comm188_final_proj

Daniel Varivoda 204755697

5/24/2020

Final project

Cleaning the Data:

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(stringr)
tweets <- read.csv("merged_tweetIDs.csv")
tweets <- tweets[,c(6,7,11,18,28)]

nrow(tweets)
```

```
## [1] 29209
```

```
#removing all non english tweets
tweets <- tweets[tweets$lang == "en",]
tweets$text <- as.character(tweets$text)
tweets$user_location <- as.character(tweets$user_location)
tweets2 <- tweets


#removing all non ASCII characters
for(i in 1:nrow(tweets)){
  tweets$text[i] <- gsub("[^\x01-\x7F]", "", tweets$text[i])
  tweets$user_location[i] <- gsub("[^\x01-\x7F]", "", tweets$user_location[i])
}

head(tweets$text)
```

```
## [1] "RT @DrGJackBrown: Tom Cotton  a U.S. Senator, stoking the flames of ignorance, racism, conspira
## [2] "RT @chayraestillore: this corona virus is giving me serious anxiety especially as someone with a
## [3] "RT @sweetkizzez504: Me after listening to @FriendZonePod wash your hands episode.  @HeyFranHey
## [4] "Underrated tweet  https://t.co/5whhQq1zaD"
## [5] "@sh_irredeemable @SenTomCotton If Nixon hadnt have gone to China, they would have collapsed und
## [6] "\"On February 10th, legislation in England was pushed through enabling authorities entry to any
```

```
#removing URLS
library(qdapRegex)
tweets$text <- rm_url(tweets$text, pattern=pastex("@rm_twitter_url", "@rm_url"))
head(tweets)
```

```
##   favorite_count           id lang
```

```
## 2              0 1.229177e+18   en
## 3              0 1.223578e+18   en
## 4              0 1.226276e+18   en
## 5              1 1.223592e+18   en
## 6              3 1.229166e+18   en
## 7              1 1.229169e+18   en
##
## 2
## 3
## 4
## 5
## 6                                                    @sh_irredeemable @SenTomCotton If Nixon hadnt have
## 7 "On February 10th, legislation in England was pushed through enabling authorities entry to any pren
##                  user_location
## 2 Undisclosed Distant Location
## 3
## 4                 Houston, TX
## 5             Lahore, Pakistan
## 6          Southern California
## 7                          NYC
```

```
#removing the retweet from the text
for(i in 1:nrow(tweets)){
  if(str_detect(tweets$text[i], "RT @.*? "))
  tweets$text[i] <- removeWords(tweets$text[i], "RT @.*? ")
}


head(tweets)
```

```
##   favorite_count           id lang
## 2              0 1.229177e+18   en
## 3              0 1.223578e+18   en
## 4              0 1.226276e+18   en
## 5              1 1.223592e+18   en
## 6              3 1.229166e+18   en
## 7              1 1.229169e+18   en
##
## 2
## 3
## 4
## 5
## 6                                                    @sh_irredeemable @SenTomCotton If Nixon hadnt have
## 7 "On February 10th, legislation in England was pushed through enabling authorities entry to any pren
##                  user_location
## 2 Undisclosed Distant Location
## 3
## 4                 Houston, TX
## 5             Lahore, Pakistan
## 6          Southern California
## 7                          NYC
```

Calculating Flesch Reading Scores

```
#function to take out the empty whitespace in words vector
keep_words <- function(words){
  words[nchar(words) > 0]
}

#check if theres a special syllable ending
is_special_ending <- function(ending) {
  is_es <- all(ending == c("e", "s"))
  is_ed <- all(ending == c("e", "d"))
  is_e_not_le <- ending[2] == "e" & ending[1] != "l"
  is_es | is_ed | is_e_not_le
}

#check if there is a special ending in the word
rm_special_endings <- function(word_letters) {
    word_tail <- tail(word_letters, n = 2)
    if (is_special_ending(word_tail)) {
      if (word_tail[2] == "e") {
        word_letters[-length(word_letters)]
      } else {
        head(word_letters, n = -2)
    }
  } else {
    word_letters
    }
}

#count the number of syllables
count_syllables <- function(word) {
  word_letters <- unlist(strsplit(word, split = ""))
  if (length(word_letters) <= 3) {
    1
  } else {
  word_letters <- rm_special_endings(word_letters)
  word_vowels <- is_vowel(word_letters)
  sum(word_vowels) - sum(diff(which(word_vowels)) == 1)
  }
}

#check if the letter is a vowel
is_vowel <- function(letter) {
  letter %in% c("a", "e", "i", "o", "u", "y")
}


#Function to actually calculate the Flesch Reading Ease Score --> For the purposes of tweets, I counted
reading_ease <- function(passage)
{
  paste(passage, collapse = " ")

  #split the passage into sentences, put in lower case,
```

```
  #and remove punctuation
  sentences <- passage
  sentences <- tolower(sentences)
  sentences <- gsub(pattern = "[[:punct:]]", replacement = "", sentences)
  sent_tot <- 1

  #split the sentences into words
  words <- strsplit(sentences, split = " ")
  words <- lapply(words, keep_words)
  words <- unname(unlist(words))
  words_tot <- length(words)

  syl_num <- 0

  #count the number of syllables in each word
  for(i in words) {
    syl_num <- syl_num + count_syllables(i)
  }

  RE <- 206.835 - (1.015 * (words_tot)) - (84.6 * (syl_num / words_tot))
  RE
}
```

```
#calculating the flesch reading ease score for each tweet
flesch <- c()
for(i in 1:nrow(tweets)){
  f <- reading_ease(tweets$text[i])
  flesch <- c(flesch, f)
}
length(flesch)
```

```
## [1] 17699
```

```
tweets <- cbind(flesch, tweets)
head(tweets)
```

```
##     flesch favorite_count           id lang
## 2 29.46786              0 1.229177e+18   en
## 3 37.25750              0 1.223578e+18   en
## 4 11.35500              0 1.226276e+18   en
## 5 35.60500              1 1.223592e+18   en
## 6 52.49658              3 1.229166e+18   en
## 7 20.64000              1 1.229169e+18   en
##
## 2
## 3
## 4
## 5
## 6                                                    @sh_irredeemable @SenTomCotton If Nixon hadnt have
## 7 "On February 10th, legislation in England was pushed through enabling authorities entry to any pren
##                   user_location
## 2 Undisclosed Distant Location
```

```
## 3
## 4                  Houston, TX
## 5             Lahore, Pakistan
## 6          Southern California
## 7                          NYC
```

Getting the states in order to sort tweets by location

```
states <- read.csv("states.csv", stringsAsFactors = FALSE)
tweets$user_location <- as.character(tweets$user_location)

states$Code <- paste(" ", states$Code, " ", sep = "")


# In order to get only tweets by state I selected for any location that either includes a state name or
is_state <- c()
state_num <- c()
issue <- c()
for(i in 1:nrow(tweets)){
  check <-  FALSE
  name <- integer(0)
  abv <- integer(0)
  if(any(str_detect(tweets$user_location[i], states$Code))||
     any(str_detect(tweets$user_location[i], states$Name))){
    is_state <- c(is_state, TRUE)
    check <- TRUE
  } else {is_state <- c(is_state, FALSE) }
  if(check){
    abv <- which(str_detect(tweets$user_location[i], states$Code))
    name <- which(str_detect(tweets$user_location[i], states$Name))
    # if(str_detect(tweets$user_location[i], "West Virginia")){
    #   name <- 48
    # } else if(str_detect(tweets$user_location[i], "Southern West Virginia")){
    # name <- 48
    #} else if(str_detect(tweets$user_location[i], "Arkansas")){
    #   name <- 4
    # }
  }

  if(length(abv) > 1 || length(name) > 1){
    issue <- c(issue,i)
    if(length(abv) >1){
      name <- abv
    }
    if(sum(name) == 63){
      name <- 35
    } else if(sum(name) == 20){
      name <- 4
    } else if(sum(name) == 94){
      name <- 48
    } else if(sum(name) == 79){
```

```
      name <- 47
    } else if(sum(name) == 49){
      name <- 6
    } else if(sum(name) == 80){
      name <- 36
    } else if(sum(name) == 39){
      name <- 7
    } else if(sum(name) == 19){
      name <- 1
    } else if(sum(name) == 51){
      name <- 18
    } else if(sum(name) == 35){
      name <- 22
    } else if(name[1] == 5 && name[2] == 27){
      name <- 27
    } else if(sum(name) == 33){
      name <- 5
    } else if(sum(name) == 79){
      name <- 32
    } else if(sum(name) == 16){
      name <- 11
    } else if(sum(name) == 21){
      name <- 9
    } else if(sum(name) == 64){
      name <- 42
    } else if(sum(name) == 53){
      name <- 25
    } else if(sum(name) == 41){
      name <- 25
    } else if(sum(name) == 65){
      name <- 33
    } else if(sum(name) == 22){
      name <- 17
    } else if(sum(name) == 11){
      name <- 10
    }
  }
  if(check == FALSE){
    state_num <- c(state_num, NA)
  } else if(length(name) > 0){
    state_num <- c(state_num, name)
  } else if(length(abv) > 0){
    state_num <- c(state_num, abv)
  } else { print(i)}
}

#Finding places where two names are dteected and changing above
#code has been commented out after being implemented to fix issues above
# for(i in 1:length(issue)){
# print(sum(which(str_detect(tweets$user_location[issue[i]], states$Code))))
# print(sum(which(str_detect(tweets$user_location[issue[i]], states$Name))))
# print(which(str_detect(tweets$user_location[issue[i]], states$Name)))
# print(which(str_detect(tweets$user_location[issue[i]], states$Code)))
```

```
# print(tweets$user_location[issue[i]])
# if(length(which(str_detect(tweets$user_location[issue[i]], states$Name))) > 0){
#   print(states$Name[which(str_detect(tweets$user_location[issue[i]], states$Name))])
# } else{
#   print(states$Name[which(str_detect(tweets$user_location[issue[i]], states$Code))])
# }
# print("************************")
# }
```

```
tweets <- cbind(is_state, state_num, tweets)

#only keeping tweets with identified states
tweets <- tweets[tweets$is_state == TRUE,]
nrow(tweets)
```

```
## [1] 1900
```

```
c_t <- tweets[,-1]

#generating a frequency table of every word in the twitter dictionary
words <- c()
for(i in 1:nrow(c_t)){
  txt <- c_t$text[i]
  #removing punctuation from words
  txt <-   gsub('[[:punct:] ]+',' ',txt)
  #splitting string into a vector
  words <- c(words, unlist(strsplit(txt, " ")))
}

#creating a frequency table of words
freq <- as.data.frame(table(words))
freq <- freq[order(freq$Freq, decreasing = T),]
head(freq,20)
```

```
##            words Freq
## 7361         the 1432
## 7492          to  981
## 227            a  667
## 5184          of  651
## 529          and  573
## 3780          in  563
## 3975          is  516
## 1799 coronavirus  417
## 3017         for  356
## 3691           I  299
## 5235          on  286
## 7356        that  264
## 626          are  252
## 8366         you  244
## 1435       China  241
## 3994          it  218
## 3429        have  215
```

```
## 7422        this  215
## 1800 Coronavirus  213
## 7644       Trump  199
```

```
#looking at frequency of words and creating a small dictionary in order to sort
big_deal <- c("pandemic", "outbreak", "epidemic", "crisis", "global", "death", "infected", "quarantine"

not_big_deal <- c("down", "flu", "fine", "support", "control", "vaccine", "nothing", "hoax","free", "cl
```

```
#adding the state name + other demographics to help with regression
states2 <- states[c(1,6,9,11)]
temp <- states2[c_t$state_num[1],]
for(i in 2:nrow(c_t)){
  temp <- rbind(temp, states2[c_t$state_num[i],])
}
head(tweets)
```

```
##    is_state state_num   flesch favorite_count           id lang
## 6      TRUE         5 52.49658              3 1.229166e+18   en
## 18     TRUE         9 78.87286              0 1.227278e+18   en
## 31     TRUE         9 42.71500              0 1.227279e+18   en
## 48     TRUE        23 49.48000              1 1.227077e+18   en
## 60     TRUE         7 42.71500              0 1.229167e+18   en
## 85     TRUE         9 72.61545              0 1.229177e+18   en
##
## 6  @sh_irredeemable @SenTomCotton If Nixon hadnt have gone to China, they would have collapsed under
## 18
## 31
## 48
## 60
## 85
##                     user_location
## 6              Southern California
## 18                    Florida, USA
## 31                    Florida, USA
## 48                  Minnesota, USA
## 60 Torrington, CT + Brooklyn, NY
## 85      Florida, Space Coast, USA
```

```
c_t <- cbind(temp, c_t)
#coronavirus case date from https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-
# combining all weekly deaths for march
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:qdapRegex':
##
##     explain

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
deaths <- read.csv("covid_deaths_by_week_per_states.csv")
deaths <- deaths[!is.na(deaths$COVID.19.Deaths),]
state_covid_deaths <- deaths %>% group_by(State) %>% summarise("march_flu_covid_pneumonia_deaths" = sum
))
state_covid_deaths$State <- as.character(state_covid_deaths$State)

c_t <- c_t %>% left_join(state_covid_deaths, by = c("Name" = "State"))

c_t <- c_t %>%
  select(march_flu_covid_pneumonia_deaths, everything())
colnames(c_t)[1] <- "total_march_covid_deaths"

head(c_t)
```

```
##   total_march_covid_deaths        Name Population Land.area.mi.2.
## 1                      814  California 39,512,223         155,779
## 2                      626     Florida 21,477,737          53,625
## 3                      626     Florida 21,477,737          53,625
## 4                       76   Minnesota  5,639,632          79,627
## 5                      477 Connecticut  3,565,278           4,842
## 6                      626     Florida 21,477,737          53,625
##   Water.area.mi.2. state_num   flesch favorite_count           id lang
## 1            7,916         5 52.49658              3 1.229166e+18   en
## 2           12,133         9 78.87286              0 1.227278e+18   en
## 3           12,133         9 42.71500              0 1.227279e+18   en
## 4            7,309        23 49.48000              1 1.227077e+18   en
## 5              701         7 42.71500              0 1.229167e+18   en
## 6           12,133         9 72.61545              0 1.229177e+18   en
##
## 1 @sh_irredeemable @SenTomCotton If Nixon hadnt have gone to China, they would have collapsed under
## 2
## 3
## 4
## 5
## 6
##                    user_location
## 1            Southern California
## 2                   Florida, USA
## 3                   Florida, USA
## 4                 Minnesota, USA
## 5 Torrington, CT + Brooklyn, NY
## 6      Florida, Space Coast, USA
```

```
write.csv(c_t, "cleaned_tweets_and_joined_data.csv", row.names = FALSE)
```

```
#using my dictionary for worried/not worried about covid to analyze sentiment by counting number of com
sentiment <- c()
for(i in 1:nrow(c_t)){
  mytxt <- c_t$text[i]
```

```
  neg <- -1*sum(str_count(mytxt, not_big_deal))
  pos <- sum(str_count(mytxt, big_deal))
  tmp <- neg + pos
  sentiment <- c(sentiment, tmp)
}
c_t <- cbind(sentiment, c_t)

head(c_t)
```

```
##   sentiment total_march_covid_deaths        Name Population
## 1         0                     814  California 39,512,223
## 2         0                     626     Florida 21,477,737
## 3         0                     626     Florida 21,477,737
## 4         1                      76   Minnesota  5,639,632
## 5         0                     477 Connecticut  3,565,278
## 6         0                     626     Florida 21,477,737
##   Land.area.mi.2. Water.area.mi.2. state_num   flesch favorite_count
## 1         155,779           7,916         5 52.49658              3
## 2          53,625          12,133         9 78.87286              0
## 3          53,625          12,133         9 42.71500              0
## 4          79,627           7,309        23 49.48000              1
## 5           4,842             701         7 42.71500              0
## 6          53,625          12,133         9 72.61545              0
##             id lang
## 1 1.229166e+18   en
## 2 1.227278e+18   en
## 3 1.227279e+18   en
## 4 1.227077e+18   en
## 5 1.229167e+18   en
## 6 1.229177e+18   en
##
## 1 @sh_irredeemable @SenTomCotton If Nixon hadnt have gone to China, they would have collapsed under
## 2
## 3
## 4
## 5
## 6
##                   user_location
## 1            Southern California
## 2                  Florida, USA
## 3                  Florida, USA
## 4                Minnesota, USA
## 5 Torrington, CT + Brooklyn, NY
## 6      Florida, Space Coast, USA
```

```
#changing the columns to numbers from chars
for(i in 1:nrow(c_t)){
  four <- c_t$Population[i]
  five <- c_t$Land.area.mi.2.[i]
  six <- c_t$Water.area.mi.2.[i]
  four <-   gsub('[[:punct:] ]+','',four)
  c_t$Population[i] <- as.numeric(four)
  five <-   gsub('[[:punct:] ]+','',five)
```

```
  c_t$Land.area.mi.2.[i] <- as.numeric(five)
  six <-   gsub('[[:punct:] ]+','',six)
  c_t$Water.area.mi.2.[i] <- as.numeric(six)
}
c_t[,4] <- as.numeric(c_t[,4])
c_t[,5] <- as.numeric(c_t[,5])
c_t[,6] <- as.numeric(c_t[,6])
```

```
library(alr3)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

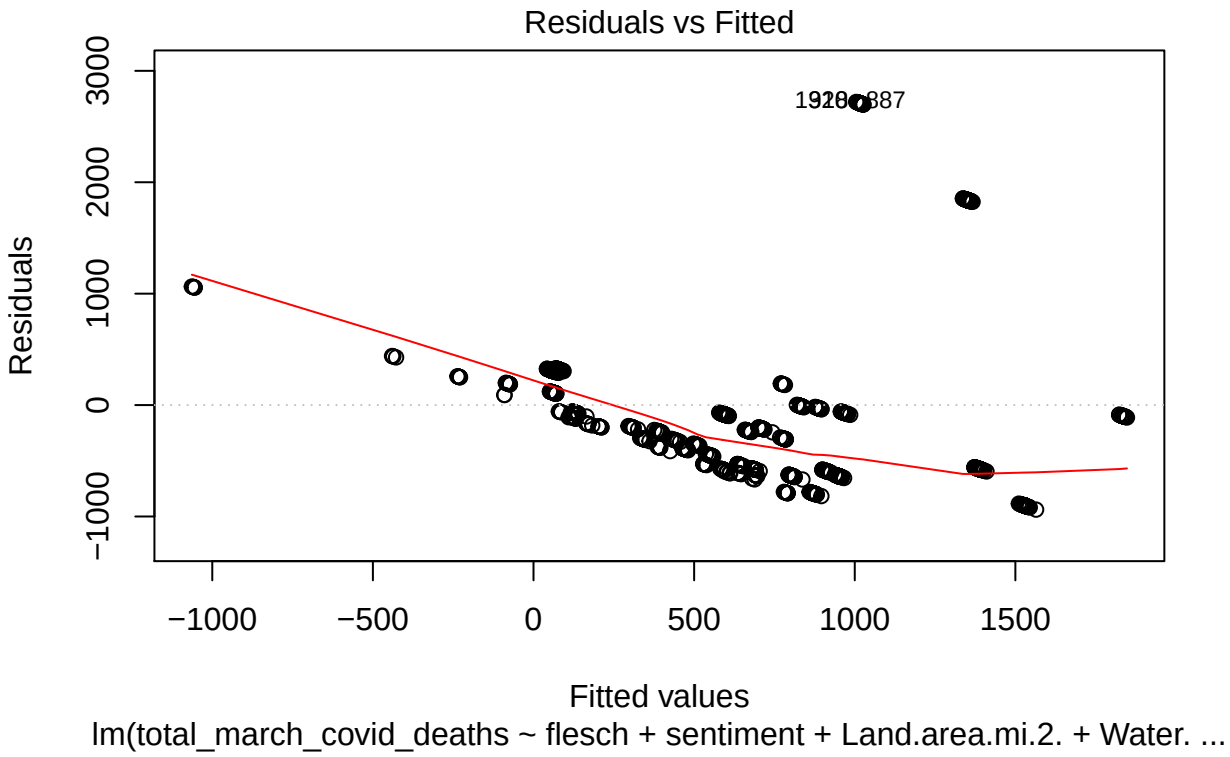
```
##
## Attaching package: 'car'
```

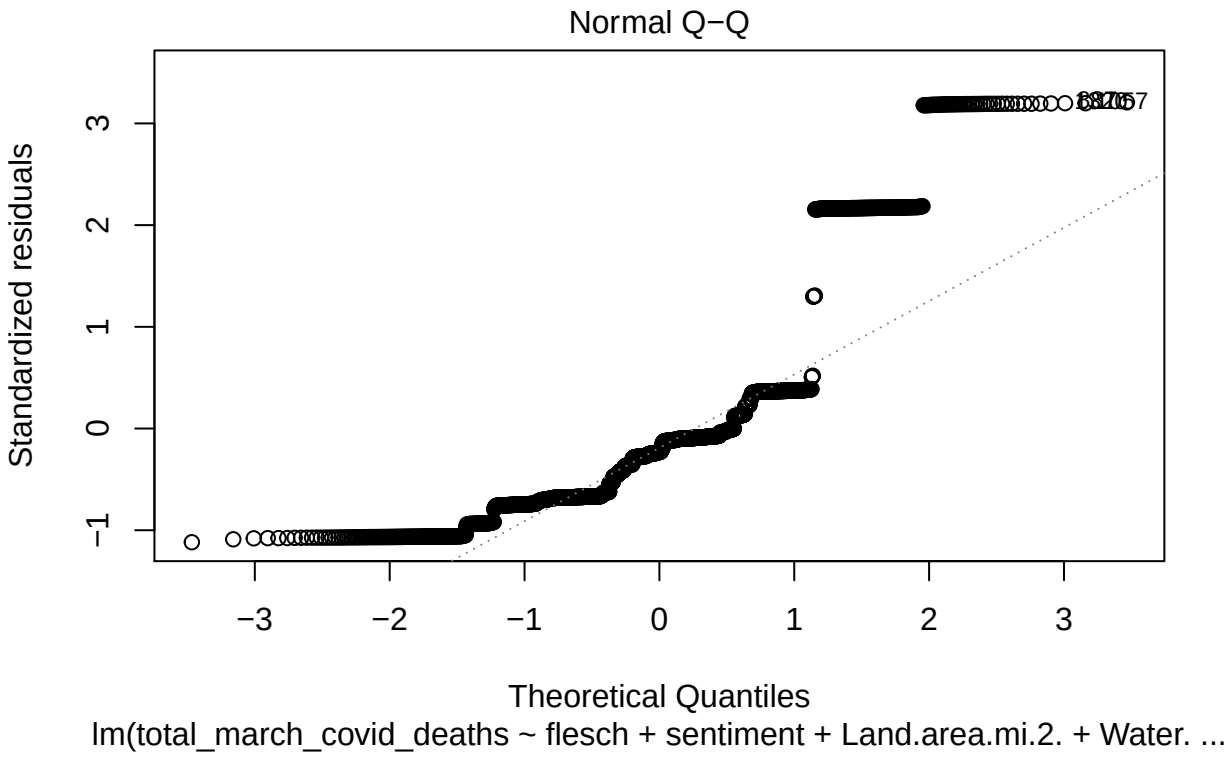
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

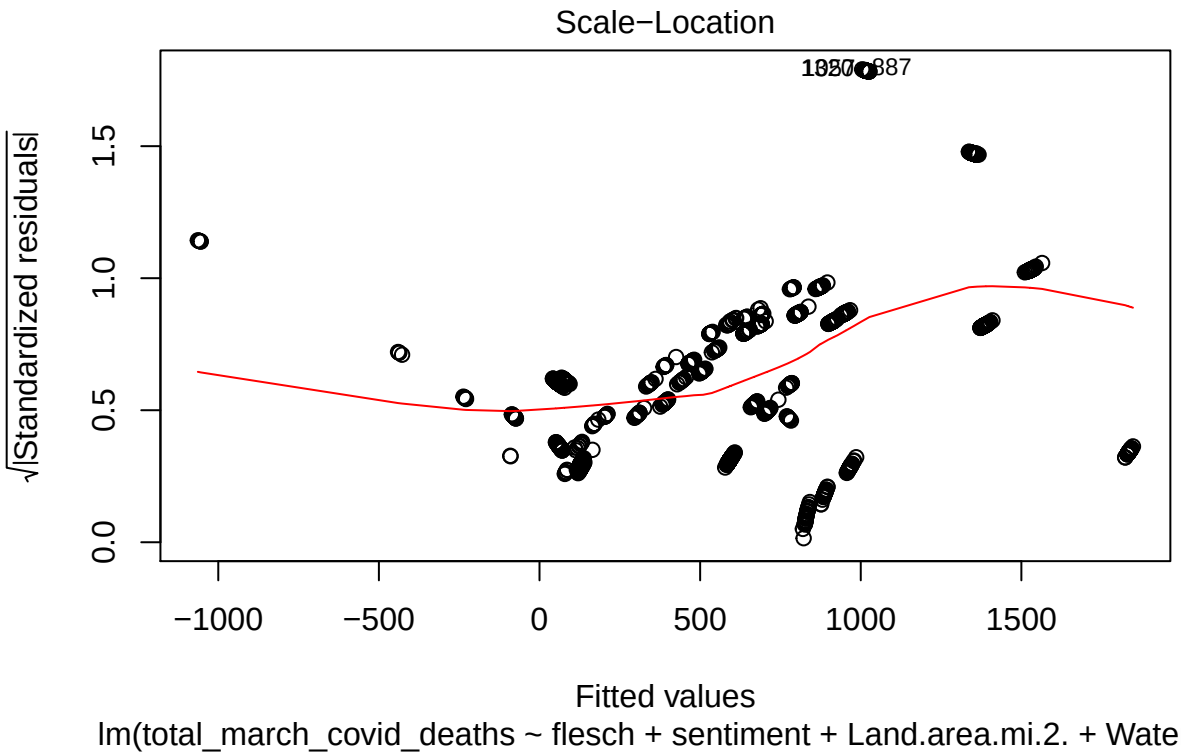
```
## The following object is masked from 'package:qdapRegex':
##
##     S
```

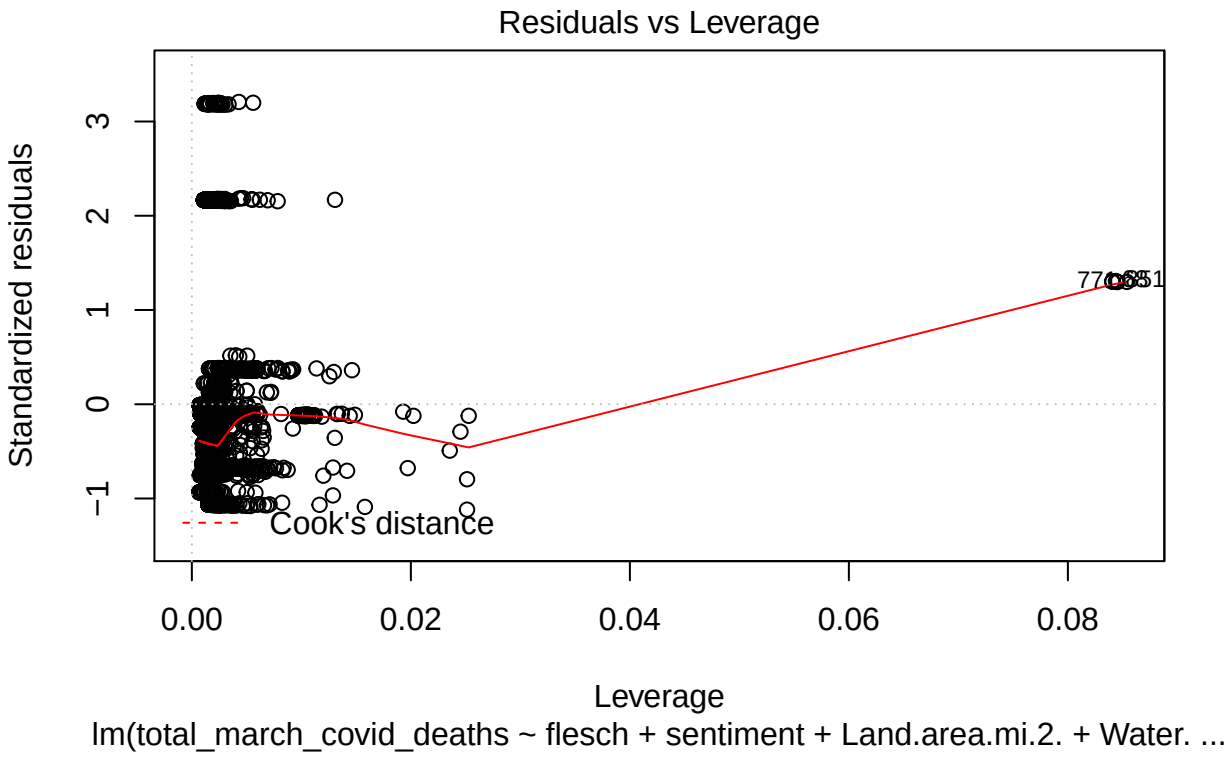
```
# Seems like Flesch Reading Score or Sentiment have no correlation to March Covid Deaths by State

#The model is not valid however power transformations inorder to attempt to create a valid model will d
fleschlm <- lm(total_march_covid_deaths ~ flesch + sentiment + Land.area.mi.2. + Water.area.mi.2. + Popu
plot(fleschlm)
```

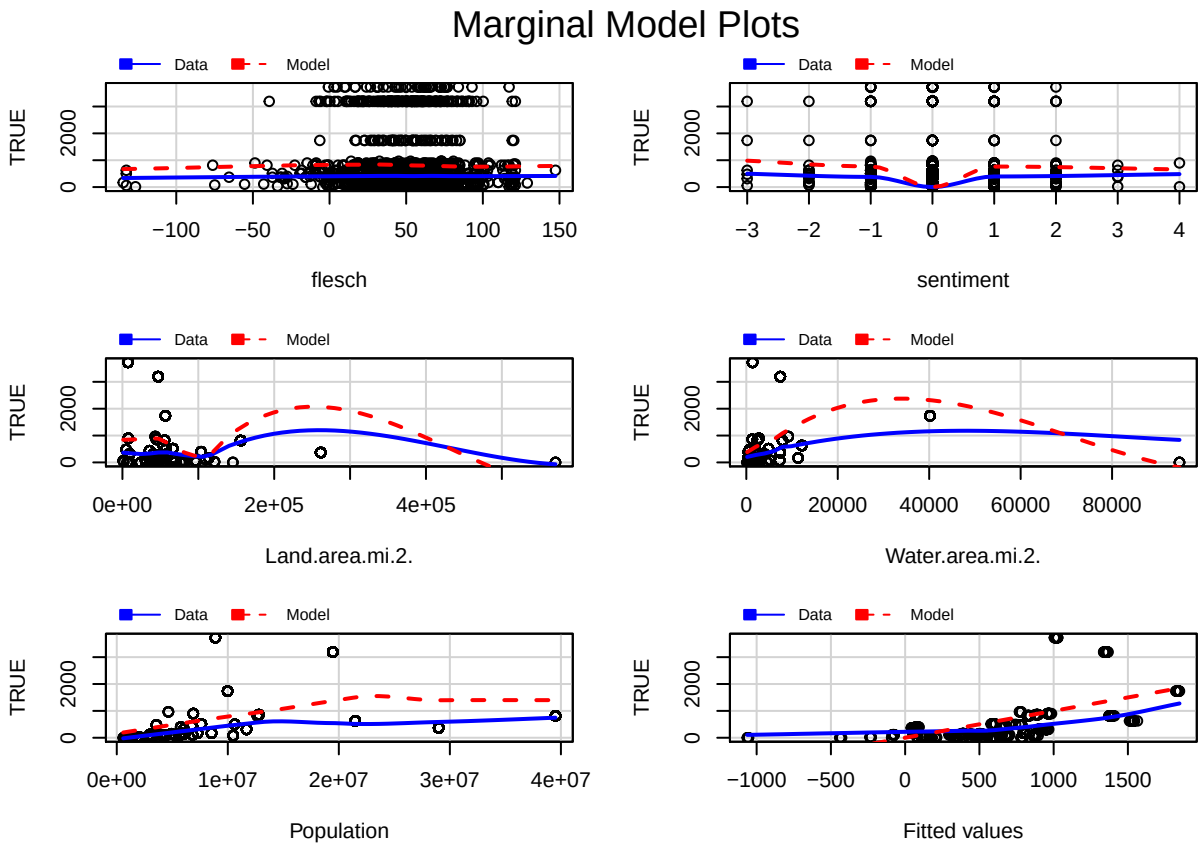

Residuals vs Fitted
Residuals
Fitted values
lm(total_march_covid_deaths ~ flesch + sentiment + Land.area.mi.2. + Water. ...
−1000
0
1000
2000
3000
−1000
−500
0
500
1000
1500


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(total_march_covid_deaths ~ flesch + sentiment + Land.area.mi.2. + Water. ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−1000
−500
0
500
1000
1500
Fitted values
lm(total_march_covid_deaths ~ flesch + sentiment + Land.area.mi.2. + Water. ...




```
mmps(fleschlm)
```



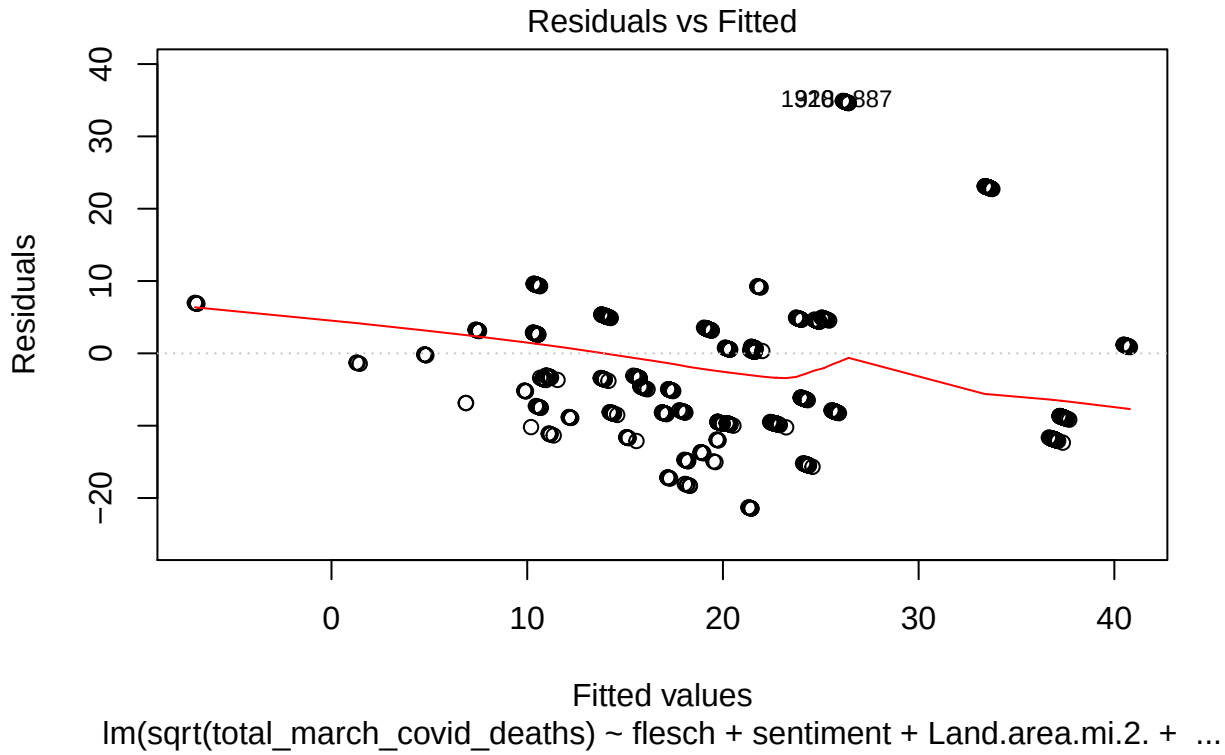

```
summary(fleschlm)
```

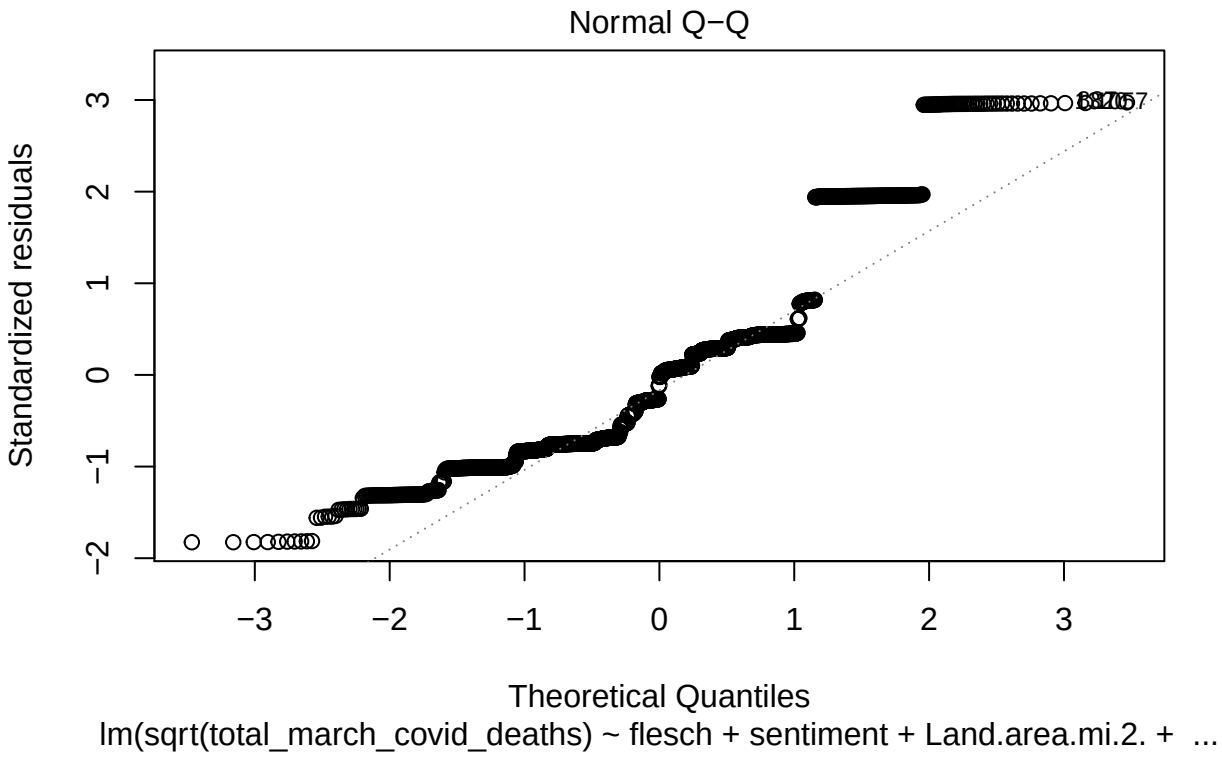
```
##
## Call:
## lm(formula = total_march_covid_deaths ~ flesch + sentiment +
##     Land.area.mi.2. + Water.area.mi.2. + Population, data = c_t)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -938.5 -574.4 -195.0  251.9 2720.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.498e+02  5.009e+01  12.972   <2e-16 ***
## flesch           -1.921e-01  7.160e-01  -0.268    0.789
## sentiment        -7.569e-01  2.922e+01  -0.026    0.979
## Land.area.mi.2.  -8.019e-03  3.308e-04 -24.245   <2e-16 ***
## Water.area.mi.2.  3.001e-02  2.327e-03  12.895   <2e-16 ***
## Population        4.447e-05  2.079e-06  21.391   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 850.1 on 1894 degrees of freedom
## Multiple R-squared:  0.2879, Adjusted R-squared:  0.2861
## F-statistic: 153.2 on 5 and 1894 DF,  p-value: < 2.2e-16
```

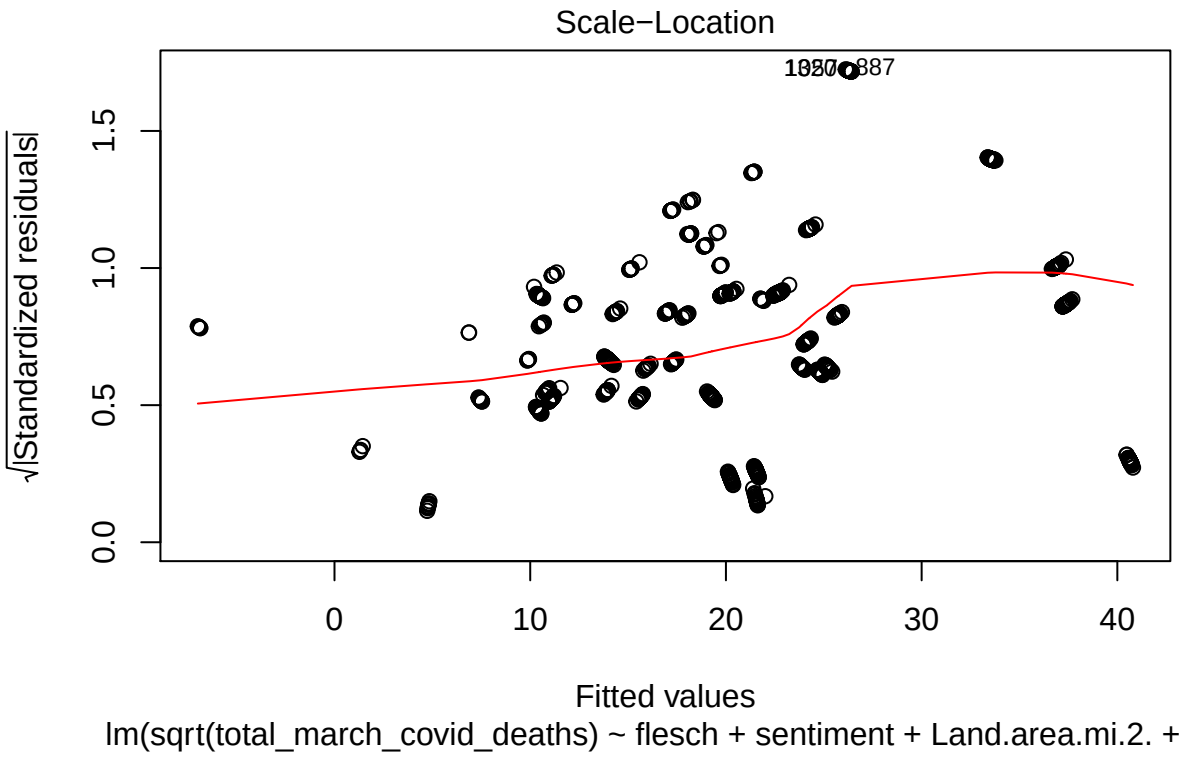
```r
anova(fleschlm)
```

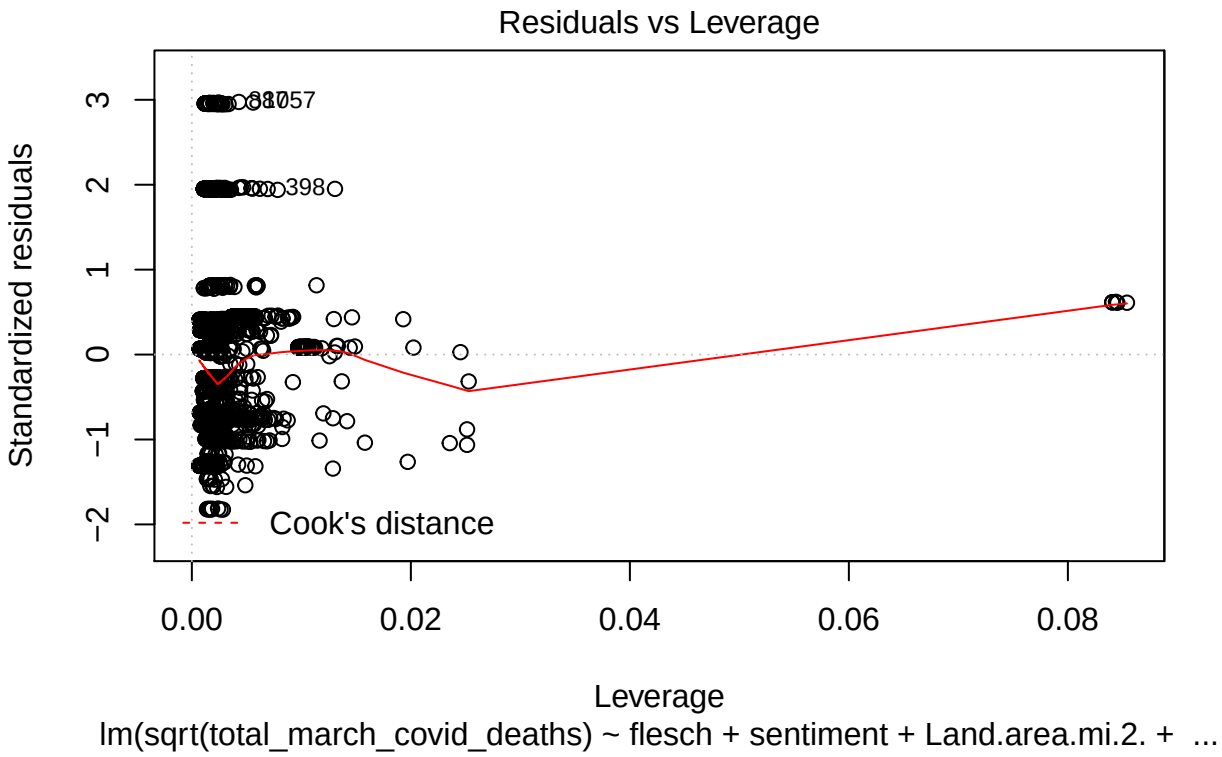
```
## Analysis of Variance Table
##
## Response: total_march_covid_deaths
##                     Df     Sum Sq   Mean Sq  F value Pr(>F)
## flesch               1     122150    122150   0.1690 0.6810
## sentiment            1     405099    405099   0.5606 0.4541
## Land.area.mi.2.      1   93900505  93900505 129.9378 <2e-16 ***
## Water.area.mi.2.     1  128376272 128376272 177.6447 <2e-16 ***
## Population           1  330660884 330660884 457.5625 <2e-16 ***
## Residuals         1894 1368713069    722657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Trying some minor changes in order to attempt at improving model, which does improve accuracy somewhat
#The data seems to be too clustered to come to any proper conclusions using this model as it does not a
fleschlm <- lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + Water.area.mi.2.
plot(fleschlm)
```

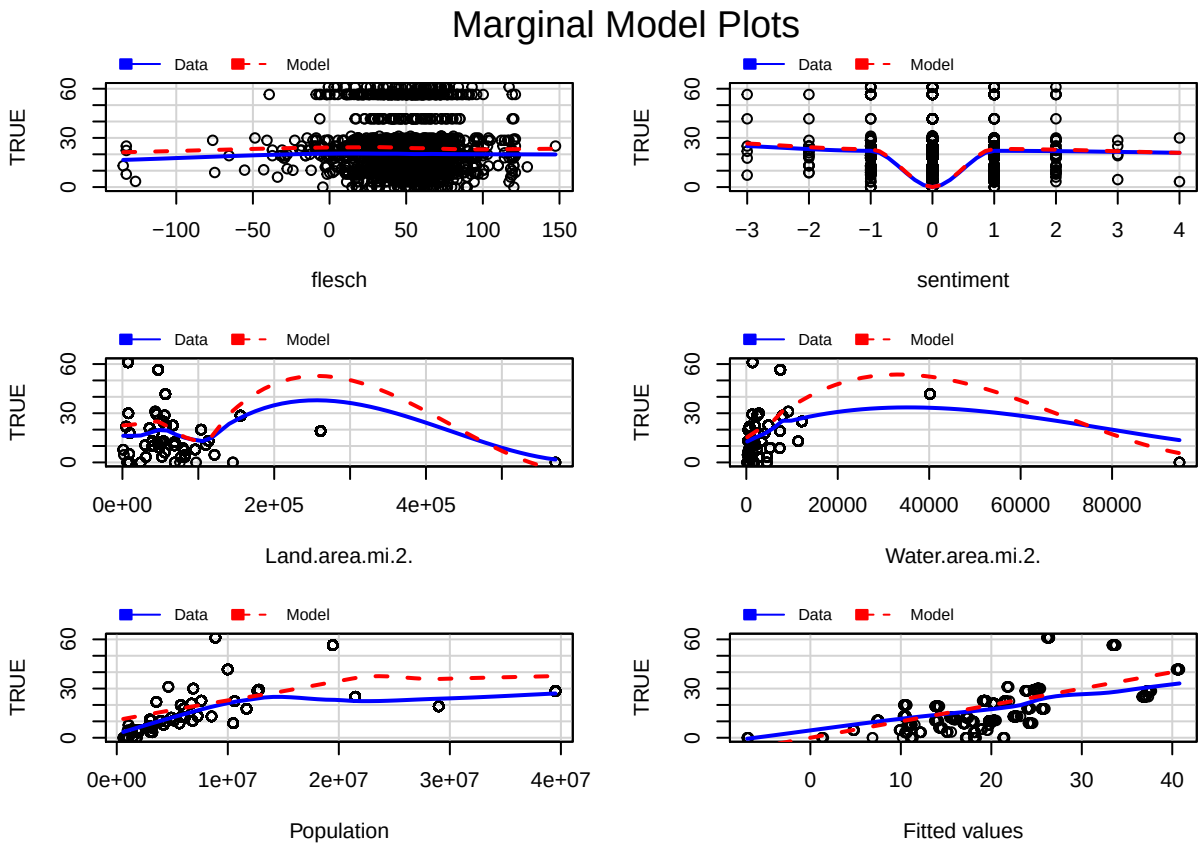



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
10
20
30
40
Fitted values
lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + Land.area.mi.2. + ...




```
mmps(fleschlm)
```




```
summary(fleschlm)
```

```
##
## Call:
## lm(formula = sqrt(total_march_covid_deaths) ~ flesch + sentiment +
##     Land.area.mi.2. + Water.area.mi.2. + Population, data = c_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.456  -8.847  -0.770   4.910  34.911
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.881e+01  6.931e-01  27.137   <2e-16 ***
## flesch           -2.541e-03  9.907e-03  -0.257    0.798
## sentiment        -1.676e-02  4.043e-01  -0.041    0.967
## Land.area.mi.2.  -1.308e-04  4.577e-06 -28.584   <2e-16 ***
## Water.area.mi.2.  5.106e-04  3.220e-05  15.859   <2e-16 ***
## Population        8.866e-07  2.877e-08  30.819   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.76 on 1894 degrees of freedom
## Multiple R-squared:  0.4082, Adjusted R-squared:  0.4066
## F-statistic: 261.2 on 5 and 1894 DF,  p-value: < 2.2e-16
```

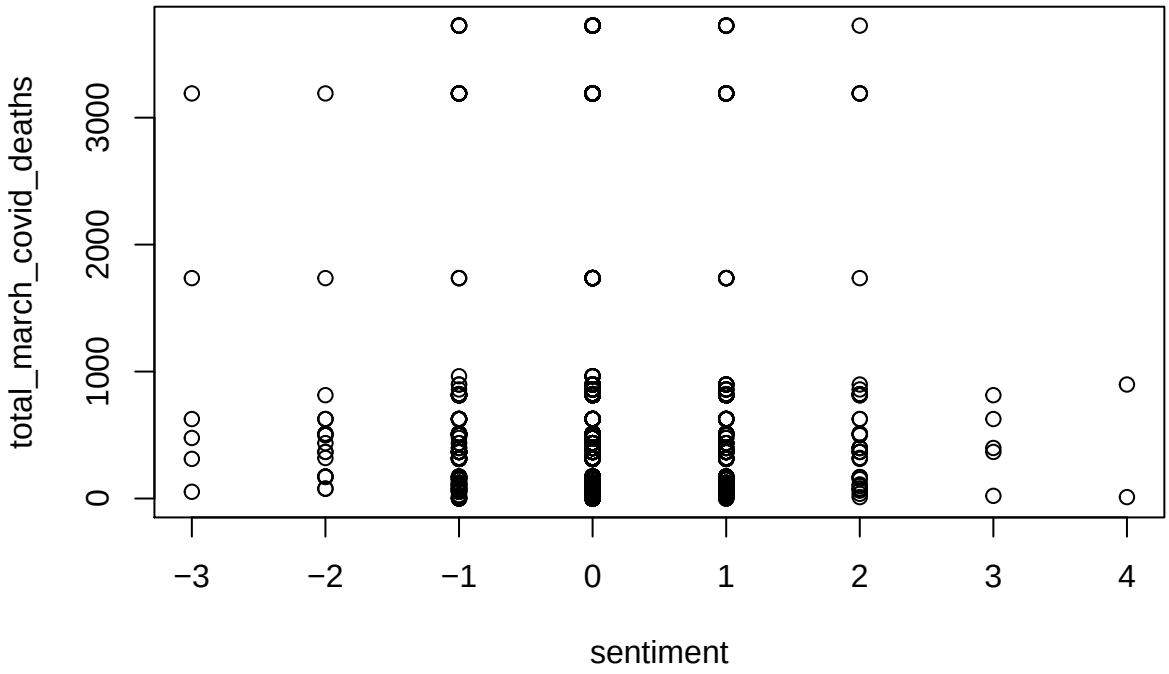
```r
anova(fleschlm)
```

```
## Analysis of Variance Table
##
## Response: sqrt(total_march_covid_deaths)
##                     Df Sum Sq Mean Sq  F value Pr(>F)
## flesch               1     43      43   0.3077 0.5791
## sentiment            1     86      86   0.6230 0.4300
## Land.area.mi.2.      1  11580   11580  83.7060 <2e-16 ***
## Water.area.mi.2.     1  37586   37586 271.6802 <2e-16 ***
## Population           1 131407  131407 949.8337 <2e-16 ***
## Residuals         1894 262029     138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Checking the actual variable correlations --> Since our model isn't the best we can try to check direc
#Theres seems to be almost no correlation between either of the variables and number of covid deaths
cor(c_t$total_march_covid_deaths, c_t$sentiment)
```

```
## [1] -0.01459617
```

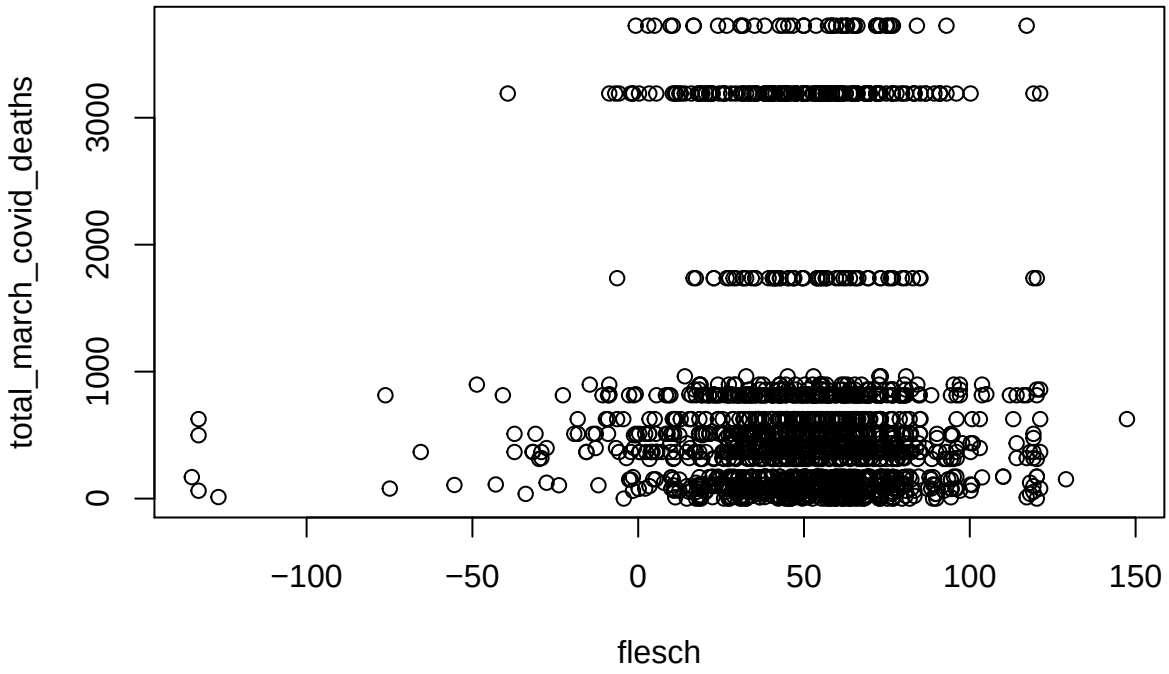
```r
plot(total_march_covid_deaths ~ sentiment, c_t)
```



```r
cor(c_t$total_march_covid_deaths, c_t$flesch)
```

```
## [1] -0.007971669
```

```r
plot(total_march_covid_deaths ~ flesch,c_t)
```




```r
#no correlation between our 2 predictors either
cor(c_t$flesch, c_t$sentiment)
```

```
## [1] 0.009991998
```

```r
#no correlation between clesch and favorite count or sentiment
cor(c_t$favorite_count, c_t$flesch)
```

```
## [1] -0.03419367
```

```r
cor(c_t$favorite_count, c_t$sentiment)
```

```
## [1] -0.003735875
```

```r
#adding airports as a predictor
airports <- read.csv("airports.csv", stringsAsFactors = F)
airports <- airports[,1:2]
c_t$Name <- toupper(c_t$Name)

c_t <- c_t %>% left_join(airports, by = c("Name" = "State"))
c_t <- c_t %>%
  select(num_Airports, everything())

# Number of airports also fails to be an accurate predictor of Covid Deaths.
# Our best predictors for number of Covid Deaths in a state are the population, land area, and water ar
fleschlm2 <- lm(sqrt(total_march_covid_deaths) ~ flesch + sentiment + num_Airports + Land.area.mi.2. + W
summary(fleschlm2)
```

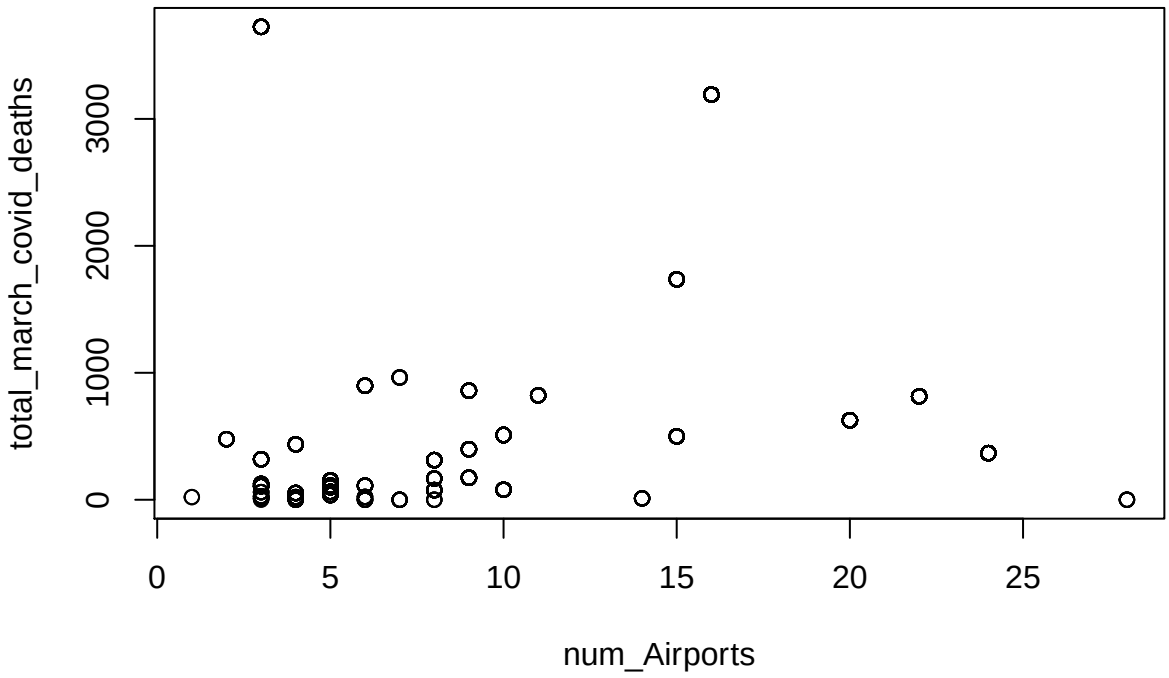
```
##
## Call:
## lm(formula = sqrt(total_march_covid_deaths) ~ flesch + sentiment +
##     num_Airports + Land.area.mi.2. + Water.area.mi.2. + Population,
##     data = c_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.160  -7.597  -2.466   4.495  36.520
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.790e+01  7.802e-01  22.937  < 2e-16 ***
## flesch           -4.982e-04  1.009e-02  -0.049    0.961
## sentiment        -8.550e-02  4.081e-01  -0.210    0.834
## num_Airports      4.903e-01  1.080e-01   4.540 5.99e-06 ***
## Land.area.mi.2.  -1.417e-04  5.016e-06 -28.248  < 2e-16 ***
## Water.area.mi.2.  4.047e-04  3.869e-05  10.460  < 2e-16 ***
## Population        6.571e-07  5.625e-08  11.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.71 on 1848 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.4171, Adjusted R-squared:  0.4152
## F-statistic: 220.4 on 6 and 1848 DF,  p-value: < 2.2e-16
```

```r
anova(fleschlm2)
```

```
## Analysis of Variance Table
##
## Response: sqrt(total_march_covid_deaths)
##                    Df Sum Sq Mean Sq  F value    Pr(>F)
## flesch              1     43      43   0.3113    0.5769
## sentiment           1    149     149   1.0855    0.2976
## num_Airports        1  37312   37312 272.2540 < 2.2e-16 ***
## Land.area.mi.2.     1 120731  120731 880.9360 < 2.2e-16 ***
## Water.area.mi.2.    1   4289    4289  31.2977 2.544e-08 ***
## Population          1  18699   18699 136.4380 < 2.2e-16 ***
```

```
## Residuals        1848 253267     137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(total_march_covid_deaths ~ num_Airports, c_t )
```

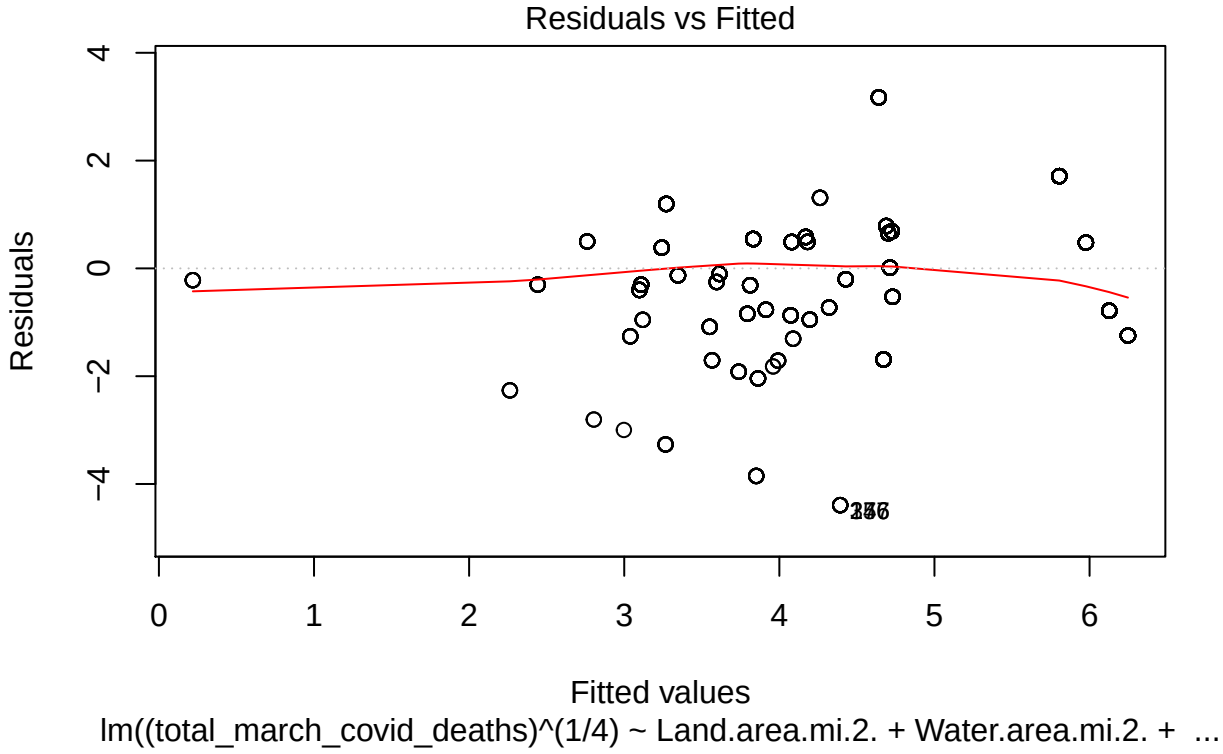


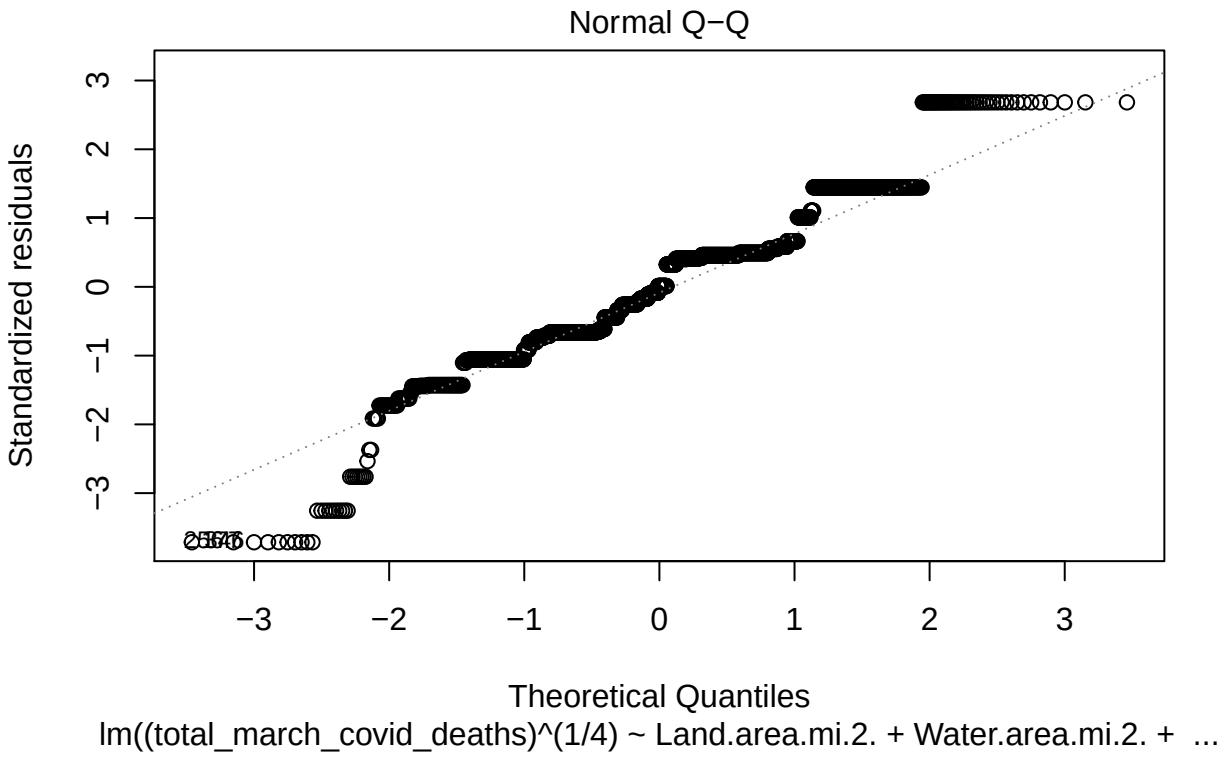

```
# This model seems to be the best predictor of Covid Cases we have as it explains ~48% of the variabili
summary(lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. + Population + num_Airp
```

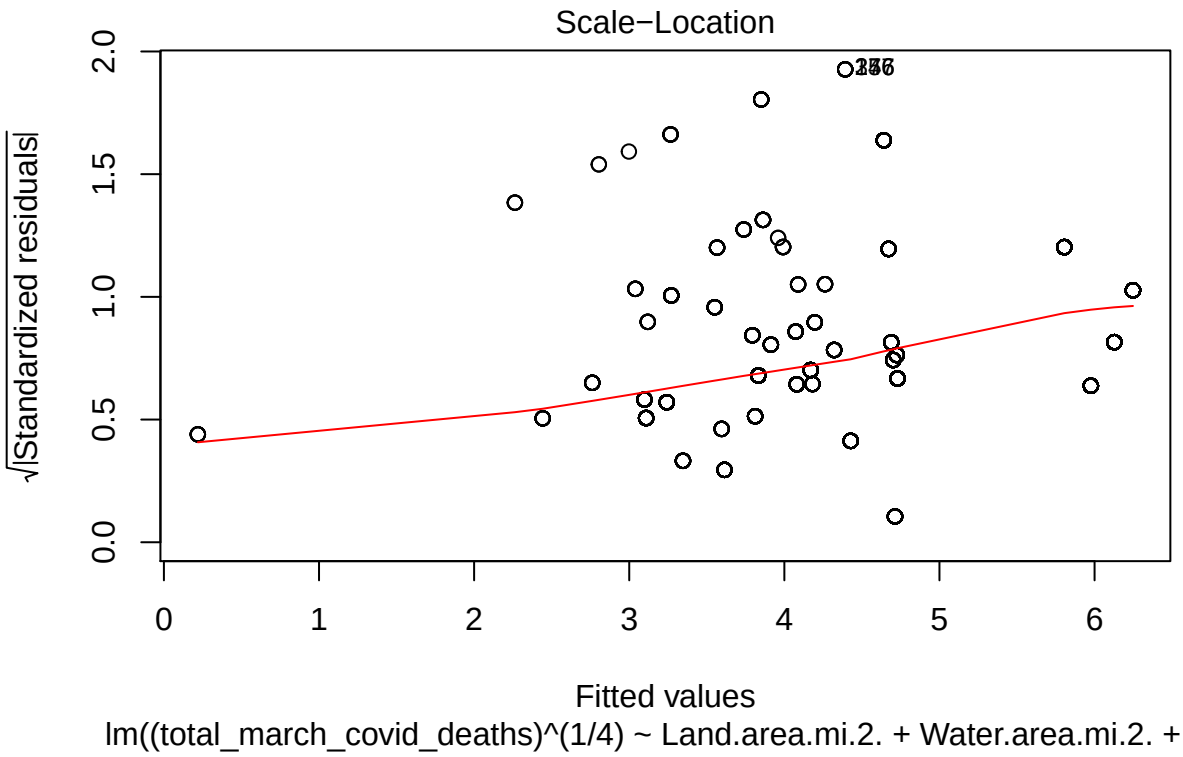
```
##
## Call:
## lm(formula = (total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. +
##     Water.area.mi.2. + Population + num_Airports, data = c_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3921 -0.7855  0.0130  0.5822  3.1713
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.839e+00  5.750e-02  66.765  < 2e-16 ***
## Land.area.mi.2.  -1.477e-05  5.070e-07 -29.126  < 2e-16 ***
## Water.area.mi.2.  3.392e-05  3.910e-06   8.676  < 2e-16 ***
## Population        7.881e-08  5.683e-09  13.869  < 2e-16 ***
## num_Airports      5.481e-02  1.091e-02   5.026 5.49e-07 ***
```

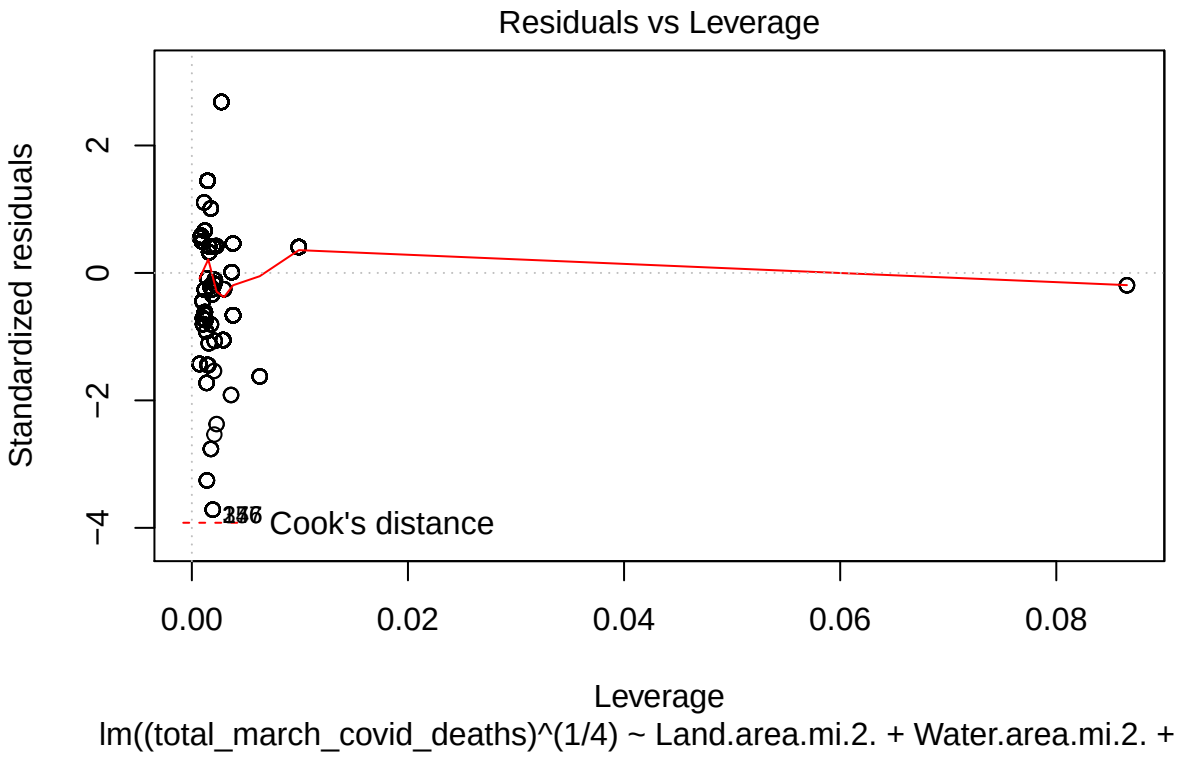
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.184 on 1850 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.4589, Adjusted R-squared:  0.4577
## F-statistic: 392.2 on 4 and 1850 DF,  p-value: < 2.2e-16
```

```
plot(lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. + Population + num_Airpor
```

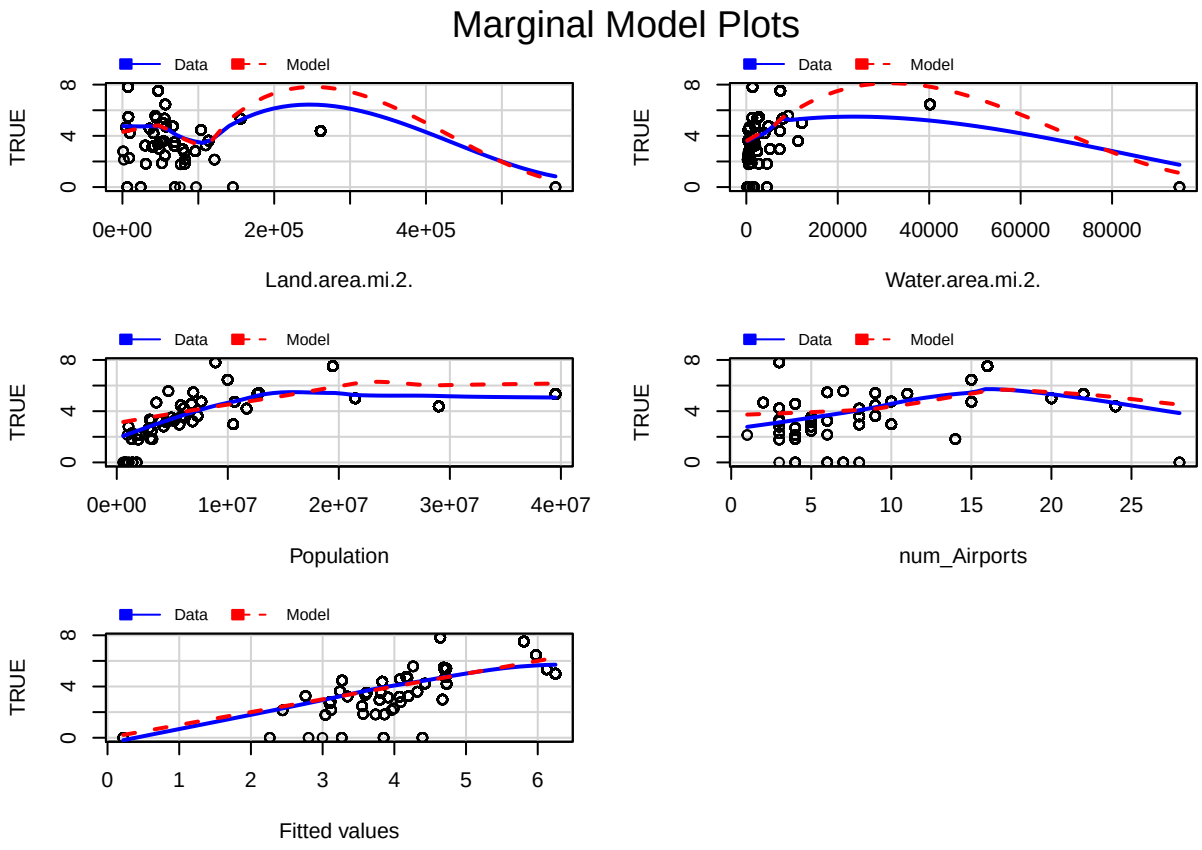



Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. + ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
6
Fitted values
lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. + ...




```
mmps(lm((total_march_covid_deaths)^(1/4) ~ Land.area.mi.2. + Water.area.mi.2. + Population + num_Airpor
```


Marginal Model Plots
Data
Model
TRUE
0
4
8
0e+00
2e+05
4e+05
Land.area.mi.2.
0
20000
40000
60000
80000
Water.area.mi.2.
0e+00
1e+07
2e+07
3e+07
4e+07
Population
0
5
10
15
20
25
num_Airports
0
1
2
3
4
5
6
Fitted values