Weather Data Analysis

Daniel Starer

2025-07-24

A bit of context first. This data originates from Zaruhi Avagyan on Kaggle and no information was provided by the creator regarding what locale it originates from. The data, nonetheless, is very detailed and I wanted to visualize the correlation between humidity and precipitation and temperature and precipitation.

First, we will load the dataset.

```
my_weather <- read.csv("weather.csv")
```

Next, to best show correlation, we will need to use ggplot2, so we'll load that library.

```
library(ggplot2)
```

Now, let's take a look at the data to get a sense of the data's structure.

```
head(my_weather)
```

```
##   MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed
## 1     8.0    24.3      0.0         3.4      6.3          NW            30
## 2    14.0    26.9      3.6         4.4      9.7         ENE            39
## 3    13.7    23.4      3.6         5.8      3.3          NW            85
## 4    13.3    15.5     39.8         7.2      9.1          NW            54
## 5     7.6    16.1      2.8         5.6     10.6         SSE            50
## 6     6.2    16.9      0.0         5.8      8.2          SE            44
##   WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm
## 1         SW         NW            6           20          68          29
## 2          E          W            4           17          80          36
## 3          N        NNE            6            6          82          69
## 4        WNW          W           30           24          62          56
## 5        SSE        ESE           20           28          68          49
## 6         SE          E           20           24          70          57
##   Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday RISK
_MM
## 1      1019.7      1015.0        7        7    14.4    23.6        No
3.6
## 2      1012.4      1008.4        5        3    17.5    25.7       Yes
3.6
## 3      1009.5      1007.2        8        7    15.4    20.2       Yes    3
9.8
## 4      1005.5      1007.0        2        7    13.5    14.1       Yes
2.8
## 5      1018.3      1018.5        7        7    11.1    15.4       Yes
0.0
## 6      1023.8      1021.7        7        5    10.9    14.8        No
```

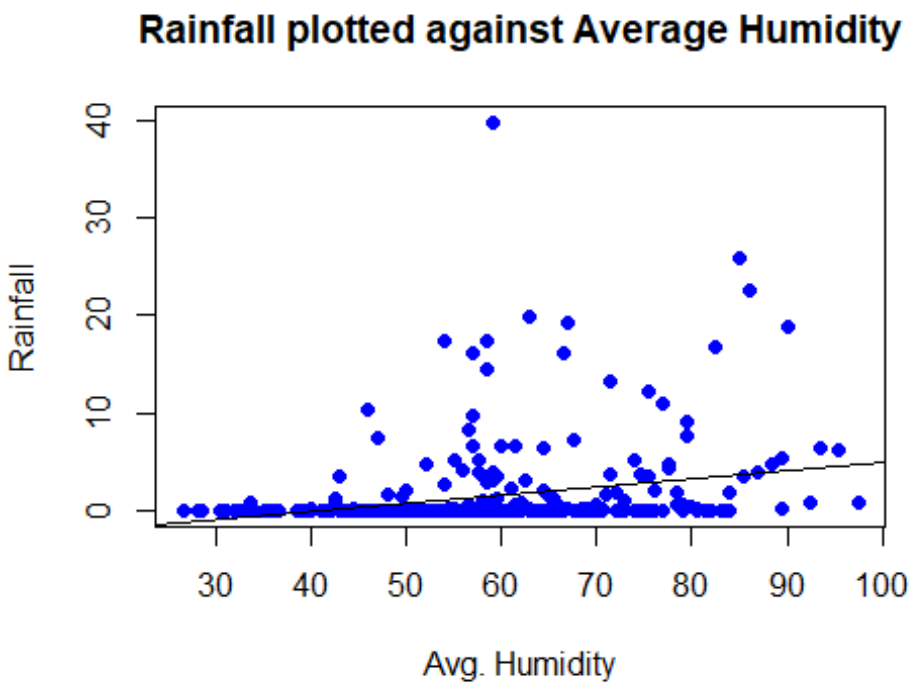
```
0.2
##   RainTomorrow
## 1          Yes
## 2          Yes
## 3          Yes
## 4          Yes
## 5           No
## 6           No
```

There are multiple columns for humidity and temperature because there are two times in which humidity and temperature were recorded: 9 AM and 3 PM. Because of this, we will need to do some cleaning to organize things a bit. We can achieve this by combining the data in both humidity columns and the data in both temperature columns and averaging them out to find the average temperature and humidity.

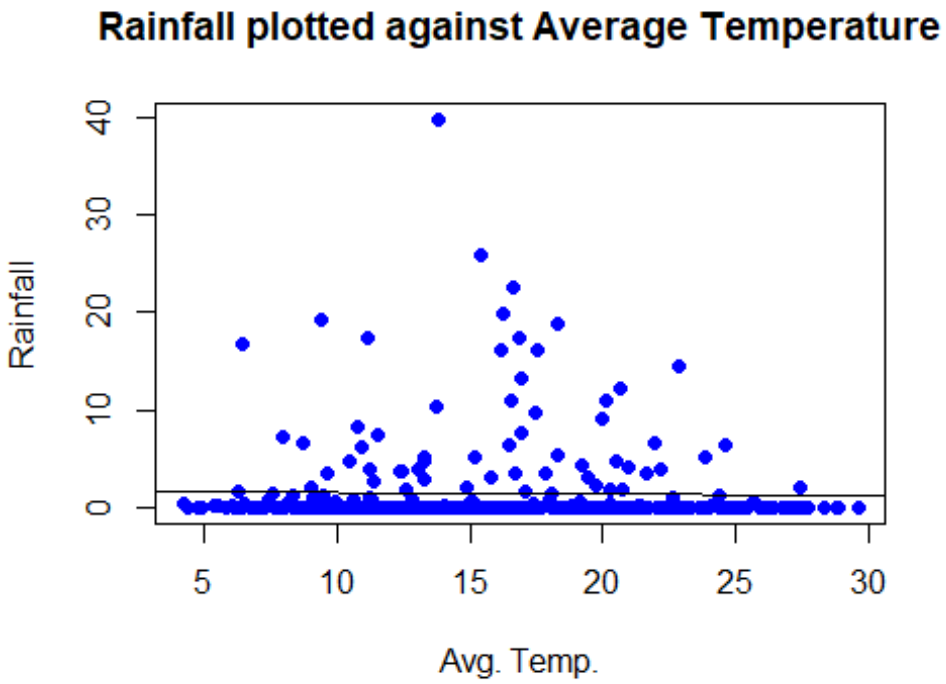
```
my_weather$AvgHumidity <- rowMeans(my_weather[, c("Humidity9am", "Humidity3pm
")], na.rm = TRUE)
my_weather$AvgTemp <- rowMeans(my_weather[, c("Temp9am", "Temp3pm")], na.rm =
TRUE)
AvgHumidity <- my_weather$AvgHumidity
AvgTemp <- my_weather$AvgTemp
```

Now that we have separated the average humidity and average temperature into their own objects, we can create an object for rainfall and test the correlation between these objects with linear models.

```
Rainfall <- my_weather$Rainfall

plot(AvgHumidity, Rainfall, pch = 16, col = "blue", main = "Rainfall plotted
against Average Humidity", xlab = "Avg. Humidity", ylab = "Rainfall")
lm(Rainfall ~ AvgHumidity)

##
## Call:
## lm(formula = Rainfall ~ AvgHumidity)
##
## Coefficients:
## (Intercept)  AvgHumidity
##    -3.39557      0.08278

abline(lm(Rainfall ~ AvgHumidity))
```




```
plot(AvgTemp, Rainfall, pch = 16, col = "blue", main = "Rainfall plotted agai
nst Average Temperature", xlab = "Avg. Temp.", ylab = "Rainfall")
lm(Rainfall ~ AvgTemp)

##
## Call:
## lm(formula = Rainfall ~ AvgTemp)
##
## Coefficients:
## (Intercept)      AvgTemp
##    1.568418    -0.008864

abline(lm(Rainfall ~ AvgTemp))
```



From these visualizations, we can see there is a general correlation between humidity and rainfall and temperature and rainfall, however there are a fair amount of outliers. Instances where there is no recorded rainfall especially conforms more to the humidity and temperature.