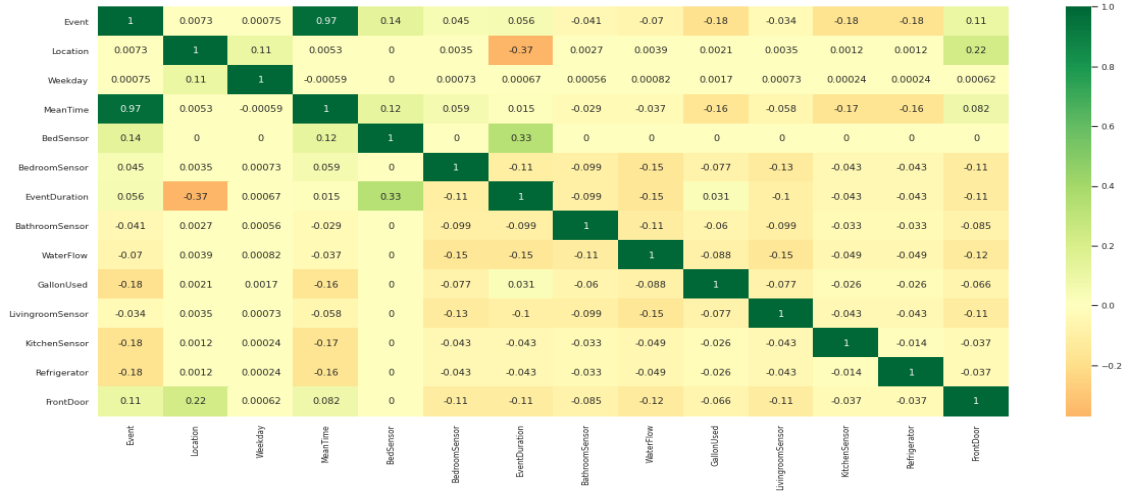

Source: Capstone Project

This is a heatmap from my graduate capstone project. The goal of this project is to predict daily activities using the data we have collected. In order to reach the target, we introduced a series of factors like “Events”, “Location” and “Weekday” etc (Shown in the x and y axis on the graph above). However, when we train our data, since there are many factors and each factor has a lot of data with it, we found that if we use all the factors in the training process, the running time of the program will be very long (needs about 1 hour to finish all the predictions). Thus, in order to reduce the running time, we decided to draw a heatmap to find the correlation between each factor and manually drop off factors that are less related.

“How to view this chart”

To find the correlation between each factor, we could first briefly check the color of each column, the greener the color, the higher the correlation between this factor and the current factor (thus, of course the correlation of the factors itself is always 1). For example, the correlation between “mean time” and “event” is very closely related(0.97). By drawing the heatmap, we found that a lot of factors do not impact the prediction accuracy at all. For example, the “Bathroom sensor” factor has negative correlation with “event”(-0.041). Therefore, during the model training procedure, we could drop off all the “less correlated” columns, by manually dropping all less related data, the model training time is greatly improved without sacrificing a lot of prediction accuracy.