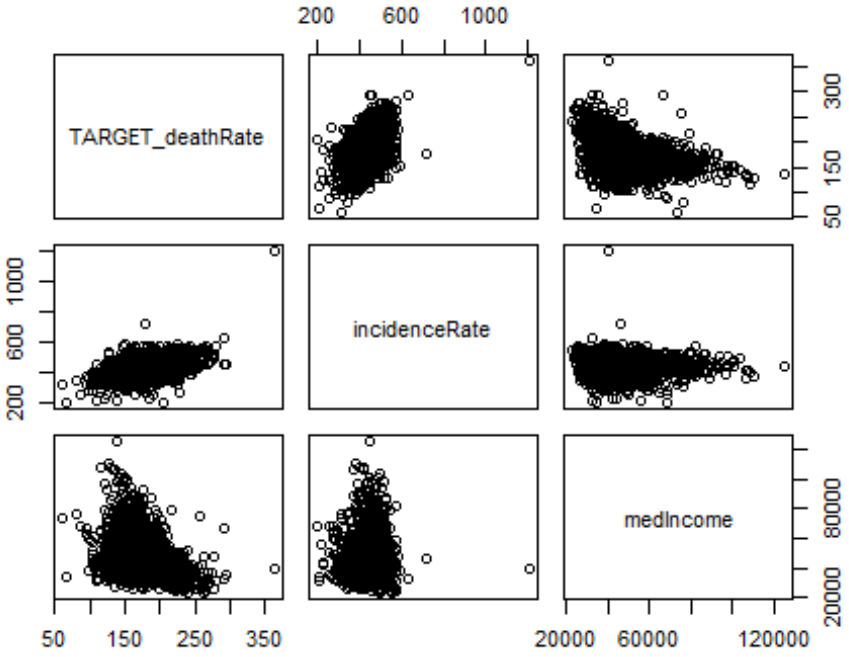
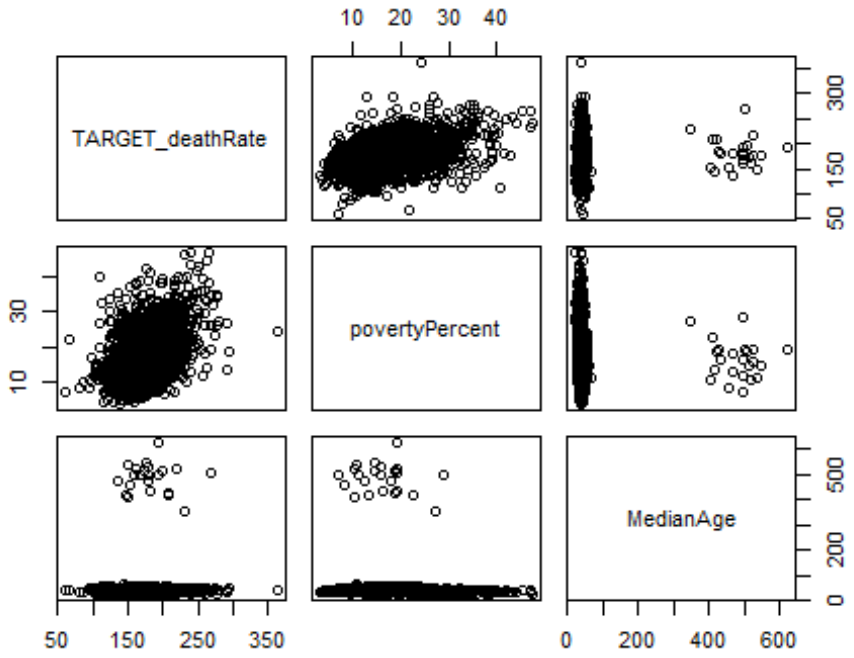
1. Exploratory Data analysis

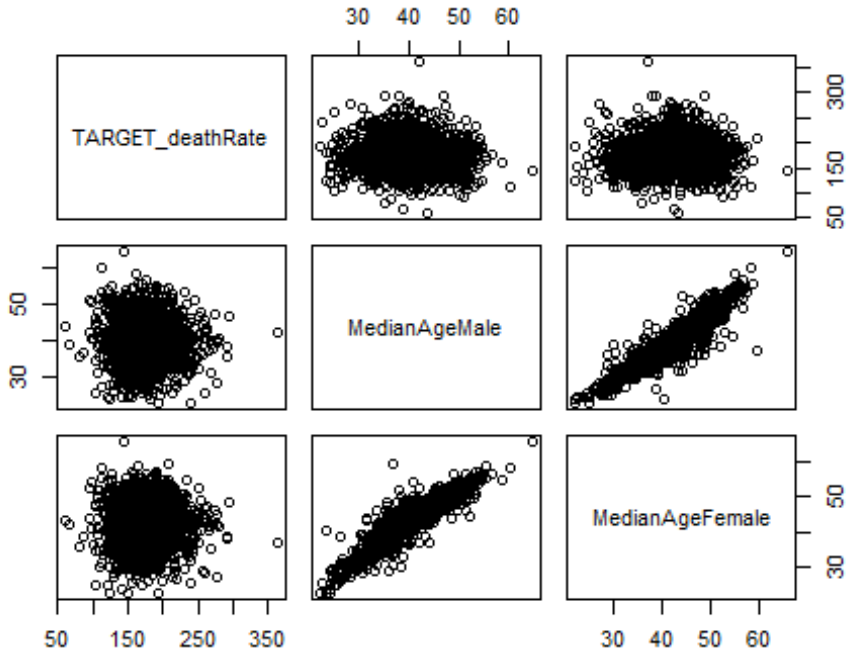
```
Train.data=read.csv("CancerData.csv",header=T)
Test.data=read.csv("CancerHoldoutData.csv",header=T)
library(ISLR)
library(car)

library(class)
library(FNN)

attach(Train.data)

pairs(Train.data[,c(1,2,3)])
```



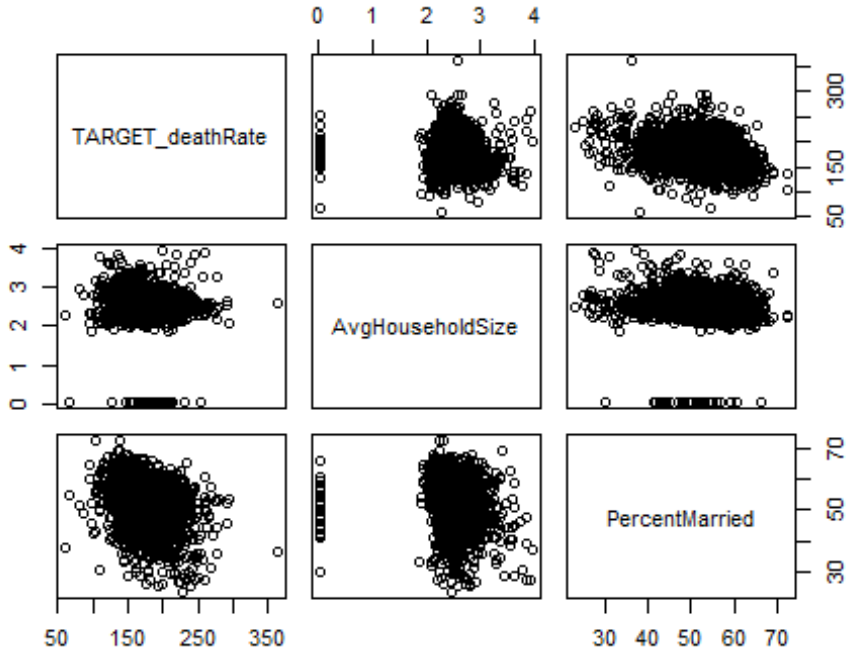

```
pairs(Train.data[,c(1,4,5)])
```



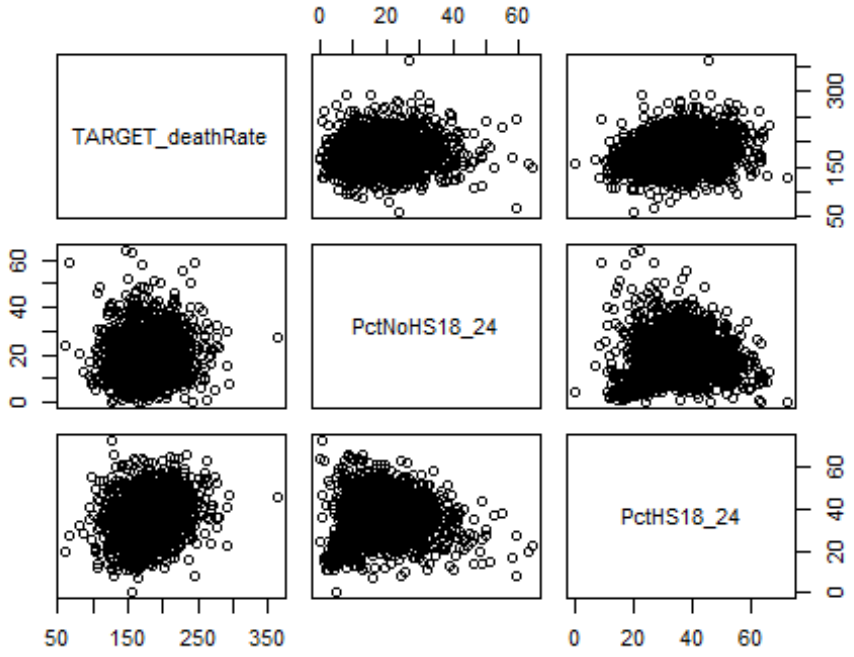

```
pairs(Train.data[,c(1,6,7)])
```



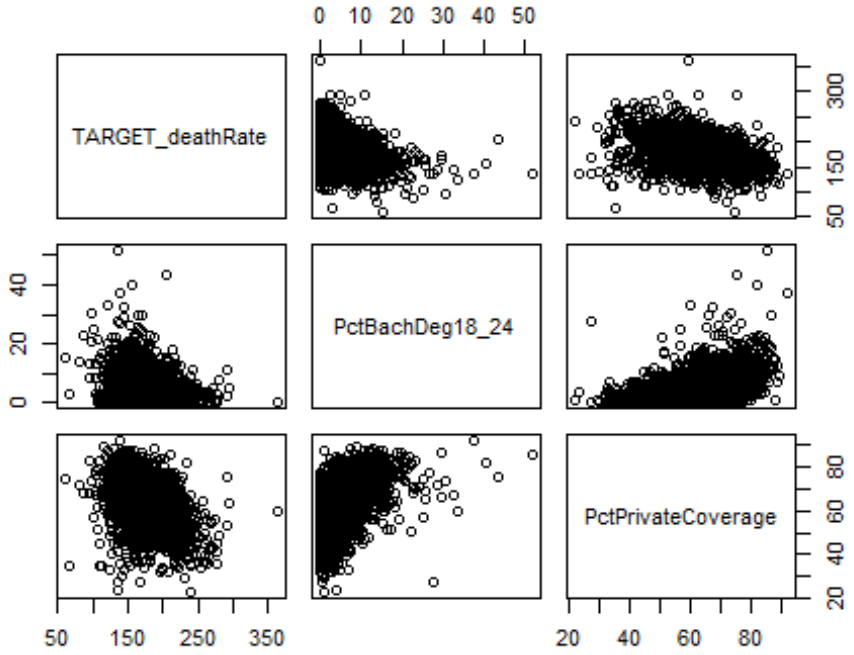

```
pairs(Train.data[,c(1,9,10)])
```



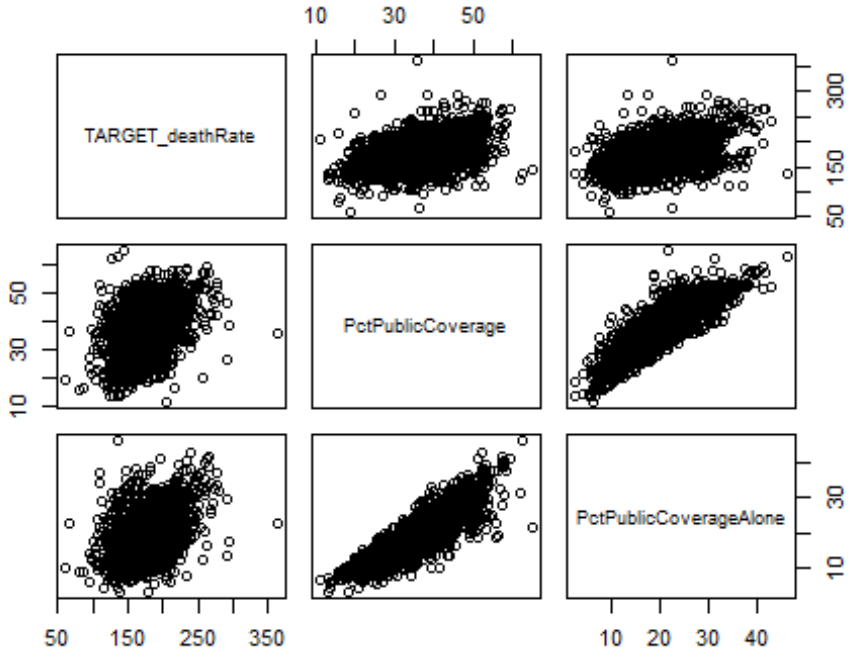

```
pairs(Train.data[,c(1,11,12)])
```



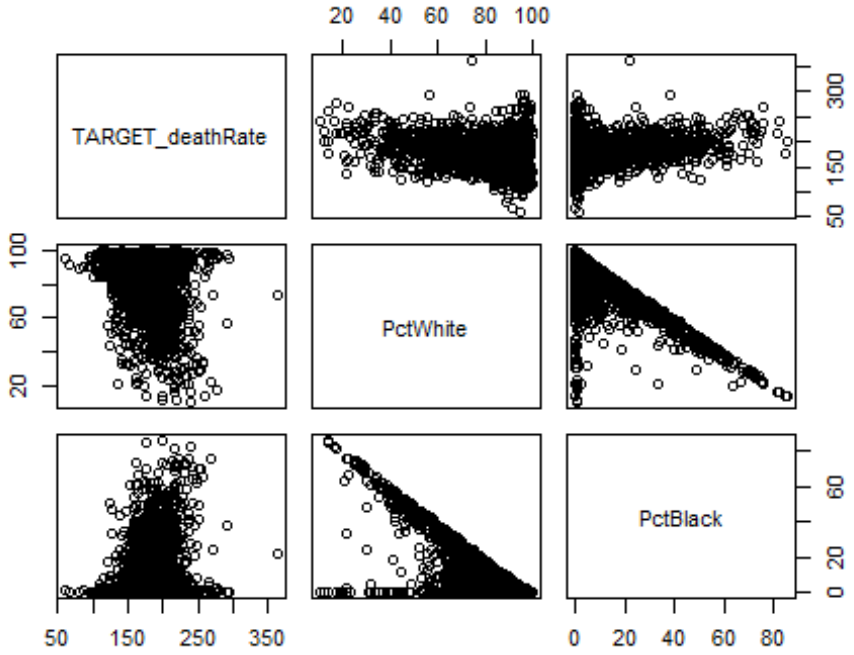

```
pairs(Train.data[,c(1,13,14)])
```



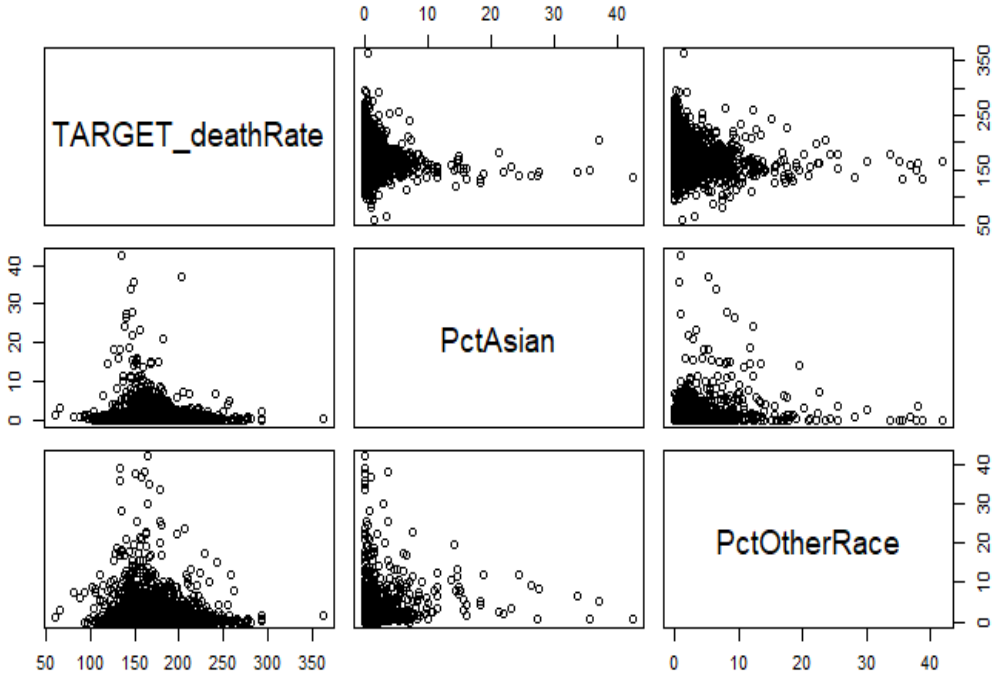

```
pairs(Train.data[,c(1,15,16)])
```




```
pairs(Train.data[,c(1,17,18)])
```




```
pairs(Train.data[,c(1,19,20)])
```




```
pairs(Train.data[,c(1,21)])
```

Explanation:

From the graphs it can be seen that there is some pattern in Target Death rate when it is plotted against these predictors: incidenceRate, medIncome, povertyPercent, PercentMarried, PctBachDeg18_24, PctPrivateCoverage, PctPublicCoverage, PctPublicCoverageAlone, PctMarriedHouseholds.

The general trend can be described as below:

Trend	**Predictors**
Linear increase	incidenceRate, povertyPercent, PctPublicCoverage, PctPublicCoverageAlone,
Linear decrease	PercentMarried, PctBachDeg18_24, PctPrivateCoverage
Non-linear decrease	medIncome, PctMarriedHouseholds

Therefore, these predictors may be significant predictors from prelimnary analysis but it is not be possible to determine their significance without linear regression.

Outlier Analysis

```
summary(Train.data)
```

```
##  TARGET_deathRate incidenceRate      medIncome      povertyPercent
##  Min.   : 59.7    Min.   : 201.3   Min.   : 22640   Min.   : 3.20
##  1st Qu.:161.0    1st Qu.: 420.2   1st Qu.: 38842   1st Qu.:12.10
##  Median :178.2    Median : 453.5   Median : 45224   Median :15.80
##  Mean   :178.6    Mean   : 447.9   Mean   : 47188   Mean   :16.85
##  3rd Qu.:195.2    3rd Qu.: 480.8   3rd Qu.: 52702   3rd Qu.:20.40
##  Max.   :362.8    Max.   :1206.9   Max.   :125635   Max.   :47.40
##
##    MedianAge      MedianAgeMale   MedianAgeFemale
##  Min.   : 22.30   Min.   :22.40   Min.   :22.30
##  1st Qu.: 37.80   1st Qu.:36.40   1st Qu.:39.10
##  Median : 41.00   Median :39.50   Median :42.40
##  Mean   : 44.73   Mean   :39.56   Mean   :42.15
##  3rd Qu.: 44.00   3rd Qu.:42.60   3rd Qu.:45.38
##  Max.   :624.00   Max.   :64.70   Max.   :65.70
##
##                              Geography    AvgHouseholdSize PercentMarried
##  Abbeville County, South Carolina:   1   Min.   :0.0222   Min.   :23.10
##  Acadia Parish, Louisiana        :   1   1st Qu.:2.3700   1st Qu.:47.90
##  Accomack County, Virginia       :   1   Median :2.5000   Median :52.40
##  Ada County, Idaho               :   1   Mean   :2.4896   Mean   :51.78
##  Adair County, Kentucky          :   1   3rd Qu.:2.6400   3rd Qu.:56.30
##  Adair County, Missouri          :   1   Max.   :3.9700   Max.   :72.50
##  (Other)                         :2584
##   PctNoHS18_24     PctHS18_24    PctSomeCol18_24 PctBachDeg18_24
##  Min.   : 0.00   Min.   : 0.00   Min.   : 7.10   Min.   : 0.00
##  1st Qu.:12.70   1st Qu.:29.20   1st Qu.:33.60   1st Qu.: 3.10
##  Median :17.10   Median :34.80   Median :40.30   Median : 5.30
##  Mean   :18.24   Mean   :34.96   Mean   :40.87   Mean   : 6.16
##  3rd Qu.:22.80   3rd Qu.:40.67   3rd Qu.:46.20   3rd Qu.: 8.20
##  Max.   :64.10   Max.   :72.50   Max.   :79.00   Max.   :51.80
##                                  NA's   :1938
##  PctPrivateCoverage PctPublicCoverage PctPublicCoverageAlone    PctWhite
##  Min.   :22.30      Min.   :11.20     Min.   : 2.60          Min.   : 10.2
0
##  1st Qu.:57.40      1st Qu.:30.80     1st Qu.:14.90          1st Qu.: 77.0
4
##  Median :65.20      Median :36.20     Median :18.70          Median : 89.9
9
##  Mean   :64.42      Mean   :36.18     Mean   :19.19          Mean   : 83.5
7
##  3rd Qu.:72.20      3rd Qu.:41.50     3rd Qu.:23.00          3rd Qu.: 95.3
6
##  Max.   :92.30      Max.   :65.10     Max.   :46.60          Max.   :100.0
0
##
##     PctBlack          PctAsian        PctOtherRace     PctMarriedHousehold
s
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :22.99
##  1st Qu.: 0.6321   1st Qu.: 0.2556   1st Qu.: 0.2899   1st Qu.:47.83
```

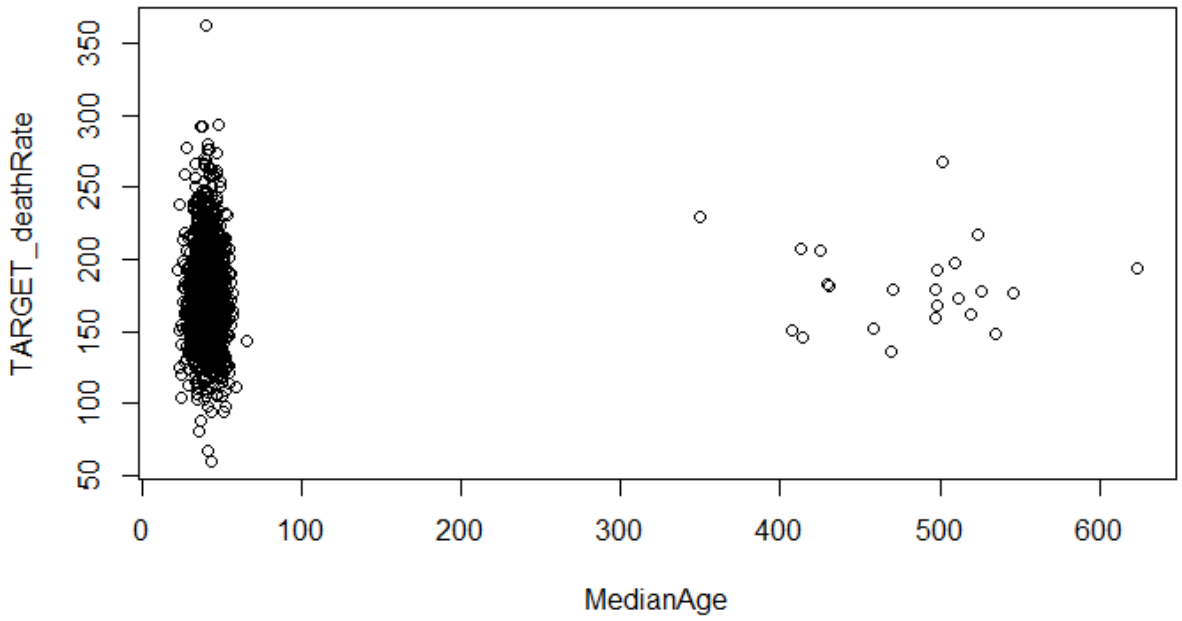
```
##  Median : 2.2692   Median : 0.5507   Median : 0.8330   Median :51.71
##  Mean   : 9.0979   Mean   : 1.2690   Mean   : 2.0311   Mean   :51.25
##  3rd Qu.:10.3528   3rd Qu.: 1.2230   3rd Qu.: 2.1823   3rd Qu.:55.33
##  Max.   :85.9478   Max.   :42.6194   Max.   :41.9303   Max.   :78.08
##
```

```
summary(Test.data)
```

```
##  TARGET_deathRate incidenceRate      medIncome      povertyPercent
##  Min.   :106.1    Min.   : 254.7   Min.   : 25807   Min.   : 3.90
##  1st Qu.:162.0    1st Qu.: 421.8   1st Qu.: 39017   1st Qu.:12.40
##  Median :177.9    Median : 453.5   Median : 45168   Median :16.30
##  Mean   :179.1    Mean   : 450.3   Mean   : 46358   Mean   :17.02
##  3rd Qu.:195.2    3rd Qu.: 482.4   3rd Qu.: 51911   3rd Qu.:20.60
##  Max.   :270.4    Max.   :1014.2   Max.   :122641   Max.   :40.60
##
##    MedianAge      MedianAgeMale   MedianAgeFemale
##  Min.   : 23.30   Min.   :23.00   Min.   :23.60
##  1st Qu.: 37.60   1st Qu.:36.30   1st Qu.:39.20
##  Median : 40.90   Median :39.70   Median :42.30
##  Mean   : 48.34   Mean   :39.62   Mean   :42.13
##  3rd Qu.: 44.00   3rd Qu.:42.30   3rd Qu.:45.20
##  Max.   :619.20   Max.   :58.60   Max.   :58.00
##
##                        Geography   AvgHouseholdSize PercentMarried
##  Adair County, Iowa        :  1   Min.   :0.0221   Min.   :26.20
##  Adair County, Oklahoma    :  1   1st Qu.:2.3600   1st Qu.:47.40
##  Adams County, Colorado    :  1   Median :2.4900   Median :52.60
##  Adams County, Indiana     :  1   Mean   :2.4235   Mean   :51.77
##  Adams County, Mississippi :  1   3rd Qu.:2.6200   3rd Qu.:56.70
##  Adams County, Pennsylvania:  1   Max.   :3.9700   Max.   :68.00
##  (Other)                   :451
##   PctNoHS18_24     PctHS18_24    PctSomeCol18_24 PctBachDeg18_24
##  Min.   : 1.50   Min.   :10.00   Min.   : 9.60   Min.   : 0.000
##  1st Qu.:13.30   1st Qu.:29.20   1st Qu.:34.73   1st Qu.: 3.200
##  Median :17.40   Median :34.50   Median :41.25   Median : 5.600
##  Mean   :18.13   Mean   :35.25   Mean   :41.58   Mean   : 6.148
##  3rd Qu.:22.00   3rd Qu.:40.80   3rd Qu.:47.25   3rd Qu.: 8.100
##  Max.   :59.70   Max.   :72.10   Max.   :78.30   Max.   :28.500
##                                  NA's   :347
##  PctPrivateCoverage PctPublicCoverage PctPublicCoverageAlone    PctWhite
##  Min.   :25.0       Min.   :11.80     Min.   : 4.60          Min.   :11.01
##  1st Qu.:56.6       1st Qu.:31.40     1st Qu.:14.80          1st Qu.:78.32
##  Median :64.0       Median :37.00     Median :19.60          Median :90.32
##  Mean   :64.0       Mean   :36.66     Mean   :19.55          Mean   :84.05
##  3rd Qu.:71.7       3rd Qu.:41.80     3rd Qu.:23.60          3rd Qu.:95.66
##  Max.   :86.9       Max.   :57.50     Max.   :39.70          Max.   :99.69
##
##     PctBlack          PctAsian        PctOtherRace     PctMarriedHousehold
s
```

```
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :24.43
##  1st Qu.: 0.5946   1st Qu.: 0.2419   1st Qu.: 0.3345   1st Qu.:47.10
##  Median : 2.2221   Median : 0.5377   Median : 0.7860   Median :51.45
##  Mean   : 9.1652   Mean   : 1.1689   Mean   : 1.7138   Mean   :51.19
##  3rd Qu.:10.7674   3rd Qu.: 1.1962   3rd Qu.: 2.0944   3rd Qu.:55.75
##  Max.   :80.6600   Max.   :33.7609   Max.   :22.4644   Max.   :68.28
```

```
plot(MedianAge,TARGET_deathRate)
```



Explanation:

- **Geography data:** It is seen that are a lot of counties in the Geography column of the dataset. It would not be possible to make a linear regression model with all the county data, without making it too complex. Therefore, the county data is converted into state data. The code for the same is shown below:

```
#Converting geography county data into state data to better interpret results
Train.data$Geography=sub(".*, ", "", Train.data$Geography)
Train.data$Geography=as.factor(Train.data$Geography)

Test.data$Geography=sub(".*, ", "", Test.data$Geography)
Test.data$Geography=as.factor(Test.data$Geography)

str(Train.data$Geography)

##  Factor w/ 51 levels "Alabama","Alaska",..: 48 48 48 48 48 48 48 48 48 48
...
```

Here, sub function is used to search for string data in the Geography column and the string after “, ” in county name is extracted and replaced in the column. This approach works since all the data is in the form of “County name, State”. This is easier to incorporate into linear regression,

since it significantly reduces the factors from 2590 county names to 51 states. After correcting this, a simple linear regression model is fit, as shown below:

Code:

```
linear.fit=lm(TARGET_deathRate~.,data=Train.data)
summary(linear.fit)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ ., data = Train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.218  -9.810  -0.705   9.843 108.870
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.926e+03  1.388e+03   1.388 0.165656
## incidenceRate             1.938e-01  2.025e-02   9.570  < 2e-16 ***
## medIncome                -2.670e-04  1.903e-04  -1.403 0.161230
## povertyPercent           -1.986e-01  3.493e-01  -0.568 0.569932
## MedianAge                -1.227e-02  1.681e-02  -0.730 0.465710
## MedianAgeMale            -1.595e-01  4.375e-01  -0.365 0.715543
## MedianAgeFemale          -1.775e-01  4.662e-01  -0.381 0.703450
## GeographyAlaska          -1.289e+01  1.145e+01  -1.126 0.260718
## GeographyArizona         -2.611e+01  1.193e+01  -2.189 0.028989 *
## GeographyArkansas         1.795e+00  7.372e+00   0.244 0.807683
## GeographyCalifornia      -1.870e+01  9.135e+00  -2.047 0.041072 *
## GeographyColorado        -2.879e+01  8.103e+00  -3.553 0.000412 ***
## GeographyConnecticut     -1.921e+01  2.016e+01  -0.953 0.341203
## GeographyDelaware        -1.970e+01  1.487e+01  -1.325 0.185604
## GeographyFlorida         -9.587e+00  8.452e+00  -1.134 0.257155
## GeographyGeorgia         -6.339e+00  6.911e+00  -0.917 0.359393
## GeographyIdaho           -3.825e+01  8.221e+00  -4.652 4.06e-06 ***
## GeographyIllinois        -1.203e+01  8.184e+00  -1.470 0.142138
## GeographyIndiana          1.742e+00  7.456e+00   0.234 0.815405
## GeographyIowa            -2.089e+01  7.542e+00  -2.770 0.005781 **
## GeographyKansas          -1.272e+01  7.436e+00  -1.710 0.087747 .
## GeographyKentucky         3.424e+00  7.160e+00   0.478 0.632677
## GeographyLouisiana        5.998e+00  7.959e+00   0.754 0.451394
## GeographyMaine           -1.105e+01  1.276e+01  -0.866 0.386919
## GeographyMaryland        -5.172e+00  8.979e+00  -0.576 0.564830
## GeographyMassachusetts   -2.852e+01  2.038e+01  -1.400 0.162102
## GeographyMichigan        -1.592e+01  8.536e+00  -1.865 0.062661 .
## GeographyMinnesota       -2.339e+01  7.354e+00  -3.181 0.001546 **
## GeographyMississippi     -1.173e-01  7.855e+00  -0.015 0.988090
## GeographyMissouri         1.590e+00  7.246e+00   0.219 0.826412
## GeographyMontana         -2.185e+01  1.030e+01  -2.121 0.034345 *
## GeographyNebraska        -1.615e+01  8.883e+00  -1.818 0.069507 .
```

```
## GeographyNevada           7.220e+00  1.096e+01   0.659 0.510341
## GeographyNew Hampshire   -1.477e+01  1.075e+01  -1.374 0.169890
## GeographyNew Jersey      -7.368e+00  9.616e+00  -0.766 0.443851
## GeographyNew Mexico      -2.558e+01  1.085e+01  -2.358 0.018682 *
## GeographyNew York        -1.827e+01  8.648e+00  -2.112 0.035077 *
## GeographyNorth Carolina  -1.124e+01  7.145e+00  -1.573 0.116334
## GeographyNorth Dakota    -1.867e+01  8.696e+00  -2.147 0.032210 *
## GeographyOhio            -9.017e-02  7.662e+00  -0.012 0.990614
## GeographyOklahoma        -6.750e+00  7.897e+00  -0.855 0.393024
## GeographyOregon          -1.694e+01  9.257e+00  -1.830 0.067688 .
## GeographyPennsylvania    -1.905e+01  7.821e+00  -2.436 0.015138 *
## GeographySouth Carolina  -2.738e+00  9.705e+00  -0.282 0.777902
## GeographySouth Dakota    -2.086e+01  8.165e+00  -2.554 0.010894 *
## GeographyTennessee        4.627e+00  7.335e+00   0.631 0.528448
## GeographyTexas           -2.713e+00  6.921e+00  -0.392 0.695145
## GeographyUtah            -3.560e+01  1.157e+01  -3.076 0.002194 **
## GeographyVermont         -2.418e+01  2.021e+01  -1.196 0.232032
## GeographyVirginia         9.010e+00  7.340e+00   1.227 0.220142
## GeographyWashington      -1.570e+01  8.589e+00  -1.828 0.068064 .
## GeographyWest Virginia   -5.328e-01  8.014e+00  -0.066 0.947011
## GeographyWisconsin       -1.999e+01  8.589e+00  -2.327 0.020310 *
## GeographyWyoming         -2.041e+01  1.265e+01  -1.613 0.107291
## AvgHouseholdSize          3.337e+00  2.394e+00   1.394 0.163796
## PercentMarried            2.890e-01  3.480e-01   0.830 0.406698
## PctNoHS18_24             -1.790e+01  1.383e+01  -1.294 0.196321
## PctHS18_24               -1.777e+01  1.385e+01  -1.283 0.199898
## PctSomeCol18_24          -1.799e+01  1.384e+01  -1.300 0.194282
## PctBachDeg18_24          -1.894e+01  1.385e+01  -1.368 0.171868
## PctPrivateCoverage        2.128e-01  3.102e-01   0.686 0.492995
## PctPublicCoverage         3.526e-01  4.751e-01   0.742 0.458217
## PctPublicCoverageAlone    6.757e-01  6.309e-01   1.071 0.284628
## PctWhite                 -1.288e-01  1.552e-01  -0.830 0.407152
## PctBlack                 -2.199e-01  1.690e-01  -1.301 0.193686
## PctAsian                  8.008e-02  4.501e-01   0.178 0.858852
## PctOtherRace             -1.060e+00  3.008e-01  -3.524 0.000458 ***
## PctMarriedHouseholds     -9.375e-01  3.297e-01  -2.844 0.004615 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.81 on 584 degrees of freedom
##   (1938 observations deleted due to missingness)
## Multiple R-squared:  0.5861, Adjusted R-squared:  0.5387
## F-statistic: 12.34 on 67 and 584 DF,  p-value: < 2.2e-16
```

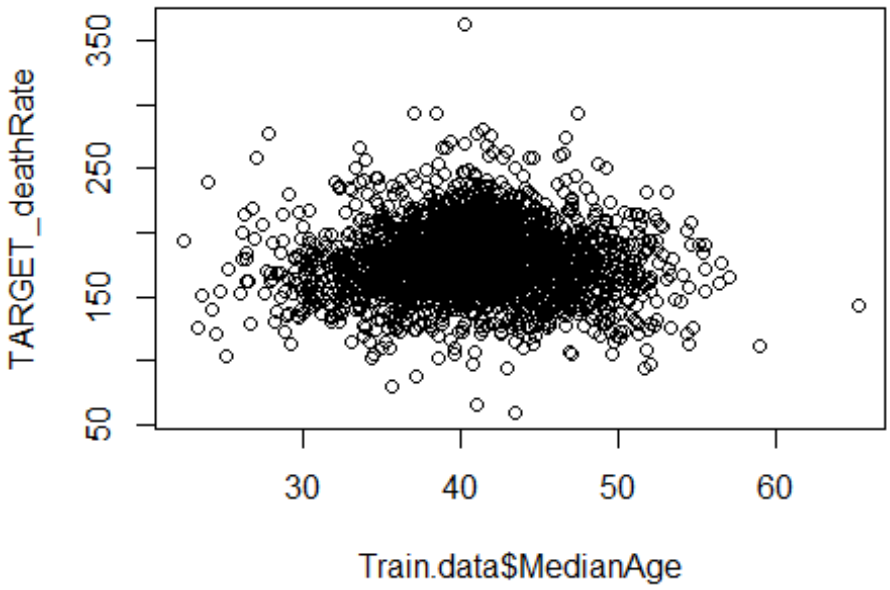
Explanation:

Model	**Adjusted R-squared**	**Residual Std Error**
Preliminary model	0.5387	18.81

- **Age data outliers:** From the summary, it is observed that in the training dataset, the maximum medianAge value is 624, which is not possible. Scatterplot for medianAge also shows that multiple values of medianAge are over 300 years, which cannot be possible. These values need to be corrected. From the dataset, it is also observed that the MedianAge of county is approximately equal to average of MedianMale and MedianFemale ages. Therefore, for counties with MedianAge over 130 years, it is replaced with average value of male and female ages, as shown below:

Code:

```
i=1
for (i in 1:2590)
{
  if (Train.data$MedianAge[i]>130)
  {
    Train.data$MedianAge[i]=(Train.data$MedianAgeMale[i]+Train.data$MedianAge
Female[i])/2
  }
}

# Correcting values in testing data
i=1
for (i in 1:457)
{
  if (Test.data$MedianAge[i]>130)
  {
    Test.data$MedianAge[i]=(Test.data$MedianAgeMale[i]+Test.data$MedianAgeFem
ale[i])/2
  }
}


# Replotting the median age graph for training data:
plot(Train.data$MedianAge,TARGET_deathRate)
```



Linear model was fitted after correcting this outlier error, as shown below:

Code:

```
attach(Train.data)

linear.fit=lm(TARGET_deathRate~.,data=Train.data) #Effect of outliers
summary(linear.fit)

##
## Call:
## lm(formula = TARGET_deathRate ~ ., data = Train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.900  -9.553  -0.663   9.997 109.090
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.960e+03  1.387e+03   1.414 0.158039
## incidenceRate            1.935e-01  2.027e-02   9.544  < 2e-16 ***
## medIncome               -2.695e-04  1.904e-04  -1.415 0.157510
## povertyPercent          -1.764e-01  3.525e-01  -0.500 0.616967
## MedianAge                1.084e+00  2.392e+00   0.453 0.650766
## MedianAgeMale           -7.158e-01  1.334e+00  -0.537 0.591678
## MedianAgeFemale         -6.854e-01  1.178e+00  -0.582 0.560942
## GeographyAlaska         -1.277e+01  1.147e+01  -1.114 0.265683
## GeographyArizona        -2.602e+01  1.194e+01  -2.180 0.029673 *
## GeographyArkansas        1.897e+00  7.381e+00   0.257 0.797304
## GeographyCalifornia     -1.886e+01  9.135e+00  -2.064 0.039455 *
## GeographyColorado       -2.934e+01  8.096e+00  -3.624 0.000316 ***
## GeographyConnecticut    -1.951e+01  2.017e+01  -0.967 0.333869
## GeographyDelaware       -1.973e+01  1.487e+01  -1.326 0.185204
```

```
## GeographyFlorida        -9.814e+00  8.458e+00  -1.160 0.246378
## GeographyGeorgia        -6.570e+00  6.906e+00  -0.951 0.341818
## GeographyIdaho          -3.839e+01  8.227e+00  -4.666  3.8e-06 ***
## GeographyIllinois       -1.233e+01  8.170e+00  -1.509 0.131921
## GeographyIndiana         1.558e+00  7.463e+00   0.209 0.834677
## GeographyIowa           -2.077e+01  7.548e+00  -2.752 0.006105 **
## GeographyKansas         -1.241e+01  7.466e+00  -1.662 0.097128 .
## GeographyKentucky        3.390e+00  7.162e+00   0.473 0.636176
## GeographyLouisiana       6.006e+00  7.962e+00   0.754 0.450925
## GeographyMaine          -1.109e+01  1.276e+01  -0.869 0.385166
## GeographyMaryland       -5.260e+00  8.982e+00  -0.586 0.558383
## GeographyMassachusetts  -2.846e+01  2.038e+01  -1.396 0.163257
## GeographyMichigan       -1.594e+01  8.539e+00  -1.867 0.062464 .
## GeographyMinnesota      -2.329e+01  7.359e+00  -3.164 0.001635 **
## GeographyMississippi     3.805e-02  7.869e+00   0.005 0.996143
## GeographyMissouri        1.661e+00  7.251e+00   0.229 0.818919
## GeographyMontana        -2.184e+01  1.031e+01  -2.120 0.034464 *
## GeographyNebraska       -1.630e+01  8.894e+00  -1.833 0.067288 .
## GeographyNevada          6.952e+00  1.098e+01   0.633 0.526785
## GeographyNew Hampshire  -1.485e+01  1.075e+01  -1.381 0.167698
## GeographyNew Jersey     -7.554e+00  9.618e+00  -0.785 0.432544
## GeographyNew Mexico     -2.538e+01  1.086e+01  -2.337 0.019775 *
## GeographyNew York       -1.837e+01  8.650e+00  -2.123 0.034161 *
## GeographyNorth Carolina -1.176e+01  7.111e+00  -1.653 0.098842 .
## GeographyNorth Dakota   -1.855e+01  8.702e+00  -2.132 0.033423 *
## GeographyOhio           -1.889e-01  7.665e+00  -0.025 0.980348
## GeographyOklahoma       -6.631e+00  7.906e+00  -0.839 0.401950
## GeographyOregon         -1.689e+01  9.263e+00  -1.824 0.068706 .
## GeographyPennsylvania   -1.952e+01  7.807e+00  -2.501 0.012653 *
## GeographySouth Carolina -2.619e+00  9.713e+00  -0.270 0.787533
## GeographySouth Dakota   -2.088e+01  8.168e+00  -2.556 0.010835 *
## GeographyTennessee       4.353e+00  7.324e+00   0.594 0.552515
## GeographyTexas          -2.670e+00  6.925e+00  -0.386 0.699937
## GeographyUtah           -3.539e+01  1.158e+01  -3.056 0.002349 **
## GeographyVermont        -2.415e+01  2.022e+01  -1.195 0.232742
## GeographyVirginia        8.919e+00  7.344e+00   1.214 0.225057
## GeographyWashington     -1.561e+01  8.597e+00  -1.815 0.069983 .
## GeographyWest Virginia  -6.732e-01  8.015e+00  -0.084 0.933085
## GeographyWisconsin      -1.994e+01  8.593e+00  -2.320 0.020674 *
## GeographyWyoming        -2.036e+01  1.266e+01  -1.609 0.108228
## AvgHouseholdSize         3.223e+00  2.398e+00   1.344 0.179421
## PercentMarried           2.342e-01  3.564e-01   0.657 0.511460
## PctNoHS18_24            -1.824e+01  1.383e+01  -1.319 0.187712
## PctHS18_24              -1.812e+01  1.384e+01  -1.309 0.191037
## PctSomeCol18_24         -1.834e+01  1.383e+01  -1.326 0.185466
## PctBachDeg18_24         -1.929e+01  1.384e+01  -1.394 0.163946
## PctPrivateCoverage       2.187e-01  3.109e-01   0.703 0.482112
## PctPublicCoverage        3.207e-01  4.787e-01   0.670 0.503188
## PctPublicCoverageAlone   6.966e-01  6.319e-01   1.102 0.270743
## PctWhite                -1.229e-01  1.558e-01  -0.789 0.430337
```

```
## PctBlack                -2.213e-01  1.690e-01  -1.309 0.190944
## PctAsian                 9.846e-02  4.505e-01   0.219 0.827071
## PctOtherRace            -1.052e+00  3.017e-01  -3.488 0.000523 ***
## PctMarriedHouseholds    -8.976e-01  3.337e-01  -2.690 0.007360 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.82 on 584 degrees of freedom
##   (1938 observations deleted due to missingness)
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5384
## F-statistic: 12.33 on 67 and 584 DF,  p-value: < 2.2e-16
```

Model	**Adjusted R-squared**	**Residual Std Error**
Preliminary model	0.5387	18.81
Model after adjusting outliers	0.5384	18.82

It can be seen that there is slight drop in the adjusted R-squared value, which means that the removed values were not too far from the linear regression line. However, the correction in data was essential since the MedianAge values were impractical.

- **Other Outliers:** From the remaining data, these is one outlier with more than 350 deaths, but this datapoint cannot be removed since it cannot be proved that it is incorrect. Thus, MedianAge was the only predictor whose values were modified.

Handling Missing Values

Code:

```
summary(Train.data$PctSomeCol18_24)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    7.10   33.60   40.30   40.87   46.20   79.00    1938

summary(Test.data$PctSomeCol18_24)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    9.60   34.73   41.25   41.58   47.25   78.30     347

#Dropping PctSomeCol18_24 due to lack of data
Train.data$PctSomeCol18_24=NULL
Test.data$PctSomeCol18_24=NULL
```

Explanation:

It can be seen that there are a lot of missing values in the PctSomeCol18_24 column. Few ways to solve this problem are:

- Calculate mean value of existing data and fill the missing values with mean value. This approach is good if the number of missing values is less. Here, since almost 50% of data is missing, this approach cannot be used.
- Other method is to predict the missing values by training a model using existing values. Again, since there is not enough training data, this method is also not used.
- The easiest but least preferred method is to entirely omit that predictor. This is used only if there is a large portion of data missing, which is the case here. Therefore, this predictor is entirely removed from the dataframe, as shown above.

The LR model was recreated after removing the PctSomeCol18_24 data, as shown below:

Code:

```
attach(Train.data)

linear.fit=lm(TARGET_deathRate~.,data=Train.data)
summary(linear.fit)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ ., data = Train.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.69 -10.50  -0.40  10.32 118.64
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.143e+02  1.598e+01   7.157 1.08e-12 ***
## incidenceRate                   2.007e-01  8.415e-03  23.854  < 2e-16 ***
## medIncome                      -1.481e-04  8.196e-05  -1.807 0.070893 .
## povertyPercent                  1.635e-01  1.658e-01   0.986 0.324019
## MedianAge                      -5.325e-01  1.104e+00  -0.482 0.629526
## MedianAgeMale                   2.293e-01  6.455e-01   0.355 0.722479
## MedianAgeFemale                -3.486e-02  5.428e-01  -0.064 0.948792
## GeographyAlaska                 8.255e+00  5.996e+00   1.377 0.168717
## GeographyArizona               -2.507e+01  6.708e+00  -3.738 0.000190 ***
## GeographyArkansas               7.115e+00  3.597e+00   1.978 0.048054 *
## GeographyCalifornia            -1.985e+01  4.292e+00  -4.625 3.94e-06 ***
## GeographyColorado              -2.734e+01  4.013e+00  -6.814 1.19e-11 ***
## GeographyConnecticut           -1.826e+01  7.750e+00  -2.356 0.018543 *
## GeographyDelaware              -1.411e+01  1.110e+01  -1.271 0.203695
## GeographyDistrict of Columbia -4.963e+00  1.900e+01  -0.261 0.793919
## GeographyFlorida               -6.565e+00  3.772e+00  -1.741 0.081874 .
## GeographyGeorgia               -8.206e+00  3.141e+00  -2.612 0.009048 **
## GeographyHawaii                -3.688e+01  1.197e+01  -3.081 0.002086 **
## GeographyIdaho                 -2.715e+01  4.085e+00  -6.645 3.70e-11 ***
## GeographyIllinois              -6.694e+00  3.526e+00  -1.899 0.057720 .
## GeographyIndiana                7.732e+00  3.598e+00   2.149 0.031746 *
```

```
## GeographyIowa                -1.631e+01  3.550e+00  -4.595 4.55e-06 ***
## GeographyKansas              -9.066e+00  3.463e+00  -2.618 0.008906 **
## GeographyKentucky             8.117e+00  3.468e+00   2.340 0.019340 *
## GeographyLouisiana           -1.485e+00  3.649e+00  -0.407 0.684017
## GeographyMaine               -8.948e+00  5.959e+00  -1.502 0.133320
## GeographyMaryland            -1.114e+00  4.978e+00  -0.224 0.822900
## GeographyMassachusetts       -1.535e+01  7.087e+00  -2.166 0.030395 *
## GeographyMichigan            -6.264e+00  3.613e+00  -1.734 0.083126 .
## GeographyMinnesota           -1.840e+01  3.678e+00  -5.005 5.99e-07 ***
## GeographyMississippi          1.463e+00  3.503e+00   0.418 0.676223
## GeographyMissouri             4.914e+00  3.355e+00   1.465 0.143144
## GeographyMontana             -2.279e+01  4.165e+00  -5.473 4.87e-08 ***
## GeographyNebraska            -1.071e+01  3.663e+00  -2.924 0.003491 **
## GeographyNevada              -2.800e+00  5.585e+00  -0.501 0.616137
## GeographyNew Hampshire       -1.068e+01  6.607e+00  -1.616 0.106252
## GeographyNew Jersey          -8.205e+00  5.195e+00  -1.579 0.114396
## GeographyNew Mexico          -2.185e+01  4.845e+00  -4.510 6.77e-06 ***
## GeographyNew York            -1.779e+01  3.991e+00  -4.459 8.61e-06 ***
## GeographyNorth Carolina      -1.202e+01  3.346e+00  -3.593 0.000334 ***
## GeographyNorth Dakota        -9.579e+00  4.086e+00  -2.344 0.019134 *
## GeographyOhio                 2.658e+00  3.536e+00   0.752 0.452360
## GeographyOklahoma             5.642e+00  3.741e+00   1.508 0.131616
## GeographyOregon              -1.417e+01  4.516e+00  -3.138 0.001719 **
## GeographyPennsylvania        -1.163e+01  3.759e+00  -3.093 0.002001 **
## GeographyRhode Island        -9.757e+00  9.818e+00  -0.994 0.320442
## GeographySouth Carolina      -3.633e+00  4.056e+00  -0.896 0.370421
## GeographySouth Dakota        -1.819e+01  3.962e+00  -4.591 4.62e-06 ***
## GeographyTennessee            3.683e+00  3.443e+00   1.070 0.284898
## GeographyTexas               -3.073e+00  3.215e+00  -0.956 0.339227
## GeographyUtah                -2.577e+01  4.892e+00  -5.268 1.50e-07 ***
## GeographyVermont             -1.499e+01  7.051e+00  -2.126 0.033634 *
## GeographyVirginia             5.282e+00  3.255e+00   1.623 0.104745
## GeographyWashington          -1.674e+01  4.315e+00  -3.879 0.000108 ***
## GeographyWest Virginia        9.441e-01  3.936e+00   0.240 0.810463
## GeographyWisconsin           -9.484e+00  3.753e+00  -2.527 0.011565 *
## GeographyWyoming             -9.611e+00  5.106e+00  -1.882 0.059932 .
## AvgHouseholdSize              1.265e+00  1.108e+00   1.142 0.253605
## PercentMarried                1.946e-01  1.686e-01   1.155 0.248403
## PctNoHS18_24                  5.314e-02  6.047e-02   0.879 0.379597
## PctHS18_24                    2.654e-01  5.104e-02   5.200 2.15e-07 ***
## PctBachDeg18_24              -4.694e-01  1.124e-01  -4.176 3.07e-05 ***
## PctPrivateCoverage           -5.566e-02  1.349e-01  -0.413 0.679881
## PctPublicCoverage             4.424e-01  2.176e-01   2.034 0.042103 *
## PctPublicCoverageAlone        4.009e-01  2.897e-01   1.384 0.166483
## PctWhite                     -1.692e-01  7.315e-02  -2.313 0.020791 *
## PctBlack                     -1.610e-01  7.946e-02  -2.026 0.042830 *
## PctAsian                     -1.633e-01  2.213e-01  -0.738 0.460597
## PctOtherRace                 -6.880e-01  1.362e-01  -5.052 4.67e-07 ***
## PctMarriedHouseholds         -4.788e-01  1.536e-01  -3.117 0.001850 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.52 on 2520 degrees of freedom
## Multiple R-squared:  0.5721, Adjusted R-squared:  0.5604
## F-statistic: 48.84 on 69 and 2520 DF,  p-value: < 2.2e-16
```

Explanation:

Model	Adjusted R-squared	Residual Std Error
Preliminary model	0.5387	18.81
Model after adjusting outliers	0.5384	18.82
Model after removing PctSomeCol18_24	0.5604	18.52

It is seen that the model performance improved after removing the PctSomeCol18_24 data. This is because since there was little data present in that column, the model was not getting enough values to determine accurate coefficient values. As a result there was more residual error, which decreased after removing the missing data column entirely.

Addressing Collinearity

Explanation:

Collinearity can be checked using correlation matrix or VIF. Using the previous LR model (Model after removing PctSomeCol18_24), collinearity was checked using the Variance Inflation Factor function. However, since there was also a categorical predictor (Geography), GVIF (Generalized Variance Inflation Factor) was calculated. GVIF^(1/(2*Df)) value less than 2 indicated VIF value less than 4, denoting that there is no significant collinearity between predictors. (Source: https://stats.stackexchange.com/questions/70679/which-variance-inflation-factor-should-i-be-using-textgvif-or-textgvif/96584#96584)

Code:

```
#Detecting collinearity
vif(linear.fit)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## incidenceRate            1.528272  1        1.236233
## medIncome                7.437125  1        2.727109
## povertyPercent           8.569118  1        2.927306
## MedianAge              246.597886  1       15.703435
## MedianAgeMale           85.134794  1        9.226852
## MedianAgeFemale         62.326513  1        7.894714
## Geography              205.203222 50        1.054683
```

```
## AvgHouseholdSize          1.500209  1        1.224830
## PercentMarried           10.218901  1        3.196702
## PctNoHS18_24              1.826803  1        1.351593
## PctHS18_24                1.589893  1        1.260910
## PctBachDeg18_24           2.007515  1        1.416868
## PctPrivateCoverage       15.515309  1        3.938948
## PctPublicCoverage        22.109302  1        4.702053
## PctPublicCoverageAlone   23.649884  1        4.863115
## PctWhite                 10.890936  1        3.300142
## PctBlack                 10.086562  1        3.175935
## PctAsian                  2.600064  1        1.612471
## PctOtherRace              1.872763  1        1.368489
## PctMarriedHouseholds      7.650046  1        2.765872
```

Explanation:

It can be seen that MedianAgeMale, MedianAgeFemale and MedianAge have GVIF^(1/(2*Df)) value greater than 4. This value is also high for PctPublicCoverage and PctPublicCoverageAlone. This indicates that there is collinearity between these predictors. To remove collinearity, MedianAgeFemale, MedianAgeMale and PctPublicCoverageAlone were removed from the model. The results are shown below:

```
#Removing collinearity
attach(Train.data)

linear.fit=lm(TARGET_deathRate~.-MedianAgeMale-MedianAgeFemale-PctPublicCover
ageAlone,data=Train.data)
summary(linear.fit)

##
## Call:
## lm(formula = TARGET_deathRate ~ . - MedianAgeMale - MedianAgeFemale -
##     PctPublicCoverageAlone, data = Train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.733 -10.400  -0.452  10.292 118.746
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.239e+02  1.444e+01   8.583  < 2e-16 ***
## incidenceRate                2.016e-01  8.371e-03  24.086  < 2e-16 ***
## medIncome                   -1.238e-04  8.023e-05  -1.543 0.122849
## povertyPercent               1.728e-01  1.655e-01   1.044 0.296361
## MedianAge                   -4.433e-01  1.432e-01  -3.095 0.001990 **
## GeographyAlaska              7.815e+00  5.967e+00   1.310 0.190461
## GeographyArizona            -2.489e+01  6.704e+00  -3.712 0.000210 ***
## GeographyArkansas            7.330e+00  3.589e+00   2.042 0.041228 *
## GeographyCalifornia         -1.961e+01  4.285e+00  -4.577 4.93e-06 ***
## GeographyColorado           -2.680e+01  3.982e+00  -6.731 2.08e-11 ***
```

```
## GeographyConnecticut          -1.746e+01  7.713e+00  -2.264 0.023647 *
## GeographyDelaware             -1.388e+01  1.109e+01  -1.251 0.211077
## GeographyDistrict of Columbia -4.173e+00  1.898e+01  -0.220 0.826020
## GeographyFlorida              -6.853e+00  3.764e+00  -1.821 0.068799 .
## GeographyGeorgia              -8.242e+00  3.137e+00  -2.627 0.008662 **
## GeographyHawaii               -3.577e+01  1.194e+01  -2.996 0.002761 **
## GeographyIdaho                -2.742e+01  4.079e+00  -6.722 2.21e-11 ***
## GeographyIllinois             -6.181e+00  3.490e+00  -1.771 0.076713 .
## GeographyIndiana               8.049e+00  3.579e+00   2.249 0.024601 *
## GeographyIowa                 -1.621e+01  3.537e+00  -4.584 4.78e-06 ***
## GeographyKansas               -9.058e+00  3.444e+00  -2.630 0.008588 **
## GeographyKentucky              8.752e+00  3.427e+00   2.554 0.010708 *
## GeographyLouisiana            -1.574e+00  3.647e+00  -0.432 0.666071
## GeographyMaine                -8.777e+00  5.957e+00  -1.473 0.140786
## GeographyMaryland             -6.686e-01  4.963e+00  -0.135 0.892848
## GeographyMassachusetts        -1.437e+01  7.044e+00  -2.040 0.041421 *
## GeographyMichigan             -6.312e+00  3.612e+00  -1.748 0.080638 .
## GeographyMinnesota            -1.791e+01  3.657e+00  -4.899 1.02e-06 ***
## GeographyMississippi           1.343e+00  3.499e+00   0.384 0.701158
## GeographyMissouri              4.884e+00  3.350e+00   1.458 0.145064
## GeographyMontana              -2.297e+01  4.160e+00  -5.522 3.69e-08 ***
## GeographyNebraska             -1.065e+01  3.638e+00  -2.928 0.003442 **
## GeographyNevada               -3.230e+00  5.575e+00  -0.579 0.562439
## GeographyNew Hampshire        -1.043e+01  6.601e+00  -1.579 0.114400
## GeographyNew Jersey           -8.032e+00  5.181e+00  -1.550 0.121160
## GeographyNew Mexico           -2.162e+01  4.839e+00  -4.467 8.28e-06 ***
## GeographyNew York             -1.738e+01  3.972e+00  -4.375 1.26e-05 ***
## GeographyNorth Carolina       -1.211e+01  3.335e+00  -3.630 0.000289 ***
## GeographyNorth Dakota         -9.615e+00  4.081e+00  -2.356 0.018543 *
## GeographyOhio                  3.362e+00  3.487e+00   0.964 0.335036
## GeographyOklahoma              5.315e+00  3.729e+00   1.425 0.154140
## GeographyOregon               -1.406e+01  4.514e+00  -3.115 0.001859 **
## GeographyPennsylvania         -1.155e+01  3.750e+00  -3.081 0.002087 **
## GeographyRhode Island         -9.756e+00  9.805e+00  -0.995 0.319812
## GeographySouth Carolina       -3.591e+00  4.050e+00  -0.887 0.375328
## GeographySouth Dakota         -1.790e+01  3.940e+00  -4.543 5.81e-06 ***
## GeographyTennessee             4.111e+00  3.420e+00   1.202 0.229476
## GeographyTexas                -3.697e+00  3.187e+00  -1.160 0.246157
## GeographyUtah                 -2.626e+01  4.878e+00  -5.384 7.97e-08 ***
## GeographyVermont              -1.370e+01  6.994e+00  -1.959 0.050223 .
## GeographyVirginia              5.342e+00  3.249e+00   1.644 0.100289
## GeographyWashington           -1.667e+01  4.314e+00  -3.864 0.000114 ***
## GeographyWest Virginia         1.435e+00  3.921e+00   0.366 0.714463
## GeographyWisconsin            -8.918e+00  3.730e+00  -2.391 0.016878 *
## GeographyWyoming              -9.863e+00  5.102e+00  -1.933 0.053328 .
## AvgHouseholdSize               1.343e+00  1.100e+00   1.222 0.221976
## PercentMarried                 2.051e-01  1.630e-01   1.259 0.208267
## PctNoHS18_24                   4.516e-02  5.902e-02   0.765 0.444307
## PctHS18_24                     2.665e-01  5.071e-02   5.256 1.60e-07 ***
## PctBachDeg18_24               -4.605e-01  1.118e-01  -4.118 3.94e-05 ***
```

```
## PctPrivateCoverage            -1.760e-01  1.062e-01  -1.658 0.097467 .
## PctPublicCoverage              6.427e-01  1.481e-01   4.341 1.48e-05 ***
## PctWhite                      -1.669e-01  7.300e-02  -2.286 0.022317 *
## PctBlack                      -1.564e-01  7.875e-02  -1.986 0.047146 *
## PctAsian                      -1.652e-01  2.212e-01  -0.747 0.455140
## PctOtherRace                  -6.794e-01  1.358e-01  -5.005 5.97e-07 ***
## PctMarriedHouseholds          -4.814e-01  1.507e-01  -3.195 0.001416 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.52 on 2523 degrees of freedom
## Multiple R-squared:  0.5717, Adjusted R-squared:  0.5605
## F-statistic: 51.03 on 66 and 2523 DF,  p-value: < 2.2e-16

vif(linear.fit)

##                           GVIF Df GVIF^(1/(2*Df))
## incidenceRate         1.512442  1        1.229814
## medIncome             7.128272  1        2.669882
## povertyPercent        8.539792  1        2.922292
## MedianAge             4.152822  1        2.037847
## Geography           141.755172 50        1.050789
## AvgHouseholdSize      1.477927  1        1.215700
## PercentMarried        9.553230  1        3.090830
## PctNoHS18_24          1.740553  1        1.319300
## PctHS18_24            1.569423  1        1.252766
## PctBachDeg18_24       1.987695  1        1.409856
## PctPrivateCoverage    9.616333  1        3.101021
## PctPublicCoverage    10.240573  1        3.200090
## PctWhite             10.849728  1        3.293893
## PctBlack              9.909250  1        3.147896
## PctAsian              2.597367  1        1.611635
## PctOtherRace          1.861756  1        1.364462
## PctMarriedHouseholds  7.359207  1        2.712786
```

Explanation:

Model	Adjusted R-squared	Residual Std Error	Collinearity
Preliminary model	0.5387	18.81	-
Model after adjusting outliers	0.5384	18.82	-
Model after removing PctSomeCol18_24	0.5604	18.52	15.7 (maximum)
Model after removing collinearity	0.5605	18.52	3.29 (maximum)

There is not much different after removing collinearity from the model. Slight improvement in Adjusted R-squared value was observed. However, the main advantage of removing dependent predictors is that we get a less complex model which gives similar performance as a complex model with dependent predictors.

2. Linear Regression

Developing a linear regression model.

Note: The model here is developed after removing PctSomeCol18_24 data and correcting the MedianAge values.

Code:

```
attach(Train.data)

linear.fit=lm(TARGET_deathRate~.,data=Train.data)
summary(linear.fit)

##
## Call:
## lm(formula = TARGET_deathRate ~ ., data = Train.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.69 -10.50  -0.40  10.32 118.64
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.143e+02  1.598e+01   7.157 1.08e-12 ***
## incidenceRate                   2.007e-01  8.415e-03  23.854  < 2e-16 ***
## medIncome                      -1.481e-04  8.196e-05  -1.807 0.070893 .
## povertyPercent                  1.635e-01  1.658e-01   0.986 0.324019
## MedianAge                      -5.325e-01  1.104e+00  -0.482 0.629526
## MedianAgeMale                   2.293e-01  6.455e-01   0.355 0.722479
## MedianAgeFemale                -3.486e-02  5.428e-01  -0.064 0.948792
## GeographyAlaska                 8.255e+00  5.996e+00   1.377 0.168717
## GeographyArizona               -2.507e+01  6.708e+00  -3.738 0.000190 ***
## GeographyArkansas               7.115e+00  3.597e+00   1.978 0.048054 *
## GeographyCalifornia            -1.985e+01  4.292e+00  -4.625 3.94e-06 ***
## GeographyColorado              -2.734e+01  4.013e+00  -6.814 1.19e-11 ***
## GeographyConnecticut           -1.826e+01  7.750e+00  -2.356 0.018543 *
## GeographyDelaware              -1.411e+01  1.110e+01  -1.271 0.203695
## GeographyDistrict of Columbia -4.963e+00  1.900e+01  -0.261 0.793919
## GeographyFlorida               -6.565e+00  3.772e+00  -1.741 0.081874 .
## GeographyGeorgia               -8.206e+00  3.141e+00  -2.612 0.009048 **
## GeographyHawaii                -3.688e+01  1.197e+01  -3.081 0.002086 **
## GeographyIdaho                 -2.715e+01  4.085e+00  -6.645 3.70e-11 ***
## GeographyIllinois              -6.694e+00  3.526e+00  -1.899 0.057720 .
## GeographyIndiana                7.732e+00  3.598e+00   2.149 0.031746 *
```

```
## GeographyIowa                -1.631e+01  3.550e+00  -4.595 4.55e-06 ***
## GeographyKansas              -9.066e+00  3.463e+00  -2.618 0.008906 **
## GeographyKentucky             8.117e+00  3.468e+00   2.340 0.019340 *
## GeographyLouisiana           -1.485e+00  3.649e+00  -0.407 0.684017
## GeographyMaine               -8.948e+00  5.959e+00  -1.502 0.133320
## GeographyMaryland            -1.114e+00  4.978e+00  -0.224 0.822900
## GeographyMassachusetts       -1.535e+01  7.087e+00  -2.166 0.030395 *
## GeographyMichigan            -6.264e+00  3.613e+00  -1.734 0.083126 .
## GeographyMinnesota           -1.840e+01  3.678e+00  -5.005 5.99e-07 ***
## GeographyMississippi          1.463e+00  3.503e+00   0.418 0.676223
## GeographyMissouri             4.914e+00  3.355e+00   1.465 0.143144
## GeographyMontana             -2.279e+01  4.165e+00  -5.473 4.87e-08 ***
## GeographyNebraska            -1.071e+01  3.663e+00  -2.924 0.003491 **
## GeographyNevada              -2.800e+00  5.585e+00  -0.501 0.616137
## GeographyNew Hampshire       -1.068e+01  6.607e+00  -1.616 0.106252
## GeographyNew Jersey          -8.205e+00  5.195e+00  -1.579 0.114396
## GeographyNew Mexico          -2.185e+01  4.845e+00  -4.510 6.77e-06 ***
## GeographyNew York            -1.779e+01  3.991e+00  -4.459 8.61e-06 ***
## GeographyNorth Carolina      -1.202e+01  3.346e+00  -3.593 0.000334 ***
## GeographyNorth Dakota        -9.579e+00  4.086e+00  -2.344 0.019134 *
## GeographyOhio                 2.658e+00  3.536e+00   0.752 0.452360
## GeographyOklahoma             5.642e+00  3.741e+00   1.508 0.131616
## GeographyOregon              -1.417e+01  4.516e+00  -3.138 0.001719 **
## GeographyPennsylvania        -1.163e+01  3.759e+00  -3.093 0.002001 **
## GeographyRhode Island        -9.757e+00  9.818e+00  -0.994 0.320442
## GeographySouth Carolina      -3.633e+00  4.056e+00  -0.896 0.370421
## GeographySouth Dakota        -1.819e+01  3.962e+00  -4.591 4.62e-06 ***
## GeographyTennessee            3.683e+00  3.443e+00   1.070 0.284898
## GeographyTexas               -3.073e+00  3.215e+00  -0.956 0.339227
## GeographyUtah                -2.577e+01  4.892e+00  -5.268 1.50e-07 ***
## GeographyVermont             -1.499e+01  7.051e+00  -2.126 0.033634 *
## GeographyVirginia             5.282e+00  3.255e+00   1.623 0.104745
## GeographyWashington          -1.674e+01  4.315e+00  -3.879 0.000108 ***
## GeographyWest Virginia        9.441e-01  3.936e+00   0.240 0.810463
## GeographyWisconsin           -9.484e+00  3.753e+00  -2.527 0.011565 *
## GeographyWyoming             -9.611e+00  5.106e+00  -1.882 0.059932 .
## AvgHouseholdSize              1.265e+00  1.108e+00   1.142 0.253605
## PercentMarried                1.946e-01  1.686e-01   1.155 0.248403
## PctNoHS18_24                  5.314e-02  6.047e-02   0.879 0.379597
## PctHS18_24                    2.654e-01  5.104e-02   5.200 2.15e-07 ***
## PctBachDeg18_24              -4.694e-01  1.124e-01  -4.176 3.07e-05 ***
## PctPrivateCoverage           -5.566e-02  1.349e-01  -0.413 0.679881
## PctPublicCoverage             4.424e-01  2.176e-01   2.034 0.042103 *
## PctPublicCoverageAlone        4.009e-01  2.897e-01   1.384 0.166483
## PctWhite                     -1.692e-01  7.315e-02  -2.313 0.020791 *
## PctBlack                     -1.610e-01  7.946e-02  -2.026 0.042830 *
## PctAsian                     -1.633e-01  2.213e-01  -0.738 0.460597
## PctOtherRace                 -6.880e-01  1.362e-01  -5.052 4.67e-07 ***
## PctMarriedHouseholds         -4.788e-01  1.536e-01  -3.117 0.001850 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.52 on 2520 degrees of freedom
## Multiple R-squared:  0.5721, Adjusted R-squared:  0.5604
## F-statistic: 48.84 on 69 and 2520 DF,  p-value: < 2.2e-16
```

Checking performance with significant and insignificant variables

Explanation:

From the above LR model, it can be seen that the following predictors have p-values less than 0.05: incidenceRate, PctHS18_24, PctBachDeg18_24, PctPublicCoverage, PctWhite, PctBlack, PctOtherRace, PctMarriedHouseholds

Thus they are statistically significant and included in the new LR model. It is to be noted that some states also have p-values<0.05. For simplicity, Geography has also been included in the new LR model shown below:

Code:

```
#LR with only significant predictors
attach(Train.data)

linear.fit=lm(TARGET_deathRate~incidenceRate+Geography+PctHS18_24+PctBachDeg1
8_24+PctPublicCoverage+PctWhite+PctBlack+PctOtherRace+PctMarriedHouseholds,da
ta=Train.data)
summary(linear.fit)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ incidenceRate + Geography + PctHS18_24 +
##     PctBachDeg18_24 + PctPublicCoverage + PctWhite + PctBlack +
##     PctOtherRace + PctMarriedHouseholds, data = Train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.757 -10.531  -0.284  10.349 114.466
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    123.086639   9.172998  13.418  < 2e-16 ***
## incidenceRate                    0.196183   0.008281  23.691  < 2e-16 ***
## GeographyAlaska                  2.315095   5.893073   0.393 0.694463
## GeographyArizona               -28.937982   6.685009  -4.329 1.56e-05 ***
## GeographyArkansas                7.641477   3.558846   2.147 0.031874 *
## GeographyCalifornia            -25.598730   4.020079  -6.368 2.27e-10 ***
## GeographyColorado              -28.641500   3.912847  -7.320 3.32e-13 ***
## GeographyConnecticut           -24.709060   7.607994  -3.248 0.001178 **
## GeographyDelaware              -18.775151  11.113442  -1.689 0.091264 .
## GeographyDistrict of Columbia -11.423675  18.991525  -0.602 0.547551
```

```
## GeographyFlorida               -9.167919   3.654842  -2.508 0.012189 *
## GeographyGeorgia               -6.707326   3.078721  -2.179 0.029453 *
## GeographyHawaii               -53.588952  10.489937  -5.109 3.49e-07 ***
## GeographyIdaho                -26.647555   4.101777  -6.497 9.86e-11 ***
## GeographyIllinois              -8.930416   3.430585  -2.603 0.009290 **
## GeographyIndiana                7.358490   3.574052   2.059 0.039609 *
## GeographyIowa                 -19.254193   3.501172  -5.499 4.19e-08 ***
## GeographyKansas               -10.771342   3.438934  -3.132 0.001755 **
## GeographyKentucky              10.256236   3.412188   3.006 0.002675 **
## GeographyLouisiana              0.119300   3.623644   0.033 0.973739
## GeographyMaine                -12.738131   5.931280  -2.148 0.031839 *
## GeographyMaryland              -6.722954   4.793973  -1.402 0.160926
## GeographyMassachusetts        -21.059184   6.879378  -3.061 0.002228 **
## GeographyMichigan             -10.289860   3.557365  -2.893 0.003854 **
## GeographyMinnesota            -21.783582   3.579407  -6.086 1.34e-09 ***
## GeographyMississippi            3.465296   3.468602   0.999 0.317867
## GeographyMissouri               5.080811   3.370936   1.507 0.131874
## GeographyMontana              -24.851037   4.132879  -6.013 2.08e-09 ***
## GeographyNebraska             -13.212216   3.620120  -3.650 0.000268 ***
## GeographyNevada                -7.903591   5.510731  -1.434 0.151634
## GeographyNew Hampshire        -15.284403   6.571983  -2.326 0.020114 *
## GeographyNew Jersey           -14.556562   4.983595  -2.921 0.003521 **
## GeographyNew Mexico           -26.340833   4.738324  -5.559 3.00e-08 ***
## GeographyNew York             -22.323210   3.839150  -5.815 6.84e-09 ***
## GeographyNorth Carolina       -13.247431   3.339178  -3.967 7.47e-05 ***
## GeographyNorth Dakota         -13.796299   3.992987  -3.455 0.000559 ***
## GeographyOhio                   2.054257   3.484878   0.589 0.555594
## GeographyOklahoma               3.938426   3.735947   1.054 0.291893
## GeographyOregon               -17.459123   4.495514  -3.884 0.000106 ***
## GeographyPennsylvania         -15.201056   3.696298  -4.113 4.04e-05 ***
## GeographyRhode Island         -16.663869   9.794622  -1.701 0.089004 .
## GeographySouth Carolina        -4.260021   4.064355  -1.048 0.294673
## GeographySouth Dakota         -20.413914   3.894978  -5.241 1.73e-07 ***
## GeographyTennessee              4.557924   3.425905   1.330 0.183497
## GeographyTexas                 -2.816951   3.115337  -0.904 0.365965
## GeographyUtah                 -23.182795   4.891117  -4.740 2.26e-06 ***
## GeographyVermont              -18.007121   6.874036  -2.620 0.008856 **
## GeographyVirginia               2.123906   3.227949   0.658 0.510615
## GeographyWashington           -21.165142   4.235462  -4.997 6.22e-07 ***
## GeographyWest Virginia          0.483515   3.916792   0.123 0.901763
## GeographyWisconsin            -12.805298   3.696200  -3.464 0.000540 ***
## GeographyWyoming              -12.628387   5.096660  -2.478 0.013285 *
## PctHS18_24                      0.222949   0.049166   4.535 6.04e-06 ***
## PctBachDeg18_24                -0.691959   0.102922  -6.723 2.19e-11 ***
## PctPublicCoverage               0.824269   0.066134  12.464  < 2e-16 ***
## PctWhite                       -0.304427   0.060750  -5.011 5.78e-07 ***
## PctBlack                       -0.293064   0.071225  -4.115 4.00e-05 ***
## PctOtherRace                   -0.555896   0.134338  -4.138 3.62e-05 ***
## PctMarriedHouseholds           -0.561039   0.083873  -6.689 2.75e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.68 on 2531 degrees of freedom
## Multiple R-squared:  0.5628, Adjusted R-squared:  0.5528
## F-statistic: 56.18 on 58 and 2531 DF,  p-value: < 2.2e-16
```

```
vif(linear.fit)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## incidenceRate         1.454683  1        1.206103
## Geography            15.223238 50        1.027602
## PctHS18_24            1.450028  1        1.204171
## PctBachDeg18_24       1.654666  1        1.286338
## PctPublicCoverage     2.008000  1        1.417039
## PctWhite              7.383899  1        2.717333
## PctBlack              7.967069  1        2.822600
## PctOtherRace          1.791779  1        1.338574
## PctMarriedHouseholds  2.240963  1        1.496985
```

Explanation:

Model	**Adjusted R-squared**	**Residual Std Error**
Model after removing PctSomeCol18_24 and adjusting MedianAge	`0.5604`	`18.52`
Model having only significant predictors	`0.5528`	`18.68`

The model was simplified considerably after including only significant factors for training. Although the number of variables reduced from 20 to 9, the model performance was not affected significantly. There was a slight drop in adjusted R-squared value (0.76%) and slight increase in residual std error (0.16), it is acceptable due to simplification of the new model.

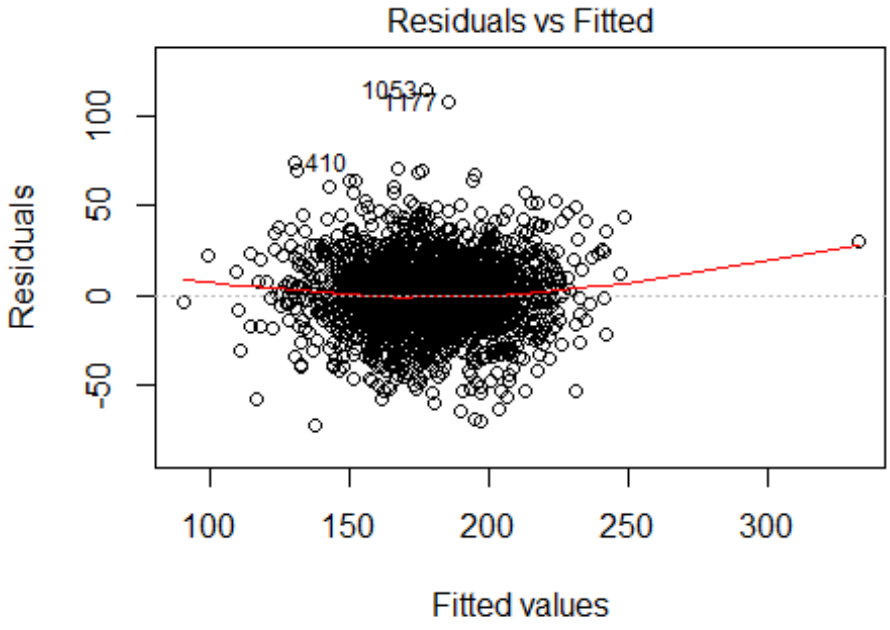
Model diagnosis

Explanation:

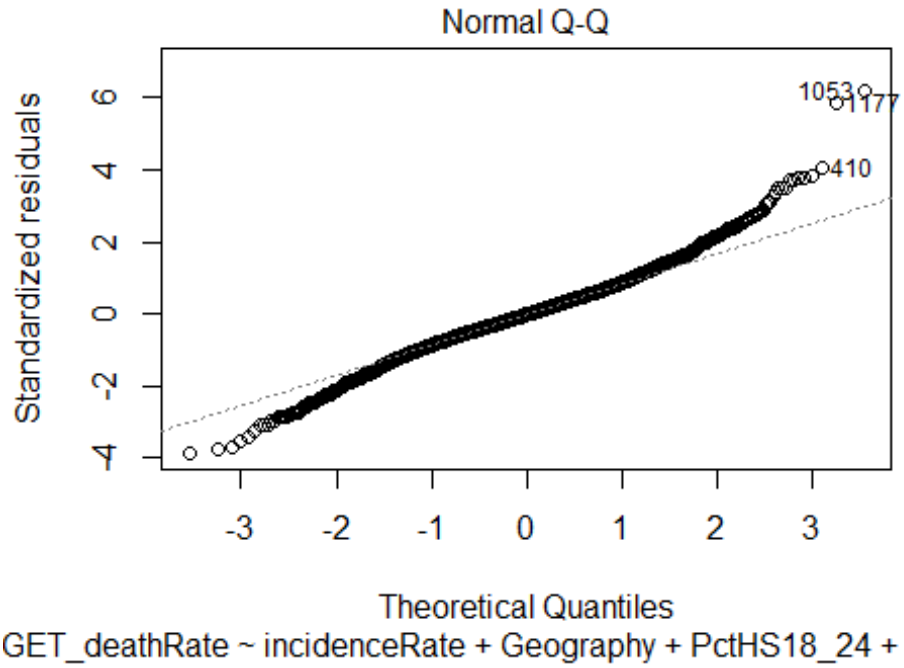
Model diagnostics was plotted for LR model with only significant variables as shown below.

Code:

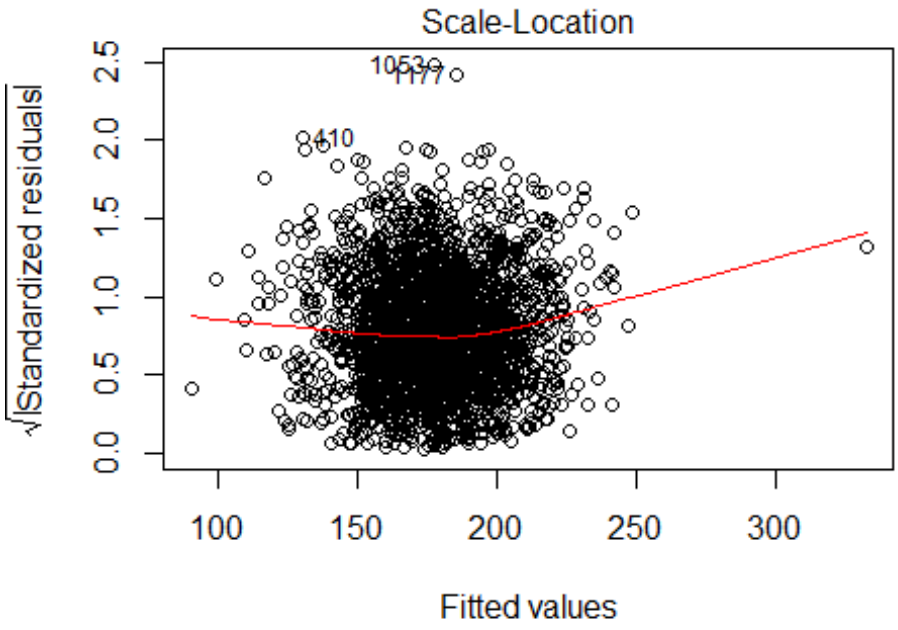
```
plot(linear.fit)
```



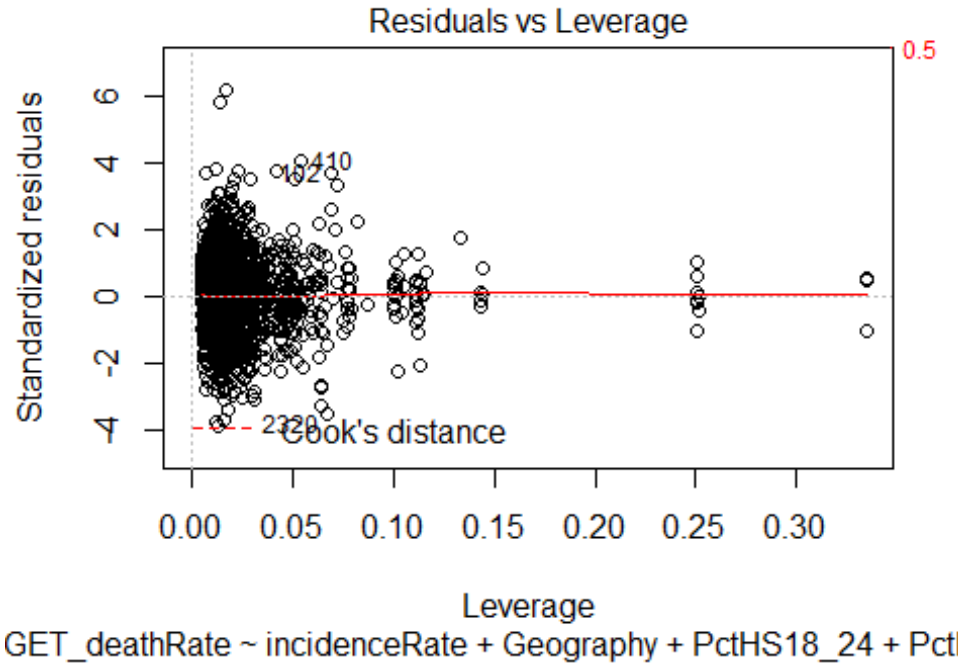
From the Residuals vs. Fitted values plot, it is evident that the residuals have linear patterns, since the plot is nearly a straight line. Therefore, there is a fairly linear relationship between the predictors and response. The curvature induced is due to a single point that has a high fitted value.



From the Normal Q-Q plot, it can be seen that the residuals are situated near the normality line. Although point 1053 and 1177 are slightly away from the line there is not much significant deviation from the line. Therefore, the data is normally distributed.



In the scale-location plot, the points appear to be uniformly distributed and there is no significant pattern across the graph. Therefore, there is uniform variance of residuals (homoscedasticity is satisfied)



All the points are within the Cook's distance lines. Therefore, there are no outliers that, if removed from the analysis would significantly change the regression model.

Inclusion of non-linear terms

```
#LR with interaction terms
attach(Train.data)

linear.fit=lm(TARGET_deathRate~incidenceRate*sqrt(medIncome)+Geography+PctPub
licCoverage+PctHS18_24+PctBachDeg18_24+PctWhite+PctBlack+PctOtherRace+I(incid
enceRate^2)*PctMarriedHouseholds,data=Train.data)
summary(linear.fit)

Call:
lm(formula = TARGET_deathRate ~ incidenceRate * sqrt(medIncome) +
    Geography + PctPublicCoverage + PctHS18_24 + PctBachDeg18_24 +
    PctWhite + PctBlack + PctOtherRace + I(incidenceRate^2) *
    PctMarriedHouseholds, data = Train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-75.012 -10.648  -0.439  10.363 119.637

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          8.868e+01  3.024e+01   2.932 0.003398 **
incidenceRate                        4.144e-01  7.243e-02   5.721 1.19e-08 ***
sqrt(medIncome)                      3.435e-01  1.364e-01   2.519 0.011847 *
GeographyAlaska                      5.411e+00  5.951e+00   0.909 0.363306
GeographyArizona                    -2.365e+01  6.735e+00  -3.511 0.000454 ***
GeographyArkansas                    9.567e+00  3.550e+00   2.695 0.007078 **
GeographyCalifornia                 -1.936e+01  4.173e+00  -4.640 3.67e-06 ***
GeographyColorado                   -2.592e+01  3.969e+00  -6.531 7.85e-11 ***
GeographyConnecticut                -1.451e+01  7.736e+00  -1.876 0.060762 .
GeographyDelaware                   -1.063e+01  1.111e+01  -0.956 0.339057
GeographyDistrict of Columbia        6.152e+00  1.907e+01   0.323 0.747040
GeographyFlorida                    -6.601e+00  3.700e+00  -1.784 0.074533 .
GeographyGeorgia                    -6.810e+00  3.063e+00  -2.223 0.026277 *
GeographyHawaii                     -5.046e+01  1.045e+01  -4.827 1.47e-06 ***
GeographyIdaho                      -2.588e+01  4.083e+00  -6.339 2.73e-10 ***
GeographyIllinois                   -4.641e+00  3.484e+00  -1.332 0.182896
GeographyIndiana                     8.820e+00  3.556e+00   2.480 0.013200 *
GeographyIowa                       -1.585e+01  3.523e+00  -4.499 7.12e-06 ***
GeographyKansas                     -9.592e+00  3.419e+00  -2.805 0.005063 **
GeographyKentucky                    1.018e+01  3.396e+00   2.997 0.002756 **
GeographyLouisiana                   1.141e+00  3.601e+00   0.317 0.751443
GeographyMaine                      -7.962e+00  5.945e+00  -1.339 0.180593
GeographyMaryland                    3.293e-01  4.941e+00   0.067 0.946863
GeographyMassachusetts              -9.863e+00  7.073e+00  -1.395 0.163272
GeographyMichigan                   -6.707e+00  3.593e+00  -1.867 0.062067 .
GeographyMinnesota                  -1.756e+01  3.631e+00  -4.836 1.40e-06 ***
GeographyMississippi                 3.206e+00  3.444e+00   0.931 0.352088
GeographyMissouri                    5.782e+00  3.353e+00   1.724 0.084801 .
GeographyMontana                    -2.438e+01  4.105e+00  -5.938 3.28e-09 ***
GeographyNebraska                   -1.163e+01  3.610e+00  -3.222 0.001290 **
GeographyNevada                     -3.172e+00  5.532e+00  -0.573 0.566416
GeographyNew Hampshire              -8.669e+00  6.613e+00  -1.311 0.190020
```

```
GeographyNew Jersey                        -4.982e+00  5.198e+00  -0.958 0.337946
GeographyNew Mexico                        -2.176e+01  4.827e+00  -4.509 6.81e-06 ***
GeographyNew York                          -1.532e+01  3.979e+00  -3.850 0.000121 ***
GeographyNorth Carolina                    -1.243e+01  3.319e+00  -3.745 0.000184 ***
GeographyNorth Dakota                      -1.177e+01  3.986e+00  -2.953 0.003175 **
GeographyOhio                               4.422e+00  3.483e+00   1.270 0.204316
GeographyOklahoma                           4.999e+00  3.714e+00   1.346 0.178398
GeographyOregon                            -1.332e+01  4.526e+00  -2.943 0.003280 **
GeographyPennsylvania                      -1.183e+01  3.709e+00  -3.190 0.001441 **
GeographyRhode Island                      -8.048e+00  9.824e+00  -0.819 0.412732
GeographySouth Carolina                    -3.380e+00  4.043e+00  -0.836 0.403146
GeographySouth Dakota                      -1.892e+01  3.879e+00  -4.877 1.14e-06 ***
GeographyTennessee                          4.843e+00  3.402e+00   1.423 0.154717
GeographyTexas                             -1.262e+00  3.115e+00  -0.405 0.685429
GeographyUtah                              -2.307e+01  4.896e+00  -4.712 2.59e-06 ***
GeographyVermont                           -1.111e+01  6.936e+00  -1.602 0.109252
GeographyVirginia                           4.838e+00  3.243e+00   1.492 0.135905
GeographyWashington                        -1.545e+01  4.333e+00  -3.566 0.000369 ***
GeographyWest Virginia                      2.262e+00  3.902e+00   0.580 0.562213
GeographyWisconsin                         -9.486e+00  3.722e+00  -2.549 0.010876 *
GeographyWyoming                           -9.953e+00  5.107e+00  -1.949 0.051449 .
PctPublicCoverage                           4.782e-01  9.620e-02   4.970 7.13e-07 ***
PctHS18_24                                  2.395e-01  4.919e-02   4.868 1.20e-06 ***
PctBachDeg18_24                            -5.576e-01  1.062e-01  -5.249 1.65e-07 ***
PctWhite                                   -3.509e-01  6.088e-02  -5.763 9.29e-09 ***
PctBlack                                   -2.836e-01  7.089e-02  -4.000 6.51e-05 ***
PctOtherRace                               -6.306e-01  1.340e-01  -4.706 2.66e-06 ***
I(incidenceRate^2)                         -1.281e-04  6.716e-05  -1.908 0.056550 .
PctMarriedHouseholds                       -9.432e-01  2.677e-01  -3.523 0.000434 ***
incidenceRate:sqrt(medIncome)              -1.153e-03  3.022e-04  -3.815 0.000140 ***
I(incidenceRate^2):PctMarriedHouseholds     3.126e-06  1.256e-06   2.488 0.012903 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.55 on 2527 degrees of freedom
Multiple R-squared:  0.5698,     Adjusted R-squared:  0.5593
F-statistic: 53.99 on 62 and 2527 DF,  p-value: < 2.2e-16
```

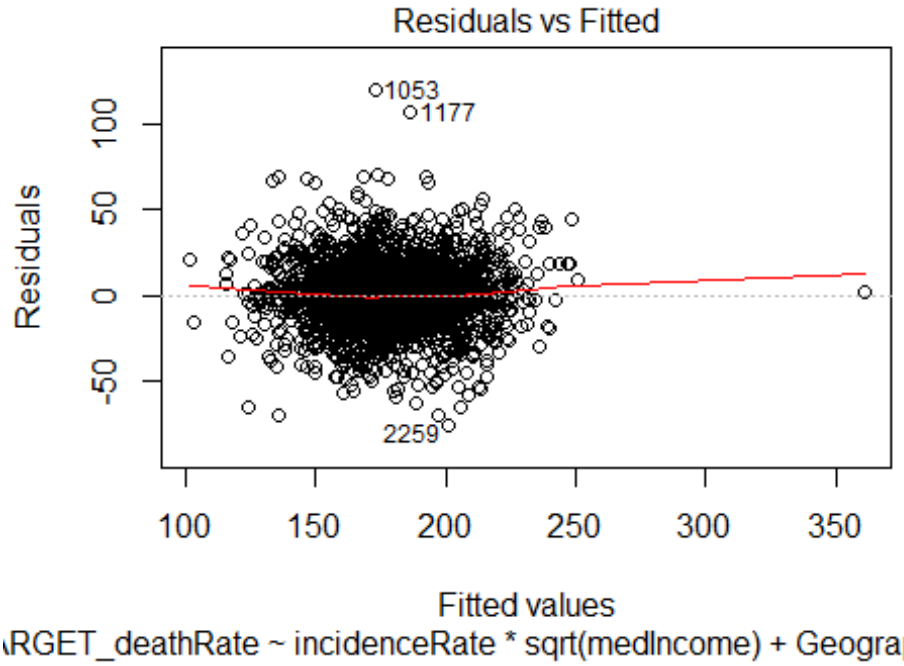
Explanation:

Non-linear terms like sqrt(medIncome), along with interaction between terms like $(\text{incidenceRate})^2$ and PctMarriedHouseholds were included in the model. It is interesting to note that there was not significant improvement in the model performance in terms of adjusted R-squared value or Residual standard error.

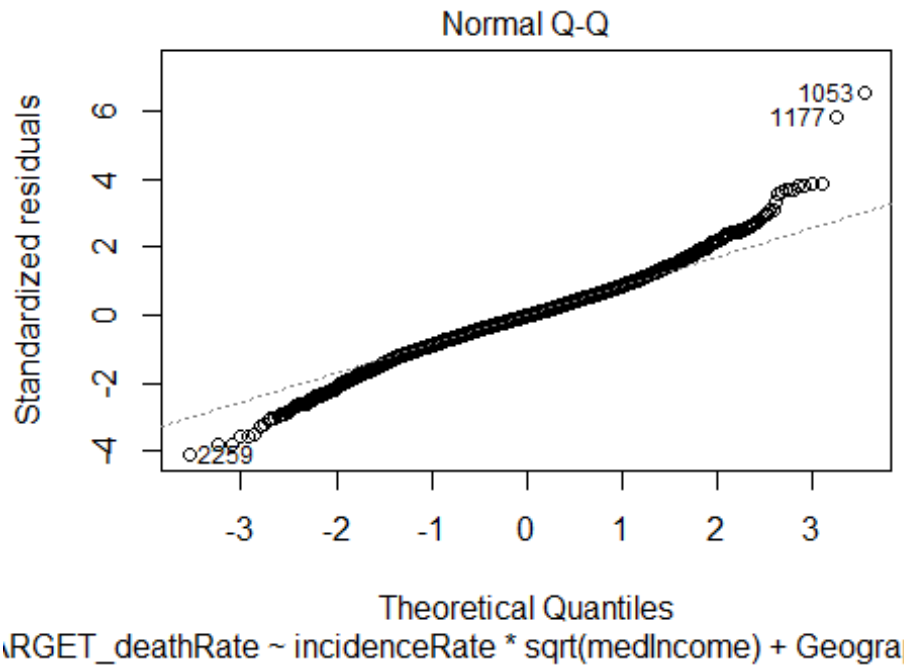
The reason behind this could be that these predictors cannot be modified any further to fit the data in a better way. New predictors may be needed to fit the data better and improve the adjusted R-squared value. The model, with and without interaction terms is summarized below:

Model	**Adjusted R-squared**	**Residual Std Error**

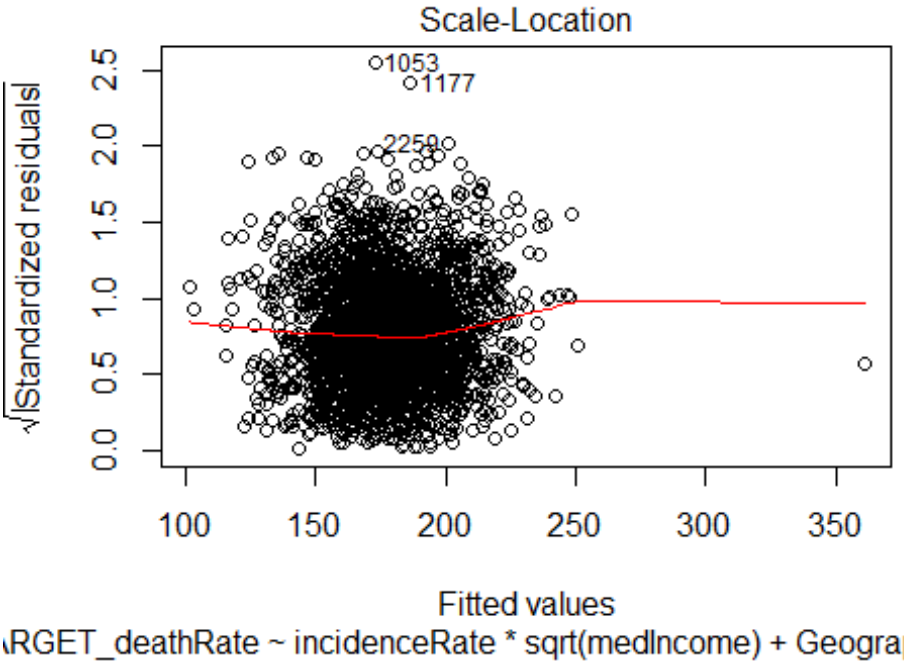
Model having only significant predictors	0.5528	18.68
Model with significant predictors and interaction terms	0.5593	18.55



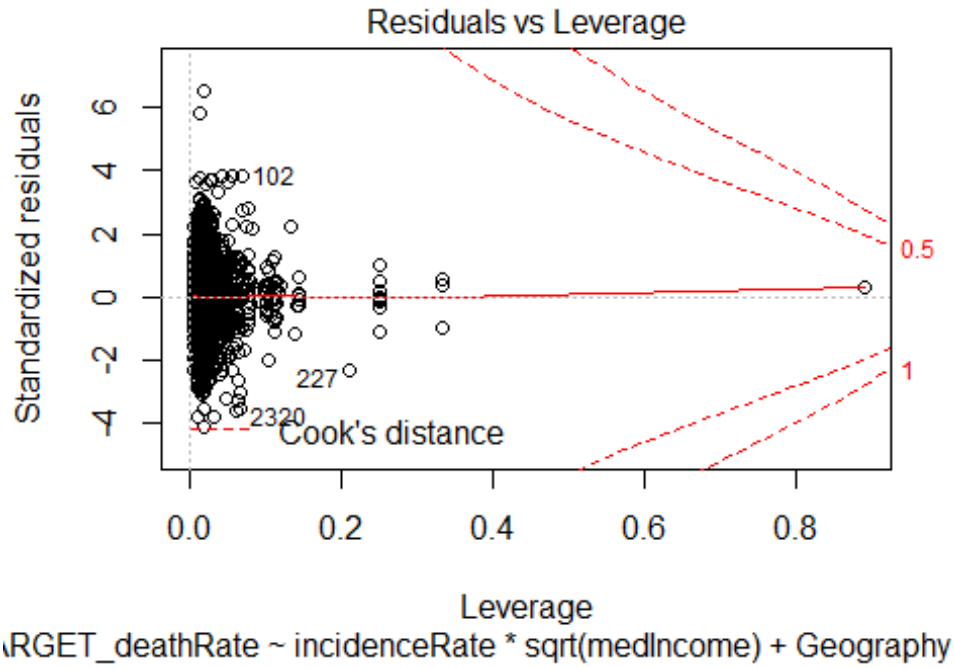
From the Residuals vs Fitted plot, it can be seen that the plot is a straight line. The earlier plot (without significant terms) had a slight curvature, which is not seen in this plot. The reason behind this is that the non-linear and interaction terms introduced above are able to fit the model slightly better and therefore have reduced non-linearity in the residual terms.



From the Normal Q-Q plot, most of the data is fairly normal. However, points 1053 and 1177 may not exhibit normal behavior and may be removed if they are within the Cook's distance lines.



The scale-location plot is nearly a straight line. Also, the points appear to be uniformly distributed and there is no significant pattern across the graph. Therefore, there is uniform variance of residuals (homoscedasticity is satisfied).



RGET_deathRate ~ incidenceRate * sqrt(medIncome) + Geography +

From the Residuals vs Leverage plot, it is seen that all the points are within Cook's distance lines. Therefore, any outliers can be removed without significantly affecting the model behavior.

3. KNN

Splitting training data into train and validation set

Code:

```
Train.data=read.csv("CancerData.csv",header=T)
Test.data=read.csv("CancerHoldoutData.csv",header=T)
library(ISLR)
library(car)
library(class)
library(FNN)

attach(Train.data)

#Dropping PctSomeCol18_24 due to lack of data
Train.data$PctSomeCol18_24=NULL
Test.data$PctSomeCol18_24=NULL

#Creating training data and testing data with all predictors except geography
and PctSomeHS18_24
attach(Train.data)

#Dividing data into training and testing
set.seed(1)
ran=sample(1:nrow(Train.data),round(0.7*nrow(Train.data)))
```

```
train.df=Train.data[ran,]
train.drate=train.df$TARGET_deathRate

test.df=Train.data[-ran,]
test.drate=test.df$TARGET_deathRate

#Removing TARGET_deathRate and geography data
train.x=data.frame(train.df[,-c(1,8)])
test.x=data.frame(test.df[,-c(1,8)])
```

Explanation:

Firstly, PctSomeCol18_24 column was dropped from the analysis. Sample function is used to divide the Train.data into 70% training and 30% testing data. Round function is used in the sample function to round off the 0.7*nrow to nearest integer.

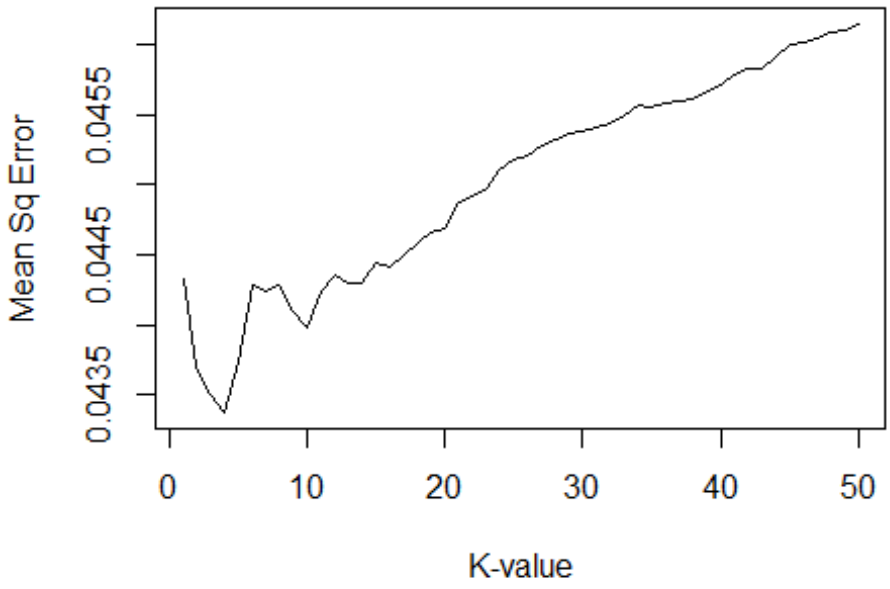
Next, deathrate data and geography data were removed from the training and testing dataframe. Deathrate was stored in a separate variable train and test drate. Geography was removed because KNN cannot accept categorical data.

KNN model development

Code:

```
#Creating a normalize function in r
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}


#Normalizing training and testing data
j=1
for (j in 1:19)
{
  train.x[,j]=normalize(train.x[,j])
  test.x[,j]=normalize(test.x[,j])
}
train.drate=normalize(train.drate)
test.drate=normalize(test.drate)

#KNN for values of k from 1 to 50
set.seed(1)
i=1
mse=matrix(,nrow=50,ncol=2)
for (i in 1:nrow(mse))
{
  knn.pred=knn.reg(train.x,test.x,train.drate,k=i)
```

```
  mse[i,1]=mean((test.drate-knn.pred$pred)^2)
  mse[i,2]=i
}
plot(mse[,2],mse[,1],type="l",xlab="K-value",ylab="Mean Sq Error")
```




```
min(mse[,1])

## [1] 0.04337501

head(mse)

##            [,1] [,2]
## [1,] 0.04433045    1
## [2,] 0.04371169    2
## [3,] 0.04351054    3
## [4,] 0.04337501    4
## [5,] 0.04373963    5
## [6,] 0.04428043    6
```

Explanation:

In techniques like KNN, data needs to be normalized first. The reason behind this is that some data like medIncome will have values like 50,000 whereas data like PctPrivateCoverage will have a maximum value of 100. Since KNN calculates distance to predict values, this would create imbalance. Therefore, entire training and testing dataset is first normalized. This is done after splitting the data to ensure that the model, in no way interacts with the testing data (called data leakage).

Once normalized, knn.reg is used to fit KNN for regression. K values from 1 to 50 are tested and mean squared error (MSE) is calculated for each k value prediction. The minimum value of MSE

is 0.043 and the corresponding K value is 4. This is the optimal value of K. Some other values of MSE and their corresponding K values are shown using head(mse)

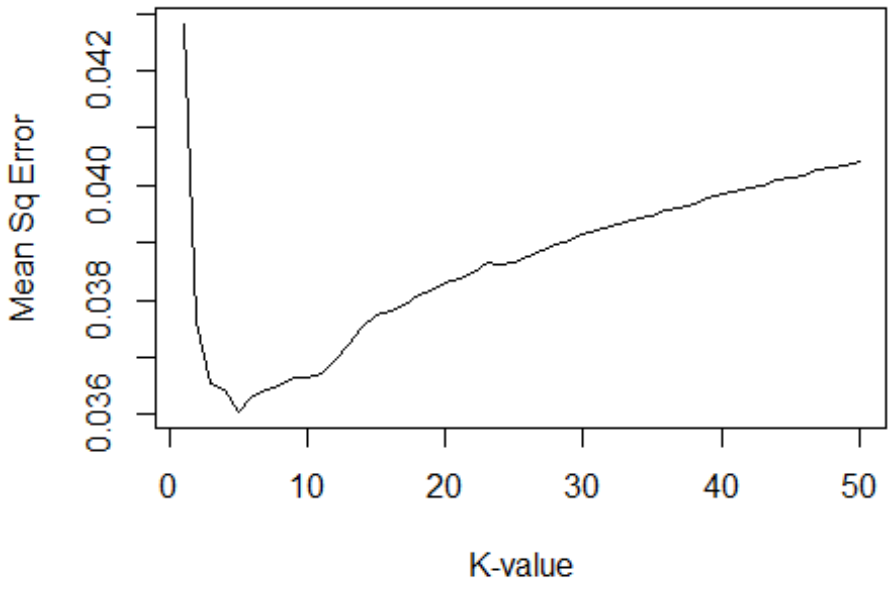
KNN model with important predictors only

Explanation:
Dataframe is created with only significant predictors as per linear regression results. Geography is not included.

Code:

```
attach(Train.data)

#Dataframe is created with only significant predictors as per linear regressi
on results. Geography is not included

df=data.frame(TARGET_deathRate,incidenceRate,PctHS18_24,PctBachDeg18_24,PctPu
blicCoverage,PctWhite,PctBlack,PctOtherRace,PctMarriedHouseholds)

#Dividing data into training and testing
set.seed(1)
ran=sample(1:nrow(df),round(0.7*nrow(df)))
train.df=df[ran,]
train.drate=train.df$TARGET_deathRate
train.drate=normalize(train.drate)
test.df=df[-ran,]
test.drate=test.df$TARGET_deathRate
test.drate=normalize(test.drate)


#Removing TARGET_deathRate from dataframe
train.x=data.frame(train.df[,-c(1)])
test.x=data.frame(test.df[,-c(1)])


#Normalizing training and testing data
j=1
for (j in 1:8)
{
  train.x[,j]=normalize(train.x[,j])
  test.x[,j]=normalize(test.x[,j])
}


#KNN for values of k from 1 to 50
set.seed(1)
i=1
mse=matrix(,nrow=50,ncol=2)
```

```
for (i in 1:nrow(mse))
{
  knn.pred=knn.reg(train.x,test.x,train.drate,k=i)


  mse[i,1]=mean((test.drate-knn.pred$pred)^2)
  mse[i,2]=i
}
plot(mse[,2],mse[,1],type="l",xlab="K-value",ylab="Mean Sq Error")
```




```
min(mse[,1])

## [1] 0.0360561

head(mse)

##            [,1] [,2]
## [1,] 0.04281496    1
## [2,] 0.03765278    2
## [3,] 0.03652809    3
## [4,] 0.03642610    4
## [5,] 0.03605610    5
## [6,] 0.03631834    6
```

Explanation:

After including only significant predictors, it can be seen that the MSE value decreases to 0.036 from 0.043 (when all predictors were used). This proves the point that KNN works better with lower dimensional data.

4. Feature Selection:

(*Interpretation made from linear regression model on page 24)*
From the analysis of death rates due to cancer in various counties, several inferences can be made. The most significant factors affecting mortality rates are:

- Incidence Rate
- Geography
- High School education
- Bachelor's degree
- % Public Coverage
- % White population
- % Black population
- % Other race
- % Married Households

An obvious conclusion is that Incidence Rate is a significant factor related to cancer mortality. High occurrence rates in counties result in higher mortalities.

A surprising observation is that income is not a significant factor in the causation of death due to cancer. Similarly, poverty % is not a significant factor, implying that counties that have more people living below the poverty line may not necessarily have more deaths due to cancer.

Geography also plays an important role. The base factor in this case is Alabama. All mortality rates are compared with Alabama as base. It is seen that places in the central US (Arkansas, Indiana, Kentucky etc) and Southern US (Louisiana, Mississippi) have higher mortality rates compared to Alabama, whereas states like Arizona, Washington, California etc have lesser mortality rates. This may be due to the healthcare infrastructure of these states. Another reason may be the presence of carcinogenic substances in those counties (sunlight intensity, water/food contaminants, etc.). Lifestyle of people also plays an important role. For instance, obesity in the Southern states may be an influencing factor resulting in higher mortality rates due to cancer in that region.

Interestingly, counties with high % of High School education have a positive correlation with mortality rate due to cancer. This may be due to the fact that most of the people have some high school education, so the chances of a person who dies due to cancer and went to high school are high

Counties with more educated population (Bachelor degree) have lower mortality rates. This may be due to awareness among population about cancer and its available treatment, better lifestyle etc.

It is seen that counties with more % of people having government provided health coverage have higher mortality rates. This may be due to a poor insurance plan with higher co-payments or deductibles, which may lead to patients getting no/inadequate treatments.

Also, counties with higher % of black and white people have higher mortality rates. This may be due to the fact that these two ethnicities constitute a major part of the US demographic.

Another observation is that counties with higher % of married households has less mortality rates. This may be attributed to the fact that the inflicted person may have additional psychological and emotional support from spouse or children, and he/she would be more regular with timely health checkups.

5. Performance on Holdout data

Summarize and compare the model performance (MSE) of LR and KNN on holdout dataset as a table.

Linear Regression Code (With Geography data):

```
Train.data=read.csv("CancerData.csv",header=T)
Test.data=read.csv("CancerHoldoutData.csv",header=T)
library(ISLR)
library(car)

library(class)
library(FNN)

attach(Train.data)

#Dropping PctSomeCol18_24 due to lack of data
Train.data$PctSomeCol18_24=NULL
Test.data$PctSomeCol18_24=NULL

#Converting geography county data into state data to better interpret results
Train.data$Geography=sub(".*, ", "", Train.data$Geography)
Train.data$Geography=as.factor(Train.data$Geography)
Test.data$Geography=sub(".*, ", "", Test.data$Geography)
Test.data$Geography=as.factor(Test.data$Geography)
i=1
for (i in 1:2590)
{
  if (Train.data$MedianAge[i]>130)
  {
    Train.data$MedianAge[i]=(Train.data$MedianAgeMale[i]+Train.data$MedianAge
Female[i])/2
  }
}
```

```
i=1
for (i in 1:457)
{
  if (Test.data$MedianAge[i]>130)
  {
    Test.data$MedianAge[i]=(Test.data$MedianAgeMale[i]+Test.data$MedianAgeFem
ale[i])/2
  }
}

attach(Train.data)

linear.fit=lm(TARGET_deathRate~.-TARGET_deathRate,data=Train.data)
summary(linear.fit)

##
## Call:
## lm(formula = TARGET_deathRate ~ . - TARGET_deathRate, data = Train.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.69 -10.50  -0.40  10.32 118.64
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.143e+02  1.598e+01   7.157 1.08e-12 ***
## incidenceRate                   2.007e-01  8.415e-03  23.854  < 2e-16 ***
## medIncome                      -1.481e-04  8.196e-05  -1.807 0.070893 .
## povertyPercent                  1.635e-01  1.658e-01   0.986 0.324019
## MedianAge                      -5.325e-01  1.104e+00  -0.482 0.629526
## MedianAgeMale                   2.293e-01  6.455e-01   0.355 0.722479
## MedianAgeFemale                -3.486e-02  5.428e-01  -0.064 0.948792
## GeographyAlaska                 8.255e+00  5.996e+00   1.377 0.168717
## GeographyArizona               -2.507e+01  6.708e+00  -3.738 0.000190 ***
## GeographyArkansas               7.115e+00  3.597e+00   1.978 0.048054 *
## GeographyCalifornia            -1.985e+01  4.292e+00  -4.625 3.94e-06 ***
## GeographyColorado              -2.734e+01  4.013e+00  -6.814 1.19e-11 ***
## GeographyConnecticut           -1.826e+01  7.750e+00  -2.356 0.018543 *
## GeographyDelaware              -1.411e+01  1.110e+01  -1.271 0.203695
## GeographyDistrict of Columbia -4.963e+00  1.900e+01  -0.261 0.793919
## GeographyFlorida               -6.565e+00  3.772e+00  -1.741 0.081874 .
## GeographyGeorgia               -8.206e+00  3.141e+00  -2.612 0.009048 **
## GeographyHawaii                -3.688e+01  1.197e+01  -3.081 0.002086 **
## GeographyIdaho                 -2.715e+01  4.085e+00  -6.645 3.70e-11 ***
## GeographyIllinois              -6.694e+00  3.526e+00  -1.899 0.057720 .
## GeographyIndiana                7.732e+00  3.598e+00   2.149 0.031746 *
## GeographyIowa                  -1.631e+01  3.550e+00  -4.595 4.55e-06 ***
## GeographyKansas                -9.066e+00  3.463e+00  -2.618 0.008906 **
## GeographyKentucky               8.117e+00  3.468e+00   2.340 0.019340 *
## GeographyLouisiana             -1.485e+00  3.649e+00  -0.407 0.684017
```

```
## GeographyMaine                -8.948e+00  5.959e+00  -1.502 0.133320
## GeographyMaryland             -1.114e+00  4.978e+00  -0.224 0.822900
## GeographyMassachusetts        -1.535e+01  7.087e+00  -2.166 0.030395 *
## GeographyMichigan             -6.264e+00  3.613e+00  -1.734 0.083126 .
## GeographyMinnesota            -1.840e+01  3.678e+00  -5.005 5.99e-07 ***
## GeographyMississippi           1.463e+00  3.503e+00   0.418 0.676223
## GeographyMissouri              4.914e+00  3.355e+00   1.465 0.143144
## GeographyMontana              -2.279e+01  4.165e+00  -5.473 4.87e-08 ***
## GeographyNebraska             -1.071e+01  3.663e+00  -2.924 0.003491 **
## GeographyNevada               -2.800e+00  5.585e+00  -0.501 0.616137
## GeographyNew Hampshire        -1.068e+01  6.607e+00  -1.616 0.106252
## GeographyNew Jersey           -8.205e+00  5.195e+00  -1.579 0.114396
## GeographyNew Mexico           -2.185e+01  4.845e+00  -4.510 6.77e-06 ***
## GeographyNew York             -1.779e+01  3.991e+00  -4.459 8.61e-06 ***
## GeographyNorth Carolina       -1.202e+01  3.346e+00  -3.593 0.000334 ***
## GeographyNorth Dakota         -9.579e+00  4.086e+00  -2.344 0.019134 *
## GeographyOhio                  2.658e+00  3.536e+00   0.752 0.452360
## GeographyOklahoma              5.642e+00  3.741e+00   1.508 0.131616
## GeographyOregon               -1.417e+01  4.516e+00  -3.138 0.001719 **
## GeographyPennsylvania         -1.163e+01  3.759e+00  -3.093 0.002001 **
## GeographyRhode Island         -9.757e+00  9.818e+00  -0.994 0.320442
## GeographySouth Carolina       -3.633e+00  4.056e+00  -0.896 0.370421
## GeographySouth Dakota         -1.819e+01  3.962e+00  -4.591 4.62e-06 ***
## GeographyTennessee             3.683e+00  3.443e+00   1.070 0.284898
## GeographyTexas                -3.073e+00  3.215e+00  -0.956 0.339227
## GeographyUtah                 -2.577e+01  4.892e+00  -5.268 1.50e-07 ***
## GeographyVermont              -1.499e+01  7.051e+00  -2.126 0.033634 *
## GeographyVirginia              5.282e+00  3.255e+00   1.623 0.104745
## GeographyWashington           -1.674e+01  4.315e+00  -3.879 0.000108 ***
## GeographyWest Virginia         9.441e-01  3.936e+00   0.240 0.810463
## GeographyWisconsin            -9.484e+00  3.753e+00  -2.527 0.011565 *
## GeographyWyoming              -9.611e+00  5.106e+00  -1.882 0.059932 .
## AvgHouseholdSize               1.265e+00  1.108e+00   1.142 0.253605
## PercentMarried                 1.946e-01  1.686e-01   1.155 0.248403
## PctNoHS18_24                   5.314e-02  6.047e-02   0.879 0.379597
## PctHS18_24                     2.654e-01  5.104e-02   5.200 2.15e-07 ***
## PctBachDeg18_24               -4.694e-01  1.124e-01  -4.176 3.07e-05 ***
## PctPrivateCoverage            -5.566e-02  1.349e-01  -0.413 0.679881
## PctPublicCoverage              4.424e-01  2.176e-01   2.034 0.042103 *
## PctPublicCoverageAlone         4.009e-01  2.897e-01   1.384 0.166483
## PctWhite                      -1.692e-01  7.315e-02  -2.313 0.020791 *
## PctBlack                      -1.610e-01  7.946e-02  -2.026 0.042830 *
## PctAsian                      -1.633e-01  2.213e-01  -0.738 0.460597
## PctOtherRace                  -6.880e-01  1.362e-01  -5.052 4.67e-07 ***
## PctMarriedHouseholds          -4.788e-01  1.536e-01  -3.117 0.001850 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.52 on 2520 degrees of freedom
```

```
## Multiple R-squared:  0.5721, Adjusted R-squared:  0.5604
## F-statistic: 48.84 on 69 and 2520 DF,  p-value: < 2.2e-16

attach(Test.data)

lm.predict=predict(linear.fit,Test.data,interval="predict")

mean((lm.predict[,1]-Test.data$TARGET_deathRate)^2)

## [1] 340.2016
```

Linear Regression Code without Geography data:

```
Train.data=read.csv("CancerData.csv",header=T)
Test.data=read.csv("CancerHoldoutData.csv",header=T)
library(ISLR)
library(car)

## Loading required package: carData

library(class)
library(FNN)

attach(Train.data)

#Dropping PctSomeCol18_24 due to lack of data
Train.data$PctSomeCol18_24=NULL
Test.data$PctSomeCol18_24=NULL


#Removing geography data
#Train.data=Train.data[,-c(8)]
attach(Train.data)

linear.fit=lm(TARGET_deathRate~.-TARGET_deathRate,data=Train.data)
summary(linear.fit)

##
## Call:
## lm(formula = TARGET_deathRate ~ . - TARGET_deathRate, data = Train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.299 -12.148  -0.113  11.640 127.367
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.060e+02  1.423e+01   7.449 1.27e-13 ***
## incidenceRate           2.181e-01  8.246e-03  26.444  < 2e-16 ***
## medIncome              -2.647e-04  7.983e-05  -3.316 0.000927 ***
## povertyPercent          3.032e-01  1.702e-01   1.781 0.074975 .
## MedianAge              -4.836e-01  1.198e+00  -0.404 0.686468
```

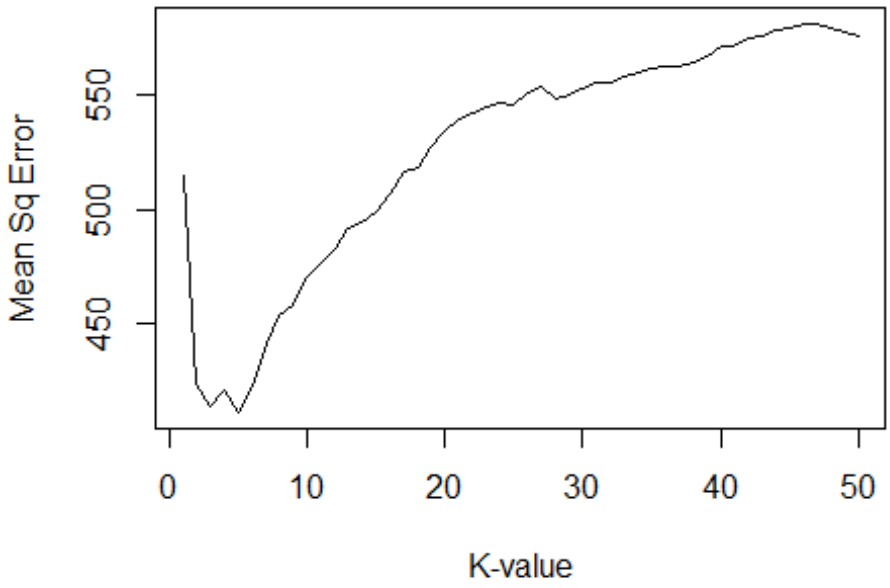
```
## MedianAgeMale            6.259e-02  6.993e-01   0.090 0.928688
## MedianAgeFemale          7.985e-02  5.861e-01   0.136 0.891647
## AvgHouseholdSize         6.479e-01  1.204e+00   0.538 0.590540
## PercentMarried           1.913e-01  1.614e-01   1.185 0.236097
## PctNoHS18_24            -4.965e-02  6.253e-02  -0.794 0.427284
## PctHS18_24               4.562e-01  5.231e-02   8.721  < 2e-16 ***
## PctBachDeg18_24         -3.489e-01  1.184e-01  -2.947 0.003242 **
## PctPrivateCoverage      -2.791e-01  1.141e-01  -2.447 0.014489 *
## PctPublicCoverage        2.638e-02  2.136e-01   0.123 0.901723
## PctPublicCoverageAlone   5.644e-01  2.781e-01   2.030 0.042463 *
## PctWhite                -4.874e-02  6.359e-02  -0.766 0.443498
## PctBlack                 3.859e-02  6.245e-02   0.618 0.536636
## PctAsian                -2.668e-01  1.990e-01  -1.341 0.180004
## PctOtherRace            -9.974e-01  1.296e-01  -7.699 1.95e-14 ***
## PctMarriedHouseholds    -3.104e-01  1.558e-01  -1.992 0.046428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.36 on 2570 degrees of freedom
## Multiple R-squared:  0.4728, Adjusted R-squared:  0.4689
## F-statistic: 121.3 on 19 and 2570 DF,  p-value: < 2.2e-16
```

```
Test.data=Test.data[,-c(8)]
attach(Test.data)

lm.predict=predict(linear.fit,Test.data,interval="predict")

mean((lm.predict[,1]-Test.data$TARGET_deathRate)^2)

## [1] 414.3202
```

KNN Code:

```
Train.data=read.csv("CancerData.csv",header=T)
Test.data=read.csv("CancerHoldoutData.csv",header=T)
library(ISLR)
library(car)

library(class)
library(FNN)

attach(Train.data)

#Dropping PctSomeCol18_24 due to lack of data
Train.data$PctSomeCol18_24=NULL
Test.data$PctSomeCol18_24=NULL

train.x=Train.data[,-c(8)]
test.x=Test.data[,-c(8)]
```

```
train.drate=train.x$TARGET_deathRate
test.drate=test.x$TARGET_deathRate

set.seed(1)
i=1
mse=matrix(,nrow=50,ncol=2)
for (i in 1:nrow(mse))
{
  knn.pred=knn.reg(train.x,test.x,train.drate,k=i)


  mse[i,1]=mean((test.drate-knn.pred$pred)^2)
  mse[i,2]=i
}
plot(mse[,2],mse[,1],type="l",xlab="K-value",ylab="Mean Sq Error")
```




```
min(mse[,1])

## [1] 411.4077

head(mse)

##          [,1] [,2]
## [1,] 514.7173    1
## [2,] 424.0900    2
## [3,] 414.3958    3
## [4,] 421.1690    4
```

```
## [5,] 411.4077    5
## [6,] 422.5839    6
```

Explanation:

Method	**Mean Squared Error**
Linear Regression (with Geography data)	**340.2**
Linear Regression (without Geography data)	414.3
KNN	411.4

Here, LR is performed with and without geography data. Also, data for KNN is not normalized so that it can be compared directly with results of Linear Regression.

It is seen that with geographic data, LR performs better than KNN as it has less MSE. But without Geography data, it is difficult to comment on the performance of models since both have nearly identical MSE values.

Reason that LR performs well with Geographic data is that dimensionality of data increases by including geography and KNN is not particularly well-suited for higher dimensional data