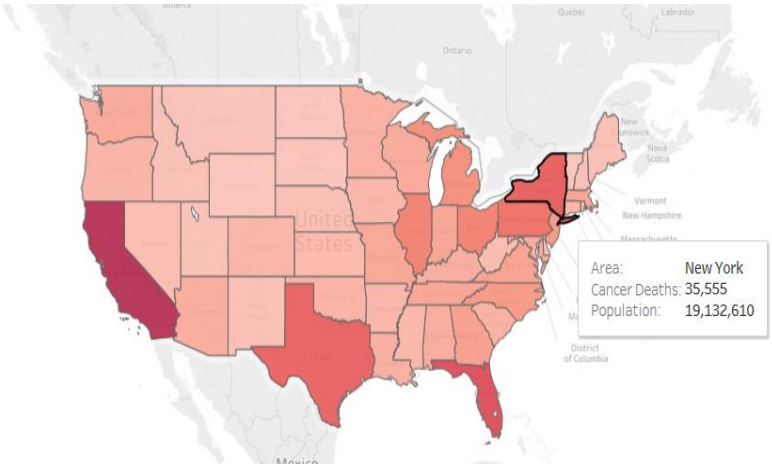


ANALYSIS IN CANCER RESEARCH

ABSTRACT

United States has the 7th highest cancer rate in the world. There are more than 100 types of cancer, including lung cancer, breast cancer, stomach cancer, prostate cancer, Colo rectum cancer, cervix cancer among several others. Generating insights through research in this field is need of the hour. This has been the motivation for us to select this project topic.

Team:
Harsh Bhotika
Neel Kanhere
Shalini Shrivatsa
BIA – 652 MULTIVARIATE ANALYISIS
Guided by:
Dr. David Belanger

TABLE

ANALYSIS IN CANCER RESEARCH

1. ABSTRACT

Cancer is the 2nd leading cause of death in the world. It is a generalized group of diseases which can be regarded as uncontrollable growth and spread of cells which can interfere with the systems of the body to cause life-threatening scenarios. Anybody can be affected by Cancer, but some individuals have different exposure to different types of Cancer based on different factors and situations they are exposed to in the journey of their life. By developing Cancer prediction model, we can provide an assessment as to how to proceed further with the diagnosis and how to be prepared and create awareness in the masses.

United States has the 7th highest cancer rate in the world. There are more than 100 types of cancer, including lung cancer, breast cancer, stomach cancer, prostate cancer, colorectal cancer, cervix cancer among several others. Generating insights through research in this field is need of the hour. This has been the motivation for us to select this project topic. The aim of this project is to predict the probability of cancer taking into consideration multiple factors of being.

2. INTRODUCTION

United States has the 7th highest cancer rate in the world. There are more than 100 types of cancer, including lung cancer, breast cancer, stomach cancer, prostate cancer, colorectal cancer, cervix cancer among several others. Generating insights through research in this field is need of the hour. This has been the motivation for us to select this project topic. The aim of this project is to predict the probability of cancer taking into consideration multiple factors of being.

Our team proposes to work on a database, having close to 1 million values which we obtained from Center for Disease Control and Prevention, United States. We plan on visualizing the cancer statistics in USA for all the 50 states with the leading types of cancer that had affected the people in each state. We would also like to relate states with high cancer rates to any external factors that might be causing high number of cancer cases in those regions. We plan to analyze these statistics using factors like geographic region, age, sex, race, ethnicity, number of people affected by cancer, number of people who lost their lives because of cancer and the total population of states. We will be analyzing cancer statistics based on these factors for the years 1999 to 2013 i.e. for 15 years. The aim of this project is to predict the probability of cancer taking into consideration multiple factors of being. After visualizing the existing scenario, we plan on predicting the cancer statistics in USA in the future. We plan on predicting the leading types of cancer in USA, the regions which will be worst affected by cancer in the future and the cancer statistics in those age groups.

3. EXPLANATION OF THE DATASET

The United States cancer statistics data set has been obtained from the Center for Disease Control & Prevention (CDC) repository. The United States Cancer Statistics data is the official federal statistics on cancer in the United States. The original dataset is multivariate in nature and has close to 1 Million observations including approximately 14,000 cases among children younger than 20 years diagnosed in each of the individual years from 1999 to 2013. The dataset also includes mortality data from malignant cancers as recorded in the National vital statistics system from the 50 states. About the incidence and mortality data, 100% of the U.S. population is covered. This data was collected by the CD from medical facilities that diagnose cancers and report information to central cancer registries in metropolitan areas, regions or states. This data is rich in information and can be used to monitor cancer trends over time, determine cancer patterns in various populations, guide planning and evaluation of cancer control programs help set priorities for allocating health resources, and provide information for a national database of cancer. The data covers of a wide range of different types of cancers classified on the age group, race, gender, state and individual years from 1999 to 2013. All the data attributes of the original dataset are described as follows:

AREA: The names of 50 states in the United States for which the data was collected.

AGE_ADJUSTED_CI_LOWER: Lowest value for the range of population age adjusted rates based on the proportion of the population in 19 specific age groups (younger than 1 year, 1–4 years, 5–9 years, 10–14 years, 15–19 years, ... 85 years and older). Lowest value in 95% confidence interval, in which the intervals reflect range of variation in the estimation of the cancer rates.

AGE_ADJUSTED_CI_UPPER: Lowest value for the range of population age adjusted rates based on the proportion of the population in 19 specific age groups (younger than 1 year, 1–4 years, 5–9 years, 10–14 years, 15–19 years, ... 85 years and older). Highest value in 95% confidence interval, in which the intervals reflect range of variation in the estimation of the cancer rates.

AGE_ADJUSTED_RATE: Average of AGE_ADJUSTED_CI_LOWER and AGE_ADJUSTED_CI_UPPER

EVENT_TYPE: The type of case i.e. Incidence or Mortality.

COUNT: The number of Cancer cases recorded for the given scenario (i.e. age group, sex, cancer type, and race).

POPULATION: The total population of the group for the given scenario (i.e. age group, sex, cancer type, and race)

RACE: 5 ethnicity classifications which are White, Black, Asian/Pacific Islander, Hispanic and American Indian/ Alaska Native.

SEX: Gender of the person i.e. Male or Female

SITE: The location/ type of cancer which are: Female_Breast1, Prostate, Brain_and_Other_NerV_System, Cervix, Colon_and_Rectum, Corpus_and_Uterus_NOS, Esophagus, Female_Breast, Hodgkin Lymphoma, Kidney_and_Renal_Pelvis, Larynx, Leukemias, Liver_Intrahepatic_Bile_Duct, Lung_and_Bronchus, Myeloma, Non_Hodgkin_Lymphoma, Oral_Cavity_and_Pharynx, Ovary, Pancreas, Stomach, Thyroid, Urinary_Bladder, Melanomas_of_the_Skin, Mesothelioma, Testis, Kaposi_Sarcoma

YEAR: The years from 1999 to 2013 for the observations.

CRUDE_RATE: The value for number of cancer cases per 100000 of population for a given set of conditions in an observation.

4. DATA QUALITY AND CHECK

1. **Removed Null and Redundant values in data**

   As a part of the data cleaning process we first removed the missing values and redundancies in data. After performing this we got a data set with 452075 observations. Example of the null values, missing data and redundant data in the dataset which were eliminated are as follows:

Alabama	22.5	26.5	24.4	595	Incidence	2293259	All Races	Female	Female Br	1999	23.9	28.1	25.9
Alabama	22.5	26.4	24.4	604	Incidence	2302835	All Races	Female	Female Br	2000	24.2	28.4	26.2
Alabama	23.7	27.8	25.7	646	Incidence	2309496	All Races	Female	Female Br	2001	25.9	30.2	28
Alabama	25.6	29.7	27.6	706	Incidence	2314370	All Races	Female	Female Br	2002	28.3	32.8	30.5
Alabama	23.2	27.1	25.1	649	Incidence	2324069	All Races	Female	Female Br	2003	25.8	30.2	27.9
Alabama	23.7	27.7	25.6	672	Incidence	2337857	All Races	Female	Female Br	2004	26.6	31	28.7
Alabama	22.2	26	24.1	645	Incidence	2354514	All Races	Female	Female Br	2005	25.3	29.6	27.4
Alabama	24.4	28.4	26.4	718	Incidence	2385480	All Races	Female	Female Br	2006	27.9	32.4	30.1
Alabama	28.1	32.3	30.1	838	Incidence	2407275	All Races	Female	Female Br	2007	32.5	37.3	34.8
Alabama	23.5	27.3	25.3	714	Incidence	2430257	All Races	Female	Female Br	2008	27.3	31.6	29.4
Alabama	25.3	29.3	27.2	772	Incidence	2448159	All Races	Female	Female Br	2009	29.3	33.8	31.5
Alabama	25.6	27.3	26.5	3898	Incidence	12356437	All Races	Female	Female Br	2009-2013	30.6	32.6	31.5
Alabama	24.5	28.3	26.4	771	Incidence	2463573	All Races	Female	Female Br	2010	29.1	33.6	31.3
Alabama	24.1	27.9	25.9	763	Incidence	2472801	All Races	Female	Female Br	2011	28.7	33.1	30.9
Alabama	24.6	28.4	26.5	801	Incidence	2481503	All Races	Female	Female Br	2012	30.1	34.6	32.3
Alabama	24.5	28.3	26.3	791	Incidence	2490401	All Races	Female	Female Br	2013	29.6	34.1	31.8
Alabama	0.9	1.8	1.3	30	Incidence	2293259	All Races	Female	Hodgkin L	1999	0.9	1.9	1.3
Alabama	~	~	~	~	Mortality	2293259	All Races	Female	Hodgkin L	1999	~	~	~
Alabama	1.7	2.9	2.2	52	Incidence	2302835	All Races	Female	Hodgkin L	2000	1.7	3	2.3
Alabama	~	~	~	~	Mortality	2302835	All Races	Female	Hodgkin L	2000	~	~	~
Alabama	1.6	2.8	2.1	49	Incidence	2309496	All Races	Female	Hodgkin L	2001	1.6	2.8	2.1
Alabama	~	~	~	~	Mortality	2309496	All Races	Female	Hodgkin L	2001	~	~	~

In the data present, above, the entries circled in red shows the types of missing data and redundant values that we first eliminated. The entries 0, ~, -, and 2009-2013 were eliminated since 0 are null values, ~ and – are missing values and 2009-2013 are redundant observations (since observations for the years 2009, 2010, 2011, 2012, 2013) are present in the dataset separately.

2. **Created dummy variables for all categorical variables in dataset**

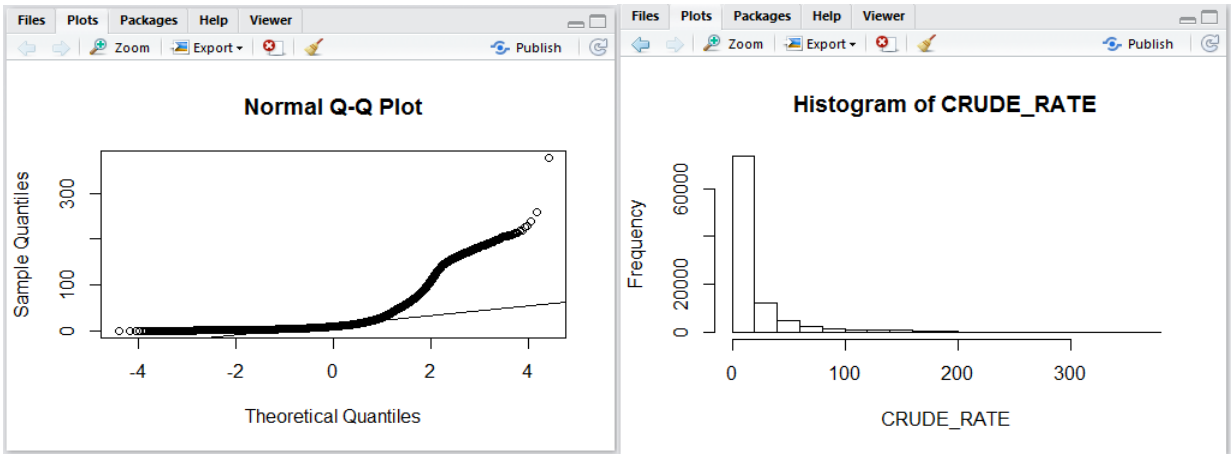
Next we created dummy variables for all categorical variables present in our dataset. Since our dependent variable is CRUDE_RATE which is a continuous variable we created dummy variables for the independent variables AREA, SEX, RACE, SITE and EVENT_TYPE. What the dataset looks like after creating dummy variables is as follows:

AGE_ADJU	AGE_ADJU	AGE_ADJU	YEAR	CRUDE_CI	CRUDE_CI	CRUDE_RA	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connectic	Delaware
48.8	151.9	88.2	2001	45.5	129.1	4.375757	1	0	0	0	0	0	0	0
47.7	139.6	83.4	2003	44.1	121.3	4.328098	1	0	0	0	0	0	0	0
35.6	107.5	63	2005	35.6	101.1	4.130355	1	0	0	0	0	0	0	0
47.7	129.3	80.5	2006	44.6	112.8	4.290459	1	0	0	0	0	0	0	0
71.4	177.8	116.2	2008	50.4	117	4.364372	1	0	0	0	0	0	0	0
49	167	97.2	2009	37.8	95.6	4.12552	1	0	0	0	0	0	0	0
68.7	149.3	102.9	2010	62.4	130.4	4.520701	1	0	0	0	0	0	0	0
37.5	102	63.8	2011	32.6	84.6	3.992681	1	0	0	0	0	0	0	0
77.2	163.5	114.6	2012	66.9	133.6	4.565389	1	0	0	0	0	0	0	0
53.9	128.5	85.7	2013	44.8	100.5	4.228293	1	0	0	0	0	0	0	0
68.5	206.9	125.7	2011	37.9	95.7	4.127134	1	0	0	0	0	0	0	0
2.1	5.3	3.5	1999	2	5	1.163151	1	0	0	0	0	0	0	0
1.7	4.7	2.9	2000	1.5	4.1	0.955511	1	0	0	0	0	0	0	0
2.3	5.7	3.7	2001	2.1	5.1	1.193922	1	0	0	0	0	0	0	0
1.6	4.4	2.7	2002	1.4	4.1	0.916291	1	0	0	0	0	0	0	0
1.7	4.6	2.9	2003	1.6	4.3	0.993252	1	0	0	0	0	0	0	0
1.8	4.6	2.9	2004	1.8	4.6	1.098612	1	0	0	0	0	0	0	0
2	5.1	3.3	2005	2	5	1.163151	1	0	0	0	0	0	0	0
1.5	4.2	2.6	2006	1.4	4	0.875469	1	0	0	0	0	0	0	0
1.5	4.2	2.6	2007	1.5	4.1	0.955511	1	0	0	0	0	0	0	0
1.6	4.5	2.8	2007	1.4	3.9	0.875469	1	0	0	0	0	0	0	0
2.9	6.3	4.4	2008	2.9	6.2	1.458615	1	0	0	0	0	0	0	0
2.4	5.6	3.8	2009	2.4	5.4	1.308333	1	0	0	0	0	0	0	0
1.4	3.9	2.5	2010	1.6	4.2	0.955511	1	0	0	0	0	0	0	0

Here all the categorical variables have been converted to binary variables having the values 1 or 0 that can be used for analysis.

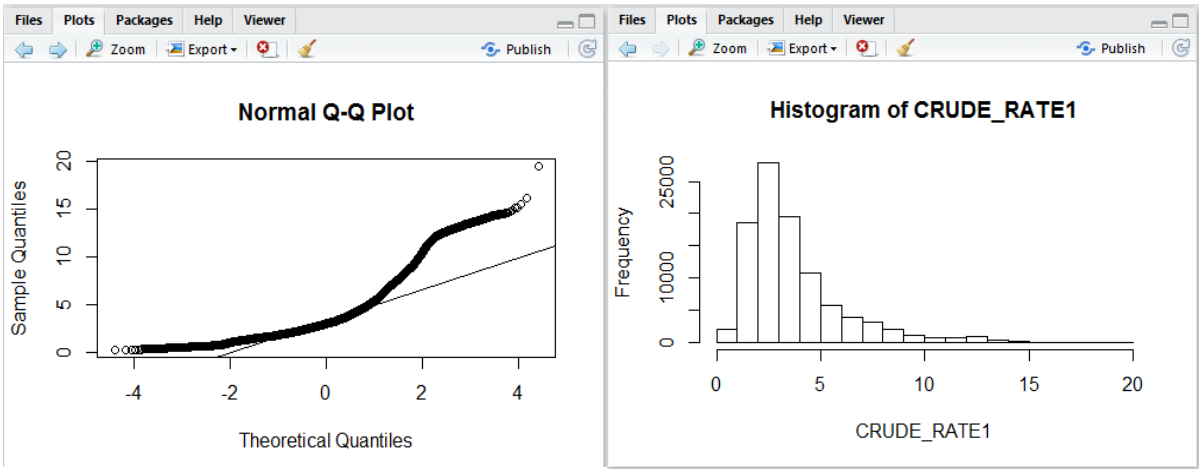
3. Square-root Transformation and Log transformation

Next we needed to check if the data was normal or not. Hence we generated the normal probability plot and histogram that are given as follows:



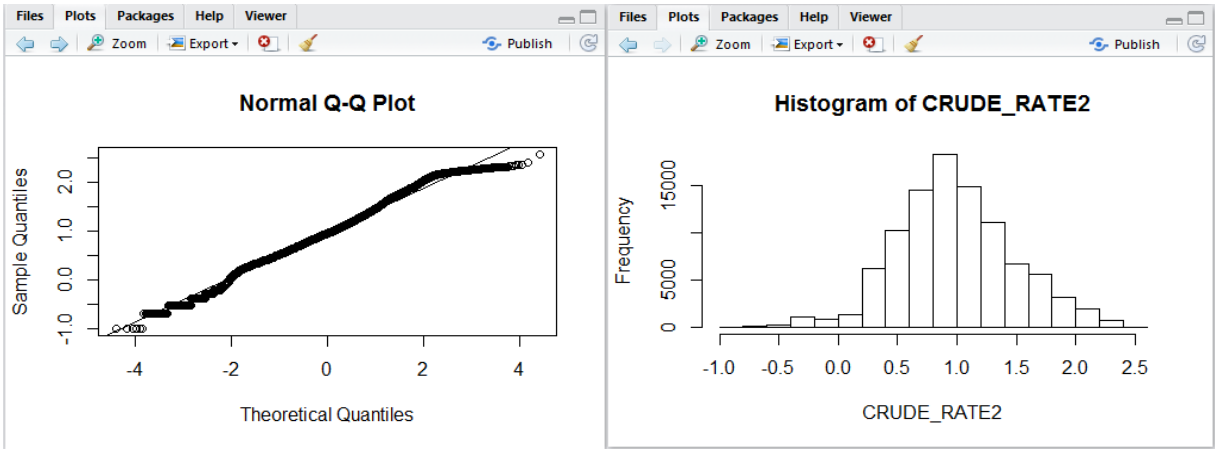
Here it can be seen that the data is not normal and hence we move on to the next step to normalize the data.

Next we performed square root transformation to normalize the data. The output that we obtained after performing square root transformation is as follows:



After performing square root transformation, the data improved but was still not in normal form. Hence we move on towards performing logarithmic transformation.

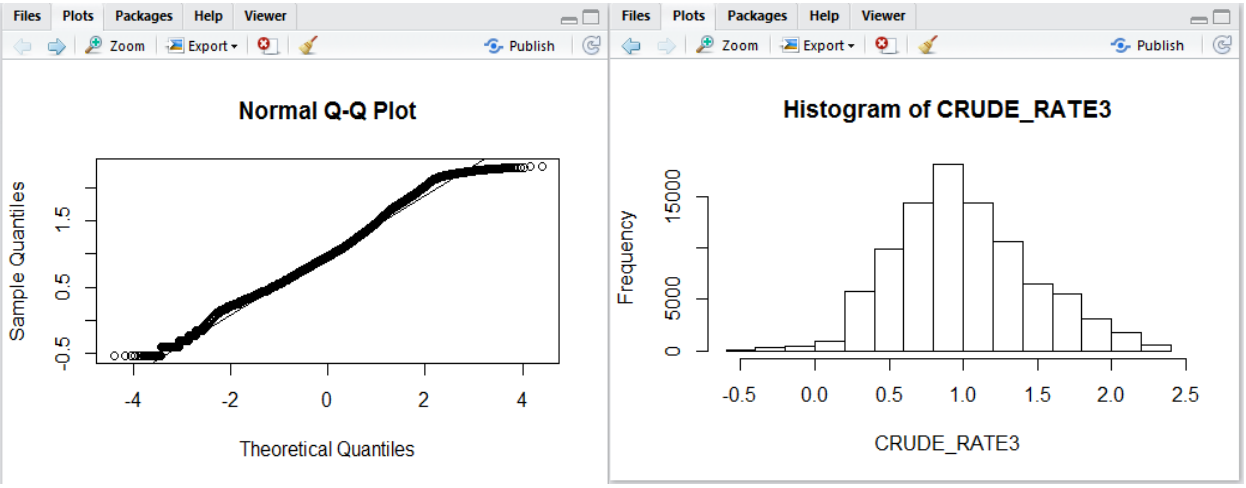
On performing logarithmic transformation, the data was obtained in normal form and the normal probability plot, histogram is as follows:



Hence after performing logarithmic transformation the data obtained was in normalized form. The next step is removal of outliers.

4. Outlier removal

As seen in the output after logarithmic transformation, the data is normalized and now the outliers are to be removed. The normal probability plot and histogram after removal of outliers is as follows:



After removal of outliers the data cleaning process is complete and the dataset obtained after completing the cleaning process has 92555 observations and we can now move on to the analysis part.

5. APPROACH AND METHODS

For the analysis part, we first divided the data into two parts; 80% as the training set and 20% as the validation set. The data set has 92 variables and 92555 observations and after dividing the dataset the training data has 74177 observations and the validation dataset has 18378 observations. After this we performed stepwise regression analysis to select and eliminate variables from the training dataset.

Stepwise regression analysis:

Stepwise regression analysis is the technique we used in the dataset to eliminate and select independent variables for analysis. The direction that we used was “both” since we want to select as well as eliminate variables. The analysis part is as follows:

```
model <- step(lm(CRUDE_RATE~1, tdata), direction='both', scope=~ AGE_ADJUSTED_CI_LOWER+AGE_ADJUSTED_CI_UPPER+
              AGE_ADJUSTED_RATE+ YEAR+ CRUDE_CI_LOWER+ CRUDE_CI_UPPER+ Alabama+ Alaska+ Arizona+ Arkansas+
              California+ Colorado+ Connecticut+ Delaware+ Florida+ Georgia+ Hawaii+ Idaho+ Illinois+
              Indiana+ Iowa+ Kansas+ Kentucky+ Louisiana+ Maine+ Maryland+ Massachusetts+ Michigan+
              Minnesota+ Mississippi+ Missouri+ Montana+ Nebraska+ Nevada+ New_Hampshire+ New_Jersey+
              New_Mexico+ New_York+ North_Carolina+ North_Dakota+ Ohio+ Oklahoma+ Oregon+ Pennsylvania+
              Rhode_Island+ South_Carolina+ South_Dakota+ Tennessee+ Texas+ Utah+ Vermont+ Virginia+
              Washington+ West_Virginia+ Wisconsin+ Wyoming+ Incidence+ Mortality+ Asian_Pacific_Islander+
              Black+ Hispanic+ White1+ American_Indian_Alaska_Native+ Female_Breast1+ Prostate+
              Brain_and_Other_NerV_System+ Cervix+ Colon_and_Rectum+ Corpus_and_Uterus_NOS+ Esophagus+
              Female_Breast+ Hodgkin_Lymphoma+ Kidney_and_Renal_Pelvis+ Larynx+ Leukemias+
              Liver_Intrahepatic_Bile_Duct+ Lung_and_Bronchus+ Myeloma+ Non_Hodgkin_Lymphoma+
              Oral_Cavity_and_Pharynx+ Ovary+ Pancreas+ Stomach+ Thyroid+ Urinary_Bladder+
              Melanomas_of_the_Skin+ Mesothelioma+ Testis+ Kaposi_Sarcoma+ Female+ Male)
```

```
Console ~/
> summary(regression)

Call:
lm(formula = CRUDE_RATE ~ CRUDE_CI_UPPER + Mesothelioma + Colon_and_Rectum +
    Incidence + Hodgkin_Lymphoma + Lung_and_Bronchus + Larynx +
    Kaposi_Sarcoma + Hispanic + Esophagus + Male + Female_Breast1 +
    Female_Breast + Pancreas + Prostate + Corpus_and_Uterus_NOS +
    Non_Hodgkin_Lymphoma + Ovary + White1 + Leukemias + Kidney_and_Renal_Pelvis +
    Urinary_Bladder + Black + Melanomas_of_the_Skin + Testis +
    California + Utah + Texas + Liver_Intrahepatic_Bile_Duct +
    Georgia + Hawaii + Cervix + Colorado + Arizona + YEAR + Thyroid +
    Illinois + Nevada + AGE_ADJUSTED_CI_UPPER + CRUDE_CI_LOWER +
    AGE_ADJUSTED_RATE + AGE_ADJUSTED_CI_LOWER + Stomach + Oral_Cavity_and_Pharynx +
    Alaska + American_Indian_Alaska_Native + Pennsylvania + Idaho +
    Florida + Connecticut + Virginia + Nebraska + Kansas + North_Carolina +
    Maryland + Minnesota + Wyoming + Rhode_Island + Washington +
    Oregon + Maine + Ohio + West_Virginia + Delaware + Missouri +
    Indiana + New_Hampshire + Michigan + South_Dakota)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17481 -0.23192  0.02221  0.24224  0.89167

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -6.0865861  0.6032153  -10.090  < 2e-16 ***
CRUDE_CI_UPPER                 0.0279954  0.0005795   48.310  < 2e-16 ***
Mesothelioma                  -1.4051577  0.0115956 -121.180  < 2e-16 ***
Colon_and_Rectum               1.1938593  0.0068230  174.976  < 2e-16 ***
Incidence                      0.5487370  0.0030129  182.127  < 2e-16 ***
Hodgkin_Lymphoma              -0.8713970  0.0098303  -88.644  < 2e-16 ***
Lung_and_Bronchus              1.3728514  0.0074752  183.654  < 2e-16 ***
Larynx                        -0.5611346  0.0086906  -64.568  < 2e-16 ***
Kaposi_Sarcoma                -1.6184027  0.0258018  -62.724  < 2e-16 ***
Hispanic                      -0.1056634  0.0067168  -15.731  < 2e-16 ***
Esophagus                     -0.1527328  0.0077430  -19.725  < 2e-16 ***
Male                           0.3139172  0.0029336  107.009  < 2e-16 ***
Female_Breast1                 1.3985270  0.0097118  144.003  < 2e-16 ***
Female_Breast                  1.1628080  0.0105644  110.068  < 2e-16 ***
Pancreas                       0.6655286  0.0068175   97.621  < 2e-16 ***
Prostate                       1.0233134  0.0101224  101.094  < 2e-16 ***
```

```
Prostate                        1.0233134  0.0101224  101.094  < 2e-16 ***
Corpus_and_Uterus_NOS           0.8166350  0.0087218   93.632  < 2e-16 ***
Non_Hodgkin_Lymphoma            0.5898935  0.0068502   86.113  < 2e-16 ***
Ovary                           0.7446401  0.0088100   84.522  < 2e-16 ***
White1                          0.3561629  0.0061452   57.958  < 2e-16 ***
Leukemias                       0.4827829  0.0068702   70.272  < 2e-16 ***
Kidney_and_Renal_Pelvis         0.3759586  0.0069632   53.992  < 2e-16 ***
Urinary_Bladder                 0.3519469  0.0071973   48.900  < 2e-16 ***
Black                           0.2026262  0.0060919   33.262  < 2e-16 ***
Melanomas_of_the_Skin           0.3309569  0.0085023   38.926  < 2e-16 ***
Testis                         -0.3167027  0.0137210  -23.082  < 2e-16 ***
California                     -0.1751270  0.0065350  -26.798  < 2e-16 ***
Utah                           -0.3397110  0.0121541  -27.950  < 2e-16 ***
Texas                          -0.1550602  0.0069369  -22.353  < 2e-16 ***
Liver_Intrahepatic_Bile_Duct    0.2037106  0.0070841   28.756  < 2e-16 ***
Georgia                        -0.1470296  0.0081895  -17.953  < 2e-16 ***
Hawaii                          0.1560303  0.0111377   14.009  < 2e-16 ***
Cervix                          0.2118120  0.0092759   22.835  < 2e-16 ***
Colorado                       -0.1465648  0.0090430  -16.208  < 2e-16 ***
Arizona                        -0.1385130  0.0083106  -16.667  < 2e-16 ***
YEAR                            0.0034365  0.0003006   11.432  < 2e-16 ***
Thyroid                         0.1514039  0.0101734   14.882  < 2e-16 ***
Illinois                       -0.0673084  0.0070937   -9.488  < 2e-16 ***
Nevada                         -0.1206564  0.0105299  -11.458  < 2e-16 ***
AGE_ADJUSTED_CI_UPPER          -0.0873341  0.0023853  -36.614  < 2e-16 ***
CRUDE_CI_LOWER                 -0.0128468  0.0007155  -17.955  < 2e-16 ***
AGE_ADJUSTED_RATE               0.1924791  0.0057926   33.228  < 2e-16 ***
AGE_ADJUSTED_CI_LOWER          -0.1071964  0.0036069  -29.720  < 2e-16 ***
Stomach                         0.0833223  0.0070441   11.829  < 2e-16 ***
Oral_Cavity_and_Pharynx         0.0744444  0.0073245   10.164  < 2e-16 ***
Alaska                         -0.1408764  0.0153762   -9.162  < 2e-16 ***
American_Indian_Alaska_Native   0.0943664  0.0119252    7.913 2.54e-15 ***
Pennsylvania                    0.0549774  0.0083281    6.601 4.10e-11 ***
Idaho                          -0.0939596  0.0126008   -7.457 8.97e-14 ***
Florida                         0.0391056  0.0070876    5.517 3.45e-08 ***
Connecticut                     0.0514871  0.0091693    5.615 1.97e-08 ***
Virginia                       -0.0554163  0.0087891   -6.305 2.90e-10 ***
Nebraska                       -0.0670646  0.0117877   -5.689 1.28e-08 ***
Kansas                         -0.0604159  0.0107526   -5.619 1.93e-08 ***
North_Carolina                 -0.0475773  0.0081417   -5.844 5.13e-09 ***
Maryland                       -0.0429727  0.0082298   -5.222 1.78e-07 ***
Minnesota                      -0.0523517  0.0104730   -4.999 5.78e-07 ***
Wyoming                        -0.0547339  0.0152538   -3.588 0.000333 ***
Rhode_Island                    0.0488712  0.0122891    3.977 6.99e-05 ***
Washington                     -0.0284523  0.0086283   -3.298 0.000976 ***
Oregon                         -0.0324529  0.0104083   -3.118 0.001822 **
Maine                           0.0401613  0.0121084    3.317 0.000911 ***
Ohio                           -0.0187002  0.0082736   -2.260 0.023809 *
West_Virginia                   0.0298694  0.0114775    2.602 0.009259 **
Delaware                        0.0307885  0.0124153    2.480 0.013145 *
Missouri                        0.0216404  0.0089930    2.406 0.016114 *
Indiana                        -0.0162695  0.0091625   -1.776 0.075794 .
New_Hampshire                  -0.0198538  0.0122932   -1.615 0.106311
Michigan                        0.0134773  0.0079232    1.701 0.088947 .
South_Dakota                    0.0210973  0.0138672    1.521 0.128167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.345 on 73879 degrees of freedom
Multiple R-squared:  0.8925,    Adjusted R-squared:  0.8924
F-statistic:  8886 on 69 and 73879 DF,  p-value: < 2.2e-16
```

Hence from the analysis above it can be seen that stepwise regression selected and eliminated variables which are then used for further analysis.

Following are the variables selected after stepwise regression:

CRUDE_CI_UPPER, Mesothelioma, Colon_and_Rectum, Mortality, Lung_and_Bronchus, Hodgkin_Lymphoma, Larynx, Kaposi_Sarcoma, Hispanic, Esophagus, Female, Female_Breast1, Female_Breast, Pancreas, Prostate, Corpus_and_Uterus_NOS, Non_Hodgkin_Lymphoma, Ovary, White1, Leukemias, Kidney_and_Renal_Pelvis, Urinary_Bladder, Black, Melanomas_of_the_Skin, Testis, Utah, California, Texas, Liver_Intrahepatic_Bile_Duct, Georgia, Hawaii, Cervix, Arizona, Colorado, YEAR, Thyroid, Nevada, Illinois, AGE_ADJUSTED_CI_UPPER, CRUDE_CI_LOWER, AGE_ADJUSTED_RATE, AGE_ADJUSTED_CI_LOWER, Stomach, Oral_Cavity_and_Pharynx, Alaska, Pennsylvania, Florida, Idaho, Asian_Pacific_Islander, North_Carolina, Kansas, Connecticut, Nebraska, Minnesota, Virginia, Wyoming, Washington, Maryland, Oregon, Maine, Rhode_Island, Indiana, New_Hampshire, Ohio, New_York, Alabama, Delaware, Vermont, Mississippi, West_Virginia, South_Dakota, Michigan, Missouri

Multiple Linear Regression:

Now we performed Multiple Regression on variables selected by Stepwise Regression on the Training Dataset. Since our dependent variable is CRUDE_RATE which is a continuous variable, multiple regression analysis is the analysis that we can perform as other analysis like log regression and clustering methods require categorical dependent variable. The analysis of multiple regression is as follows:

```
> summary(regression)

Call:
lm(formula = CRUDE_RATE ~ CRUDE_CI_UPPER + Mesothelioma + Colon_and_Rectum +
    Mortality + Lung_and_Bronchus + Hodgkin_Lymphoma + Larynx +
    Kaposi_Sarcoma + Hispanic + Esophagus + Female + Female_Breast1 +
    Female_Breast + Pancreas + Prostate + Corpus_and_Uterus_NOS +
    Non_Hodgkin_Lymphoma + Ovary + White1 + Leukemias + Kidney_and_Renal_Pelvis +
    Urinary_Bladder + Black + Melanomas_of_the_Skin + Testis +
    Utah + California + Texas + Liver_Intrahepatic_Bile_Duct +
    Georgia + Hawaii + Cervix + Arizona + Colorado + YEAR + Thyroid +
    Nevada + Illinois + AGE_ADJUSTED_CI_UPPER + CRUDE_CI_LOWER +
    AGE_ADJUSTED_RATE + AGE_ADJUSTED_CI_LOWER + Stomach + Oral_Cavity_and_Pharynx +
    Alaska + Pennsylvania + Florida + Idaho + Asian_Pacific_Islander +
    North_Carolina + Kansas + Connecticut + Nebraska + Minnesota +
    Virginia + Wyoming + Washington + Maryland + Oregon + Maine +
    Rhode_Island + Indiana + New_Hampshire + Ohio + New_York +
    Alabama + Delaware + Vermont + Mississippi + West_Virginia +
    South_Dakota + Michigan + Missouri)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18800 -0.23333  0.02272  0.24349  0.91322

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 -5.1836224  0.6039924   -8.582  < 2e-16 ***
CRUDE_CI_UPPER               0.0281758  0.0005842   48.228  < 2e-16 ***
Mesothelioma                -1.4003052  0.0117028 -119.655  < 2e-16 ***
Colon_and_Rectum             1.1928555  0.0067992  175.440  < 2e-16 ***
Mortality                   -0.5456158  0.0030135 -181.054  < 2e-16 ***
Lung_and_Bronchus            1.3707670  0.0074533  183.914  < 2e-16 ***
Hodgkin_Lymphoma            -0.8673968  0.0098700  -87.882  < 2e-16 ***
Larynx                      -0.5702427  0.0086955  -65.579  < 2e-16 ***
Kaposi_Sarcoma              -1.6155387  0.0255779  -63.161  < 2e-16 ***
Hispanic                    -0.1917286  0.0118023  -16.245  < 2e-16 ***
Esophagus                   -0.1539673  0.0076645  -20.088  < 2e-16 ***
Female                      -0.3143839  0.0029342 -107.143  < 2e-16 ***
Female_Breast1               1.3944307  0.0096841  143.992  < 2e-16 ***
Female_Breast                1.1600893  0.0106009  109.433  < 2e-16 ***
Pancreas                     0.6658243  0.0068431   97.299  < 2e-16 ***
Prostate                     1.0209869  0.0100550  101.540  < 2e-16 ***
Corpus_and_Uterus_NOS        0.8151183  0.0087422   93.239  < 2e-16 ***
Non_Hodgkin_Lymphoma         0.5868113  0.0068347   85.858  < 2e-16 ***
Ovary                        0.7436500  0.0088456   84.070  < 2e-16 ***
White1                       0.2659920  0.0118134   22.516  < 2e-16 ***
Leukemias                    0.4787948  0.0068995   69.395  < 2e-16 ***
Kidney_and_Renal_Pelvis      0.3737998  0.0069895   53.480  < 2e-16 ***
Urinary_Bladder              0.3537616  0.0072096   49.068  < 2e-16 ***
Black                        0.1137784  0.0114370    9.948  < 2e-16 ***
Melanomas_of_the_Skin        0.3424578  0.0084650   40.456  < 2e-16 ***
Testis                      -0.3098899  0.0137781  -22.491  < 2e-16 ***
Utah                        -0.3393403  0.0121121  -28.017  < 2e-16 ***
California                  -0.1734725  0.0067515  -25.694  < 2e-16 ***
Texas                       -0.1542812  0.0070964  -21.741  < 2e-16 ***
Liver_Intrahepatic_Bile_Duct 0.2044549  0.0070711   28.914  < 2e-16 ***
Georgia                     -0.1493510  0.0083316  -17.926  < 2e-16 ***
Hawaii                       0.1518595  0.0113033   13.435  < 2e-16 ***
Cervix                       0.2136401  0.0092229   23.164  < 2e-16 ***
Arizona                     -0.1346926  0.0085478  -15.758  < 2e-16 ***
Colorado                    -0.1356200  0.0091527  -14.817  < 2e-16 ***
YEAR                         0.0034608  0.0003008   11.507  < 2e-16 ***
Thyroid                      0.1560504  0.0101154   15.427  < 2e-16 ***
```

```
Nebraska                        -0.0657823   0.0117223    -5.612 2.01e-08 ***
Minnesota                       -0.0554462   0.0105325    -5.264 1.41e-07 ***
Virginia                        -0.0471229   0.0089998    -5.236 1.65e-07 ***
Wyoming                         -0.0668159   0.0155099    -4.308 1.65e-05 ***
Washington                      -0.0434840   0.0087909    -4.947 7.57e-07 ***
Maryland                        -0.0401184   0.0082467    -4.865 1.15e-06 ***
Oregon                          -0.0423545   0.0105343    -4.021 5.81e-05 ***
Maine                            0.0502095   0.0122361     4.103 4.08e-05 ***
Rhode_Island                     0.0400896   0.0124036     3.232  0.00123 **
Indiana                         -0.0250638   0.0092648    -2.705  0.00683 **
New_Hampshire                   -0.0304028   0.0124577    -2.440  0.01467 *
Ohio                            -0.0181003   0.0084000    -2.155  0.03118 *
New_York                        -0.0125821   0.0069505    -1.810  0.07026 .
Alabama                         -0.0161637   0.0088249    -1.832  0.06701 .
Delaware                         0.0337248   0.0125813     2.681  0.00735 **
Vermont                          0.0309405   0.0141469     2.187  0.02874 *
Mississippi                      0.0210945   0.0098020     2.152  0.03140 *
West_Virginia                    0.0222466   0.0117968     1.886  0.05932 .
South_Dakota                     0.0232759   0.0141094     1.650  0.09901 .
Michigan                         0.0147242   0.0080654     1.826  0.06791 .
Missouri                         0.0155734   0.0090623     1.718  0.08571 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3455 on 74102 degrees of freedom
Multiple R-squared:  0.8923,    Adjusted R-squared:  0.8922
F-statistic:  8411 on 73 and 74102 DF,  p-value: < 2.2e-16
```

Here, on performing regression analysis, the adjusted R- squared value obtained is 0.89 i.e. 89%. Hence, the independent variables explain/fit 89% of our model which is a good value.

Predicting the validation data:

Now we use the predict function to predict CRUDE_RATE in the validation dataset using the analysis of training data and observations of validation data. This gives the following result:

```
> View(vdata)
> prediction <- predict(regression,vdata)
> head(prediction)
       1        2        3        4        5        6
5.015041 1.589197 1.606030 1.658283 1.643804 1.323747
> head(vdata)
   X AGE_ADJUSTED_CI_LOWER AGE_ADJUSTED_CI_UPPER AGE_ADJUSTED_RATE YEAR CRUDE_CI_LOWER CRUDE_CI_UPPER
1  7                 68.7                 149.3              102.9 2010           62.4          130.4
2 15                  1.6                   4.4                2.7 2002            1.4            4.1
3 16                  1.7                   4.6                2.9 2003            1.6            4.3
4 23                  2.4                   5.6                3.8 2009            2.4            5.4
5 24                  1.4                   3.9                2.5 2010            1.6            4.2
6 45                  3.4                   7.2                5.0 2007            3.0            6.4
  CRUDE_RATE Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida Georgia Hawaii
1  4.5207010       1      0       0        0          0        0           0        0       0       0      0
2  0.9162907       1      0       0        0          0        0           0        0       0       0      0
3  0.9932518       1      0       0        0          0        0           0        0       0       0      0
4  1.3083328       1      0       0        0          0        0           0        0       0       0      0
5  0.9555114       1      0       0        0          0        0           0        0       0       0      0
6  1.5040774       1      0       0        0          0        0           0        0       0       0      0
```

Here it can be seen that the predicted values of CRUDE_RATE and the actual values of CRUDE_RATE are not very different. Hence to check the correctness of prediction we calculate the mean absolute error.

Accuracy Metrics:

```
Wyoming                -0.0547339  0.0152538   -3.588 0.000333 ***
Rhode_Island            0.0488712  0.0122891    3.977 6.99e-05 ***
Washington             -0.0284523  0.0086283   -3.298 0.000976 ***
Oregon                 -0.0324529  0.0104083   -3.118 0.001822 **
Maine                   0.0401613  0.0121084    3.317 0.000911 ***
Ohio                   -0.0187002  0.0082736   -2.260 0.023809 *
West_Virginia           0.0298694  0.0114775    2.602 0.009259 **
Delaware                0.0307885  0.0124153    2.480 0.013145 *
Missouri                0.0216404  0.0089930    2.406 0.016114 *
Indiana                -0.0162695  0.0091625   -1.776 0.075794 .
New_Hampshire          -0.0198538  0.0122932   -1.615 0.106311
Michigan                0.0134773  0.0079232    1.701 0.088947 .
South_Dakota            0.0210973  0.0138672    1.521 0.128167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.345 on 73879 degrees of freedom
Multiple R-squared:  0.8925,    Adjusted R-squared:  0.8924
F-statistic:  8886 on 69 and 73879 DF,  p-value: < 2.2e-16
```

```
18360    2.917771
18361    4.689511
18362    2.721295
18363    2.667228
18364    2.028148
18365    1.974081
18366    1.856298
18367    1.824549
18368    2.406945
18369    2.066863
18370    1.757858
18371    3.380995
18372    3.597312
18373    2.014903
18374    3.653252
18375    2.066863
18376    2.091864
18377    1.987874
Name: CRUDE_RATE, dtype: float64
Mean Absolute error 0.276133291821
```

Here the mean absolute error is 0.276 i.e. 27.6% and the accuracy of our prediction is 72% which is a good and acceptable number.

6. CONCLUSION

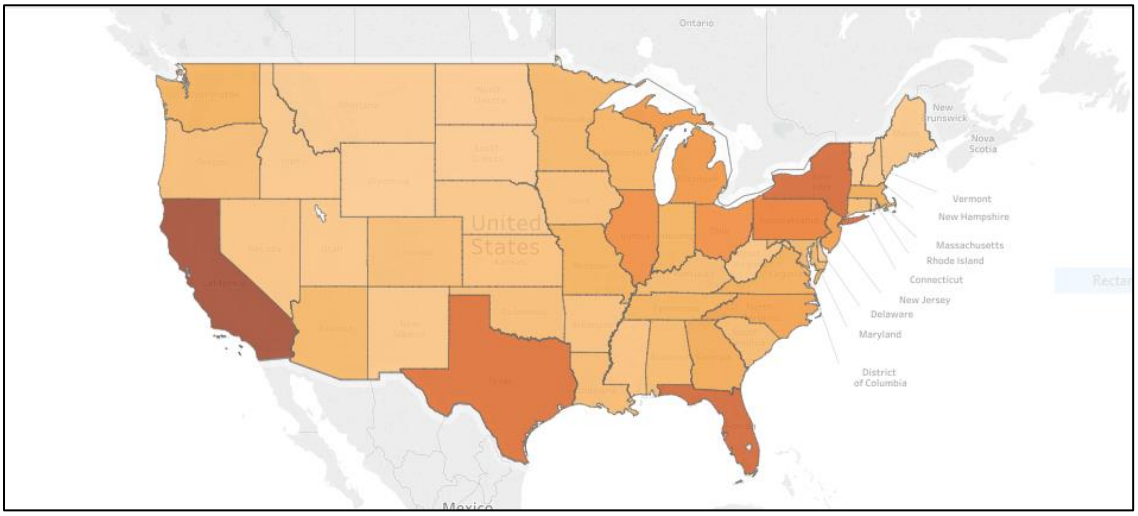
Our model predicts crude rate under specific clauses like gender, race, type of cancer and area of living. Crude rates will be helpful in determining the cancer burden and specific needs for services for given area and race. Also, Crude-rate will be used to plan for population-based cancer prevention and control interventions.

We trained our dataset using multiple regression model with R-square value as 0.892. Using Stepwise regression and multiple linear regressions, we could predict the probability of cancer for an individual of race, from a certain state with a specific cancer type with a mean absolute error of 27.6%. The accuracy we achieved was almost 74% which is a pretty high. Our model is successful and almost accurate in predicting the probability of cancer in under specific clauses like gender, race, type of cancer and area of living.

Our way forward will be to fine tune our model and make it more robust to handle any random observation. A better understanding of the independent variables and using other approaches like Cart Algorithm based on Classification and Regression Trees might lead towards better results with high accuracy.

7. ANAYSIS OF DATA VISUALIZED IN TABLEAU:

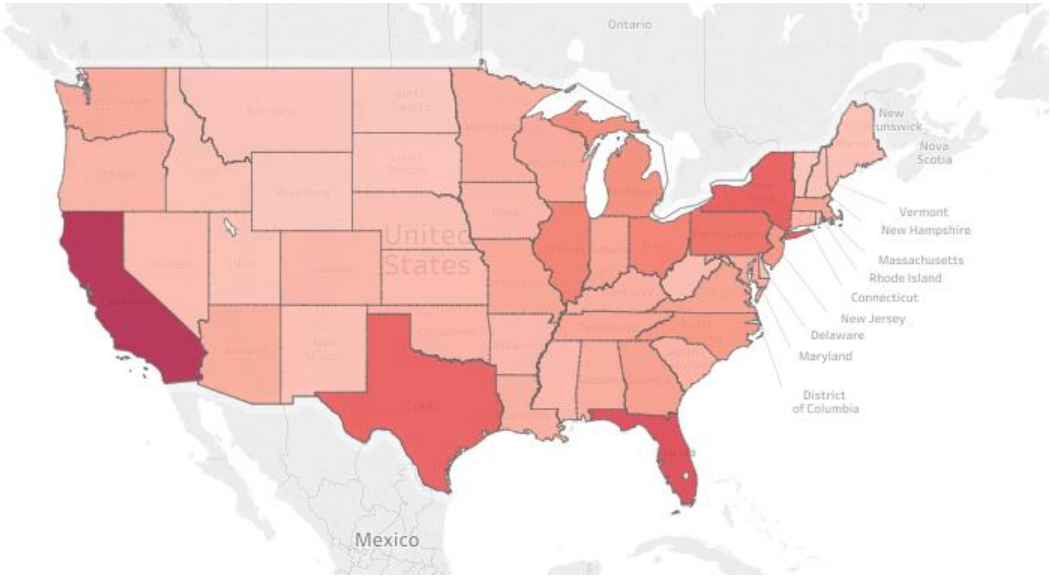
1. Number of Cancer Cases in 2005 of people from all races, both genders and all type of cancers



From the plot, we can make out that the maximum numbers of cases in the year 2005 were presented by the following states:

- California
- New York
- Florida
- Texas

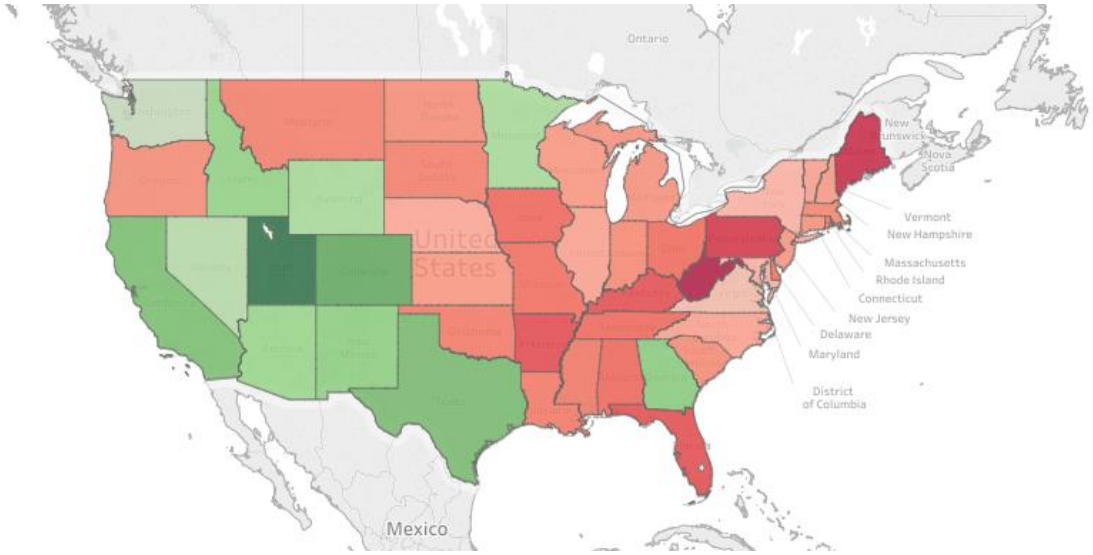
2. **<u>Number of Cancer Deaths in 2005 of people from all races, both genders and all type of cancers</u>**



From the plot, we can make out that the maximum numbers of deaths in the year 2005 were also in the same states:

- California
- New York
- Florida
- Texas

3. **Population % Deaths caused by cancer in 2005 of people from all races, both genders and all type of cancers**



From the plots, we can instantly find out that the states with maximum number of Cancer cases and deaths are not the same as Population % of deaths caused by cancer. The states were:

- West Virginia (0.25%)
- Maine (0.24%)
- Pennsylvania (0.23%)
- Arkansas (0.22%)

The number of cases and deaths were higher in the other states as their population was high. But the percentage with respect to their population gives a different outcome.

4. **<u>Total number of Cancer Cases Registered through 1999-2013 of people from all races, both genders and all type of cancers</u>**

Through the years, we observe that, the Cancer cases have increased in different states as well. They are:

- Maine
- West Virginia
- Pennsylvania
- Florida
- Michigan

8. REFERENCES

- https://nccd.cdc.gov/uscs/
- https://www.cancer.gov/about-cancer/understanding/statistics
- https://www.cancer.org/research/cancer-facts-statistics.html