Q2 Report

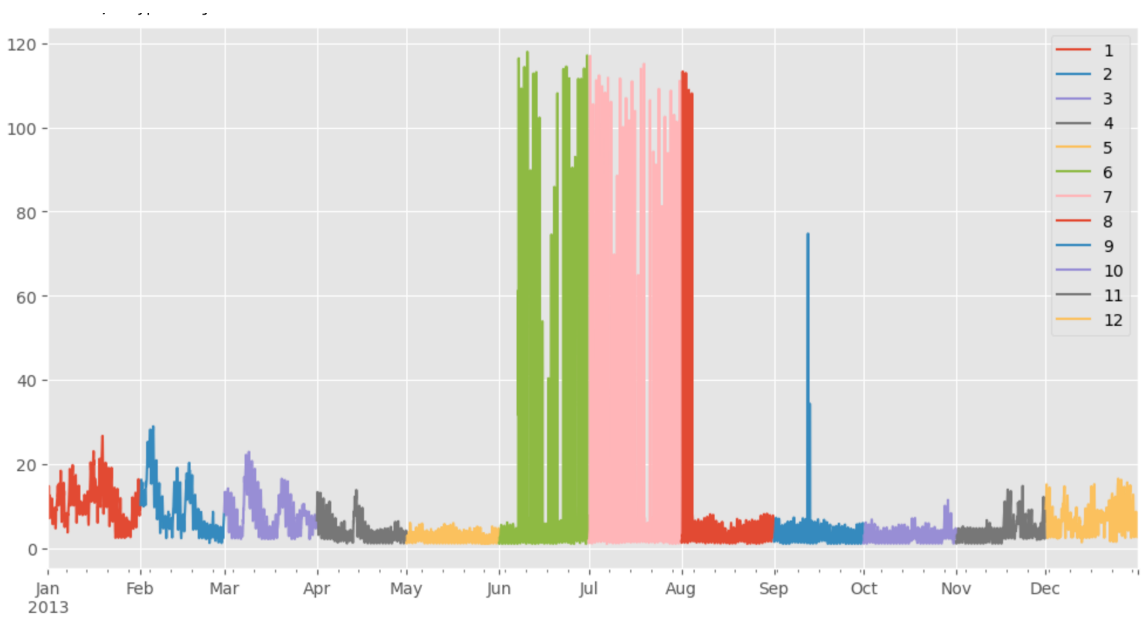
For this assignment, I run into some troubles with the DateTime format, which took a lot of effort to straighten out. We have two data files here, **resident consumption** and **one appliance consumption.**

The first thing I noticed was that the DateTime format in **resident consumption** and **one appliance consumption is** different, so we need to unify the DateTime in order to merge the two datasets. I first changed the format in the resident dataset and subtracted 1 from all hours, now that 1:00 means the consumption between 1:00-2:00 (previously, 1:00 represented the consumption between 0:00 and 1:00). This also changed all 24:00 to 23:00, which removed the ambiguity of which day the consumption belongs to.

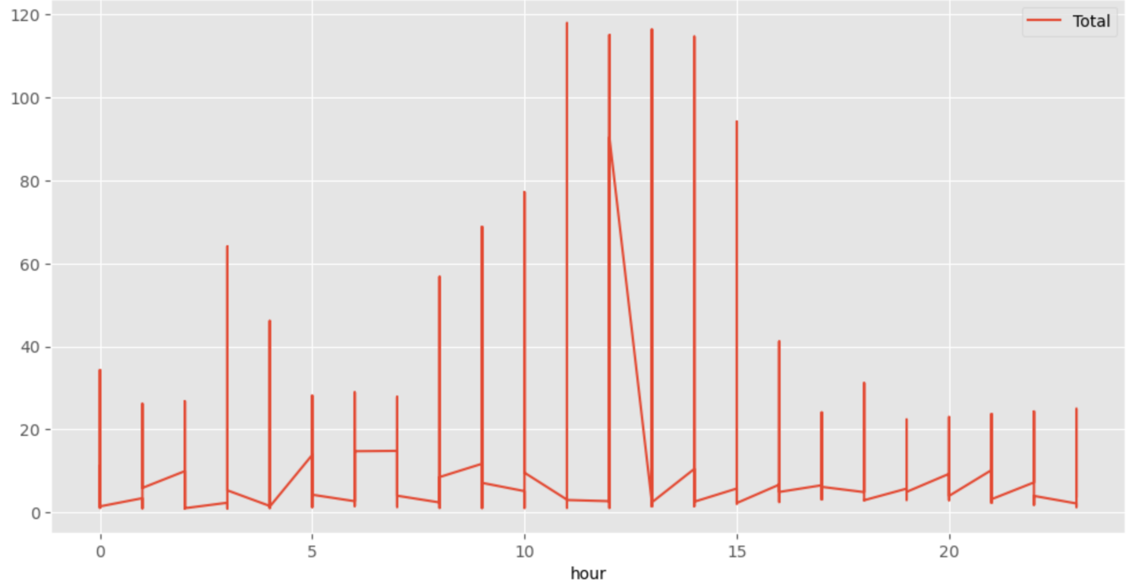
Next, I calculated the difference in days between the first date in the **resident dataset (**1900-01-01, 00:00:00) and (2013-01-01, 00:00:00) and add that difference in days to unify the year to 2013. The reason is that when you only give the dates without the year to the datetime, they assume it is 1900. Here I kept the year because, without the year in the date, it is difficult to determine which day of the week it is, and I think the day of the week can be an important piece of information. The reason is that the total electricity consumption can be highly correlated with the days of the week.

Then I processed the datetime for the **one appliance consumption.** Here the most important thing is to remove all the minute-level and second-level data so that 11:04:00 become 11. By doing this, we unify this format with the resident consumption data. Then we can use the pandas **groupby** function to group all the same hours and sum up the total usage by each hour. Note that the usage here is in Watts, so we need to divide by 1000 to get the total usage in Kw.

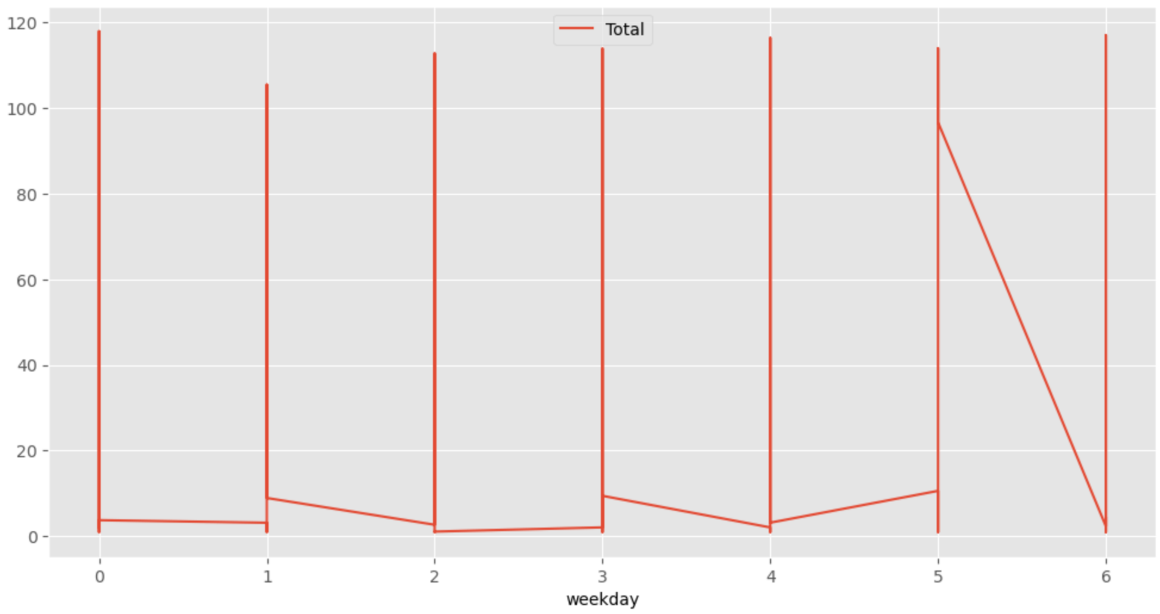
Now that the formats and date index are unified, the pandas **join** function can help us merge the two dataframes based on the index. One thing to note here is we need to fill all the "Nan" values and make them 0 to avoid errors. Now we can again sum up all the usage by every hour and plot the usage with respect to hour/month/weekday:



From the month graph we can observe that the highest usage of electricity is in the summer, specifically from early June to early august, due to the high usage of the one appliance. What's abnormal for me is there is no data for any usage of the on appliance between early August and mid-September, which caused the gap in peaks. Also, even within the month of June and July there are some small down times due to missing data. In mid-September we observe another short peak of the usage.

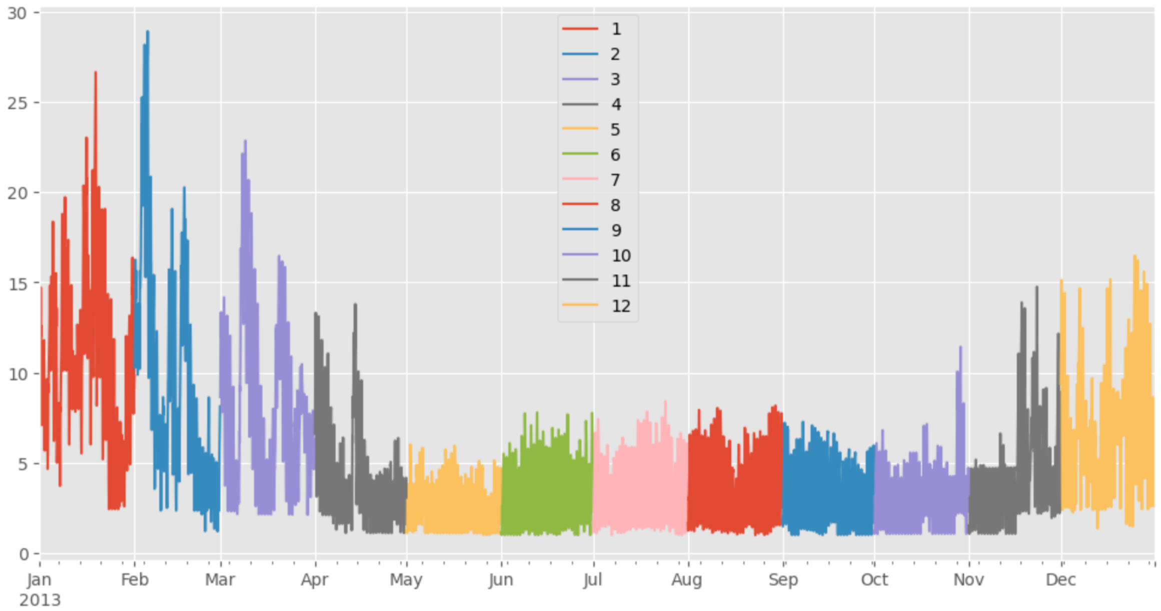


From the hourly usage graph we can observe that most of the usage are concentrated between 8:00 and 15:00 during the day. Which corresponding the general working/business hours of a business day.

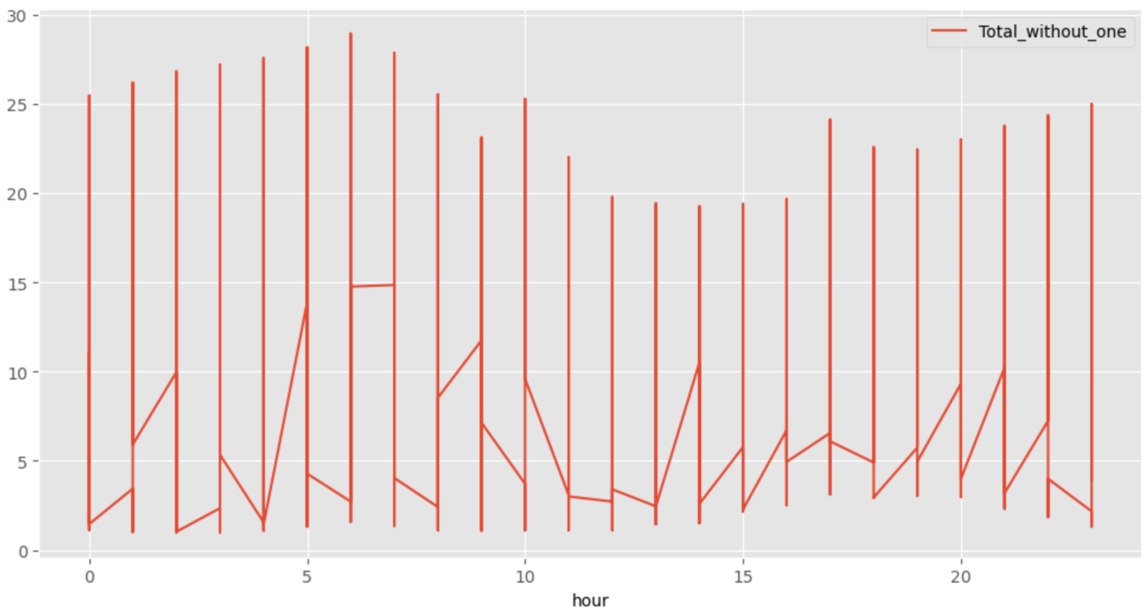


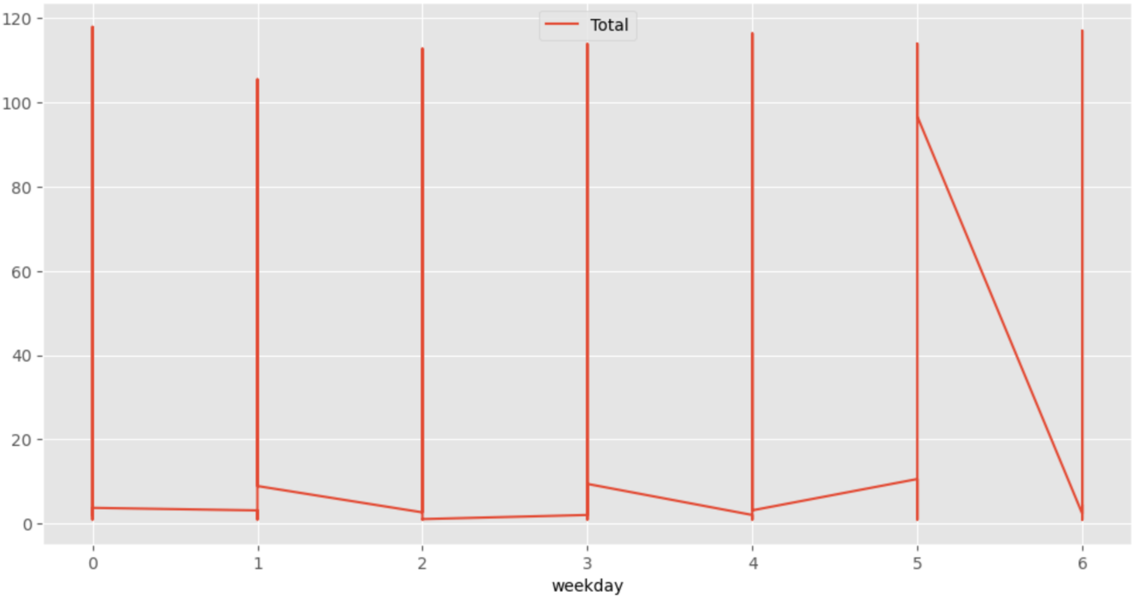
From the weekday graph there does not appear to be a significant correlation between weekdays and power usage.

Additionally, I have also plotted the total usage of all appliances **without the one appliance, which shows:**



From the monthly graph, it appears the peak usage of electricity, without the one appliance us concentrated in the end of the year and beginning of the year (mid-November to Mid-April), and relatively low in other months.





From the hourly and weekday usage graphs, there is no significant pattern in high and low times correlated with hours or weekdays.