TURN ON THE LIGHTS

Group 2 | Final Project

ABSTRACT

Using methods of Data Science to improve Schneider Electric's estimation of future consumption of electricity consumed across various commercial real estate properties for two of their clients.

Jo Vivian
Andrew Miller

David Molenda,

Michael Morales

Jo Vivian

IST 718 | Big Data Analytics

Contents

Introduction

Schneider Electric's Energy and Sustainability Services (ESS) is an 1,800-employee global team of energy management experts. The team manages an energy spend of more than $30 billion across 500,000 monthly energy invoices at over 430 client sites. As industrial and commercial companies begin to integrate their use and procurement of energy with sustainability goals and initiatives, ESS provides resources for active energy management to help clients realize increased efficiency, financial savings, and more sustainable operations. One of several vertical markets that can benefit from the ESS value proposition is the commercial real estate sector. Commercial real estate has flourished since the establishment of real estate investment trusts (REIT) in 1960.

While real estate investment had been primarily limited to wealthy individuals prior to 1960, the REIT Act allowed for special partnerships to be formed that combined attributes of real estate and stock investment, creating a new approach to wealth creation via passive income. Several refinements later, REITs have become large corporations with total holdings in excess of $3 trillion of gross real estate assets. In large cities such as New York, Los Angeles, and Chicago REITs own large portfolios of office real estate. These holdings are occupied by tenants of various types along with data centers, cafes, gyms, and sometimes even small retail shops. REITs need to know how to bill their tenants and how to forecast cost and usage by occupancy. Of the many services offered by ESS, intelligent budget forecasts are some of the most important.

While market-driven changes can be tracked via commodity markets and public utility knowledge, occupancy is more difficult to pin down. As efficiency increases over time with building automation, on-site generation, and other similar measures, occupancy may move the usage needle in different ways. Historically, it has been generally accepted that for every percent increase in occupancy, there is half a percent increase in electricity usage regardless of region, weather, or other external forces. This may still be the correct ratio, but it is important to analyze the data as regularly as possible to ensure that efficiencies are considered as they evolve. This is the goal of this analysis.

Analysis & Models

About the data

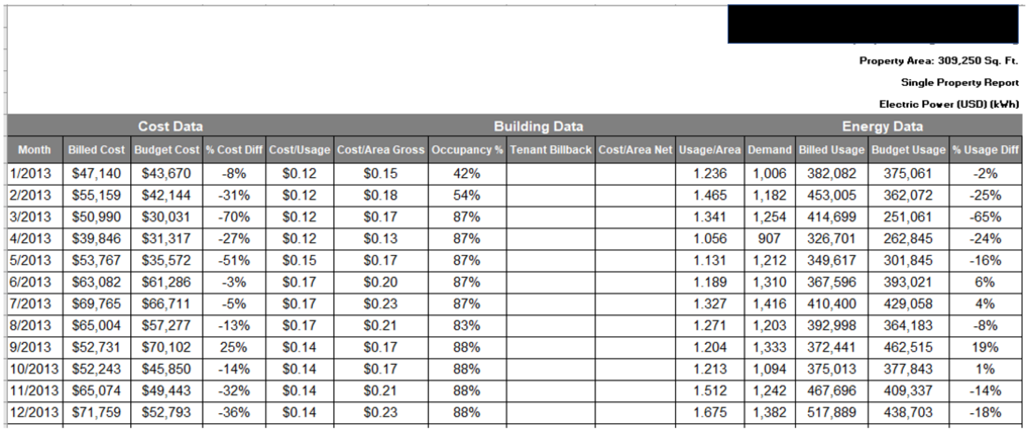
The data used in this analysis is cost (US dollars), usage (kWh), demand (kw), area (square feet), and percent occupied data from 161 commercial real estate buildings across the United States collected over more than 6 years.

Data consists of information from two key clients of Schneider Electric. There were 104 separate files used for client A (Copt) and 57 for client B (Piedmont).

There were properties in 64 different zip codes.

The files for both clients were similar in output. The data consisted of the following:

- Information relating to location (addresses, zip codes, and property size in square feet)
- Occupancy rate by month
- Cost/Usage by month
- Cost/Area by month
- Usage/Area by month
- Energy data demand by month
- Billed usage by month
- Budget used by month
- % Difference of budget vs. billed usage by month



[redacted]

Property Area: 309,250 Sq. Ft.

Single Property Report

Electric Power (USD) (kWh)

Cost Data					Building Data					Energy Data			
Month	Billed Cost	Budget Cost	% Cost Diff	Cost/Usage	Cost/Area Gross	Occupancy %	Tenant Billback	Cost/Area Net	Usage/Area	Demand	Billed Usage	Budget Usage	% Usage Diff
1/2013	$47,140	$43,670	-8%	$0.12	$0.15	42%			1.236	1,006	382,082	375,061	-2%
2/2013	$55,159	$42,144	-31%	$0.12	$0.18	54%			1.465	1,182	453,005	362,072	-25%
3/2013	$50,990	$30,031	-70%	$0.12	$0.17	87%			1.341	1,254	414,699	251,061	-65%
4/2013	$39,846	$31,317	-27%	$0.12	$0.13	87%			1.056	907	326,701	262,845	-24%
5/2013	$53,767	$35,572	-51%	$0.15	$0.17	87%			1.131	1,212	349,617	301,845	-16%
6/2013	$63,082	$61,286	-3%	$0.17	$0.20	87%			1.189	1,310	367,596	393,021	6%
7/2013	$69,765	$66,711	-5%	$0.17	$0.23	87%			1.327	1,416	410,400	429,058	4%
8/2013	$65,004	$57,277	-13%	$0.17	$0.21	83%			1.271	1,203	392,998	364,183	-8%
9/2013	$52,731	$70,102	25%	$0.14	$0.17	88%			1.204	1,333	372,441	462,515	19%
10/2013	$52,243	$45,850	-14%	$0.14	$0.17	88%			1.213	1,094	375,013	377,843	1%
11/2013	$65,074	$49,443	-32%	$0.14	$0.21	88%			1.512	1,242	467,696	409,337	-14%
12/2013	$71,759	$52,793	-36%	$0.14	$0.23	88%			1.675	1,382	517,889	438,703	-18%

Cleaning the data

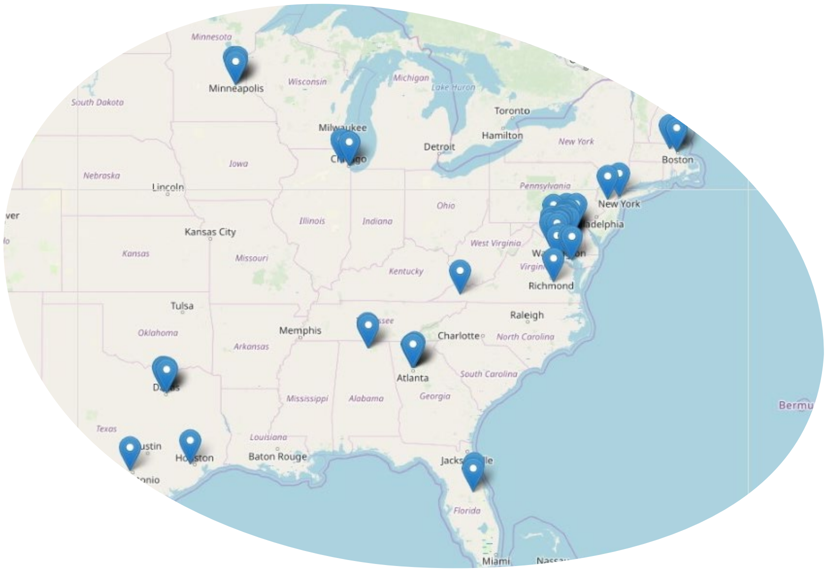
Steps required for cleaning the data included:

- Merging the files for each client
- Separated out the occupancy, zip codes, square footage, and energy data from the property data.
- For client A, address data includes a "Site Reference No" that is blended into the "Property Address", whereas the file for client B file does not have any data related to this. Thus, the column was split into two.

Explore the data

The data represented information from two clients using electricity from ESS. It offered a similar information and structure for both organizations. The only differences were that Client A included a datapoint called “Site Reference No" that was blended into the specific field "Property Address" and Client B had a separate breakout of the Energy Star information.

There was a total of 64 unique zip codes across the two files.



Client A:

There are 94 properties that have billing information (there are 210 properties in the zip code section, 272 in the SQFT section, and 149 in the occupancy section)

The Occupancy tab has 9,672 rows of data and the properties tab has 7,941

Data came from 104 separate files

Client B:

This file was fairly clean and contained 90 addresses.

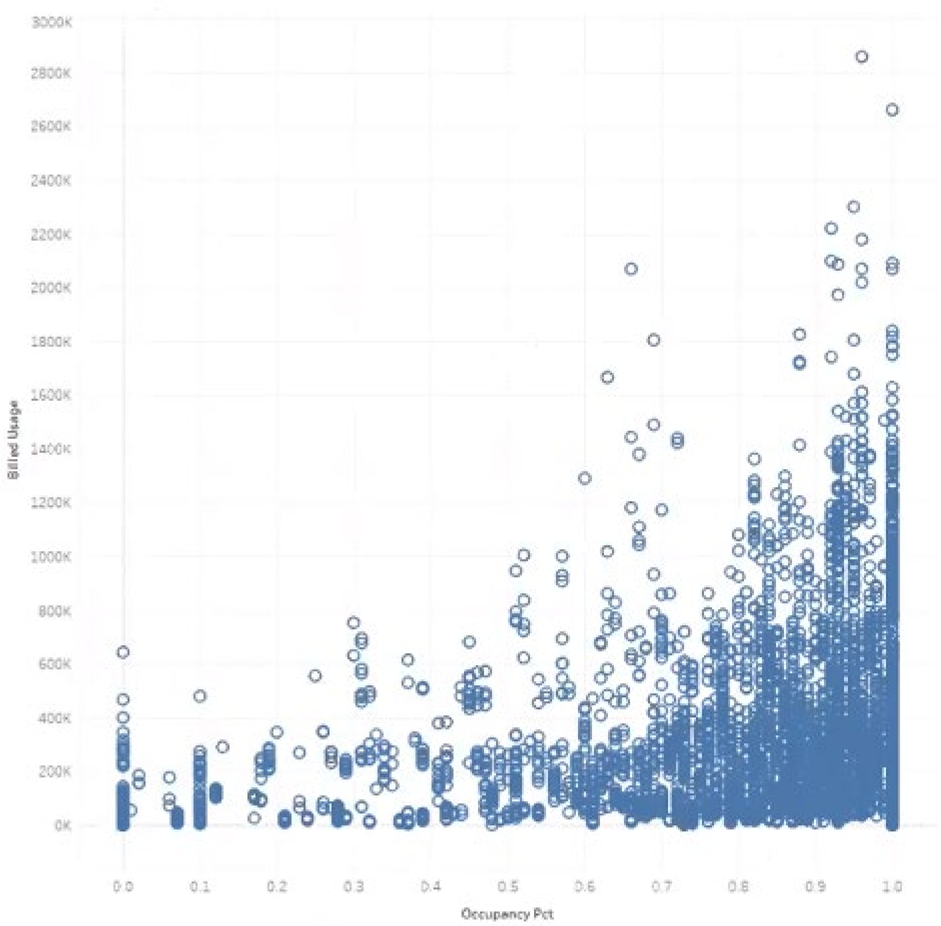
The Occupancy tab has 4,629 rows of data and the properties tab has 3,855

Data came from 57 separate files

Visualizing the data

Below is a scatter plot chart after removing the outliers showing billed usage and occupancy rate.

Most properties had a higher level of occupancy (over 80%) and it appears that there billed usage rises when the occupancy rate goes up.



Libraries used

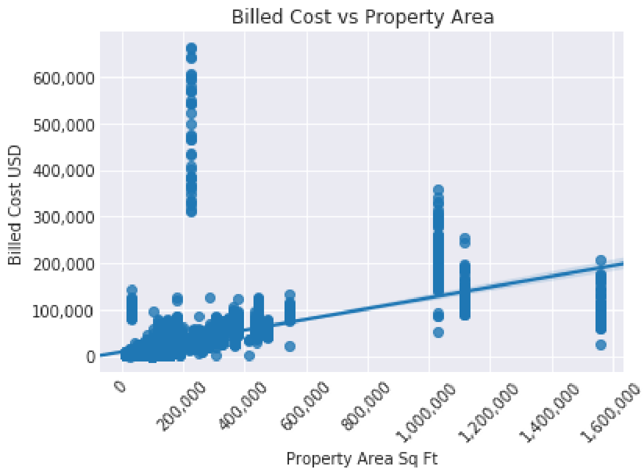
- bokeh – used for interactive dashboard.
- Leaflet – generate map to show location of building by zip code.
- Matplotlib – charts and graphs.
- NumPy – adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Pandas – data manipulation and analysis. Used for data structures and operations for manipulating numerical tables and time series.
- Prophet – used for statistical analysis of 161 different properties.
- Seaborn – visualization of statistical reporting.
- Statsmodels – module that provides classes and functions for the estimation of many different statistical models

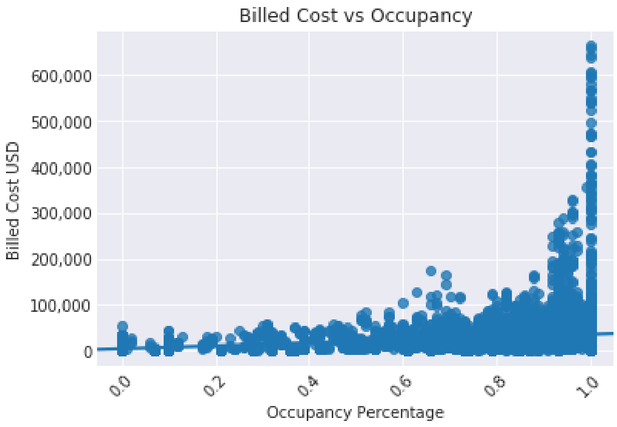
Results

Models

Regression Model

The linear regression model was first used to forecast billed cost based on property size (square feet) and occupancy rate (removed outliers prior to running the model) of all the data in aggregate. The visuals below seem to make sense as billed cost generally grows as the occupancy rate increases (more energy is used as there are more people in the space).



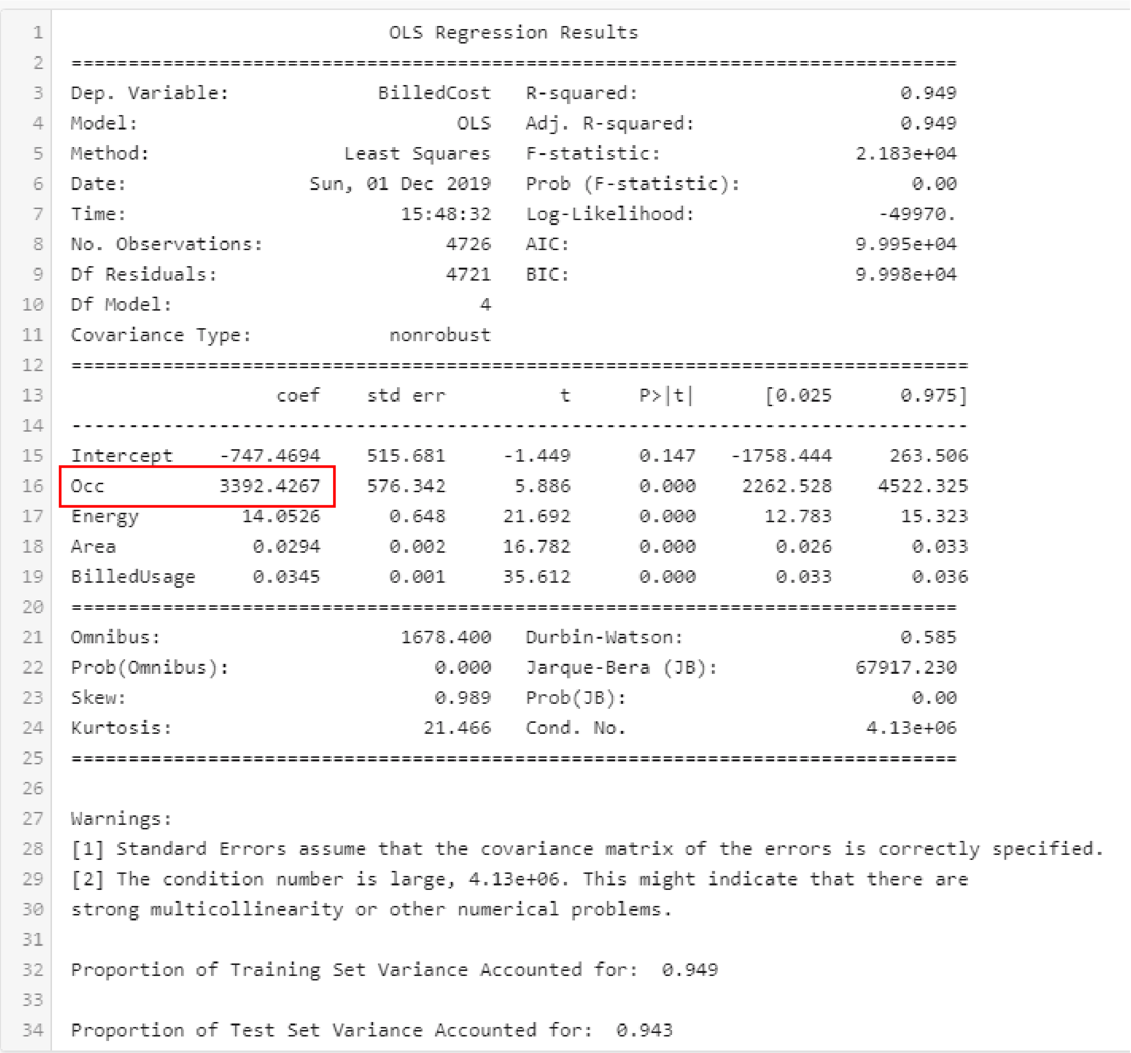


Regression results

The results seem to be positive. The R-squared value at 94% was good. The variables (Occupancy, Energy certified, Area size (sqft), and Billed usage) all showed significant results statistically speaking.

The occupancy goes from 0-1 (0-100%). Thus, the interpretation is that full occupancy will add ~$3,400 to the billed cost.

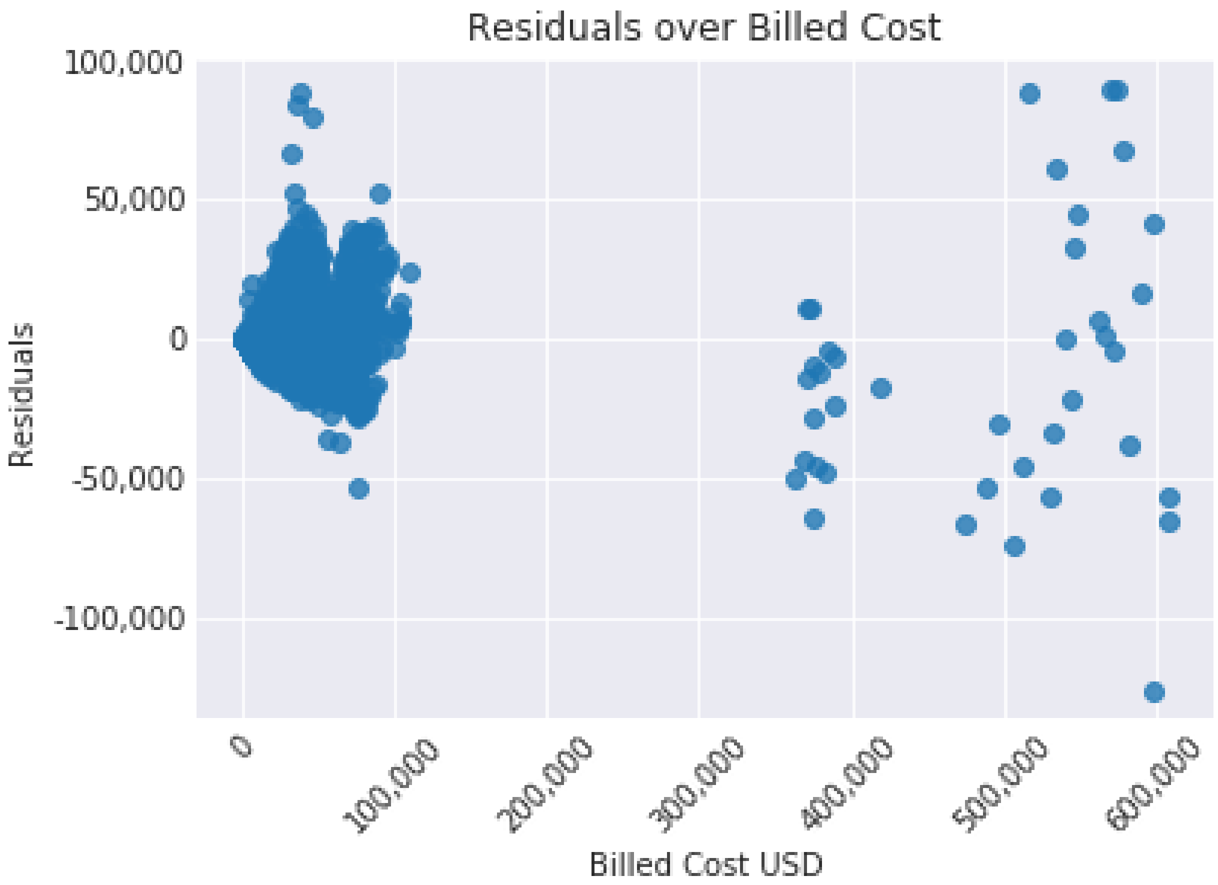
Training set was used to ensure that there was not any overfitting with equally good results.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             BilledCost   R-squared:                       0.949
Model:                            OLS   Adj. R-squared:                  0.949
Method:                 Least Squares   F-statistic:                 2.183e+04
Date:                Sun, 01 Dec 2019   Prob (F-statistic):               0.00
Time:                        15:48:32   Log-Likelihood:                -49970.
No. Observations:                4726   AIC:                         9.995e+04
Df Residuals:                    4721   BIC:                         9.998e+04
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept    -747.4694    515.681     -1.449      0.147   -1758.444     263.506
Occ          3392.4267    576.342      5.886      0.000    2262.528    4522.325
Energy         14.0526      0.648     21.692      0.000      12.783      15.323
Area            0.0294      0.002     16.782      0.000       0.026       0.033
BilledUsage     0.0345      0.001     35.612      0.000       0.033       0.036
==============================================================================
Omnibus:                     1678.400   Durbin-Watson:                   0.585
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            67917.230
Skew:                           0.989   Prob(JB):                         0.00
Kurtosis:                      21.466   Cond. No.                     4.13e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.13e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Proportion of Training Set Variance Accounted for:  0.949

Proportion of Test Set Variance Accounted for:  0.943
```

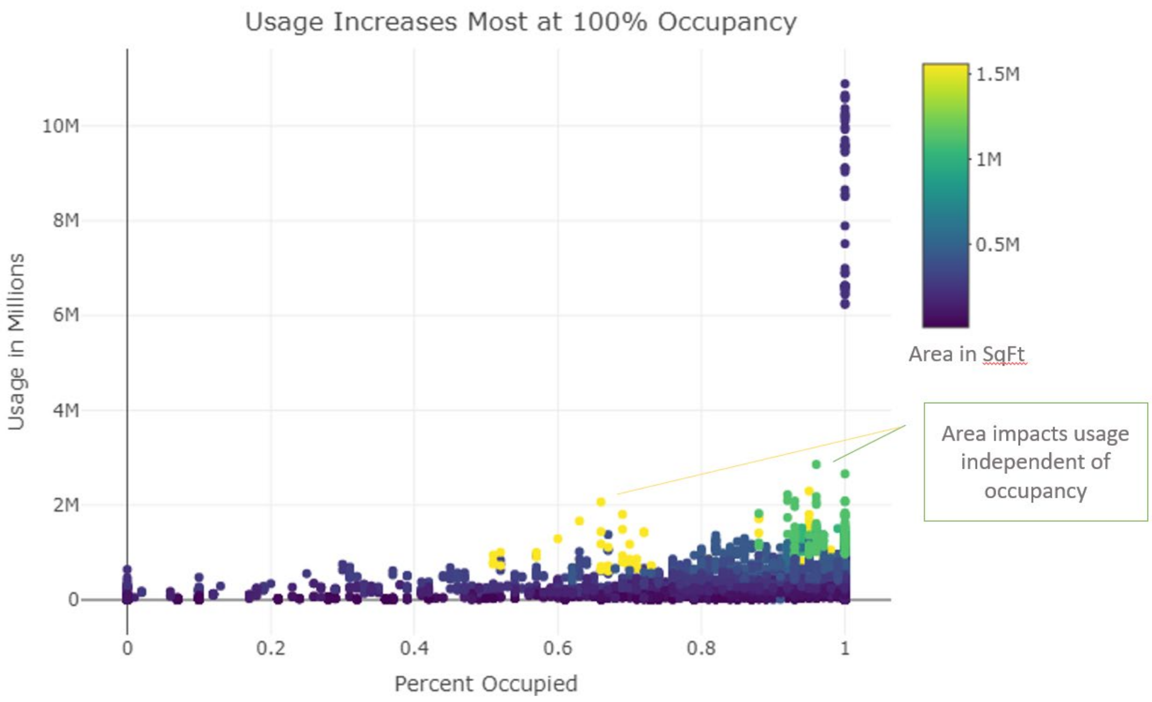
Heteroskedasticity

While some of the early indicators looked promising, it is apparent that there is heteroskedasticity in the data when after exploring the residuals over billed cost. Unfortunately, the dataset that was provided was limited in the information that was offered. Perhaps, the regression model would be more reliable if there were a categorical that differentiated the type of business (for example, ordinary office space, call centers, data centers, etc…) that operated within each building.



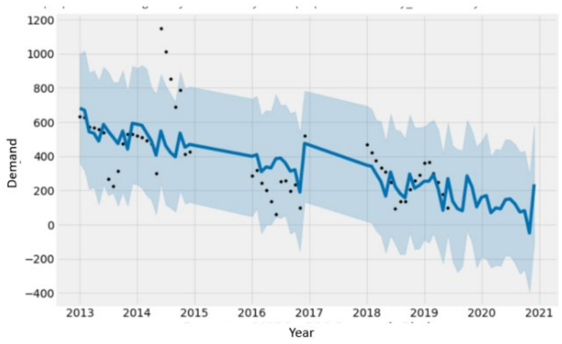
Interpretation

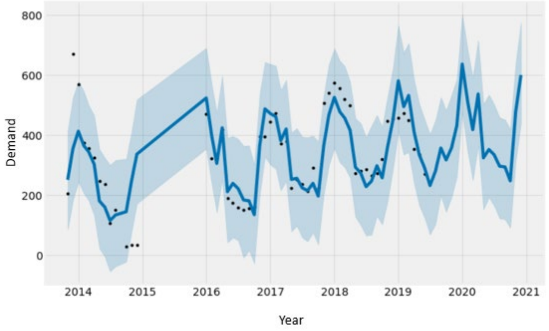
The regression analysis was able to validate that occupancy is significant. Below is another visualization that helps explore the data. In this plot, the larger areas (square footage) are independent in the consumption of occupancy.

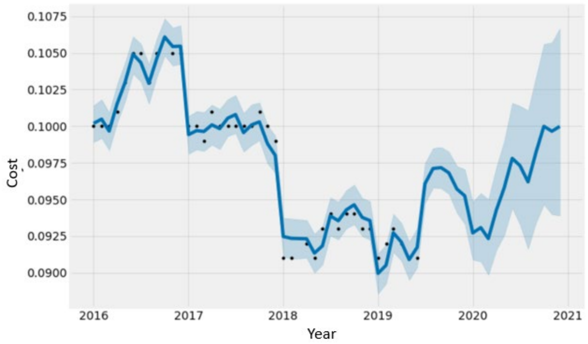


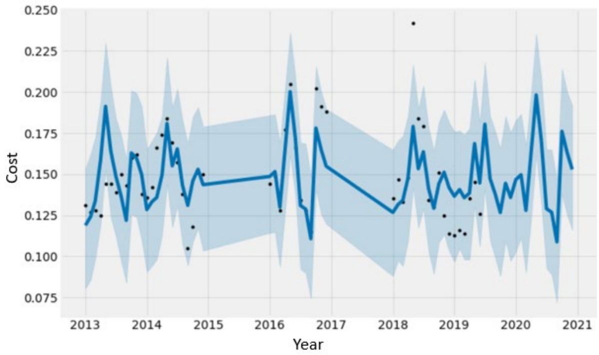
Prophet

Because of the Heteroskedasticity, we could not use the regression analysis to predict future costs. Thus we used the Prophet, which looked at time series predictions for cost. The results were very positive. Below are some examples of results for a few properties:







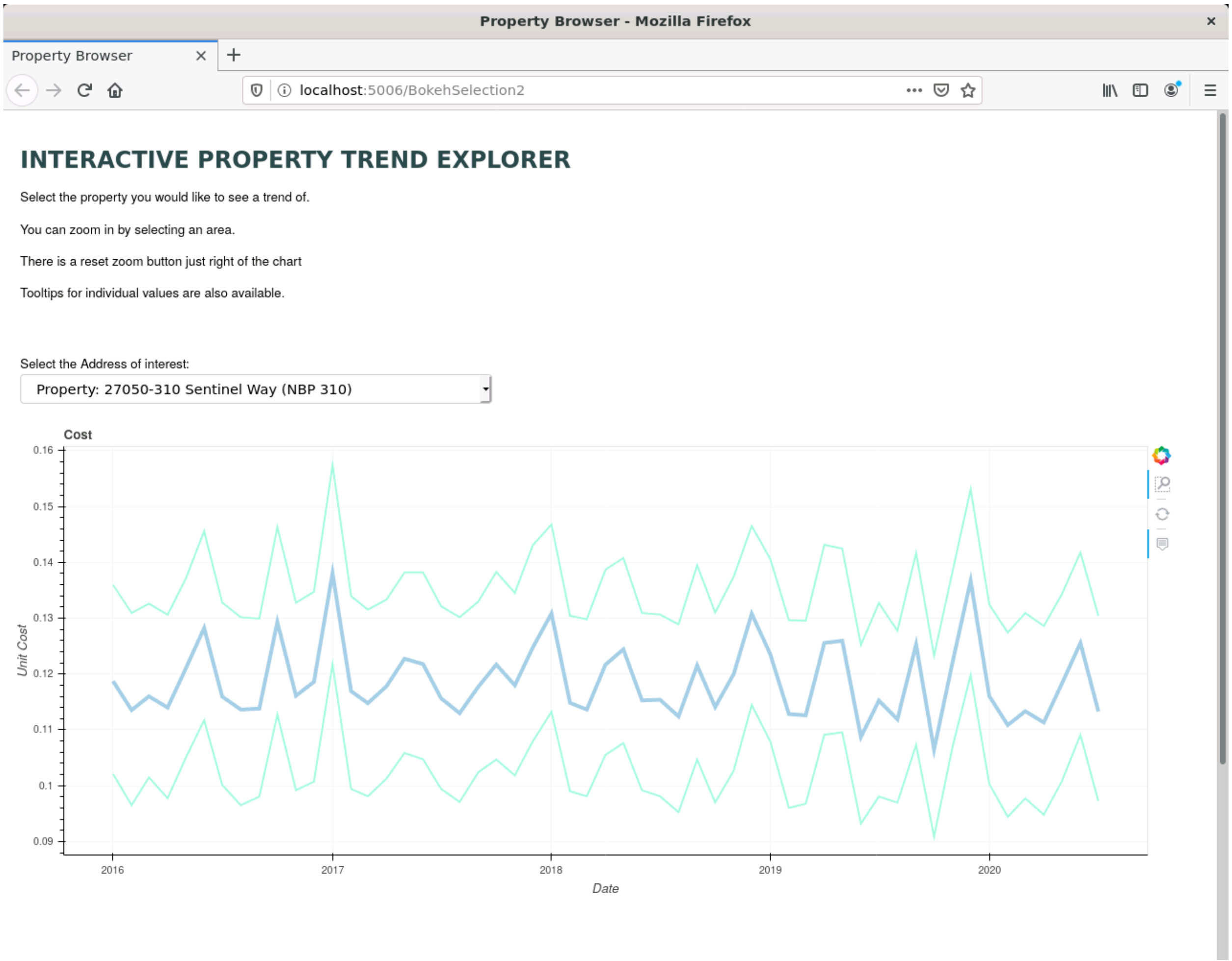


Dashboard

The challenge that comes with interpreting the results from the Prophet is that each property is analyzed separately. Thus, a general conclusion is difficult to make in aggregate of all the properties.

In order to explore the results for each property, it became necessary to create a dynamic dashboard that allows the end user to filter through each property as needed.

Below is a screenshot of the dashboard showing one random property from the data set. This interactive tool was created using python and the bokeh library to plot the results of the time series estimates by property that were created with Prophet.



INTERACTIVE PROPERTY TREND EXPLORER

Select the property you would like to see a trend of.

You can zoom in by selecting an area.

There is a reset zoom button just right of the chart

Tooltips for individual values are also available.

Select the Address of interest:

Property: 27050-310 Sentinel Way (NBP 310)

Conclusion

The app and the interpretation of the Prophet was the most valuable outcome from this project as it relates to Schneider Electric. It helps offer an easy way to look at historical performance, while giving an easy means of accessing future predictions.

Next steps would be to identify a way to differentiate each building into a meaning category of how the property is being used or what type of business is occupying it. This will help improve the way we predict into the future. However, the challenge here will involve identifying those categories and assigning it to the data for future use.

File Manifest

File Name	Description
RegressionCode1.py	Python code for importing the data from multiple excel files, plotting, and runs the regression.
BokehSelection2.py	Interactive web application written in Python to display results for properties from the Prophet time series prediction
AppTitle.html	Required by BokehSelection2.py, The descriptive header for the Interactive web application
ReadFilesandProphet.ipnyb	Reads in all spreadsheets and maintains them individually from a directory, then runs Prophet Time Series based on Demand and Cost Output.
CombinedData2.csv	Output spreadsheet after aggregating all properties and proprietary data removed
Turn On The Lights Group_FInal_Jo_Mike_David_Andrew.ppt	Presentation delivered in class