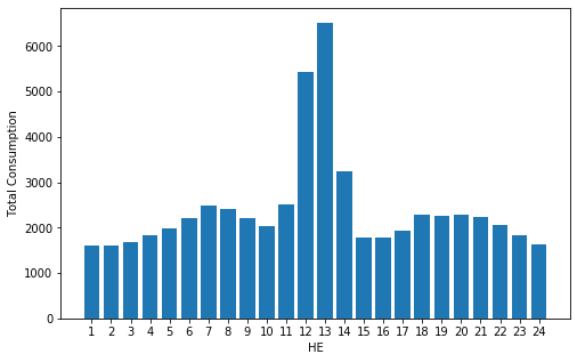
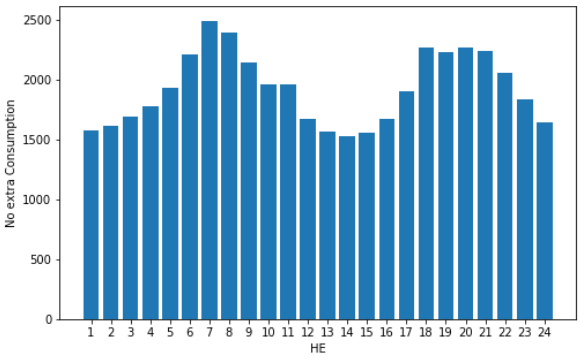
Assignment 2

Since the data in the first file is already hourly, I first convert the second file into hourly data. For each timestamp, I extract the date and hour, by which the dataset is grouped. I then sum up the consumption and divide by 1000 to convert W into kW, so that the unit is matched with the first file. Since the first file has data on every hour of a year, I left join the newly formatted second dataset onto the first dataset to merge these two files. I assume the timestamps in the first file represents hour ending, so 11:06 in the second file should match 12:00 (HE12) in the first file. To get the total consumption, simply fill NAs with 0 and sum up all the columns.

I find it very challenging to not to hardcode column names and numbers. The dfply package has similar syntax to SQL, and in SQL one almost always have to hardcode column names because the manipulation of data is purely situational. If the timestamp comes in a different format, we have to parse it through with a different approach. The dataset can have more columns that pose new problems. Unless a new dataset is come in the exact same format, I cannot think of a good way to generalize the formatting process, especially with SQL-like language.

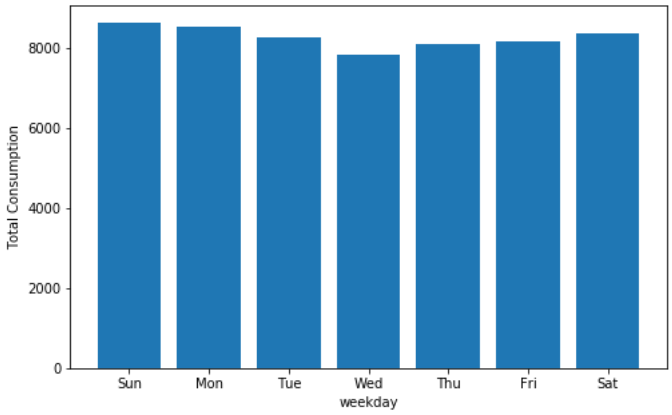
The graph on the left is the total consumption by hours. The graph is actually surprising since it shows more consumptions in the middle of the day. One would expect consumption to be lowest during that period since most people are out working. The graph on the right shows the total consumption by hours excluding the extra usage in the second file. The behavior is anticipated, since more electricity is consumed from 7am to 9am and from 6pm to 10pm, when people are either getting ready to leave for work or staying at home after work.





The figure below is the total consumption by weekdays. The pattern is somewhat expected, since people are home more often on weekends and will consume more power. I actually

expect to see larger discrepancies between workdays and weekends, but this chart is also reasonable.



The figures below are the total consumption with or without the extra data from the second file by months. The one on the left is the total consumption with the extra data, and the pattern is expected. People will use more power in winter and summer for AC, but interestingly, the consumption is low in August when it is still arguably very hot. The graph on the right side is surprising; without the extra data, the consumption is very low in summer months. AC is a big part of electricity consumption, not to mention that kids are usually home for summer break. Maybe the extra data represents the electricity used by AC's cooling function.