CSE 6242 Activity 1

Vincent La (Georgia Tech ID - vla6)

September 12, 2017

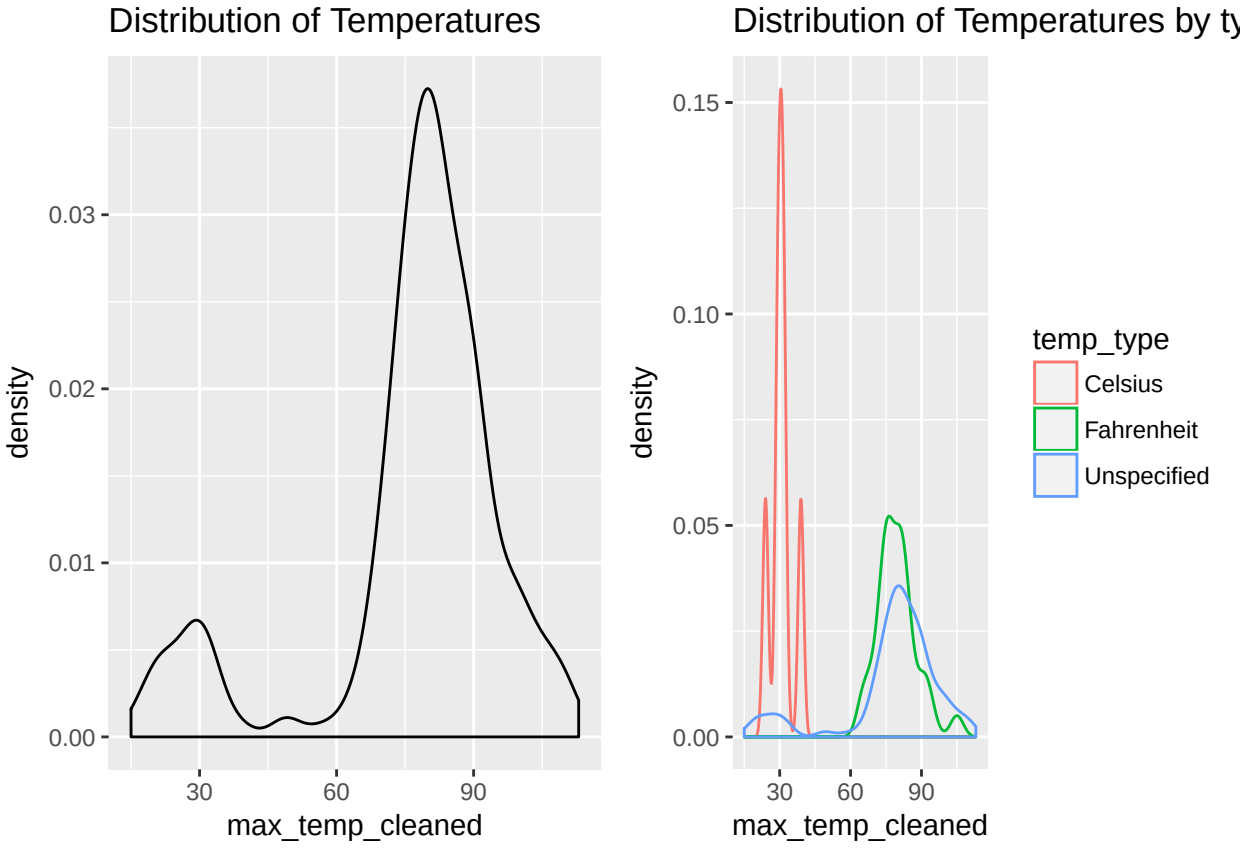
Introduction

First, we load the data and clean a few of the fields. We don't show it in this report, but a couple of things I did was clean the timestamps, extract numerics from temperature, and extract the temperature type (Celsius vs Fahrenheit).

Exploring Temperature

The first thing we're going to look at is temperature. I find this interesting because I have this hypothesis that students in the United States of America are more likely to record their temperatures in Fahrenheit. Students in other parts of the world are more likely to record their temperatures in Celsius. I want to better explore the temperatures recorded in this data set.

After cleaning the data and converting to a numeric, we can plot some distributions.



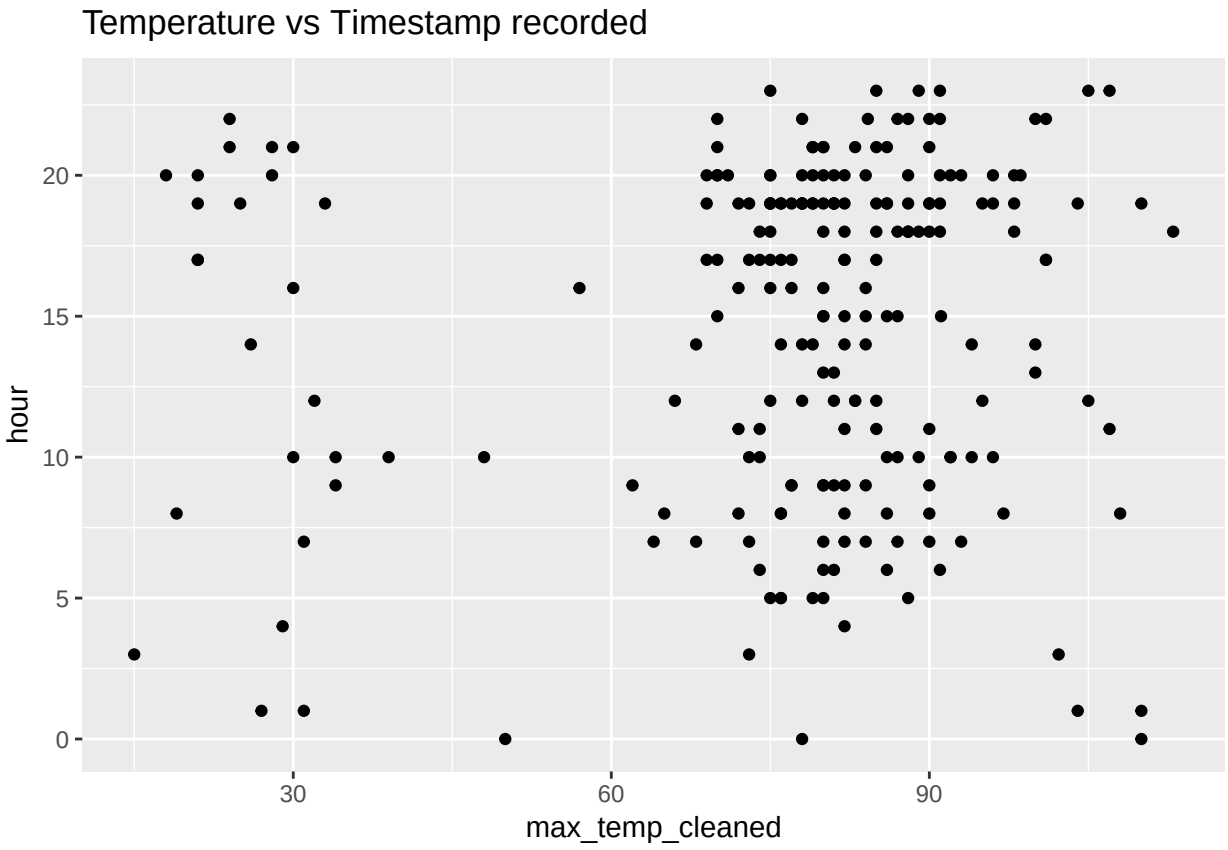
The first plot is a plot of the distribution of all the temperatures recorded. The second plot is a plot grouped by the temperature type recorded (e.g. Fahrenheit or Celsius). If the temperature type was not inputted, we default to a value "Unspecified".

I find the first plot interesting because we see that the distribution is very slightly bimodal, but mostly concentrated on the higher end. More specifically, there is a small bump at around 30, but most of the data points are again between 70 to 80 degrees. This probably means that most of the temperatures are recorded in Fahrenheit, but some are recorded in Celsius. We can explore this more by looking at the second plot.

In the second plot, we can look more directly at the temperatures by type. For the temperatures that were recorded as Celsius, most of the temperatures recorded are around 30 degrees. For Fahrenheit, most of the temperatures recorded are around 70 to 80 degrees. This makes sense as it is the summer for the northern hemisphere. For the data points that are "Unspecified", the distribution looks similar to the aggregate distribution.

Relationship Between Time and Temperature

I also wanted to explore the relationship between time and temperature recorded. The hypothesis here would be that those who recorded in Celsius maybe from non-USA countries, whereas those who recorded in Fahrenheit may be from USA, which we might be able to intuit from the timestamps.



Superficially, it doesn't look like there is much of a relationship between time and temperature recorded.