**A REPORT**

**ON**

**Project Name**

**BY**

**Aman Vohra, Simran Lalwani,Anush Minocha, Jugal Patel**

**Prepared in Fulfillment of**
**Project Course: MGIS**

**AT**

RIT

**Rochester Institute of Technology,**
**Rochester, NY,**
**United States of America.**

# CONTENTS

# ACKNOWLEDGEMENT

We would sincerely like to thank Prof. Ali Tosyali for assigning us this project to apply our knowledge and also for his constant supervision and mentorship. This allowed us to gain a significant amount of knowledge in the topic. We would also like to thank him for teaching the course with such clarity as we were able to apply the concepts with much better understanding.

# I. Introduction

As one of the leading causes of death in the United States, trailing heart disease, cancer is considered a serious issue in the population. According to surveys, one out of four deaths in the United States is due to cancer. In recent years, the improvement of big data and technology has seen a significant amount of data collected on the incidence of cancer and its relation to other variables. It is important to view trends in this data to gain a comprehensive understanding of the Cancer situation in the U.S. Looking at the 2009-2014 survey data on cancer incidence by state, it is clear that some states reported significantly higher cancer rates than others.

## Executive Summary

The research goal is to obtain an exhaustive understanding of the cancer condition in the United States. The research is motivated by cancer and trailing heart diseases that have become a leading cause of death, recording one in at least four deaths reported in the United States. The research aims to understand states with highest and lowest incidence rates and the average death rates in every state, which would help comprehend the states that are more prone to cancer and attributes linked to the incident rates. The research integrates data with categorical and numeric variables and consists of 3072 observations and 16 attributes. The first step was conducting an elementary analysis, and it was achieved using a map that plots the average death rates in every state. When evaluated per 100 000 people in the USA, Kentucky records the highest incident rate of 512.74 cases, while Arizona has the lowest average incident rate at 354.00 cases. When the analysis is further narrowed to focus on specific countries, the Union County in Florida records the highest cancer rates at 1206.9 cases while the Aleutians West Census Area in Alaska records the lowest incidence rate at 201.3 cases. The target variable for the report was incident rates, and when checking for normality using a histogram, the distribution of data skewed to the left, which was an indication that incident rates have a less arithmetic average (448.28742) than the median (453.55). As the distribution of data is skewed to the left, it indicates that the highest number of data points used in the analysis are relatively dissimilar, showing outliers in the data. Correlation analysis is conducted using a heatmap, which indicates the target variable 'incidence rate', highly correlated to variables such as 'deathRate' and 'countyCode.' Based on the results obtained from correlation analysis, the research uses the inference to create a multilinear regression model which takes the variables correlated to the target variable (Incident rate) as inputs. The study also assesses multicollinearity, and the results indicate that it arises in regression when independent variable predictors correlate with the target variable and amongst themselves.

## II. Objectives

1) To determine which states of the country are most prone to cancer.
2) Also to create a four level indicator variable for the attribute 'medIncome'. We divide this into four levels that are 'VeryLow', 'Low', 'High', 'VeryHigh'.
3) To understand what attributes are highly correlated with 'incidenceRate' and also determine which attributes are correlated amongst each other.
4) Creating a regression model with the target variable being 'incidenceRate'. The model should have a high $R^2$ value as the higher the value, the greater variability of the model will be determined.
5) Finally, to write an executive summary of our results that could be understood by a non-technical person.

## III. Data Set Description

Our data stems from a cancer related study.It consists of 16 attributes and 3072 observations. There is a mix of categorical and numeric variables. Our target variable is **'incidenceRate'**.

1) **CountyCode** (**Quantitative - Discrete**) :
   2-digit State FIPS + 3-digit county FIPS
2) **State**(**Qualitative - Nominal**) :
   The states of the United States of America
3) **PovertyEst**(**Quantitative - Discrete**) :
   The sum of people below poverty line.
4) **PovertyPercent**(**Quantitative - Continuous**) :
   Percent of population below poverty line.
5) **medIncome**(**Quantitative - Discrete**) :
   The median of income by household counties.
6) **Name**(**Qualitative - Nominal**) :
   Names of counties.
7) **popEst2015**(**Quantitative - Discrete**) :
   Estimated population by county in 2015.
8) **County**(**Qualitative - Nominal**) :
   The list of different counties.
9) **incidenceRate**(**Quantitative - Continuous**) : **Target Variable**
   Cancer (all cancers) age-adjusted incidence per 100,000.
10) **avgAnnCount**(**Quantitative - Discrete**) :
    2009-2013 mean incidences per county.
11) **recentTrend**(**Qualitative - Ordinal**) :
    Incidence recent trend.
12) **fiveYearTrend**(**Quantitative - Discrete**) :
    Incidence five year trend.
13) **countyName**( **Qualitative - Nominal**) :
    The list of different counties.
14) **deathRate**(**Quantitative - Continuous**) :
    Deaths from cancer per 100,000 per county
15) **avgDeathsPerYear**(**Quantitative - Discrete**) :
    Average number of deaths per county per year(2009-2013).
16) **recTrend**(**Qualitative - Ordinal**) :
    The recent trend of cancer mortality.

# IV. Analysis

## 4.a)Elementary Analysis

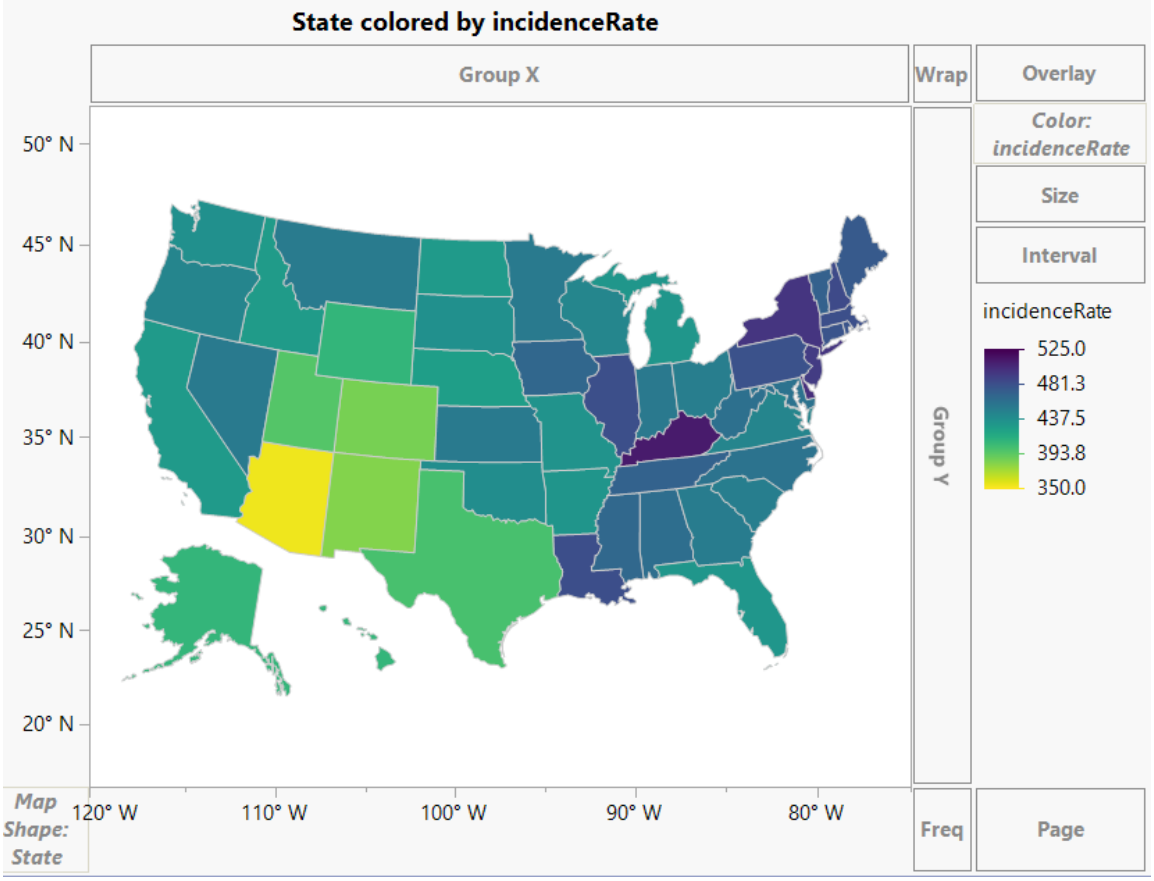


**Figure 4.1**

This Map in **Figure 4.1** is plotted with the average of death rates in each state. We can see what states have the highest and lowest average death rates in the country.

A higher incidence rate implies that a particular region has a high incidence of cancer per 100,000 people.

## State-Wise Analysis

The map tells us that the state of **Kentucky** has the highest average incidence rate at **512.74** cases per 100,000 people, followed by Delaware(502.13) and NewYork(498.28).

| State | Average Incidence Rate |
|---|---|
| KY | 512.74 |
| DE | 502.1333333 |
| NY | 498.2854839 |
| NJ | 494.747619 |
| NH | 486.14 |

The map tells us that the state of **Arizona** has the lowest average incidence rate at **354.00 per 100,000 people**, followed by New Mexico(382.4875) and Colorado(385.91).

| State | Average Incidence Rate |
|---|---|
| AZ | 354 |
| NM | 382.4875 |
| CO | 385.9116667 |
| UT | 396.8111111 |
| TX | 400.9108333 |

## County-Wise Analysis

Further we do analysis of Highest and lowest incidence rates by counties to get the following results.

Based on our analysis, the county **Aleutians West Census Area** in **Alaska** has the lowest incidence rate at **201.3** cases per 100,000 people, followed by Presidio County, Texas(211.1) and Hudspeth County, Texas(214.8).

| County | incidenceRate |
|---|---|
| Aleutians West Census Area, Alaska(6,10) | 201.3 |
| Presidio County, Texas(6,10) | 211.1 |
| Hudspeth County, Texas(6,10) | 214.8 |
| Stanton County, Nebraska(6,10) | 221.5 |
| Dolores County, Colorado(6,10) | 234 |

The **Union County** in **Florida** has the highest cancer incidence rate at **1206.9 cases ,** followed by Williamsburg City, Virginia (1014.2) and Charlottesville City, Virginia(718.9).

| County | incidenceRate |
|---|---|
| Union County, Florida(6,10) | 1206.9 |
| Williamsburg City, Virginia(6,10) | 1014.2 |
| Charlottesville City, Virginia(6,10) | 718.9 |
| Petersburg City, Virginia(6,10) | 651.3 |
| Bracken County, Kentucky(7,9) | 639.7 |

## 4.b)Checking for normality of our target variable

On basis of the histogram in **Figure 4.2** the data seems to be skewed to the left.

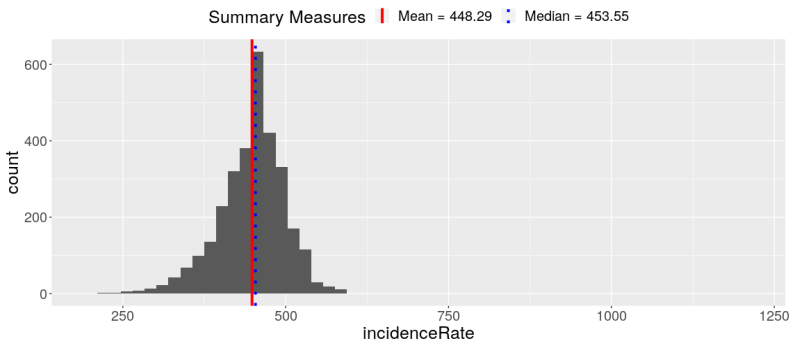


**Figure 4.2**

The boxplot in **Figure 4.3** validates our observation that the data of our target variable is left skewed.

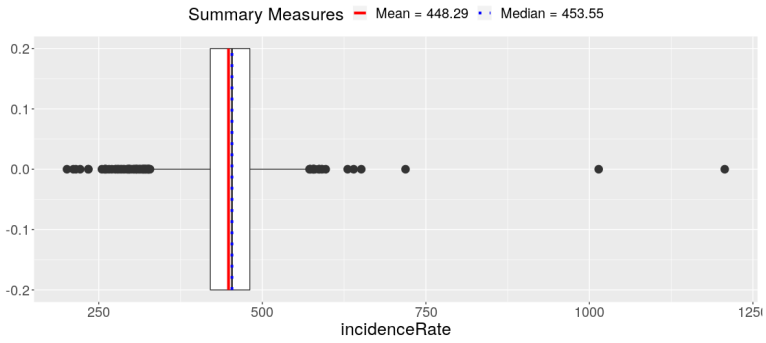


**Figure 4.3**

The scatter plot in **Figure 4.4** confirms that the data is not normally distributed (p-value < 0.05).Therefore, we can conclude that the data of our target variable is left skewed.

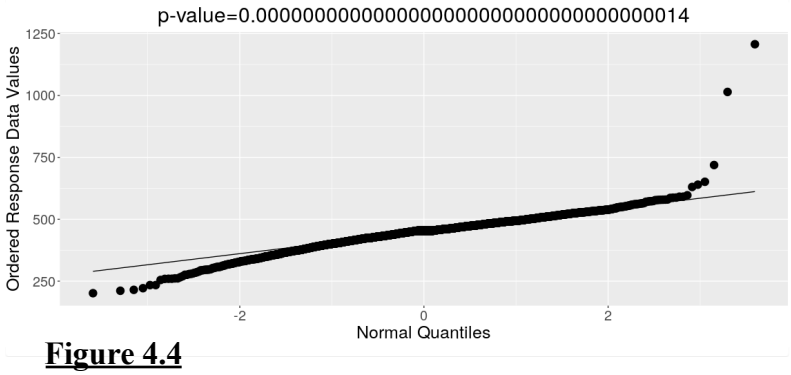


**Figure 4.4**

Also, the **Mean** = 448.28742 is less than the **Median** = 453.55.

Therefore the data is **skewed to the left**.This implies that some data points in our data are more extreme as compared to others. This could indicate outliers. We go ahead and analyze this further.

We investigated the data to have some sort of junk values in order to eliminate the skewness.But since the data is correct and credible , we continue further with the skewness.

## 4.c) Comparison of incidence Rates by means of Median Income.

| Median Income Levels | Average Incidence Rate |
|---|---|
| **Very Low** (22537.004 to 48388.75) | **448.23** |
| **Low** (48388.75 to 74137.5) | **448.52** |
| **High** (74137.5 to 99886.25) | **448.89** |
| **Very High** (99886.25 to 125635.0) | **428.65** |

We notice a slight increase in average incidence rates from **Very Low** to **Low** median income level. It seems that the average incidence rates tend to decrease as the median income increases. We can therefore make the inference that **medIncome** is a significant factor and can impact the cancer incidence per 100,000 people.

Practically thinking, people who have **Very Low** or **Low** income are more likely to be diagnosed with cancer than those who have a **High** or **Very High** income. This could be due to a lot of factors. Some might be ,the poorer quality medical care or older equipment used or the type of facilities available ,could all be potential causes for the disparity between **medIncome** and cancer incidence rate per 100,000 people in each county.

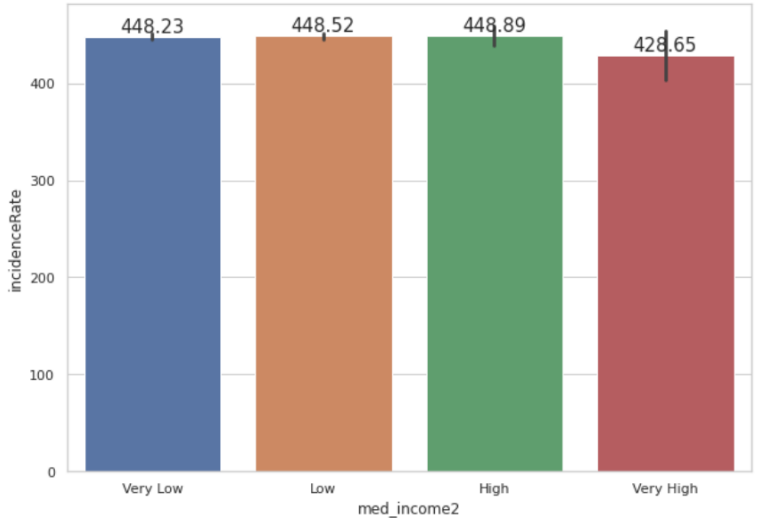


**Figure 4.5**

On the basis of the bar plot in **Figure 4.5,** we can see that the incidenceRate for the Very Low, Low and High median Income levels are somewhat similar. But the incidenceRate for the places with VeryHigh median income levels is lower, by about 5 %. This implies that with a significant increase in the median income level the incidence of Cancer per 100,000 people would come down by a decent amount.

Further, we can see the counts of each class after creating the 4 level indicators.

```
Very Low     1882
Low          1077
High          102
Very High      11
Name: medIncome, dtype: int64
```

## 4.d) Correlation Analysis

### Correlation Heatmap

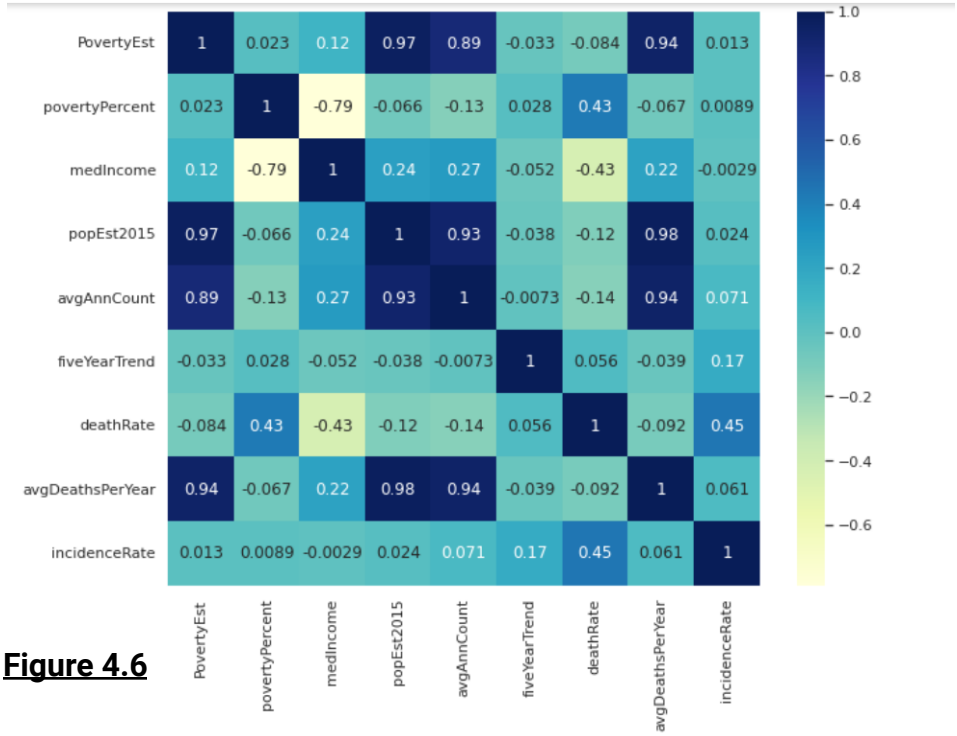


**Figure 4.6**

The correlation heatmap in **Figure 4.6** gives a graphical representation of the values with high and low correlation. A lighter color implying a lesser amount of correlation.

The variable countyCode has decent correlation with our target variable incidenceRate, but we do not take it into account because it just consists of the code of a specific county, a factor that would not have any effect on the incidence Rate of cancer .

a).

On the basis of the correlation matrix in **Figure 4.6**, we can deduce that the following features have decent correlation with the target variable - **incidenceRate.**

PovertyEst, popEst2015, avgAnnCount,deathRate, avgDeathsPerYear,

We will use the features with high correlation to our target variable , provided they do not have correlation between themselves .

b).

On basis of the correlation matrix in **Figure 4.6**, we can deduce that the following features have significant correlation between each other:

| Variable | Correlated Variable | Correlation |
|---|---|---|
| povertyEst | popEst2015 | 0.97 |
| povertyEst | avgDeathsPerYear | 0.94 |
| popEst2015 | avgAnnCount | 0.93 |
| popEst2015 | avgDeathsPerYear | 0.98 |
| avgDeathsPerYear | avgAnnCount | 0.94 |

We can get rid of one of these features as they are so strongly related and a change in one variable would definitely cause a change in another variable which would result in an inaccurate prediction made by our model.

A positive correlation would mean that an increase in the value of one variable would cause an increase in the value of the other. A negative correlation would mean that an increase in the value of one variable would cause a decrease in the value of the other.

## 4.e) Models

### Regression Model -1

We use the inferences drawn by us in our correlation analysis and create a Multilinear Regression Model with all the variables that are correlated with incidence Rate.

```
Call:
lm(formula = incidenceRate ~ povertyPercent + medIncome + popEst2015 +
    avgAnnCount + deathRate, data = Cancer)

Residuals:
    Min      1Q  Median      3Q     Max
-288.90  -24.88    2.06   27.70  589.73

Coefficients:
                  Estimate    Std. Error t value             Pr(>|t|)
(Intercept)    233.533297897  10.826524767  21.570 < 0.0000000000000002 ***
povertyPercent  -0.873462004   0.223695994  -3.905      0.00009638191911 ***
medIncome        0.000578970   0.000120001   4.825      0.00000147039688 ***
popEst2015      -0.000047058   0.000006885  -6.835      0.00000000000984 ***
avgAnnCount      0.014193053   0.001602013   8.860 < 0.0000000000000002 ***
deathRate        1.111050099   0.034076934  32.604 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.54 on 3066 degrees of freedom
Multiple R-squared:  0.2716,    Adjusted R-squared:  0.2704
F-statistic: 228.7 on 5 and 3066 DF,  p-value: < 0.00000000000000022
```

The Regression Equation is as follows -:

incidenceRate = 233.53 - 0.8734 * **povertyPercent** + 0.00057 * **medIncome**

-0.00004 * **popEst2015** + 0.0141 * **avgAnnCount**

+ 1.1110* **deathRate**

We get a **R Squared** value = **0.2716**

The **p- value** of the model is < **0.00000000000000022**.

Standard error is the calculated value of the distance between the actual points verses the regression line.

Precisely, the smaller the standard error value , the closer the observations are to the regression line. This means that the model explains the data accurately.

**Standard Error** = 46.54 on 3066 degrees of freedom.This implies that the model is significant as a whole.

This means that this model explains **27.16 %** of the variability in our target variable.

## Regression Model- 2

We check for **MultiColinearity** before training this model. MultiColinearity occurs in Regression when the independent variable predictor variables have significant correlation amongst themselves in addition to the correlation with the target variable.This causes our model to perform poorly. We aim to identify this MultiColinearity and get rid of the variables that cause this.On basis of our analysis in **Part 4.c,** we get rid of the following variables from our Regression Model:

povertyPercent ,popEst2015

```
Call:
lm(formula = incidenceRate ~ medIncome + avgAnnCount + deathRate,
    data = Cancer)

Residuals:
    Min      1Q  Median      3Q     Max
-291.65  -24.74    2.39   28.21  585.68

Coefficients:
              Estimate   Std. Error t value             Pr(>|t|)
(Intercept) 208.02303930   8.38371132  24.813 < 0.0000000000000002 ***
medIncome     0.00094382   0.00008016  11.774 < 0.0000000000000002 ***
avgAnnCount   0.00361941   0.00062063   5.832        0.00000000605 ***
deathRate     1.08387423   0.03398025  31.897 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.1 on 3068 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2525
F-statistic: 346.9 on 3 and 3068 DF,  p-value: < 0.00000000000000022
```

The Regression Equation is as follows -:

incidenceRate = 208.02 + 0.00094 * **medIncome**+ 0.0036 * **avgAnnCount**

+ 1.0838 * **deathRate**

We get a **R-Squared Value** = **0.2533**

The **p- value** of the model is < 0.00000000000000022.

**Standard Error** = 47.1 on 3068 degrees of freedom.

This implies that the model is significant as a whole.

This means that this model explains **25.33** % of the variability in our target variable.

**Regression Model- 3**

We create another regression model using the variable fiveYearTrend along with the variables that we took in the previous part. This variable has almost all values as numerical. But a few values are marked as '*' which is not categorical but just an anomaly . So,we replace these values with a 0 and convert the data type of this variable into numerical.

We train the model with the following variables -

The regression equation is as follows -:

incidenceRate = 208.02 + 0.0009688 * **medIncome**+ 0.00357 * **avgAnnCount**

+ 1.071* **deathRate** + 2.018 * **fiveYearTrend**

We get a **R-Squared Value** = 0.2757

The **p- value** of the model is < **0.00000000000000022**.

This implies that the model is significant.

This means that this model explains 27.57 % of the variability in our target variable.

**K-NeighborsRegressor**

Neighbors-based regression is used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors.

KNeighborsClassifier implements learning based on the k nearest neighbors of each query point.

The regression equation is not obtained using this method, but we are able to report the R-Squared value.

We get an **R-Squared** value = 0.3345

This means that this model explains 33.45 % of the variability in our target variable.

# V. Key Findings and Recommendations

**Model Comparison**

| Model Used | R-Squared Value |
|---|---|
| Regression Model -1 | 0.2716 |
| Regression Model -2 | 0.2533 |
| Regression Model -3 | 0.2757 |
| KNRegressor Model | 0.3345 |

**Figure 5.1**

On the basis of the analysis done by us in Figure 5.1 , we can conclude that the **KN Regressor Model** is the best model as it explains the most amount of variability(33.45%) in our target variable. The next best model is our Regression Model-3 with an R-Squared Value of 0.2757(27.57% variation explained).

**Recommendations:**

Although, we get a decent value of R square from our various models, but we noticed that the variables in the given data were not very significantly correlated with our target variable; So it would help the model a lot if more categories of data are explored and identified for changes caused to the target variable.

Further, a larger dataset would aid in better identifying the correlation of attributes with the target feature and among the attributes themselves. It would also train the model exponentially better and give us a better understanding of the variance hence increasing the r-squared value. It would also widen the range of models we could train to ultimately get better insights about the target variable ‘incidenceRate’.

## VI. Conclusion

Our team met the basic requirements of the assignment by reporting the solutions/analysis, they are specified below:

- These states have a higher incidence rate as compared to others - Kentucky, Delaware, New York.
- These states have a lower incidence rate as compared to others -
  - Arizona, New Mexico, Colorado.
- These counties have a higher incidence rate -
  - Aleutians West Census Area, Presidio County, Hudspeth County
- These counties have a lower incidence rate -
  - Union County ,Williamsburg City, Charlottesville City
- The median income level Very High has people with the lowest incidence rate of cancer.
- Based on the median Level Income Vs incidenceRate indicator, we know that the cancer incidence Rates generally increase as income decreases.
- There are not many attributes with considerable correlation with the target feature 'incidence rate'. Although, some features have significant correlation with each other.

Our team exceeds the basic requirements by multiple factors listed below:

- We used Python to perform the Exploratory Data Analysis and to implement the regression models on the dataset. We've also reported inferences from these models along with the insights they provide.
- We have used multiple tools like MS Excel , JMP Pro , R and Python to obtain insights that we've reported. This enabled us to report multiple visualizations on the dataset provided.
- We identified the skewness of the data and inferred why that was occurring and concluded that it was valid data points.
- We did data processing by identifying a variable 'fiveYearTrend' that had numerical data.
- Using python, along with implementing linear regression models, we trained a k-neighbors regressor model and compared the R-sqaured value thus obtained. We also experimented with a few other models like the Random Forest Regressor.
- We trained multiple models for Multiple linear regression using multiple combinations of attributes based on our understanding of their correlation and significance on the target feature.

## VII. References


- scikit-learn.org
- towardsdatascience.com
- realpython.com
- pandas.pydata.org


Tools Used : Microsoft Excel, R Studio, Jupyter notebook, JMP Pro