EAS 508 (Section 002): Homework 1

Gurteg Sawhney

1. What is your suggested model for predicting TARGET_deathRate? Give equation and method with final parameters.

The Suggested model for predicting TARGET_deathRate is Multiple Linear Regression.

Final Features -

```
'studyPerCap'
'incidenceRate'
'PctBachDeg25_Over'
'MedianAgeMale'
'PctHS18_24 '
'PctUnemployed16_Over'
'PctOtherRace'
'PctMarriedHouseholds'
'BirthRate'
```

Equation-

TARGET_deathRate = 144.21683+ (-0.01459)*studyPerCap + (0.20740)*incidenceRate + (-1.81558)*PctBachDeg25_Over + (-0.54007)*MedianAgeMale + (0.39306)*PctHS18_24 + (1.04836)*PctUnemployed16_Over + (-0.64385)*PctOtherRace + (-0.53268)*PctMarriedHouseholds + (-1.04198)*BirthRate +Error

```
Error=(Residual standard error: 19.72 on 1718 degrees of freedom)
```

2. What descriptors did you select and why? What impact should that have on the model?

The descriptors for the final Model are as below:

'studyPerCap' - Per capita number of cancer-related clinical trials per county

'incidenceRate' - Mean per capita (100,000) cancer diagoses

'PctBachDeg25_Over' - Percent of county residents ages 25 and over highest education attained: high school diploma

'MedianAgeMale'- Median age of male county residents

'PctHS18_24 '- Percent of county residents ages 18-24 highest education attained: high school diploma

'PctUnemployed16_Over' - Percent of county residents ages 16 and over unemployed

'PctOtherRace' - Percent of county residents who identify in a category which is not White, Black, or Asian

'PctMarriedHouseholds' - Percent of married households

'BirthRate' - Number of live births relative to number of women in county

Using Backward elimination of the features, the above 9 descriptors were the most significant ones

```
> step.model$results
   nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
1      1 25.34354 0.1963378 19.79407 1.557905 0.04672969 0.6883422
2      2 21.08765 0.4411741 16.14623 1.090445 0.05433466 0.9095715
3      3 20.38325 0.4791271 15.56393 1.171225 0.04766307 1.0618113
4      4 20.37870 0.4799843 15.54043 1.161744 0.04888705 1.0443050
5      5 20.07610 0.4954177 15.26245 1.130568 0.05606828 1.0412979
6      6 20.02322 0.4976461 15.15528 1.145977 0.05220690 1.1065838
7      7 19.95017 0.5009297 15.09286 1.113417 0.04957956 1.0101564
8      8 19.81108 0.5079642 14.91491 1.102271 0.05119039 0.9942433
9      9 19.77072 0.5096335 14.89235 1.117736 0.05266578 0.9827868
10    10 19.83331 0.5062445 14.93141 1.092822 0.05281075 0.9785122
11    11 19.82184 0.5068718 14.91557 1.053763 0.05302166 0.9546798
12    12 19.83678 0.5060972 14.94828 1.035554 0.05181342 0.9134489
13    13 19.86591 0.5046654 14.95997 1.043675 0.05204763 0.9275844
14    14 19.87423 0.5043165 14.96919 1.047434 0.05172690 0.9226566
15    15 19.86303 0.5048270 14.95779 1.039701 0.05156415 0.9122461
16    16 19.85029 0.5054355 14.95330 1.036135 0.05109667 0.9135934
17    17 19.85343 0.5052748 14.95728 1.032753 0.05122772 0.9129907
18    18 19.85642 0.5051128 14.95672 1.034294 0.05128909 0.9125689
19    19 19.85642 0.5051128 14.95672 1.034294 0.05128909 0.9125689
> step.model$bestTune
  nvmax
9     9
```

The above 9 descriptors were used because at n=9 with these features, the R Squared value was the maximum and MAE was the minimum. So these features would give the best performance. Also, there was no collinearity between these features, so no problem for Variance.

Feature Importance

```
                       Overall
studyPerCap           2.886580
incidenceRate        23.341771
PctBachDeg25_Over    16.606810
MedianAgeMale         5.289265
PctHS18_24            6.754969
PctUnemployed16_Over  6.131748
PctOtherRace          4.913550
PctMarriedHouseholds  6.210802
BirthRate             4.358739
```

Impact of these Descriptors on the Model

The impact of the descriptors on the model can be explained by the following equation

TARGET_deathRate = 144.21683+ (-0.01459)*studyPerCap + (0.20740)*incidenceRate + (-1.81558)*PctBachDeg25_Over + (-0.54007)*MedianAgeMale + (0.39306)*PctHS18_24 + (1.04836)*PctUnemployed16_Over + (-0.64385)*PctOtherRace + (-0.53268)*PctMarriedHouseholds + (-1.04198)*BirthRate +Error

Adjusted R Squared Value - 0.514

Had we used some other descriptors which were not significant, there would have been Variance, while validating on the test data. The results could have been good on the train data but would have differed while fitting on Test data

3. Why did you select the model as your final answer?

Model 1- Used all Features

- Step 1. Removed the column 'PctSomeCol18_24 ' which had 2285 null values (as there were a lot of missing values ~80% of rows). Imputed the missing values in 'PctEmployed16_Over' and the 'PctPrivateCoverageAlone' with the median values as the distribution of these variables followed a normal distribution
- Step 2. Removed features - Geography as it would not add any value, (can be used for descriptive analysis),Binned Income as we already have medIncome column
- Step 3 Split the remaining dataset into Test and Train (25:75), to train the model on train and then validate it on test
- Step 4 All the features remaining - n=31 were used for linear regression:

Results-

```
Multiple R-squared: 0.4132

Adjusted R-squared: 0.4138

p-value: < 2.2e-16
```

- Step 5 the same model was fit on the Test data

Results-

```
RMSE - 29.5

MAPE (Mean Average Oercent Deviation) - Fitted vs acatual - 12.36%
```

Model 2- Removed Highly Correlated Features

- Step 1. Removed the column 'PctSomeCol18_24 ' which had 2285 null values (as there were a lot of missing values ~80% of rows). Imputed the missing values in 'PctEmployed16_Over' and the 'PctPrivateCoverageAlone' with the median values as the distribution of these variables followed a normal distribution
- Step 2. Removed features - Geography as it would not add any value, (can be used for descriptive analysis),Binned Income as we already have medIncome column
- Step 3 Created a correlation matrix plot for all numerical features and removed highly correlated features (cor>0.75)
- Step 4 Split the remaining dataset into Test and Train (25:75), to train the model on train and then validate it on test
- Step 5 Now the obtained 21 features were used for linear regression:

Results-

```
Multiple R-squared:  0.4765
Adjusted R-squared:  0.4842
p-value: < 2.2e-16
```

- Step 7 the same model was fit on the Test data

Results-

```
RMSE - 22.8
MAPE (Mean Average Oercent Deviation) - Fitted vs acatual - 10.24%
```

Model 3 - Removed Outliers and Highly Correlated Features and Used Backward Elimination

- Step 1. Removed the column 'PctSomeCol18_24 ' which had 2285 null values (as there were a lot of missing values ~80% of rows). Imputed the missing values in 'PctEmployed16_Over' and the 'PctPrivateCoverageAlone' with the median values as the distribution of these variables followed a normal distribution
- Step 2. Removed outliers from the data using boxplot - 'avgDeathsPerYear','incidenceRate','studyPerCap'. Median Age column- removed all values >100 .Also removed features - Geography as it would not add any value, (can be used for descriptive analysis),Binned Income as we already have medIncome column
- Step 3 Created a correlation matrix plot for all numerical features and removed highly correlated features (cor>0.75)
- Step 4 Split the remaining dataset into Test and Train (25:75), to train the model on train and then validate it on test
- Step 5 Applied Backward Elimination on the train dataset(not applied on whole dataset as it would cause variance if test data is also used for Backward Elimination)
- Step 6 On the basis of results of Backward elimination - for n=9 features, the R squared value was the highest and p value was smallest compared to other n values
- Step 7 Now the obtained 9 features were used for linear regression:

Results-

```
Residual standard error: 19.72 on 1718 degrees of freedom
Multiple R-squared:  0.5165
Adjusted R-squared:  0.514
F-statistic: 203.9 on 9 and 1718 DF
p-value: < 2.2e-16
```

- Step 7 the same model was fit on the Test data

Results-

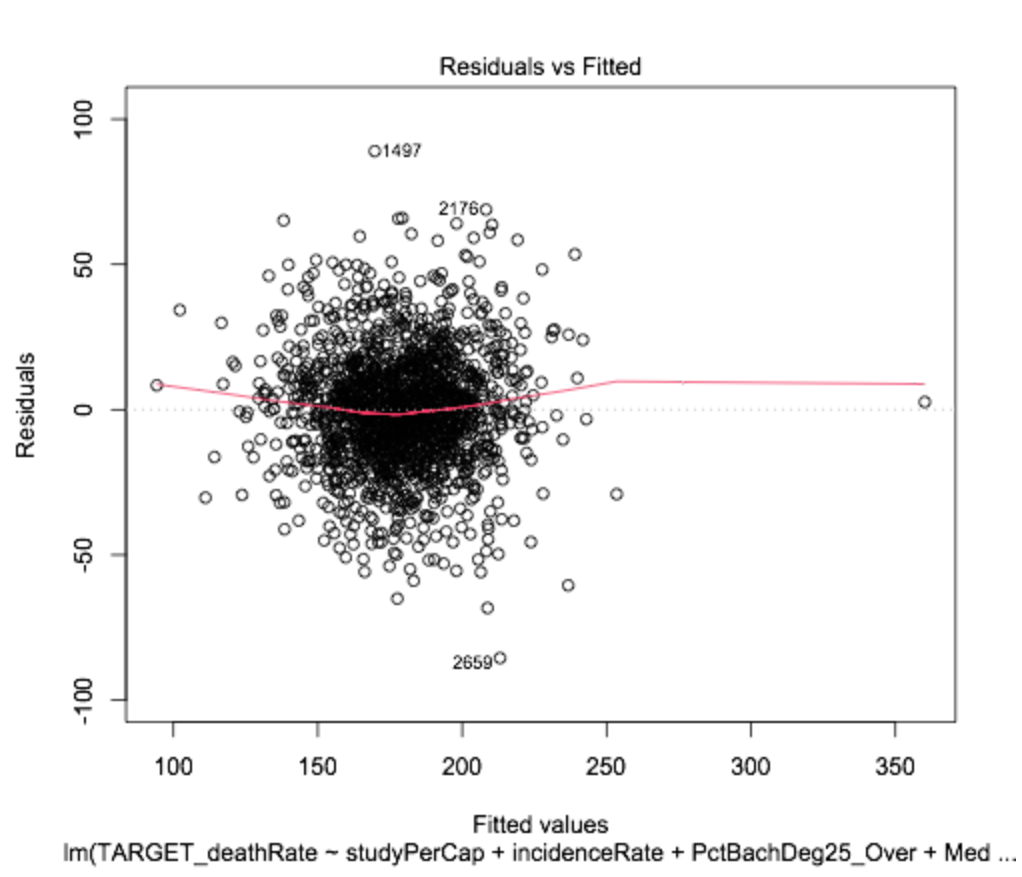
```
RMSE - 21.5
MAPE (Mean Average Oercent Deviation) - Fitted vs actual - 9.36%
```

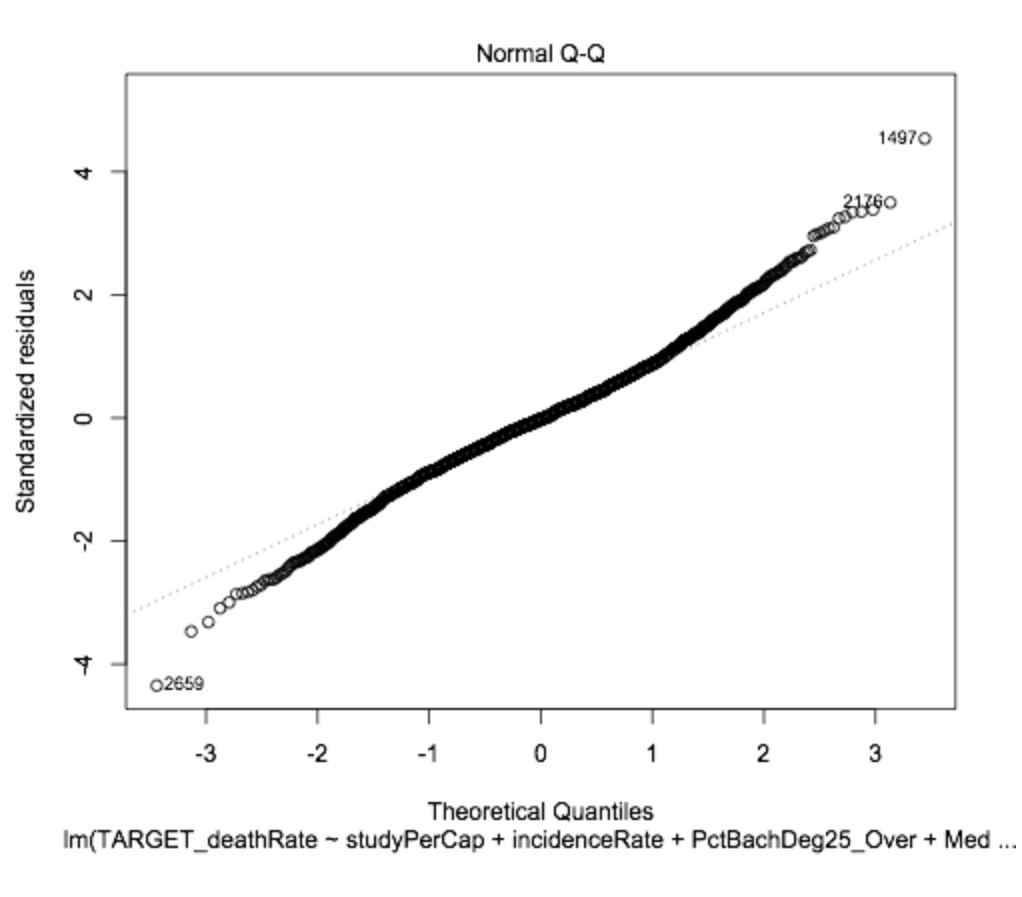
As the adjusted R squared value was the highest in Model 3 and the RMSE was the lowest in model 3, Model 3 was chosen as the final Model. Also there was no Multicollinearity between the features hence no problem of Variance while fitting on test data

Linear Regression Plots for the Final Model

Linear regression model assumptions:

- The errors has normal distribution
- The errors has mean 0
- Homoscedasticity of errors or equal variance
- The errors are independent.





Residual vs Fitted Values:

- The residuals spread randomly around the 0 line indicating that the relationship is linear.
- The residuals roughly horizontal placed around the 0 line indicating homogeneity of error

Normal Q-Q Plot:

- The residual points roughly lie within the lines and suggests that the error terms are indeed normally distributed.

variance(constant variance)

- No residuals are away from random pattern of residuals indicating no outliers.

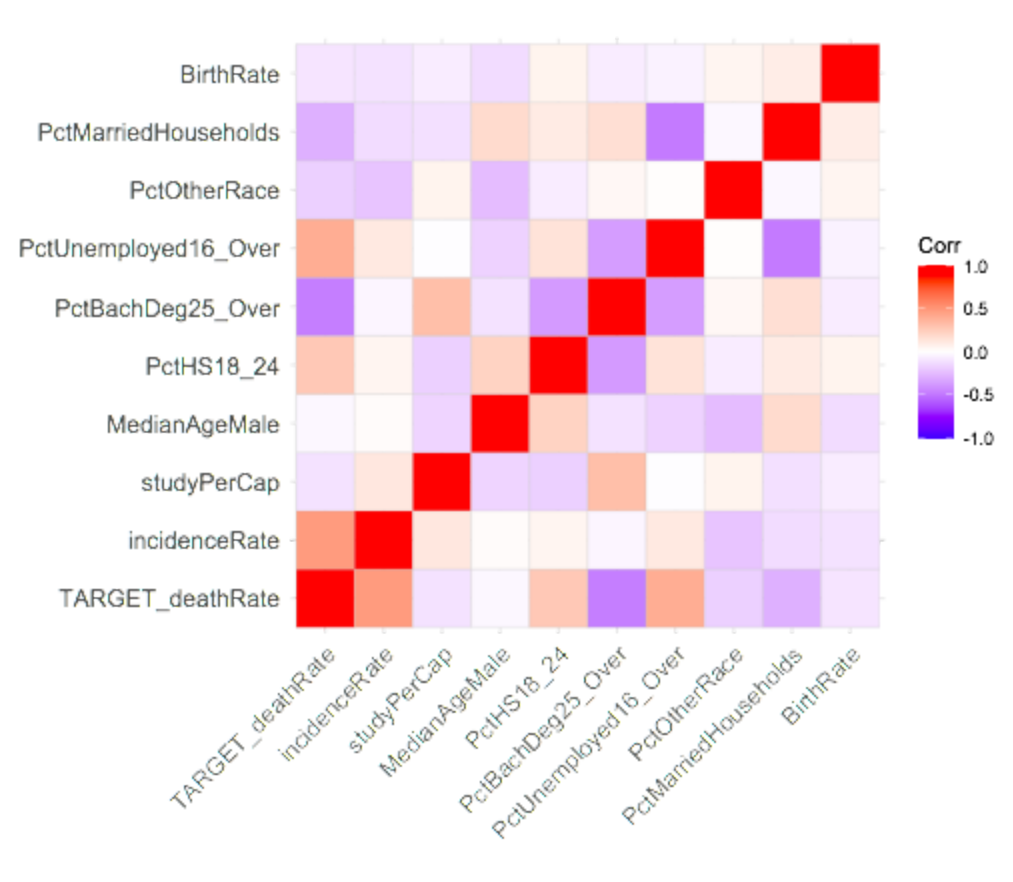
Therefore,the final Model- Model 3 satisfies all the assumptions of the linear regression model and is valid

4. Discuss your result in a maximum of 500 words and 3 supporting plots. The discussion needs to be coherent and understandable to a non-expert (Assuming that we are communicating our analysis to a higher official who is not a data scientist).

Result

TARGET_deatRate= 144.21683+ (-0.01459)*studyPerCap + (0.20740)*incidenceRate + (-1.81558)*PctBachDeg25_Over + (-0.54007)*MedianAgeMale + (0.39306)*PctHS18_24 + (1.04836)*PctUnemployed16_Over + (-0.64385)*PctOtherRace + (-0.53268)*PctMarriedHouseholds + (-1.04198)*BirthRate + Error

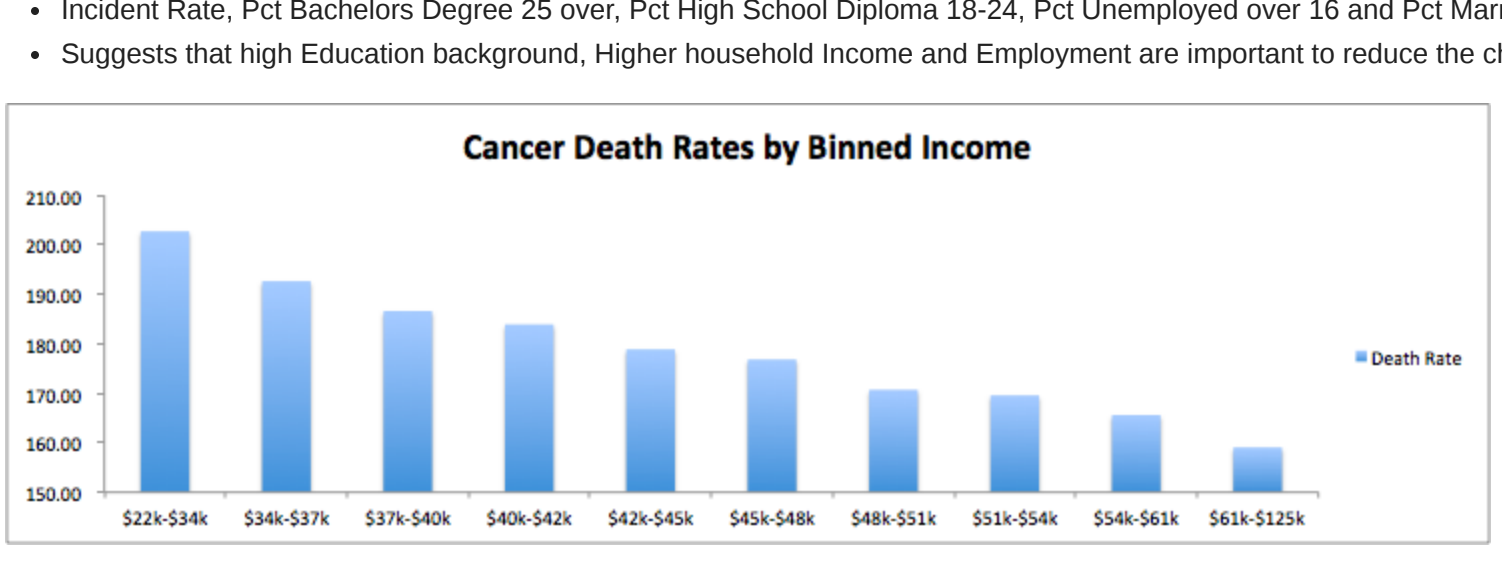
```
Feature               Importance
studyPerCap            2.886580
incidenceRate         23.341771
PctBachDeg25_Over     16.606810
MedianAgeMale          5.289265
PctHS18_24             6.754969
PctUnemployed16_Over   6.131748
PctOtherRace           4.913550
PctMarriedHouseholds   6.210802
BirthRate              4.358739
```



The above matrix shows the correlation between the features of the Model. Shows how the Incidence Rate and pct Unemployed over 16 are highly positively correlated with Death Rate, while Pct Bachelors Degree 25 over is negatively correlated.

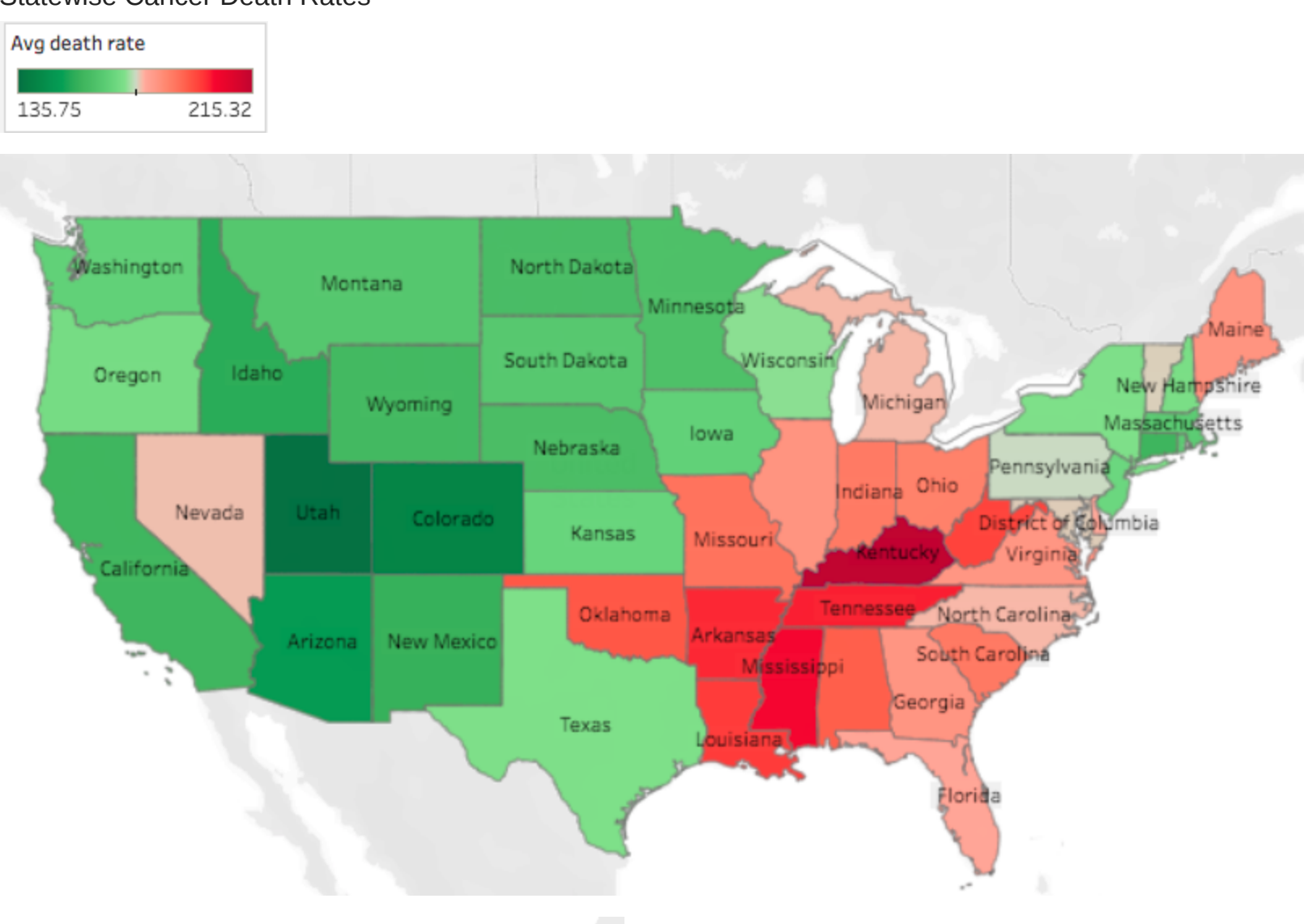
Summary

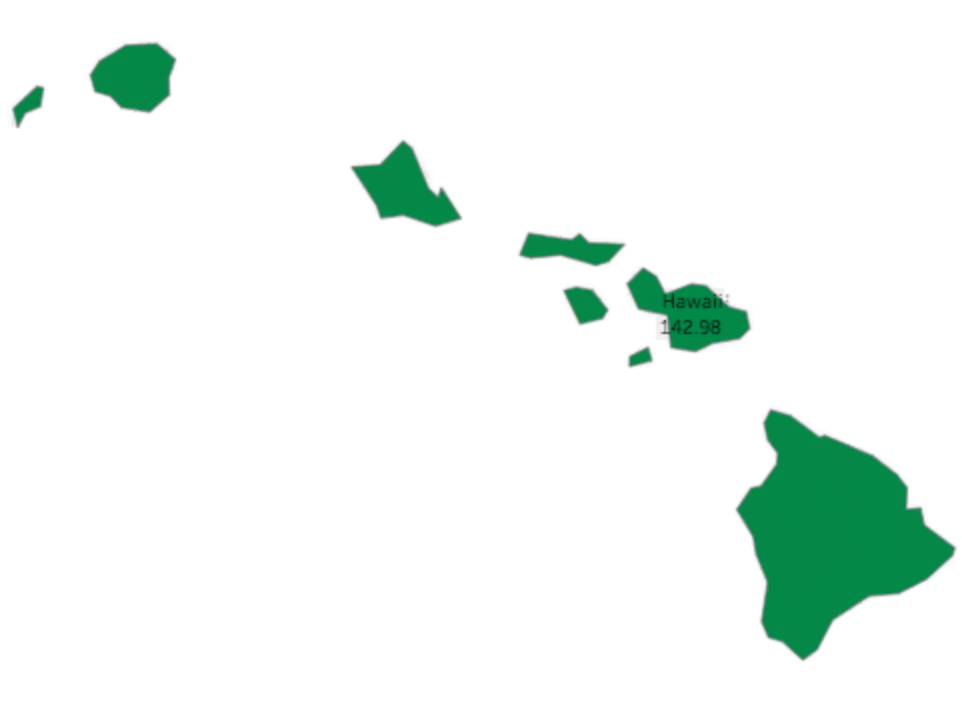
- Incident Rate, Pct Bachelors Degree 25 over, Pct High School Diploma 18-24, Pct Unemployed over 16 and Pct Married Househols are the most important features
- Suggests that high Education background, Higher household Income and Employment are important to reduce the chances of death rate due to Cancer



The graph above suggests that Higher the Income lower the Cancer Death Rate

Statewise Cancer Death Rates





Top 5 States with the highest Cancer Death Rates:

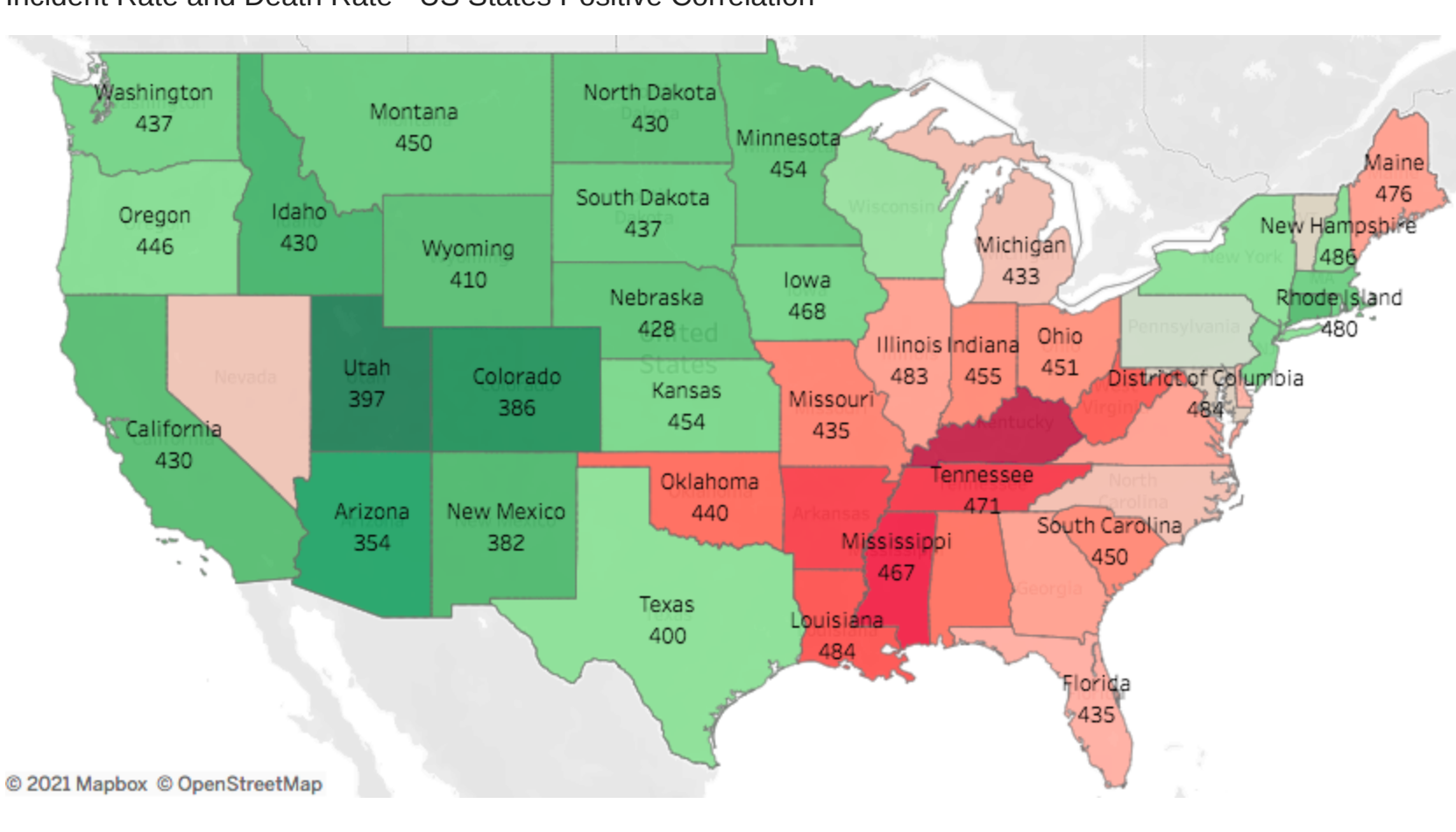
```
Kentucky
Mississippi
Tennessee
Arkansas
Louisiana
```

Top 5 States with the lowest Cancer Death Rates:

```
Utah
Colorado
Hawaii
Arizona
Idaho
```

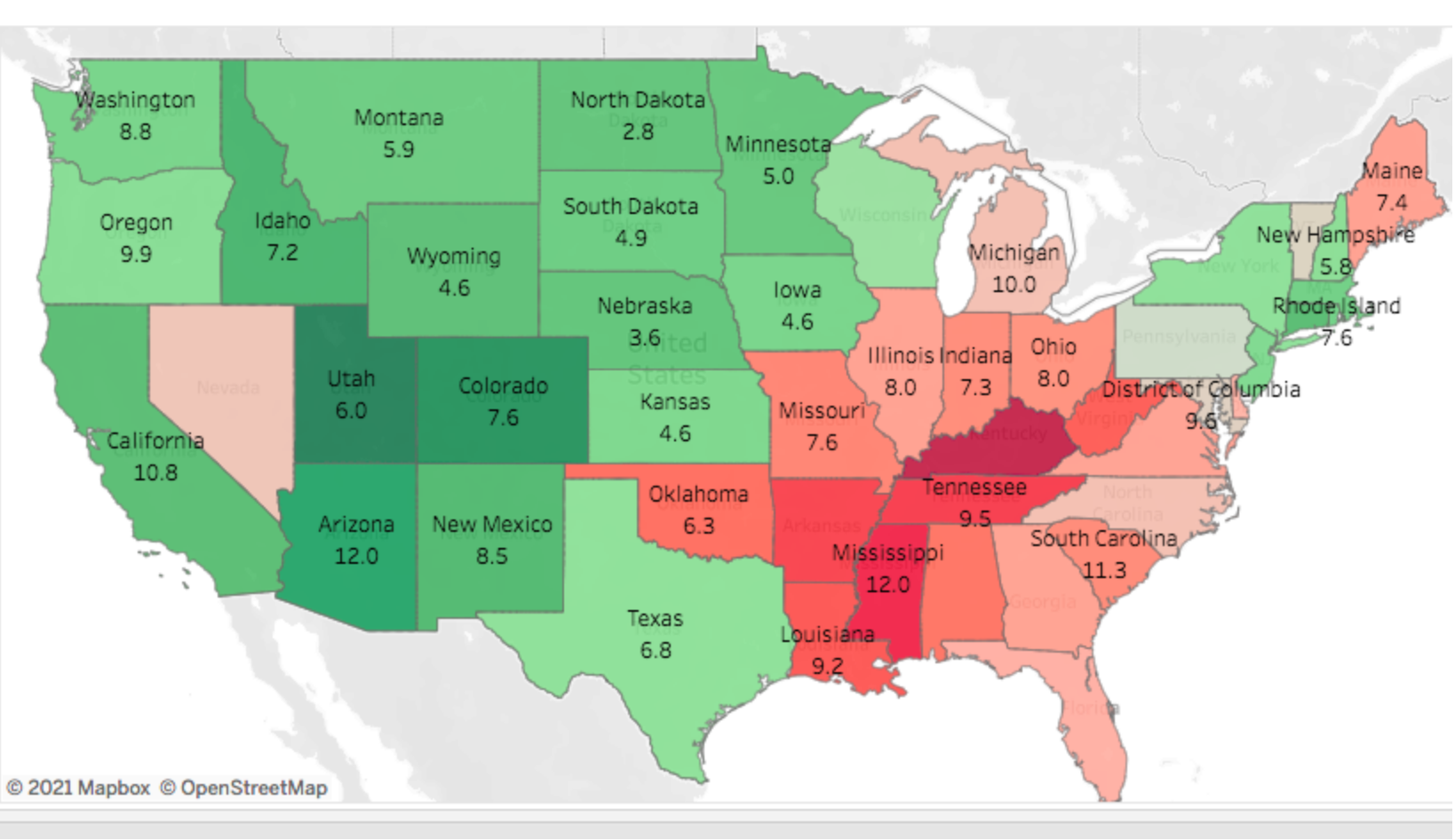
Avg death rate
135.75 215.32

Incident Rate and Death Rate - US States Positive Correlation



The US map above shows the Avg Incidence Rate(marked on the plot within the states) vs Avg Death Rate correlation, it can be seen that generally wherever the AvgIncident rate is high, the Avg Death Rate is high. For eg - Mississippi and Tenesse with one of the highest death rates have high avg incident rates (467 and 471 respectively) while Utah, with very Low death rate also has low Incident Rate 397

Percent Unemployed Over 16 and Death Rate - US States Positive Correlation



The US map above shows the Avg percent of Unemployed over 16(written in the plot on the states) in the state vs the Avg Death Rate correlation. It is seen that wherever the the Avg Unemployed percentage is higher, the Death Rate is higher. For eg , Mississipi and Tenesse with the highest death rates, have very high avg percent of unemployed over 16, 12% and 9.2% respectively while Utah with the lowest death rate also has one of the lowest avg unemployment percentage 6%

Deep Dive into Features

- 'studyPerCap' - Per capita number of cancer-related clinical trials per county. As studyPerCap increases by 1 unit, the Target death rate decreases by 0.014. So more the cancer related clinical trials ,lesser seems to be the death rate(marginal though), which is expected as more chances of being treated early
- 'incidenceRate' - Mean per capita (100,000) cancer diagoses.As incidenceRate increases by 1 unit, the Target death rate increases by 0.20. So more the incidencerate, death rate increases, which is expected as more chances of death with increasing number of diagnosis.
- 'PctBachDeg25_Over' - Percent of county residents ages 25 and over highest education attained: bachelor's degree.As PctBachDeg25_Over increases by 1 unit, the Target death rate decreases by 1.8 units. So more the percent of people with bachelors degree over 25 years of age ,lesser seems to be the death rate, which is expected as they would have better jobs or incomes to support the treatment.They would be more aware of the treatment options.
- 'MedianAgeMale'- Median age of male county residents.As Median age of males in a county increase by 1 unit, death rate seems to decrease by 0.54 units. Could be due to the reason that they get more time to be diagnosed and treated. This feature does not contribute much to the prediction model though.
- 'PctHS18_24 '- Percent of county residents ages 18-24 highest education attained: high school diploma . As the percentage of county residents within ages 18-24 with highest education as high School Diploma increases by 1 unit,the death rate increases by 0.39 units.Again hints at education being important factor in the death rate.
- 'PctUnemployed16_Over' - Percent of county residents ages 16 and over unemployed. As the percentage of unemployed residents in a county increase by 1unit, the death rate increases by more than 1 unit, which is expected as they do not have the income or health coverage plans to support the treatment.
- 'PctOtherRace' - Percent of county residents who identify in a category which is not White, Black, or Asian.As the percent of non Whites/Blacks/Asians increase by 1 unit, the death rate decreases by 0.64, suggesting that other races do not die die of cancer as compared to Whites/Asians/Blacks.
- 'PctMarriedHouseholds' - Percent of married households. As the percent of married households increase by 1 unit,the death rate decreases by 0.53. suggests that married households would have more income and means to tackle with cancer treatment.