Hendel_FINAL

Dalit Hendel

12/11/2021

```
## load libraries
install.packages("dplyr")
```

```
## Installing package into 'C:/Users/dhende01/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\dhende01\AppData\Local\Temp\RtmpsfdAzz\downloaded_packages
```

```
install.packages("readxl")
```

```
## Installing package into 'C:/Users/dhende01/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'readxl' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\dhende01\AppData\Local\Temp\RtmpsfdAzz\downloaded_packages
```

```
install.packages("gapminder")
```

```
## Installing package into 'C:/Users/dhende01/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'gapminder' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\dhende01\AppData\Local\Temp\RtmpsfdAzz\downloaded_packages
```

```
install.packages('data.table')
```

```
## Installing package into 'C:/Users/dhende01/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'data.table' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\dhende01\AppData\Local\Temp\RtmpsfdAzz\downloaded_packages
```

```
library("gapminder")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.2     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readxl)
library(data.table)

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose

library(dplyr)
library(ggplot2)
library(ggthemes)
```

Loading in and cleaning the AQI data

```
#reading in AQI data by State and county 2015-2018
aqi2015 <- read.csv('FINALproject/AQI_2011-
2021/annual_aqi_by_county_2015.csv')
aqi2016 <- read.csv('FINALproject/AQI_2011-
2021/annual_aqi_by_county_2016.csv')
aqi2017 <- read.csv('FINALproject/AQI_2011-
2021/annual_aqi_by_county_2017.csv')
aqi2018 <- read.csv('FINALproject/AQI_2011-
2021/annual_aqi_by_county_2018.csv')

#filtering for only California in AQI
aqi2015c <- aqi2015 %>% filter(State == 'California')
aqi2016c <- aqi2016 %>% filter(State == 'California')
aqi2017c <- aqi2017 %>% filter(State == 'California')
aqi2018c <- aqi2018 %>% filter(State == 'California')

#dropping columns we do not need for the merge
aqi2015cc <- aqi2015c %>% select(-c(State, Year, X90th.Percentile.AQI,
Days.CO, Days.NO2, Days.Ozone, Days.SO2, Days.PM2.5, Days.PM10))
aqi2016cc <- aqi2016c %>% select(-c(State, Year, X90th.Percentile.AQI,
```

```
Days.CO, Days.NO2, Days.Ozone, Days.SO2, Days.PM2.5, Days.PM10))
aqi2017cc <- aqi2017c %>% select(-c(State, Year, X90th.Percentile.AQI,
Days.CO, Days.NO2, Days.Ozone, Days.SO2, Days.PM2.5, Days.PM10))
aqi2018cc <- aqi2018c %>% select(-c(State, Year, X90th.Percentile.AQI,
Days.CO, Days.NO2, Days.Ozone, Days.SO2, Days.PM2.5, Days.PM10))

table(aqi2018cc$Days.with.AQI) # there are for each about 4 values with low
counts i wont address it to save time and I need it for every county ~ will
fix  by finding perportion later

##
## 180 291 348 350 357 363 364 365
##   1   1   1   1   1   2   4  42

#combine aqi years 2015-2016 and 2017-2018
aqi2015_2016 <- bind_rows(aqi2015cc, aqi2016cc)
aqi2017_2018 <- bind_rows(aqi2017cc, aqi2018cc)

#grouping by county and summing the days values and averaging AQI values
  #for 2015-2016
aqi2015_2016_county <- aqi2015_2016 %>%
  group_by(County) %>%
  summarise(days = sum(Days.with.AQI),
            sum(Good.Days),
            sum(Moderate.Days),
            sum(Unhealthy.for.Sensitive.Groups.Days),
            sum(Unhealthy.Days),
            sum(Very.Unhealthy.Days),
            sum(Hazardous.Days),
            BADdays = sum(c(Unhealthy.Days, Very.Unhealthy.Days)),
            mean(Max.AQI),
            mean(Median.AQI))

## `summarise()` ungrouping output (override with `.groups` argument)

aqi2015_2016_county_bad <- data.frame(aqi2015_2016_county$County,
aqi2015_2016_county$BADdays, aqi2015_2016_county$days)

#now all that for 2017-2018
aqi2017_2018_county <- aqi2017_2018 %>%
  group_by(County) %>%
  summarise(days = sum(Days.with.AQI),
            sum(Good.Days),
            sum(Moderate.Days),
            sum(Unhealthy.for.Sensitive.Groups.Days),
            sum(Unhealthy.Days),
            sum(Very.Unhealthy.Days),
            sum(Hazardous.Days),
            BADdays = sum(c(Unhealthy.Days, Very.Unhealthy.Days)),
```

```
            mean(Max.AQI),
            mean(Median.AQI))

## `summarise()` ungrouping output (override with `.groups` argument)

aqi2017_2018_county_bad <- data.frame(aqi2017_2018_county$County,
aqi2017_2018_county$BADdays, aqi2017_2018_county$days)

#merging into one df for simplicity
  #renaming first column
names(aqi2015_2016_county_bad)[names(aqi2015_2016_county_bad) ==
'aqi2015_2016_county.County'] <- 'County'
names(aqi2017_2018_county_bad)[names(aqi2017_2018_county_bad) ==
'aqi2017_2018_county.County'] <- 'County'
  #finding perportion of bad days over days recorded
aqi2015_2016_county_bad$proportion.bad.2015_2016 <-
(aqi2015_2016_county_bad$aqi2015_2016_county.BADdays /
aqi2015_2016_county_bad$aqi2015_2016_county.days) * 100

aqi2017_2018_county_bad$proportion.bad.2017_2018 <-
(aqi2017_2018_county_bad$aqi2017_2018_county.BADdays /
aqi2017_2018_county_bad$aqi2017_2018_county.days) * 100

 #joining them into one df
aqi2015_2018_bad <- inner_join(aqi2015_2016_county_bad,
aqi2017_2018_county_bad, by='County')
aqi2015_2018_bad$proportion.bad <- NULL
```

Loading in and cleaning the AQI data

```
#reading in asthma data for all age groups 2015-2018 (2 year periods)
asthma <- read_excel('FINALproject/current-asthma-prevalence-by-county-
2015_2018.xlsx')
asthma <- as.data.frame(asthma)
  #drop columns and change names
asthma <- asthma[, -c(3, 4, 6, 7, 8)]
colnames(asthma)[3] <- 'asthma.PREVALENCE'
colnames(asthma)[1] <- 'County'
#group by county and years ## AND REPLACE THE MANY NA VALUES WITH MEAN
asthma_c_d_1516 <- asthma %>% filter(YEARS == '2015-2016') %>%

mutate(asthma.PREVALENCE=ifelse(is.na(asthma.PREVALENCE),mean(asthma.PREVALEN
CE,na.rm=T),asthma.PREVALENCE)) %>% group_by(County) %>% summarise(MEAN_15_16
= mean(asthma.PREVALENCE))

## `summarise()` ungrouping output (override with `.groups` argument)

asthma_c_d_1718 <- asthma %>% filter(YEARS == '2017-2018')%>%

mutate(asthma.PREVALENCE=ifelse(is.na(asthma.PREVALENCE),mean(asthma.PREVALEN
CE,na.rm=T),asthma.PREVALENCE)) %>% group_by(County) %>% summarise(MEAN_17_18
= mean(asthma.PREVALENCE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

  #combine the two into one df
asthma_c_d <- inner_join(asthma_c_d_1516, asthma_c_d_1718, by = 'County')
```

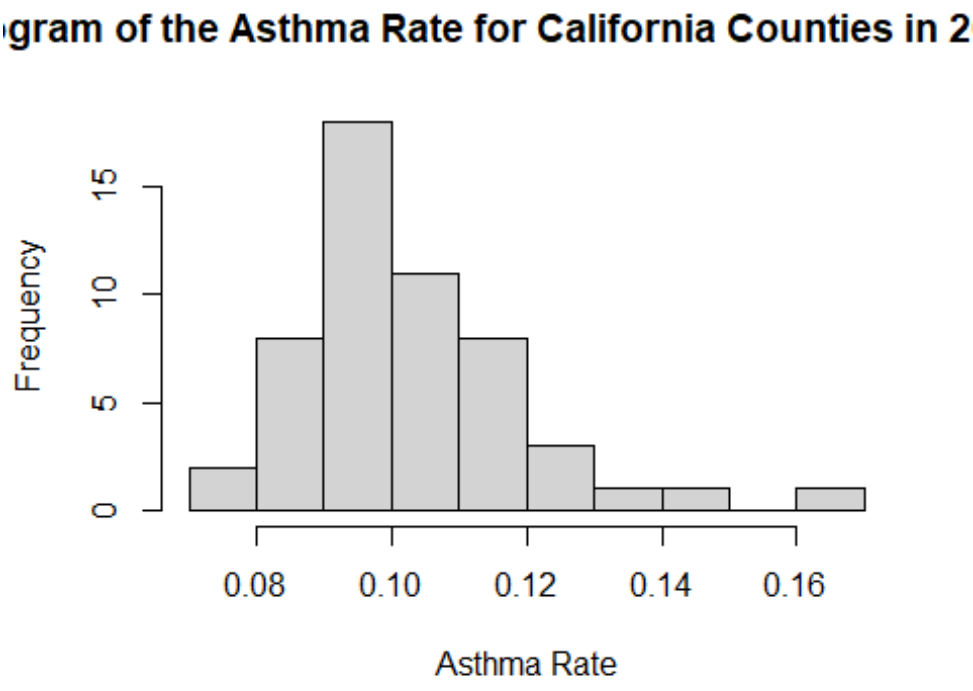
Merging AQI and aesthma dfs

```
air <- inner_join(aqi2015_2018_bad, asthma_c_d, by = 'County')
head(air)

##         County aqi2015_2016_county.BADdays aqi2015_2016_county.days
## 1      Alameda                           2                       731
## 2       Amador                           0                       730
## 3        Butte                           1                       731
## 4    Calaveras                           3                       731
## 5       Colusa                           2                       731
## 6 Contra Costa                           0                       731
##   proportion.bad.2015_2016 aqi2017_2018_county.BADdays
aqi2017_2018_county.days
## 1                0.2735978                          17
730
## 2                0.0000000                           0
726
## 3                0.1367989                          15
730
## 4                0.4103967                           8
718
## 5                0.2735978                          19
730
## 6                0.0000000                          15
730
##   proportion.bad.2017_2018 MEAN_15_16 MEAN_17_18
## 1                 2.328767 0.11180944 0.09732992
## 2                 0.000000 0.10315052 0.14088203
## 3                 2.054795 0.10355637 0.11917956
## 4                 1.114206 0.10315052 0.14088203
## 5                 2.602740 0.09049931 0.12619956
## 6                 2.054795 0.10020780 0.12063617

str(air)

## 'data.frame':    53 obs. of  9 variables:
##  $ County                    : chr  "Alameda" "Amador" "Butte"
"Calaveras" ...
##  $ aqi2015_2016_county.BADdays: int  2 0 1 3 2 0 0 14 49 0 ...
##  $ aqi2015_2016_county.days   : int  731 730 731 731 731 731 291 731 731
727 ...
##  $ proportion.bad.2015_2016   : num  0.274 0 0.137 0.41 0.274 ...
##  $ aqi2017_2018_county.BADdays: int  17 0 15 8 19 15 2 10 46 7 ...
##  $ aqi2017_2018_county.days   : int  730 726 730 718 730 730 478 727 730
728 ...
##  $ proportion.bad.2017_2018   : num  2.33 0 2.05 1.11 2.6 ...
```

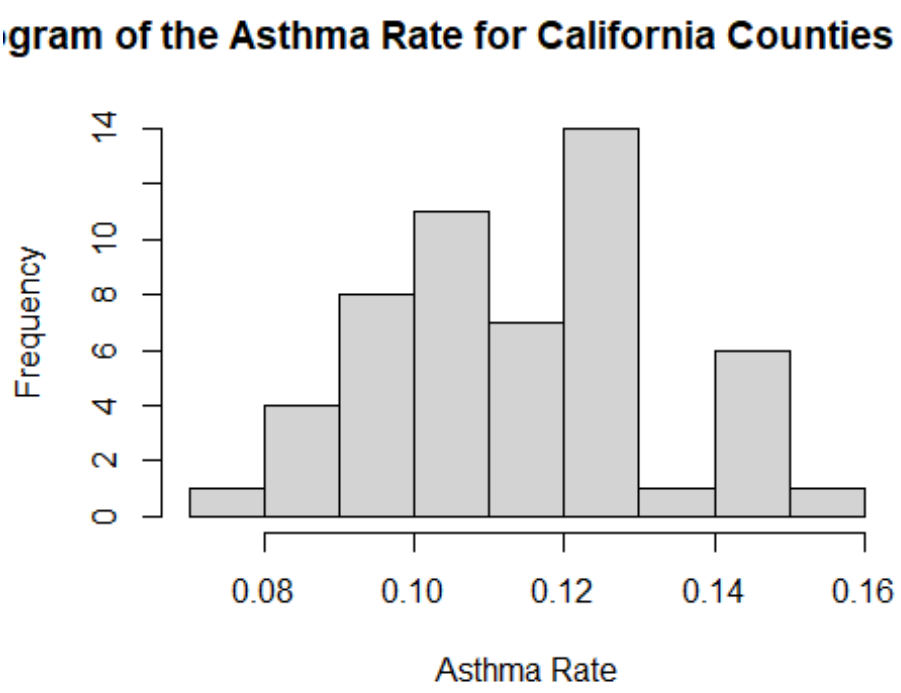
```
##  $ MEAN_15_16                : num  0.1118 0.1032 0.1036 0.1032 0.0905
...
##  $ MEAN_17_18                : num  0.0973 0.1409 0.1192 0.1409 0.1262
...

dim(air)

## [1] 53  9
```

Looking at Histograms of the data

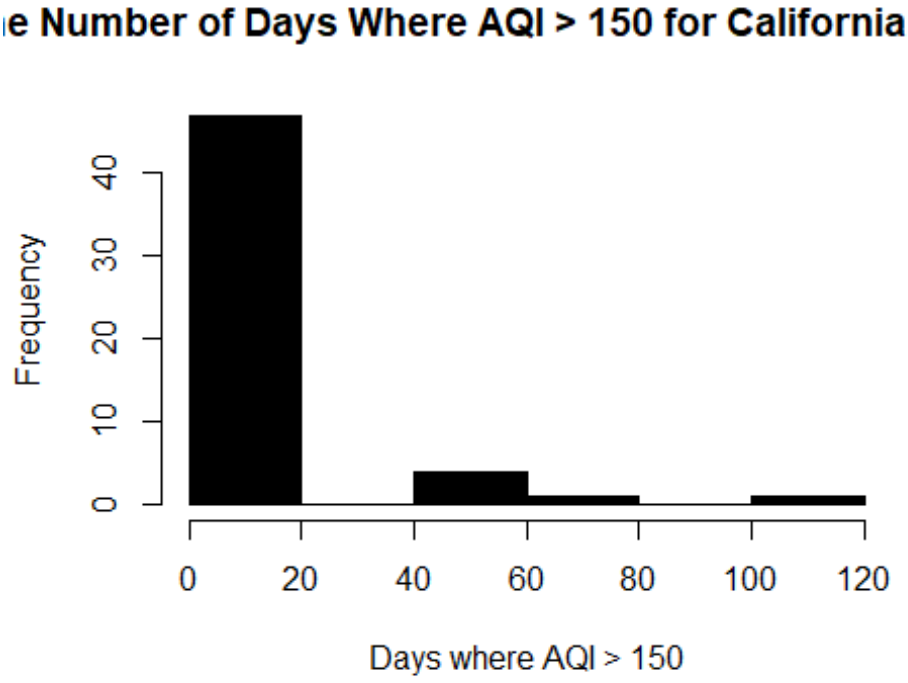
```
#checking distributions
hist(air$MEAN_15_16, main = 'Histogram of the Asthma Rate for California
Counties in 2015-2016', xlab='Asthma Rate') #normal
```



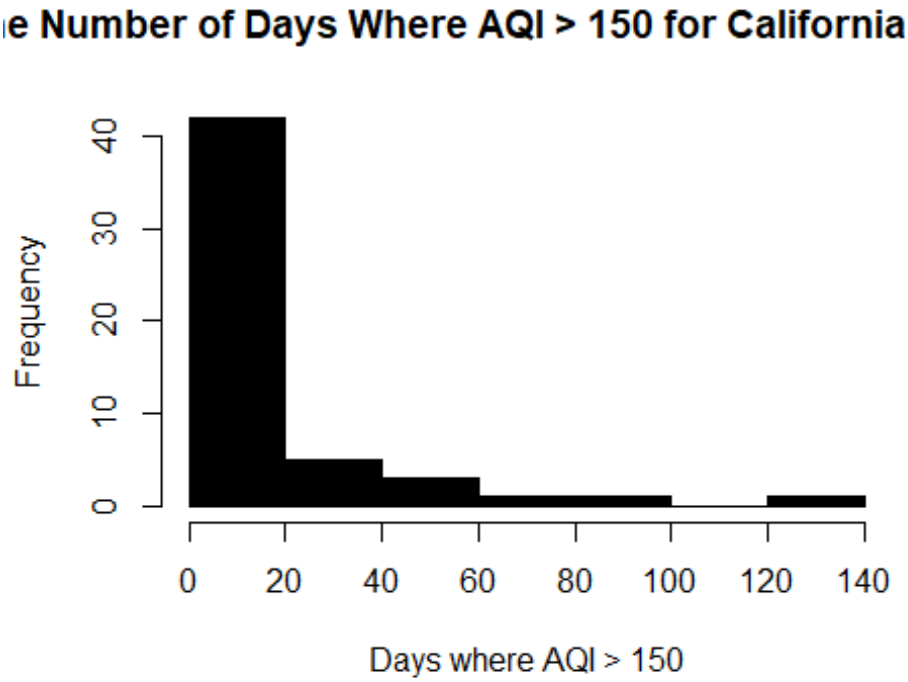

```
hist(air$MEAN_17_18,  main = 'Histogram of the Asthma Rate for California
Counties in 2017-2018', xlab='Asthma Rate') #normal
```



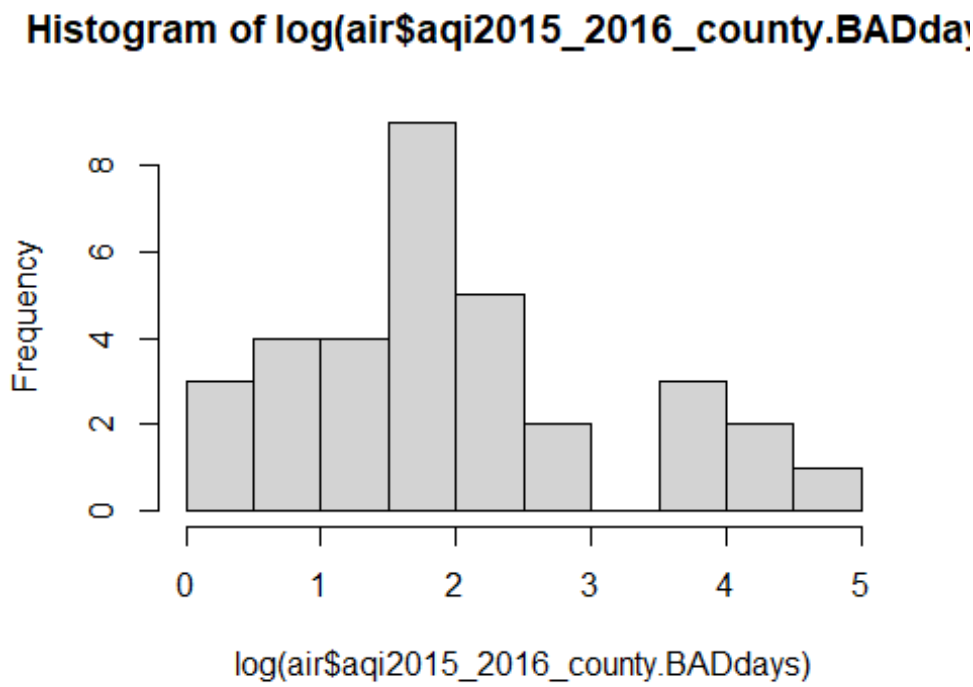

```
hist(air$aqi2015_2016_county.BADdays, main = 'Histogram of the Number of Days
Where AQI > 150 for California Counties in 2015-2016', xlab='Days where AQI >
150', col = 'black') #not normal
```



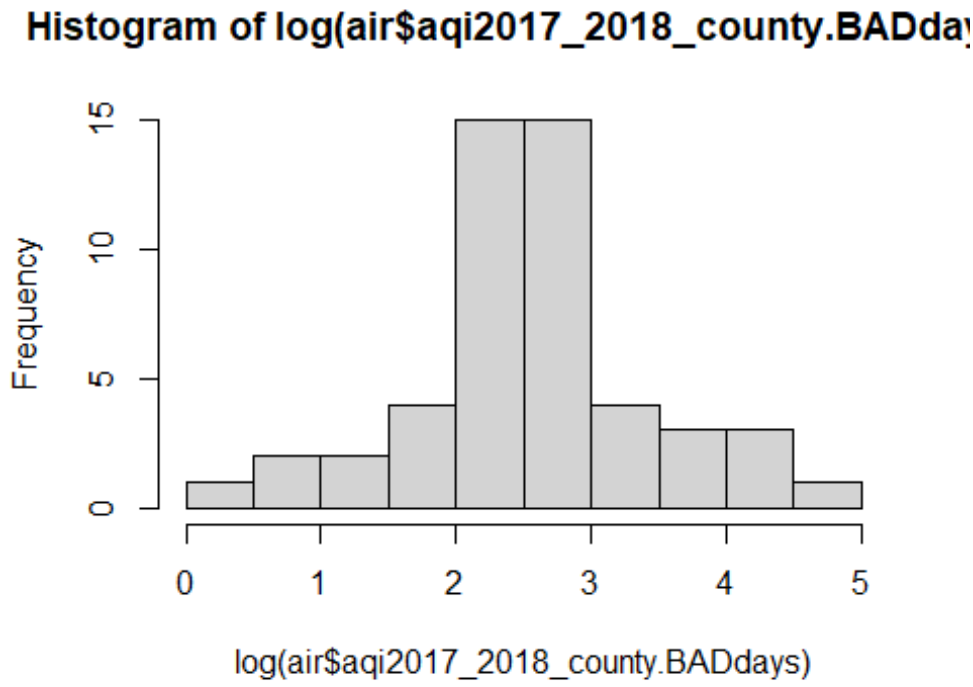
```r
hist(air$aqi2017_2018_county.BADdays, main = 'Histogram of the Number of Days Where AQI > 150 for California Counties in 2017-2018', xlab='Days where AQI > 150', col = 'black') #not normal
```




```r
hist(log(air$aqi2015_2016_county.BADdays)) #log helps
```

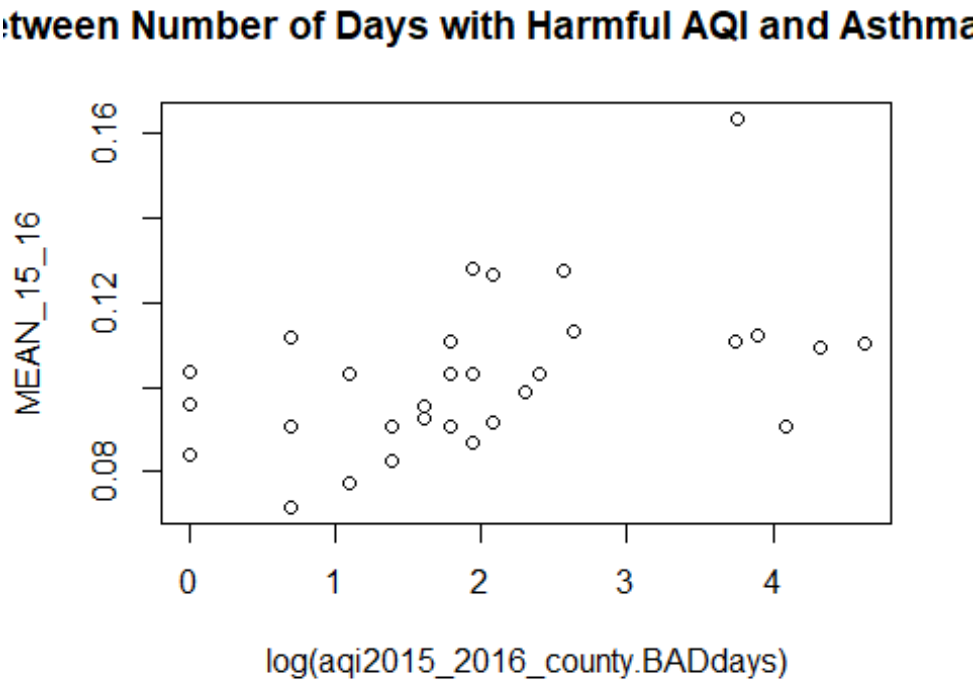



```
hist(log(air$aqi2017_2018_county.BADdays)) #Log helps
```

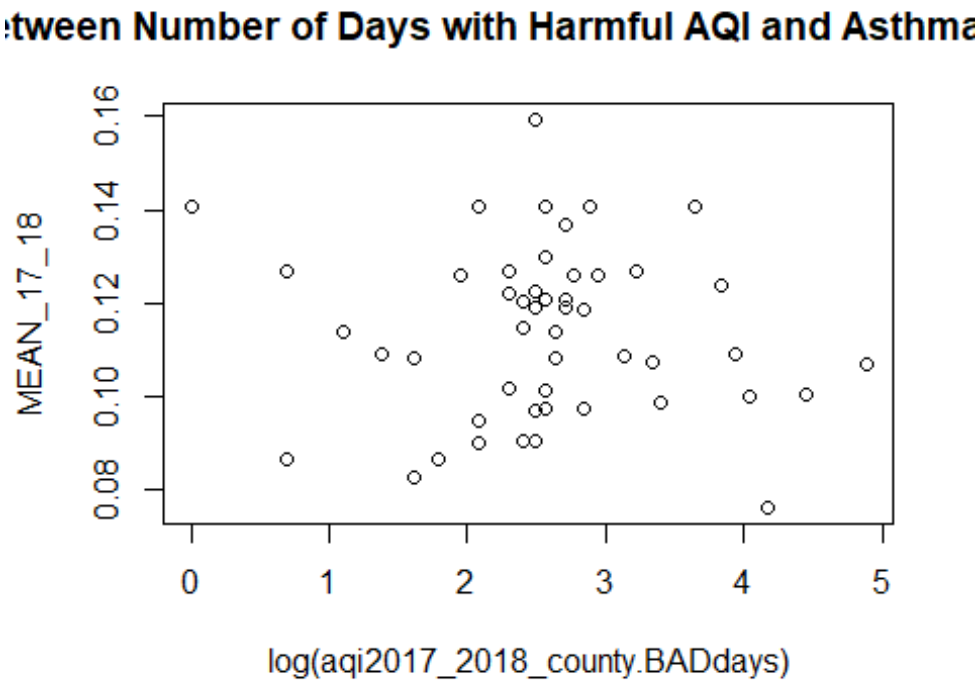


Plotting the data

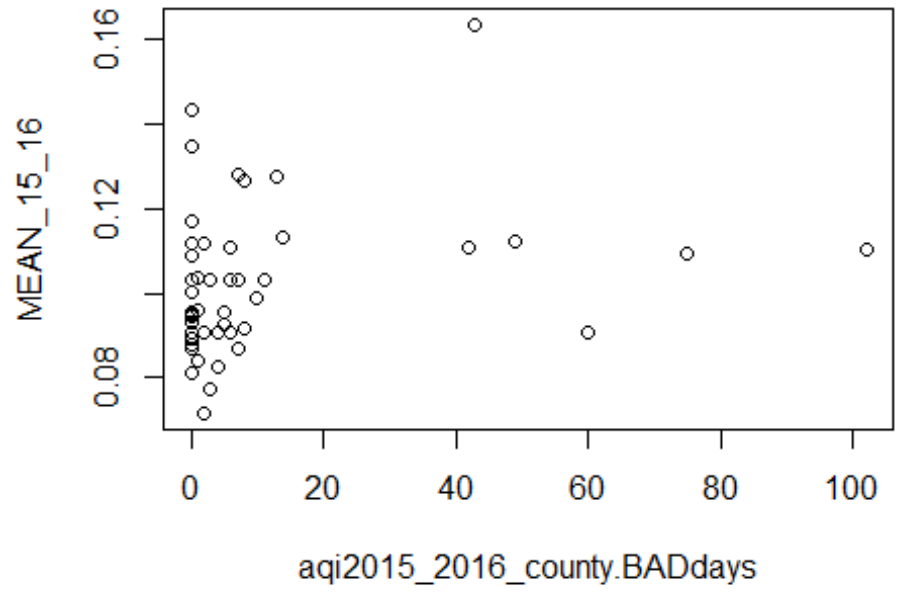
```
plot(MEAN_15_16~log(aqi2015_2016_county.BADdays), data = air, main=
'Relationship between Number of Days with Harmful AQI and Asthma rates in
2015-2016')
```



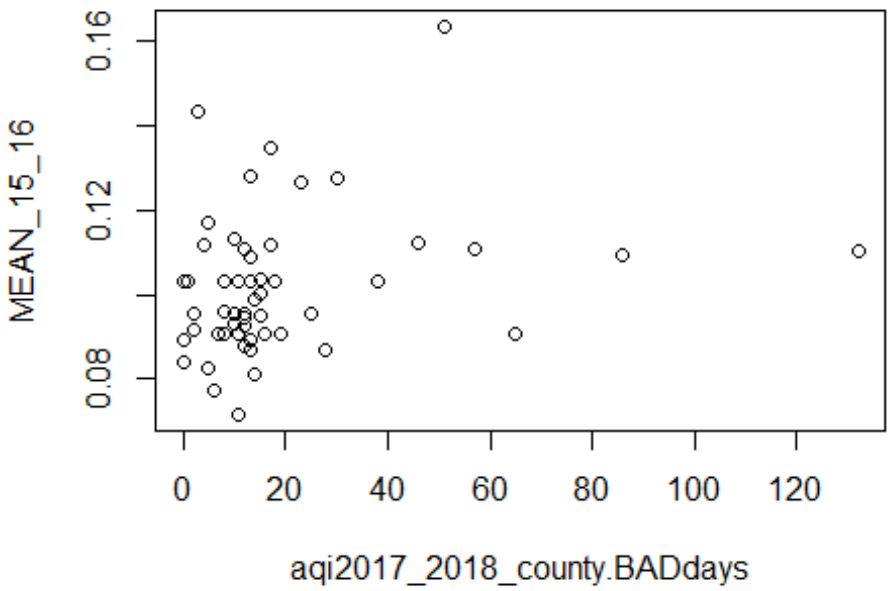

```
plot(MEAN_17_18~log(aqi2017_2018_county.BADdays), data = air, main=
'Relationship between Number of Days with Harmful AQI and Asthma rates in
2015-2016')
```




```
plot(MEAN_15_16~(aqi2015_2016_county.BADdays), data = air)
```




```
plot(MEAN_15_16~(aqi2017_2018_county.BADdays), data = air)
```



Correlations of Asthma and AQI

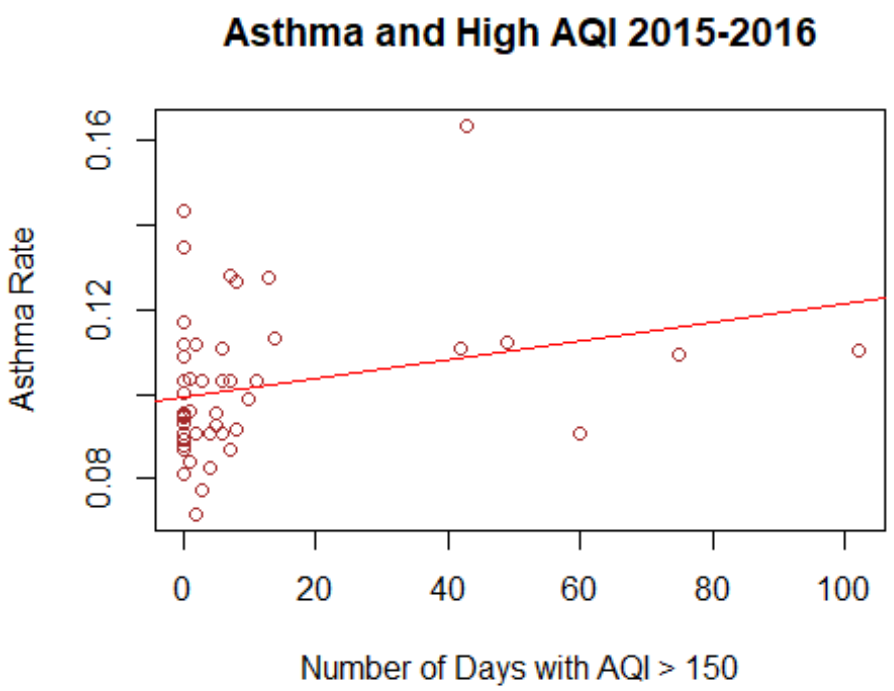
```
cor.test(air$MEAN_15_16, (air$aqi2015_2016_county.BADdays)) #cor: 0.2675652

##
##  Pearson's product-moment correlation
##
## data:  air$MEAN_15_16 and (air$aqi2015_2016_county.BADdays)
## t = 1.9831, df = 51, p-value = 0.05275
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002941326  0.501583862
## sample estimates:
##       cor
## 0.2675652

cor.test(air$MEAN_15_16, log(air$aqi2015_2016_county.BADdays), method =
"spearman", exact=FALSE) #rho: 0.3039764

##
##  Spearman's rank correlation rho
##
## data:  air$MEAN_15_16 and log(air$aqi2015_2016_county.BADdays)
## S = 17264, p-value = 0.02691
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3039764
```

```
cor.test(air$MEAN_17_18, (air$aqi2017_2018_county.BADdays)) #cor:

##
##  Pearson's product-moment correlation
##
## data:  air$MEAN_17_18 and (air$aqi2017_2018_county.BADdays)
## t = -1.0572, df = 51, p-value = 0.2954
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4008734  0.1289504
## sample estimates:
##        cor
## -0.1464478

cor.test(air$MEAN_17_18, log(air$aqi2017_2018_county.BADdays), method =
"spearman", exact=FALSE) #rho:

##
##  Spearman's rank correlation rho
##
## data:  air$MEAN_17_18 and log(air$aqi2017_2018_county.BADdays)
## S = 24522, p-value = 0.9356
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.0113765
```

Regressions of the data

```
#REGRESSION
airmod1 <- lm(MEAN_15_16~aqi2015_2016_county.BADdays, data=air)
summary(airmod1) #p for AQI is 0.0528

##
## Call:
## lm(formula = MEAN_15_16 ~ aqi2015_2016_county.BADdays, data = air)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028343 -0.009803 -0.003949  0.003721  0.054750
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.0994297  0.0024971  39.819   <2e-16 ***
## aqi2015_2016_county.BADdays 0.0002198  0.0001108   1.983   0.0528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01631 on 51 degrees of freedom
## Multiple R-squared:  0.07159,    Adjusted R-squared:  0.05339
## F-statistic: 3.933 on 1 and 51 DF,  p-value: 0.05275
```

```
plot(MEAN_15_16~(aqi2015_2016_county.BADdays), data = air, main = 'Asthma and
High AQI 2015-2016', xlab = 'Number of Days with AQI > 150', ylab = 'Asthma
Rate', col = 'brown')
abline(airmod1, col = 'red')
```




```
airmod2 <- lm(MEAN_17_18~aqi2017_2018_county.BADdays, data=air)
summary(airmod2) #p for AQI is  0.2954

##
## Call:
## lm(formula = MEAN_17_18 ~ aqi2017_2018_county.BADdays, data = air)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032919 -0.013375  0.000052  0.012089  0.044612
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.1161323  0.0032546  35.682   <2e-16 ***
## aqi2017_2018_county.BADdays -0.0001161  0.0001098  -1.057    0.295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0183 on 51 degrees of freedom
## Multiple R-squared:  0.02145,    Adjusted R-squared:  0.00226
## F-statistic: 1.118 on 1 and 51 DF,  p-value: 0.2954
```

```
airmod3 <-
lm(MEAN_17_18~aqi2017_2018_county.BADdays+aqi2015_2016_county.BADdays,
data=air)
summary(airmod3) #p for 17-18 is 0.348  #p for 15-19 is 0.169

##
## Call:
## lm(formula = MEAN_17_18 ~ aqi2017_2018_county.BADdays +
aqi2015_2016_county.BADdays,
##     data = air)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030129 -0.015680  0.001641  0.010818  0.042460
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.1132663  0.0038243  29.618   <2e-16 ***
## aqi2017_2018_county.BADdays  0.0003021  0.0003191   0.947    0.348
## aqi2015_2016_county.BADdays -0.0005038  0.0003613  -1.394    0.169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01813 on 50 degrees of freedom
## Multiple R-squared:  0.05807,    Adjusted R-squared:  0.02039
## F-statistic: 1.541 on 2 and 50 DF,  p-value: 0.2241

airmod4 <- lm(abs(MEAN_17_18 - MEAN_15_16)~abs(aqi2017_2018_county.BADdays-
aqi2015_2016_county.BADdays), data = air)
summary(airmod4) # p is 0.489 for the delta for AQI

##
## Call:
## lm(formula = abs(MEAN_17_18 - MEAN_15_16) ~
abs(aqi2017_2018_county.BADdays -
##     aqi2015_2016_county.BADdays), data = air)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.022558 -0.011883 -0.003483  0.011623  0.042755
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                       0.019358
## abs(aqi2017_2018_county.BADdays - aqi2015_2016_county.BADdays) 0.000211
##                                                                   Std. Error
## (Intercept)                                                         0.003605
## abs(aqi2017_2018_county.BADdays - aqi2015_2016_county.BADdays)   0.000303
##                                                                   t value
Pr(>|t|)
```

```
## (Intercept)                                                    5.369
1.96e-06
## abs(aqi2017_2018_county.BADdays - aqi2015_2016_county.BADdays)   0.696
0.489
##
## (Intercept)                                                   ***
## abs(aqi2017_2018_county.BADdays - aqi2015_2016_county.BADdays)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01538 on 51 degrees of freedom
## Multiple R-squared:  0.009417,   Adjusted R-squared:  -0.01001
## F-statistic: 0.4849 on 1 and 51 DF,  p-value: 0.4894
```

Plotting Variables on Map

```
#plorring cali asthma data
library(grid)
cal_counties <-map_data('county', 'california') %>%
  select(lon =long, lat, group, id=subregion)
head(cal_counties)
```

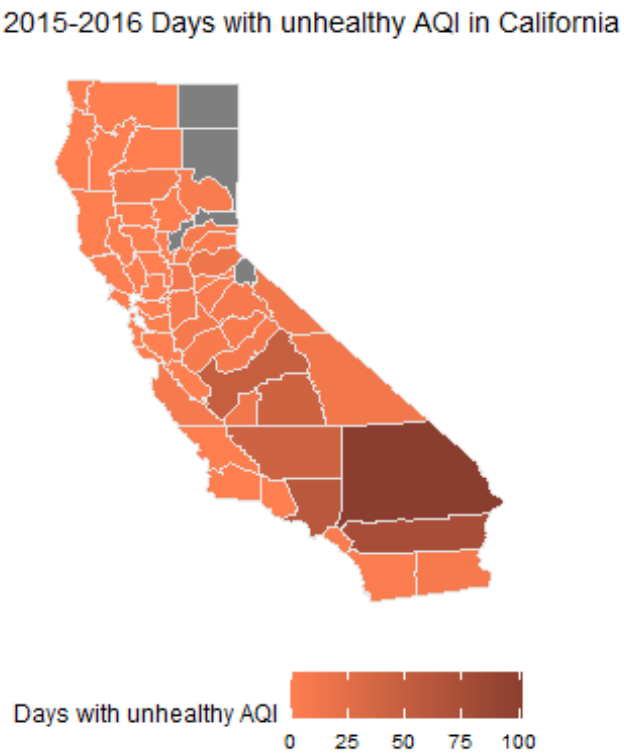
```
##         lon      lat group      id
## 1 -121.4785 37.48290     1 alameda
## 2 -121.5129 37.48290     1 alameda
## 3 -121.8853 37.48290     1 alameda
## 4 -121.8968 37.46571     1 alameda
## 5 -121.9254 37.45998     1 alameda
## 6 -121.9483 37.47717     1 alameda
```

```
  #rename county col to merge later
names(cal_counties)[names(cal_counties) == 'id'] <- 'County'
cal_counties$County <- str_to_title(cal_counties$County)
cal_df <- left_join(cal_counties, air, by = 'County')
names(cal_df)
```

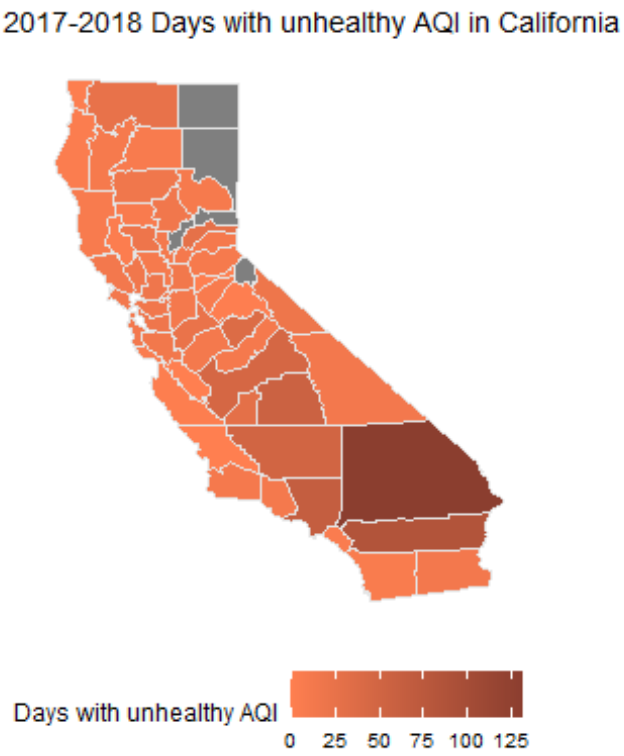
```
##  [1] "lon"                          "lat"
##  [3] "group"                        "County"
##  [5] "aqi2015_2016_county.BADdays" "aqi2015_2016_county.days"
##  [7] "proportion.bad.2015_2016"    "aqi2017_2018_county.BADdays"
##  [9] "aqi2017_2018_county.days"    "proportion.bad.2017_2018"
## [11] "MEAN_15_16"                  "MEAN_17_18"
```

```
#aqi2015_2016
p_aqi2015_2016_county.BADdays <- ggplot(cal_df, aes(x=lon, y=lat,
group=group, fill= aqi2015_2016_county.BADdays)) +
geom_polygon(color='gray90', size=0.1) + coord_map(projection = 'albers',
lat0=39, lat1=45)

pcal_aqi2015_2016_county.BADdays <- p_aqi2015_2016_county.BADdays +labs(title
= '2015-2016 Days with unhealthy AQI in California') +
  scale_fill_gradient(low = "coral", high = "coral4") +
```

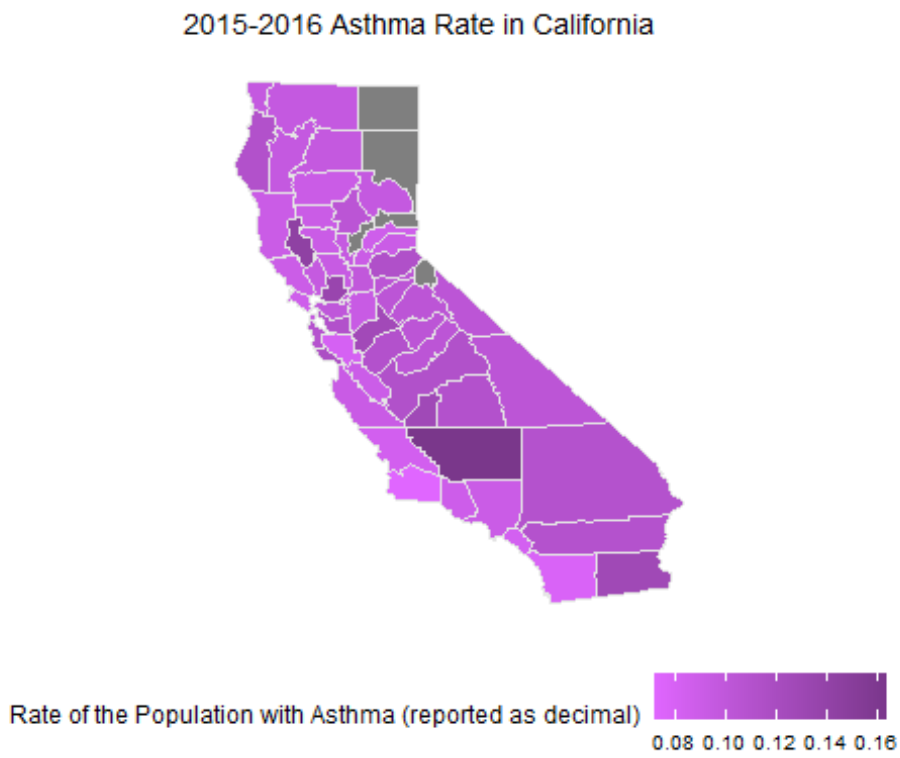
```
  theme_map() +labs(fill = "Days with unhealthy AQI" ) +
theme(legend.position = 'bottom')

pcal_aqi2015_2016_county.BADdays
```



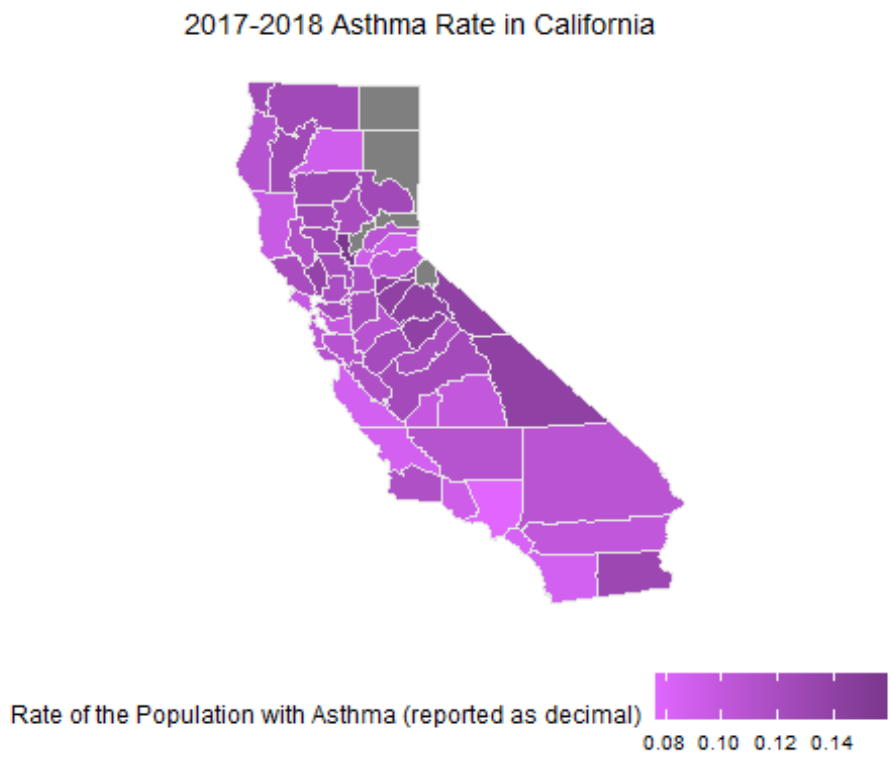

```
  #aqi2017_2018
p_aqi2017_2018_county.BADdays <- ggplot(cal_df, aes(x=lon, y=lat,
group=group, fill= aqi2017_2018_county.BADdays)) +
geom_polygon(color='gray90', size=0.1) + coord_map(projection = 'albers',
lat0=39, lat1=45)

pcal_aqi2017_2018_county.BADdays <- p_aqi2017_2018_county.BADdays +labs(title
= '2017-2018 Days with unhealthy AQI in California') +
  scale_fill_gradient(low = "coral", high = "coral4") +
  theme_map() +labs(fill = "Days with unhealthy AQI" ) +
theme(legend.position = 'bottom')

pcal_aqi2017_2018_county.BADdays
```




```
  #ASTHMA rates 2015-2016
p_MEAN_15_16 <- ggplot(cal_df, aes(x=lon, y=lat, group=group, fill=
MEAN_15_16)) + geom_polygon(color='gray90', size=0.1) + coord_map(projection
= 'albers', lat0=39, lat1=45)

pcal_MEAN_15_16 <- p_MEAN_15_16 +labs(title = '2015-2016 Asthma Rate in
California') +
  theme_map() +labs(fill = "Rate of the Population with Asthma (reported as
decimal)" ) + scale_fill_gradient(low = "mediumorchid1", high =
"mediumorchid4") + theme(legend.position = 'bottom')

pcal_MEAN_15_16
```




```
#ASTHMA rates 2017-2018
p_MEAN_17_18 <- ggplot(cal_df, aes(x=lon, y=lat, group=group, fill=
MEAN_17_18)) + geom_polygon(color='gray90', size=0.1) + coord_map(projection
= 'albers', lat0=39, lat1=45)

pcal_MEAN_17_18 <- p_MEAN_17_18 +labs(title = '2017-2018 Asthma Rate in
California') +
  theme_map() +labs(fill = "Rate of the Population with Asthma (reported as
decimal)" ) + scale_fill_gradient(low = "mediumorchid1", high =
"mediumorchid4") + theme(legend.position = 'bottom')

pcal_MEAN_17_18
```

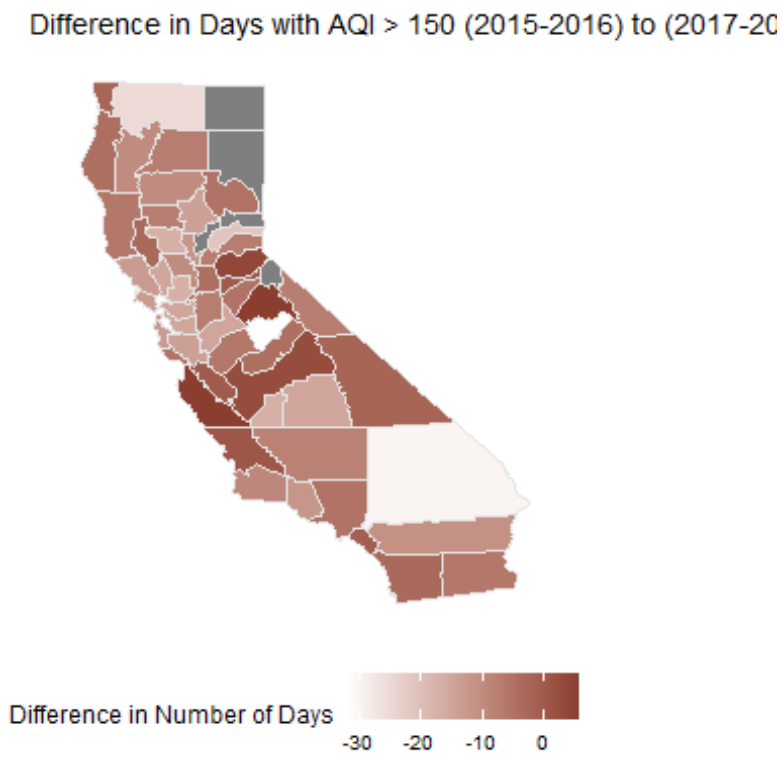



```
  #diff in AQI between years
cal_df$diff_aqi <- cal_df$aqi2015_2016_county.BADdays -
cal_df$aqi2017_2018_county.BADdays
cal_df$diff_asthma <- cal_df$MEAN_15_16 - cal_df$MEAN_17_18

head(cal_df)

##         lon      lat group  County aqi2015_2016_county.BADdays
## 1 -121.4785 37.48290     1 Alameda                           2
## 2 -121.5129 37.48290     1 Alameda                           2
## 3 -121.8853 37.48290     1 Alameda                           2
## 4 -121.8968 37.46571     1 Alameda                           2
## 5 -121.9254 37.45998     1 Alameda                           2
## 6 -121.9483 37.47717     1 Alameda                           2
##   aqi2015_2016_county.days proportion.bad.2015_2016
aqi2017_2018_county.BADdays
## 1                      731               0.2735978
17
## 2                      731               0.2735978
17
## 3                      731               0.2735978
17
## 4                      731               0.2735978
17
## 5                      731               0.2735978
17
## 6                      731               0.2735978
```

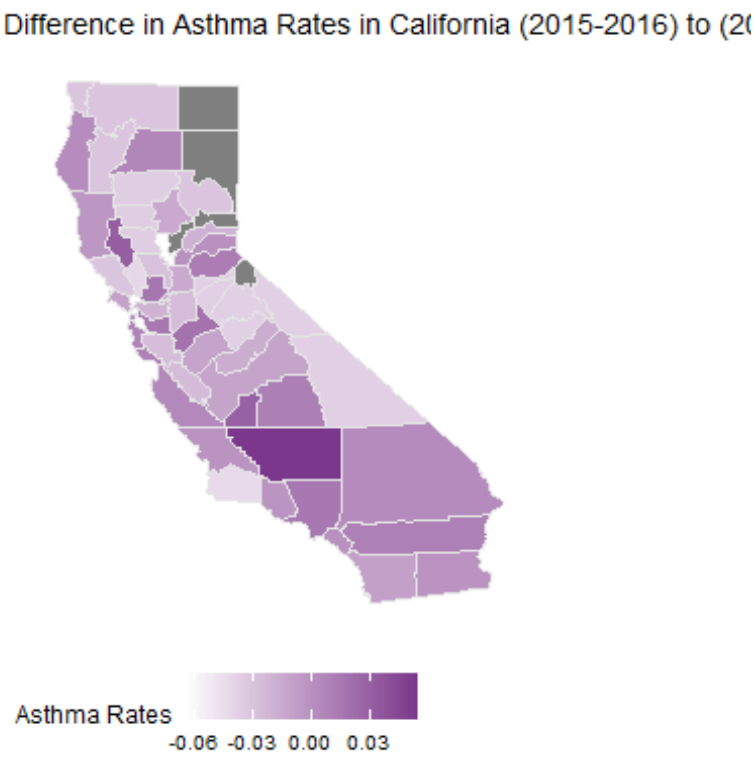
```
17
##   aqi2017_2018_county.days proportion.bad.2017_2018 MEAN_15_16 MEAN_17_18
## 1                      730                 2.328767  0.1118094 0.09732992
## 2                      730                 2.328767  0.1118094 0.09732992
## 3                      730                 2.328767  0.1118094 0.09732992
## 4                      730                 2.328767  0.1118094 0.09732992
## 5                      730                 2.328767  0.1118094 0.09732992
## 6                      730                 2.328767  0.1118094 0.09732992
##   diff_aqi diff_asthma
## 1      -15  0.01447951
## 2      -15  0.01447951
## 3      -15  0.01447951
## 4      -15  0.01447951
## 5      -15  0.01447951
## 6      -15  0.01447951
```

Plotting diffrences

```
p_diffAQI <- ggplot(cal_df, aes(x=lon, y=lat, group=group, fill= diff_aqi)) +
geom_polygon(color='gray90', size=0.1) + coord_map(projection = 'albers',
lat0=39, lat1=45)

pcal_diffAQI <- p_diffAQI +labs(title = 'Difference in Days with AQI > 150
(2015-2016) to (2017-2018)') +
  theme_map() +
  scale_fill_gradient(low = "white", high = "coral4") +
  labs(fill = "Difference in Number of Days" ) + theme(legend.position =
'bottom')

pcal_diffAQI
```




```
  #diff in asthma
p_diff_asthma <- ggplot(cal_df, aes(x=lon, y=lat, group=group, fill=
diff_asthma)) + geom_polygon(color='gray90', size=0.1) + coord_map(projection
= 'albers', lat0=39, lat1=45)

pcal_diff_asthma <- p_diff_asthma +labs(title = 'Difference in Asthma Rates
in California (2015-2016) to (2017-2018)') +
  theme_map() + scale_fill_gradient(low = "white", high = "mediumorchid4") +
  labs(fill = "Asthma Rates" ) + theme(legend.position = 'bottom')

pcal_diff_asthma
```



Trying different way to clean data and running regressions

```
asthma_c <- asthma %>% drop_na() %>% group_by(County, YEARS) %>%
  summarise(asthma = mean(asthma.PREVALENCE))

## `summarise()` regrouping output by 'County' (override with `.groups`
argument)

head(asthma_c)

## # A tibble: 6 x 3
## # Groups:   County [3]
##   County  YEARS     asthma
##   <chr>   <chr>      <dbl>
## 1 Alameda 2015-2016 0.116
## 2 Alameda 2017-2018 0.0903
## 3 Alpine  2015-2016 0.104
## 4 Alpine  2017-2018 0.160
## 5 Amador  2015-2016 0.104
## 6 Amador  2017-2018 0.160

aqi_c <- aqi2015_2018_bad[, -c(3,4,6,7)]
aqi_c <- aqi_c %>% drop_na() %>% group_by(County) %>%
  summarise('2015-2016' = sum(aqi2015_2016_county.BADdays),
            '2017-2018' = sum(aqi2017_2018_county.BADdays))

## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
aqi_c <- aqi_c %>%
  pivot_longer(!County, names_to = 'YEARS', values_to = 'BAD.days')

dim(aqi_c)
```

```
## [1] 106   3
```

```r
airdrop <- inner_join(asthma_c, aqi_c, by = c('County', 'YEARS'))
head(airdrop)
```

```
## # A tibble: 6 x 4
## # Groups:   County [3]
##   County  YEARS     asthma BAD.days
##   <chr>   <chr>      <dbl>    <int>
## 1 Alameda 2015-2016 0.116        2
## 2 Alameda 2017-2018 0.0903      17
## 3 Amador  2015-2016 0.104        0
## 4 Amador  2017-2018 0.160        0
## 5 Butte   2015-2016 0.105        1
## 6 Butte   2017-2018 0.125       15
```

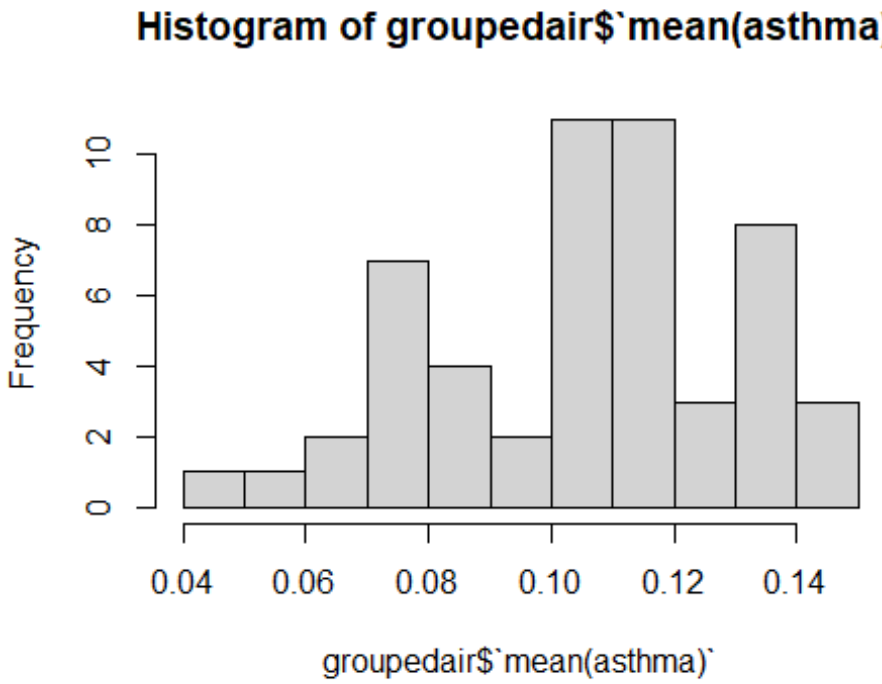
```r
groupedair <- airdrop %>% group_by(County) %>%
  summarise(mean(asthma),
            sum(BAD.days))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

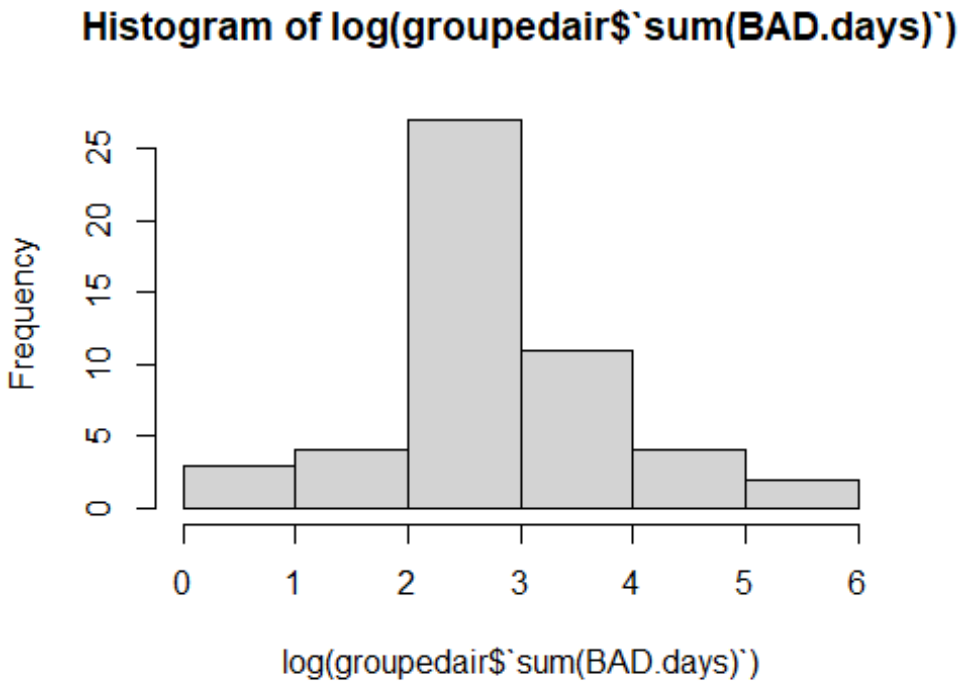
```r
head(groupedair)
```

```
## # A tibble: 6 x 3
##   County       `mean(asthma)` `sum(BAD.days)`
##   <chr>                 <dbl>           <int>
## 1 Alameda               0.103              19
## 2 Amador                0.132               0
## 3 Butte                 0.115              16
## 4 Calaveras             0.132              11
## 5 Colusa                0.109              21
## 6 Contra Costa          0.112              15
```

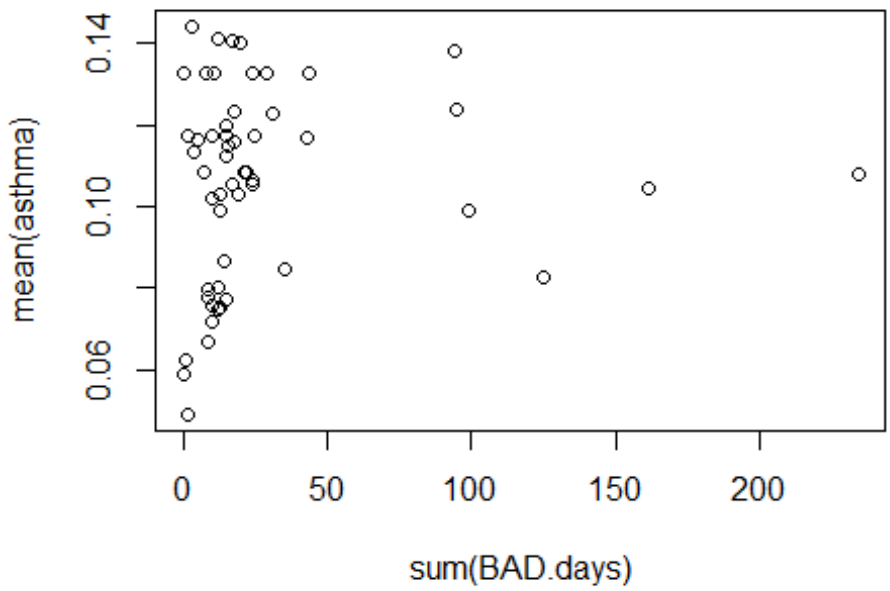
```r
#CORRELATION on grouped aqi and asthma data
hist(groupedair$`mean(asthma)`)
```




```
hist(log(groupedair$`sum(BAD.days)`)) #log normalizes it
```




```
plot(`mean(asthma)`~`sum(BAD.days)`, data = groupedair)
```




```
cor.test(groupedair$`mean(asthma)`, groupedair$`sum(BAD.days)`) #cor
0.08661648

##
##  Pearson's product-moment correlation
##
## data:  groupedair$`mean(asthma)` and groupedair$`sum(BAD.days)`
## t = 0.6209, df = 51, p-value = 0.5374
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1880807  0.3487455
## sample estimates:
##        cor
## 0.08661648

cor.test(groupedair$`mean(asthma)`, log(groupedair$`sum(BAD.days)`), method =
"spearman", exact = FALSE) #rho 0.2384135

##
##  Spearman's rank correlation rho
##
## data:  groupedair$`mean(asthma)` and log(groupedair$`sum(BAD.days)`)
## S = 18890, p-value = 0.08558
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2384135
```

```r
groupedmod1<-lm(`mean(asthma)`~`sum(BAD.days)`, data=groupedair)
summary(groupedmod1) #p-value: 0.5374
```

```
##
## Call:
## lm(formula = `mean(asthma)` ~ `sum(BAD.days)`, data = groupedair)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055464 -0.021571  0.003111  0.014803  0.039440
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.045e-01  3.978e-03  26.264   <2e-16 ***
## `sum(BAD.days)` 4.818e-05  7.760e-05   0.621    0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02405 on 51 degrees of freedom
## Multiple R-squared:  0.007502,   Adjusted R-squared:  -0.01196
## F-statistic: 0.3855 on 1 and 51 DF,  p-value: 0.5374
```

```r
#REGRESSTION on data with dropped values
head(airdrop)
```

```
## # A tibble: 6 x 4
## # Groups:   County [3]
##   County  YEARS     asthma BAD.days
##   <chr>   <chr>      <dbl>    <int>
## 1 Alameda 2015-2016 0.116        2
## 2 Alameda 2017-2018 0.0903      17
## 3 Amador  2015-2016 0.104        0
## 4 Amador  2017-2018 0.160        0
## 5 Butte   2015-2016 0.105        1
## 6 Butte   2017-2018 0.125       15
```

```r
  #create dummy gor years
airdrop$YEARS <-  factor(airdrop$YEARS)
head(airdrop)
```

```
## # A tibble: 6 x 4
## # Groups:   County [3]
##   County  YEARS     asthma BAD.days
##   <chr>   <fct>      <dbl>    <int>
## 1 Alameda 2015-2016 0.116        2
## 2 Alameda 2017-2018 0.0903      17
## 3 Amador  2015-2016 0.104        0
## 4 Amador  2017-2018 0.160        0
## 5 Butte   2015-2016 0.105        1
## 6 Butte   2017-2018 0.125       15
```

```
yearmod1<-lm(asthma~BAD.days*YEARS, data=airdrop)
summary(yearmod1) #p-value YEARS: 0.00483  ## p val for BAD.days:YEARS
0.08716

##
## Call:
## lm(formula = asthma ~ BAD.days * YEARS, data = airdrop)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.058050 -0.021409 -0.004666  0.019760  0.078983
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.0963543  0.0046514  20.715  < 2e-16 ***
## BAD.days                 0.0003288  0.0002065   1.593  0.11436
## YEARS2017-2018           0.0208651  0.0072378   2.883  0.00483 **
## BAD.days:YEARS2017-2018 -0.0004775  0.0002764  -1.728  0.08716 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03039 on 100 degrees of freedom
## Multiple R-squared:  0.085,  Adjusted R-squared:  0.05755
## F-statistic: 3.097 on 3 and 100 DF,  p-value: 0.03028
```

Regressions

```
modI <- lm(asthma~BAD.days, data=airdrop)
summary(modI)

##
## Call:
## lm(formula = asthma ~ BAD.days, data = airdrop)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.056109 -0.023862 -0.004035  0.020988  0.086175
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1049528  0.0036718  28.583   <2e-16 ***
## BAD.days    0.0001298  0.0001383   0.939     0.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03132 on 102 degrees of freedom
## Multiple R-squared:  0.008564,   Adjusted R-squared:  -0.001156
## F-statistic: 0.881 on 1 and 102 DF,  p-value: 0.3501

modII <- lm(asthma~BAD.days+YEARS, data=airdrop)
summary(modII)
```

```
##
## Call:
## lm(formula = asthma ~ BAD.days + YEARS, data = airdrop)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055021 -0.023187 -0.001581  0.018000  0.078801
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.900e-02  4.434e-03  22.327   <2e-16 ***
## BAD.days       6.242e-05  1.386e-04   0.450   0.6534
## YEARS2017-2018 1.413e-02  6.158e-03   2.295   0.0238 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03068 on 101 degrees of freedom
## Multiple R-squared:  0.05769,    Adjusted R-squared:  0.03903
## F-statistic: 3.092 on 2 and 101 DF,  p-value: 0.04974
```

THE END