MP2

Gwynnie and Aria

Motivations Behind our Data

After doing our amazing research on salmon we wanted to look at more data from Washington, and one thing that is becoming very prevalent is wildfires. Both of us have been in very close proximity to wildfires and have seen the aftermath of many of them. We want to look at the amount of forest land that was lost or affected by forest fires. Thinking through this data we quickly found a data set of the land area affected by fires, we then realized that it would be interesting to see the amount of forested area that was affected by the fires and then also the total land area that was affected in the county as a whole.

Research Questions

We are interested in looking at the total land area that is affected by forest fires in our home state. We wanted to not only look at the forest land that was affected by also the amount of land area that is forested and then also affected by the fire.

```
# Our first table came from wikipedia, which is an allowed source
is_valid_robotstxt("https://en.wikipedia.org/wiki/List_of_Washington_wildfires")
```

```
[1] TRUE
```

```
#reading the html of the website
wildfires <- read_html("https://en.wikipedia.org/wiki/List_of_Washington_wildfires")

#scraping the table
wildfiretables <- html_nodes(wildfires, css = "table")

#our first raw set of tables
html_table(wildfiretables, header = TRUE, fill = TRUE)
```

```
[[1]]
# A tibble: 0 x 2
# i 2 variables:  <lgl>,
#   This list is incomplete; you can help by adding missing items.  (August 2015) <lgl>

[[2]]
# A tibble: 11 x 11
    Year `Fire name`       `Complex name` County `Start dateCause` `Size(acres)`
   <int> <chr>             <chr>          <chr>  <chr>             <chr>
 1  2024 Beam Road Fire[2] ""             Yakima "June 15"         8,542 acres ~
 2  2024 Big Horn Fire[3]~ ""             Klick~ "July 22, unknow~ 51,569 acres~
 3  2024 Black Canyon Fir~ ""             Yakima "July 22, unknow~ 9,211 acres ~
 4  2024 Cougar Creek Fir~ ""             Asoti~ "July 15, unknow~ 20,699 acres~
 5  2024 Pioneer Fire[8]   ""             Chelan "June 8, human c~ 36,763 acres~
 6  2024 Retreat Fire[9][~ ""             Yakima "July 23, cause ~ 44,588 acres~
 7  2024 Swawilla Fire[11~ ""             Ferry~ "July 17, Lightn~ 53,462 acres~
 8  2023 Oregon Fire[13]   ""             Spoka~ ""                10,817 acres~
 9  2023 Gray Fire[15]     ""             Spoka~ ""                10,085[15][1~
10  2020 Cold Springs Can~ "Labor Day fi~ Okano~ ""                Over 410,000~
11  2020 Whitney Fire      ""             Linco~ "September 7"     127,430
# i 5 more variables: Structureslost <chr>, Deaths <chr>, Injuries <int>,
#   Notes <chr>, Image <chr>

[[3]]
# A tibble: 66 x 11
    Year `Fire name`           `Complex name`  County `Start date` `Size(acres)`
   <int> <chr>                 <chr>           <chr>  <chr>        <chr>
 1  2019 243 Command Fire[18]  ""              Grant  "June 3"     20,380 acres~
 2  2019 Cold Creek Fire[19]   ""              Benton ""           42,000 acres~
 3  2019 Pipeline Fire         ""              Kitti~ ""           6,515 acres ~
 4  2019 Powerline Fire[20]    ""              Grant  ""           7,800 acres ~
 5  2019 Williams Flats Fire   ""              Okano~ ""           44,446 acres~
 6  2016 Hart Fire             ""              Linco~ ""           18,220
 7  2016 Range 12 Fire[21]     ""              Yakima ""           177,210
 8  2016 2016 Snake River Fire ""              Garfi~ ""           11,452 acres~
 9  2016 Spokane Complex Fire  "Spokane Compl~ Spoka~ ""           7,251 acres ~
10  2015 Black Canyon Fire[22] "Chelan Comple~ Chelan "August 14"  6,761
# i 56 more rows
# i 5 more variables: Structureslost <chr>, Deaths <int>, Injuries <int>,
#   Notes <chr>, Image <chr>

[[4]]
# A tibble: 55 x 11
```

```
    Year `Fire name`            `Complex name` County `Start date` `Size(acres)`
   <int> <chr>                  <chr>          <chr>  <chr>        <chr>
 1  2009 Dry Creek Complex[50]  "Dry Creek Co~ Bento~ ""           48,902
 2  2009 Oden Road Fire[50]     ""             Okano~ ""           9,607
 3  2008 Badger Mountain Fire[~ ""             Chela~ ""           15,023
 4  2008 Cold Springs Fire      ""             Klick~ ""           7,729
 5  2008 Columbia River Road F~ ""             Okano~ ""           22,115
 6  2008 Smith Lake Fire[64]    ""             Dougl~ ""           12,513
 7  2008 Spokane Valley Fire[6~ ""             Spoka~ ""           1,008
 8  2008 Swanson Lake Fire[50]  ""             Linco~ ""           19,090
 9  2007 Domke Lake Fire[50]    ""             Okano~ ""           11,900
10  2007 Easy Street Fire[50]   ""             Chelan ""           5,209
# i 45 more rows
# i 5 more variables: Structureslost <int>, Deaths <int>, Injuries <chr>,
#   Notes <chr>, Image <chr>

[[5]]
# A tibble: 28 x 11
    Year `Fire name`            `Complex name` County `Start date` `Size(acres)`
   <int> <chr>                  <chr>          <chr>  <chr>        <chr>
 1  1998 Cleveland Fire[84]     ""             Klick~ ""           18,500
 2  1998 Rattle Snake Ridge Fi~ ""             Yakima ""           18,000
 3  1997 Olympia Command Fire[~ ""             Benton ""           5,500
 4  1997 Pow Wah Kee Fire[1]    "August 3"     Asotin ""           8,000
 5  1996 Baird Springs Fire[1]  ""             Grant  "August 2"   14,000
 6  1996 Cold Creek Fire[50]    ""             Bento~ ""           57,000
 7  1994 Copper Butte Fire[96]  ""             Ferry  ""           10,473
 8  1994 Rat Creek / Hatchery ~ ""             Chelan ""           43,000
 9  1994 Tyee Creek Fire[98][9~ ""             Chelan ""           135,000
10  1992 Castlerock Fire[1]     ""             Wenat~ ""           3,500[100]
# i 18 more rows
# i 5 more variables: Structureslost <chr>, Deaths <chr>, Injuries <chr>,
#   Notes <chr>, Image <chr>

[[6]]
# A tibble: 39 x 10
    Year `Fire name`            `Complex name` County `Start date` `Size(acres)`
   <int> <chr>                  <chr>          <chr>  <chr>        <chr>
 1  2024 Bridge Creek Fire      ""             Ferry  "July 19"    3,998 acres ~
 2  2016 Buck Creek             ""             Chelan "July 22"    1,987 acres ~
 3  2015 231 Fire               ""             Steve~ ""           1,138
 4  2015 Twenty-One Mile Grade~ ""             Ferry  ""           2,250
 5  2014 Hansel Fire            ""             Chelan ""           1,016
```

```
 6  2014 Little Bridge Fire     ""             Okano~ "August 2"   4,896
 7  2014 Lone Mountain Fire     ""             Chelan "July 14"    2,770
 8  2012 Cashmere Fire          "Wenatchee Co~ Chelan ""           2,651
 9  2012 Highway 141 Fire[84]   ""             Klick~ ""           1,644
10  2011 Salmon Fire[50]        ""             Okano~ ""           1,631
# i 29 more rows
# i 4 more variables: Structureslost <int>, Injuries <int>, Notes <chr>,
#   Image <chr>

[[7]]
# A tibble: 0 x 2
# i 2 variables:  <lgl>,
#   This list is incomplete; you can help by adding missing items.  (September 2015) <lgl>

[[8]]
# A tibble: 24 x 10
   ``     Totalfires `Total area burned` `Total area burned` Structureslost
   <chr>  <chr>      <chr>               <chr>               <chr>
 1 ""     Totalfires Acres               Hectares            "Structureslost"
 2 "2002" 1,285      92,742              37,531              ""
 3 "2003" 1,373      200,517             81,146              ""
 4 "2004" 1,674      92,617              37,481              ""
 5 "2005" 998        185,748             75,170              ""
 6 "2006" 1,579      410,060             165,950             ""
 7 "2007" 1,268      214,925             86,977              ""
 8 "2008" 1,303      147,264             59,596              ""
 9 "2009" 1,976      77,250              31,260              ""
10 "2010" 870        56,820              22,990              ""
# i 14 more rows
# i 5 more variables: Fatalities <chr>, Injuries <chr>, Totalcost <chr>,
#   Notes <chr>, Source <chr>

[[9]]
# A tibble: 12 x 2
   .mw-parser-output .navbar{display:inline;font-size:8~1 .mw-parser-output .n~2
   <chr>                                                <chr>
 1 "Pre-2014"                                           "Yacolt Burn (1902)\n~
 2 "2014"                                               "Carlton Complex"
 3 "2015"                                               "Okanogan Complex"
 4 "2016"                                               "Range 12"
 5 "2017"                                               "Diamond Creek\nJack ~
 6 "2018"                                               "Soap Lake\nMaple Fir~
 7 "2019"                                               "243 Command Fire\nLe~
```

```
 8 "2020"                                                    "Evans Canyon\nLabor ~
 9 "2021"                                                    "Schneider Springs Fi~
10 "2023"                                                    "Eagle Bluff Fire\nGr~
11 "2024"                                                    "Pioneer Fire\nRetrea~
12 "Category\n Commons"                                      "Category\n Commons"
# i abbreviated names:
#   1: `.mw-parser-output .navbar{display:inline;font-size:88%;font-weight:normal}.mw-parser-
#   2: `.mw-parser-output .navbar{display:inline;font-size:88%;font-weight:normal}.mw-parser-

[[10]]
# A tibble: 3 x 2
  `vteWildfires in the United States` `vteWildfires in the United States`
  <chr>                               <chr>
1 "States"                            "Alabama\nAlaska\nArizona\nArkansas\nCali~
2 "Territories"                       "American Samoa\nGuam\nNorthern Mariana I~
3 "Category\n Commons"                "Category\n Commons"
```

```r
# Since we had so many tables from one scrape to use, we created a small
# function to choose the table from the list using its subset number, cleaned
# the names, remove unnecessary columns, and rename a common variables. Due to
# inconsistency, all variables were set set as character and then parsed for
# numbers.

cleaninggg <- function(table, i) {
  html_table(table, header = TRUE, fill = TRUE)[[i]]|>
    janitor::clean_names() |>
    select(-notes, -image, -injuries, -complex_name) |>
    mutate(across(c(structureslost, size_acres), as.character),
           across(c(structureslost, size_acres), parse_number)) |>
    rename("fire_size_acres" = "size_acres")
}

# Running the function for each of the times to
# pull the data out of the list from wikipedia into 5 (nearly) uniform datasets
twenty <- cleaninggg(wildfiretables, 2) |> rename("start_date" = "start_date_cause")
ten <- cleaninggg(wildfiretables, 3)
thousand <- cleaninggg(wildfiretables, 4)
nines <- cleaninggg(wildfiretables, 5)
minors <- cleaninggg(wildfiretables, 6)

# Binds all of the major fires into one dataset and removes deaths for
```

```
# consistency with the minor fires
majors <- rbind(twenty, ten, thousand, nines) |> select(-deaths)

# Adds a column that identifies is a fire was major or minor
minors['fire_type'] = "Minor"
majors['fire_type'] = "Major"

# Joins all fires together
fires <- rbind(majors, minors)

head(fires)
```

```
# A tibble: 6 x 7
   year fire_name     county start_date fire_size_acres structureslost fire_type
  <int> <chr>         <chr>  <chr>                <dbl>          <dbl> <chr>
1  2024 Beam Road Fi~ Yakima June 15                8542              0 Major
2  2024 Big Horn Fir~ Klick~ July 22, ~            51569              0 Major
3  2024 Black Canyon~ Yakima July 22, ~             9211              0 Major
4  2024 Cougar Creek~ Asoti~ July 15, ~            20699              4 Major
5  2024 Pioneer Fire~ Chelan June 8, h~            36763              0 Major
6  2024 Retreat Fire~ Yakima July 23, ~            44588              5 Major
```

```
# As most major fires burn throughout forests, we wanted to add in a dataset
# about forest coverage per county, we were planning to make a for-loop for
# this, but all of the websites we tried to scrape weren't reading the actual
# number as it was stored as an image? So we found this website that stores it
# all as a list
is_valid_robotstxt("https://data.workingforests.org/#")
```

```
[1] TRUE
```

```
session <- bow("https://data.workingforests.org/#")

# Scraped the county names as one list
county_title <- scrape(session) |>
  html_nodes(".countyName") |>
  html_text()
```

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

```
county_title
```

```
 [1] "Statewide"           "Adams County"        "Asotin County"
 [4] "Benton County"       "Chelan County"       "Clallam County"
 [7] "Clark County"        "Columbia County"     "Cowlitz County"
[10] "Douglas County"      "Ferry County"        "Franklin County"
[13] "Garfield County"     "Grant County"        "Grays Harbor County"
[16] "Island County"       "Jefferson County"    "King County"
[19] "Kitsap County"       "Kittitas County"     "Klickitat County"
[22] "Lewis County"        "Lincoln County"      "Mason County"
[25] "Okanogan County"     "Pacific County"      "Pend Oreille County"
[28] "Pierce County"       "San Juan County"     "Skagit County"
[31] "Skamania County"     "Snohomish County"    "Spokane County"
[34] "Stevens County"      "Thurston County"     "Wahkiakum County"
[37] "Walla Walla County"  "Whatcom County"      "Whitman County"
[40] "Yakima County"
```

```
# Scraped the forest coverage as another list
forest_cov <- scrape(session) |>
  html_nodes(".dataValueEmphasized") |>
  html_text()
forest_cov
```

```
 [1] "22,983,438" "1,452"      "103,022"    "351"        "1,392,891"
 [6] "1,034,606"  "251,273"    "203,917"    "657,909"    "16,983"
[11] "1,072,722"  "1,733"      "100,933"    "6,706"      "1,120,182"
[16] "86,883"     "1,064,350"  "1,003,402"  "187,620"    "783,309"
[21] "516,397"    "1,374,647"  "69,114"     "552,926"    "1,982,401"
[26] "534,690"    "787,506"    "800,881"    "85,258"     "890,416"
[31] "996,021"    "1,065,150"  "318,506"    "1,149,289"  "329,638"
[36] "147,694"    "30,934"     "1,033,817"  "26,889"     "1,201,021"
```

```
# Brought the 2 lists together as one tibble with 2 columns, removed " County"
# from name to synchronize with main table
forest_cover <- tibble(county = county_title,
                       forest_coverage_acres = forest_cov) |>
  mutate(county = str_remove(county, " County"),
         forest_coverage_acres = parse_number(forest_coverage_acres))
```

```
# Joins this forest coverage with our fire data by county. For ease of analysis
# at this stage without knowing string analysis in detail (yet!), we removed all
# rows that contained 2 counties by dropping NA's in forest coverage. This way
# all rows should have a complete collection of county name, forest size, and
# fire size.
fullfires <- fires |> left_join(forest_cover) |>
  drop_na(forest_coverage_acres)
```

```
Joining with `by = join_by(county)`
```

```
head(fullfires)
```

```
# A tibble: 6 x 8
   year fire_name     county start_date fire_size_acres structureslost fire_type
  <int> <chr>         <chr>  <chr>                <dbl>          <dbl> <chr>
1  2024 Beam Road Fi~ Yakima "June 15"             8542              0 Major
2  2024 Big Horn Fir~ Klick~ "July 22,~           51569              0 Major
3  2024 Black Canyon~ Yakima "July 22,~            9211              0 Major
4  2024 Pioneer Fire~ Chelan "June 8, ~           36763              0 Major
5  2024 Retreat Fire~ Yakima "July 23,~           44588              5 Major
6  2023 Gray Fire[15] Spoka~ ""                   10085            259 Major
# i 1 more variable: forest_coverage_acres <dbl>
```

```
# Lastly, we also thought it would be good to include the size of the counties
# themselves as a comparison to the size of the forest its fires, so we scraped
# this table
counties <- read_html("https://en.wikipedia.org/wiki/List_of_counties_in_Washington")
countytable <- html_nodes(counties, css = "table")
countytable
```

```
{xml_nodeset (8)}
[1] <table class="infobox vevent" style="float: right; width: ;"><tbody>\n<tr ...
[2] <table class="wikitable sortable sticky-header" style="text-align: center ...
[3] <table class="nowraplinks mw-collapsible mw-collapsed navbox-inner" style ...
[4] <table class="nowraplinks mw-collapsible autocollapse navbox-inner" style ...
[5] <table class="nowraplinks hlist mw-collapsible autocollapse navbox-inner" ...
[6] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
[7] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
[8] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
```

```r
# This identifies the table we want, cleans the names, removes part of the name
# ' County' for consistency, parses the sq. mi. and converts it to acres, and
# selects just county and county size

countysize <- html_table(countytable, header = TRUE, fill = TRUE)[[2]] |>
  janitor::clean_names() |>
  mutate(county = str_remove(county, " County"),
         county_size_acres = parse_number(land_area_11) * 640) |>
    select(county, county_size_acres)

# Finally ! We join this last table with the main dataset
final_fires <- fullfires |> left_join(countysize)
```

```
Joining with `by = join_by(county)`
```

```r
final_fires
```

```
# A tibble: 170 x 9
    year fire_name    county start_date fire_size_acres structureslost fire_type
   <int> <chr>        <chr>  <chr>                <dbl>          <dbl> <chr>
 1  2024 Beam Road F~ Yakima "June 15"             8542              0 Major
 2  2024 Big Horn Fi~ Klick~ "July 22,~           51569              0 Major
 3  2024 Black Canyo~ Yakima "July 22,~            9211              0 Major
 4  2024 Pioneer Fir~ Chelan "June 8, ~           36763              0 Major
 5  2024 Retreat Fir~ Yakima "July 23,~           44588              5 Major
 6  2023 Gray Fire[1~ Spoka~ ""                   10085            259 Major
 7  2020 Whitney Fire Linco~ "Septembe~          127430             NA Major
 8  2019 243 Command~ Grant  "June 3"             20380              0 Major
 9  2019 Cold Creek ~ Benton ""                   42000             NA Major
10  2019 Pipeline Fi~ Kitti~ ""                    6515             NA Major
# i 160 more rows
# i 2 more variables: forest_coverage_acres <dbl>, county_size_acres <dbl>
```

```r
head(final_fires)
```

```
# A tibble: 6 x 9
   year fire_name     county start_date fire_size_acres structureslost fire_type
  <int> <chr>         <chr>  <chr>                <dbl>          <dbl> <chr>
1  2024 Beam Road Fi~ Yakima "June 15"             8542              0 Major
```

```
2  2024 Big Horn Fir~ Klick~ "July 22,~          51569              0 Major
3  2024 Black Canyon~ Yakima "July 22,~           9211              0 Major
4  2024 Pioneer Fire~ Chelan "June 8, ~          36763              0 Major
5  2024 Retreat Fire~ Yakima "July 23,~          44588              5 Major
6  2023 Gray Fire[15] Spoka~ ""                  10085            259 Major
# i 2 more variables: forest_coverage_acres <dbl>, county_size_acres <dbl>
```

Future Uses of this Data

For future uses of this data we have a lot of things that we want to clean with string functions. We were also looking into census data for each county in Washington, which would be interesting to see if there is higher population in a county that has more forest fire activity. It would also be interesting to add spatial data to this to map the percentage of forest area affected by fires or other percentage maps.