Regular Expressions and Webscraping Part 2

Grant Esparza

October 8, 2018

Incomplete Links

The goal here is to find AQI web links from the EPA website and complete them with the base url.

```
## Grab html
get_request <- GET("https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI")
html <- content(get_request, "text")

## No encoding supplied: defaulting to UTF-8.

## Find links
incomplete_links <- as.data.frame(str_match_all(html, pattern="<a href=\"daily.*county_[0-9]{4}\\.zip\"

## Split
incomp_links <- as.data.frame(str_split_fixed(incomplete_links[,1], pattern="\"", 3))
colnames(incomp_links) <- c("href", "link", "end")

## Add home url
complete_links <- paste("https://aqs.epa.gov/aqsweb/airdata/", incomp_links$link, sep="")
head(complete_links, 3)

## [1] "https://aqs.epa.gov/aqsweb/airdata/daily_aqi_by_county_2018.zip"
## [2] "https://aqs.epa.gov/aqsweb/airdata/daily_aqi_by_county_2017.zip"
## [3] "https://aqs.epa.gov/aqsweb/airdata/daily_aqi_by_county_2016.zip"
```

Using `str_match_all()` with an appropriate regular expression allows us to find the links and concatenate them with the correct base url.

Adding to Directories

Suppose we had actually downloaded the AQI data and we wanted to store it. Using the function `file.path()` we can construct filepaths that point to the location we want to store our data in.

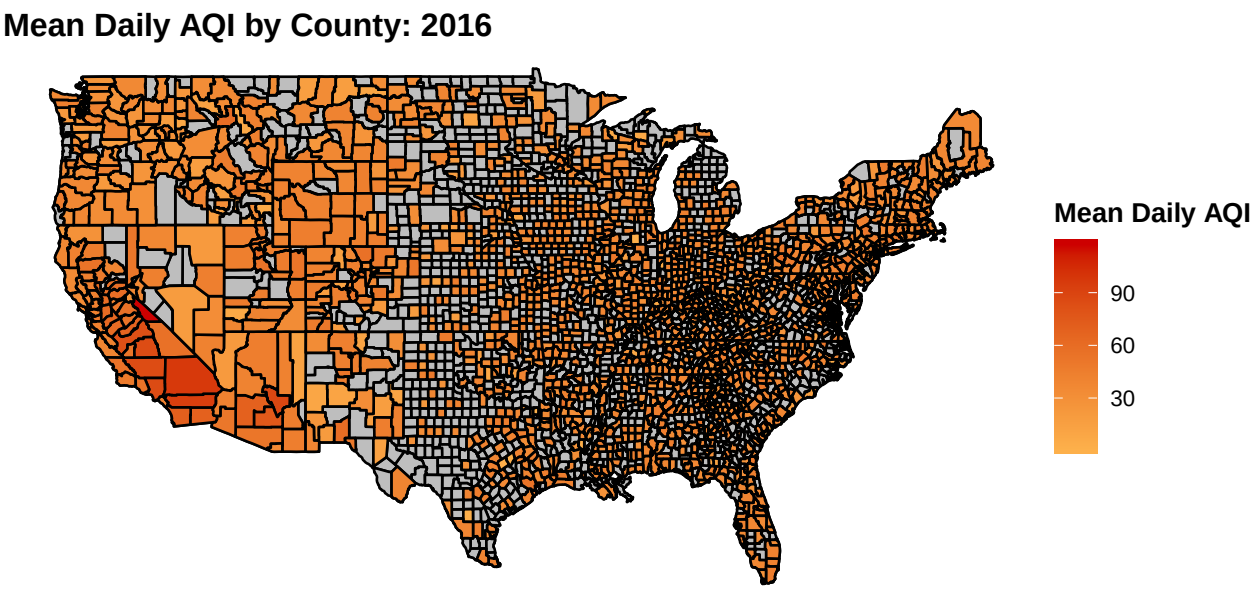
```
## Add urls to directory
folder <- '~/home/gesparza3/MATH385/labs/regex02/aqi_data'
r <- character()
for (i in incomp_links$link) {
  r <- c(r, file.path(folder, i))
}
head(r, 3)
```

```
## [1] "~/home/gesparza3/MATH385/labs/regex02/aqi_data/daily_aqi_by_county_2018.zip"
## [2] "~/home/gesparza3/MATH385/labs/regex02/aqi_data/daily_aqi_by_county_2017.zip"
## [3] "~/home/gesparza3/MATH385/labs/regex02/aqi_data/daily_aqi_by_county_2016.zip"
```

Working With Downloads

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.

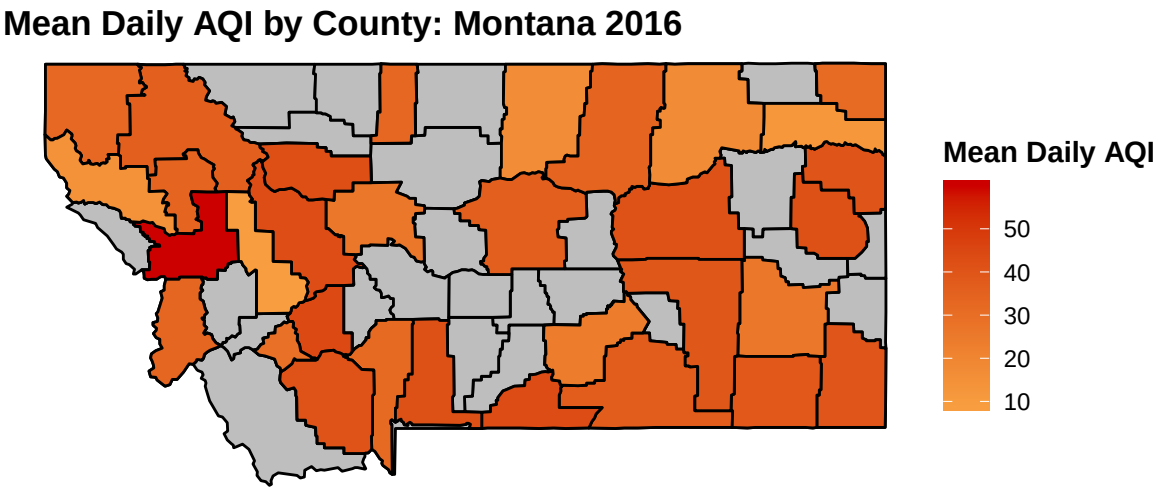
```
## Get data
get_request <- GET("https://aqs.epa.gov/aqsweb/airdata/daily_aqi_by_county_2016.zip")
bin_data <- content(get_request, "raw")
writeBin(bin_data, "daily-county-aqi")
unzip(zipfile="daily-county-aqi")

## Read in data
aqi.df <- read.csv("daily_aqi_by_county_2016.csv")

## Fix aqi names
colnames(aqi.df) <- c("region", "subregion", "state.code", "county.code", "date",
                      "aqi", "category", "defining.parameter", "defining.site",
                      "number.of.sites.reporting")
aqi.df$region <- tolower(aqi.df$region)
aqi.df$subregion <- tolower(aqi.df$subregion)

## 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: 2016") +
    labs(fill="Mean Daily AQI") +
    theme_void() +
    theme(title = element_text(face="bold"))
```



Missoula County, MT does not have the same AQI as in the 2017 analysis. This seems to support my idea that 2017's high AQI was due to an abnormal year of wildfires. However, it will still be interesting to see how Missoula compares to other counties in Montana.

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 2016") +
    labs(fill="Mean Daily AQI") +
    theme_void() +
    theme(title = element_text(face="bold"))
```



When plotting Montana you can see that Missoula still has a higher AQI than the surrounding counties. More data would be needed to determine if this is still due to fires or if there is another variable that increases the AQI.