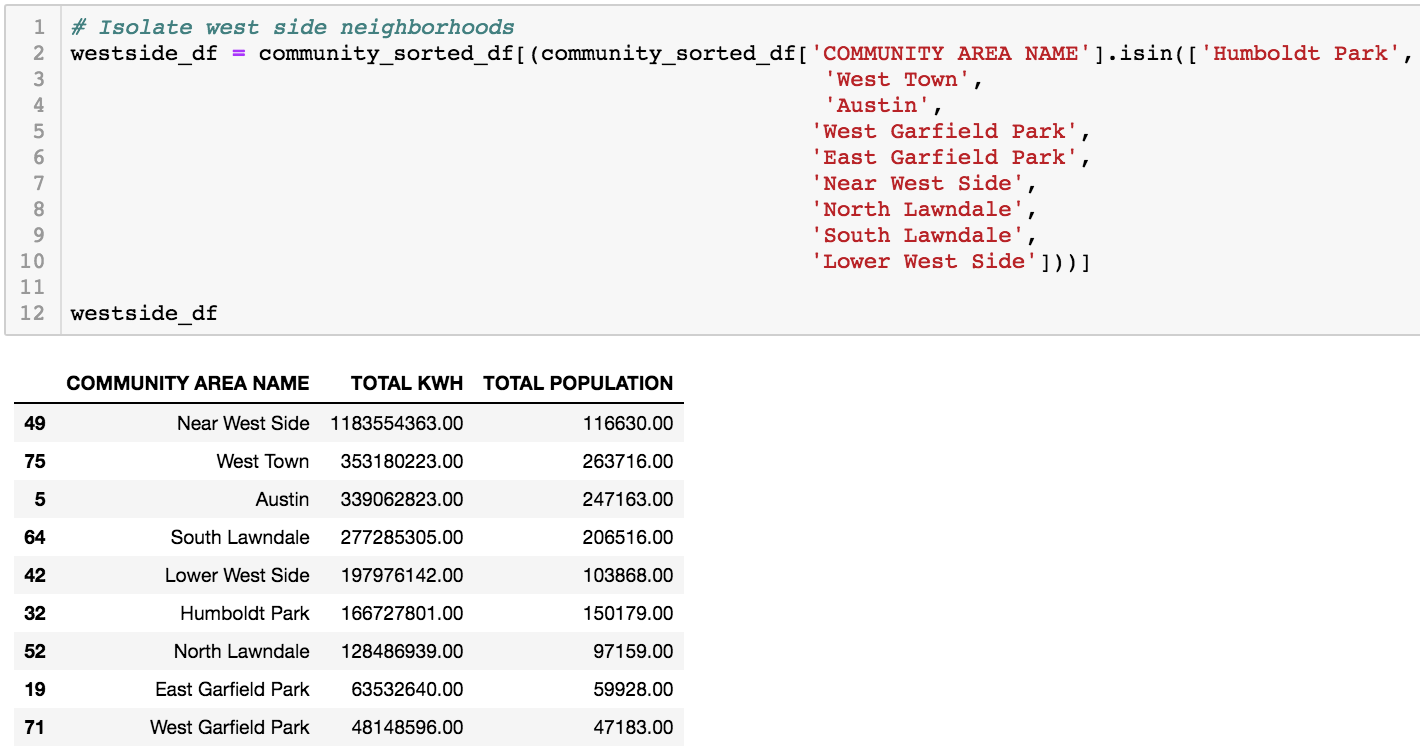
**Project 1**

**2010 Chicago Energy Insights:**

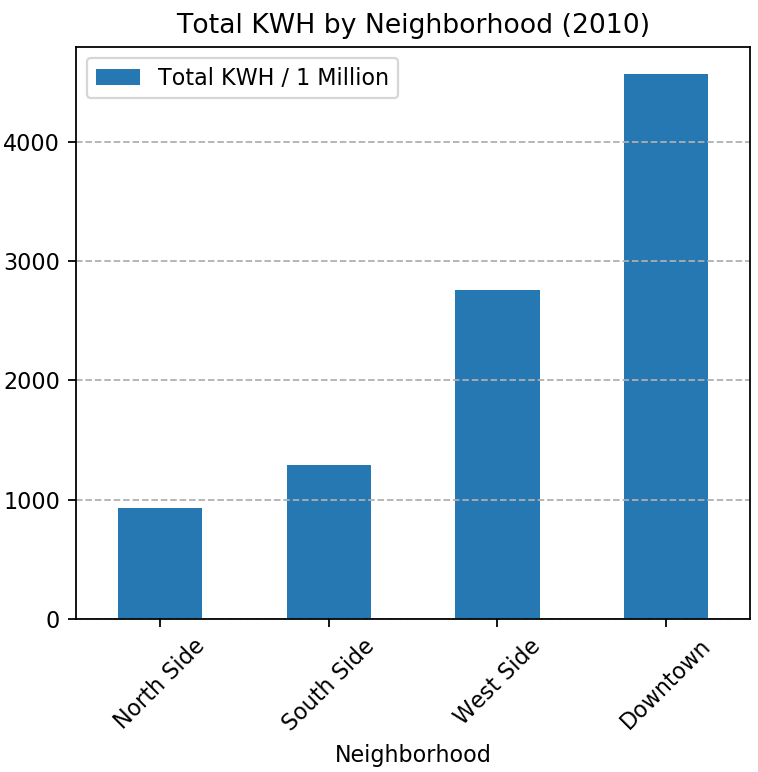
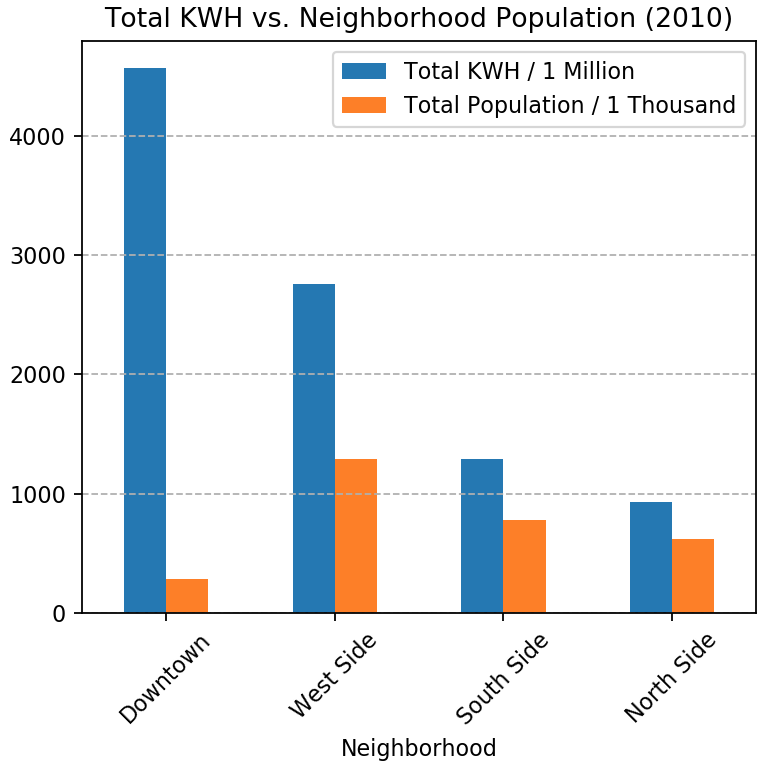
Arthur Velez, Olivia Martin, Victor Birza, Shabreen Khan

**1.How does energy usage differ based on neighborhood?**

When attempting to look at this data based on different neighborhood, it was first necessary to figure out exactly how the city of Chicago categorizes its wide range of neighborhoods. From the csv we pulled from Chicago data portal and information provided by the city I was able to see that that Chicago is officially divided into 77 different “Community Areas” and those area are grouped more generally by “sides”. Approaching this data set I decided it would be pertinent to set me scope around the four major areas of the city’s center. In doing this I initially ended up creating four different data frames (see West Side example below).



From these data frames I found the sum for energy usage and population, and went ahead with charting and analysis -- first by KWH and then a multiplot including population.

Interestingly, though the Downtown “side” has the lowest population of the four areas, it uses more energy than the South Side and the West Side combined. My cleaning and analysis offers a first pass at this data set, and sets up our presentation to be furthered by analysing by economic, design and seasonal factors.

**2. How does energy usage differ based on income level?**

The comparison of energy data, electricity and gas consumption, and income data includes only energy readings for residential building types. The electricity data of interest was KWH MEAN 2010, mean kilowatt hours in 2010. The gas data of interest was THERM MEAN 2010, mean therms in 2010. The respective energy DataFrames and per capita income (PCI) DataFrame were merged along Community Area Names. The PCI information came from the Public Health Statistics data set (1). The next step was to group each energy and income DataFrame by community area (df.groupby().mean()). The following analysis is based on this community area grouped DataFrame.

The PCI brackets were defined in increments of $10,000 from 0 to $90,000. When grouping the PCI bracket data by mean of KWH MEAN 2010 or THERM MEAN 2010, the data represents the average of the average electricity or gas consumption per community area. Finding the COMMUNITY AREA COUNT was done by grouping by PCI bracket (df.groupby().count()). Similar to the energy mean data, the MEAN TOTAL POPULATION is also an average of the average total population of a residential building per community area. The final DataFrame contains these summaries for each PCI bracket.

The PCI bracket that consumed the most kWH (kilowatt hours) and natural gas (therms) was the 60k to 70k PCI bracket, 811539 kWH and 31730 therms, respectively. The 60k to 70k PCI bracket represents two community areas, Near South Side and Loop. Even compared to the next highest energy consuming PCI bracket, 80k to 90k, the 60k to 70k PCI bracket consumes five times more kWH (811,539 kWH vs 144,505 kWH) and three times as more natural gas (31,730 therms vs 8,935 therms). The population difference may also explain this variance in usage. For the 60k to 70k PCI bracket, the average total population (per residential building) is also five times the average total population of 80k to 90k PCI bracket. These comparisons suggest that both PCI brackets, 60k to 70k and 80k to 90k, consumed on average the same amount of electricity per person, but consumed on average more natural gas per person in 2010.

Also interestingly, comparing the 60k to 70k PCI bracket to the 10k to 20k PCI bracket, the difference in electricity usage is 799,872 kWH and 29,634 therms. Those community areas’ buildings with a PCI of 60k to 70k use an average of 70 times more electricity and 15 times more natural gas than community areas’ buildings with PCI of 10k to 20k. (However, both of these income brackets have similar mean total populations of residential buildings, 60k to 70k: 3317 vs 10k to 20k: 3524.

Looking back, creating a DataFrame that took the mean of energy readings and the sum of total populations per building, when grouping by COMMUNITY AREA NAMES, would provide a better scope of total population. Future interests include comparing the relationships amongst PCI brackets’ energy consumption during each month or season. It would be good to compare the data throughout history too; assuming the community area boundaries have not varied significantly over the years. Also with data from the next census in 2020, new comparisons across potentially changing community areas and the affect on PCI groups would be great to investigate as well.

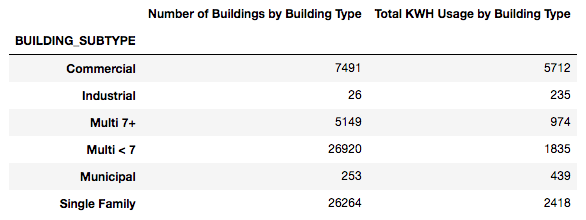
|  |
| --- |
| **Chicago Average Electricity Usage (kWH) 2010 by Per Capita Income Brackets** |
|  |
| **Chicago - Average Natural Gas Usage (Therm) 2010 by Per Capita Income Brackets** |
|  |

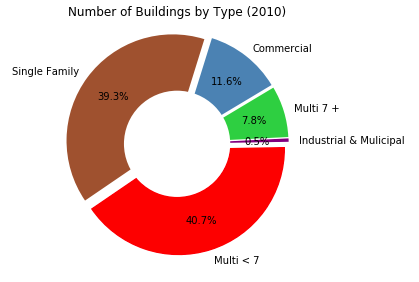
* 1. Illinois Department of Public Health (IDPH) and U.S. Census Bureau. “Public Health Statistics- Selected Public Health Indicators by Chicago Community Area.” Https://Data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-Public-Health-in/Iqnk-2tcu, 18 May 2012.

**3.How does energy usage differ based on building type?**

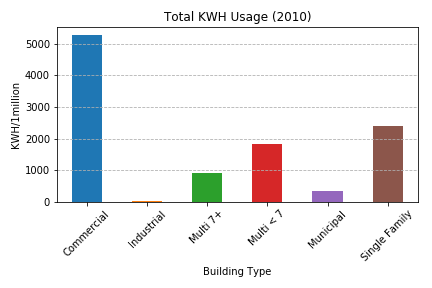
The comparison of electricity and gas usage based on building type was categorized by six different building types. This information was located in the “BUILDING\_SUBTYPE” column. Buildings were classified as commercial, industrial, municipal, multi story over seven units, multi story under 7 units, or single family. Electricity usage was found in the “TOTAL KWH.” This column combined the total kwh usage for every month in 2010. Gas usage was found in the “TOTAL THERMS” column which also combined 2010’s monthly usage.

For the first comparison we look at the number of buildings separated into the six types and the total electricity usage. The dataframe below was used to create the two charts.



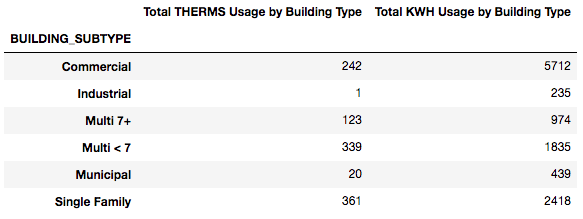
This dataframe show that there was a large difference in the type of buildings in Chicago. Only 26 industrial building were listed. This number was so low compared to the other building types that it caused an issue when the graphs were created. 

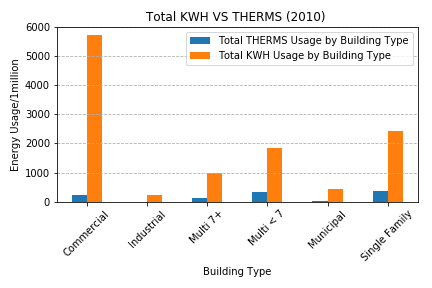
For this reason Industrial and Municipal buildings were grouped in the chart to the right. Even with this grouping we can see that they only make up .5% of the total buildings in Chicago. This graph also shows that the Multi story under 7 (Multi <7) units and single family building were almost evenly divided.



The second graph shows us the total electricity usage based on the building type. We can see that commercial buildings consumed the most electricity in 2010 even though they only accounted for 11.6% of the total buildings in Chicago. This is more than *double* compared with single family buildings. Additionally single family buildings consumed more electricity than multi < 7 building even though the total number of buildings were almost evenly split (first graph).

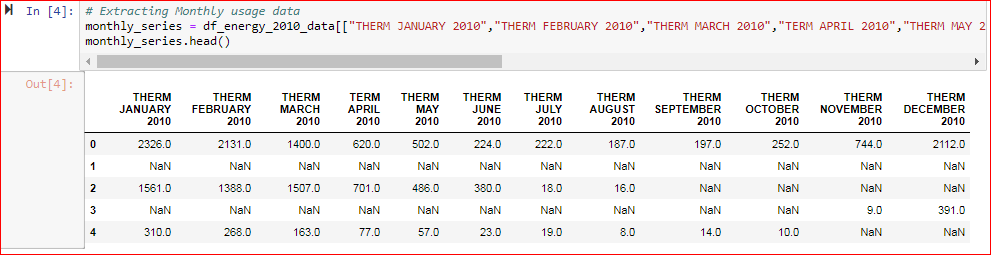
For the second comparison we look at the total electricity usage compared with total gas usage grouped by building type. The numbers illustrated below are divided by 1 million to make the data more readable. As we can see Industrial and Municipal buildings once again have very little gas usage compared with the other building types.

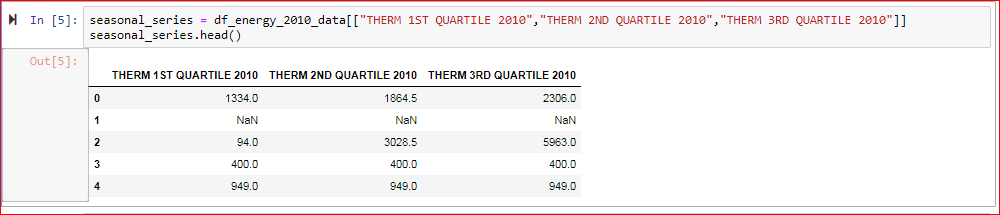


This dataframe was used to create the graph below. This graph shows the total electricity vs gas usage based on building type. We can see that much more electricity was used than gas in 2010. Another observation is that single family buildings used the most amount of gas. However, commercial buildings even though they only account for 11.6% of buildings consume about the same amount of gas as Multi <7 and Single Family buildings. 

How does energy usage differ based on the time of the year?

Electricity usage varies with the weather, as changes in the temperature and humidity affect the demand for heating and cooling. For this part of the analysis I used the “The Therms by Months” columns. For months I assumed that the “Therm per Quartile” columns represent the seasons.



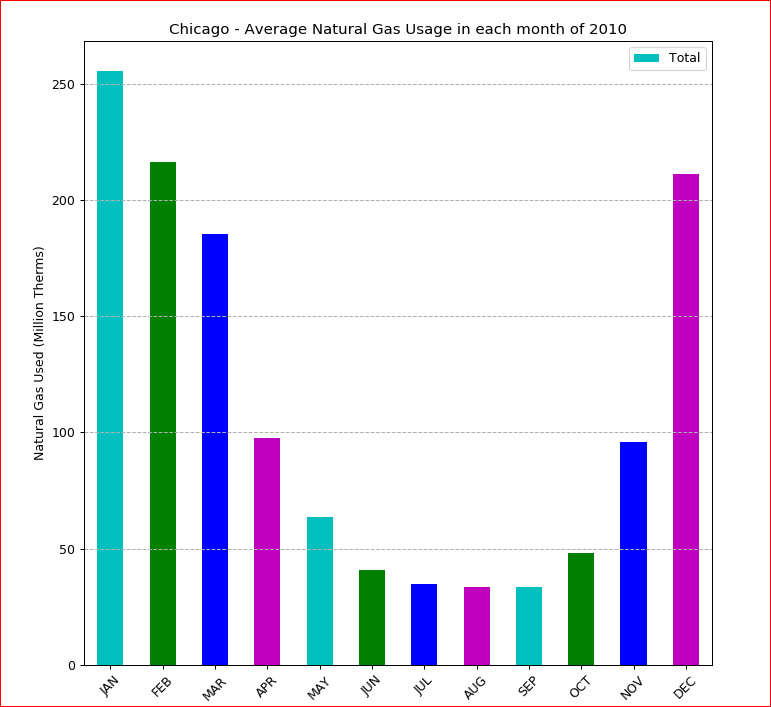


I realized that I will not be able to define the seasons with the data in the “Therm 1st Quartile 2010” columns.

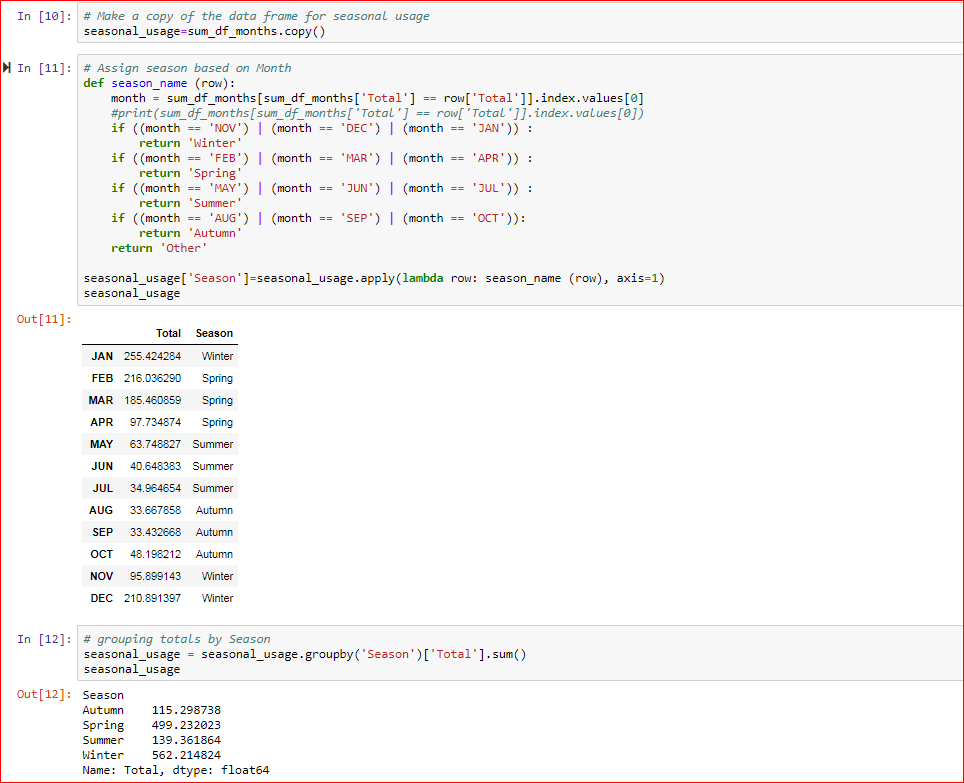
I tried renaming the Months columns to shorter names. Such as Jan, Feb, Mar etc. Transposed the table to make each row a month. Also, divided the sum by a million to show in the graph as a million therm value instead of therms because they are huge in numbers.



Plotted all above in graph



There after I made a copy of data frame for seasonal usage



Then plotted a pie-chart for seasonal usage

