Modeling Cancer Death Rate

Team: CTRL + ALT + DEL
Members: Jeremiah Joseph, Samuel La, Jay Shah

Introduction:

The goal of the analysis is to determine which factors are strong predictors of cancer death rate. The dataset chosen was extracted from the “data.world” website, which were aggregated from census.gov, clinicaltrials.gov, and cancer.gov; it contains 3047 observations/rows and 34 variables, covering multiple subjects such as educational, employment, health, economic, geographical, and racial information for each county. Of the 34 variables, TARGET_deathRate, avgAnnCount, avgDeathsPerYear, and incidenceRate relies on the years 2010 to 2016; the rest is based on the 2013 Census Estimates.

Data Description:

Our dataset contained many demographic metrics, most of which measured economic, educational, insurance and health information for each county in the U.S. as well as the cancer death rate which was our dependent variable.

The raw variables in the dataset were listed and defined as:

TARGET_deathRate: The dependent variable, number of cancer death per capita (per 100,000 residents) in the county
avgAnnCount - Mean number of cancer cases reported per year
avgDeathsPerYear - Mean number of cancer death reported per year
incidenceRate - Mean number cancer cases per year per capita (100,000)
medianIncome - Median income per capita per county
popEst2015 - Population of the county in 2015
povertyPercent - Percent of people in the county in poverty
studyPerCap - number of cancer related clinical trials per capita
binnedInc - decile scale from 1 to 10 binning the median income per capita of the county
MedianAge - Median age of the county’s residents
MedianAgeMale - Median Age of the county’s male residents
MedianAgeFemale - Median Age of the county’s female residents
Geography - Name of the county
AvgHouseholdSize - Mean household size in the county

PercentMarried - Percent of married residents in the county
PctNoHS18_24 - Percent of county residents aged 18-24 who have not attained high school diploma or higher
PctHS18_24 - Percent of county residents aged 18-24 whose highest education attained was high school diploma
PctSomeCol18_24 - Percent of county residents aged 18-24 whose highest education attained was some college
PctBachDeg25_Over - Percent of county residents aged 25+ whose highest level of education attained was bachelor's degree
PctEmployed16_Over - Percent of county residents aged 16+ that were employed
PctUnemployed16_Over - Percent of county residents aged 16+ that were unemployed
PctPrivateCoverage - Percent of county residents with private health coverage
PctPrivateCoverageAlone - Percent of county residents with only private health coverage (no public assistance)
PctEmpPrivCoverage - Percent of county residents with employee provided health coverage
PctPublicCoverage - Percent of county residents with government provided health coverage
PctPublicCoverageAlone - Percent of county residents with government provided health coverage only
PctWhite - Percent of county residents that were white
PctBlack - Percent of county residents that were black
PctAsian - Percent of county residents that were asian
PctOtherRace - Percent of residents that were not black, white, or asian
PctMarriedHouseholds - Percent of married households in the county
BirthRate - Number of live births relative to the number of women in the county

Data Cleaning:

In our data cleaning process, most of the data was given in a way that was easy to work with. There were only two main aspects of our data cleaning process. The first was dealing with null values as the second was dealing with the issues in the binned income variable.

First when looking at the NA values, we found that there were three variables that contained null values. The first was the PctSomeCol18_24 variable. There were 2,285 null values of this variable, constituting around 75% of the data. Due to having such a high percentage we decided this variable would not really give us any useful information and decided to delete it. The next variable that had null values was PctPrivateCoverageAlone. This had 609 null values, accounting for around 20% of the observations. Though this was not as high of a percentage as PctSomeCol18_24, we noticed that there were many other similar variables. Due to this we decided it would be best to remove this variable as well. Finally, we had PctEmployed16_Over with null values. This only 152 null values, meaning that the null values only were a very small percentage of the overall data. Due to the fact however that we had the PctUnemployed16_Over with no null values, we decided it did not make sense to keep both variables and decided to get rid of the PctUnemployed16_Over as well.

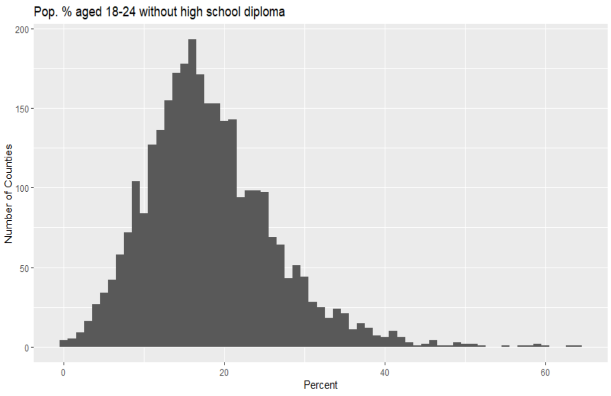
The other aspect of our data cleaning process was making the binnedInc variable easier to use. The variable initially was a character that consisted of a range of incomes encapsulating a certain bin. For example, (48021.6, 51046.4] was one of the ranges. What we did to make this easier to understand and graph was we first turned this into a factor. This then arranged the variables into 10 distinct categories. From there we were able to order the ranges from 1 to 10. This allowed for a much more interpretable variable.

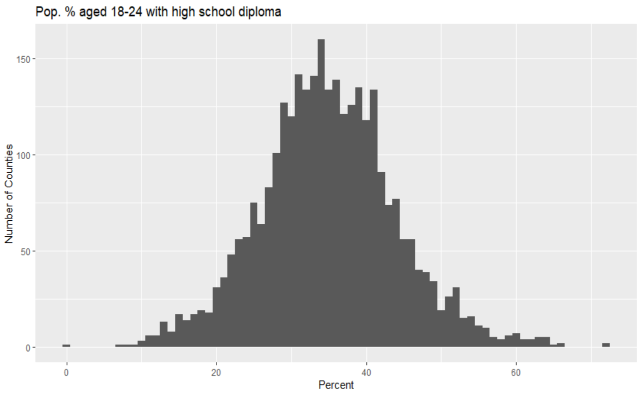
Data Visualization:

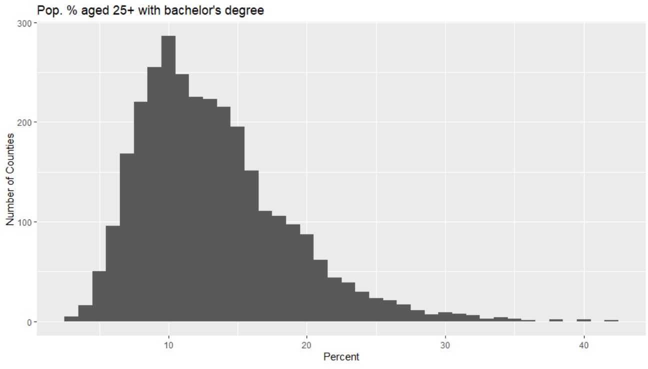
To better understand our data we plotted some of our variables and their relationship to cancer death rate. We have separated a few categories in which we saw patterns in the data.

Educational Data -

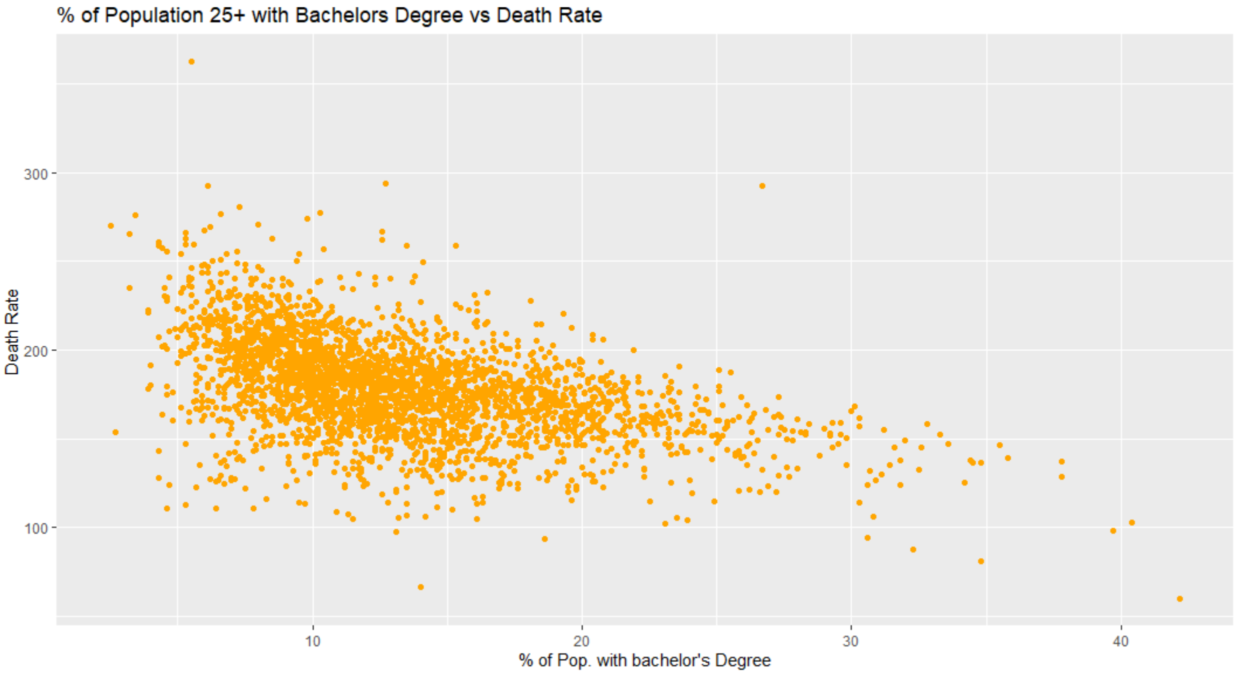
When looking at the educational data, we wanted to see how the education data was generally distributed. Looking at the population aged 18-24 without a highschool diploma histogram, we can see that the counties tend to have low percentages in this category. In contrast, the percentage of population aged between 18-24 with high school diploma (as the highest education) has a bell curve placed more so in the center, which is significantly more numerous than the counties' population without high school diploma (as the highest education). Finally when looking at the population percentage of people aged 25+ with bachelor degree as their highest degree, we can see that most counties have a fairly small percentage of people achieving this with the highest density occurring around the 10-12 percent range.





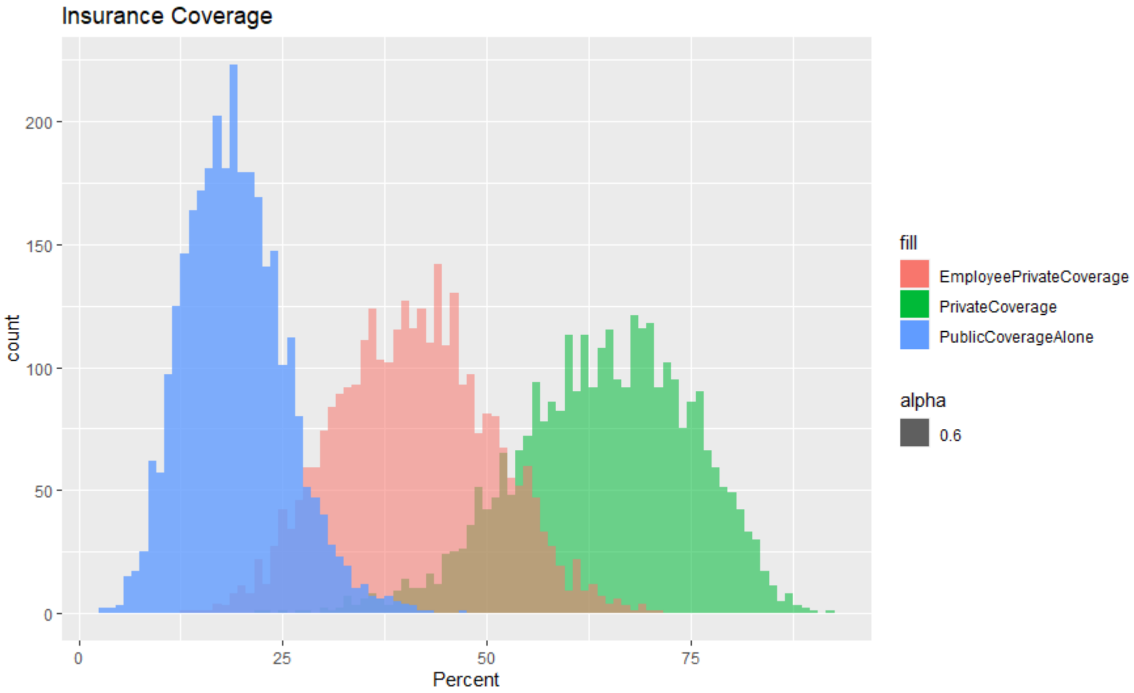


In order to understand how education level affects the overall target death rate we plotted a graph of bachelor's degree 25+ as the highest education with the death rate. This graph shows that as the education level increased the death rate actually tended to go down within the counties.

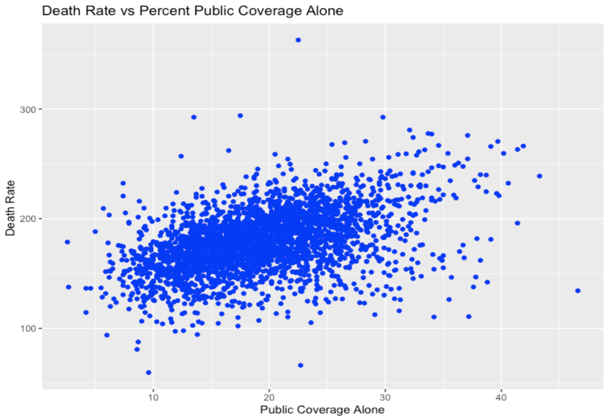


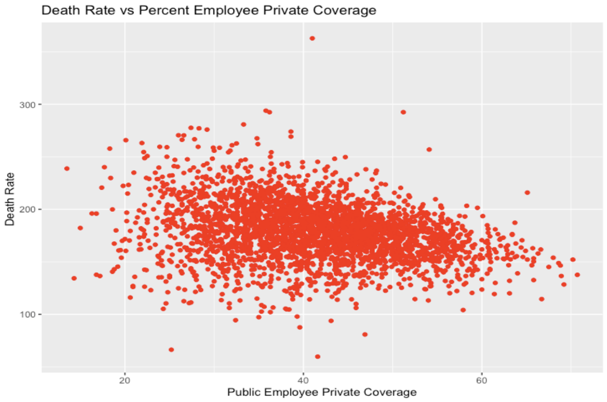
<u>Insurance Data -</u>

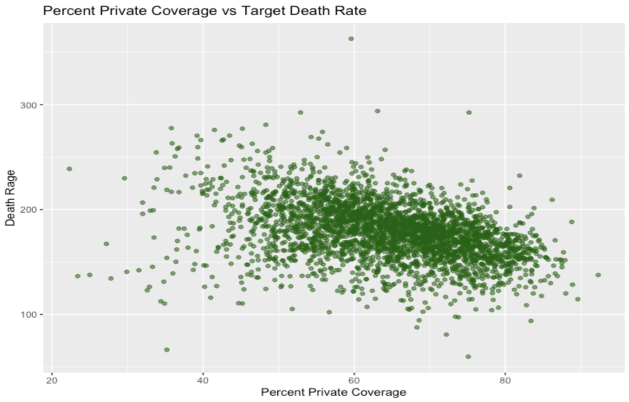
Since Private and public coverage rates are related, we graphed the insurance coverage data variables that we did not remove initially from our model to visualize their relationship.



There was a positive correlation between the death rate and the percent of people on only government health coverage. However there appeared to be a negative correlation between both employer-provided-private coverage and private coverage to cancer death rate.

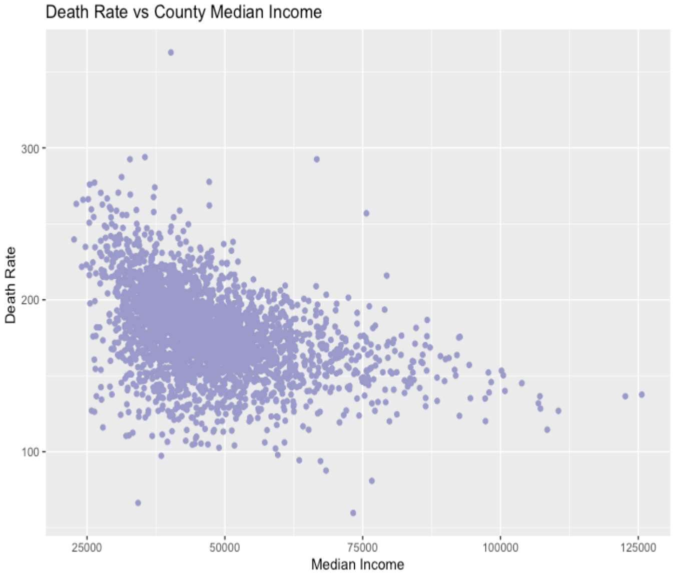

Death Rate vs Percent Public Coverage Alone

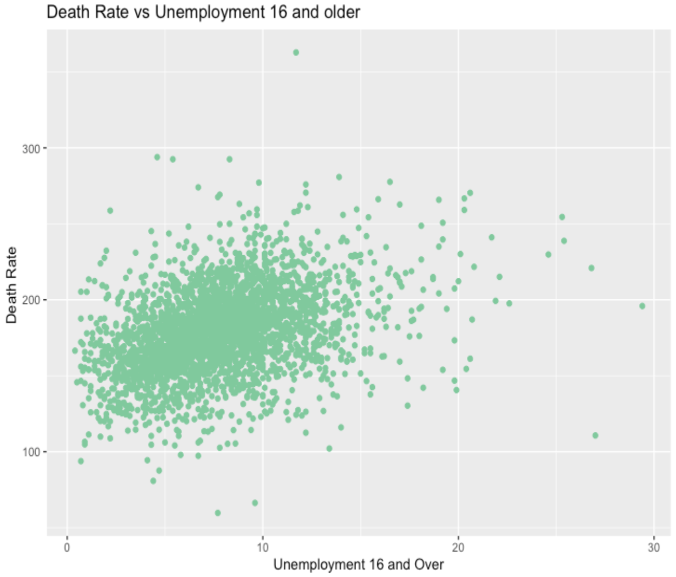

Death Rate vs Percent Employee Private Coverage


Percent Private Coverage vs Target Death Rate

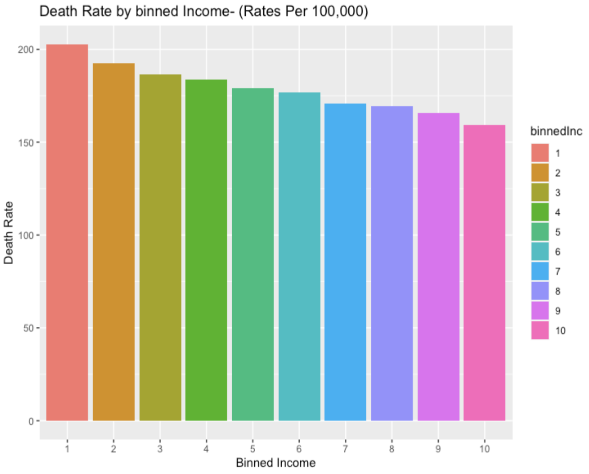
Income Data -

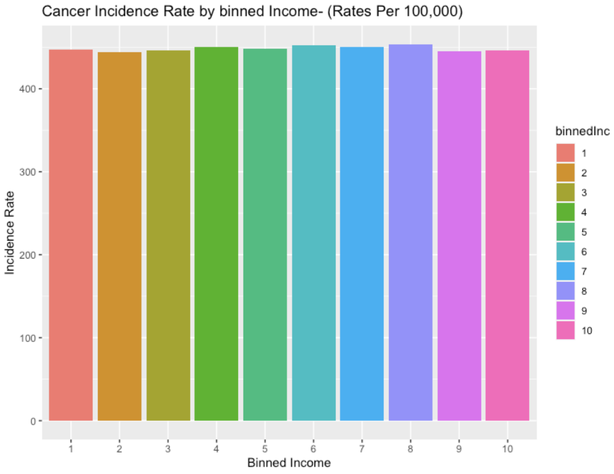
When looking at income we wanted to see how some of the general factors such as median income and unemployment rate affected the cancer death rate. Looking at the graphs we can see that as the median income increased the death rate increased and as the unemployment went up so did the death rate.


Death Rate vs County Median Income


Death Rate vs Unemployment 16 and older

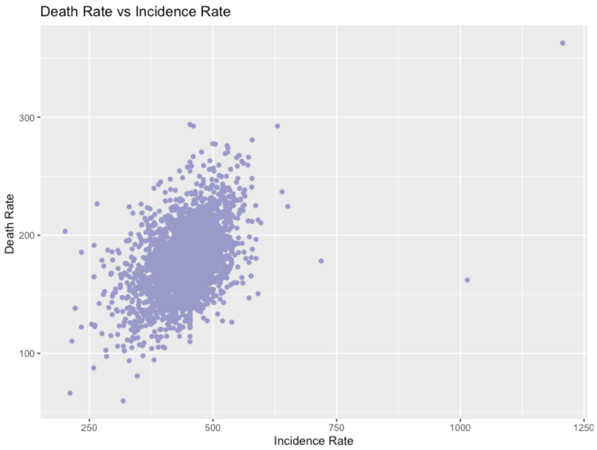
Another thing we looked at is the binned income and how it changed with the death rate. In this we also clearly saw that as income went up the death rate decreased. What was also interesting is that income did not seem to really affect the actual rates of the incidents of cancer. Due to this, we can see despite counties of various incomes having very similar incidents of cancer, the rates of death still changed.

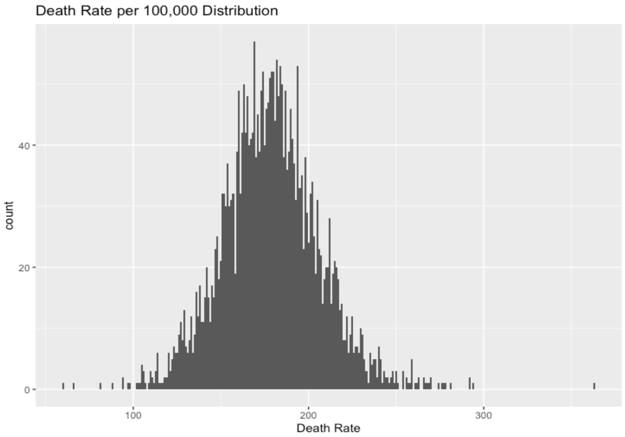




Misc. -

Another relationship observed is Incidence Rate vs Death Rate; the graph shows a positive correlation, where higher death rate occurs as incidence rate increases. Another illustration is the average death rate in each county, and on average, there is approximately 180 deaths per 100k people (for each county):





Variable Selection:

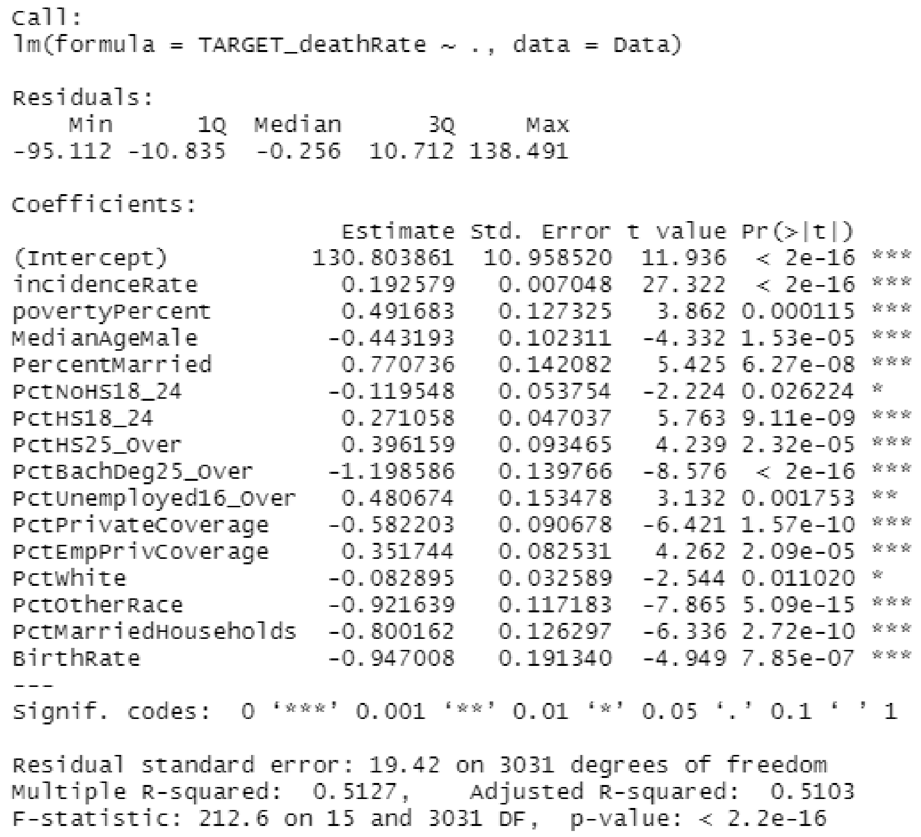
We had two main steps for variable selection. The first step is we took out variables with too many null values, similar definitions, or variables that could not be used for other reasons. The next is we used a backwards stepwise approach to find the rest of our variables.

First let us look at the initial step. In our case the variables with null values we had removed were PctSomeCol18_24, PctPrivateCoverageAlone, and PctUnemployed16_Over as specified in our data cleaning process. The variable that we took out for similar definitions was binned income. We noticed that we had a lot of variables reference to income. The reason why we decided to take out binned income specifically was that this was a factor variable, so we concluded that the other variables would give a more specific picture while also allowing the model to be more interpretable. Finally we took out a few variables that could not be used in our model. The first that followed under this was Geography. This was just a character variable that had the name of the county that would obviously not be useful in our model. We also took out avgDeathsPerYear and popEst2015. This is because this was used in the calculation of the death rate in our model and we concluded that this would thus dominate our model too much. Finally we took out avgAnnCount as we noticed this was too highly dependent on the population and did not really tell us anything constructive.

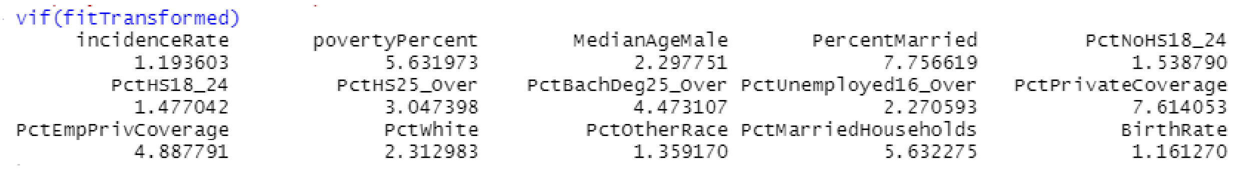
The next part of our variable selection process was the backwards stepwise approach to getting significant variables in our model. The specific approach we did was we initially fit a regression model with all the predictors. We would then take one predictor out of the model with the highest p-value. We would then continue this till we had all p values in the model below .05. The reason why we chose that this was this seemed like a good enough level to indicate that the model was likely being predicted by this variable. By completing this step we were able to get our initial model fit.

Model Fitting:

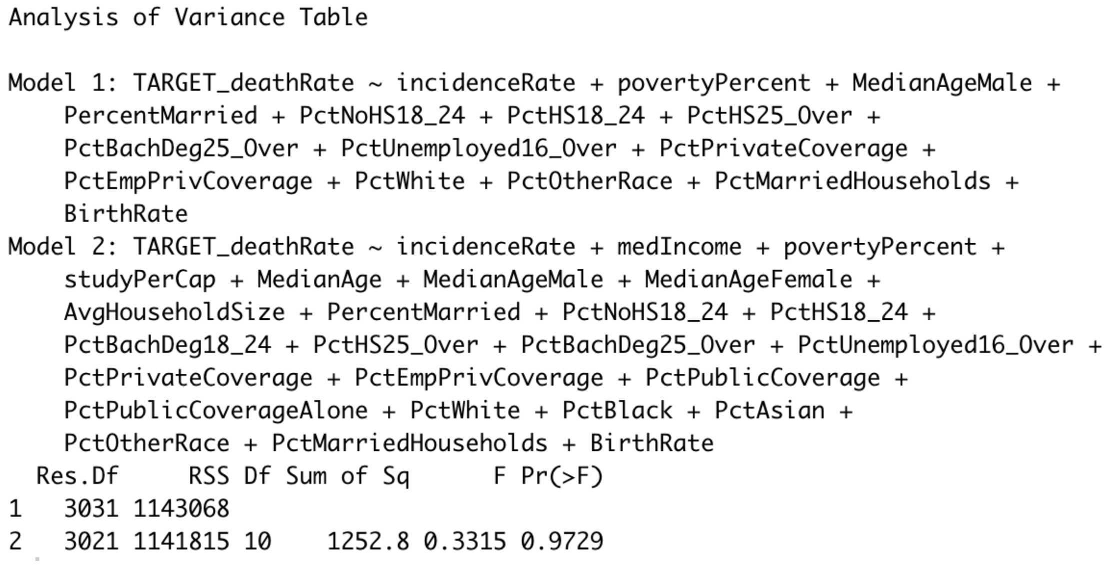
Taking the remaining variables, a linear fit was made in respect to death rate, and it was found that all of the p-values for all the variables had less than 0.05 value. However, we noticed that PctWhite and PctNoHS18_24 were closer to 0.05 value than other variables. Nonetheless, using the Variation Inflation Factor for multicollinearity, all of the VIF values were less than 10; due to this, we decided not to remove any of the variables for multicollinearity. Then, the ANOVA was used to compare the cleaned-model (model 1) versus the raw-model (model 2) and found that the p-value was higher than 0.05 (was shown 0.9729), therefore, the cleaned-model was significantly better.



```
Call:
lm(formula = TARGET_deathRate ~ ., data = Data)

Residuals:
    Min      1Q  Median      3Q     Max 
-95.112 -10.835  -0.256  10.712 138.491 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           130.803861  10.958520  11.936  < 2e-16 ***
incidenceRate           0.192579   0.007048  27.322  < 2e-16 ***
povertyPercent          0.491683   0.127325   3.862 0.000115 ***
MedianAgeMale          -0.443193   0.102311  -4.332 1.53e-05 ***
PercentMarried          0.770736   0.142082   5.425 6.27e-08 ***
PctNoHS18_24           -0.119548   0.053754  -2.224 0.026224 *  
PctHS18_24              0.271058   0.047037   5.763 9.11e-09 ***
PctHS25_Over            0.396159   0.093465   4.239 2.32e-05 ***
PctBachDeg25_Over      -1.198586   0.139766  -8.576  < 2e-16 ***
PctUnemployed16_Over    0.480674   0.153478   3.132 0.001753 ** 
PctPrivateCoverage     -0.582203   0.090678  -6.421 1.57e-10 ***
PctEmpPrivCoverage      0.351744   0.082531   4.262 2.09e-05 ***
PctWhite               -0.082895   0.032589  -2.544 0.011020 *  
PctOtherRace           -0.921639   0.117183  -7.865 5.09e-15 ***
PctMarriedHouseholds   -0.800162   0.126297  -6.336 2.72e-10 ***
BirthRate              -0.947008   0.191340  -4.949 7.85e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.42 on 3031 degrees of freedom
Multiple R-squared:  0.5127,	Adjusted R-squared:  0.5103 
F-statistic: 212.6 on 15 and 3031 DF,  p-value: < 2.2e-16
```



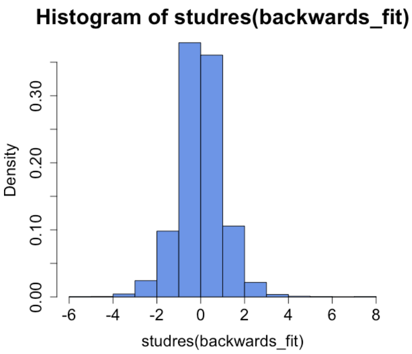
```
vif(fitTransformed)
       incidenceRate       povertyPercent        MedianAgeMale       PercentMarried         PctNoHS18_24 
            1.193603             5.631973             2.297751             7.756619             1.538790 
          PctHS18_24         PctHS25_Over    PctBachDeg25_Over PctUnemployed16_Over   PctPrivateCoverage 
            1.477042             3.047398             4.473107             2.270593             7.614053 
  PctEmpPrivCoverage             PctWhite         PctOtherRace PctMarriedHouseholds            BirthRate 
            4.887791             2.312983             1.359170             5.632275             1.161270 
```

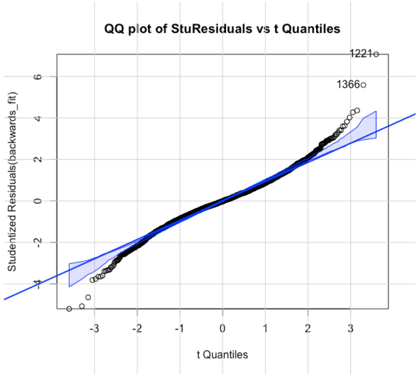


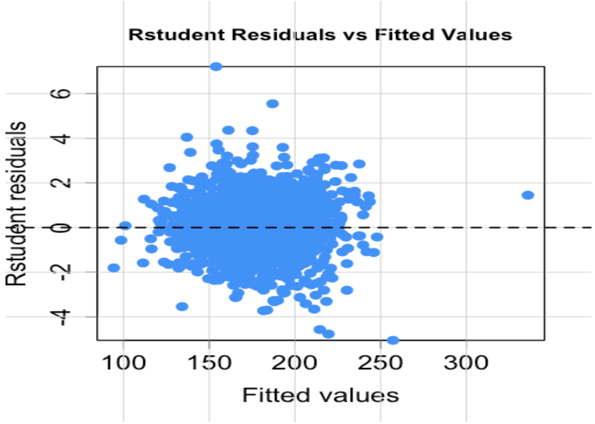
```
Analysis of Variance Table

Model 1: TARGET_deathRate ~ incidenceRate + povertyPercent + MedianAgeMale + 
    PercentMarried + PctNoHS18_24 + PctHS18_24 + PctHS25_Over + 
    PctBachDeg25_Over + PctUnemployed16_Over + PctPrivateCoverage + 
    PctEmpPrivCoverage + PctWhite + PctOtherRace + PctMarriedHouseholds + 
    BirthRate
Model 2: TARGET_deathRate ~ incidenceRate + medIncome + povertyPercent + 
    studyPerCap + MedianAge + MedianAgeMale + MedianAgeFemale + 
    AvgHouseholdSize + PercentMarried + PctNoHS18_24 + PctHS18_24 + 
    PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + PctUnemployed16_Over + 
    PctPrivateCoverage + PctEmpPrivCoverage + PctPublicCoverage + 
    PctPublicCoverageAlone + PctWhite + PctBlack + PctAsian + 
    PctOtherRace + PctMarriedHouseholds + BirthRate
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1   3031 1143068                           
2   3021 1141815 10    1252.8 0.3315 0.9729
```

Residual Analysis:

Looking at the residual graphs we can see that the residuals have an approximately normal distribution. We can see the residuals seem to be randomly distributed. We noticed that there were several leverage points that we addressed later, but the majority of points were centered in the graph.

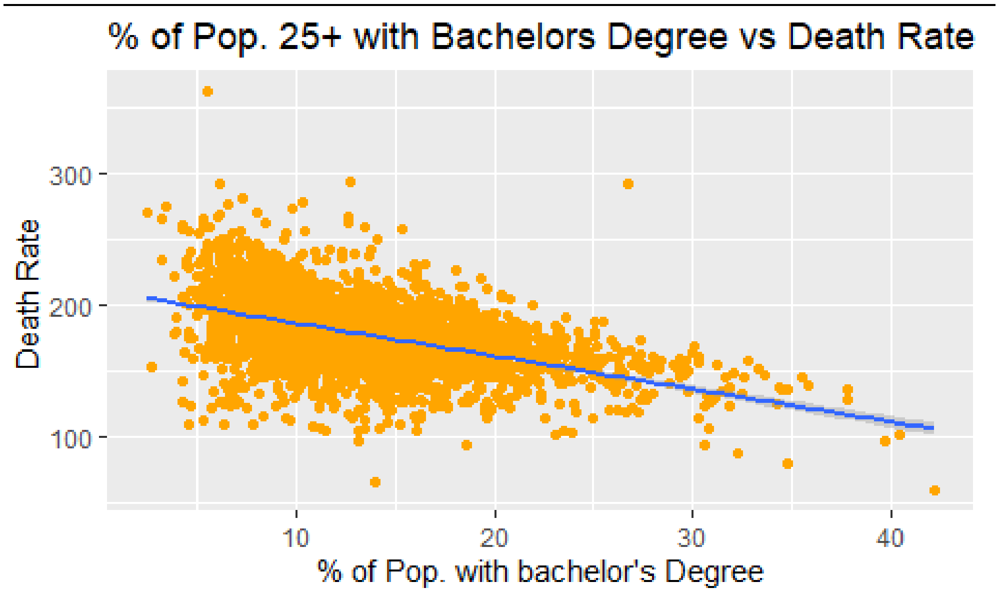






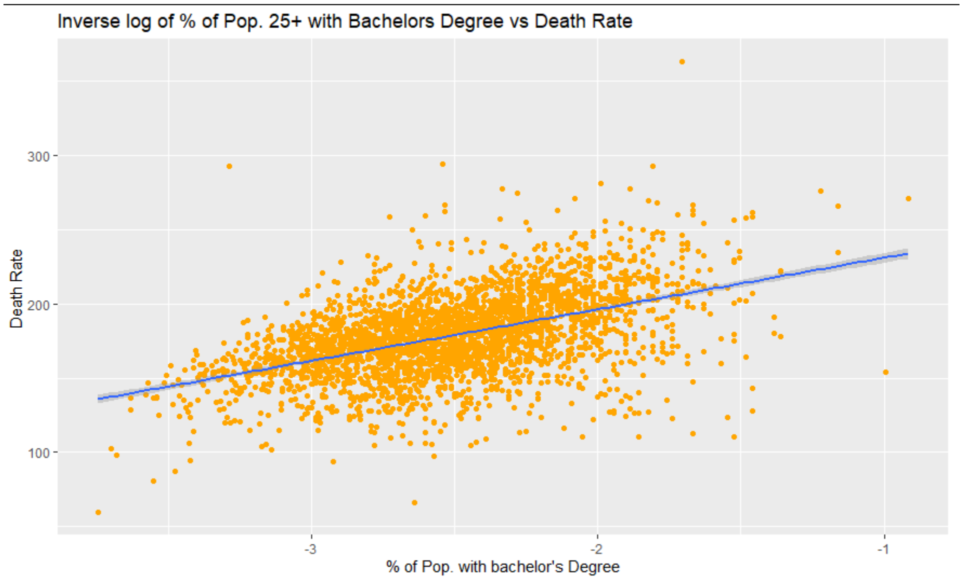
Transformation:

We reviewed our variable's relationships with death rate and found 1 variable that we could apply a transformation to make more linear, which was the percent of the population 25+ with a bachelor's degree vs death rate.



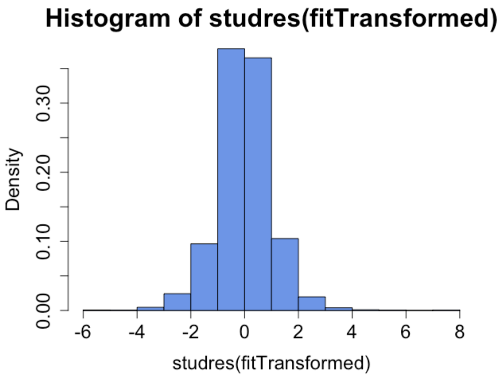
(r^2 = 0.234)

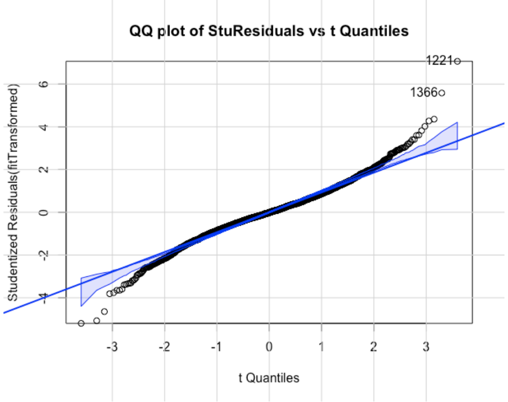
After applying the inverse log transformation, the graph became more linear, as shown below.

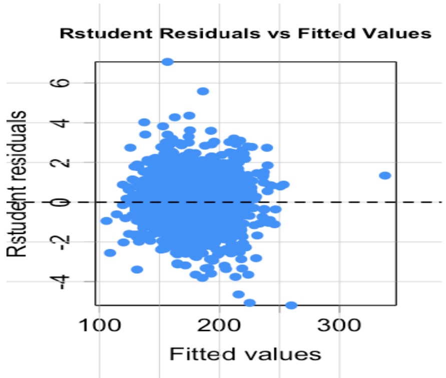


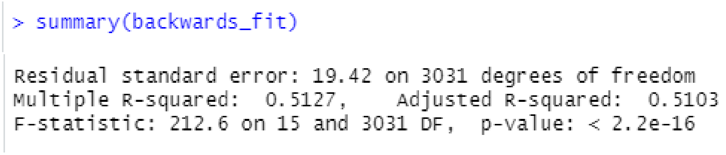
(r^2 = 0.246)

After doing the transformation, we looked at the residuals of the model and the fit summary.

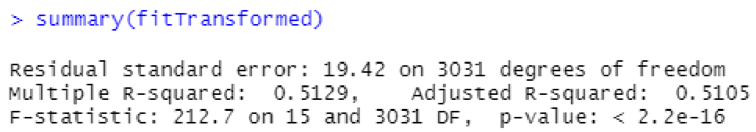








```
> summary(backwards_fit)

Residual standard error: 19.42 on 3031 degrees of freedom
Multiple R-squared:  0.5127,    Adjusted R-squared:  0.5103
F-statistic: 212.6 on 15 and 3031 DF,  p-value: < 2.2e-16
```

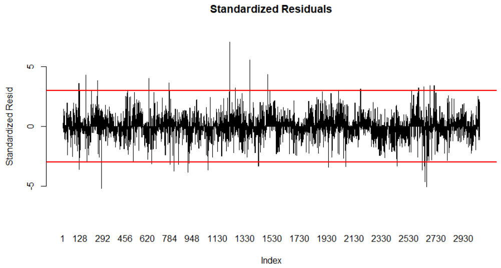


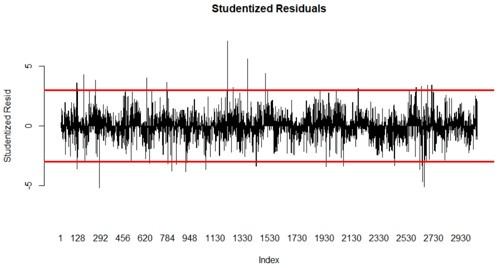
```
> summary(fitTransformed)

Residual standard error: 19.42 on 3031 degrees of freedom
Multiple R-squared:  0.5129,    Adjusted R-squared:  0.5105
F-statistic: 212.7 on 15 and 3031 DF,  p-value: < 2.2e-16
```

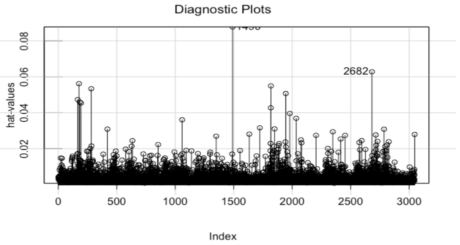
The summary shows that the transformation resulted in a slight increase in multiple and adjusted R-squared values. The residuals in the transformation continue to be approximately normal and randomly distributed.

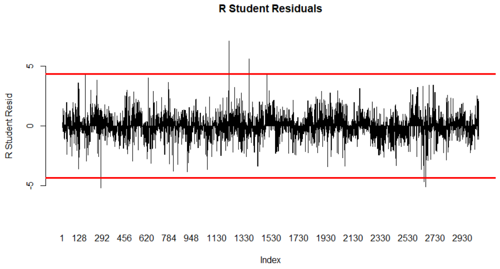
Influential Analysis:

The first part of our influential analysis was to try to tell if we had any possible influential points that were possible outliers and influencing the model heavily. In order to do this we decided to look at the hat values, the standardized residuals, the studentized residuals and the r student residuals. We used 3 as our cutoff point of the red line in our standardized residuals and studentized, and we used a 95% confidence for our r student residuals. Looking at the graphs it was clear that we had many outliers.



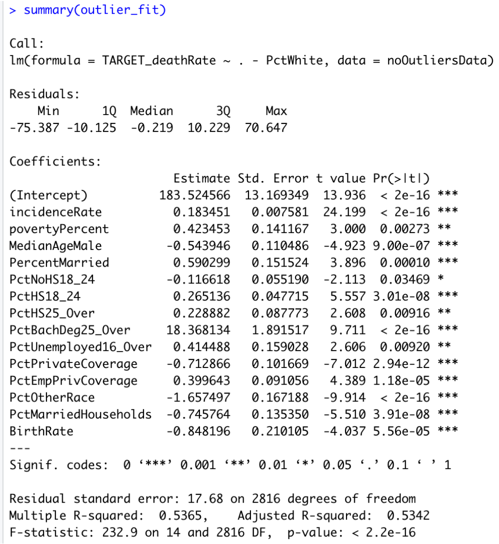




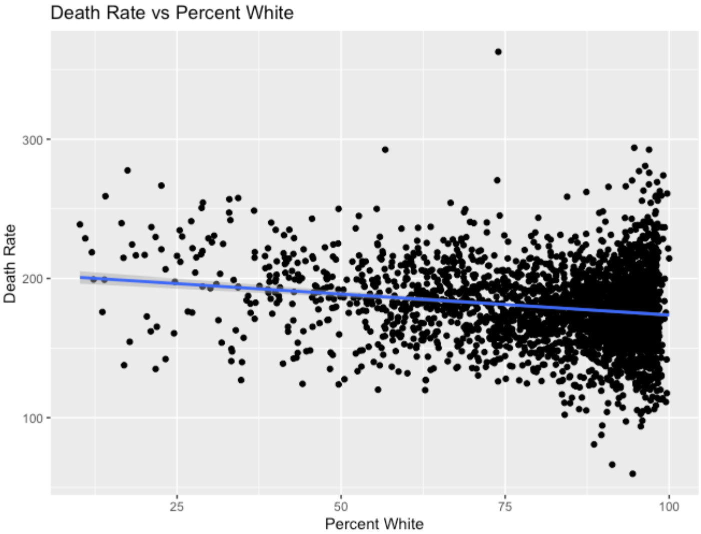


After seeing these graphs we decided to try to identify the potential outliers, remove them from the model the data, and evaluate how the model had changed. Initially we had to identify the possible outliers. To detect outliers we used the criteria that the r student residuals were greater than the 95% range or have hat values greater than the 2p/n standard cutoff range. We chose these two values as we knew these were good at finding outlier points. After doing this we were given 216 outlier points. We then removed them from the data.

After removing the outliers from the data we ran our model with the transformations on the smaller dataset. When doing this we noticed that the percent white variable was no longer significant in our model. When looking at the scatterplot of the death rate vs the percent of white people in a county, we noticed there was basically no correlation between the two points. We then decided that significance that we found in our previous models was likely due to outlier points and we removed it from the model. Below we have the scatterplot mentioned above as well as the summary of the new model. Looking at the summary we can see that both the multiple R-squared and adjusted R-squared values increased, as well as the F statistic from our previous models.

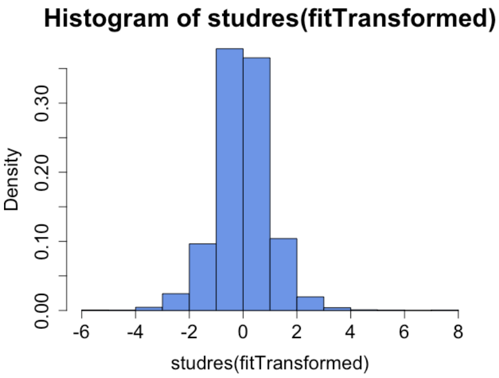


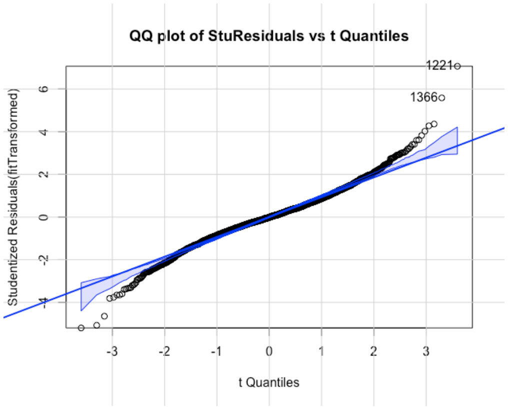
```
> summary(outlier_fit)

Call:
lm(formula = TARGET_deathRate ~ . - PctWhite, data = noOutliersData)

Residuals:
    Min      1Q  Median      3Q     Max 
-75.387 -10.125  -0.219  10.229  70.647 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          183.524566  13.169349  13.936  < 2e-16 ***
incidenceRate          0.183451   0.007581  24.199  < 2e-16 ***
povertyPercent         0.423453   0.141167   3.000  0.00273 ** 
MedianAgeMale         -0.543946   0.110486  -4.923 9.00e-07 ***
PercentMarried         0.590299   0.151524   3.896  0.00010 ***
PctNoHS18_24          -0.116618   0.055190  -2.113  0.03469 *  
PctHS18_24             0.265136   0.047715   5.557 3.01e-08 ***
PctHS25_Over           0.228882   0.087773   2.608  0.00916 ** 
PctBachDeg25_Over     18.368134   1.891517   9.711  < 2e-16 ***
PctUnemployed16_Over   0.414488   0.159028   2.606  0.00920 ** 
PctPrivateCoverage    -0.712866   0.101669  -7.012 2.94e-12 ***
PctEmpPrivCoverage     0.399643   0.091056   4.389 1.18e-05 ***
PctOtherRace          -1.657497   0.167188  -9.914  < 2e-16 ***
PctMarriedHouseholds  -0.745764   0.135350  -5.510 3.91e-08 ***
BirthRate             -0.848196   0.210105  -4.037 5.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.68 on 2816 degrees of freedom
Multiple R-squared:  0.5365,	Adjusted R-squared:  0.5342 
F-statistic: 232.9 on 14 and 2816 DF,  p-value: < 2.2e-16
```

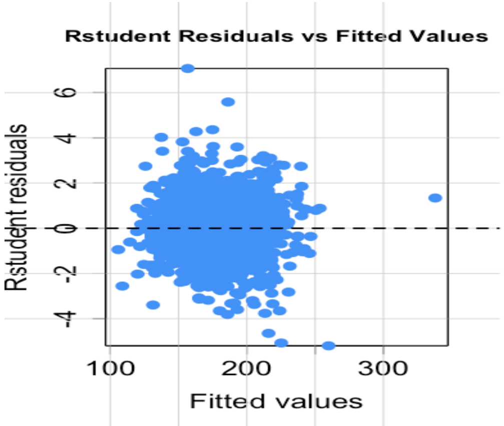


Residual Analysis on Post-Outlier Model -

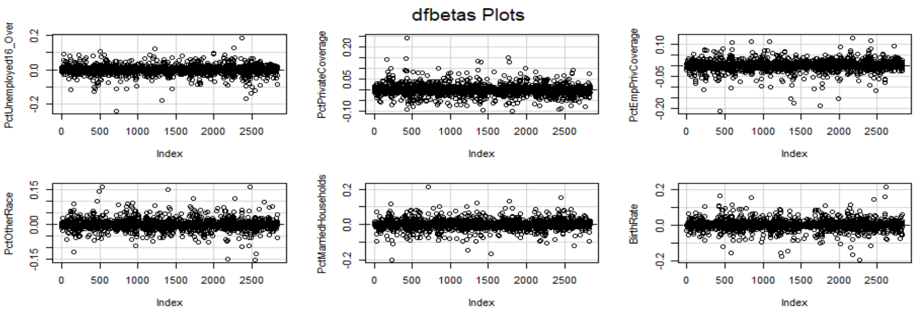
When examining the histograms of the student residuals and the QQ plot of the student residuals vs the student residuals vs the t quantities we can see that there seems to be a little improvement in our post-outlier model. The points seem to follow the diagonal line much more closely and the histogram shows an approximately normal distribution. The r student residuals vs fitted values seem to be more rectangular, indicating randomly distributed residuals.

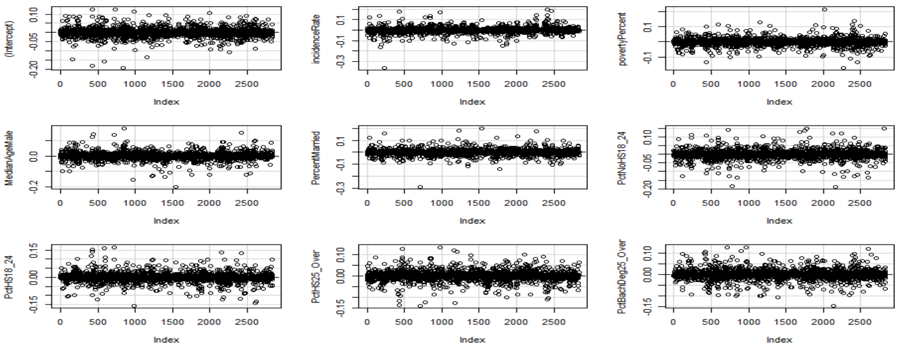


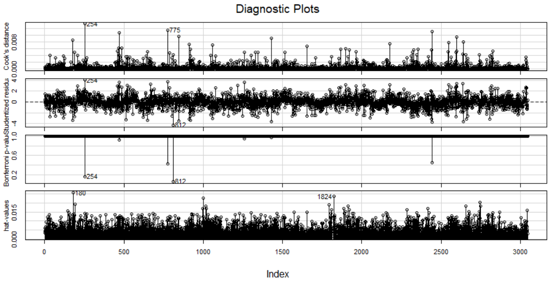




We then decided to look at the diagnostic plots as well as the dfbeta plots. When looking at the diagnostic plots we can see that there were still a few possible influential points, but not too many. Also when looking at the dfbeta plots most tend to be centered at 0, but there are still some that go off. Further research might be able to reveal some answers to why this is the case.

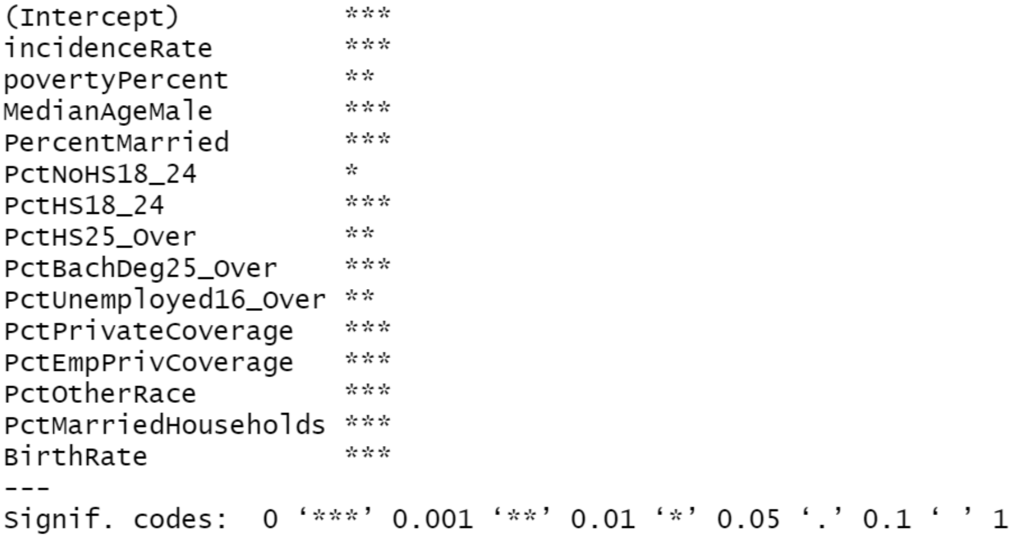






Conclusion:

From our initial data exploration we found that income, education and insurance data are best correlated to the death rate. After all the model analysis steps, we were left with 14 significant remaining predictor variables.



```
(Intercept)           ***
incidenceRate         ***
povertyPercent        **
MedianAgeMale         ***
PercentMarried        ***
PctNoHS18_24          *
PctHS18_24            ***
PctHS25_Over          **
PctBachDeg25_Over     ***
PctUnemployed16_Over **
PctPrivateCoverage    ***
PctEmpPrivCoverage    ***
PctOtherRace          ***
PctMarriedHouseholds ***
BirthRate             ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This indicates that those variables are the best correlated to cancer death rate, and our model suggests that further study could be done on the relationship of those variables and cancer death rate. Another interesting trend we found was that despite median income having a strong correlation with death rate, there was little correlation between median income and incidence rate. Also, we found that most of the demographic variables describing racial data were significant in the model except for PctOtherRace, which the dataset was not clear on the meaning of, so it could be topic for further research on why it affects the cancer death rate. We also noticed that median age of males was a significant variable in our final model, but we found median female age to be non-significant and removed it during our backward selection process. This may be due to the fact that Males have a higher cancer death rate than Females.

Reflection:

Ultimately, the analysis went well; however, there were some limitations that prevented us from accomplishing a comprehensive, accurate understanding on the relationship between the predictors and the target response. The most effective parts of the data analysis were the backwards selection process and the outlier removal as they significantly improved the model and gave us the most insight into which variables were detrimental to the model. However, certain steps of the analysis weren't as effective such as the multicollinearity analysis by looking at the variance inflation factors. Although several of the variables had higher VIF values close to 10, there were none above our threshold, there were several variables in our final model that we suspect could have high multicollinearity, such as percent married individuals and percent married households, or percent of residents aged 18-24 with highest education achieved being high school diploma and percent of residents aged 25+ with highest education achieved high school diploma. Further analysis would be needed to determine if those were skewing the model. Also, another data analysis step that we found ineffective was applying transformations to our variables to improve the linearity of our data and reduce outliers. Due to the large number of data points in our variables, the large number of outliers for most variables meant we couldn't handpick a transformation to address them, and the distribution of most of our data didn't fit any transformation we could apply.

Another limitation of the model was that we were limited to the selection of variables in the dataset, which left out potentially important factors related to cancer prevalence that could have helped us draw more conclusions about cancer death rate. Although there were some health-related variables, such as incidence rate and average deaths per year, there were no specific bodily-health-related variables to relate to the target response, such as obesity rate, smoking rate, alcoholic rate, etc. Furthermore, the dataset we're given was based on a single census year and six years of health statistics; so, a more concrete analysis would've been made had the health statistics recorded for a decade long with two census data being taken both on 2010 and 2020. Specifically in regards to data cleaning, there were variables that had significant amounts of empty or null values that simply could not provide enough data points for analysis; therefore, this whole project was left with fewer variables than we had anticipated. Overall, income data, education data, and insurance data were the better predictors of the target death rate, not the racial data, the marital-related data, geographic data, and so on. A more complete and reliable dataset would be something to aim for.

Appendix:

Team roles -

Jay Shah: Data Visualization, Data Cleaning, Programming, Writing, Variable Analysis

Samuel La: Data Visualization, Programming, Writing

Jeremiah Joseph: Data Visualization, Variable Selection, Influential Analysis, Writing, Programming

References -

Dataset:

https://data.world/nrippner/ols-regression-challenge/workspace/project-summary?agentid=nrippner&datasetid=ols-regression-challenge

Cancer Statistics information:

https://www.cancer.gov/about-cancer/understanding/statistics

R Script-

```
library("ggplot2")
library(car)
library(MASS)



# DATA CLEANING
#-----------------------------------
Data <- read.csv("cancer_reg.csv")
str(Data)

#Finding amount of Na's per column
for (i in colnames(Data)) {
  x <- sum(is.na(Data[,i]))
  print(i)
  print(x)
}

#Getting rid of columns with Na's
Data$PctPrivateCoverageAlone <- NULL
Data$PctEmployed16_Over <- NULL
Data$PctSomeCol18_24 <- NULL

#Separting Income Decile
Data$binnedInc <- as.factor(Data$binnedInc)
str(Data$binnedInc)
levels(Data$binnedInc) <- c("2","3", "4", "5","6","7","8","9","10","1")
Data$binnedInc <- relevel(Data$binnedInc, "1")
str(Data$binnedInc)
```

```
#INITIAL DATA EXPLORATION
#----------------------------------

#Economic Exploration
ggplot ( data = Data) + geom_point(mapping=aes(x =medIncome, y =
TARGET_deathRate),color="#9999CC")   +
  labs(x="Median Income", y= "Death Rate ", title = "Death Rate vs County Median Income")
ggplot ( data = Data ) + geom_point(mapping=aes(x =povertyPercent, y = TARGET_deathRate),
color="#CC6666") +
  labs(x="Poverty Percent", y= "Death Rate ", title = "Death Rate vs County Poverty Percent")
ggplot ( data = Data ) + geom_point(mapping=aes(x =PctUnemployed16_Over, y =
TARGET_deathRate), color="#66CC99") +
  labs(x="Unemployment 16 and Over", y= "Death Rate ", title = "Death Rate vs Unemployment
16 and older")

ggplot(Data, aes(x = binnedInc, y = TARGET_deathRate, fill=binnedInc)) +
  geom_bar(stat = "summary", fun = "mean") + labs(title = "Death Rate by binned Income-
(Rates Per 100,000)", y= "Death Rate", x = "Binned Income")
ggplot(Data, aes(x = binnedInc, y = incidenceRate, fill=binnedInc)) +
  geom_bar(stat = "summary", fun = "mean") + labs(title = "Cancer Incidence Rate by binned
Income- (Rates Per 100,000)", y= "Incidence Rate", x = "Binned Income")


#Health Exploration
ggplot ( data = Data ) + geom_point(mapping=aes(x =incidenceRate, y = TARGET_deathRate),
color="#9999CC") +
  labs(x="Incidence Rate", y= "Death Rate ", title = "Death Rate vs Incidence Rate")
ggplot ( data = Data ) + geom_point(mapping=aes(x =avgAnnCount, y = TARGET_deathRate))
+
  labs(x="Avg Annual Count", y= "Death Rate ", title = "Death Rate vs Avg Annual Count")


#Age Exploration
ggplot ( data = Data ) + geom_point(mapping=aes(x =MedianAge, y = TARGET_deathRate)) +
  labs(x="Median Age", y= "Death Rate ", title = "Death Rate vs Median Age")


#Educational Data overview plots
#PctNoHS18_24, PctBachDeg18_24, PctHS18_24
NoHsPlot <- ggplot(Data, aes(x = PctNoHS18_24)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Pop. % aged 18-24 without high school diploma", x  =  "Percent", y = "Number of
Counties")
```

```
NoHsPlot
HsPlot <- ggplot(Data, aes(x = PctHS18_24)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Pop. % aged 18-24 with high school diploma", x  =  "Percent", y = "Number of
Counties")
HsPlot
BachPlot <- ggplot(Data, aes(x = PctBachDeg25_Over)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Pop. % aged 25+ with bachelor's degree", x  =  "Percent", y = "Number of
Counties")
BachPlot

#Insurance Exploration
ggplot() +
  labs(title = "Insurance Coverage", x = "Percent") +
  geom_histogram(binwidth = 1, data = Data, aes(x=PctPrivateCoverage, fill = 'PrivateCoverage',
alpha = 0.6)) +
  geom_histogram(binwidth = 1, data = Data, aes(x=PctEmpPrivCoverage, fill =
'EmployeePrivateCoverage', alpha = 0.6)) +
  geom_histogram(binwidth = 1, data = Data, aes(x=PctPublicCoverageAlone, fill =
'PublicCoverageAlone', alpha = 0.6)) +
  geom_histogram(binwidth = 1, data = Data, aes(x=PctPublicCoverageAlone, fill =
'PublicCoverageAlone', alpha = 0.6))

ggplot(Data)+
  geom_point(mapping = aes(x = PctPrivateCoverage, y = TARGET_deathRate), color = "dark
green", alpha = 0.6) +
  labs( title = "Percent Private Coverage vs Target Death Rate", x= "Percent Private Coverage",
y="Death Rage")
ggplot ( data = Data ) + geom_point(mapping=aes(x =PctPublicCoverageAlone, y =
TARGET_deathRate), color="blue") +
  labs(x="Public Coverage Alone", y= "Death Rate ", title = "Death Rate vs Percent Public
Coverage Alone")
ggplot ( data = Data ) + geom_point(mapping=aes(x =PctEmpPrivCoverage, y =
TARGET_deathRate), color="red") +
  labs(x="Public Employee Private Coverage", y= "Death Rate ", title = "Death Rate vs Percent
Employee Private Coverage ")

#Overall deathrate Exploration
ggplot() + geom_histogram(binwidth = 1, data = Data, aes(x=TARGET_deathRate)) + labs(title=
"Death Rate per 100,000 Distribution", x= "Death Rate")



#INITIAL MODEL
```

```
#----------------------------------
Data$Geography <- NULL
Data$avgDeathsPerYear <- NULL
Data$binnedInc <- NULL
Data$avgAnnCount <- NULL
Data$popEst2015 <- NULL

#Full Model
full_model <- lm(TARGET_deathRate ~., data = Data)
summary(full_model)

#Backwards Selection
Data$PctPublicCoverageAlone <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data = Data)
summary(backwards_fit)
Data$MedianAgeFemale <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data = Data)
summary(backwards_fit)
Data$studyPerCap <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$MedianAge <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$AvgHouseholdSize <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$PctAsian <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$PctBlack <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$PctPublicCoverage <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$PctBachDeg18_24 <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)
Data$medIncome <- NULL
backwards_fit <- lm(TARGET_deathRate ~., data=Data)
summary(backwards_fit)

#running anova on the model to see if there was improvment
anova(backwards_fit, full_model)
```

```
#preforming vif to look for multicolinearity
vif(backwards_fit)

#Residual Analysis
par(mfrow=c(1,2))
hist(studres(backwards_fit), breaks=10, freq=F, col="cornflowerblue",
    cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(backwards_fit, main= "QQ plot of StuResiduals vs t Quantiles")
par(mfrow=c(1,1))

residualPlot(backwards_fit, type="rstudent", quadratic=F, col = "dodgerblue",
          pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, main="Rstudent Residuals vs Fitted
Values")

#TRANSFORMATION ON MODEL
#---------------------------------

ggplot ( data = Data) + geom_point(mapping=aes(x =incidenceRate, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =povertyPercent, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =MedianAgeMale, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PercentMarried, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctHS25_Over, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctBachDeg25_Over, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =log(1/(PctBachDeg25_Over)), y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctHS18_24, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctUnemployed16_Over, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctPrivateCoverage, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctEmpPrivCoverage, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctWhite, y = TARGET_deathRate)) +
geom_smooth(method = "lm", aes(x = PctWhite, y = TARGET_deathRate)) + labs(title =
"Death Rate vs Percent White", x= "Percent White", y="Death Rate")
ggplot ( data = Data) + geom_point(mapping=aes(x =PctOtherRace, y = TARGET_deathRate)) +
geom_smooth(method = "lm", aes(x = PctOtherRace, y = TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =PctMarriedHouseholds, y =
TARGET_deathRate)) + geom_smooth(method = "lm", aes(x = PctMarriedHouseholds, y =
TARGET_deathRate))
ggplot ( data = Data) + geom_point(mapping=aes(x =BirthRate, y = TARGET_deathRate)) +
geom_smooth(method = "lm", aes(x = BirthRate, y = TARGET_deathRate))
```

```
Data$PctBachDeg25_Over <- log(1/(Data$PctBachDeg25_Over))
fitTransformed <- lm(TARGET_deathRate ~., data=Data)
summary(fitTransformed)

#Code for transformation:
ggplot ( data = Data) + geom_point(mapping=aes(x =log(1/PctBachDeg25_Over), y =
TARGET_deathRate), color = "orange") +
  labs(x="% of Pop. with bachelor's Degree", y= "Death Rate ", title = "Inverse log of % of Pop.
25+ with Bachelors Degree vs Death Rate") +
  geom_smooth(method = "lm", aes(x = log(1/PctBachDeg25_Over), y = TARGET_deathRate))
summary(lm(log(1/Data$PctBachDeg25_Over) ~ Data$TARGET_deathRate))


anova(fitTransformed, backwards_fit)

vif(fitTransformed)

#residual analysis
par(mfrow=c(1,2))
hist(studres(fitTransformed), breaks=10, freq=F, col="cornflowerblue",
     cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(fitTransformed, main= "QQ plot of StuResiduals vs t Quantiles")

par(mfrow = c(1, 1))

residualPlot(fitTransformed, type="rstudent", quadratic=F, col = "dodgerblue",
             pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5, main="Rstudent Residuals vs Fitted
Values")

#INFLUENTIAL ANALYSIS
#-----------------------------------
#Graphing standardized residuals
#Find the range of the values.
range(stdres(fitTransformed))
#Set the range of y axis with argument ylim. Centering to zero is recomended.
barplot(height = stdres(fitTransformed), names.arg = 1:3047,
        main = "Standardized Residuals", xlab = "Index",
        ylab = "Standardized Resid", ylim=c(-8,8))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


#Graphing studentized residuals
```

```
range(studres(fitTransformed))
#Set the range of y axis with argument ylim. Centering to zero is recomended.
barplot(height = studres(fitTransformed), names.arg = 1:3047,
        main = "Studentized Residuals", xlab = "Index",
        ylab = "Studentized Resid", ylim=c(-8,8))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=3)
abline(h=-3, col = "Red", lwd=3)


#Graphing R student residuals
RStudent <- rstudent(fitTransformed)
#Find the range of the values.
range(RStudent)
#Set the range of y axis with argument ylim. Centering to zero is recommended.
barplot(height = RStudent, names.arg = 1:3047,
        main = "R Student Residuals", xlab = "Index",
        ylab = "R Student Resid", ylim=c(-8,8))
cor.level <- 0.05/(2*3047)
cor.qt <- qt(cor.level, 3031, lower.tail=F)
abline(h=cor.qt , col = "Red", lwd=3)
abline(h=-cor.qt , col = "Red", lwd=3)

#influential measures
myInf <- influence.measures(fitTransformed)
summary(myInf)

#plot of hat values
influenceIndexPlot(fitTransformed, vars = c("hat"))

hatCutoff <- (2 * 17) / 3047
outliers_index <- which((RStudent <= -cor.qt | RStudent >= cor.qt) | myInf$infmat[,'hat'] > hatCutoff)




#MODEL POST OUTLIERS
#-----------------------------------
#removing outliers
noOutliersData <- Data[-outliers_index, ]

#fitting model without outliers
outlier_fit <- lm(TARGET_deathRate ~., data=noOutliersData)
```

```
summary(outlier_fit)
outlier_fit <- lm(TARGET_deathRate ~.-PctWhite, data=noOutliersData)
summary(outlier_fit)

#residual analysis
par(mfrow=c(1,2))
hist(studres(outlier_fit), breaks=10, freq=F, col="cornflowerblue",
    cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(outlier_fit)

par(mfrow = c(1, 1))

residualPlot(outlier_fit, type="rstudent", quadratic=F, col = "dodgerblue",
           pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)

#influence graphs
influenceIndexPlot(outlier_fit)
dfbetasPlots(outlier_fit, intercept = T)
```