

ERHS 535 Final Project: Prediction of Human Cases of West Niles Virus in Southern California by Precipitation and Temperature (2006-2010)

Devin, Anastasia, Grant, Ana

11/7/2018

1. Packages used in the final project

```
library(dplyr)
library(stringr)
library(tidyselect)
library(ggplot2)
library(rmarkdown)
library(RColorBrewer)
library(knitr)
library(forcats)
library(readr)
library(tidyr)
library(broom)
library(purrr)
library(scales)
library(lubridate)
library(viridis)
library(RColorBrewer)
library(scales)
library(tidyverse)
library(tigris)
library(sf)
library(cowplot)
library(gridExtra)
```

2. Datasets description

Different data sets were collected from CHHS and CDC Wonder. The report of human cases for California, had data arrange by week and county for the years 2006 to 2018. The data is available in:

[CHHS] <https://data.chhs.ca.gov/dataset/west-nile-virus-cases-2006-present/resource/6ef33c1b-9f54-49f2-a92e-51a1b78f0a06>

Daily precipitation and temperature in California were downloaded from CDC Wonder website. Every year was pulled independently. The request form might be found in the following website.

[CDC Wonder]<https://wonder.cdc.gov/>

```
report <- read_csv("../Data/wnv_human_cases.csv")
temp_2006 <- read_csv("../Data/temp/temp_2006.csv")
temp_2007 <- read_csv("../Data/temp/temp_2007.csv")
temp_2008 <- read_csv("../Data/temp/temp_2008.csv")
temp_2009 <- read_csv("../Data/temp/temp_2009.csv")
temp_2010 <- read_csv("../Data/temp/temp_2010.csv")
```

```
precip_2006 <- read_csv("../Data/precip/precip_2006.csv")
precip_2007 <- read_csv("../Data/precip/precip_2007.csv")
precip_2008 <- read_csv("../Data/precip/precip_2008.csv")
precip_2009 <- read_csv("../Data/precip/precip_2009.csv")
precip_2010 <- read_csv("../Data/precip/precip_2010.csv")
```

3. Tyding and organizing the original data sets

3.1. Human cases dataset

The main objective was change the *weekly* report to a *year* report for the number of cases by county

```
cases <- report %>%
  select(Year,
         county = County,
         positive_cases = "Positive Cases") %>%
  mutate(date = Year) %>%
  filter(date %in% c("2006", "2007", "2008", "2009", "2010")) %>%
  group_by(date, county) %>%
  summarize(positive_cases = sum(positive_cases)) %>%
  ungroup %>%
  arrange(date)

kable(head(cases))
```

date	county	positive_cases
2006	Alameda	1
2006	Butte	31
2006	Colusa	3
2006	Contra Costa	8
2006	El Dorado	2
2006	Fresno	11

```
kable(tail(cases))
```

date	county	positive_cases
2010	Sacramento	12
2010	San Bernardino	5
2010	San Francisco	1
2010	San Joaquin	6
2010	Stanislaus	12
2010	Tulare	12

3.2. Temperature dataset

The main objective was join all the temperature data in one data set and select the relevant variables.

```
temp <- rbind(temp_2006, temp_2007, temp_2008, temp_2009, temp_2010)

temp <- temp %>%
  select(County, "County Code", "Month Day, Year Code", "Day of Year",
```

```

    "Avg Daily Max Air Temperature (F)",
    "Avg Daily Min Air Temperature (F)") %>%
  rename(county = County,
         fips = "County Code",
         date = "Month Day, Year Code",
         day_year = "Day of Year",
         max_temp = "Avg Daily Max Air Temperature (F)",
         min_temp_f = "Avg Daily Min Air Temperature (F)") %>%
  mutate(date = mdy(date))
kable(head(temp))

```

county	fips	date	day_year	max_temp	min_temp_f
Alameda County, CA	06001	2006-01-01	1	54.56	48.00
Alameda County, CA	06001	2006-01-02	2	51.49	46.46
Alameda County, CA	06001	2006-01-03	3	57.01	45.33
Alameda County, CA	06001	2006-01-04	4	62.25	49.38
Alameda County, CA	06001	2006-01-05	5	65.99	50.21
Alameda County, CA	06001	2006-01-06	6	67.36	51.97

3.3. Precipitation dataset

Same as the previous objceyive, precipitation datasets were joined and the relevant information was selected.

```

precip <- rbind(precip_2006, precip_2007, precip_2008, precip_2009, precip_2010)

precip <- precip %>%
  select(County, "Month Day, Year Code", "Avg Daily Precipitation (mm)")%>%
  rename(county = County,
         date = "Month Day, Year Code",
         avg_precip = "Avg Daily Precipitation (mm)") %>%
  mutate(date = mdy(date))

kable(head(precip))

```

county	date	avg_precip
Alameda County, CA	2006-01-01	3
Alameda County, CA	2006-01-02	31
Alameda County, CA	2006-01-03	1
Alameda County, CA	2006-01-04	0
Alameda County, CA	2006-01-05	0
Alameda County, CA	2006-01-06	0

4. Merging cleaned datasets

Two approaches were considered.

- First, precipitation and temperature datasets were merged. Right after that averages by year were calculated.
- Second,the yearly human West Niles Virus cases dataset was joined with weather data.

4.1. Weather data

In this step, temperature and precipitation data sets were merged. Additionally, calculation of yearly averages

```
ca_weather <- merge(temp, precip, by = c("county", "date")) %>%
  select(county, date, fips, avg_precip, max_temp) %>%
  separate(county, c("county", "state"), sep = " County, CA") %>%
  select(county, date, fips, avg_precip, max_temp) %>%
  mutate(date = year(date)) %>%
  group_by(county, fips, date) %>%
  summarise(avg_precip = mean(avg_precip), avg_max_temp = mean(max_temp)) %>%
  ungroup

kable(tail(ca_weather))
```

county	fips	date	avg_precip	avg_max_temp
Yolo	06113	2010	1.747945	73.43753
Yuba	06115	2006	2.838356	71.49586
Yuba	06115	2007	1.868493	72.56929
Yuba	06115	2008	1.792350	72.60995
Yuba	06115	2009	2.238356	72.58660
Yuba	06115	2010	3.378082	69.89458

4.2. Cases and weather data

In this step cases and weather dataset were joined

```
ca_weather_cases <- full_join(ca_weather, cases, by = c('date', 'county')) %>%
  mutate(positive_cases = ifelse(!is.na(positive_cases), positive_cases, 0)) %>%
  arrange(positive_cases)

kable(head(ca_weather_cases))
```

county	fips	date	avg_precip	avg_max_temp	positive_cases
Alameda	06001	2007	0.7506849	71.00534	0
Alameda	06001	2009	1.0739726	71.20956	0
Alpine	06003	2006	2.6575342	50.47945	0
Alpine	06003	2007	1.6082192	51.38460	0
Alpine	06003	2008	2.0327869	52.24068	0
Alpine	06003	2009	2.3671233	51.35307	0

5. Creating the final data set

To create the final dataset and in order to be able to create a Shiny App, geographical data from California counties was joined with human cases and weather data (joined in part 4.2). In addition, for the structure of the data just counties that are part of Southern California will be analyzed.

```
ca_counties <- counties(state = "CA", cb = TRUE, class = "sf")
```

```
##
```

```
|  
|  
|
```

```
| 0%
```

	1%
=	1%
=	2%
==	2%
==	3%
==	4%
===	4%
===	5%
====	5%
====	6%
====	7%
=====	7%
=====	8%
=====	9%
=====	10%
=====	10%
=====	11%
=====	12%
=====	13%
=====	13%
=====	14%
=====	15%
=====	15%
=====	16%
=====	16%
=====	17%
=====	18%

=====	18%
=====	19%
=====	19%
=====	20%
=====	21%
=====	21%
=====	22%
=====	22%
=====	23%
=====	24%
=====	24%
=====	25%
=====	26%
=====	27%
=====	27%
=====	28%
=====	28%
=====	29%
=====	30%
=====	30%
=====	31%
=====	32%
=====	33%
=====	33%
=====	34%
=====	35%
=====	35%

=====	36%
=====	36%
=====	37%
=====	38%
=====	38%
=====	39%
=====	39%
=====	40%
=====	41%
=====	41%
=====	42%
=====	42%
=====	43%
=====	44%
=====	44%
=====	45%
=====	45%
=====	46%
=====	47%
=====	47%
=====	48%
=====	48%
=====	49%
=====	50%
=====	50%
=====	51%
=====	52%

=====	53%
=====	53%
=====	54%
=====	55%
=====	55%
=====	56%
=====	56%
=====	57%
=====	58%
=====	58%
=====	59%
=====	59%
=====	60%
=====	61%
=====	61%
=====	62%
=====	62%
=====	63%
=====	64%
=====	64%
=====	65%
=====	65%
=====	66%
=====	67%
=====	67%
=====	68%
=====	69%

=====	70%
=====	70%
=====	71%
=====	72%
=====	72%
=====	73%
=====	73%
=====	74%
=====	75%
=====	75%
=====	76%
=====	76%
=====	77%
=====	78%
=====	78%
=====	79%
=====	79%
=====	80%
=====	81%
=====	81%
=====	82%
=====	82%
=====	83%
=====	84%
=====	84%
=====	85%
=====	85%

=====	86%
=====	87%
=====	87%
=====	88%
=====	89%
=====	90%
=====	90%
=====	91%
=====	92%
=====	93%
=====	93%
=====	94%
=====	95%
=====	95%
=====	96%
=====	96%
=====	97%
=====	98%
=====	98%
=====	99%
=====	99%
=====	100%

```

ca_county_cases <- ca_counties %>%
  mutate(fips = paste(STATEFP, COUNTYFP, sep = "")) %>%
  full_join(ca_weather_cases, by = "fips") %>%
  filter(county %in% c("Imperial" , "Kern", "Los Angeles", "Orange",
    "Riverside", "San Bernardino", "San Diego",
    "San Luis Obispo", "Santa Barbara", "Ventura"))

#kable(head(ca_county_cases, 3))

unique(ca_county_cases$county)

```

```
## [1] "Los Angeles"      "Santa Barbara"    "San Bernardino"
## [4] "Ventura"          "Orange"           "Imperial"
## [7] "San Luis Obispo"  "Riverside"        "Kern"
## [10] "San Diego"
```

```
unique(ca_county_cases$date)
```

```
## [1] 2010 2006 2009 2007 2008
```