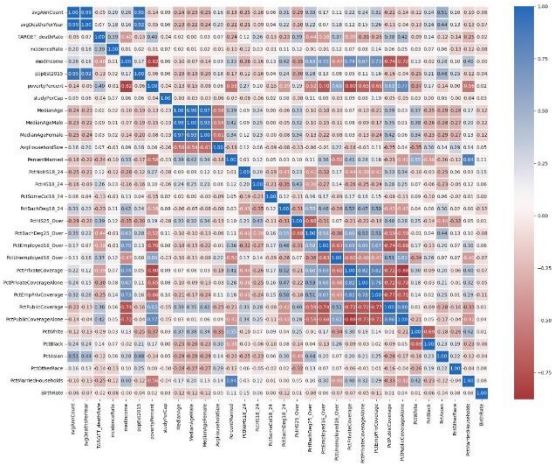


Figure 8, Heat map for the correlation between each attribute and other attributes.

Based on this heatmap, we selected the attributes that have a high correlation. So we chose a range of

$$[-1: -0.3] \cup [0.3: 1]$$

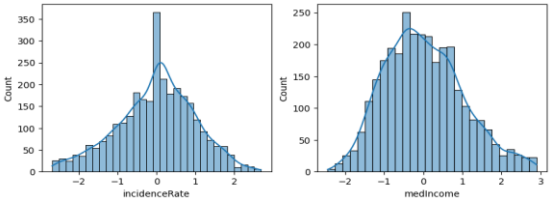
Finally, we started to make all the needed statistics for the selected data. That is like measuring the mean, median, mode, standard deviation, and range of each attribute. And we did that also by using *Pandas*.

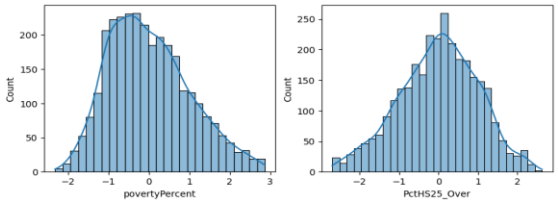
- Standardization:

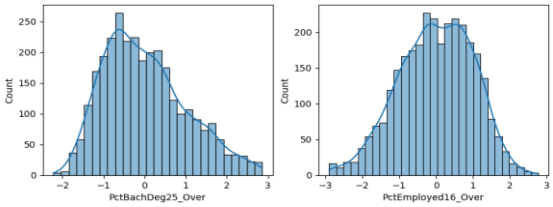
Before we proceed with our study and perform the linear regression model, we must ensure that all the newly selected attributes are on a similar scale to prevent attributes with large scales from dominating the analysis. And that is called data standardization or Z-score normalization, where we find the "Z-score" by subtracting the mean from each point and dividing the result by the standard deviation. The procedure transforms variables so that they have a mean of zero and a standard deviation of one.

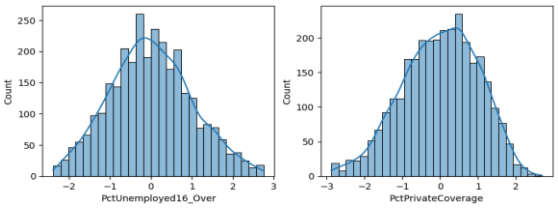
$$\boldsymbol{Z = \frac{(x - mean)}{standard\ deviation}}$$

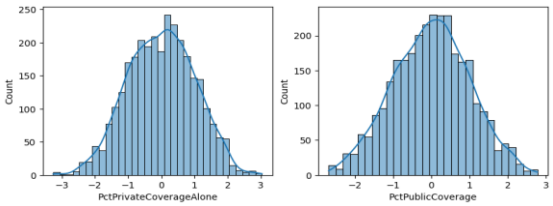
Then, we re-perform the forward fill method on the newly standardized data to make sure there aren't any null values formed during the standardization process. As the forward fill can miss the first value of the attribute if it is null, we made a statement to fill them with 1, if they exist.

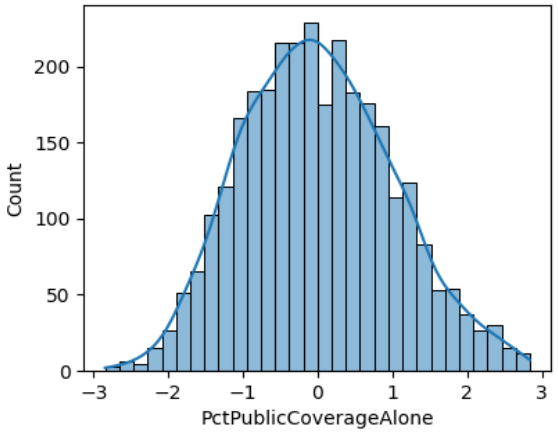












- Linear Regression coefficient for each feature:

Starting by calculating it from scratch. To do that, calculate the mean of the feature and the target provided in the dataset. After that, calculate the standardized covariance and the variance of the feature. That is done to find the regression coefficients and be able to calculate and predict the slope and intercept of the best-fit line.

$$Sxy = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The slope (b_1) can be calculated by dividing the covariance of the target feature and the current feature by the variance of the current feature:

$$b_1 = \frac{S_{xy}}{S_x^2}$$

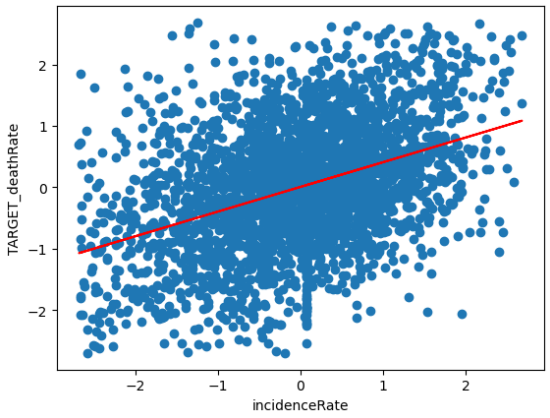
To calculate the intercept, multiply b_1 by the feature's mean, then subtract it from the target's mean:

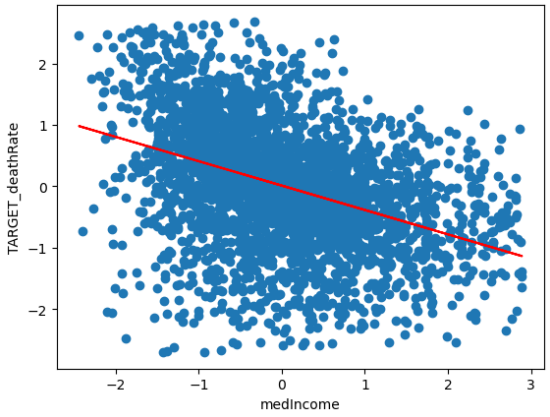
$$b_0 = \bar{y} - b_1\bar{x}$$

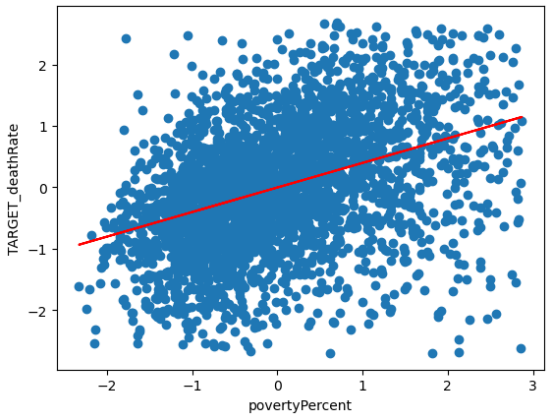
The best-fit line is found by the following formula:

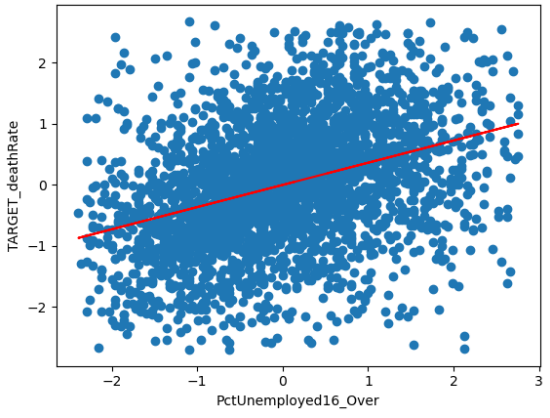
$$\hat{y} = b_0 + b_1 x$$

After we calculated the best-fit line, we can benefit from visualizing it. What follows are some examples of the linear regression model between some features and the target death rate.









After we calculated the model from scratch, we can start using the available functions in the *Scikit-learn* python package and compare the two results to determine the accuracy of our model.

```
RC for incidenceRate is : bo = 0.0070103701509411 66, b1 = 0.4020450378823925
using standard python libraries: bo = 0.007010370150941163, b1 = 0.40204503788239226
----------------------------------------------------------------------------------------
RC for medIncome is : bo = 0.0115308581831625 42, b1 = -0.3970150205858868
using standard python libraries: bo = 0.011530858183162539, b1 = -0.3970150205858866
----------------------------------------------------------------------------------------
```

- Multivariable Regression:

In this step, the *Scikit-learn* package is used to make a multivariable regression model between the target death rate and all other features.

- Hypothesis Testing:

Assuming the null hypothesis $H_\circ$ is the distribution of each feature is normal. Consequently, the alternative hypothesis H_a is the distribution is not normal. And to check this, we followed the ***Shapiro-Wilk*** test:

Using Shapiro-Wilk test

We used the available functions in the *Scipy* Python package to perform this test. In this test, we measure the discrepancy between the observed data and the expected data under the assumption of normality. In other words.

It examines how close the sample data fit a normal distribution [3]

Using *Scipy's* available function, we calculate the p-value, which represents the strength of the evidence against the null hypothesis. So if the p-value is less than or equal to the significance level "Alpha", this would mean that the alternative hypothesis is true and the feature's distribution is not normal. And if the p-value is greater than the significance level, it would mean that the null hypothesis is true and the feature's distribution is normal.

The previous test results stated that the distribution of all features is not normally distributed. So we tried to perform statistical techniques to transform each feature distribution into a normal distribution and re-check the normality using the *Shapiro-Wilk* test.

For this, we used the ***Yeo-Johnson transformation*** [4] test and the **modified *Box-Cox*** test [5][6]. After applying both techniques and re-checking for normality, the features' distribution didn't change and remained non-normally distributed.

- Assessment the multivariable regression quality:

The goal of this process is to evaluate the performance of the model and determine how well it fits the data. It is also carried out to determine the strength and significance of the relationships between the predictors and the dependent variable. The assessment will be done in terms of:

Individual Regression Coefficients:

The regression coefficients represent the strength and direction of the relationship between the predictor and the dependent variable. There are multiple ways to assess this:

- **Using a t-test or z-test:**
  - By setting a null hypothesis $H_\circ$ that the coefficient is zero or no relation between the predictor and the dependent variable.
  - Setting the alternative hypothesis H_a that the coefficient is not zero or that there is a significant relationship.
  - Select the significance level (Alpha α), such as 5%
  - Calculate the T-statistic or Z-statistic based on the available information. If the standard deviation is known and the dataset is large, we can use the Z-statistic. And if the standard deviation is

unknown and the dataset is small, we can use the T-statistic.

$$T - statistic = \frac{coefficient - 0}{standard\ error\ of\ coefficient}$$

$$Z - statistic = \frac{coefficient - 0}{standard\ deviation\ of\ coefficient}$$

- Find the critical value and the p-value
- Compare the p-value to the significance level (α)
- Interpret the results.
  If H_0 is rejected, it indicates that the coefficient is significantly different from zero. If not, it means that there is no strong evidence of a relationship between the predictor and the dependent variable.

- **Using the confidence level:**
- To measure the precision of the coefficient estimate, calculate the standard error of the coefficient.
- Determine the desired level of confidence, such as 95% or 99% confidence levels.
- Calculate the margin of error by multiplying the standard error by the critical value.
- Calculate the confidence interval:

Coefficient estimate - Margin of error: Coefficient estimate + Margin of error

The resulting range represents the likely range within which the true population value of the coefficient lies.

- Interpret the results:
  A confidence interval can be used to describe how reliable survey results are.

Regression Model as a whole:

This process is carried out to determine how well the model fits the data and makes accurate predictions. There are multiple ways to assess this:

- **R-squared:**
- R-squared tells us how much of the variation in the outcome we can explain using the predictors we have in our model. If the R-squared value is higher (close to 1), it means that our model fits the data better. But R-squared alone doesn't tell us how good our model is at predicting new data or if it will work well with different information.

- **Adjusted R-squared:**
- This variant of R-squared adjusts for the number of predictors in the model, penalizing the inclusion of unnecessary variables. It is a more conservative measure that accounts for model complexity.

Results:

Our study has shown that the eleven following features have the highest impact on the target death rate:

Attribute	Correlation Coefficient
Incidence Rate (x_1)	0.393332
Median income per county (x_2)	-0.400594
Percent of populace in poverty (x_3)	0.400301
The percentage of county residents aged 25 and above with a high school diploma (x_4)	0.389953
The percentage of county residents aged 25 and above with a bachelor's degree (x_5)	-0.435021
Percent of county residents ages 16 and over employed (x_6)	-0.381089
Percent of county residents ages 16 and over unemployed (x_7)	0.367553
Percent of county residents with private health coverage (x_8)	-0.386774
Percent of county residents with private health coverage alone (no public assistance) (x_9)	-0.302710
Percent of county residents with government-provided health coverage(x_{10})	0.376922
Percent of county residents with government-provided health coverage alone (x_{11})	0.424369

Note that negative correlation coefficient means negative impact and positive coefficient means positive impact.

The following are the equations of the impact of each attribute on the target death rate ($\hat{y}$):

$$\hat{y} = 0.007 + (0.402) * (incidence\ rate)$$

$$\hat{y} = 0.016 + (-0.397) * (medIncome)$$

$$\hat{y} = -0.003 + (0.4) * (povertyPct)$$

$$\hat{y} = 0.004 + (0.391) * (PctHS24)$$

$$\hat{y} = 0.007 + (0.433) * (PctBachDeg25)$$

$$\hat{y} = 0.001 + (-0.393) * (PctEmployed16\_over)$$

$$\hat{y} = -0.002 + (0.364) * (PctUnemployed16\_over)$$

$$\hat{y} = 0.001 + (-0.380) * (PctPrivateCoverage)$$

$$\hat{y} = 0.002 + (-0.306) * (PctPrivateCoverageAlone)$$

$$\hat{y} = 0.002 + (0.387) * (PctPublicCoverage)$$

$$\hat{y} = -0.0001 + (0.422) * (PctPublicCoverageAlone)$$

The multivariable regression model formula can be expressed as follows:

$$model = 0.3717 * x1 - 0.0789 * x2 + 0.093 * x3 + 0.2291 * x4 - 0.0744 * x5 - 0.0443 * x6 + 0.1036 * x7 - 0.1312 * x8 + 0.0219 * x9 - 0.0991 * x10 + 0.065 * x11 + 0.0074$$

To assess the accuracy of the models, the root mean squared error is used.

$$error = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n}}$$

Where x is the actual value and $\bar{x}$ is the calculated value.

The error in the individual regression coefficients model:

Error in Attribute	Error in %
incidenceRate	1.67
medIncome	1.67
povertyPercent	1.67
PctHS25_Over	1.68
PctBachDeg25_Over	1.65
PctEmployed16_Over	1.68
PctUnemployed16_Over	1.70
PctPrivateCoverage	1.69
PctPrivateCoverageAlone	1.74
PctPublicCoverage	1.68
PctPublicCoverageAlone	1.66

Error in the multivariable regression model 1.304%

- Comments on the distribution of each feature

Distribution Type	Attribute
likely Gaussian	TARGET_deathRate IncidenceRate medIncome povertyPercent MedianAge MedianAgeMale MedianAgeFemale PctNoHS18_24 PctHS18_24, PctHS25_over PctUnemployed16_over PctPrivateCoverage PctPrivateCoverageAlone PctEmpPrivCoverage PctPublicCoverage PctPublicCoverageAlone PctMarriedHouseholds BirthRate
Skewed	avgAnnCount avgDeathPerYear PctWhite PctBlack PctAsian PctOtherRace
Exponential	popEst2015 studyPerCap
Multimodal	avgHouseHoldsize PercentMarried PctSomeCol18_24 PctBachDeg18_24 PctEmployed16_over PctBachDeg25_over

Conclusion:

From the results, we can conclude the following:

- Increasing the median income of a county can reduce the cancer mortalities. Since the resident can access better or more healthcare services.
- The higher the percent of people with higher education in a county, the less the cancer mortalities. As the educated individual can follow a lifestyle that protects him from having cancer such as avoiding tobacco or alcohol. Consequently, this will reduce the incidence rate.
- The type of the health coverage which a resident in a county receives has a significant impact on the cancer mortalities in a county. This health coverage can help the resident access better healthcare services.
- The more unemployed people in a county, the more the probability of having high cancer mortalities. As an unemployed resident won't have sufficient money to access several health services. Also, he could suffer from bad lifestyle such as low-quality nutrition that increases the cancer incidence rate such as colorectal cancer.

Member Contribution:

~ Ahmed Kamal
- Calculating RCs from scratch
- Designing the presentation

~ Zeyad Hossam
- Data Cleaning
- Removing the outliers from the dataset
- Calculating RCs from scratch
- Formatting the notebook

~ Abdulrahman Shawqy
- Removing the outliers from the dataset
- Designing the presentation
- Finding the type of distribution

~ Kareem Salah
- Writing and formatting the report
- Assessment the multivariable regression quality

~ Mohamed Ibrahim
- Cleaning and visualization of the data
- Standardization
- Calculate a set of descriptive statistics
- Calculating RCs from scratch
- Calculating RCs using standard Python libraries
- Hypothesis testing
- Formatting the notebook

References:

[1] N. Rippner, "OLS Regression Challenge," data.world. [Online]. Available: https://data.world/nrippner/ols-regression-challenge.

[2] Centers for Disease Control and Prevention, "U.S. Cancer Statistics Data Brief: No. 3, USCS—Highlights" [Online]. Available: https://www.cdc.gov/cancer/uscs/about/data-briefs/no3-USCS-highlights-2015-incidence.htm.

[3] "Wilk Test," in ScienceDirect, [Online]. Available: https://www.sciencedirect.com/topics/mathematics/wilk-test.

[4] "Box-Cox Transformation," in Statistics How To, [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/normal-distributions/box-cox-transformation/.

[5] "Shapiro-Wilk Test," in StatsKingdom, [Online]. Available: https://www.statskingdom.com/doc_shapiro_wilk.html.

[6] "Power transform," in Wikipedia, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Power_transform.