# stat final project workplace

```
suppressMessages(library(tidyverse))
suppressMessages(library(usmap))
suppressMessages(library(scales))
suppressMessages(library(mice))
suppressMessages(library(glmnet))
```

```
propo = read.csv("cancer_reg.csv")
```

```
testpropo = propo
testpropo <- testpropo %>% mutate(Target_div_Income = TARGET_deathRate/medIncome)
```

```
testpropo1 = cbind(testpropo, str_match(testpropo$Geography,"(.+), (.+)")[ ,-1])
colnames(testpropo1)[37] ="State"
colnames(testpropo1)[36] = "County"
testpropo1[167,36] <- "Dona Ana County"
testpropo1[821,36] <- "La Salle Parish"
```

```
codes <- rep(NULL, length(testpropo1$County))

for (i in 1:length(testpropo1$avgAnnCount)){
 codes[i] = fips(state = testpropo1$State[i], county = testpropo1$County[i])
}
```

```
testpropo2 = cbind(testpropo1, fips = codes)
graphdata = data.frame(fips = testpropo2$fips, values = scale(testpropo2$Target_div_Income
```

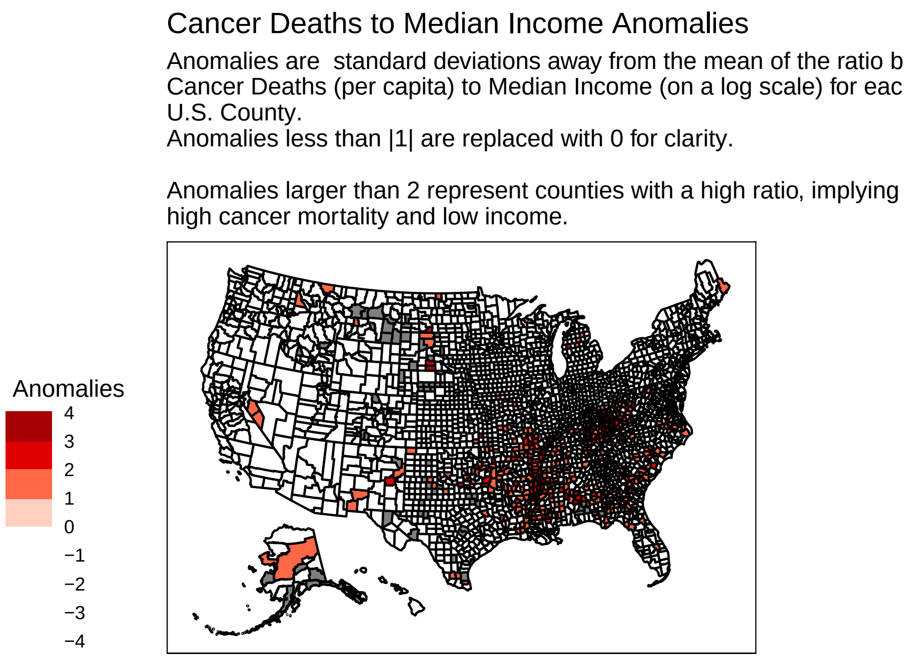
```
newbie <- graphdata %>% mutate(anomalies = ifelse(abs(values) > 1, values, 0))
newbie <- newbie[,c(1,3)]
```

New attempt, log ratio things instead of using scale

```
testpropo3 <- testpropo2 %>% mutate(Target_div_LogIncome = TARGET_deathRate/log(medIncome)
testpropolog = cbind(testpropo3, fips = codes)
graphdatalog = data.frame(fips = testpropolog$fips, values = testpropo3$Target_div_LogInco
```

```
newbieLOG <- graphdata %>% mutate(anomalies = ifelse(abs(scale(values)) > 1, values, 0))
newbieLOG <- newbieLOG[,c(1,3)]
```

```
plot_usmap(data = newbieLOG, values = "anomalies") +
  scale_fill_stepsn(breaks= -4:4, limits = c(-4,4),
                    colors=c("white","white", "white","red","dark red"),
                    guide = guide_colorsteps(even.steps = FALSE), name = " Anomalies") +
  theme(panel.background = element_rect(color = "black")) +
  theme(legend.position = "left") + labs(title = "Cancer Deaths to Median Income Anomalies
       subtitle = "Anomalies are  standard deviations away from the mean of the ratio betw
       \nAnomalies larger than 2 represent counties with a high ratio, implying \nhigh can
```



### Data Dictionary

**TARGET_deathRate:** Dependent variable. Mean *per capita* (100,000) cancer mortalities($a$)

**avgAnnCount:** Mean number of reported cases of cancer diagnosed annually($a$)

**avgDeathsPerYear:** Mean number of reported mortalities due to cancer($a$)

**incidenceRate:** Mean *per capita* (100,000) cancer diagoses($a$)

**medianIncome:** Median income per county ($b$)

**popEst2015:** Population of county ($b$)

**povertyPercent:** Percent of populace in poverty ($b$)

**studyPerCap:** *Per capita* number of cancer-related clinical trials per county ($a$)

**binnedInc:** Median income per capita binned by decile ($b$)

**MedianAge:** Median age of county residents ($b$)

**MedianAgeMale:** Median age of male county residents ($b$)

**MedianAgeFemale:** Median age of female county residents ($b$)

**Geography:** County name ($b$)

**AvgHouseholdSize:** Mean household size of county ($b$)

**PercentMarried:** Percent of county residents who are married ($b$)

**PctNoHS18_24:** Percent of county residents ages 18-24 highest education attained: less than high school ($b$)

**PctHS18_24:** Percent of county residents ages 18-24 highest education attained: high school diploma ($b$)

**PctSomeCol18_24:** Percent of county residents ages 18-24 highest education attained: some college ($b$)

**PctBachDeg18_24:** Percent of county residents ages 18-24 highest education attained: bachelor's degree ($b$)

**PctHS25_Over:** Percent of county residents ages 25 and over highest education attained: high school diploma ($b$)

**PctBachDeg25_Over:** Percent of county residents ages 25 and over highest education attained: bachelor's degree ($b$)

**PctEmployed16_Over:** Percent of county residents ages 16 and over employed ($b$)

**PctUnemployed16_Over:** Percent of county residents ages 16 and over unemployed ($b$)

**PctPrivateCoverage:** Percent of county residents with private health coverage ($b$)

**PctPrivateCoverageAlone:** Percent of county residents with private health coverage alone (no public assistance) ($b$)

**PctEmpPrivCoverage:** Percent of county residents with employee-provided private health coverage ($b$)

**PctPublicCoverage:** Percent of county residents with government-provided health coverage ($b$)

**PctPubliceCoverageAlone:** Percent of county residents with government-provided health coverage alone ($b$)

**PctWhite:** Percent of county residents who identify as White ($b$)

**PctBlack:** Percent of county residents who identify as Black ($b$)

**PctAsian:** Percent of county residents who identify as Asian ($b$)

**PctOtherRace:** Percent of county residents who identify in a category which is not White, Black, or Asian ($b$)

**PctMarriedHouseholds:** Percent of married households ($b$)

**BirthRate:** Number of live births relative to number of women in county ($b$)

($a$): years 2010-2016

($b$): 2013 Census Estimates

Data Pre processing - include everything up to testpropo3

```r
moddat <- testpropo3

(colMeans(is.na(moddat)))*100
```

```
        avgAnnCount      avgDeathsPerYear        TARGET_deathRate
           0.000000              0.000000                0.000000
      incidenceRate             medIncome              popEst2015
           0.000000              0.000000                0.000000
     povertyPercent           studyPerCap               binnedInc
           0.000000              0.000000                0.000000
          MedianAge         MedianAgeMale         MedianAgeFemale
           0.000000              0.000000                0.000000
          Geography      AvgHouseholdSize          PercentMarried
           0.000000              0.000000                0.000000
```

```
           PctNoHS18_24             PctHS18_24        PctSomeCol18_24
               0.000000               0.000000              74.991795
        PctBachDeg18_24            PctHS25_Over        PctBachDeg25_Over
               0.000000               0.000000               0.000000
      PctEmployed16_Over   PctUnemployed16_Over      PctPrivateCoverage
               4.988513               0.000000               0.000000
PctPrivateCoverageAlone      PctEmpPrivCoverage        PctPublicCoverage
              19.986872               0.000000               0.000000
 PctPublicCoverageAlone                PctWhite                PctBlack
               0.000000               0.000000               0.000000
               PctAsian            PctOtherRace    PctMarriedHouseholds
               0.000000               0.000000               0.000000
              BirthRate       Target_div_Income                  County
               0.000000               0.000000               0.000000
                  State                    fips    Target_div_LogIncome
               0.000000               0.000000               0.000000
```

Since PctSomeCol18_24 has a NA rate of 74.99%, and represents the inbetween between high school diploma and bachelors, we can justify excluding it.

PctEmployed16_Over has only a 4.99% NA rate, and PctPublicCoverageAlone, which is the percentage of county residents with government-provided health coverage alone, has a 19.99% NA rate, but seems too important to ignore if we wish to consider the status of coverage as a variable(s).

Let us do MICE (Multiple Imputation by Chained Equations) to replace these NA values with very likely substitutions. MICE operates under the assumption that the data missing is MAR (Missing at Random).

Due to the data collection process (each row represents a county), the likely possible bias is that certain states refuse or fail to collect these variables in a systematic way, and thus the data is no longer MAR. We will check this assumption towards the end of the modelling by considering our finalized model on both the imputated and original dataset (rows including NA's will be removed), and assess their similarities. Regardless, modelling will be done using the imputed dataset, assuming MAR.

```r
trim = moddat[,-18]
imp <- mice(trim, m = 5, maxit = 50, meth = "pmm")
```

```
Warning: Number of logged events: 505
```

```r
complete(imp)
```

```r
imputed <- complete(imp)
```

Initial variable selection for our model will be informed by domain knowledge and insight gained from prior visualization of the data.

Literature on socioeconomic factors affecting cancer mortality point to poverty, education, and race as some of the most important factors. In the 2017 paper "Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States, 1950–2014: Over Six Decades of Changing Patterns and Widening Inequalities," the authors concluded that individuals in lower income and education groups had significantly higher mortality and incidence rates. The authors also noted that Blacks had significantly higher mortality and incidence rates than other races. In the 2021 paper "Leading cancers contributing to educational disparities in cancer mortality in the US, 2017," the authors concluded that there was a significant difference between the mortality rate between individuals with a bachelors degree and higher, and all education levels below that. Since both these studies use data exclusively from the U.S., and are within the the time frame of interest to us, we are comfortable using these conclusions to guide our variable selection.

The visualizations of our own data support these conclusions as well as suggest a categorical variable indicating whether a given county is in the Southwest region.

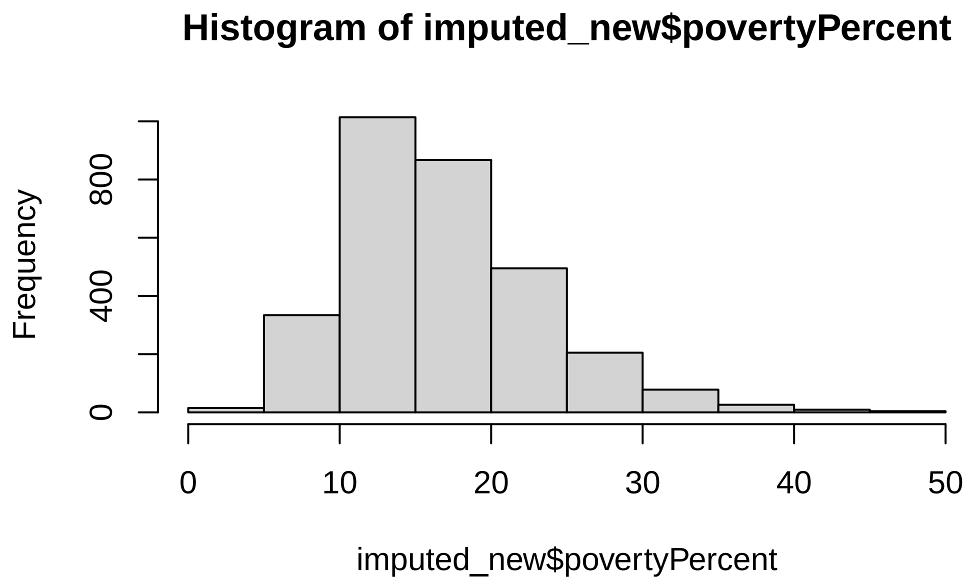
Additionally, the conclusions from the second paper suggest two new variables, **PctNoHS18_24** and **PctHS18_24**, which represent thepercent of county residents ages 18-24 whose highest education attained is less than a high school degree, and then a high school degree, respectively. While there are several other variables related to educational goals, such as percentage of county residents ages 18-24 who have attained a bachelors, the literature above suggests that residents with lower educational achievements have a higher cancer mortality, while the opposite is not necessarily true.

The variables for the initial model will be povertyPercent (Percent of populace in poverty), PctBlack, and PctNoHS18 and PctHS18_24. For future investigation we will consider isSouthEast (a categorical variable created later), as well as the variables related to healthcare coverage (**PctPrivateCoverage, PctPrivateCoverageAlone, PctEmpPrivCoverage, PctPublicCoverage, PctPubliceCoverageAlone).**

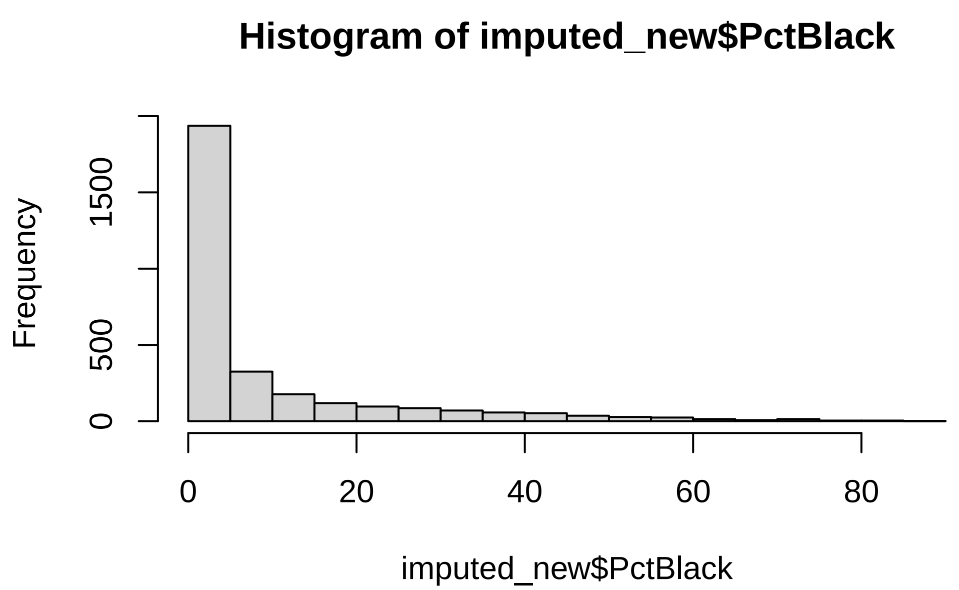
#work in progress belowwwwwwww

Firstly, let us examine the variables selected to see if any transformations would be appropriate.

```r
imputed_new <- imputed
hist(imputed_new$povertyPercent)
```




```
hist(imputed_new$PctBlack)
```



#okay continue with finished workkkkkkkk

```
mod1 <- lm(data = imputed_new, TARGET_deathRate ~ povertyPercent + PctBlack + PctNoHS18_24
```

```
summary(mod1)
```

```
Call:
lm(formula = TARGET_deathRate ~ povertyPercent + PctBlack + PctNoHS18_24 +
    PctHS18_24, data = imputed_new)

Residuals:
     Min       1Q   Median       3Q      Max
-106.595  -13.332    1.245   14.515  164.404

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    127.58734    2.10954  60.481  < 2e-16 ***
povertyPercent   1.66957    0.08312  20.087  < 2e-16 ***
PctBlack         0.13644    0.03527   3.869 0.000112 ***
PctNoHS18_24    -0.17345    0.05673  -3.058 0.002251 **
PctHS18_24       0.70898    0.04883  14.518  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.21 on 3042 degrees of freedom
Multiple R-squared:  0.2402,    Adjusted R-squared:  0.2392
F-statistic: 240.4 on 4 and 3042 DF,  p-value: < 2.2e-16
```

The initial fit is rather weak, with a R-squared of 0.24. Let us code and add isSoutheast as a categorical variable. To define which states belong to the Southeast, we will be using the regions specified by the Bureau of Economic Analysis, who divide the United States into 8 regions. Finally, we will use ANOVA to discern whether this suggested categorical variable is significant to our regression.

```
new_england <- c("Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island",
mideast <- c("Delaware", "District of Columbia", "Maryland", "New Jersey", "New York", "Pe
great_lakes <- c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin")
plains <- c("Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska", "North Dakota", "South
southeast <- c("Alabama", "Arkansas", "Florida", "Georgia", "Kentucky", "Louisiana", "Miss
southwest <- c("Arizona", "New Mexico", "Oklahoma", "Texas")
```

```r
rocky_mountain <- c("Colorado", "Idaho", "Montana", "Utah", "Wyoming")
far_west <- c("Alaska", "California", "Hawaii", "Nevada", "Oregon", "Washington")

get_region <- function(state) {
  if (state %in% new_england) {
    return("New England")
  } else if (state %in% mideast) {
    return("Mideast")
  } else if (state %in% great_lakes) {
    return("Great Lakes")
  } else if (state %in% plains) {
    return("Plains")
  } else if (state %in% southeast) {
    return("Southeast")
  } else if (state %in% southwest) {
    return("Southwest")
  } else if (state %in% rocky_mountain) {
    return("Rocky Mountain")
  } else if (state %in% far_west) {
    return("Far West")
  } else {
    return(NA)
  }
}

imputed_new$Region <- sapply(imputed_new$State, get_region)

imputed_new$isSoutheast <- ifelse(imputed_new$Region == "Southeast", "Yes", "No")

anova_result <- aov(data = imputed_new, TARGET_deathRate ~ povertyPercent + PctBlack + Pct

summary(anova_result)
```

```
                 Df  Sum Sq Mean Sq F value  Pr(>F)    
povertyPercent    1  432519  432519 764.933 < 2e-16 ***
PctBlack          1    4439    4439   7.851 0.00511 ** 
PctNoHS18_24      1    2909    2909   5.145 0.02338 *  
PctHS18_24        1  123510  123510 218.434 < 2e-16 ***
isSoutheast       1   63005   63005 111.428 < 2e-16 ***
Residuals      3041 1719484     565                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the ANOVA results, isSoutheast is as significant as povertyPercent and bachDiff, both variables informed by domain knowledge. Thus, we feel comfortable adding this variable to our model. Interestingly, PctNoHS18_24 appears to have the lowest F values, and in our last regression, had a negative coefficient, which is difficult to interpret in face of literature suggesting the opposite. For this reason, we will be omitting it in our next model.

```
mod2 <- lm(data = imputed_new, TARGET_deathRate ~ povertyPercent + PctBlack + PctHS18_24 +
summary(mod2)
```

```
Call:
lm(formula = TARGET_deathRate ~ povertyPercent + PctBlack + PctHS18_24 +
    isSoutheast, data = imputed_new)

Residuals:
     Min       1Q   Median       3Q      Max
-110.725  -13.083    1.391   14.825  159.778

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    129.08605    2.01556  64.045   <2e-16 ***
povertyPercent   1.43749    0.08036  17.889   <2e-16 ***
PctBlack        -0.04313    0.03854  -1.119    0.263
PctHS18_24       0.61491    0.04856  12.663   <2e-16 ***
isSoutheastYes  12.20437    1.12638  10.835   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.79 on 3042 degrees of freedom
Multiple R-squared:  0.2661,    Adjusted R-squared:  0.2652
F-statistic: 275.8 on 4 and 3042 DF,  p-value: < 2.2e-16
```

Similar to the selection of education-related variables above, for insurance-related variables we will only consider **PctPublicCoverage and PctPubliceCoverageAlone,** as these variables correspond to the the percentages of each county. Using cross-validation to find the optimal Lambda value, we will use Lasso regression to choose a model using either or these two variables, or neither. Additionally, this will help us choose whether or not to remove PctBlack, as it had the lowest t value in our previous regression.

```
y = imputed_new$TARGET_deathRate
x = data.matrix(imputed_new[, c('povertyPercent', 'PctBlack', 'PctHS18_24', 'isSoutheast',
```

```r
cv_model <- cv.glmnet(x, y, alpha = 1)
best_lambda <- cv_model$lambda.min
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
```

```
7 x 1 sparse Matrix of class "dgCMatrix"
                               s0
(Intercept)           112.5259029
povertyPercent          0.6354839
PctBlack                .
PctHS18_24              0.4738554
isSoutheast            11.9045537
PctPublicCoverage       0.2495975
PctPublicCoverageAlone  0.7167604
```

Check below for 28 variable regression results :(

```r
#allcheck
all <- imputed_new[,c(-1,-2,-4,-9,-13,-34,-35,-36,-37,-38,-39,-40,-41)]

summary(lm(data = all, TARGET_deathRate ~ .))
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-97.396 -11.709   0.321  11.503 169.747

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.467e+02  1.741e+01  14.169  < 2e-16 ***
medIncome                2.097e-04  8.891e-05   2.358 0.018415 *
popEst2015              -3.993e-07  1.403e-06  -0.285 0.775927
povertyPercent           1.302e-01  1.811e-01   0.719 0.472174
studyPerCap              9.471e-04  7.486e-04   1.265 0.205918
MedianAge                1.841e-03  8.668e-03   0.212 0.831794
MedianAgeMale           -5.201e-01  2.307e-01  -2.254 0.024239 *
MedianAgeFemale         -4.598e-01  2.399e-01  -1.917 0.055350 .
AvgHouseholdSize        -3.709e-01  1.062e+00  -0.349 0.726895
```

```
PercentMarried           1.726e+00  1.888e-01   9.139  < 2e-16 ***
PctNoHS18_24            -2.207e-01  6.201e-02  -3.559 0.000377 ***
PctHS18_24               2.625e-01  5.472e-02   4.798 1.68e-06 ***
PctBachDeg18_24         -8.858e-02  1.199e-01  -0.739 0.460023
PctHS25_Over             5.547e-01  1.069e-01   5.191 2.22e-07 ***
PctBachDeg25_Over       -1.283e+00  1.716e-01  -7.475 1.00e-13 ***
PctEmployed16_Over      -8.510e-01  1.210e-01  -7.033 2.50e-12 ***
PctUnemployed16_Over     4.097e-01  1.850e-01   2.214 0.026878 *
PctPrivateCoverage       6.178e-02  2.876e-01   0.215 0.829921
PctPrivateCoverageAlone -1.949e-01  3.457e-01  -0.564 0.572858
PctEmpPrivCoverage       5.930e-01  1.359e-01   4.365 1.32e-05 ***
PctPublicCoverage       -4.904e-01  3.454e-01  -1.420 0.155713
PctPublicCoverageAlone   1.232e+00  3.956e-01   3.113 0.001867 **
PctWhite                -7.357e-02  6.359e-02  -1.157 0.247366
PctBlack                 3.889e-02  6.137e-02   0.634 0.526337
PctAsian                -2.248e-01  2.087e-01  -1.077 0.281559
PctOtherRace            -1.284e+00  1.369e-01  -9.383  < 2e-16 ***
PctMarriedHouseholds    -1.985e+00  1.798e-01 -11.043  < 2e-16 ***
BirthRate               -1.280e+00  2.136e-01  -5.990 2.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.43 on 3019 degrees of freedom
Multiple R-squared:  0.4091,    Adjusted R-squared:  0.4038
F-statistic: 77.41 on 27 and 3019 DF,  p-value: < 2.2e-16
```