Exploratory Data Analysis

Ian Bachli

10/9/2019

Load the required libraries for exploratory analysis.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ readr 1.3.1  
## ✔ tibble 2.1.3 ✔ purrr 0.3.2  
## ✔ tidyr 1.0.0 ✔ stringr 1.4.0  
## ✔ ggplot2 3.2.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(forcats)  
library(ggthemes)  
library(knitr)  
library(naniar)  
library(broom)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

Load the processed data from the RDS.

load("~/Desktop/epid8060fall2019/IanDBachli-Project/data/processed\_data/colmozzie.rda")

Then take a look at the data.

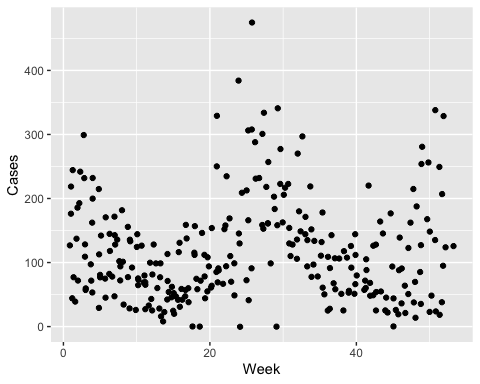
glimpse(colmozzie)

## Observations: 279  
## Variables: 12  
## $ Cases <int> 44, 39, 57, 53, 29, 45, 47, 34, 28, 26, 27, 25, 28, 23, 20…  
## $ Year <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009…  
## $ Week <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,…  
## $ TEM <dbl> 27.3000, 26.4000, 27.1400, 26.8000, 26.7750, 26.8500, 26.2…  
## $ TMAX <dbl> 32.7000, 29.7667, 32.0200, 31.0000, 30.0750, 30.5500, 30.9…  
## $ Tm <dbl> 23.6000, 23.9000, 23.5200, 23.3500, 23.6000, 23.0000, 22.1…  
## $ SLP <chr> "1010.7", "1010.7", "1012.58", "1009.9", "1010.1", "1012.0…  
## $ H <dbl> 68.0000, 78.6667, 67.0000, 68.0000, 78.0000, 72.5000, 65.2…  
## $ PP <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 17.2100, 0.0000, 0.0000, 0…  
## $ VV <dbl> 13.0000, 18.3333, 20.0000, 20.0000, 18.9750, 20.0000, 20.0…  
## $ V <dbl> 11.1000, 5.8333, 6.1200, 7.1500, 3.2000, 5.5500, 4.7200, 4…  
## $ VM <dbl> 20.6000, 9.2000, 10.4600, 11.1000, 6.8250, 10.3000, 9.2800…

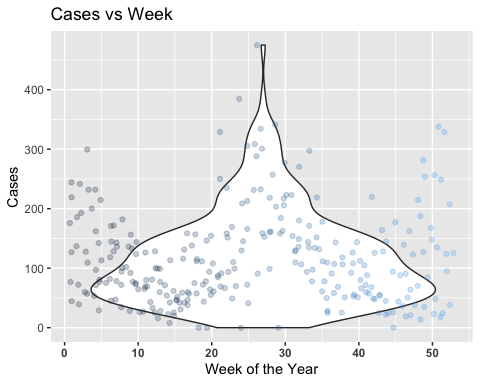
Once the data has been inspected, we can now start taking a look at some variables and how they stack up in an exploratory analysis. To do this, we will need to generate some plots.

For my first plot, I would like to investigate the Number of Cases of Dengue per Week in the Year to see if there is any correlation to seasonal changes. I will attempt to use a Jitter Plot to look for anything intertesting.

colmozzie %>% ggplot() +   
 geom\_jitter (aes(x = Week, y = Cases))



Cases\_v\_Week <- ggplot(colmozzie, aes(x = Week, y = Cases, color = Week)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Week") + xlab("Week of the Year") + ylab("Cases") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_Week

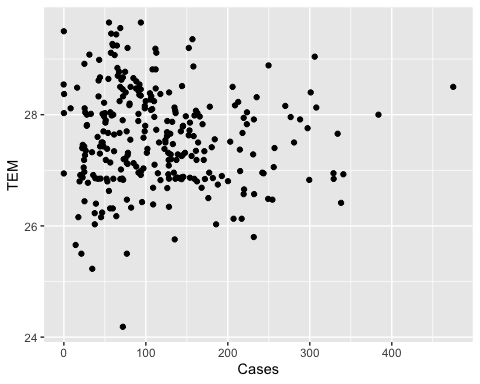


ggsave(filename = "../../results/Cases\_v\_Week.png",plot = Cases\_v\_Week)

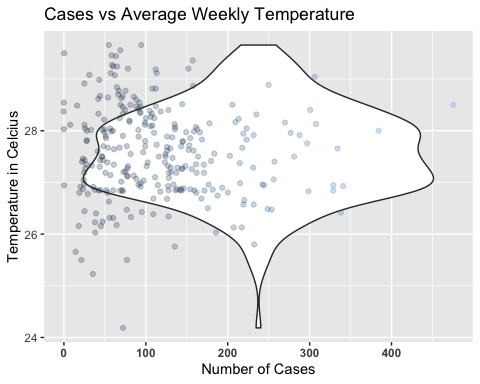
## Saving 5 x 4 in image

Next we will take a look at the number of Cases present when comparing it to the Average Temperature of the Week.

colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = TEM))



Cases\_v\_Temperature <- ggplot(colmozzie, aes(x = Cases, y = TEM, color = Cases)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Average Weekly Temperature") + xlab("Number of Cases") + ylab("Temperature in Celcius") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_Temperature

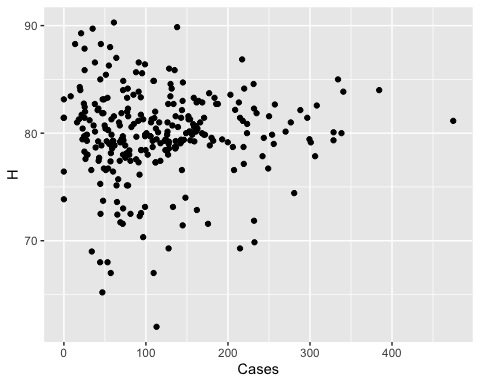


ggsave(filename = "../../results/Cases\_v\_Temperature.png",plot = Cases\_v\_Temperature)

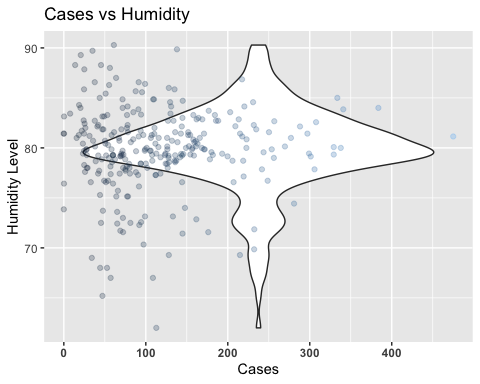
## Saving 5 x 4 in image

Now here is a graph comparing the Number of Cases of Dengue in comparison to the Humidity.

colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = H))



Cases\_v\_Humidity <- ggplot(colmozzie, aes(x = Cases, y = H, color = Cases)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Humidity") + xlab("Cases") + ylab("Humidity Level") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_Humidity

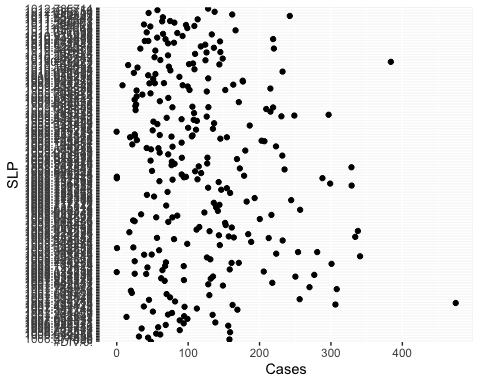


ggsave(filename = "../../results/Cases\_v\_Humidity.png",plot = Cases\_v\_Humidity)

## Saving 5 x 4 in image

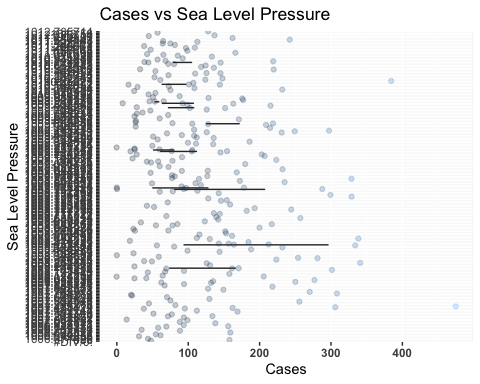
Let’s check out the number of Cases in Comparison to the Sea Level Pressure.

colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = SLP))



Cases\_v\_SLP <- ggplot(colmozzie, aes(x = Cases, y = SLP, color = Cases)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Sea Level Pressure") + xlab("Cases") + ylab("Sea Level Pressure") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_SLP

## Warning: position\_dodge requires non-overlapping x intervals



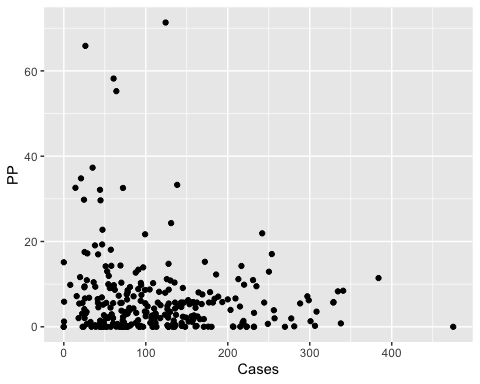
ggsave(filename = "../../results/Cases\_v\_SLP.png",plot = Cases\_v\_SLP)

## Saving 5 x 4 in image

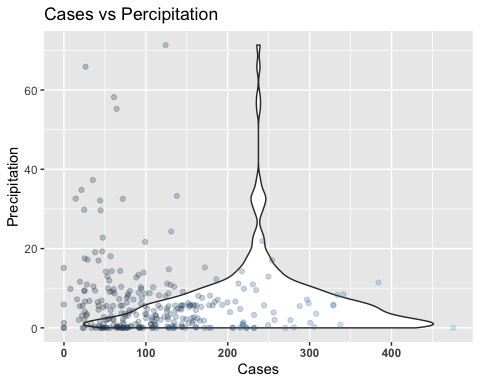
## Warning: position\_dodge requires non-overlapping x intervals

Next let’s look at the comparison of the Number of Cases against the Precipitation Level.

colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = PP))



Cases\_v\_Precipitation <- ggplot(colmozzie, aes(x = Cases, y = PP, color = Cases)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Percipitation") + xlab("Cases") + ylab("Precipitation") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_Precipitation

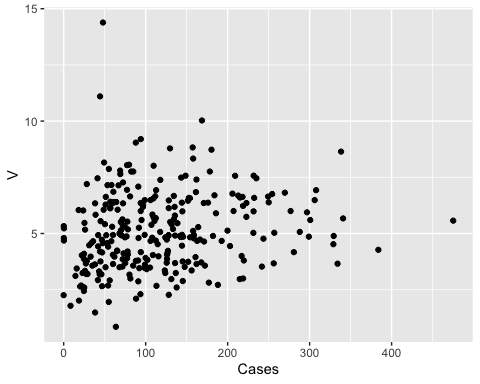


ggsave(filename = "../../results/Cases\_v\_Precipitation.png",plot = Cases\_v\_Precipitation)

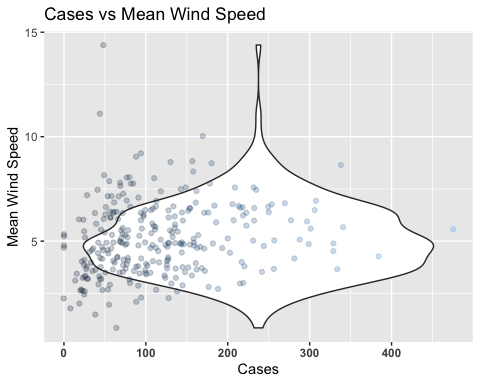
## Saving 5 x 4 in image

A big interest of mine was examining the number of cases of Dengue with the Maximum Sustained Wind Speed and Average Wind Speed. My interest in this comes from my theory that because Mosquitos have a harder time flying around in higher wind speeds due to their body mass and strength, there will be less Dengue cases because they cannot locate targets and position themselves in places as easily to spread the disease. First we will examine the Average Wind Speed, then the Maximum Sustained Windspeed.

colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = V))



Cases\_v\_MeanWindspeed <- ggplot(colmozzie, aes(x = Cases, y = V, color = Cases)) + geom\_violin(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Mean Wind Speed") + xlab("Cases") + ylab("Mean Wind Speed") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))  
  
Cases\_v\_MeanWindspeed

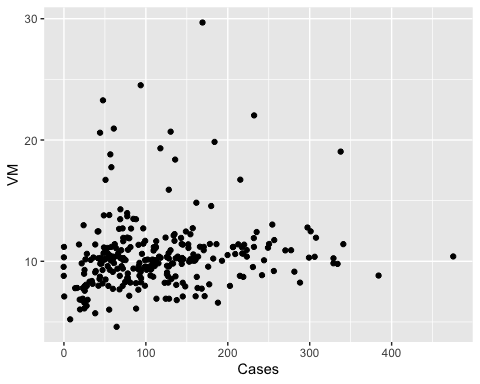


ggsave(filename = "../../results/Cases\_v\_MeanWindspeed.png",plot = Cases\_v\_MeanWindspeed)

## Saving 5 x 4 in image

Now here is Maximum Sustained Windspeed.

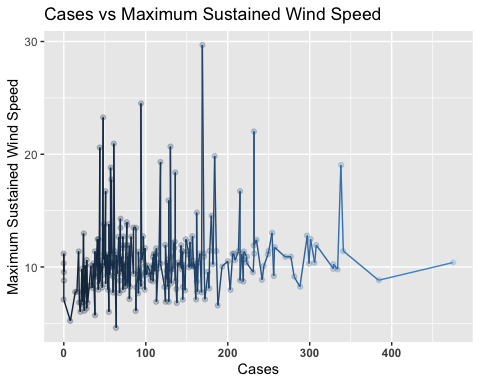
colmozzie %>% ggplot() +   
 geom\_jitter(aes(x = Cases, y = VM))



Cases\_v\_MaxWindspeed <- ggplot(colmozzie, aes(x = Cases, y = VM, color = Cases)) + geom\_line(width = 1) + geom\_jitter(alpha = 0.25) + ggtitle("Cases vs Maximum Sustained Wind Speed") + xlab("Cases") + ylab("Maximum Sustained Wind Speed") + theme(legend.position = "none", axis.text.x = element\_text(face = "bold"))

## Warning: Ignoring unknown parameters: width

Cases\_v\_MaxWindspeed

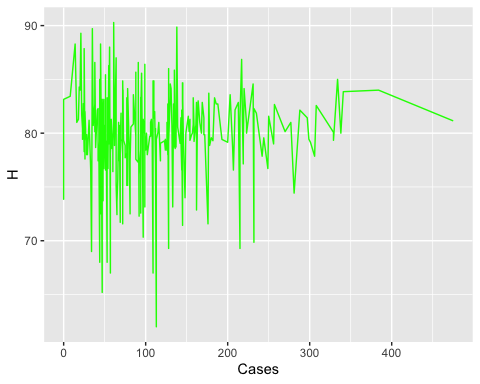


ggsave(filename = "../../results/Cases\_v\_MaxWindspeed.png",plot = Cases\_v\_MaxWindspeed)

## Saving 5 x 4 in image

Below are just some test plots that I just wanted to test out in order to to see what they would look like in other graph types.

ggplot(colmozzie) +   
 geom\_line(mapping = aes(x = Cases, y = H), color = "green")



ggplot(colmozzie) +  
geom\_bar(mapping = aes(x = Week, y = Cases), color = "Navy", stat = "identity") +   
 ylab("Cases")

