Classroom Temperature
Exploratory Analysis

Brett Tully, btully@bellarmine.edu

I. INTRODUCTION

I am a huge believer in automation and artificial intelligence, but more importantly, improving the efficiency/convenience of our lives. With the IOT, we have unlimited possibilities of what we can do with technology. For my Senior Project, I wanted to do something different from what most people do and I really wanted to work with live data because I think live data is fascinating and the most interesting to people. My dataset is put together by the use of three Raspberry Pi's scattered between three different classrooms, all with a temperature sensor connected to them. These sensors report Room Number, Temperature, Humidity, and a Timestamp every 2 seconds to an SQL database in the cloud.

II. DATA SET DESCRIPTION

As of 3/29/2023 2:50 PM, we have a total of 1,330,901 rows and 4 columns in the raw data format. I have created a script that when I pull the data into a data frame from the SQL database, it creates a separate column for Time, Weekday, and Date from the datetime column. This then fills out the 7 columns below.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
V1: Device	Object	0%
V2: Datetime	Datetime64	0%
V3: Temp F	Float64	0%
V4: Humidity	Float64	0%
V5: Time	Object	0%
V6: Weekday	Int64	0%
V7: Date	object	0%

III. Data Set Summary Statistics

With this dataset, I am limited to very little features. My main feature will come from assigning a 0 or 1 to mark when the classes are currently in session versus not being in session. These marks should help the model's ability to predict occupancy based on the temperature changes.

With the variable "Temp F", we have seen some odd outliers in the data. Most notable being the min/max of the temperature. This indicates that we have a sensor issue, because there will never be a time when a room is -4.18 degrees F or 800.96 degrees F.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Temp F	*1,330,901*	67.02	2.89	-4.18	65.30	67.28	68.72	800.96
Humidity	*1,330,901*	42.33	14.00	1.40	31.30	41.80	51.00	84.70

There should be a table for **EACH** categorical variable.

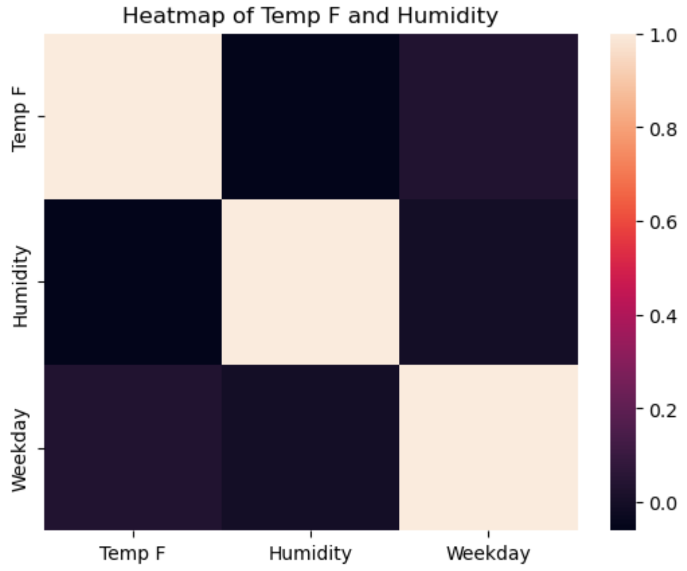
Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
P006B	*344011*	*32.32%*
P002	*344794*	*32.39%*
P004	*375535*	*35.28%*

In an ideal situation, the frequency and proportion of the rooms would be equal, but due to various issues, such as the disconnection of one of the devices in the room, it will cause the data to not be reported to the sql database.

Table 4: Correlation Table/Tables

	Temp F	Humidity	Weekday
Temp F	1.0000	-0.060589	0.034679
Humidity	-0.060589	1.0000	-0.006124
Weekday	0.034679	-0.006124	1.0000



IV. DATA SET GRAPHICAL EXPLORATION

One thing very exciting about this dataset is although the scope of my data is narrow, my abilities with visualizations are great. I can view dating pertaining to particular days of the week, times of the day, months of the year, specific classrooms, and much more. I include findings/descriptions with each graph below if there is anything worth noting.

A. Distributions

i. Histograms of Temp F & Humidity

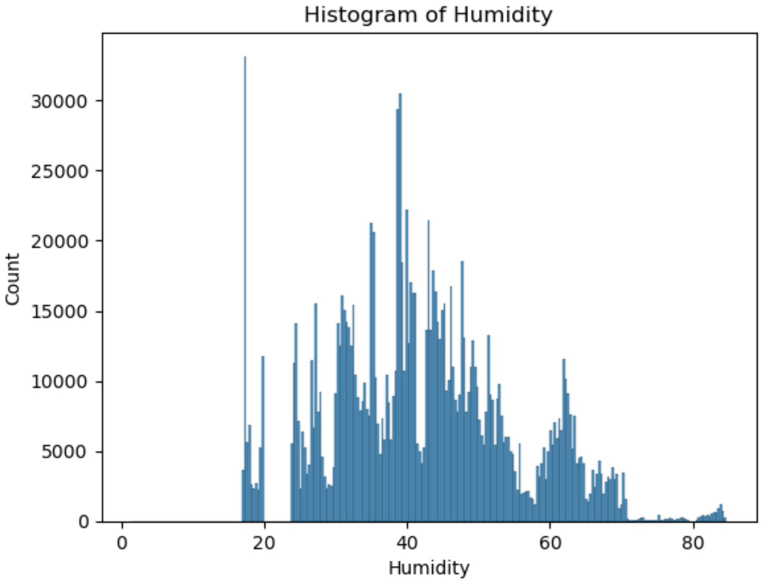


Figure 1: Histogram of Humidity from dataset: Humidity seems to be slightly skewed left, with most of the distribution being around 40%.

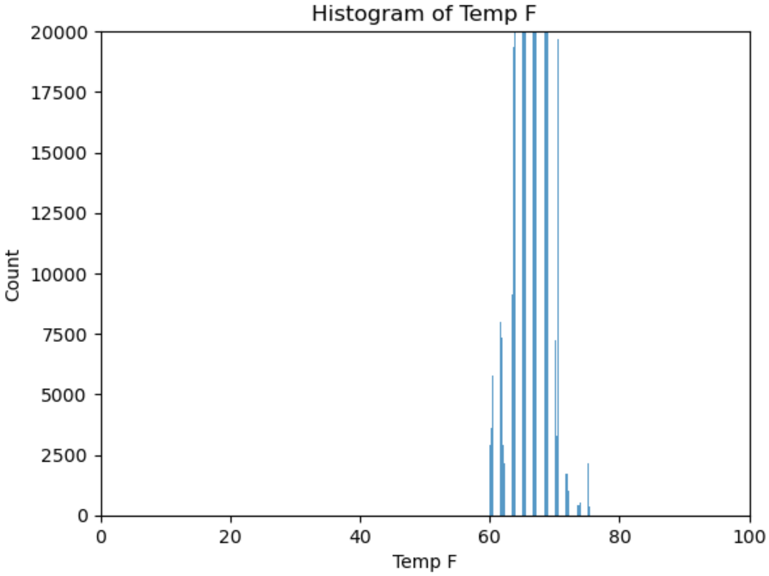


Figure 2: Histogram of Temp F from dataset: Our Temp F data seems to be very consistent in the temperatures

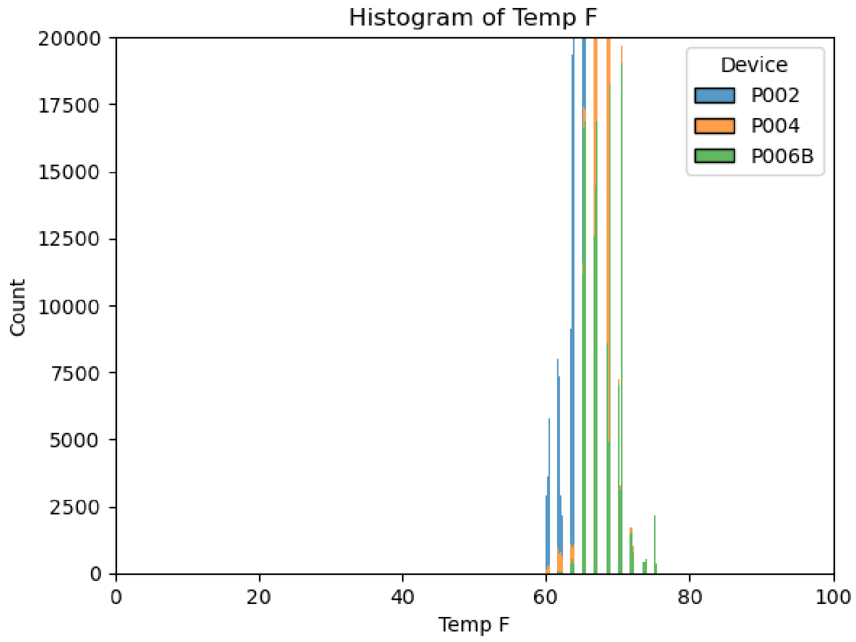


Figure 3: Histogram of Temp F with Device as the hue from dataset

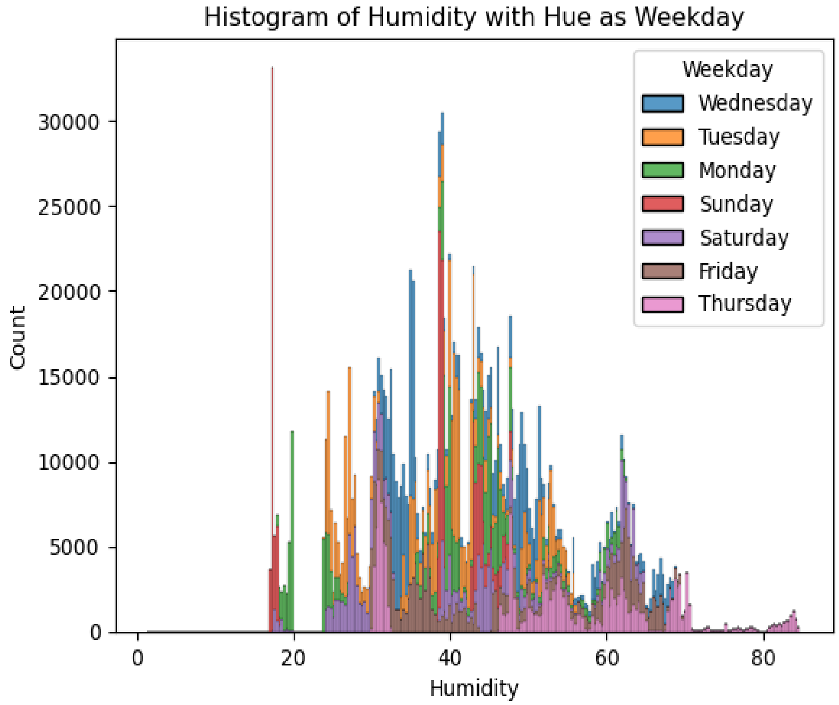


Figure 4: Histogram of Humidity with Weekday as the hue from dataset

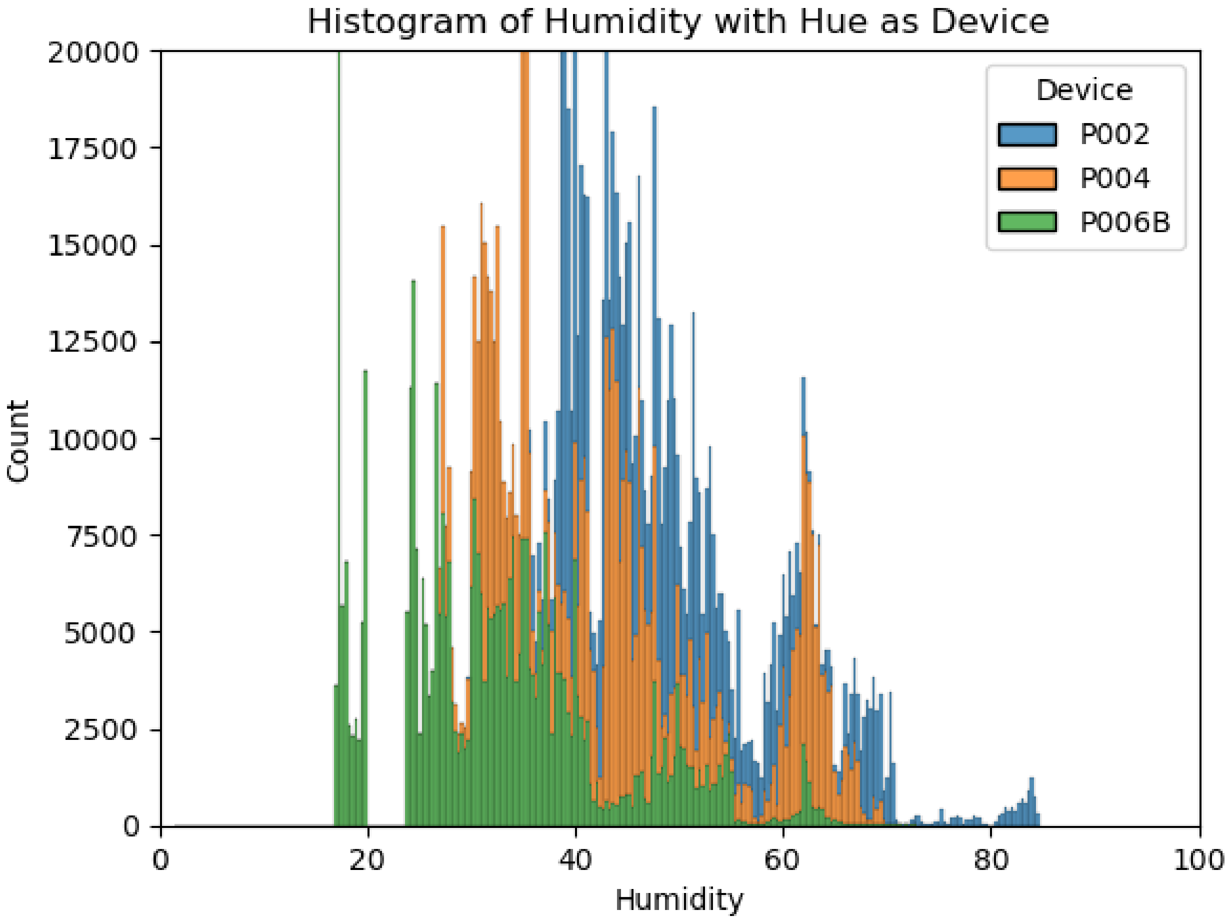


Figure 4: Histogram of Humidity with Device as the hue from dataset: Humidity for P002 seems to skew left more than the other rooms. P006B also seems to have an odd distribution.

B. ScatterPlots / Pairwise Plots (continuous variables)

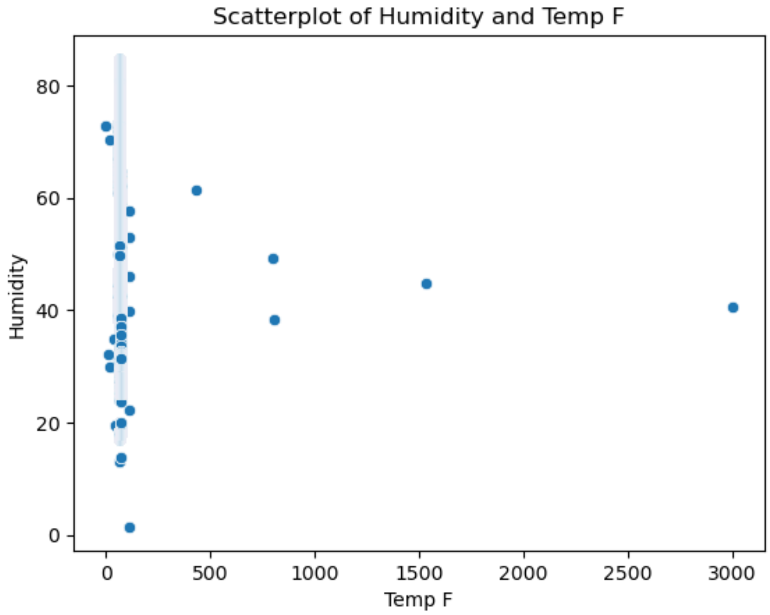


Figure 4: Scatterplot of Humidity/Temp from dataset (With Outliers): With the heavy outliers included, we can easily see the Temperature sensor issues I spoke about previously

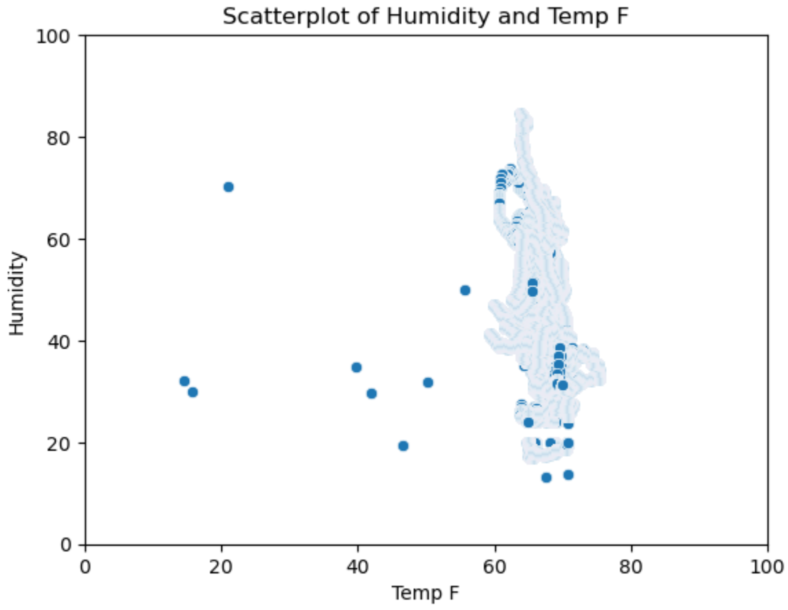


Figure 4: Scatterplot of Humidity/Temp from dataset (Without Outliers): This visual is much better without the outliers. We can clearly see clusters between 60-80 in Temp F while the Humidity ranges from 20%-90%

C. *Bar charts (categorical variables)*

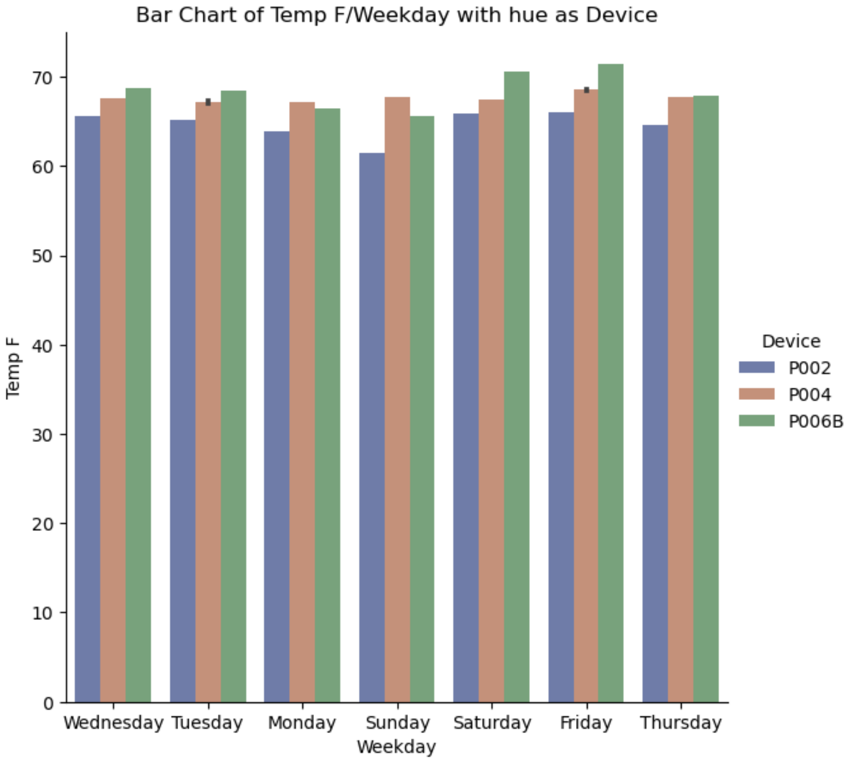


Figure 4: Categorical Bar Chart of Temp F/Weekday with Device as the hue from dataset

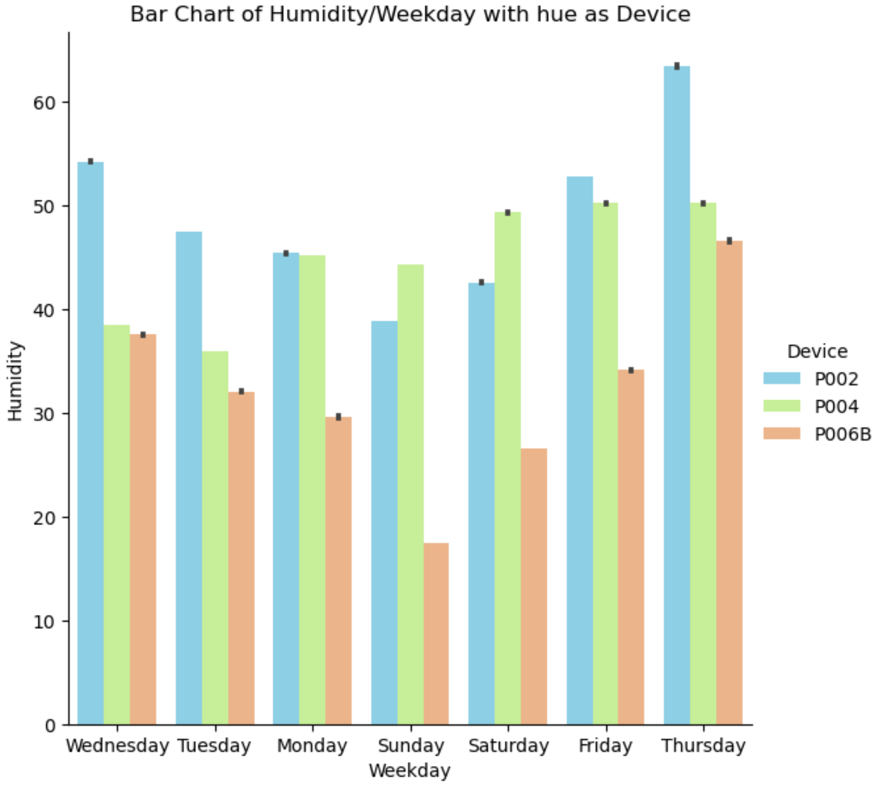


Figure 4: Categorical Bar Chart of Humidity/Weekday with Device as the hue from dataset: P006B seems to maintain lower humidity levels than the other rooms throughout the week

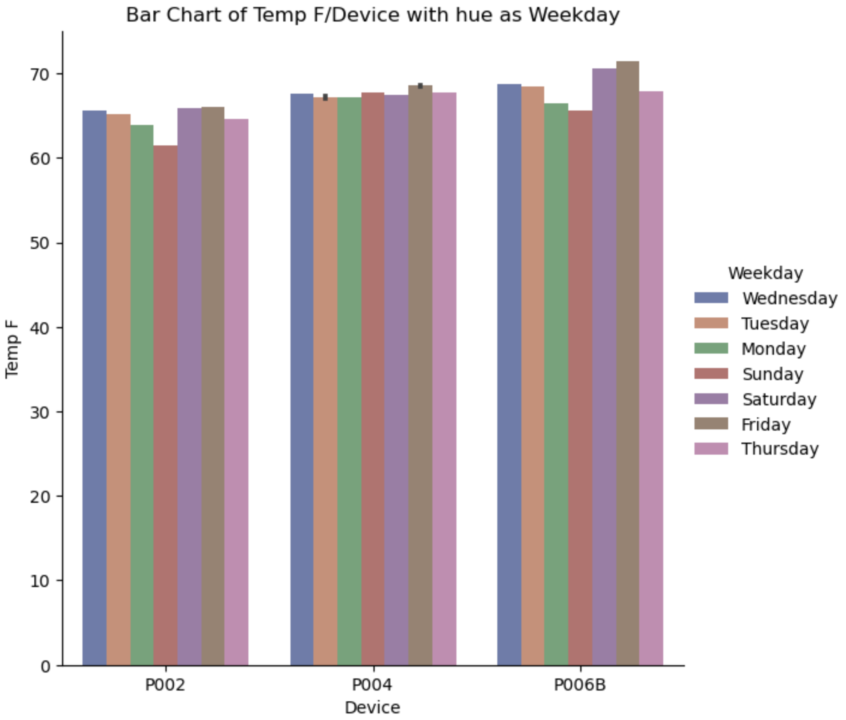


Figure 4: Categorical Bar Chart of Temp F/Device with Weekday as the hue from dataset: P006B seems, on average, to have the highest temperatures throughout the week. It is important to note that this room has no windows.

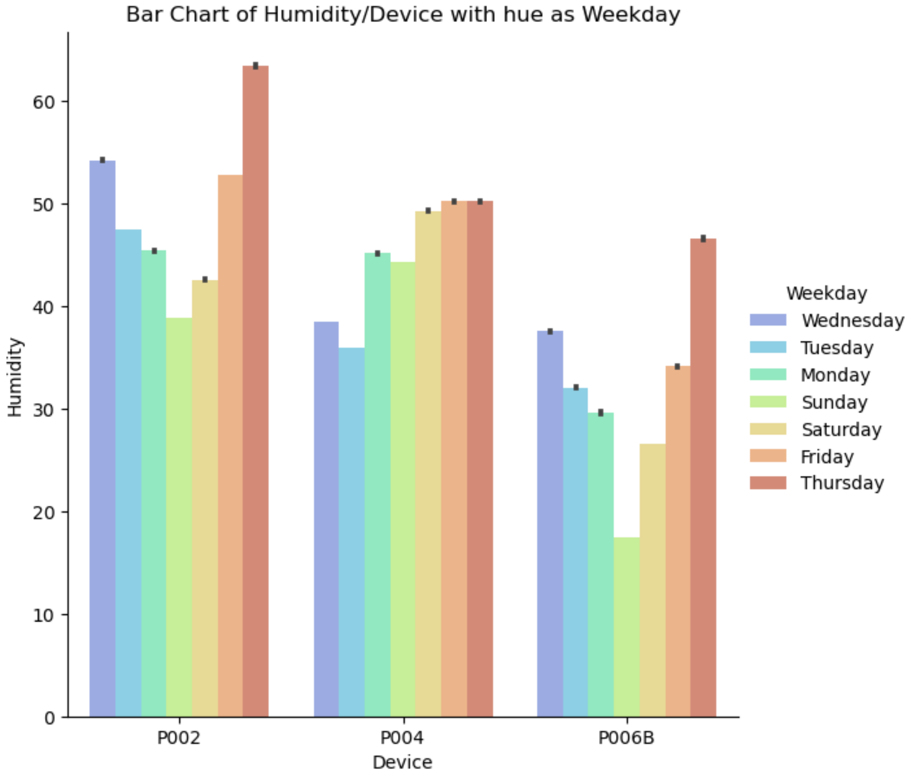


Figure 4: Categorical Bar Chart of Humidity/Device with Weekday as the hue from dataset:

D. Other Plots

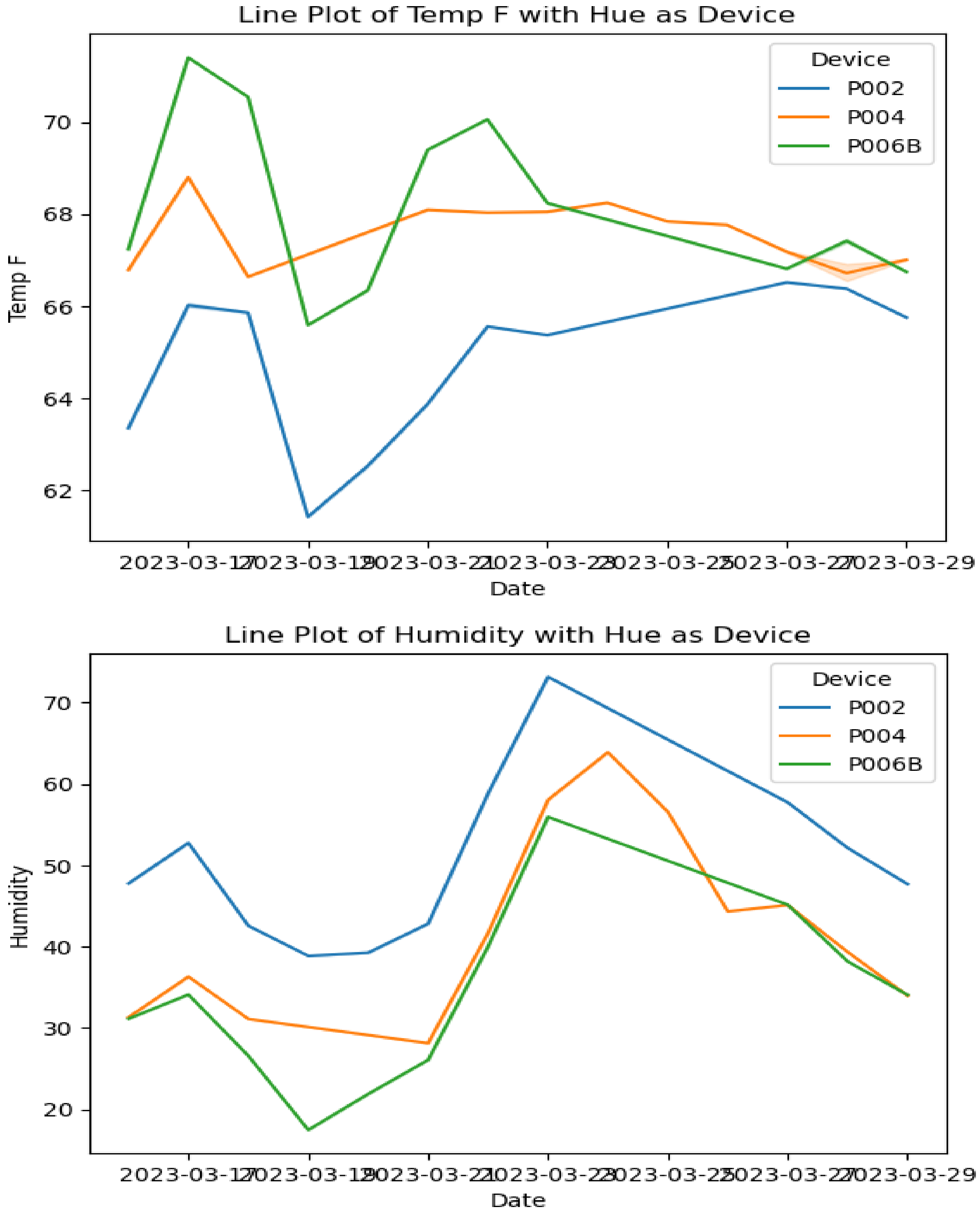


Figure 2: (a) Line plot of Humidity/Date with hue as Device (b) Line plot of Temp F/Date with hue as Device (multiple plots): Although the date can be slightly hard to ready, the starting date is March 12. It is quite interesting to see the high correlation between Temp and Humidity. From these graphs, we most likely can concur that we suffered some cold few days in the month of march, which explains the dip in Temp and Humidity. Clearly, these two variables rise and fall together. To support this claim, humidity is relatively low during cold temperatures.

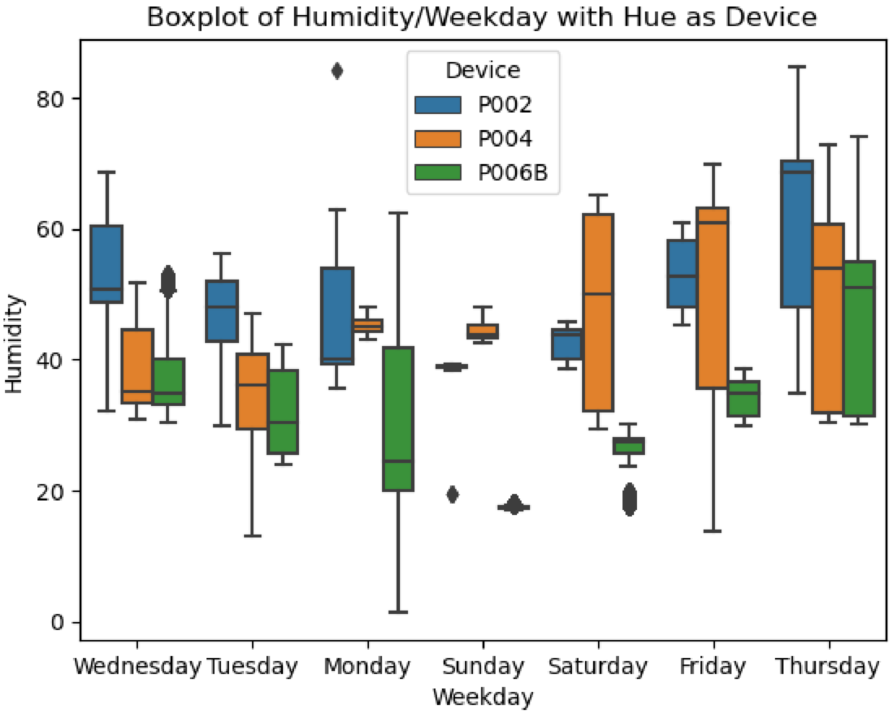


Figure 4: Box Chart of Humidity/Weekday with Device as the hue from dataset

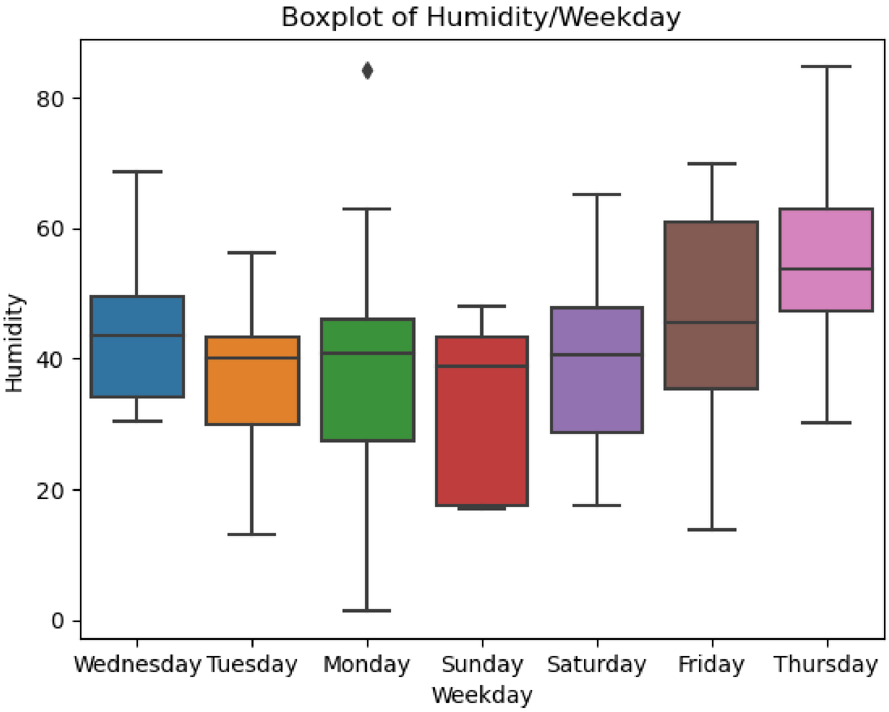


Figure 4: Box Chart of Humidity/Weekday from dataset: Thursday seems to typically have the highest levels of humidity

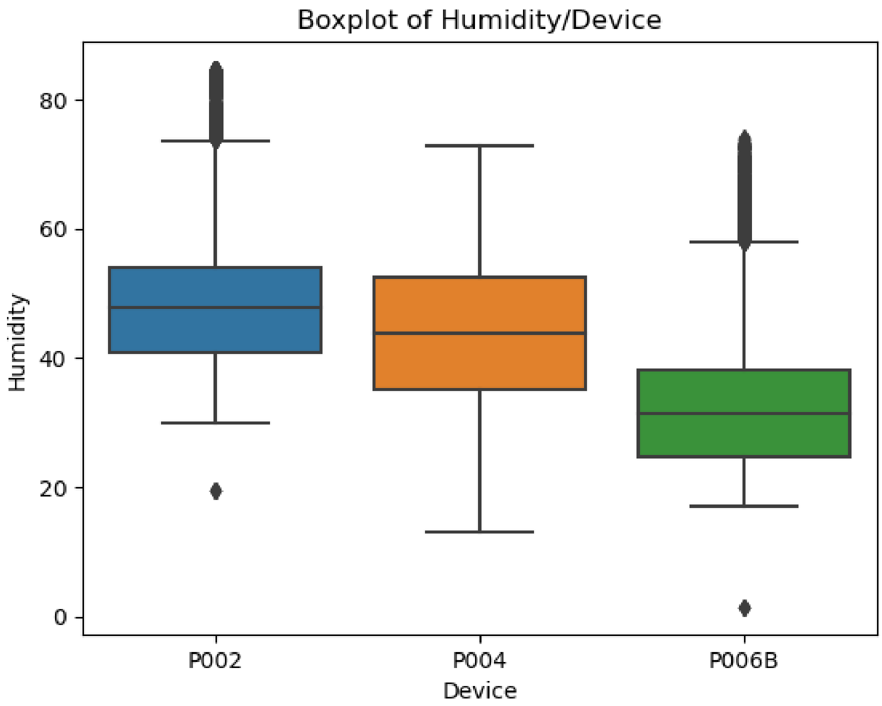


Figure 4: Box Chart of Humidity/ Device from dataset: Overall, P002 seems to have the highest peaks in Humidity

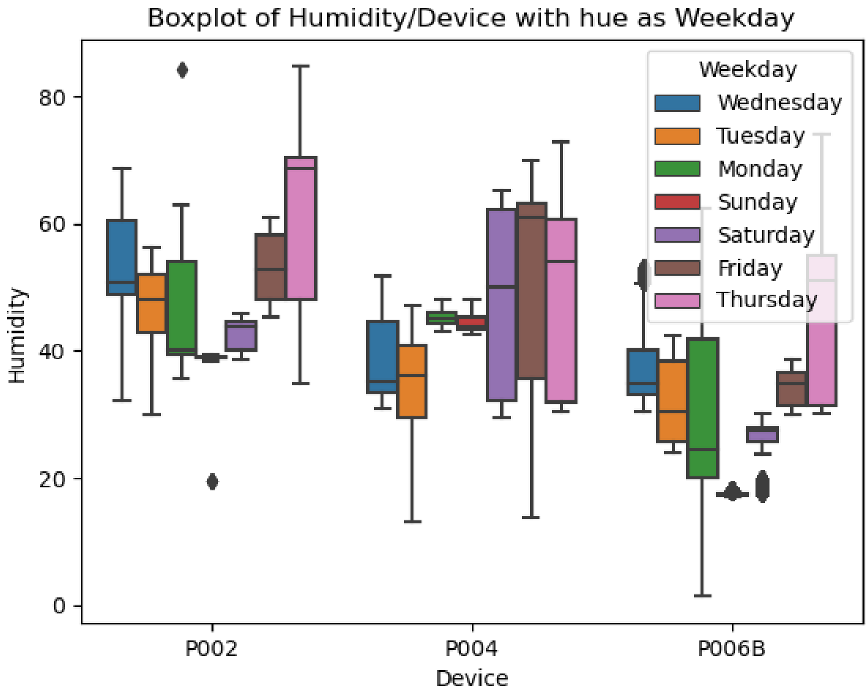


Figure 4: Box Chart of Humidity/Device with Weekday as the hue from dataset: Interesting to see that Thursdays seem to be the highest in humidity for each device

V. SUMMARY OF FINDINGS

I found very interesting bits of information through my EDA, such as humidity being higher on Thursdays versus the rest of the week or viewing such high correlation between Temp and Humidity, even though that was expected. I think the device that will cause me the most trouble, in terms of making predictions, will be P006B, as it is a room with no windows and having only 1 vent. Along with that, Dr. K is the only person that goes into that room, as it is his office. Even though I have a general time frame of when he is usually in his office, the times can vary greatly. Another issue I will run into will be dealing with the outliers for the Temp data. It seems most of the outliers come from P004, so I will have to make adjustments in my model to account for these outliers. In terms of making this project smoother/easier, one thing I wish I would have added to my script on the Raspberry Pi's would be to automatically create an indicator of when classrooms are supposed to be occupied, rather than having this process being done in my analysis code. Overall, I think I am seeing great data with a high indication that I will be able to accurately predict occupancy once the model is up and running.