# HOMEWORK 1

DENNIS MULUMBI KYALO

2021-10-01

## 1. INTRODUCTION

We will analyze the cancer dataset to find the most valuable descriptors in predicting the Mean per capita (100,000) cancer mortalities (TARGET_deathRate) in the US. The analysis uses the backward stepwise selection method, a comprehensive approach, to develop a multiple linear regression model to predict the TARGET_deathRate.This method begins by applying all the predictors to form a multiple linear regression model. Then it iteratively reduces the predictors by removing the least useful ones, a step at a time. Based on the dimension of the data set, this approach is highly suitable since the number of samples (rows) is larger than the number of variables (columns), as evident in the summary table below.

We begin by loading the dataset and skimming through it to have a better understanding of the variables.

Table 1: Data summary

| | |
|---|---|
| Name | cancer_data |
| Number of rows | 3047 |
| Number of columns | 34 |
| Column type frequency: | |
| character | 2 |
| numeric | 32 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| binnedInc | 0 | 1 | 14 | 16 | 0 | 10 | 0 |
| Geography | 0 | 1 | 16 | 42 | 0 | 3047 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| avgAnnCount | 0 | 1.00 | 606.34 | 1416.36 | 6.00 | 76.00 | 171.00 | 518.00 | 38150.00 |
| avgDeathsPerYear | 0 | 1.00 | 185.97 | 504.13 | 3.00 | 28.00 | 61.00 | 149.00 | 14010.00 |
| TARGET_deathRate | 0 | 1.00 | 178.66 | 27.75 | 59.70 | 161.20 | 178.10 | 195.20 | 362.80 |
| incidenceRate | 0 | 1.00 | 448.27 | 54.56 | 201.30 | 420.30 | 453.55 | 480.85 | 1206.90 |

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| medIncome | 0 | 1.00 | 47063.28 | 12040.09 | 22640.00 | 38882.50 | 45207.00 | 52492.00 | 125635.00 |
| popEst2015 | 0 | 1.00 | 102637.37 | 329059.22 | 827.00 | 11684.00 | 26643.00 | 68671.00 | 10170292.00 |
| povertyPercent | 0 | 1.00 | 16.88 | 6.41 | 3.20 | 12.15 | 15.90 | 20.40 | 47.40 |
| studyPerCap | 0 | 1.00 | 155.40 | 529.63 | 0.00 | 0.00 | 0.00 | 83.65 | 9762.31 |
| MedianAge | 0 | 1.00 | 45.27 | 45.30 | 22.30 | 37.70 | 41.00 | 44.00 | 624.00 |
| MedianAgeMale | 0 | 1.00 | 39.57 | 5.23 | 22.40 | 36.35 | 39.60 | 42.50 | 64.70 |
| MedianAgeFemale | 0 | 1.00 | 42.15 | 5.29 | 22.30 | 39.10 | 42.40 | 45.30 | 65.70 |
| AvgHouseholdSize | 0 | 1.00 | 2.48 | 0.43 | 0.02 | 2.37 | 2.50 | 2.63 | 3.97 |
| PercentMarried | 0 | 1.00 | 51.77 | 6.90 | 23.10 | 47.75 | 52.40 | 56.40 | 72.50 |
| PctNoHS18_24 | 0 | 1.00 | 18.22 | 8.09 | 0.00 | 12.80 | 17.10 | 22.70 | 64.10 |
| PctHS18_24 | 0 | 1.00 | 35.00 | 9.07 | 0.00 | 29.20 | 34.70 | 40.70 | 72.50 |
| PctSomeCol18_24 | 2285 | 0.25 | 40.98 | 11.12 | 7.10 | 34.00 | 40.40 | 46.40 | 79.00 |
| PctBachDeg18_24 | 0 | 1.00 | 6.16 | 4.53 | 0.00 | 3.10 | 5.40 | 8.20 | 51.80 |
| PctHS25_Over | 0 | 1.00 | 34.80 | 7.03 | 7.50 | 30.40 | 35.30 | 39.65 | 54.80 |
| PctBachDeg25_Over | 0 | 1.00 | 13.28 | 5.39 | 2.50 | 9.40 | 12.30 | 16.10 | 42.20 |
| PctEmployed16_Over | 152 | 0.95 | 54.15 | 8.32 | 17.60 | 48.60 | 54.50 | 60.30 | 80.10 |
| PctUnemployed16_Over | 0 | 1.00 | 7.85 | 3.45 | 0.40 | 5.50 | 7.60 | 9.70 | 29.40 |
| PctPrivateCoverage | 0 | 1.00 | 64.35 | 10.65 | 22.30 | 57.20 | 65.10 | 72.10 | 92.30 |
| PctPrivateCoverageAlone | 609 | 0.80 | 48.45 | 10.08 | 15.70 | 41.00 | 48.70 | 55.60 | 78.90 |
| PctEmpPrivCoverage | 0 | 1.00 | 41.20 | 9.45 | 13.50 | 34.50 | 41.10 | 47.70 | 70.70 |
| PctPublicCoverage | 0 | 1.00 | 36.25 | 7.84 | 11.20 | 30.90 | 36.30 | 41.55 | 65.10 |
| PctPublicCoverageAlone | 0 | 1.00 | 19.24 | 6.11 | 2.60 | 14.85 | 18.80 | 23.10 | 46.60 |
| PctWhite | 0 | 1.00 | 83.65 | 16.38 | 10.20 | 77.30 | 90.06 | 95.45 | 100.00 |
| PctBlack | 0 | 1.00 | 9.11 | 14.53 | 0.00 | 0.62 | 2.25 | 10.51 | 85.95 |
| PctAsian | 0 | 1.00 | 1.25 | 2.61 | 0.00 | 0.25 | 0.55 | 1.22 | 42.62 |
| PctOtherRace | 0 | 1.00 | 1.98 | 3.52 | 0.00 | 0.30 | 0.83 | 2.18 | 41.93 |
| PctMarriedHouseholds | 0 | 1.00 | 51.24 | 6.57 | 22.99 | 47.76 | 51.67 | 55.40 | 78.08 |
| BirthRate | 0 | 1.00 | 5.64 | 1.99 | 0.00 | 4.52 | 5.38 | 6.49 | 21.33 |

The dataset comprises 3047 samples and 34 variables. We, therefore, decided to remove the PctSomeCol18_24 (Percent of county residents ages 18-24 highest education attained: some college) variable as 75% of its data is missing.

## 2. ANALYSIS AND RESULTS

We now apply the backward stepwise selection method to find the best variables that will help us develop the most suitable multiple linear regression model.

Table 4: Best Number of Variables

| | nvmax |
|---|---|
| 19 | 19 |

The algorithm chooses nineteen variables as the most suitable number of descriptors that best fit the model.

Table 5: Number of Variables Used and Metrics

| nvmax | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|
| 1 | 24.24921 | 0.2384889 | 18.30278 | 1.145893 | 0.0465072 | 0.7753477 |
| 2 | 21.10370 | 0.4203313 | 15.98932 | 1.061179 | 0.0701523 | 0.5560704 |
| 3 | 20.44730 | 0.4555658 | 15.46460 | 1.195745 | 0.0667451 | 0.6803303 |
| 4 | 20.13708 | 0.4719312 | 15.11746 | 1.352081 | 0.0699925 | 0.8011677 |
| 5 | 20.05164 | 0.4766205 | 15.01135 | 1.354834 | 0.0709618 | 0.7782178 |
| 6 | 19.95432 | 0.4815938 | 14.88869 | 1.307665 | 0.0697909 | 0.8086007 |
| 7 | 19.80390 | 0.4896107 | 14.75518 | 1.312898 | 0.0687577 | 0.8609882 |
| 8 | 19.79559 | 0.4894933 | 14.75296 | 1.291123 | 0.0676517 | 0.8540133 |
| 9 | 19.78708 | 0.4901420 | 14.73712 | 1.344076 | 0.0707649 | 0.8842798 |
| 10 | 19.70987 | 0.4942940 | 14.66012 | 1.367816 | 0.0719481 | 0.8851551 |
| 11 | 19.73457 | 0.4931045 | 14.66601 | 1.351145 | 0.0713661 | 0.8889487 |
| 12 | 19.67630 | 0.4959279 | 14.61515 | 1.358006 | 0.0705398 | 0.8808894 |
| 13 | 19.58100 | 0.5008022 | 14.52852 | 1.320399 | 0.0693665 | 0.8392938 |
| 14 | 19.61007 | 0.4988743 | 14.53081 | 1.321865 | 0.0681912 | 0.8372212 |
| 15 | 19.57734 | 0.5002854 | 14.48158 | 1.318707 | 0.0681909 | 0.8386580 |
| 16 | 19.57424 | 0.5004611 | 14.49685 | 1.336266 | 0.0689853 | 0.8585954 |
| 17 | 19.59171 | 0.4995104 | 14.50399 | 1.306099 | 0.0687557 | 0.8464782 |
| 18 | 19.54724 | 0.5018940 | 14.46140 | 1.308206 | 0.0676250 | 0.8460585 |
| 19 | 19.54410 | 0.5020028 | 14.45823 | 1.327458 | 0.0686320 | 0.8529369 |
| 20 | 19.54452 | 0.5019477 | 14.44746 | 1.333554 | 0.0694234 | 0.8600773 |
| 21 | 19.56254 | 0.5011202 | 14.46858 | 1.320457 | 0.0688756 | 0.8470675 |
| 22 | 19.58558 | 0.5000049 | 14.47514 | 1.309000 | 0.0690169 | 0.8423427 |
| 23 | 19.60463 | 0.4990313 | 14.48646 | 1.312524 | 0.0693234 | 0.8479446 |
| 24 | 19.60734 | 0.4988603 | 14.48846 | 1.310954 | 0.0692585 | 0.8466045 |
| 25 | 19.60976 | 0.4987246 | 14.49318 | 1.307522 | 0.0691509 | 0.8441854 |
| 26 | 19.61426 | 0.4985073 | 14.49578 | 1.308295 | 0.0691774 | 0.8427704 |
| 27 | 19.61429 | 0.4984799 | 14.49640 | 1.306959 | 0.0691975 | 0.8421445 |
| 28 | 19.61770 | 0.4983150 | 14.49748 | 1.307009 | 0.0691944 | 0.8434438 |
| 29 | 19.61790 | 0.4983052 | 14.49775 | 1.304436 | 0.0690936 | 0.8406189 |
| 30 | 19.61859 | 0.4982729 | 14.49804 | 1.303855 | 0.0690484 | 0.8407818 |
| 31 | 19.61943 | 0.4982276 | 14.49889 | 1.305011 | 0.0691174 | 0.8410893 |

The table above shows the number of variables and the resulting metrics after applying cross-validation on the dataset using k = 10 folds. The application of cross-validation on the dataset computed the cross-validation error for each model using the backward selection method, then selected the model with the least test error. The above table shows that the nineteen chosen variables had the least RMSE of 19.54410 compared to the others.

Table 6: Coefficients of the Best Model

| | x |
|---|---|
| (Intercept) | 132.7706252 |
| avgAnnCount | -0.0034304 |
| avgDeathsPerYear | 0.0184398 |
| incidenceRate | 0.1929741 |
| popEst2015 | -0.0000156 |
| povertyPercent | 0.4984929 |
| MedianAgeMale | -0.4896163 |
| PercentMarried | 0.8913298 |
| PctNoHS18_24 | -0.1169021 |
| PctHS18_24 | 0.2581844 |
| PctHS25_Over | 0.3600992 |
| PctBachDeg25_Over | -1.2045271 |
| PctEmployed16_Over | -0.0549043 |
| PctUnemployed16_Over | 0.3848801 |
| PctPrivateCoverage | -0.5569978 |
| PctEmpPrivCoverage | 0.3675523 |
| PctWhite | -0.0967972 |
| PctOtherRace | -0.8556766 |
| PctMarriedHouseholds | -0.8540822 |
| BirthRate | -0.8726283 |

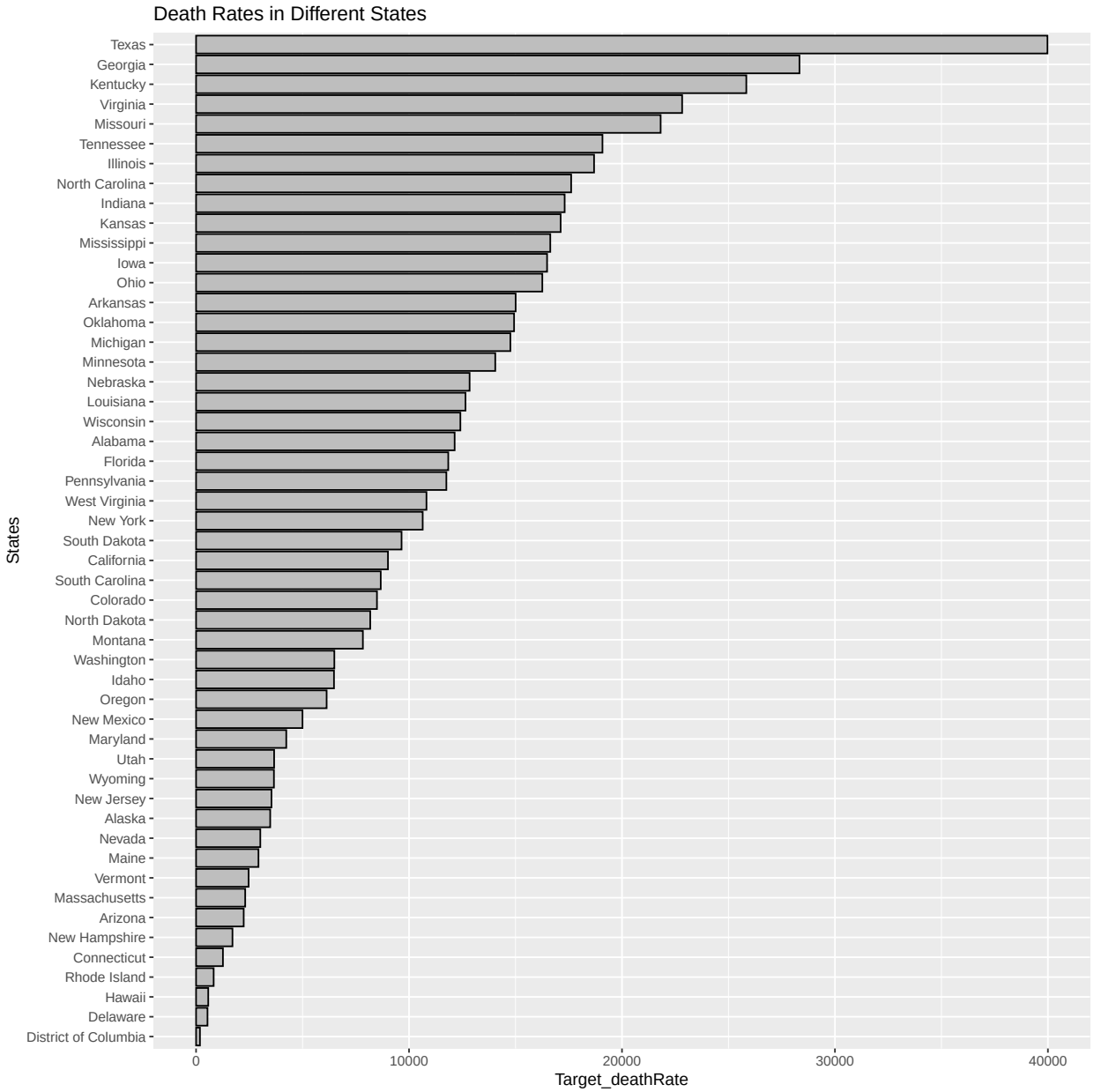
After applying the algorithm to the dataset, the analysis produced the following multiple linear regression model with nineteen significant variables.

*TARGET_deathRate = 132.7706252 - 0.0034304 avgAnnCount + 0.0184398 avgDeathsPerYear + 0.1929741 incidenceRate - 0.0000156 popEst2015 + 0.4984929 povertyPercent - 0.4896163 MedianAgeMale + 0.8913298 PercentMarried - 0.1169021 PctNoHS18_24 + 0.2581844 PctHS18_24 + 0.3600992 PctHS25_Over - 1.2045271 PctBachDeg25_Over - 0.0549043 PctEmployed16_Over + 0.3848801 PctUnemployed16_Over - 0.5569978 PctPrivateCoverage + 0.3675523 PctEmpPrivCoverage - 0.0967972 PctWhite - 0.8556766 PctOtherRace - 0.8540822 PctMarriedHouseholds - 0.8726283 BirthRate*

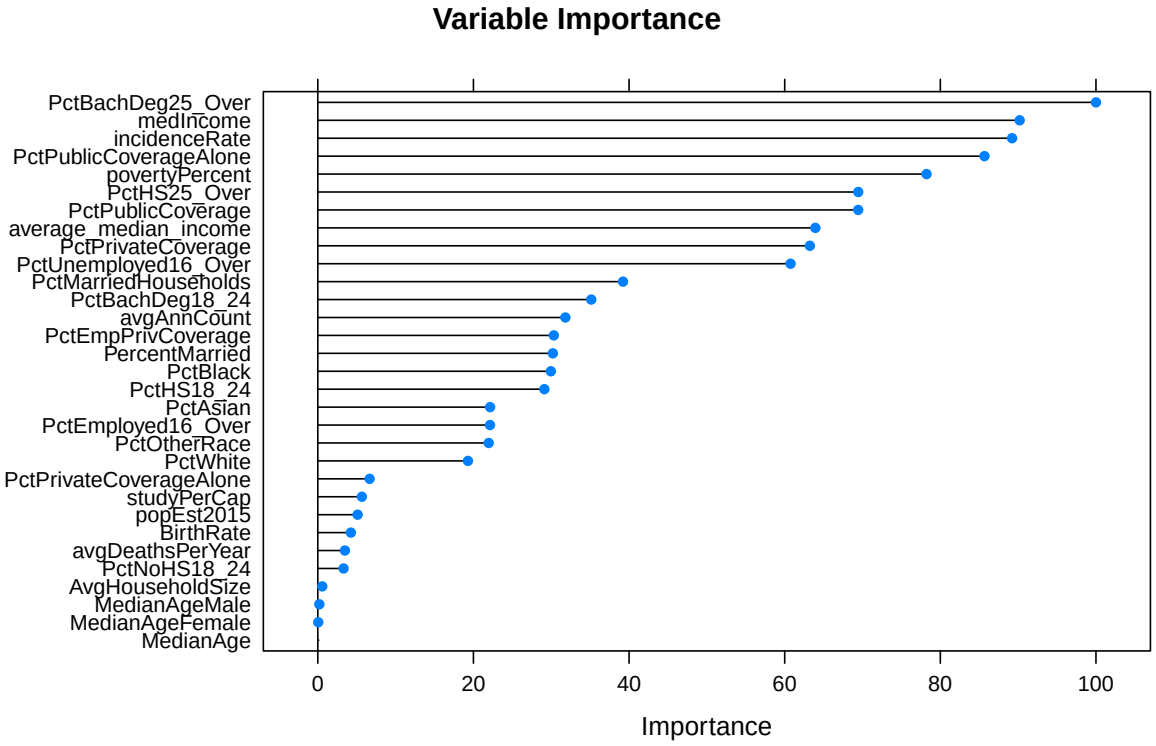
The model selected the following descriptors as they were the most significant descriptors from the rest. These descriptors have a high impact on accurately predicting the TARGET_deathRate of the dataset.

*avgAnnCount: Mean number of reported cases of cancer diagnosed annually*
*avgDeathsPerYear: Mean number of reported mortalities due to cancer*
*incidenceRate: Mean per capita (100,000) cancer diagoses*
*popEst2015: Population of county*
*povertyPercent: Percent of populace in poverty*
*MedianAgeMale: Median age of male county residents*
*PercentMarried: Percent of county residents who are married*
*PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school*
*PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma*
*PctHS25_Over: Percent of county residents ages 25 and over highest education attained: high school diploma PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree PctEmployed16_Over: Percent of county residents ages 16 and over employed*
*PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed PctPrivateCoverage: Percent of county residents with private health coverage*
*PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage PctWhite: Percent of county residents who identify as White PctOtherRace: Pct of county residents who identify in a category which is'nt White, Black, or Asian*

*PctMarriedHouseholds: Percent of married households*
*BirthRate: Number of live births relative to number of women in county*

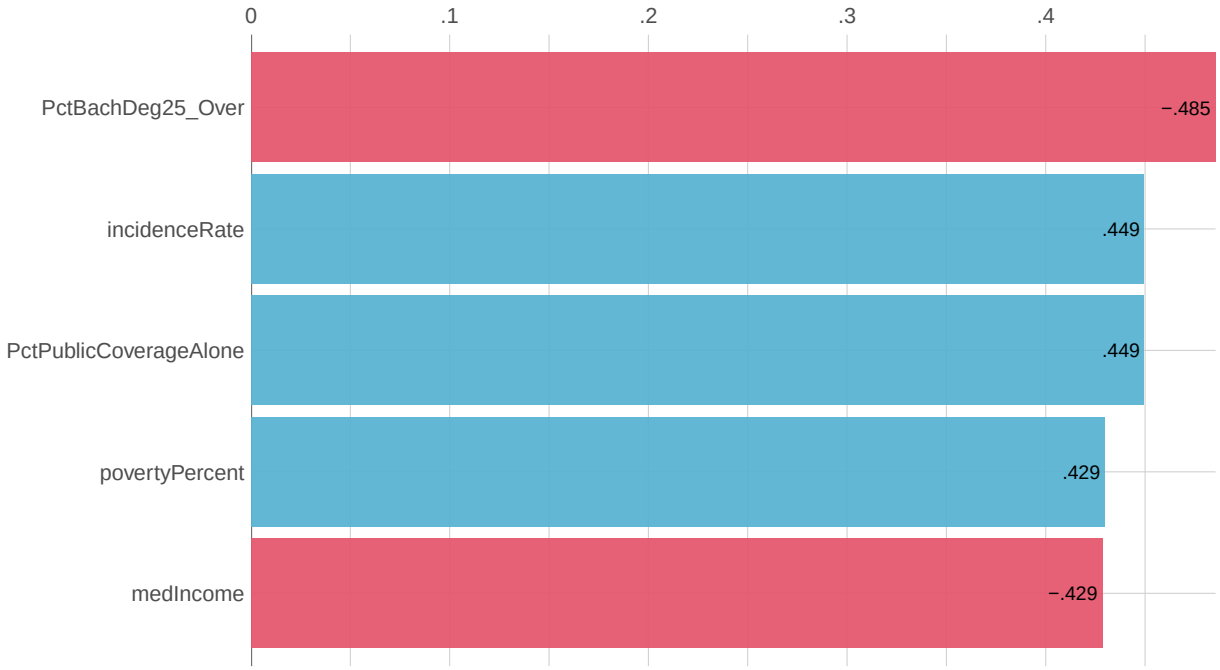



From the above graph, Texas had the highest Mean per capita (100,000) cancer mortalities (TARGET_deathRate), while the District of Columbia had the least Mean per capita (100,000) cancer mortalities (TARGET_deathRate).




## Correlations of TARGET_deathRate

*Top 5 out of 54 variables (original & dummy)*



# 3. CONCLUSION

Cancer is among the deadliest diseases in the US, killing so many people annually based on different factors. Our analysis develops a solid model that helps us determine the most crucial factors that lead to cancer deaths in the US. From the study, we have created a multiple linear regression model comprising the best and most effective nineteen descriptors that help us predict the Mean per capita (100,000) cancer mortalities (TARGET_deathRate). Among the chosen variables, the five most essential descriptors in the building of the model include *Percent of county residents ages 25 and over highest education attained: bachelor's degree (PctBachDeg25_Over), Median income per county (medianIncome), Mean per capita (100,000) cancer diagnoses (incidenceRate), Percent of county residents with government-provided health coverage alone (PctPubliceCoverageAlone), and Percent of populace in poverty (povertyPercent)* as evidenced in the variable importance graph above. It is also evident that the same variables tend to have a high correlation to the TARGET_deathRate as shown in the correlation graph above. The *Percent of county residents ages 25 and over highest education attained: bachelor's degree (PctBachDeg25_Over) variable and the Median income per county (medianIncome)* variable have a negative correlation of -0.485 and -0.429 respectively to the TARGET_deathRate variable. Meanwhile, the *Mean per capita (100,000) cancer diagnoses (incidenceRate), Percent of county residents with government-provided health coverage alone (PctPubliceCoverageAlone), and Percent of populace in poverty (povertyPercent)* variable have a positive correlation of 0.449, 0.449, and 0.429, respectively to the TARGET_deathRate.

# 4. REFERENCE

An Introduction to Statistical Learning 1st edition, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, New York.