Regular Expressions and Webscraping

Grant Esparza

October 5, 2018

Regular Expressions

The goal here is to list all `csv` files located under `R.home()`. Using the function `list.files()`, we can use a regular expression to specify the type of files we want.

```
## List csv files in R.home()
list.files(path=R.home(), pattern=".*csv", recursive=TRUE, include.dirs=TRUE)
```

```
## [1] "library/utils/misc/exDIF.csv"
```

Interestingly, I only seem to have one `csv` file in `R.home()`. I'm not sure why other files that belong in packages don't show up. However, searching through other folders yield the correct results:

```
list.files(path="~/repos/MATH385/", pattern=".*csv", recursive=TRUE, include.dirs=TRUE)
```

```
## [1] "homework/simple_analysis/lobbyist-data-compensation.csv"
## [2] "homework/simple_analysis/lobbyist-data-contributions.csv"
## [3] "homework/simple_analysis/lobbyist-data-gifts.csv"
## [4] "labs/daily_aqi_by_county_2017.csv"
```

Webscraping

Sometimes the only way to obtain data is to tell `R` to go get it from a webpage. While we could simply click and download this EPA data, it'll be good practice to use the package `httr`.

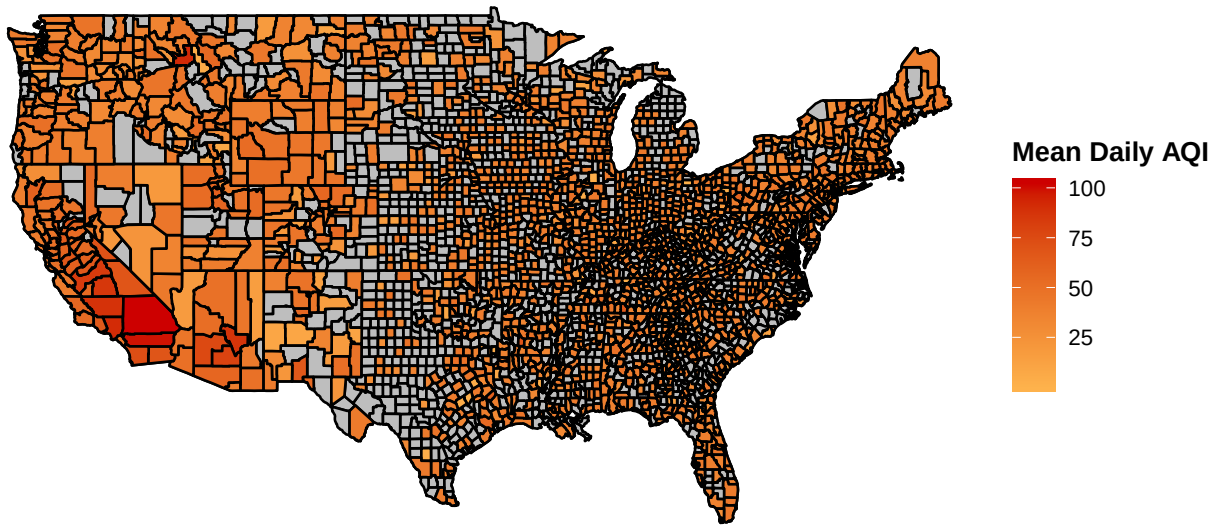
```
## Using httr
get_request <- GET("https://aqs.epa.gov/aqsweb/airdata/daily_aqi_by_county_2017.zip")
bin_data <- content(get_request, "raw")
writeBin(bin_data, "daily-county-aqi")
unzip(zipfile="daily-county-aqi")

## Read in data
aqi.df <- read.csv("daily_aqi_by_county_2017.csv")
```

After using the function `GET`, we can retrieve the raw data and write it to a folder in our current directory. After using `unzip` we can the read the extracted zip as normal.

Analysis

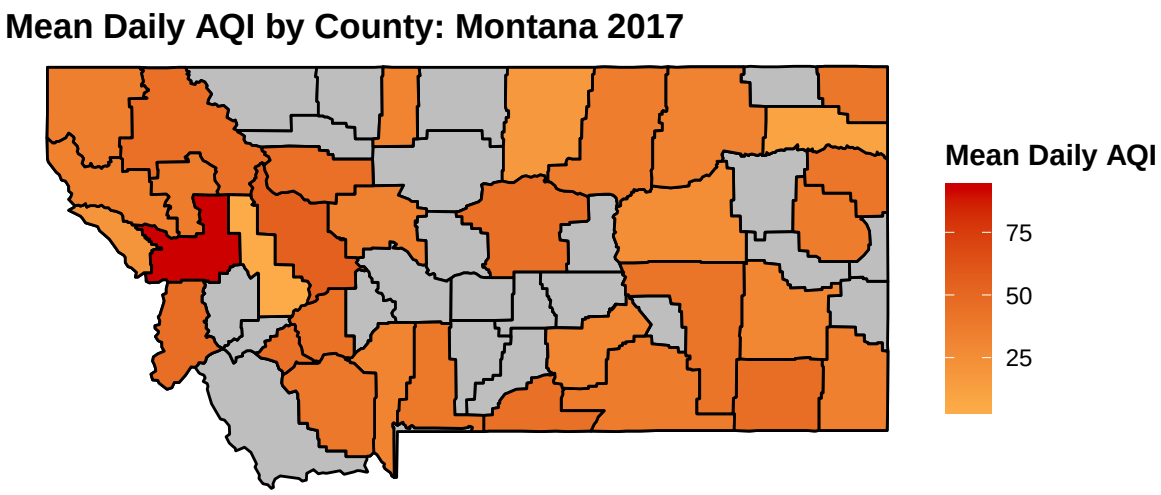
This data contains records of the daily Air Quality Index throughout the US. When dealing with geographic data, I find it easiest to interpret a map.

```
## Create map of US
counties <- map_data("county")

## Generate mean aqi per county
aqi.county <- aqi.df %>%
  group_by(subregion) %>%
  summarise(mn_aqi=mean(aqi), md_aqi=median(aqi))

## Join stat data
aqi.map <- inner_join(counties, aqi.county, by="subregion")

## Plot US map with data
ggplot(data=counties, aes(x=long, y=lat, group=group)) +
    coord_fixed(1.3) +
    geom_polygon(color="black", fill="gray") +
    geom_polygon(data=aqi.map, aes(fill=mn_aqi), color="black") +
    scale_fill_gradient2(low="#FFFFE0", mid="#FEB24C", high="#CD0000") +
    ggtitle("Mean Daily AQI by County: 2017") +
    labs(fill="Mean Daily AQI") +
    theme_void() +
    theme(title = element_text(face="bold"))
```



Missoula County, MT

```
## Look at Montana
montana.dat <- aqi.map %>% filter(region == "montana")

ggplot(data=counties[counties$region == "montana",], aes(x=long, y=lat, group=group)) +
    coord_fixed(1.3) +
    geom_polygon(color="black", fill="gray") +
    geom_polygon(data=montana.dat, aes(fill=mn_aqi), color="black") +
    scale_fill_gradient2(low="#FFFFE0", mid="#FEB24C", high="#CD0000") +
    ggtitle("Mean Daily AQI by County: Montana 2017") +
    labs(fill="Mean Daily AQI") +
    theme_void() +
    theme(title = element_text(face="bold"))
```



An interesting piece of the previous plot was the county in Montana with an abnormally high AQI. Upon closer inspection we can find the name of the county:

```
## Order by mean AQI
tail(montana.dat[order(montana.dat$mn_aqi),], 1)
```

```
##           long      lat group order  region subregion   mn_aqi md_aqi
## 1106 -113.6061 47.58987  1596 47980 montana  missoula 92.16164     61
```

For a state known for its low population and a supposedly lower AQI value, its surprising to find such a high value here. A quick google search reveals a possible reason. It turns out in 2017 Montana experienced **twenty-one** wildfires that consumed over **438,000** acres. Missoula county is the second largest in the state hosting many parks, trails, and forest that are susceptible to fires. Therefore my hypothesis is that Missoula has such a high mean AQI due to an abnormal year of fires. My next step would be to find wildfire data in Monatana to see if I could determine if there is a correlation.