EO88 Energy Analytics & Data Cleansing

Dan Smilowitz

Historical energy usage and building information is used to analyze performance of New York State agencies towards an Executive Order requiring an improvement in energy efficiency. Missing values are analyzed and imputed to increase accuracy of reported results. Values are found to closely match previous results, allowing the New York Power Authority to proceed with in-house calculation of performance using the framework established through this effort.

Problem Description

Background

On December 28, 2012, New York Governor Andrew Cuomo issued Executive Order 88 requiring that

> By April 1, 2020 ... State Entities shall collectively reduce the average EUI [Energy Use Intensity] ... by at least 20% from a baseline of the average EUI ... for State fiscal year 2010/2011

The New York Power Authority (NYPA) was charged with establishing a management and implementation team to administer the executive order, directed to ensure agencies' compliance by implementing reporting requirements to document progress toward the target. In order to track performance against baselines, NYPA's BuildSmart NY team collects utility bill information for all fuels (e.g. electricity, natural gas, and water) for all covered facilities as compared to the square footage of each facility. This information is reported by each agency and submitted to NYPA in the form of an Excel spreadsheet template.

Challenges

Agencies are responsible for performing their own data validation — the performance of this task has proved incomplete. Many facilities have missing data for the baseline year, as well as showing large spikes or dips in reported usage believed to be due to data entry errors. NYPA must ensure accurate data is used for the establishment of baselines and the tracking and reporting of performance.

EO88 reporting data has been hosted by a software vendor responsible for the the creation and maintenance of NYPA's New York Energy Manager (NYEM) platform. The current structure of the data provides a great deal of redundancy and unused columns, harming the efficiency of computations based on this data. Finally, the vendor's platform has been unable to properly handle the quality issues of reported data.

Finally, the NYEM platform and the annual EO88 report produced by the BuildSmart NY team identify performance at agency and facility levels, but do not sufficiently highlight trends in the data, nor serve to identify performance by facility characteristics to allow for the development of new energy service offerings to help customer meet their required energy reductions.

Data Preparation

An extract of the database was provided in two files – one containing building data, and one containing reported consumption data. The two extract files are imported and their structures investigated.

Building Data

The building data set contains 126 columns, many of which correspond to reporting fields for each fiscal year. There also a fair number of fields corresponding to building information independent of fiscal year. To make more efficient use of storage, the data was converted using Hadley Wickham's tidy data principles:

- Data corresponding to each State Fiscal Year (SFY) was "gathered" to be represented as a field-value pair
- SFY was extracted from each field name & stored as the SFY for each measure
- Any duplicated entries were removed
- SFY-dependent fields were "spread" so that each field is a column and has a value for each SFY
- Spaces were removed from all field names to make references simpler

Building metadata that remains constant and building data that may change by fiscal year were separated into two tables to join like operations and further optimize use of storage. The primary key for each building is `ESPLocationID`; as such, this field is contained in both tables.

Fixed Building Metadata

The building metadata required only minor cleanup – removing duplicate entries (arising from the repeating of these fields across fiscal years) and the renaming of fields for enhanced clarity in naming convention.

This fixed building metadata table is now far more compact, containing only 12 variables.

Variable Building Data

Building data varying across fiscal years required additional cleanup, as the names of fields are lengthy and not all variables are stored as the correct data type due to earlier transformations. The following transformations were performed:

- The redundant characters "SFY" were removed from the `SFY` field
- Field names were adjusted:
  - Parentheses were removed from all field names
  - Long names were shortened to be more manageable
- The unused `Type` field was removed
- SFY-reported fields were converted from character to numeric

This table is now far more reasonable for processing, and is properly tided and normalized, containing no fixed metadata fields.

Consumption Data

The reported consumption dataset contains roughly 177,000 rows and 16 columns. As with the two building datasets above, field names were modified to remove any spaces or parentheses and shortened to be more concise. Fuel and units of measure of reported usage are reported as a single field; these were separated into separate fields to enable later analysis.

Conversion to Source kBtu

Energy use intensity is calculated using a simple formula:

$$EUI = \frac{Energy\ Use}{Area}$$

Reported usage values span seven different units. In order to create a unified view of EUI across all fuels, all reported usage must be converted to a single unit of measure. The standard units for measuring EUI are thousands of British thermal units (kBtu) – Energy Star Portfolio Manager provides extensive Thermal Energy Converstions technical reference to handle these conversions.

Progress towards EO88 goals is measured using *source EUI*, which takes into account the total energy needed to make the energy used available (i.e. the amount of raw fuel needed to meet the consumption need). Therefore, the kBtu must be converted from site energy use to source energy use. Energy Star Portfolio Manager also provides a Source Energy technical reference, which shows that these conversion factors vary by fuel type. These conversion factors vary over time based on changes in power generation. To avoid changes in conversion of reported values between fiscal years, the BuildSmart NY team uses a fixed conversion table for all years; this table was used for conversion.

Removal of Redundant Data

The first seven of the 16 variables in the consumption data are also included in the building data. Following conversion to source kBtu, these variables were removed (with the exception of `ESPLocationID`) to enforce normalization. Additionally, the empty `Rate/ServiceClassification` was removed, as was the redundant variable containing the fuel & units (in favor of the separated `Fuel` and `Units` variables).

Changes in Campus Reporting

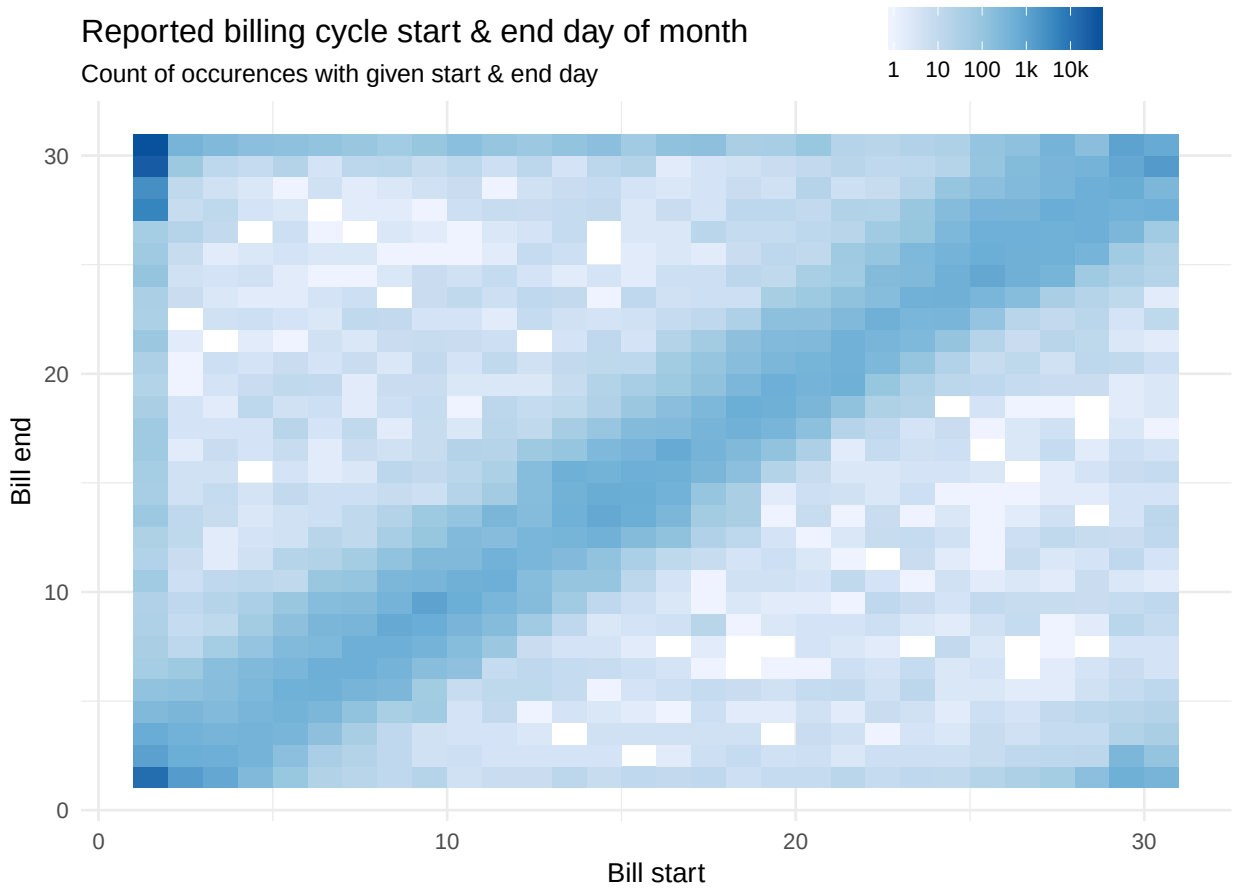
One of the fields in the consumption data is `Parent` – this indicates relationships between facilities where one facility may be a sub-facility of another (such as a college campus or office plaza). Investigation of the consumption data revealed that not all `ESPLocationID` values supplied in the consumption data have associated values in the building data (fixed or variable). Discussion with the BuildSmart NY team revealed that this occurs in two cases:

- One facility with a number of sub-facilities was exempt from EO88 reporting & compliance
- Many large college campuses moved from site-level reporting to campus-level reporting

Those sites that changed to campus-level reporting and are not included in the building metadata were changed to the `ESPLocationID` of the parent facility so that their information can be accurately included in calculation of performance in earlier fiscal years. The `Parent` field was then removed, as it is no longer relevant.

Mapping to Calendar Month

As outlined above, reporting for EO88 is based on New York State Fiscal Years, which span from April 1 to March 31 of the following year. In order to accurately report on each State Fiscal Year's source EUI, reported usage must align with each State Fiscal Year, particularly the calendar months at the start and end of each SFY. Investigation of billing period start & end dates shows that while an bills most commonly start at the beginning of the month and end at the end of the month, many bills start and/or end at other points in the month, with a strong trend towards roughly matching start and end days:



These values must be converted to contain the reported usage in a given month so that they can be mapped to State Fiscal years. In order to accomplish this, a function is created that extracts, for each reported value:

- Each calendar month the value encompasses
- The number of days the billed value overlaps each month encompassed
- The prorated share of usage per month, assuming flat usage through the period

This function was applied to each reported consumption value. During application, the State Fiscal Year for each month is also added, and the bill start & end dates removed. Finally, months falling outside the required reporting period (SFY 2010-11 through SFY 2016-17) were excluded.

Following this conversion to calendar dates, the building and consumption data are sufficiently prepared for analysis.

Data Cleansing: Initial Approach

As identified in the *Problem Description* section above, a major issue encountered in analyzing and reporting EO88 performance is the presence of missing data, as well as possible abberant reported data. In order to identify reported values in need of correction, a function is created to flag data points that met any of three criteria:

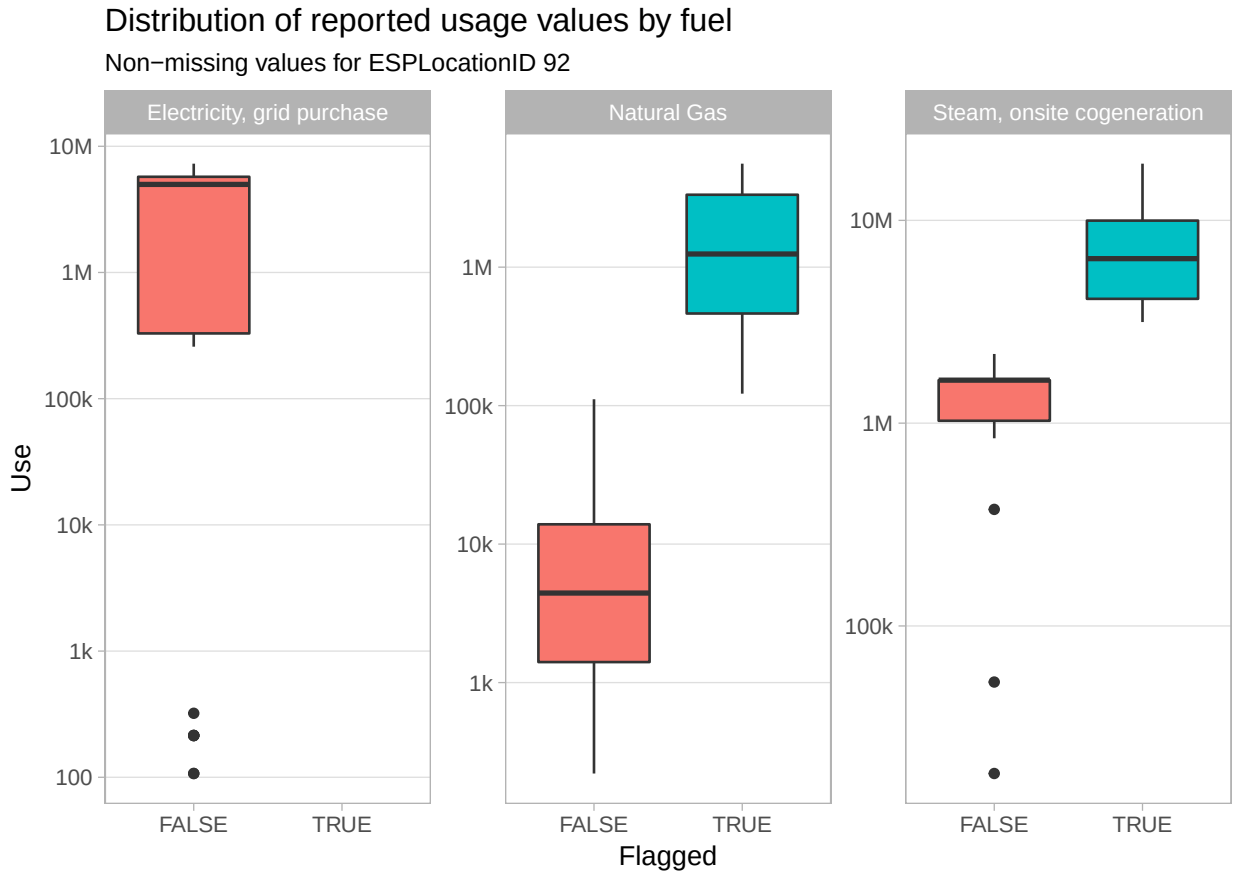
- Missing usage data
- Reported usage < 0
- Reported usage that is an outlier for a given facility & fuel

The `caret` package was utilized to model more appropriate values for those values flagged. Weather data from the National Climatic Data Center's Divisional Data was utilized to model this data using an array of methods:

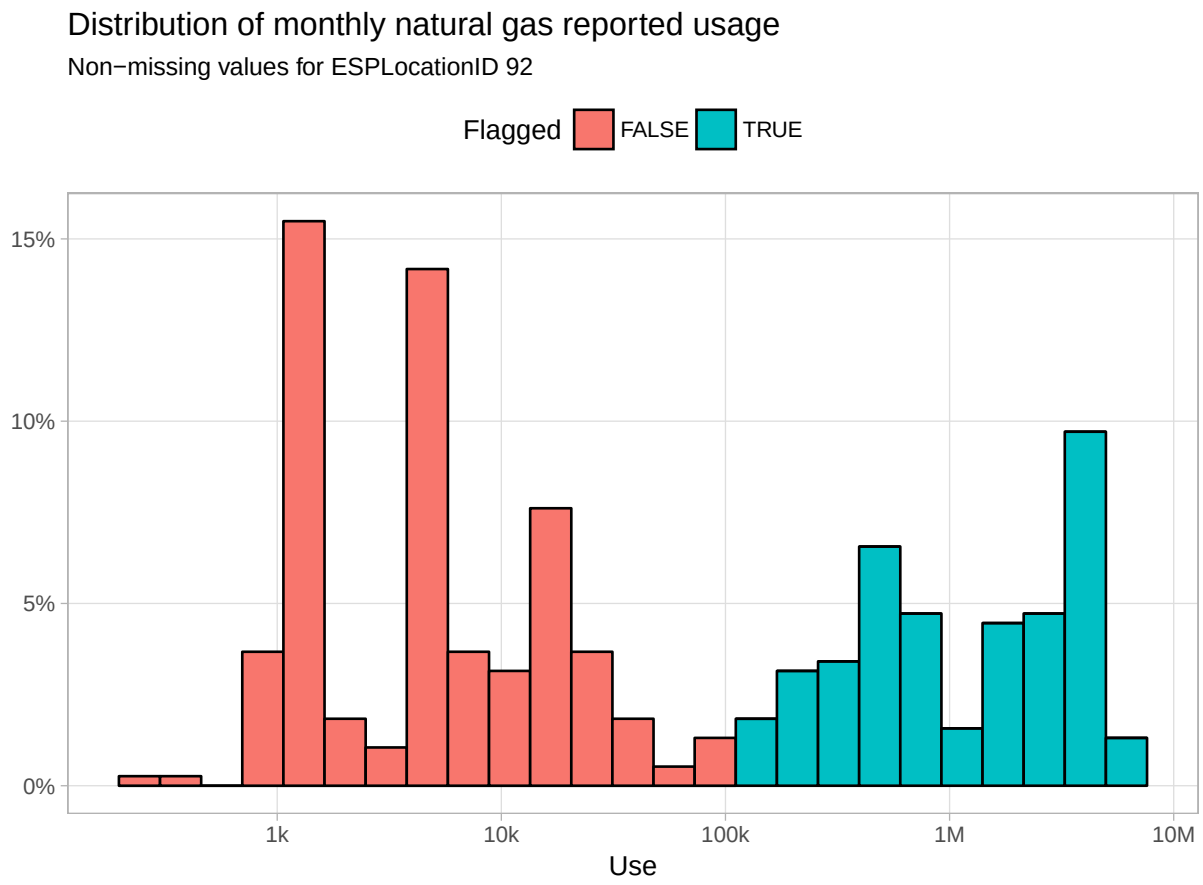
- Training a number of models on the entire as-is dataset
  - 4 linear models
  - 6 nonlinear models
  - 5 tree-based models
- Training a k-nearest neighbors (KNN) model on the entire dataset
  - Using dummy variables for fuel & building type
  - Using transformed predictors
  - Centered & scaled
  - KNN-imputed
  - Highly-correlated & zero-variance predictors removed
- Training a KNN model for each location
- Training a KNN model for each fuel
- Training a Cubist model for each location & fuel combination

Additionally, imputation was attempted using only data for a given facility & fuel in a given calendar month (with a window of $\pm$ 1 month added).

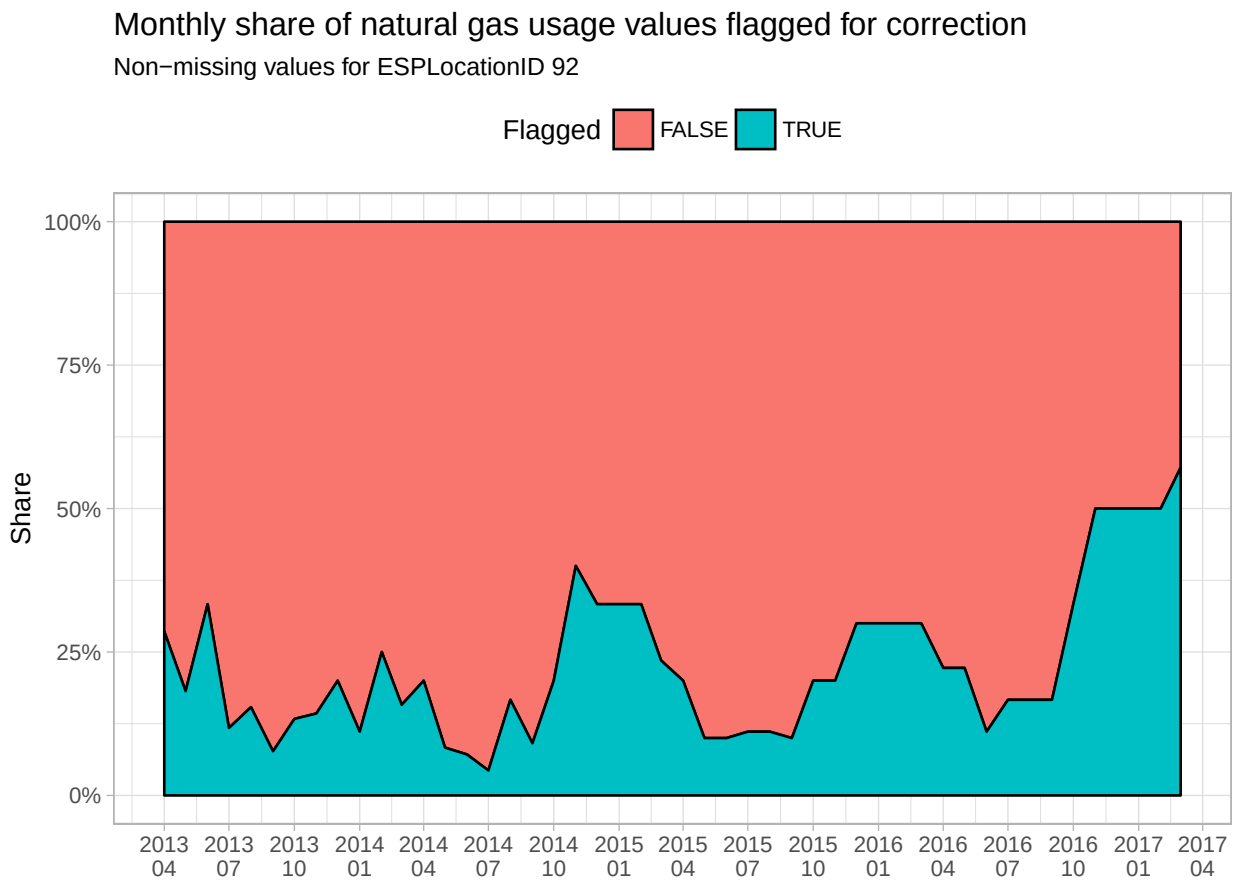
Despite the varied approaches employed and hundreds of models fit to the data, the errors returned by resampling and evaluation against test sets were very high, often as high as the order of 1×10^9 kBtu. To identify possible causes of this error, a facility with particularly high error across the models fit (`ESPLocationID` 92) was selected for investigation. First, a boxplot of the non-missing usage values by fuel was created, separated by flagged & un-flagged values:



This plot appears to show a relationship between fuel type, usage, and flagged status. For electricity, all but 3 values were over 100,000 kBtu, and none were flagged for correction. For natural gas and steam, all values over roughly 100,000 kBtu were flagged for correction, and none below this threshold were flagged. This was further investigated by creating a histogram of non-missing natural gas usage, colored by flag value;



This appears to confirm the bias towards flagging larger reported usage values. This was interpreted as a side effect of the function used for flagging non-missing values – since the scale of values varies so widely for some fuels, values two orders of magnitude higher than others will be flagged as outliers given the standard criteria of $x < Q1(X) - 1.5IQR(X)$ or $x > Q3(X) + 1.5IQR(X)$. In order to address this, a log-transformation of reported usage values is considered prior to flagging data for correction. Given that the two fuels identified above are thermal fuels with expected seasonal usage, the flagged status for natural gas usage is also investigated by month:

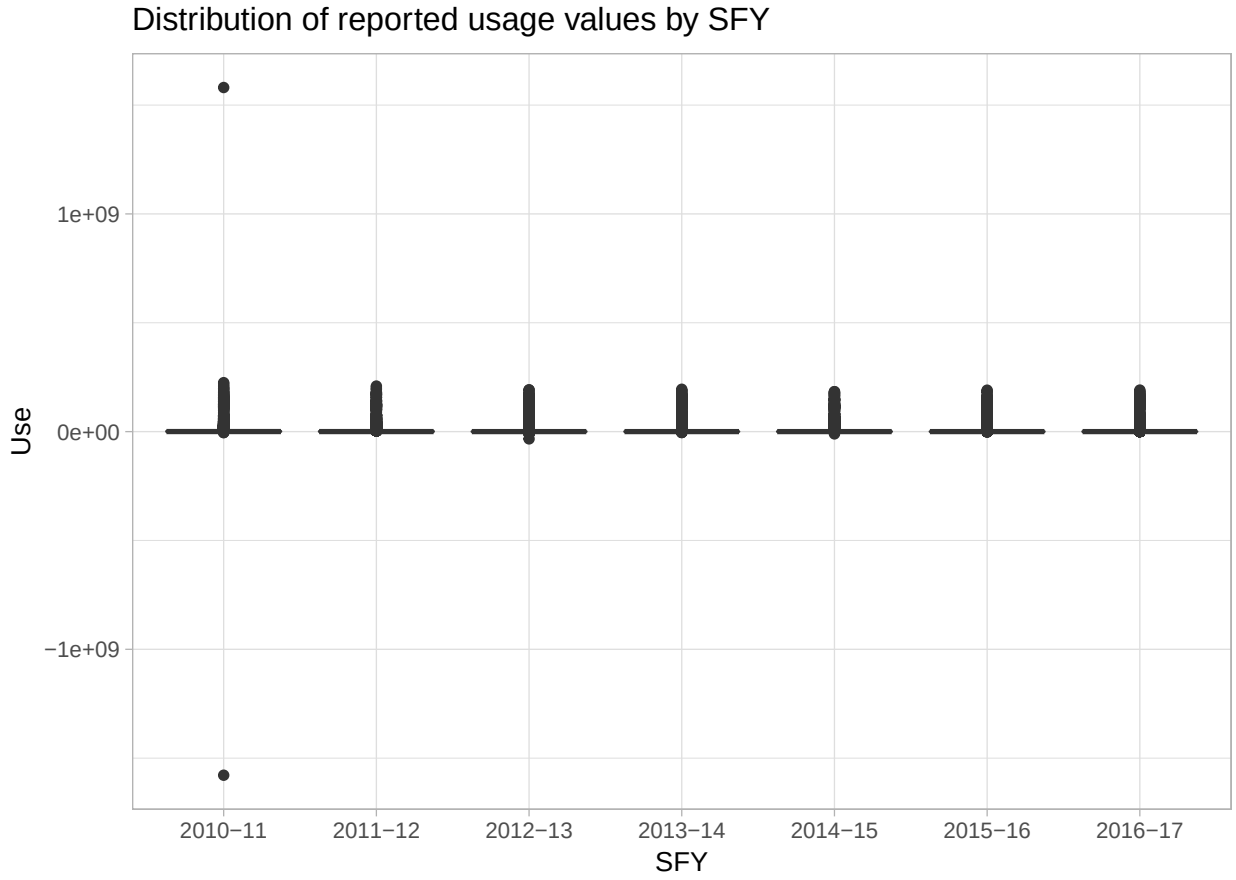


This plot shows clear seasonality, with far more values flagged for correction in winter months. This makes sense given the earlier observation of higher values being flagged, as natural gas is intuitively expected to have higher usage in winter months given its use for heating.

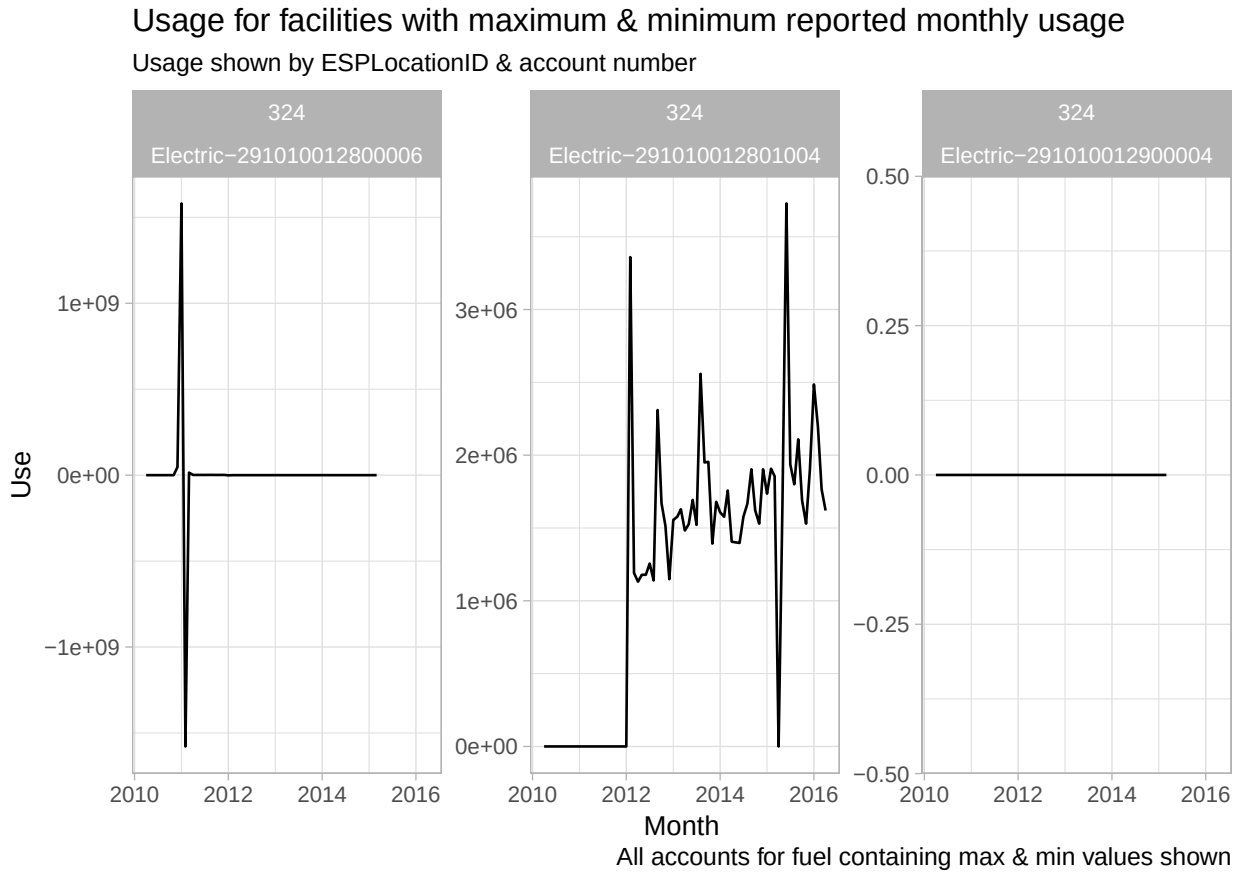
These patterns, as well as the poor accuracy of attempted data correction, were brought to the BuildSmart NY and NYEM teams to help identify a solution to improve accuracy and reduce any bias. The NYEM team, after conversation with the vendor that provided the data extract, revealed that **all provided values has already been verified by the reporting agencies.** This information changed the direction of the approach: data no longer needed to be flagged for imputation; only missing values needed to be filled.

Data Cleansing: Revised Approach

The redefined scope of this project called for a focus on only missing values. Prior to investigating these missing values, the overall distribution of reported usage values was investigated:



There are two values in SFY 2010-11, one positive and one negative, that are well beyond the scale of all other reported values. These two extreme values (the overall maximum and minimum) are investigated; usage for the facilities and fuels corresponding to the extreme values are plotted by account number:

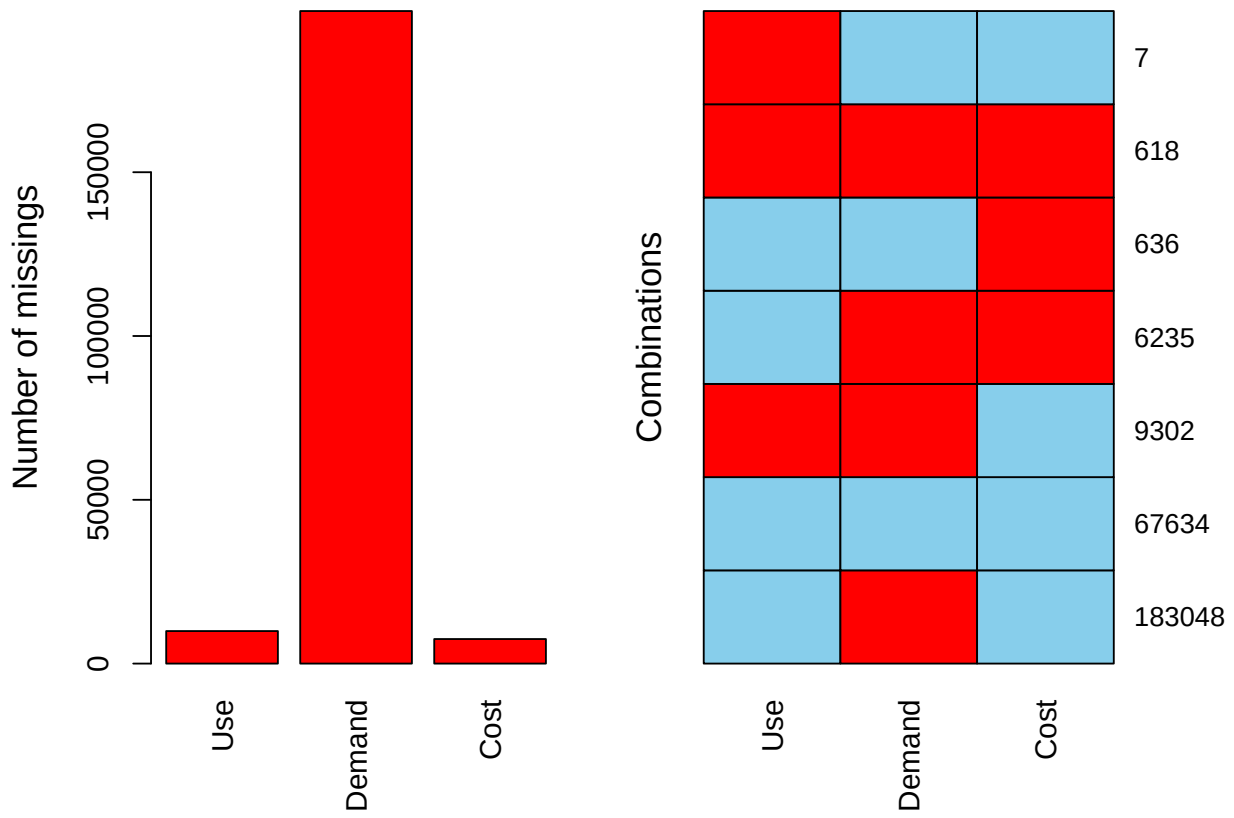


The maximum ($\sim 1.6 \times 10^9$) and minimum ($\sim -1.6 \times 10^9$) reported monthly kBtu occur for the same facility (`ESPLocationID` 324), fuel (Electricity, grid purchase), and account number (*Electric-291010012800006*) in consecutive months (January-February 2011). While these numbers have been verified by the reporting agency (per the NYEM team), it is likely that these two errors represent a billing or metering error followed by a corresponding correction. To avoid any errors in reporting that may be caused by these extreme values, both were replaced by the mean use of the two months; the same is performed for cost and demand. A modified version of the consumption data set was created with an additional variable `Imputed` added to track all imputations performed; this change is recorded as *manual*.

To understand missing values and how they may be imputed, the frequency and patterns of missingness were investigated using the `mice` package (note that the `Imputed` variable was created as an empty character string to avoid obfuscating missingness patterns):

	67634	119	7	182929	636	3	9302	6232	618	
ESPLocationID	1	1	1	1	1	1	1	1	1	0
Month	1	1	1	1	1	1	1	1	1	0
Share	1	1	1	1	1	1	1	1	1	0
Cost	1	1	1	1	0	0	1	0	0	7489
Use	1	1	0	1	1	1	0	1	0	9927
Demand	1	0	1	0	1	0	0	0	0	199203
Utility	0	1	0	0	0	1	0	0	0	267358
AccountNumber	0	0	0	0	0	0	0	0	0	267480
Fuel	0	0	0	0	0	0	0	0	0	267480
Units	0	0	0	0	0	0	0	0	0	267480
SFY	0	0	0	0	0	0	0	0	0	267480
Imputed	0	0	0	0	0	0	0	0	0	267480
	6	6	7	7	7	7	8	8	9	1821377

The table shows that only three reported variables contain missing values – `Use`, `Demand`, and `Cost`. The patterns in missing data between these variables were visualized using the `VIM` package:



The left segment of this plot shows that there are roughly 10,000 cases missing `Use`, roughly 8,000 missing `Cost`, and over 200,000 missing `Demand` – this last number is sensible since the original name of the field specifically referenced kW and therefore would only apply to electricity. The right segment shows that there are six patterns in which data is missing some or all of these variables, with the most common being missing only demand.

For the purposes of Executive Order 88 analysis and reporting, demand and cost are of no immediate import; only use is of concern, as it is used to calculate source EUI. There are only three patterns of missingness which involve missing use:

- Missing use; reported demand; reported cost (7 instances)
- Missing use; missing demand; missing cost (618 instances)
- Missing use; missing demand; reported cost (9302 instances)

Each of these patterns was investigated individually, as the approach to imputing the missing values was expected to vary.

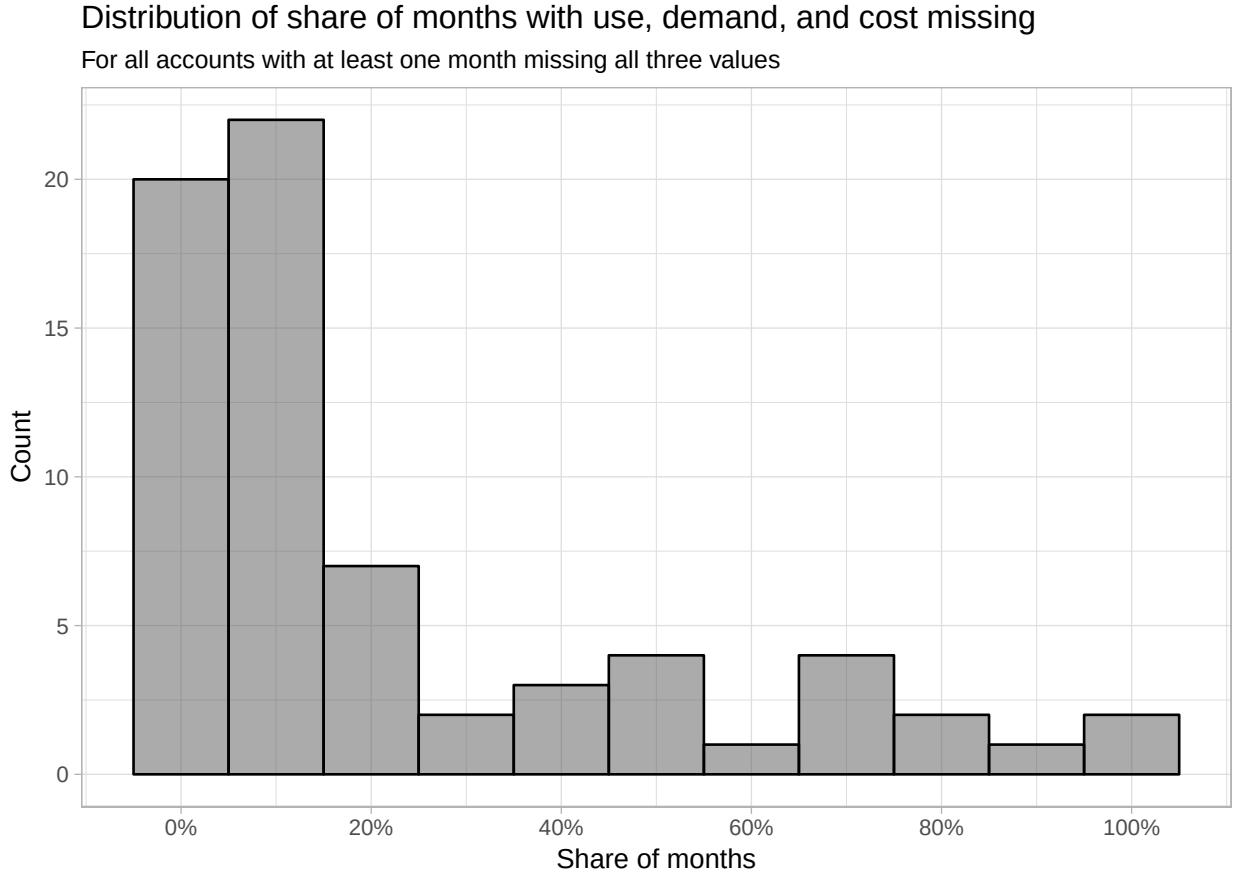
Reported Demand & Cost

The instances of missing reported use with reported demand and cost all occur for the same account (*Electric-495520212030000*) in consecutive months (April-October 2012); the demand and cost values for each of these months is zero. Investigation with NYPA staff reveals that

this account did not exist prior to November 2012 – it is likely that the reported zero demand and cost were back-filled to populate the State Fiscal Year with data. These use values are imputed with zero and flagged with an `Imputed` tag of *zero*.

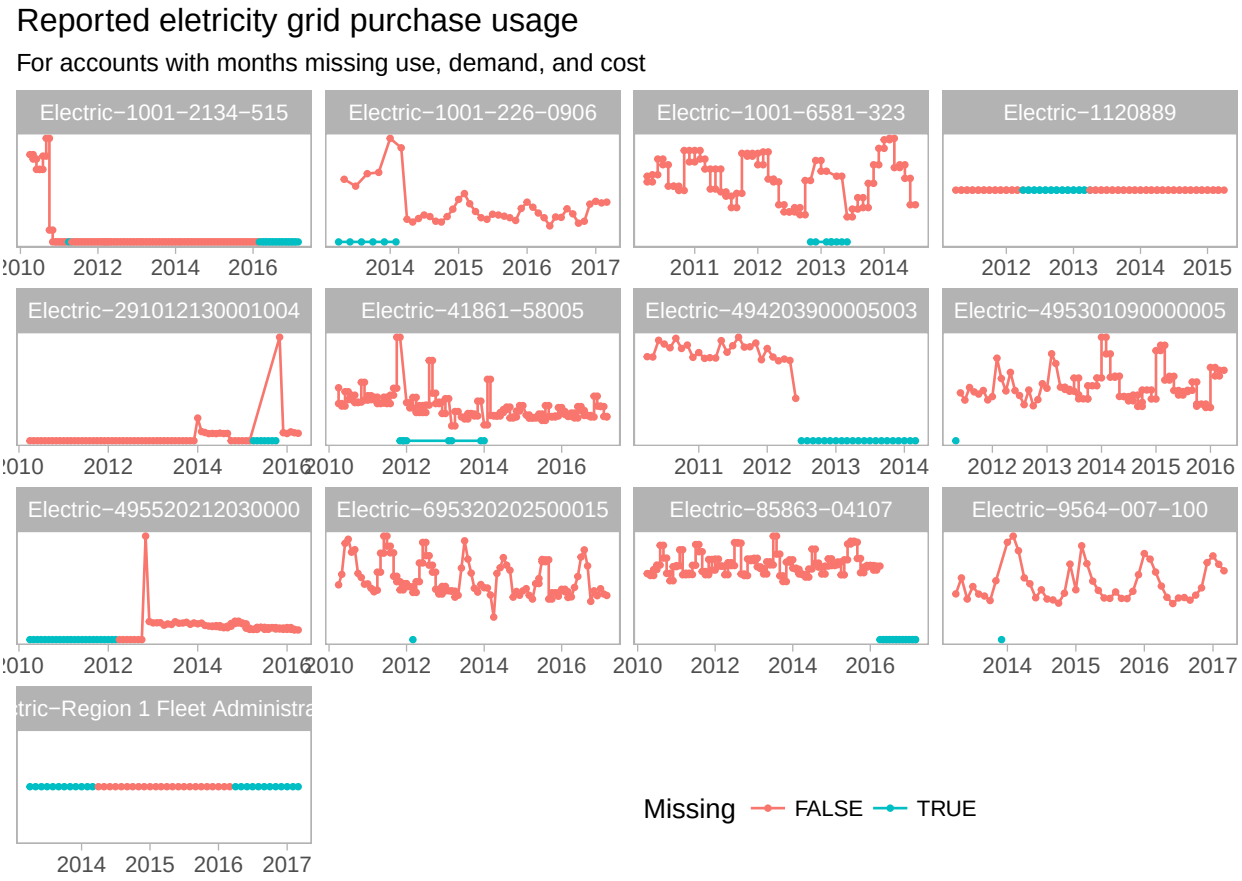
No Demand or Cost

For accounts containing entries missing each of cost, demand, and cost, the share of months containing all missing values was calculated, and the distribution of accounts by share investigated:



The two account numbers with 100% all-missing values both belong to the same facility and span only April 2011 through April 2012. Investigation of the variable building data revealed that the `Status` of this facility in SFY 2011-12 and SFY 2012-13 was *excluded*, meaning that its usage is not reported as part of EO88 compliance. To investigate if this explained all missing values, the consumption and building data sets were joined and data missingness reexamined. Over 98% of missing values still remained, so further investigation was continued, and excluded values were considered throughout.

Electricity grid purchases were investigated first by visualizing the accounts containing at least one month missing all three values:



Many of these accounts appear to have missing values before and/or after data was reported with non-missing values. Given the back-filling seen in the above section to make reported values line up with State Fiscal Years, it was assumed that these accounts followed the same pattern. Missing use values for these accounts in months also missing demand and cost were filled with zero and tagged as *zero* imputation.

Six of the accounts do not appear to follow this pattern – their all-missing values do not fall before or after periods with non-missing values. These values were flagged for further imputation – given the lack of other reported values (demand and cost), an approach that incorporated other data was selected. The weather data that was acquired for the initial approach was identified as data for use in imputation, and random forest imputation was selected as the preferred method due to its low variance and bias. These values were tagged with *rf* imputation but left as missing until all investigation is complete.

A large number of account number for thermal fuels (fuel oil #2, natural gas, kerosene, propane, and steam) contained missing use, cost, and demand. Inspection of the patterns of these missing values reflected periods of usage followed by no use as well as periodic purchases. Both of these patterns make intuitive sense for thermal fuels, as they are highly seasonal; further, some fuels are reserve-based – a consumer purchases a reserve of fuel and uses it in a non-metered fashion. Based on this, the "triple-missing" values for thermal fuels were imputed with zero.

Reported Cost; No Demand

For accounts with entries missing use and demand with reported cost, the number of months between instances of this missingness pattern was investigated. It was found that many accounts showed zero months between instances, suggesting that these accounts *never* had any use but *always* had a cost – this is reasonable, as some cost-only accounts exist for non-metered fuels. These instances represented 83% (419/503) of accounts and 90% (8341/9302) of entries showing this pattern. These cost-only accounts were imputed with zero and flagged accordingly.

The number of accounts and instances of each gap between missing values are investigated following the removal of cost-only accounts:

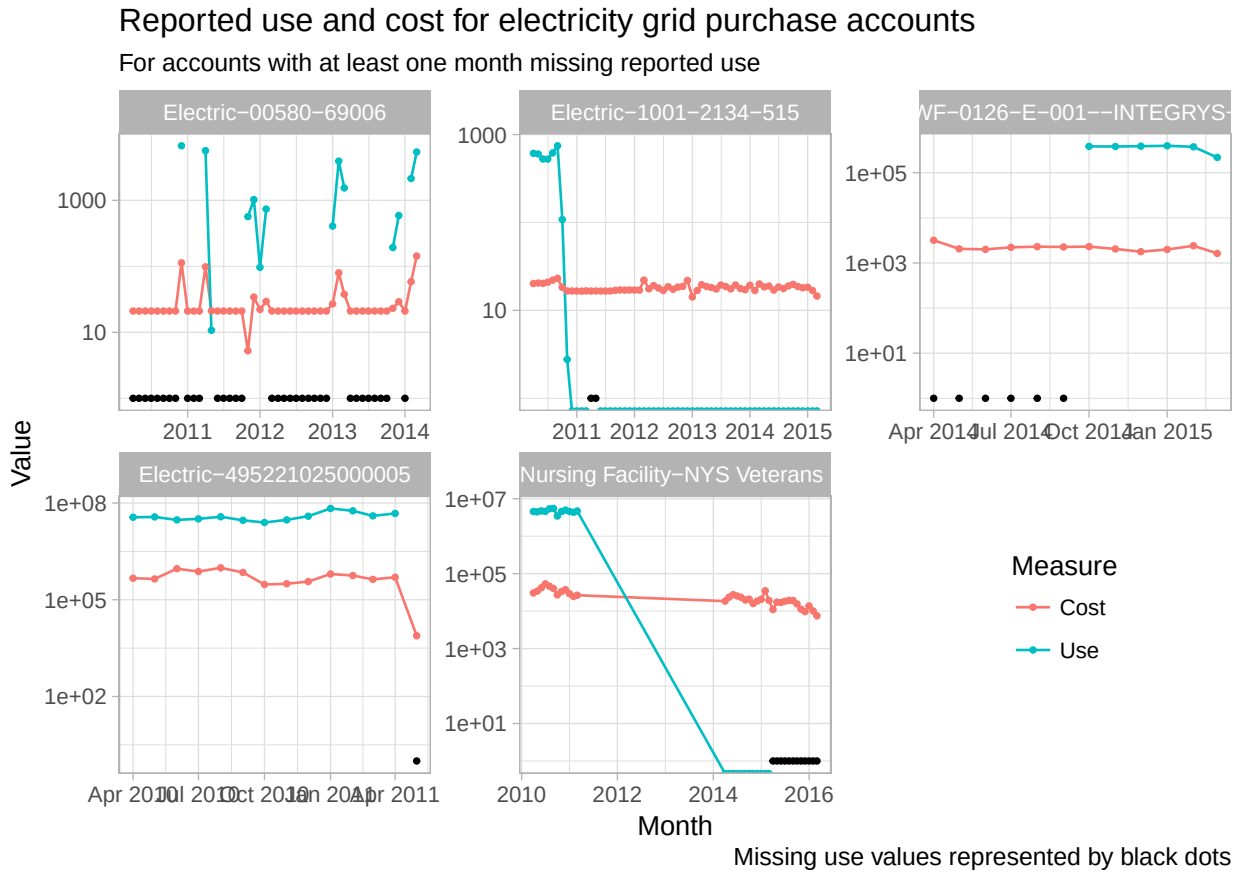
MonthDiff	n_accounts	n_instances
33	1	1
13	1	1
10	1	1
9	1	1
8	1	1
7	1	1
6	4	4
5	8	8
4	5	5
3	1	2
2	8	8
1	77	561
0	52	283

There are still many accounts and instances showing consecutive months with missing use and reported cost. The data set was checked for duplicates, but that was revealed not to be the cause. Remaining accounts and instances missing are investigated by fuel:

Fuel	n_accounts	n_missing
Natural Gas	72	831
Fuel Oil #2	6	69
Electricity, grid purchase	5	60
Propane and Liquid Propane	1	1

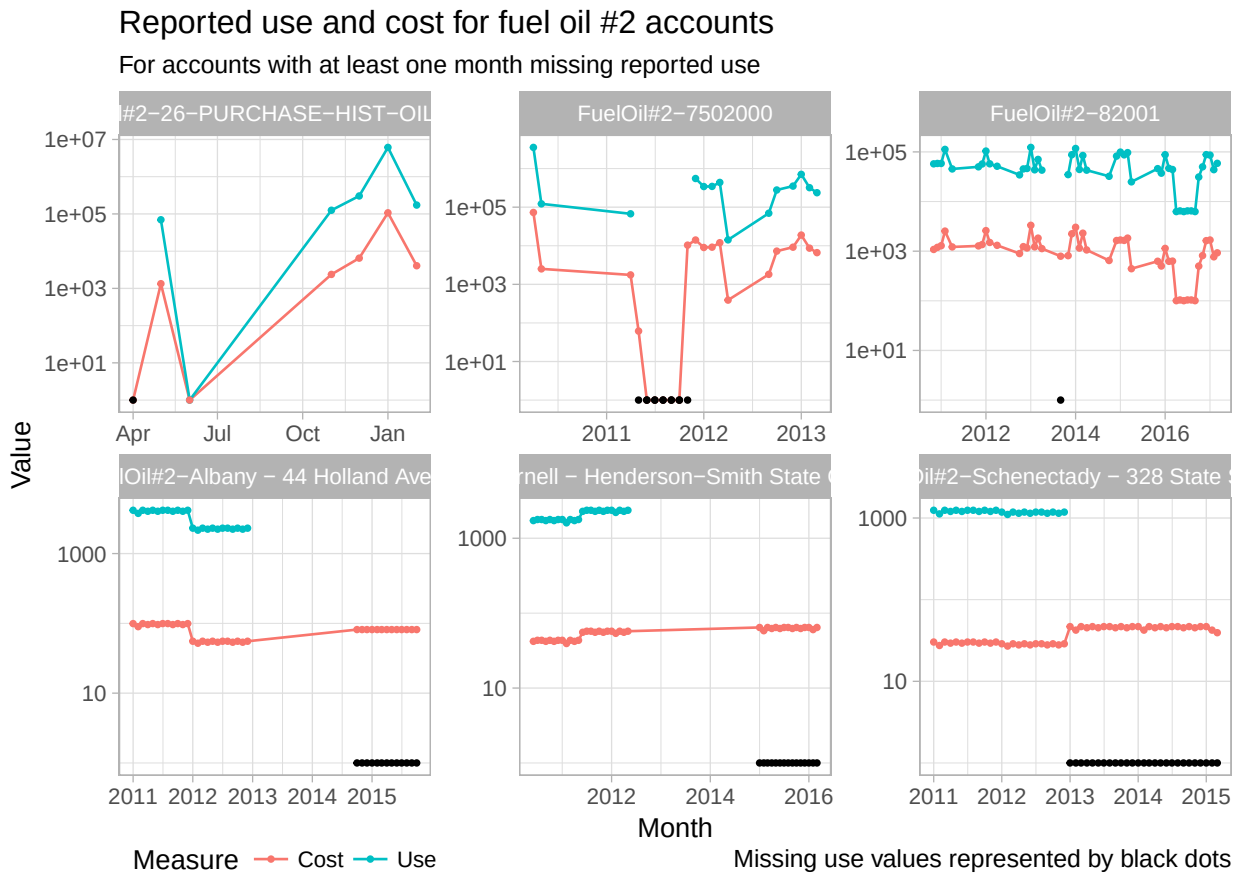
The single missing value for a propane account occurred 7 months after the prior reported value and was followed by a zero 6 months later. This missing value was filled with zero.

Electricity values were next inspected graphically:



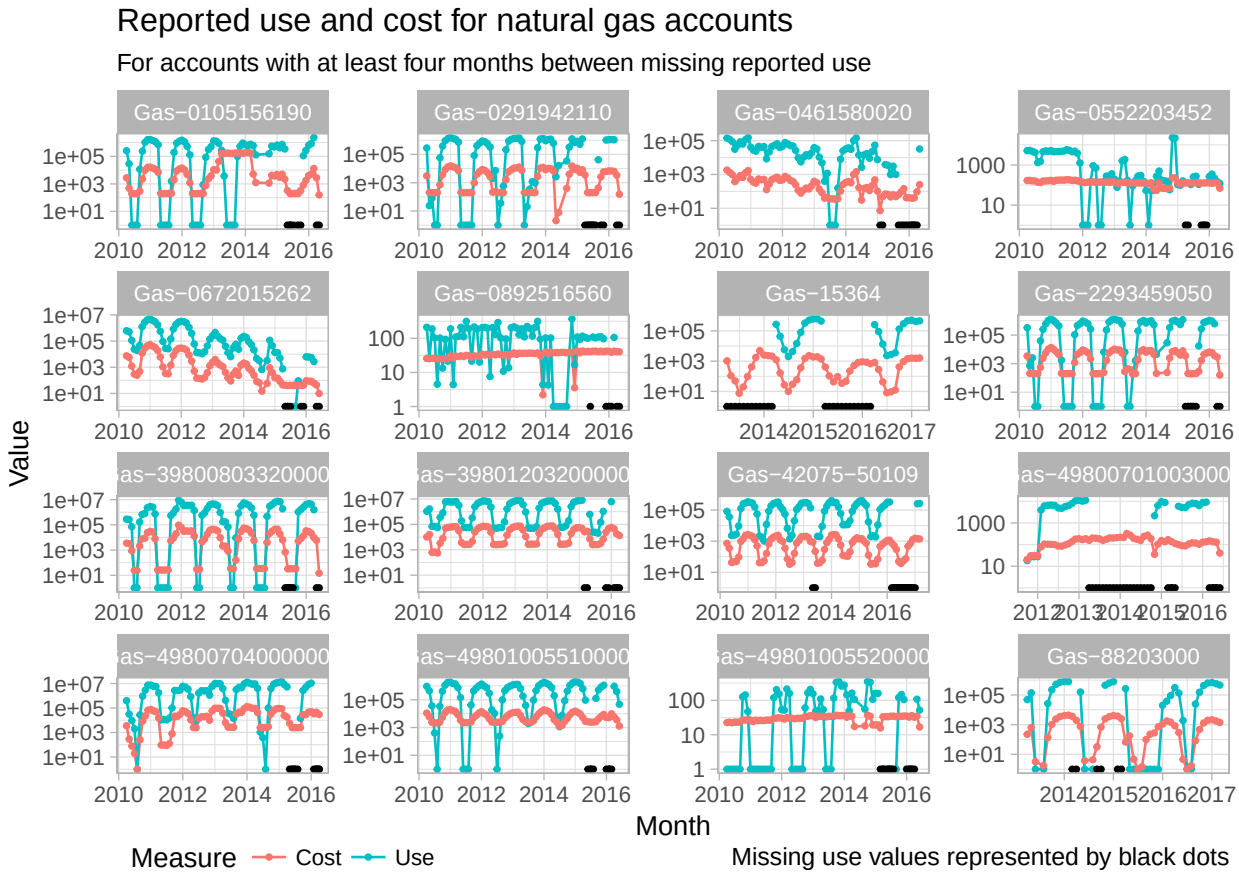
Three of the five accounts show missing values either surrounded by zeroes or in the last months of their reporting. These accounts were assumed to be zero, possibly representing cancelled accounts with remaining charges. The remaining two accounts were identified for more detailed imputation – given the clear relationship between cost and use, linear regression was selected as the imputation method for these values.

Fuel oil was investigated next. All accounts were shown to either be missing a single value or have missing values at most one month apart. Accounts were again investigated graphically:



The first two accounts have missing values that appear to correspond to zero based on other points in the plots; these were filled with zero. The remaining four accounts were flagged for imputation with regression.

The large number of natural gas accounts were prioritized based on the time between missing values – these values ranged from 0 to 33 months. Accounts with time difference between missing values over 4 months were investigated first:

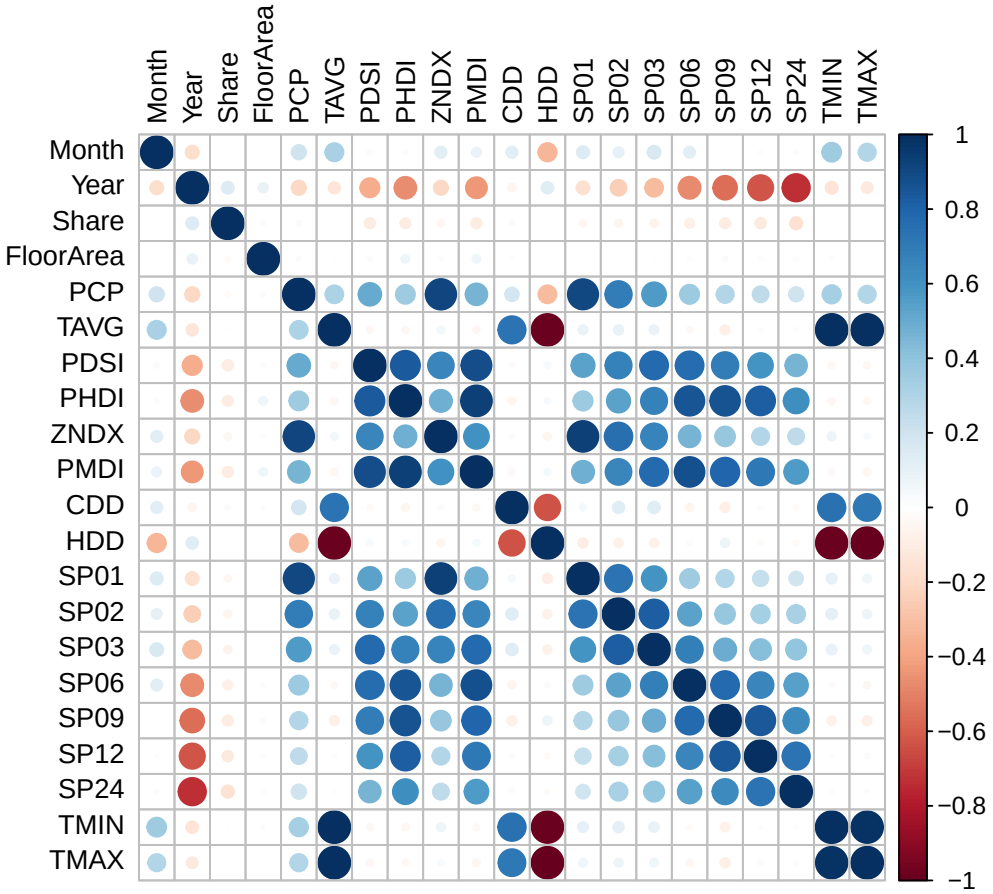


Some of these accounts appear to represent "minimum bill" accounts – accounts that have a minimum cost even with zero use. Programmatic identification of these accounts was attempted but was unsuccessful; graphical investigation of the 72 natural gas accounts was instead performed. 20 of the accounts were found to be likely minimum bill accounts or have missing values before or after any reported values – these were filled with zero. Missing values from the remaining 52 accounts were flagged for imputation with linear regression.

Data Imputation

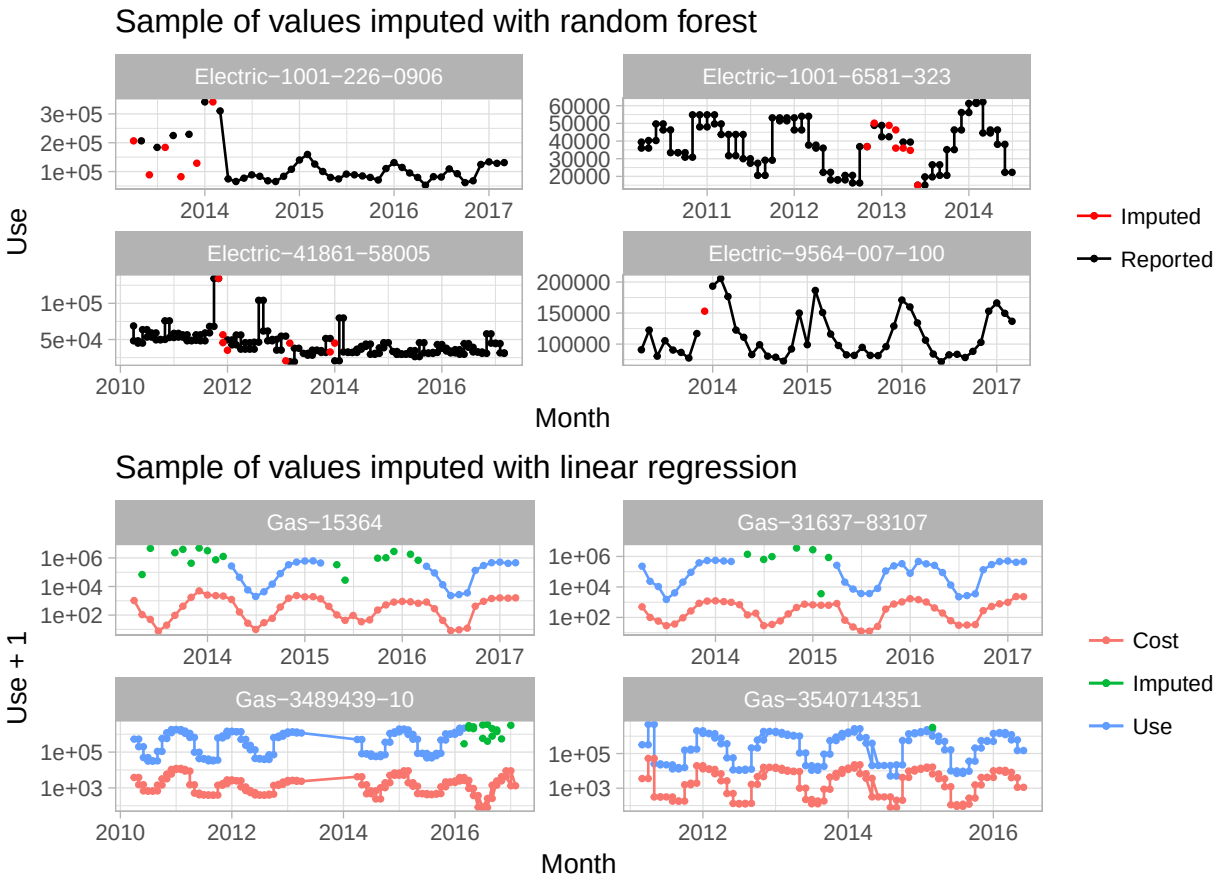
All remaining missing values have been flagged for imputation with random forest (25 values) or linear regression (672 values) models. Each imputation method is implemented separately using the `mice` package. The `AccountNumber` field is converted to a factor and included in the imputation predictor matrix – this leads to the creation of an individual fit for each account.

For accounts containing values missing use, demand, and cost, reported data was joined with the National Climatic Data Center weather data, as well as variable building data fields. When attempting random forest imputation, an error was returned related to computational singularity of the predictor matrix. The correlation of the matrix revealed that the minimum temperature `TMIN` and maximum temperature `TMAX` were the most highly correlated variables (as shown below); imputation was successful after removal of `TMAX`.



For accounts containing values missing use with reported cost, data was processed by the `mice` package to perform Bayesian linear regression. The variable `Share` created during the mapping of reported values to calendar months was removed from the predictor matrix prior to imputation.

With imputation complete, inspection of samples of imputed values were inspected to confirm validity:



As both imputation methods appear to have returned reasonable values, these values were used to fill the missing values in the consumption data set. The total use by calendar month (resulting from multiple bill cycles existing within a month) was calculated by multiplying the use by the share for each month and summing by month for each account number.

Data Management

Data was written to a Microsoft SQL Server database to allow for analysis in the database or in other environments outside of R. These include the following:

- Building data
  - Fixed metadata (`building_metadata`)
  - Variable data (`building_filingdata`)
- Weather data
  - Climate region lookup (`noaa_regions`)
  - Monthly weather data (`weather_monthly`)
- Consumption data
  - Data as reported, mapped to calendar months (`consumption_filingdata_asis`)
  - Data as reported, summed by calendar month (`consumption_filingdata_asis_summed`)
  - Imputed data (`consumption_filingdata_imputed`)
  - Imputed data, summed by calendar month (`consumption_filingdata_imputed_summed`)
  - The same table, as used for reporting (`consumption_filingdata_final`)

Performance Analysis

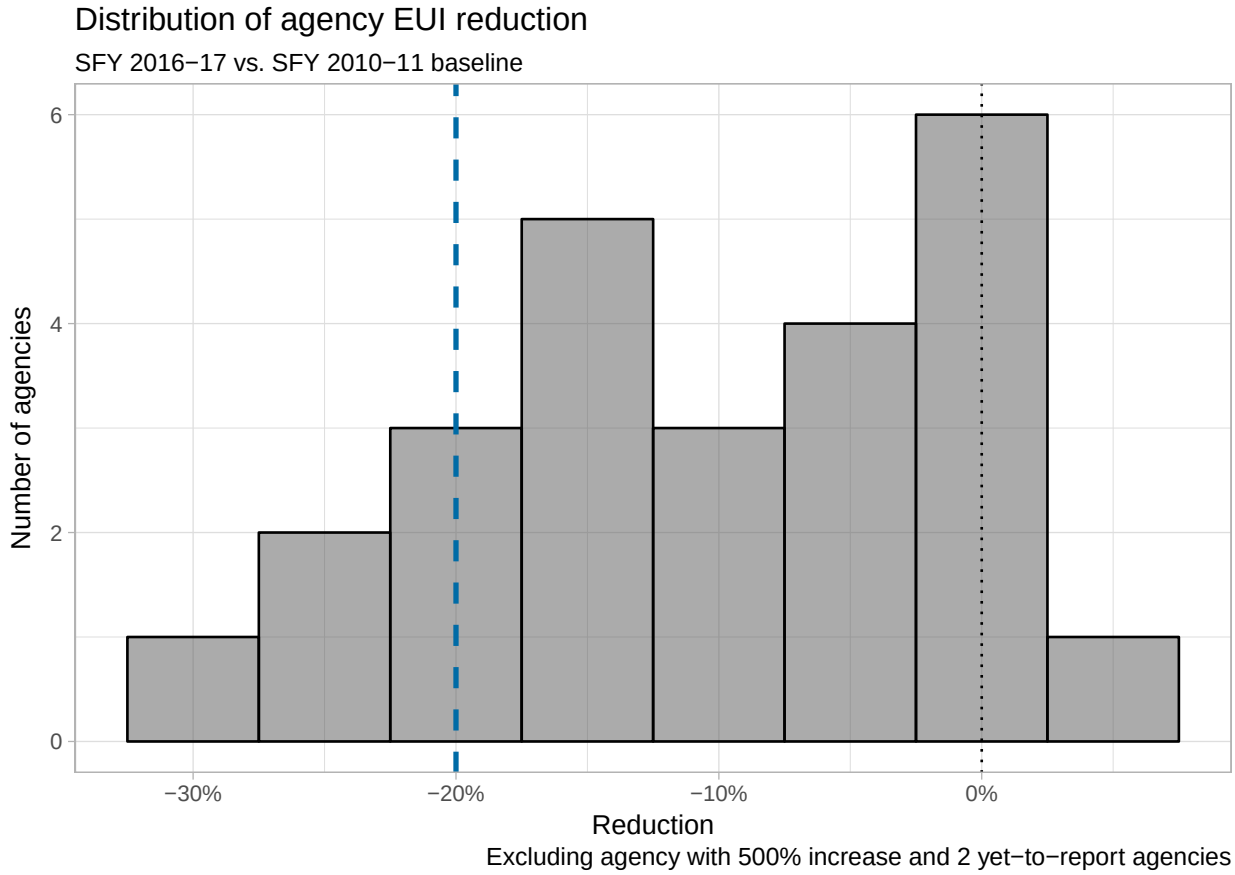
The data, populated with imputed values, was used to calculate performance towards the EUI reduction targets of Executive Order 88. This was initially attempted using the `dbplyr` and `dbplot` package to perform joins, visualizations, and calculations in the SQL Server database. The transit time for these calculations were long, and the size of the tables used (roughly 170,000 rows for consumption data) made in-memory calculation feasible, so in-database calculations were abandoned.

The overall statewide reduction in source EUI is presented in the table below:

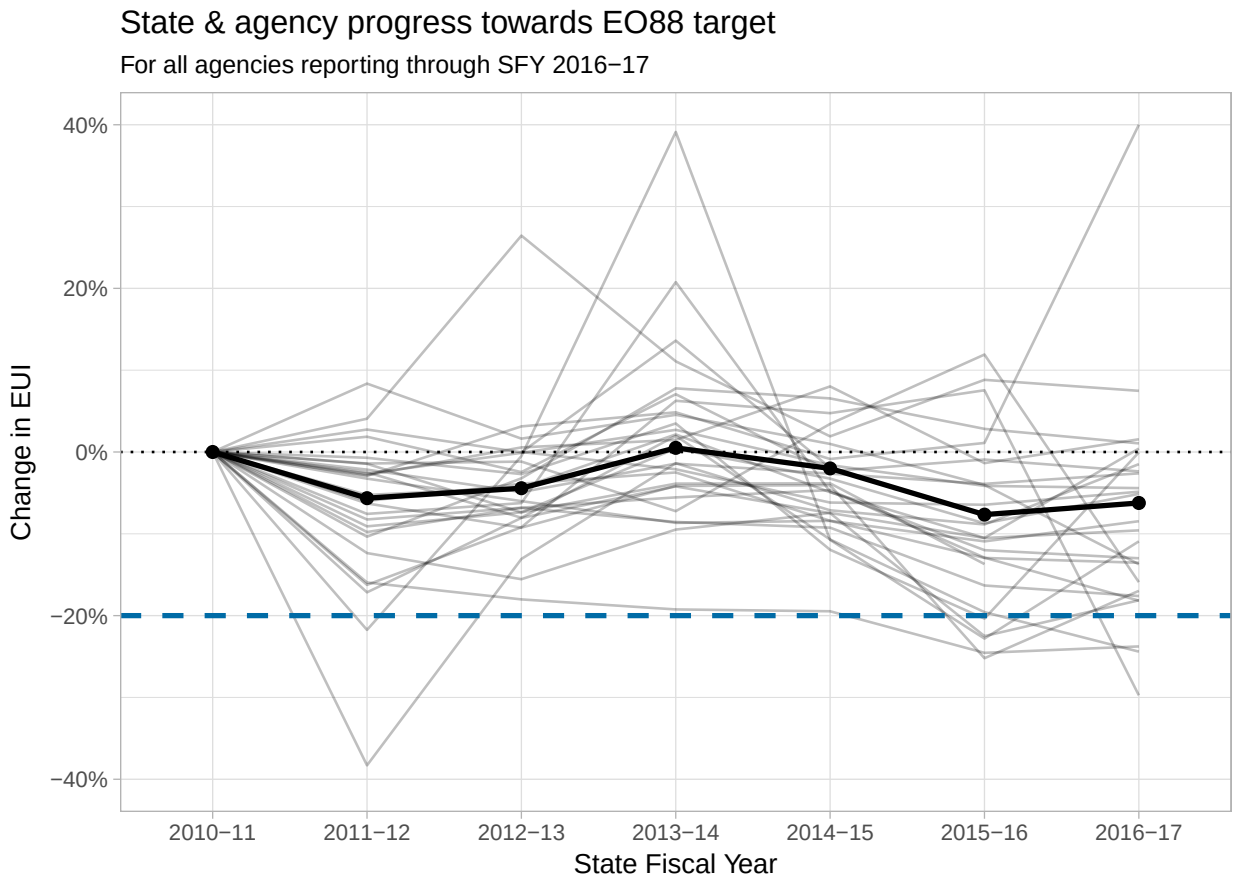
SFY	kBtu	sf	EUI	Reduction
2010-11	5.123e+10	208081544	246.2	0%
2011-12	4.937e+10	212527523	232.3	-5.64%
2012-13	5.038e+10	214100709	235.3	-4.43%
2013-14	5.356e+10	216441266	247.5	0.51%
2014-15	5.27e+10	218450906	241.3	-2.01%
2015-16	4.972e+10	218670194	227.4	-7.66%
2016-17	4.709e+10	2.04e+08	230.9	-6.23%

This reduction closely aligns with the (non weather-normalized) reduction previously calculated by NYPA's software vendor. To validate the overall calculation approach, the `eo88_summed` data set was compared to reports currently available in the NYEM platform based on the non-imputed data – all values were found to be within 1% of reported total source kBtu and total square footage.

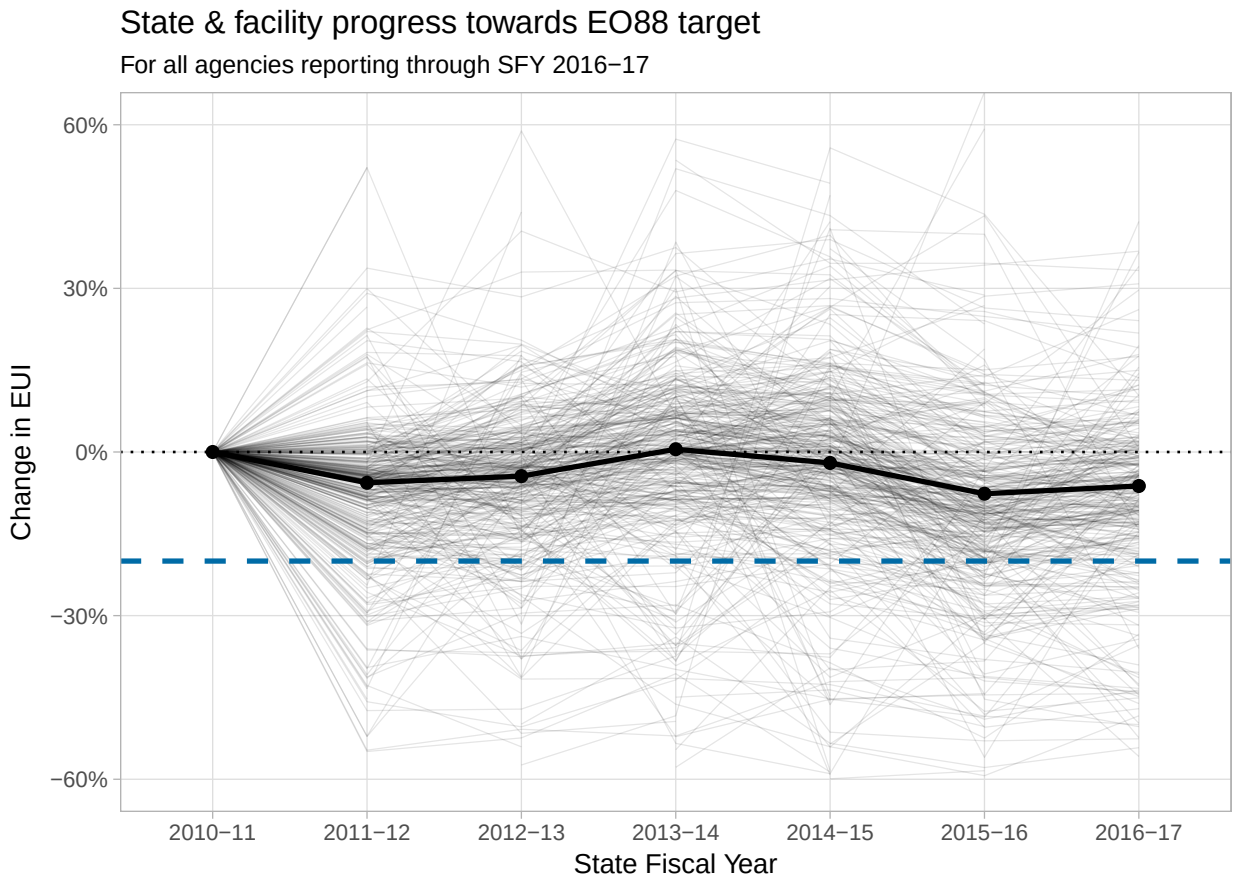
The distribution of agencies' reduction in SFY 2016-17 is shown below, excluding one agency showing a 500+% increase, which is excluded from further analyses:



Six agencies showed a reduction of at least the required 20%, with an additional 12 agencies showing a reduction from the baseline year. The remaining agencies show a near-zero change or an increase in energy intensity since the baseline year. Agency and state reduction by SFY were visualized jointly:



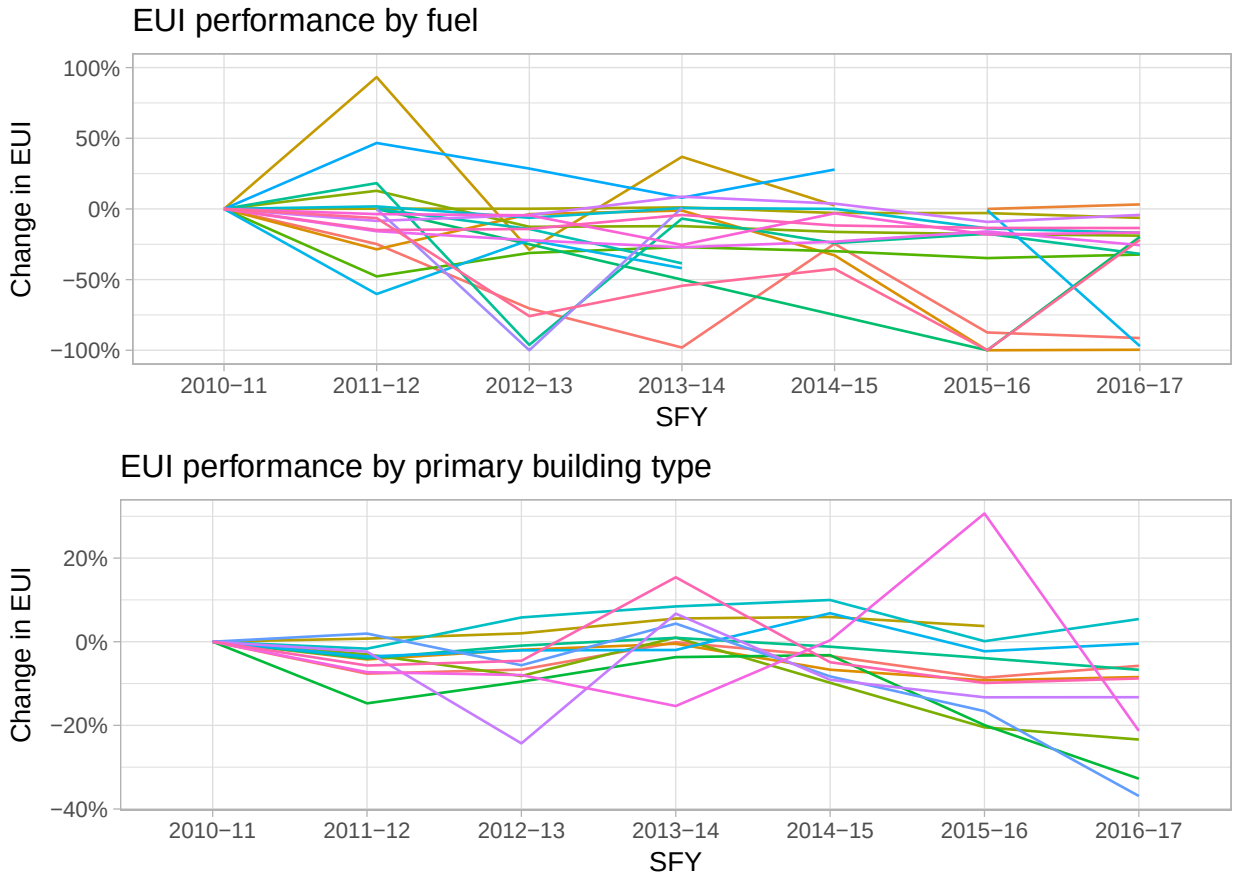
Finally, the EUI reduction of each facility was visualized alongside the state performance – while there is no facility-specific requirement for EUI reduction, this level of analysis may help the BuildSmart NY & NYEM teams identify facilities that may help improve agency or overall state reduction through energy efficiency efforts.



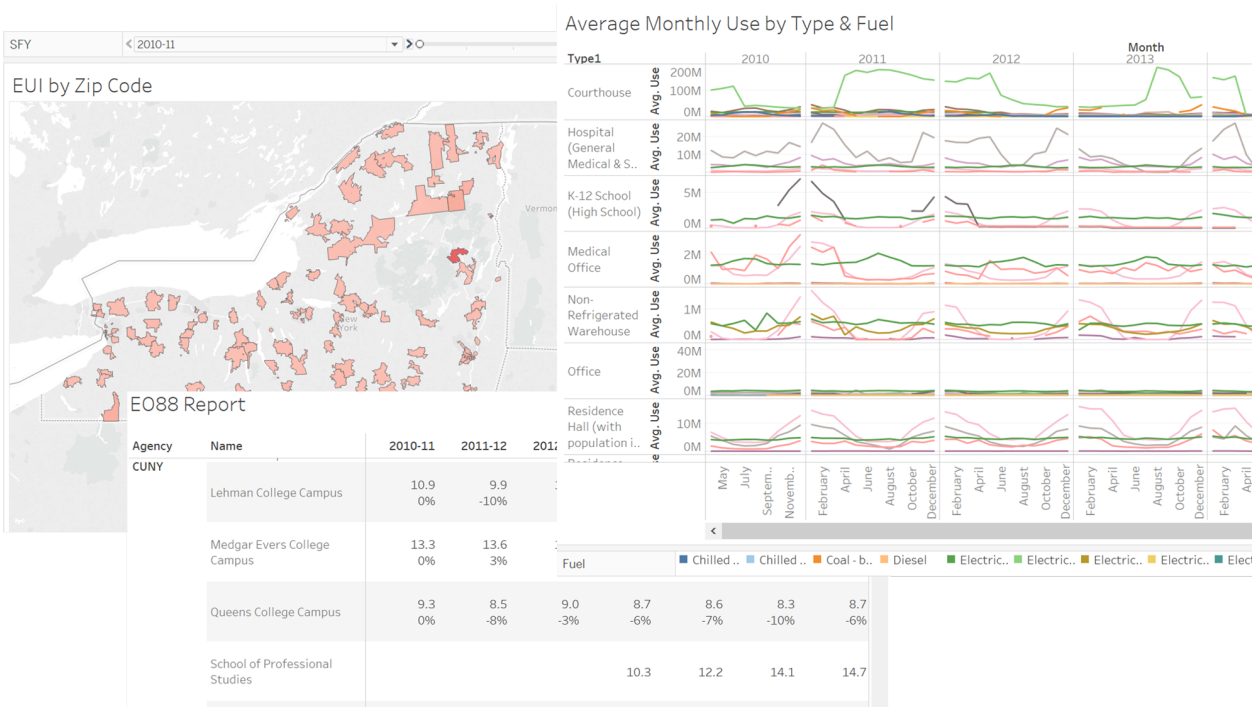
This plot shows that there is a far greater variance in the performance of individual facilities than there is in the performance of agencies – this is sensible, as the aggregation of individual facilities into an agency benchmark will naturally smooth some of the variance in facility performance. A concentration of lines around the overall state performance can also be observed, indicating two things: first, overall state performance may be a reflection of macro trends that can also be observed at a more granular level; and second, there is a large concentration of performance (at the state and facility level) around zero.

Analysis by Segments

As outlined in the problem statement, the analysis of EUI reduction by building type and/or fuel type will be useful to the BuildSmart NY and NYEM teams. To this end, EUI reduction by fuel of measure and primary building type were calculated and visualized:



Both plots show a large amount of noise and a significant number of different paths, making the information difficult to interpret. To allow greater transparency & interactivity, a Tableau workbook was created. The views within this workbook provide many ways of viewing the data: EO88 progress with drilldown capability into the agency & facility performance; total EUI by Zip code per SFY; average use by fuel & facility type per month. Example screenshots of these views are presented below:



Additional Considerations

The Tableau views shown above have values that do not tie exactly to those calculated in R – this is due to a difference in the mechanics of aggregation in both programs' functionalities. If the Tableau dashboard is to be deployed as the method of data consumption by business users in the BuildSmart NY and NYEM teams, these values must be reconciled. Additionally, many reported Zip codes are not recognized by Tableau; discussion with the NYEM team revealed that this is due to some State agencies having dedicated Zip codes, which must be related to physical Zip codes in order for informative mapping to be possible.

Filed consumption data at the time of this report included 26 of 28 Affected State Agencies. One of the remaining agencies has failed to report consumption in many years of EO88 and is not expected to file data. The filing data for the remaining agency for SFY 2016-17 must be processed, imputed (if necessary), and its performance calculated. This provides an opportunity to utilize the retroactive process illustrated by this project in a real-time workflow. If successful, the data ingestion & transformation utilized may be operationalized by NYPA using data integration tools to automate the process.

As mentioned in the *Data Cleansing: Initial Approach* section of this report, the data received was potentially modified through collaboration between the NYEM & BuildSmart NY teams, the software vendor, and reporting agencies from original reported values through methods not reflected in the reported data. It will be essential to understand any processes used to modify any historical reported data for future implementation of this reporting. Any changes made to data will need to be documented and accurately reflected in the data store of record to guarantee accuracy and defensibility of NYPA's EO88 reporting.

Discussion with business stakeholders following production of the results presented in this report identified challenges faced by some agencies in meeting EO88 goals; future work will aim to alleviate these challenges. The most universal of these challenges is the timeline to implement projects improving the energy efficiency (i.e. EUI) of buildings – many energy efficiency projects require a number of years to complete (due largely to the scope, schedule, cost, and procurement requirements involved), leading to the impact of these projects not registering before the end of SFY 2019-20. This may be addressed by crediting agencies for pledged energy efficiency reductions committed to prior to the end of the EO88 period. This incorporation will be challenging, but may be relatively straightforward for projects implemented by NYPA. Additionally, some agencies have constructed new facilities in response to increased need by their stakeholders – the nature of these additional facilities (for example, construction of technologically-advanced laboratories) can often have an adverse affect on total source EUI. A method for avoiding penalizing agencies for construction that supports Governor Cuomo's economic development goals may also be developed. Finally, facilities that underwent energy efficiency improvements prior to the start of EO88 may be challenged to find additional opportunities to reduce EUI by 20%. A method of identifying high-performing facilities in the baseline year (e.g. by comparing to Energy Star Portfolio Manager benchmarks) and excluding them from EO88 scoring may be implemented.

All R code used to perform the analyses explained in this report, as well as the source code for this report, can be found online at https://github.com/dsmilo/DATA698. The original data used are not provided for confidentiality concerns, but an anonymized version of the data is included in a compressed format.

Acknowledgements

Thank you to Honey Berk for her guidance in overcoming the myriad challenges encountered throughout the course of this project. Thank you also to Nate Anctil, Emilie Bolduc, David Benjamin, Gabe Cowles, Ryan Droege, Srini Pusuluri, Joe O'Connor, Julie Reardon, Prasad Surapaneni, and Godly Varghese for their assistance in the acquisition, understanding, and deployment of the data used for this project.

References

Augie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics.* R package version 2.3. https://CRAN.R-project.org/package=gridExtra

BuildSmart NY (2016). *EO88 2016 Annual Report.* https://www.nypa.gov/-/media/nypa/documents/document-library/operations/eo88-annualreport-2016.pdf

Cuomo, Governor Andrew M. (2012). *Executive Order No. 88: Directing State Agencies and Authorities to Improve the Energy Efficiency of State Buildings.* http://www.governor.ny.gov/news/no-88-directing-state-agencies-and-authorities-improve-energy-efficiency-state-buildings.

Daróczi, Gergely & Tsegelskyi, Roman (2017). *pander: An R 'Pandoc' Writer.* R package version 0.6.1. https://CRAN.R-project.org/package=pander

Energy Star Portfolio Manager (2013). *Technical Reference: Source Energy.* https://portfoliomanager.energystar.gov/pdf/reference/Source%20Energy.pdf

Energy Star Portfolio Manager (2015). *Technical Reference: Thermal Energy Conversions.* https://portfoliomanager.energystar.gov/pdf/reference/Thermal%20Conversions.pdf

Grolemund, Garrett & Wickham, Hadley (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Hester, Jim & Wickham, Hadley (2017). *odbc: Connect to ODBC Compatible Databases (using the DBI Interface).* R package version 1.1.1.9000. https://github.com/rstats-db/odbc

Kowarik, Alexander & Templ, Matthias (2016). "Imputation with the R Package VIM." *Journal of Statistical Software*, 74(7), 1-16. doi:10.18637/jss.v074.i07

Kuhn, Max (2017). *caret: Classification and Regression Training.* R package version 6.0-78. https://CRAN.R-project.org/package=caret

National Climatic Data Center (2017). *Divisional Data Select.* https://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp

New York Power Authority (2017). *BuildSmart NY.* https://www.nypa.gov/innovation/programs/buildsmart-ny

New York Power Authority (2017). *NY Energy Manager.* https://www.nypa.gov/services/digital-energy-services/ny-energy-manager

R Special Interest Group on Databases (R-SIG-DB), Wickham, Hadley & Müller, Kirill (2017). *DBI: R Database Interface.* R package version 0.7. https://CRAN.R-project.org/package=DBI

Ruiz, Edgar (2017). *dbplot: Simplifies Plotting Data Inside Databases.* R package version 0.1.1. https://CRAN.R-project.org/package=dbplot

van Buuren, Stef & Groothuis-Oudshoorn, Karin (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3), 1-67. http://www.jstatsoft.org/v45/i03/.

Wei, Taiyun & Simko, Viliam (2017). *R package "corrplot": Visualization of a Correlation Matrix (Version 0.84).* Available from https://github.com/taiyun/corrplot

Wickham, Hadley (2017). *scales: Scale Functions for Visualization.* R package version 0.5.0. https://CRAN.R-project.org/package=scales

Wickham, Hadley (2017). *tidyverse: Easily Install and Load the 'Tidyverse'.* R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Wickham, Hadley & Ruiz, Edgar (2017). *dbplyr: A 'dplyr' Back End for Databases.* R package version 1.1.0.9000. https://github.com/tidyverse/dbplyr