Data analysis

Maarten Donkersloot

December 8, 2021

1 Introduction

To better understand our data and use it in meaningful ways we did a data analysis on the most important columnsw.

2 Data gathering

Our data was sourced from an already existing study, with our own generate acceptability columns using the ASHREA 55 standard.

3 Data explanation

- original-entry-id: Data is concatenated from multiple days/rooms, these are the original entry id's from those sources.
- node-id: At what place in the room the data was collected from
- room: Which room was used to take the reading
- datetime: datetime of reading
- relative-time: The time between this and the previous reading
- date: Date of reading
- temperature: Temperature in Celsius
- mean-temp-day: The mean temperature from the day of the reading in Mannheim Germany (place where the study took place)
- heatindex: What the temperature feels like
- relative-humidity: Percentage of humidity in the air
- light-sensor-one-wavelength: Wavelength in nm at reading time
- light-sensor-two-wavelength: Wavelength in nm at reading time
- number-occupants: Amount of people in the room at the time of the reading
- activity-occupants: Activity the occupants were doing at the time off the reading (0 = n/a, 1 = read, 2 = stand, 3 = walk, 4 = work)
- door-state: Was the door open or closed
- window-state: was the window open or closed
- tmp-cmf: Comfort temperature a that specific running mean temperature, in °C

- tmp-cmf-80-low: Lower acceptable comfort temperature for 80% occupants, in ˚C
- tmp-cmf-80-up: Upper acceptable comfort temperature for 80% occupants, in ˚C
- tmp-cmf-90-low: Lower acceptable comfort temperature for 90% occupants, in ˚C
- tmp-cmf-90-up: Upper acceptable comfort temperature for 90% occupants, in ˚C
- acceptability-80: Acceptability for 80% occupants
- acceptability-90: Acceptability for 90% occupants

4 Univariate analysis

First we wanted to do a univariate analysis to get a better understanding of the individual features.

4.1 Feature: Room

We found that the room feature is positively skewed towards A making room C less prominent in the dataset.

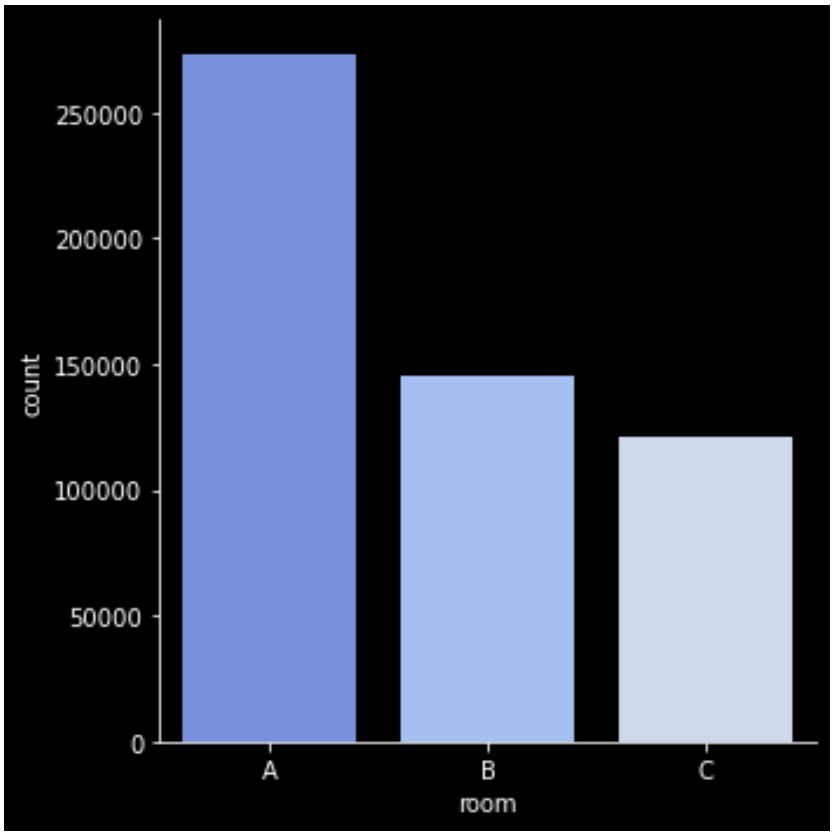

Figure 1: Record count per room

4.2 Feature: Date

We found that the Date feature is positively skewed towards 2016. 2017 makes up 23 percent off all data. Including 2017 may lead to potential overfitting.

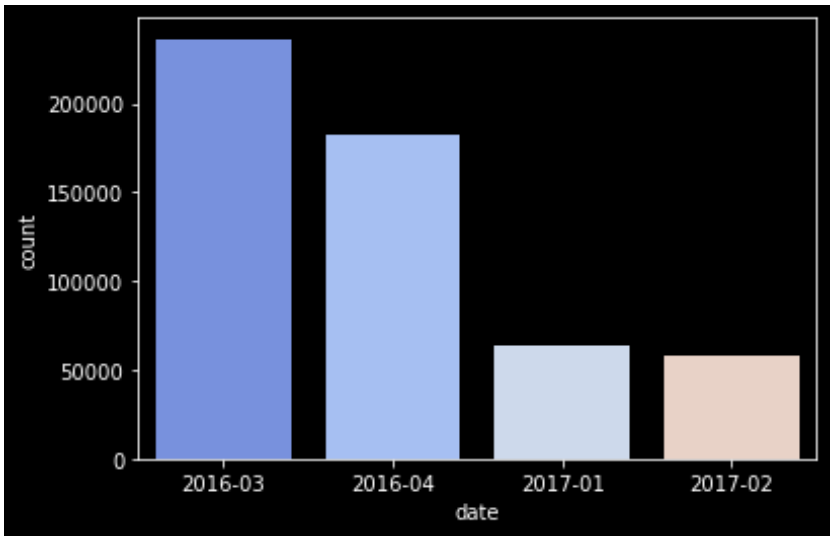

Figure 2: Record count per month

4.3 Feature: Temperature

We found that the Temperature feature has overfitting towards the upper end of the dataset, we don't know where this comes from yet and will investigate this further in the multivariate analysis.

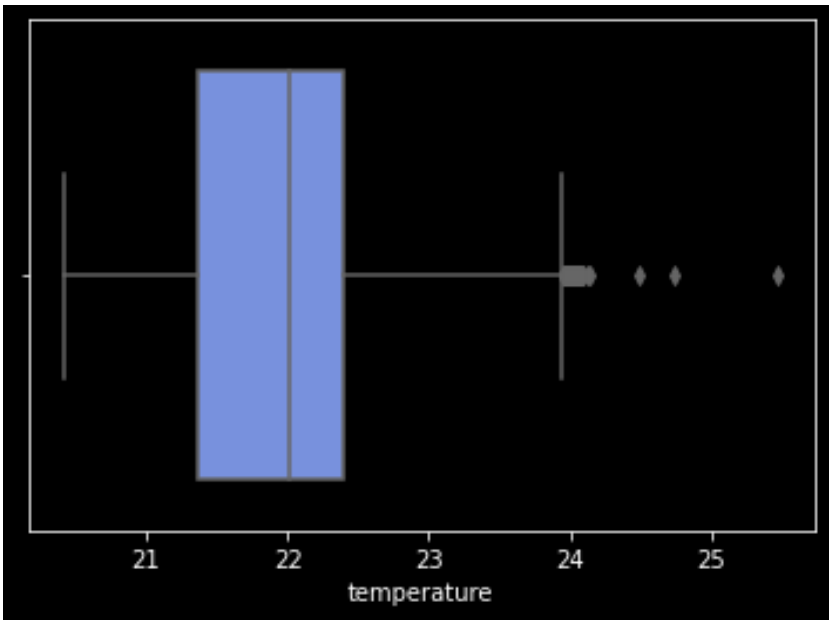

Figure 3: Temperature boxplot

4.4 Feature: Activity occupants

We found that the Tactivity occupants feature is heavily skewed towards no activities, putting all activities together gives us a better distribution but we'll have to decide if we keep it splitted, put it together or remove it entirely.

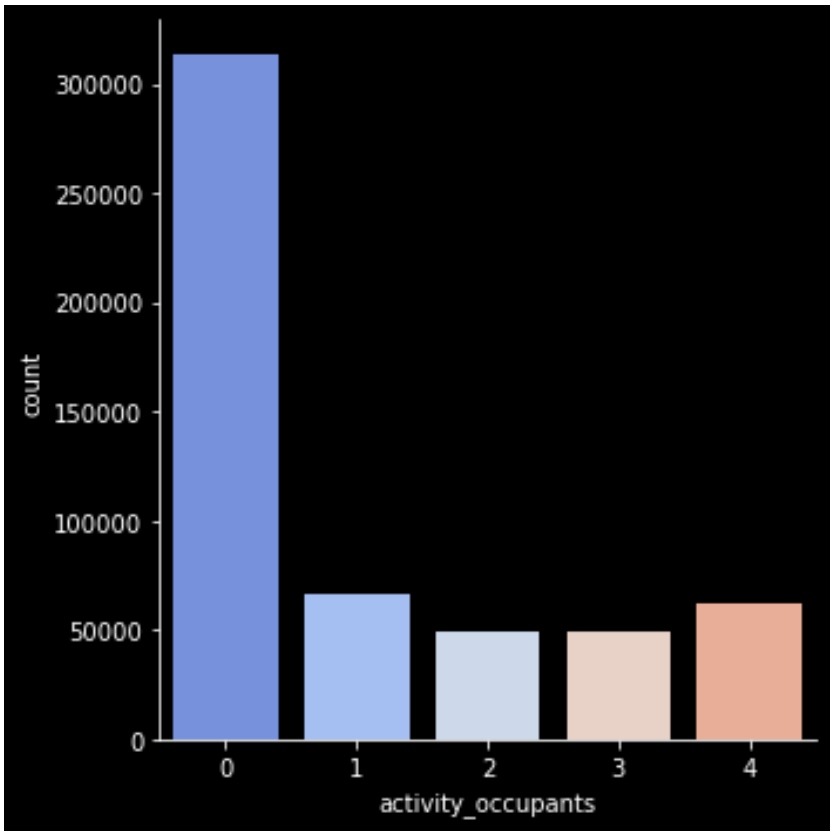

Figure 4: Record count per activity

4.5 Feature: Door state

We found that the door state feature is heavily skewed towards the closed state. If this leads to outliers must be investigated.

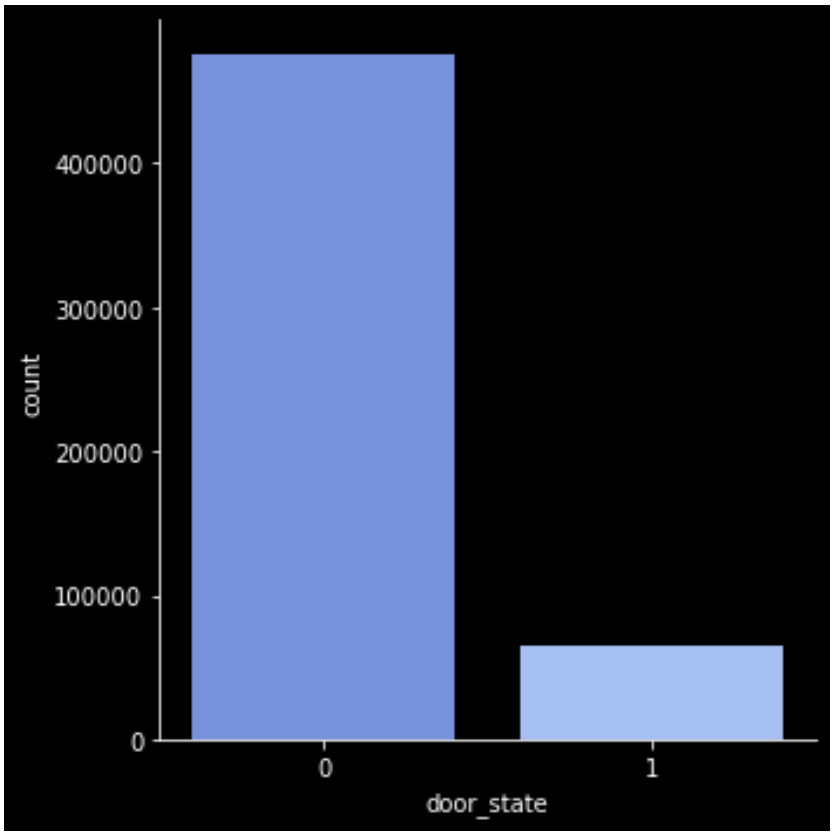

Figure 5: Door state count per state

4.6 Feature: Window state

We found that the window state is completely full off the closed state, this means it will be useless for the prediction.

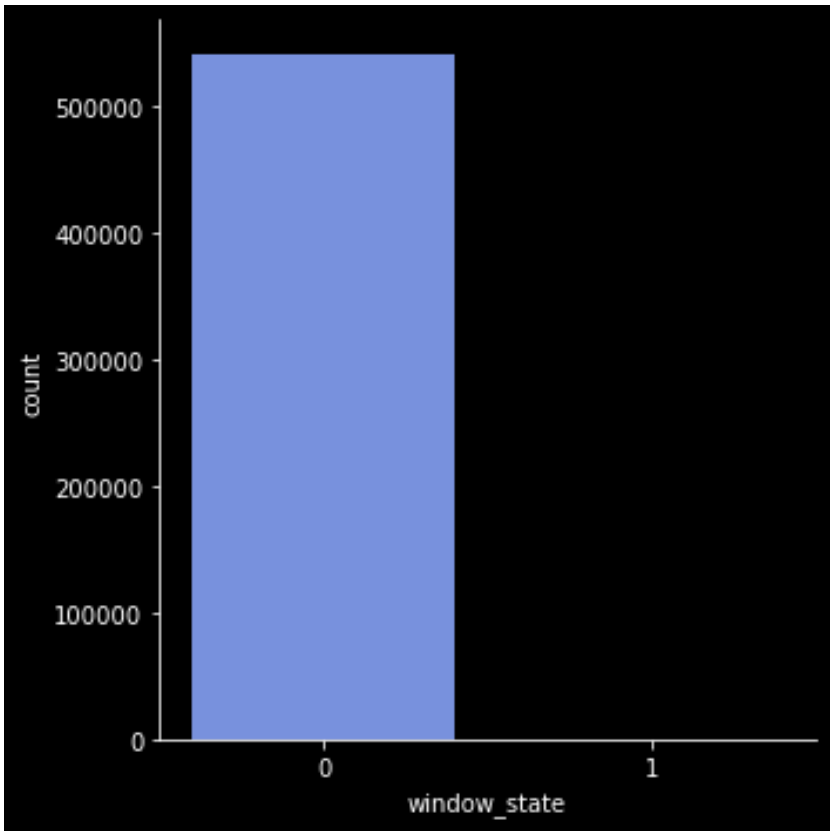

Figure 6: Window state count per state

5 Multivariate analysis

With the discoveries from the univariate analysis we went forward with the multivariate analysis.

5.1 Feature: Room

We found that when we plot temperature and room together there are a bunch of outliers in room 1. These will have to be removed at a later moment.

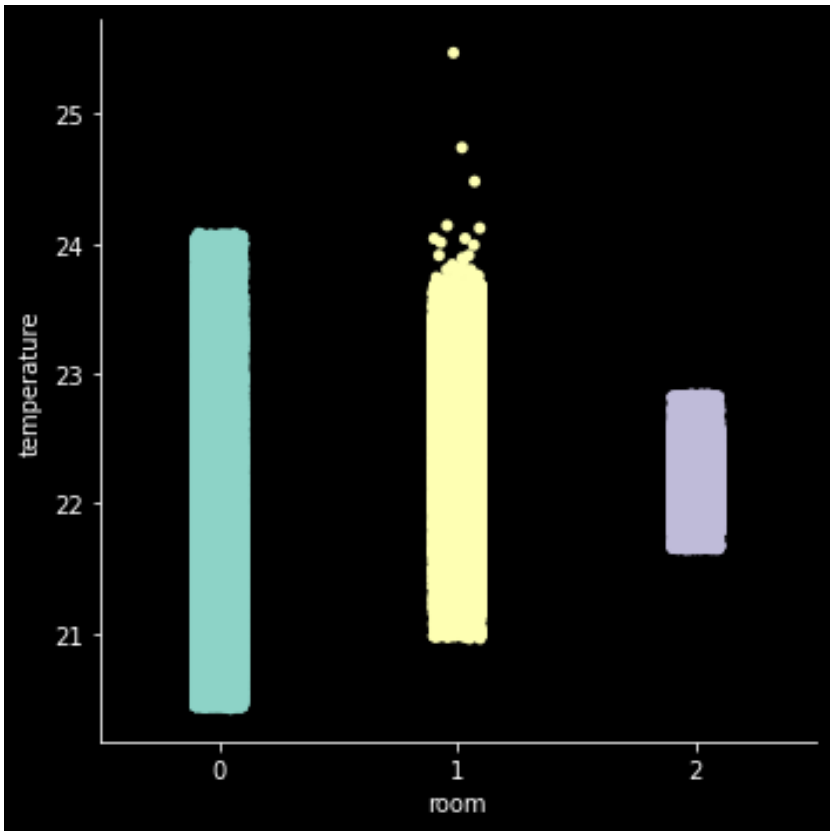

Figure 7: Record count per room

Next we plotted room on hour to see the distribution in records per room per hour.

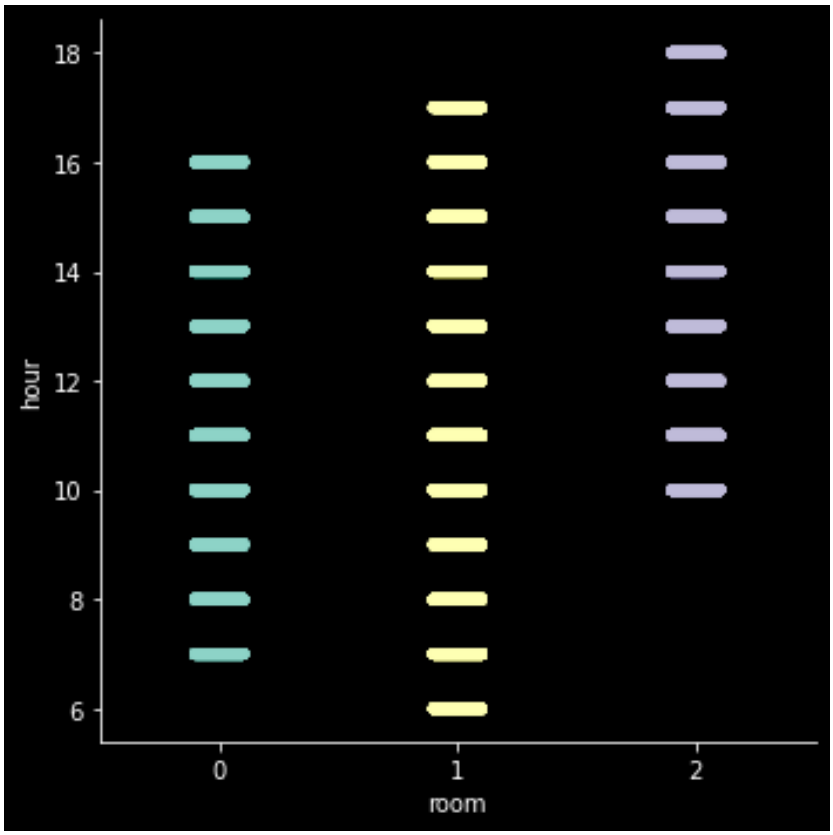

Figure 8: Record count per room

After that we also plotted the room and the mean-temp-day to see which records are from what year and what the mean temperature was that day. From this we can see a big difference in 2017 compared to 2016.

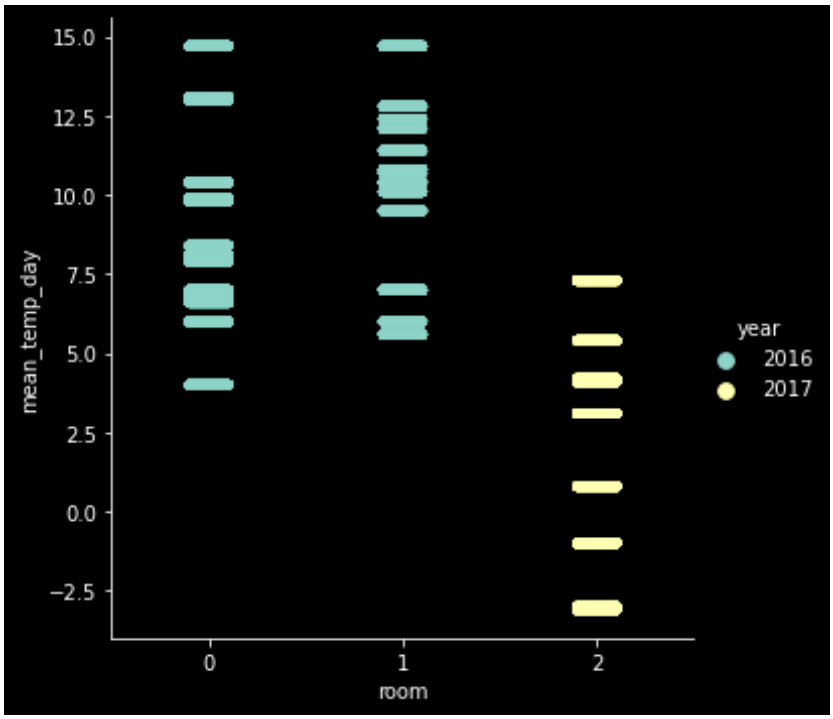

Figure 9: Record count per room

To confirm our suspicions about room 2 only having data from 2017 we plot the record count per year with hue being the room. From this we find that indeed room 2 only has data from 2017.

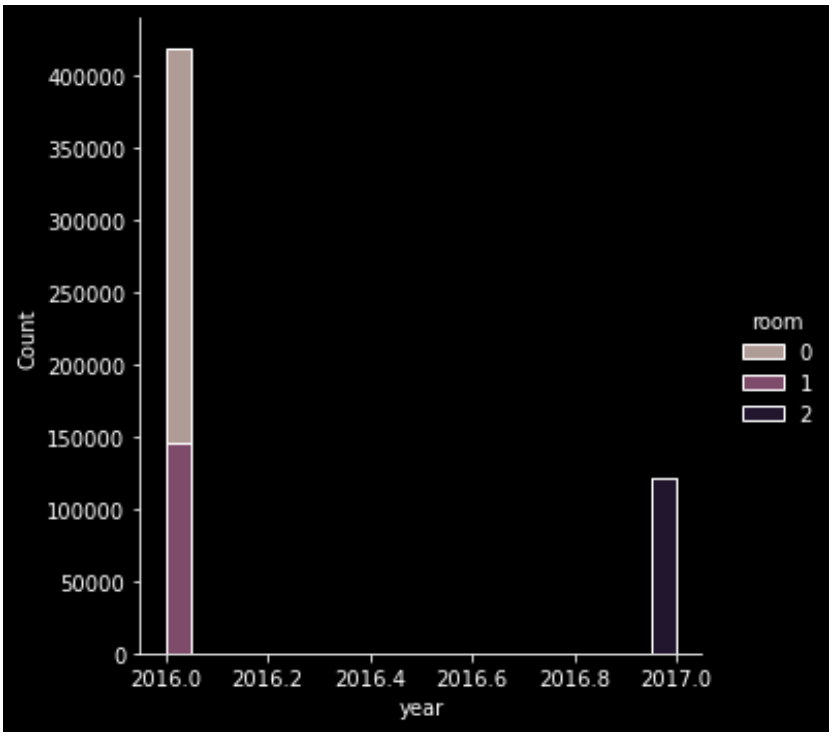

Figure 10: Record count per room

5.2 Feature: Temperature

We plotted temperature and meantemp along with their bloxplots to see where the outliers lay. We found that most outliers lay below 0 degrees meantemp and above 24 degrees temperature.

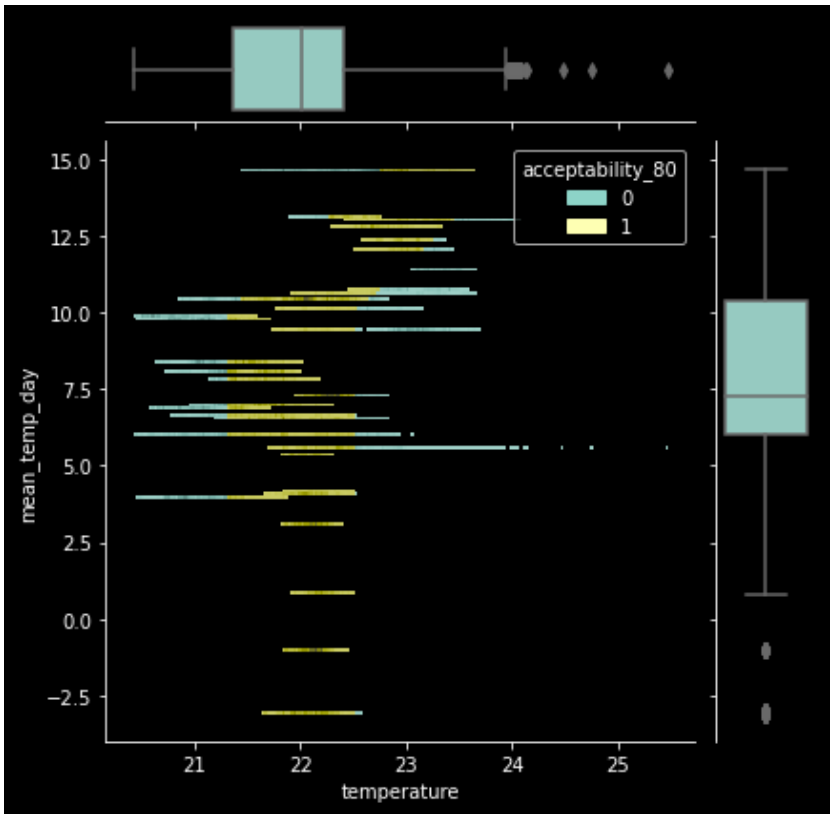

Figure 11: Record count per room

We wanted to plot temperature and heatindex to see if we have any outlying data. And indeed we have outlier data on the top end of heatindex and temperature.

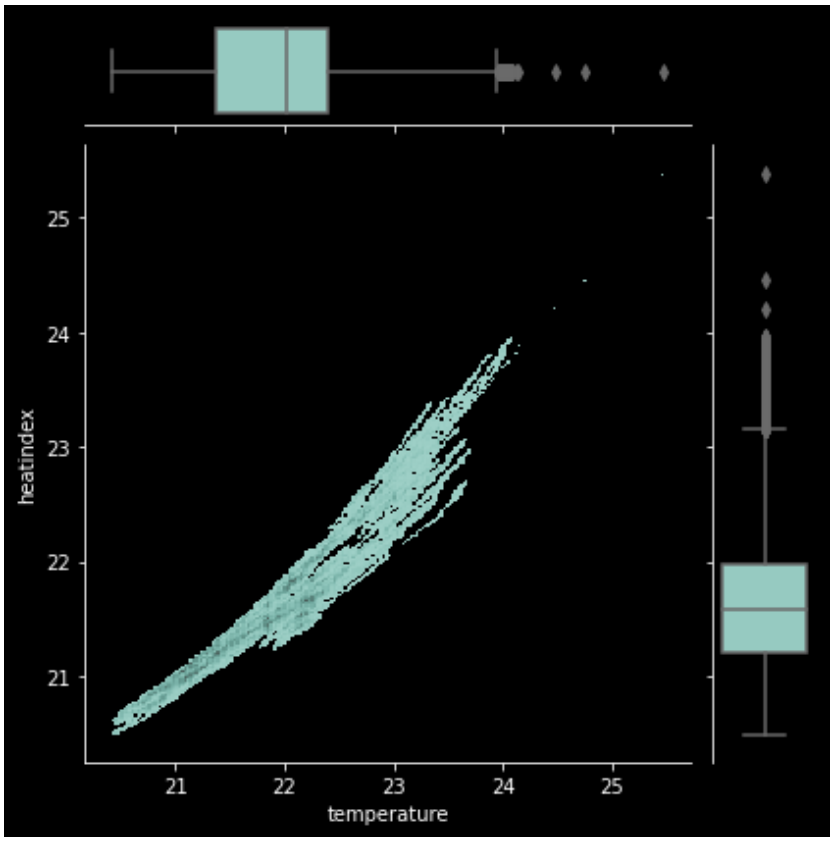

Figure 12: Record count per room

Too further confirm our suspicions on the temperature outliers we plot the temperature in a boxplot with year as hue. Furthermore we plotted that same data per room with the hue as year.

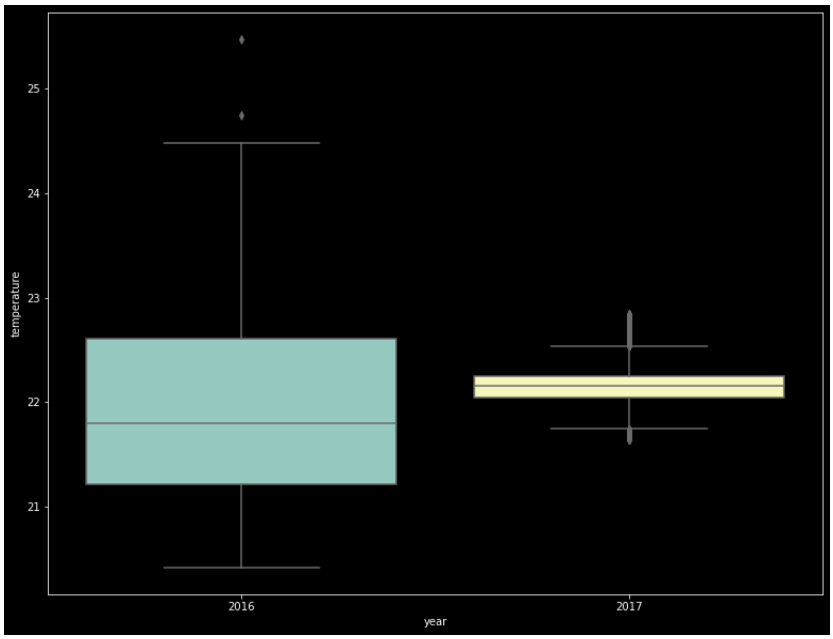

Figure 13: Record count per room

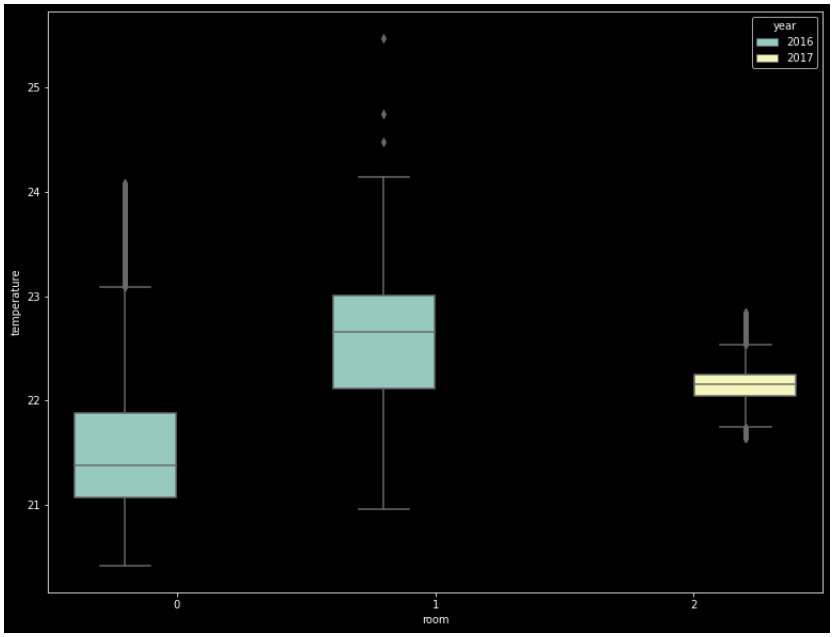

Figure 14: Record count per room