STA 141 Fall 2016

Project Report

Analysis of Annual Energy Consumption of UC Davis Campus

Group Member	Email	Student ID
Jiewei Chen	cjwchen@ucdavis.edu	999 494 235
Da Xue	xdxue@ucdavis.edu	999 450 295
Lingjun Ma	lma@ucdavis.edu	999 475 319

CONTENTS

Project division:

The majority of this project is done by all group members together, including coding, analysis, discussion, and drafting this report. Some parts are performed individually as follows:

- Lingjun Ma: generating map plots and information reflection on maps
- Da Xue: building a linear regression model to fill in missing values for "Annual Usage"
- Jiewei Chen: validating the predictive model and predicting energy usage for 2017

1 INTRODUCTION

There are currently 87 buildings on UC Davis campus whose energy consumption is a major component of the university operating cost. In recent years, sustainable development and clean energy have received increasing attention, and a transition in the energy resource types of the buildings on campus has been proposed and implemented in a certain number of buildings. Motivated by these considerations, the objective of this project is to analyze the past three years' energy usage of buildings on campus, investigate its influencing factors, and develop a statistical prediction model to predict the annual energy usage for a certain building. To achieve these purposes, data on the CEED website is utilized for statistical analysis and the model development. Information of the monthly usage of four types of energy resources (including chilled water, steam, natural gas, and electricity), footage, the year constructed, annual total energy usage and cost, and the functionality of each building is recorded, together with Davis weather report.

The rest of the report is organized as follows. Detailed data description is provided in Section 2 followed by data cleaning and pre-processing procedures. The correlations between different variables are analyzed in Section 3, qualitatively investigating relationships between the response variable, the annual energy usage, and the predictor variables are investigated. The predictive model is formulated and validated in Section 4 using linear regression and machine learning techniques. Finally, conclusions are given in Section 5.

2 DATA DESCRIPTION & DATA PRE-PROCESSING

2.1 DATA DESCRIPTION

The data obtained from the CEED website are separated into two categories. The first is a summary, including building names, abbreviations of the bulding names, year of construction, square footage, annual energy usage, annual energy cost, primary and secondary use of the buildings. Fig.2.1 shows the first five rows of the summary data, containing the information of "Academic Surge", "Activities and Recreation Center", etc., for example. The second part is a list of files containing the highest, lowest and average temperature, as well as the monthly usage of four different types of energy resources, namely, chilled water, steam, natural gas, and electricity. Each file includes the records of the past three years for a certain building. Fig4.2 shows the first few rows of the record for Mathematical Sciences Building as an example.

2.2 CHECK MISSING VALUE

Missing values are a major obstruct in our future model building. The numbers and frequencies of missing values are summarized in the table below.

	Construction year	Square footage	Annual cost	Annual usage
Number of NA's	2/87	1/87	60/87	56/87
Frequency of NA's	2.3%	1.1%	69.0%	64.6%

Below is a snapshot of the orinal data.

Name	Abbreviation	Year.Constructed	Square.Footage.FT.2.	Annual.Usage.KBTU.YEAR.	Annual.Cost...	Primary.Use	Secondary.Use
Academic Surge Building	academicsurge	1992	125810	NA	NA	LAB (66%)	OFFICE (19%)
Activities and Recreation Center	arc	2002	158120	18500040	230812	RECREATION (73%)	GENERAL (16%)
Advanced Materials Research Laboratory	amrl	2008	7560	NA	NA	LAB (71%)	GENERAL (19%)
Aggie Stadium	aggiestadium	2007	34538	NA	NA	ATHLETICS (72%)	GENERAL (27%)
Art Building	art	1966	57665	NA	NA	LAB (67%)	OFFICE (23%)

Figure 2.1: Snapshot of original data

2.3 DATA PRE-PROCESSING

To better utilize the data, pre-processing is first performed in the following steps.

- Calculate "year of service" defined as how long a building has been put into use, and add a column into the original data set.
- Based on the primary use, regroup the functionality of the buildings into 5 categories: "classroom", "community", "lab", "office", and "others", using regular expressions.
- Use geocode to extract location information of each building and add into the original data set. Manual correction was involved for missing or incorrect values.

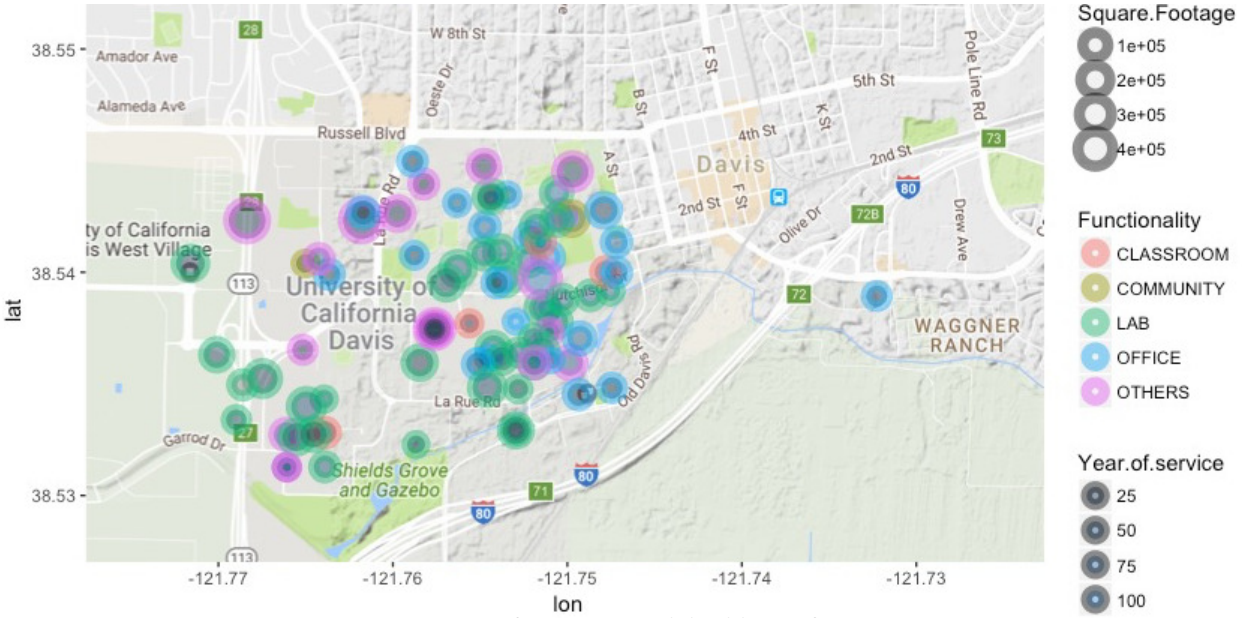


Figure 2.2: Map of UC Davis with building information

The results of the above data pre-processing and analysis are graphically represented in the map in Fig.2.2. Each point represents a building, where the circle size indicates the square footage, the color of the outter circle represents its functionality, and the color of the inner circle reflects its year of service. It can be observed that most buildings are used as labs and offices. A few newly constructed buildings can be identified whose colors of the inner circle are relatively darker. There is no significant difference in the square footage of different buidlings.

3 QUALITATIVE RELATIONSHIPS AMONG VARIABLES

Before constructing the predictive model of annual energy usage, a brief analysis on the relationship among predictor and response variables is performed. Some major observations are presented in the rest of this section. Other results are attached in the Appendix.

3.1 EFFECTS OF PREDICTOR VARIABLES ON USAGE OF FOUR ENERGY RESOURCES

As described in Section2, the major response variable is the annual energy usage, which is made up of four energy resources. The purpose of this section is to explore the impact of predictor variables on the usage of the four energy resources.

3.1.1 EFFECT OF FUNCTION ON NATURAL GAS

Refer to Fig.3.1(3), there are clear natural clusters on natural gas consumption separated by building functionalities. High natural gas consumption majorly comes from offices and low natural gas consumption is from communities, while classrooms and labs do not use natual gas as an energy resource. There is a trend that the natural gas consumption decreases as temperature goes up

observed for both communities and offices. This trend can be explained that the purpose of natural gas is heating, which has a lower demand when temperature is high.

3.1.2 Effect of temperature on four energy resources

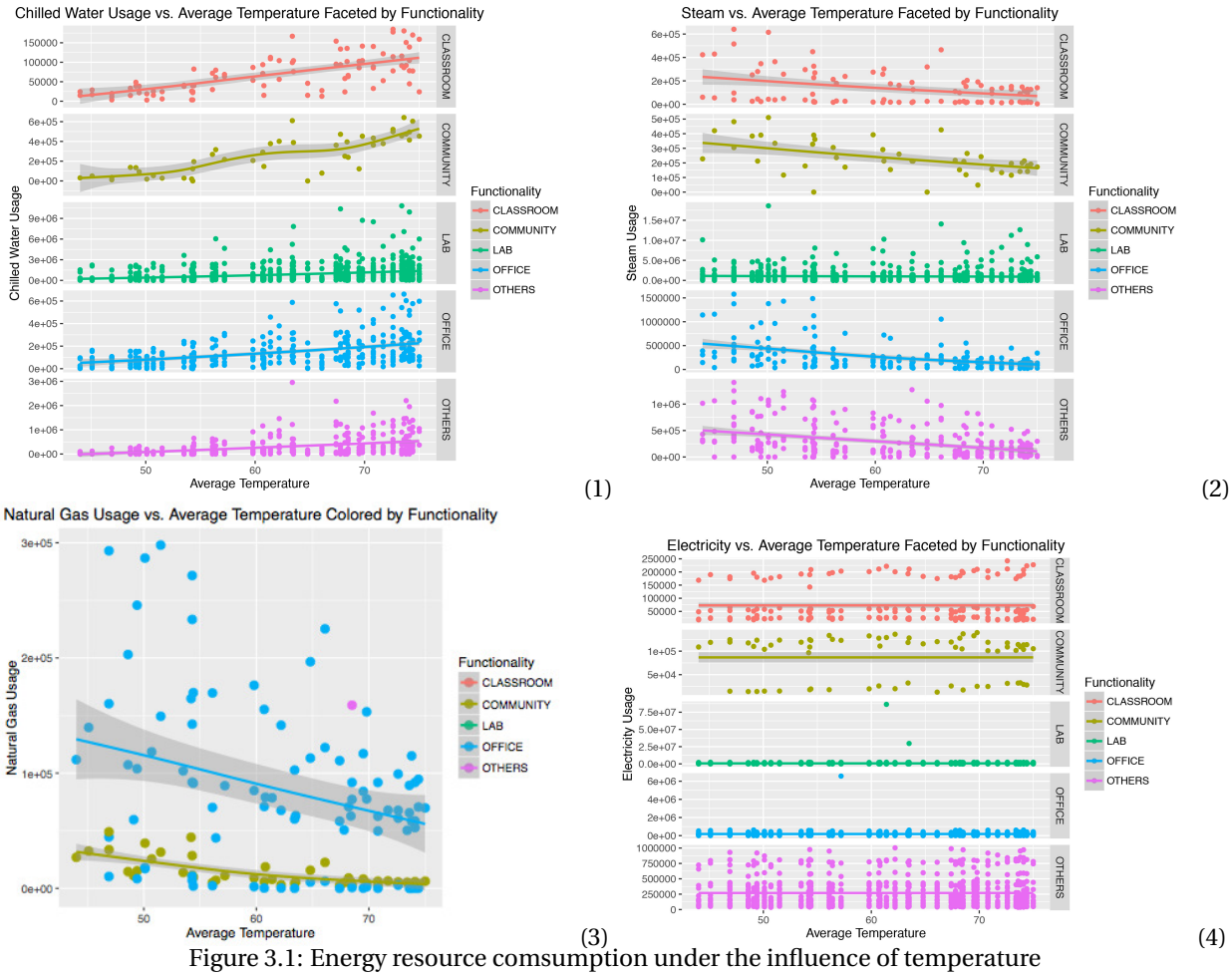


Figure 3.1: Energy resource comsumption under the influence of temperature

Besides the consumption of natural gas, the impact of temperature on the consumption of other types of energy resources is shown in Fig.3.1. It can be seen that as temperture increases, the usage of chilled water increases and steam increases decreases, while the consumption of electricity does not change across a wide range of temperature. These observations imply that steam and natural gas are supplementary energy resources used mainly for heating. Buildings choose either one as the heating method. Chilled water and natural gas or steam, on the other hand, are complementary, as it is used for cooling. Electricity consumption is not affected by temperature, which might not be used for heating or cooling.

3.1.3 Effect of seasonality on four energy resources

Fig.3.2 shows the energy consumption of the four resources over the past three years. Each line in the plots represents the record of one building, and the number of lines in each plot denotes the number of buildings that use the specific energy resource. It reveals that the usage of chilled water, steam and natural gas exhibit cyclic behavior when plotted against time. However, the consumption of electricity varied little with time, which is consistent with the findings of temperature effects.

3.1.4 Effect of foot square on four energy resources

Fig.3.3 shows that medium range footage buildings have wider chilled water and steam usage ranges. Buildings whose footage is relatively small, tend to use natural gas instead of steam. All the buildings

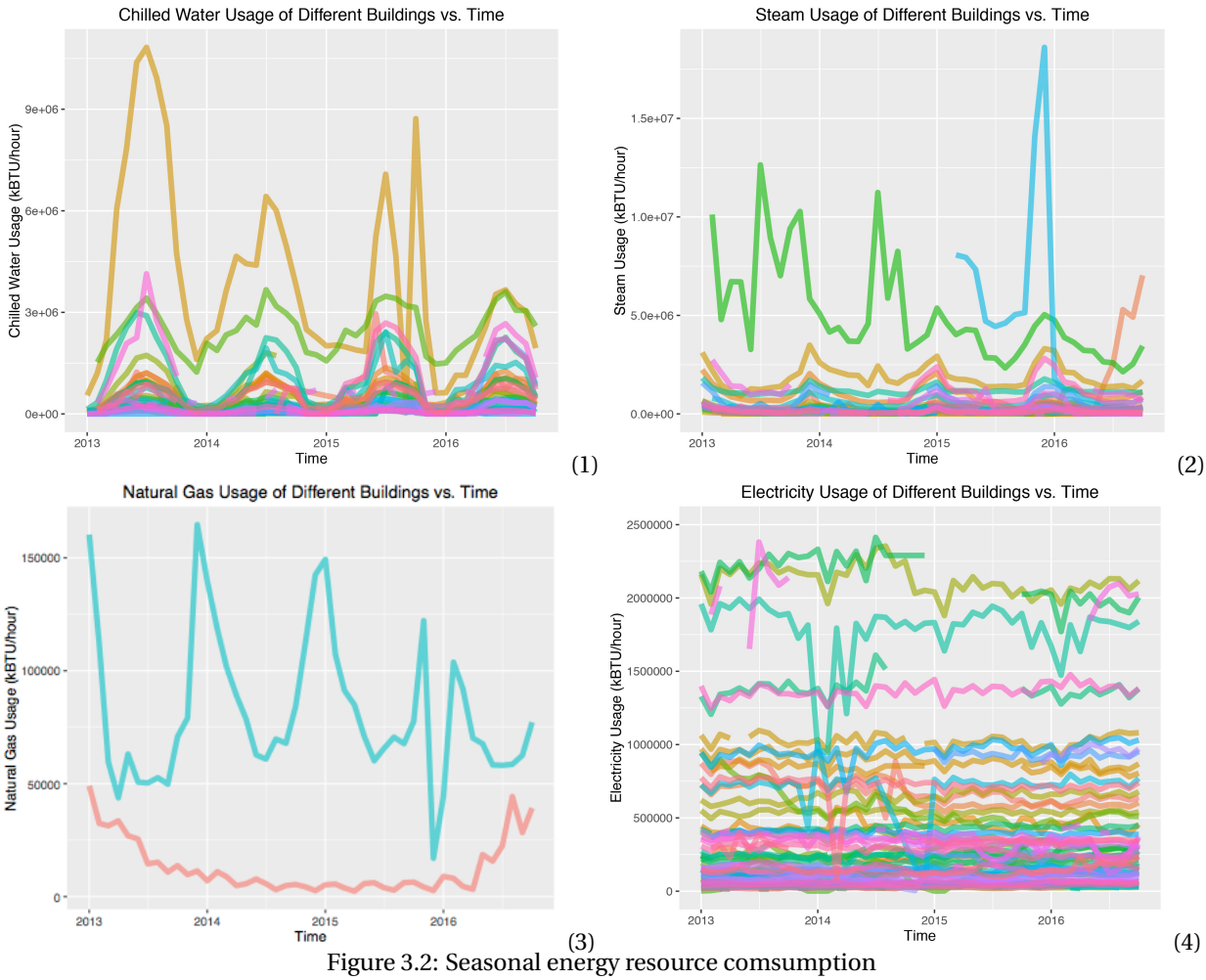


Figure 3.2: Seasonal energy resource comsumption

use electricity regardless of footage. The colors in Fig.3.3 correspond to building functionalities, same as Fig.3.1. The legends in the rest of this section are omitted for better visualization.

3.2 Relationship between Annual Usage and Predictor Variables

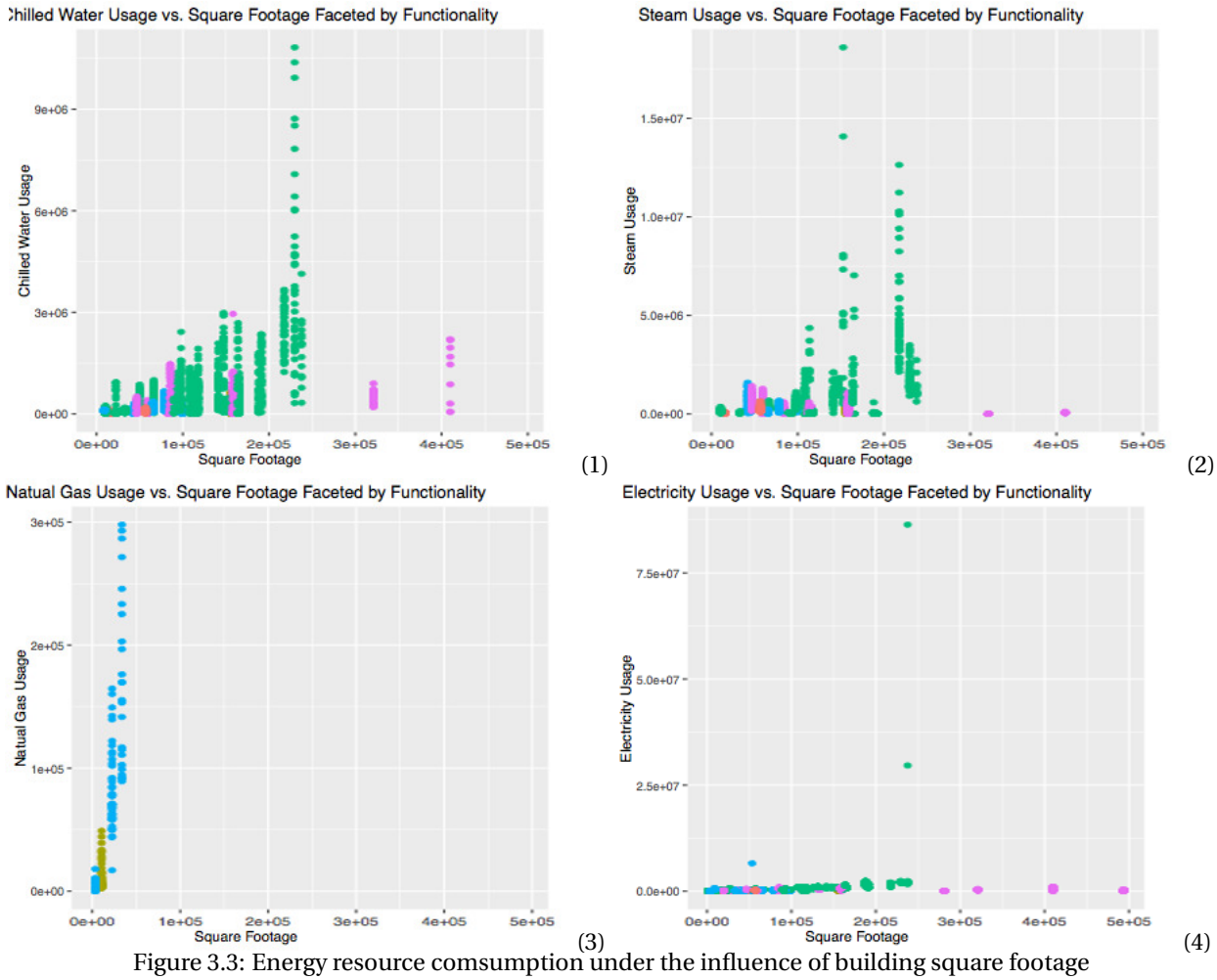
The main objective of this project is to develop a predictive model for the annual energy usage of a building. After exploring the relationship between the predictor variables and the four specific energy resources, in this section, the relationship between the annual usage and predictor variables is investigated. Referring to Fig.3.4, the median of annual usage of labs is much higher than those of buildings with other functionalities, leading to the conclusion that labs is a major contributor to the annual usage. Median of annual usage of office is slightly higher than those of classroom and community. Meanwhile, it is observed that lab has a steeper increase of annual usage as square footage goes up, and thus lab and square footage might be major contributors to annual usage.

4 Predictive Model of Annual Energy Usage

After pre-processing all the variables in our data set, and quanlitatively analyzing the relationship among variables, a linear model is built for the prediction of the annual usage in this section.

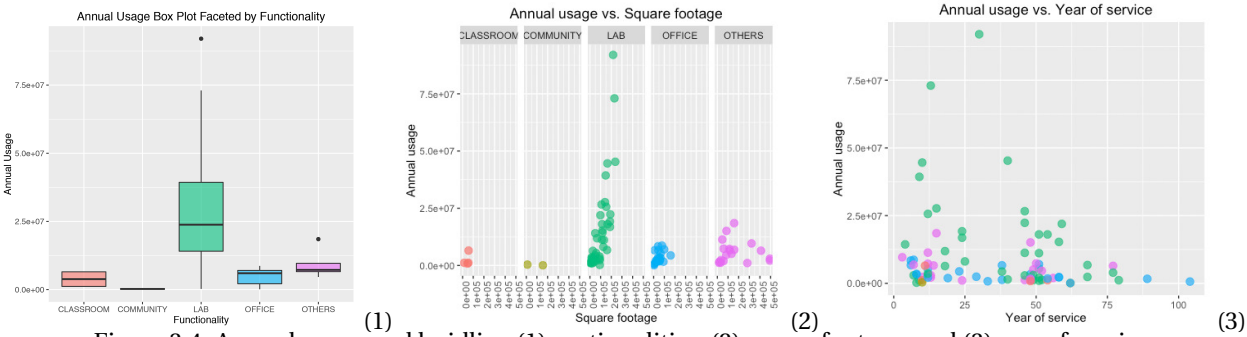
4.1 Split Training and Validation Dataset

To build the model, first the dataset is split into two parts. There are in total 87 buildings on our current UC Davis campus. The two buildings constructed in 2016 are removed from the dataset, due to the lack of data from a full year cycle. Within the remaining 85 buildings, 80% of them are set to be the training data, and the rest 20% are set to be the validation data.



(1) (2) (3) (4)

Figure 3.3: Energy resource comsumption under the influence of building square footage



(1) (2) (3)

Figure 3.4: Annual usage and buidling (1)unctionalities, (2)square footage, and (3)year of service.

4.2 Fill Annual Usage Missing Values

As discussed in Section2.3, there are 56 missing values in the response variable "Annual Usage", making up over 64% of the data. The high proportion of missing values, or the limited availability of the data, resists our ability to develop the predictive model. It is desired that the missing values of the "Annual Usage" could be filled using available data based on a regression model, instead of simply removing the missing values. It is intuitive to assume that the annual energy usage of a certain building is the summation of the energy usage of each individual energy resource. To this end, the monthly energy usage record of each building is obtained from the CEED website for the past three years. Below is an example of the first five records of MSB.

To predict the annual usage for buildings with missing values, simple machine learning method is used. The subset of data in the predictive model training data set containing 27 buildings is initially randomly split into two sub-data sets, the training set with 22 buildings information, and the validation data set with 5 buildings.

DateTime	CHILLED WATER (kBTU/hour)	STEAM (kBTU/hour)	NATURAL GAS (kBTU/hour)	ELECTRICITY (kBTU/hour)	Temperature Range (low)	Temperature Range (high)	Average Temperature
2013-01-01 00:00:00	91446	217655		221219	33.3	61.9	46.9
2013-02-01 00:00:00	91375	145286		196674	31.2	63	44
2013-03-01 00:00:00	136803	81404		207735	34.3	68.3	49.1
2013-04-01 00:00:00	204001	61692		205976	38.6	75.7	56.4
2013-05-01 00:00:00	260631	28564		213551	46	91.7	63.5

Figure 4.1: Snapshot of the monthly energy usage of the Mathematical Sciences Building

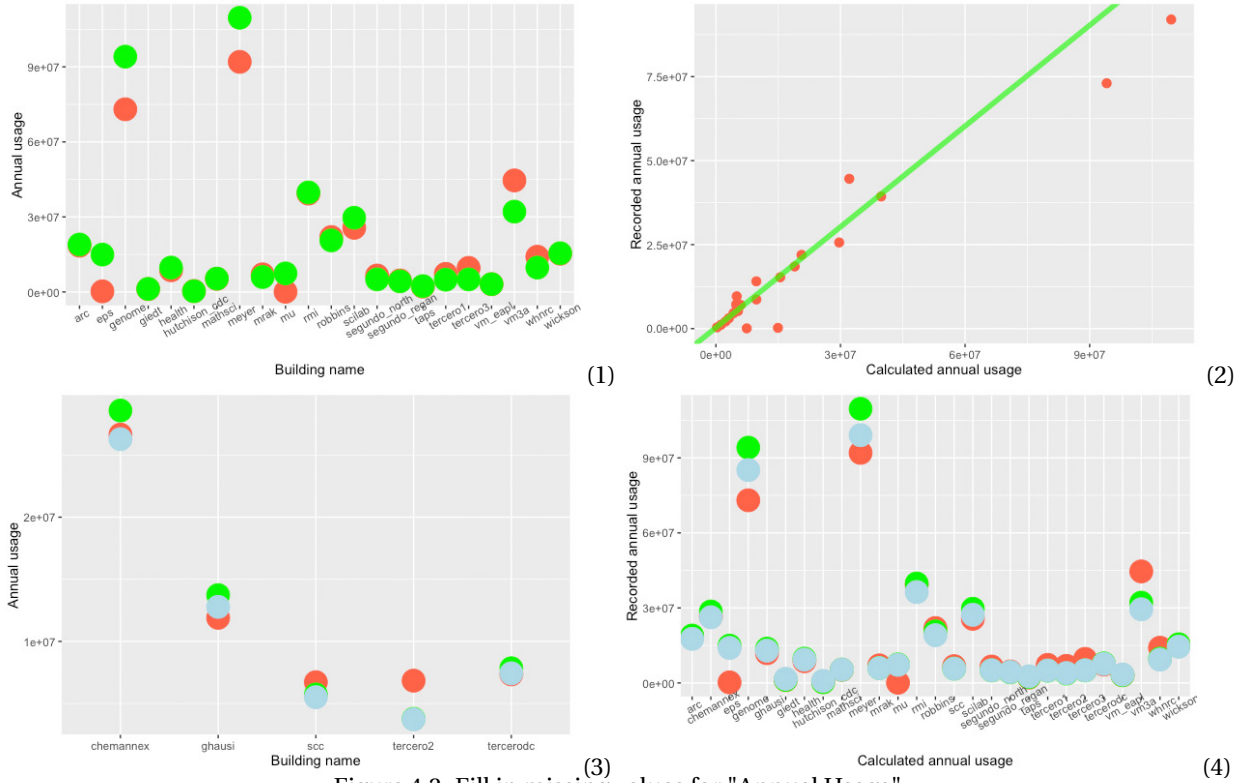


Figure 4.2: Fill in missing values for "Annual Usage"

Fig.4.2(1) shows the comparison of the annual usage recorded (in red) with the summation of the annual usage (in green) of the four types of energy resources. The annual usage values for each building is relatively close except for a few outliers. A linear model is then regreesed for annual usage on the usage of the individual energy resource. Fig.4.2(2) shows that the regression line captures the majority of the data. The linear regression model is then test on the validation data set containing 5 buildings, and Fig.4.2(3) shows the results of model fitting. The light blue dots representing the fitted values match the recorded data (in red) much better than the simple summations (in green).

The validation of the model confirms the linear relationship between the annual energy usage and the usage of the individual energy resource. The training and validation data sets are then combined together to fit the final model for the annual usage. The results are shown in Fig.4.2(4) and is used to fill in the missing values for the rest buildings in the data set.

4.3 Linear Model of Annual usage on Predictor variables

A linear model is built by using "All Subsets" method with AIC criterion. The best model chosen is the linear model with "Annual Usage" as response variable, and "Year of Service", "Square Footage", "Functionality-Lab" as predictor variables. Functionality is a variable with five classes, including "Classroom", "Community", "Lab", "Office" and "Others". In the best model we selected, the functionality has been reduced to "Functionality-Lab".

4.4 Diagnosis and Model Validation

The diagnostic plot of the model built upon the training data set is shown in Fig.4.4. It can be seen that the linearity assumption holds, since the red line is almost horizontal at around zero in the

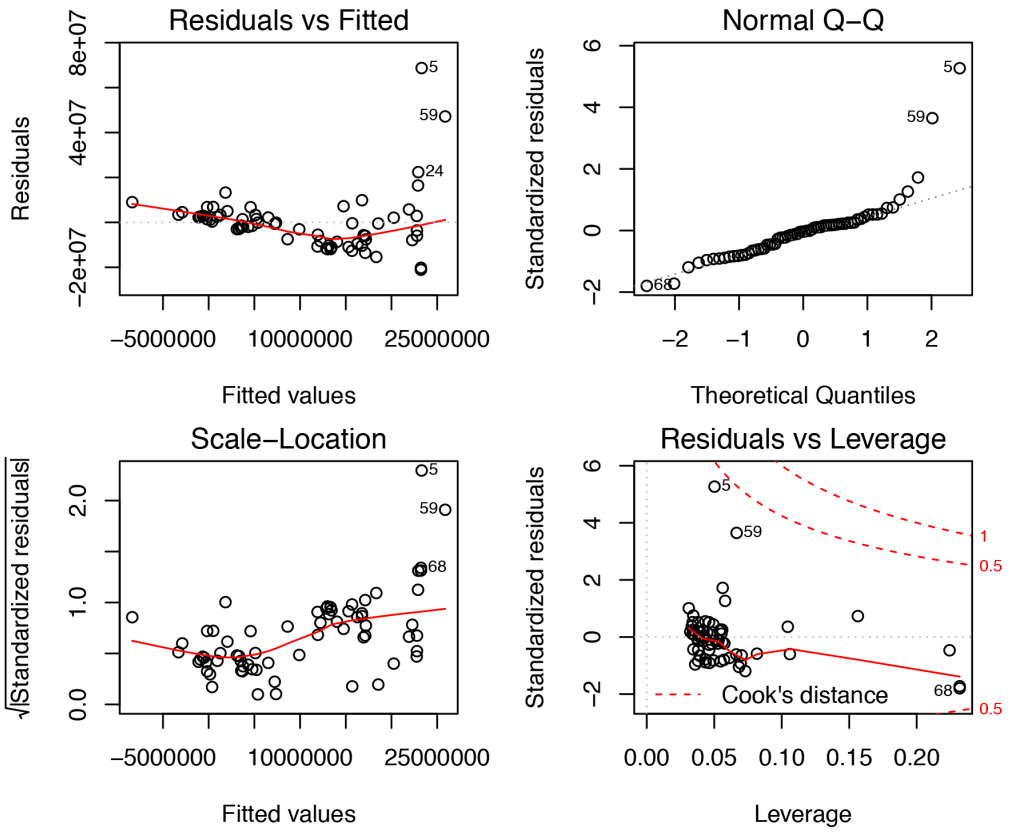


Figure 4.3: Diagnostic plots for the linear regression model

the "Residuals vs Fitted" plot. However, the normality assumption does not hold very well. The "Normal Q-Q" plot shows that the right-side of the distribution has heavier tail than the normal distribution. However, since we are not doing any hypothesis test, this violation will not affect our final model. The equal variance assumption is ok, but not strictly followed. There is no outlier in our data set, considering that all points are located inside the red dashed line with Cook's distance of 0.5, shown in "Residuals vs. Leverage". Generally speaking, this model is a good regression model.

The validation of our model is conducted both internally and externally.
The predicted and actual "Annual Usage" are plotted in Fig.4.4 for each building. It can be observed that our model works well in the prediction of most of the buildings, as the "red" and "green" points are almost overlapped on each other for both training and testing data set.

4.5 Prediction of Annual Usage of Each Building in 2017

Finally, same linear regression is carried out on the whole data set. The diagnosis plots (omitted here, but included in the appendix), show that all assumptions hold. This model is used for the prediction of the annual usage of each building on 2017. The results are shown in Fig.4.5.

5 Conclusion

In this report, data cleaning and pre-processing methods were involved. Statistical and graphical analysis on the annual energy consumption, including four categories, i.e., chilled water, steam, natural gas, and electricity of 87 buildings on UC Davis campus was performed by evaluating the influence of various factors. Interesting results were observed, for example, steam and natural gas are supplementary energy resources mainly used for heating; electricity is generally used without the influence of temperature, seasonality and square footage, which might be used for any purpose besides heating or cooling, such as lighting and instrument operation.

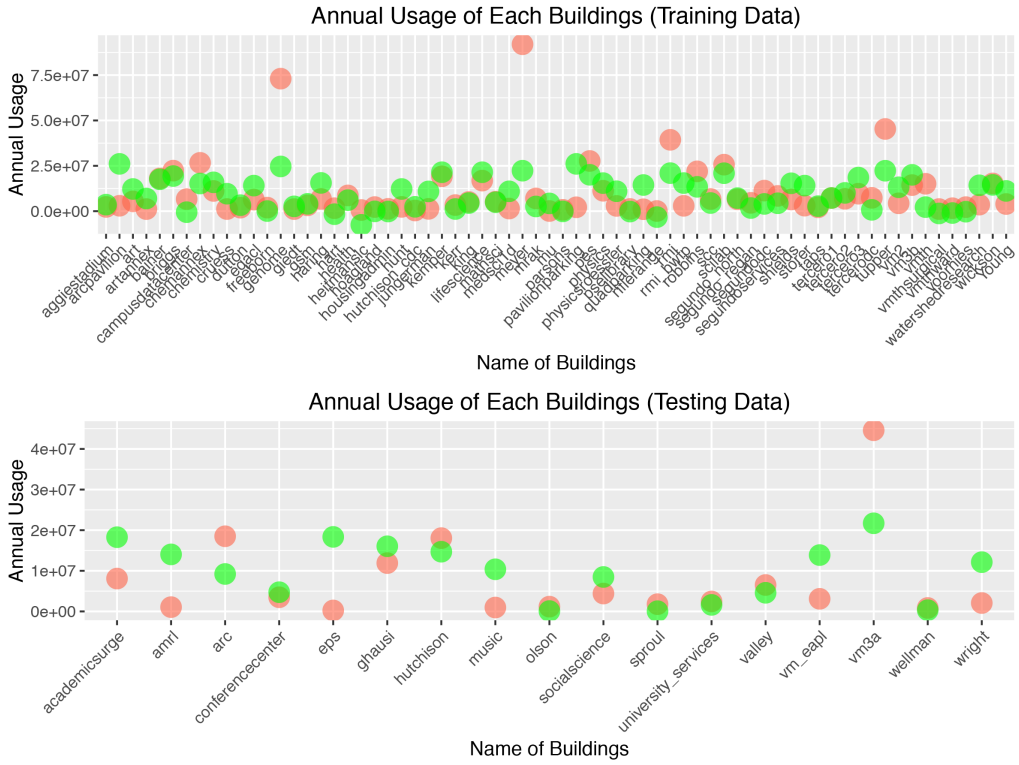


Figure 4.4: Model validation

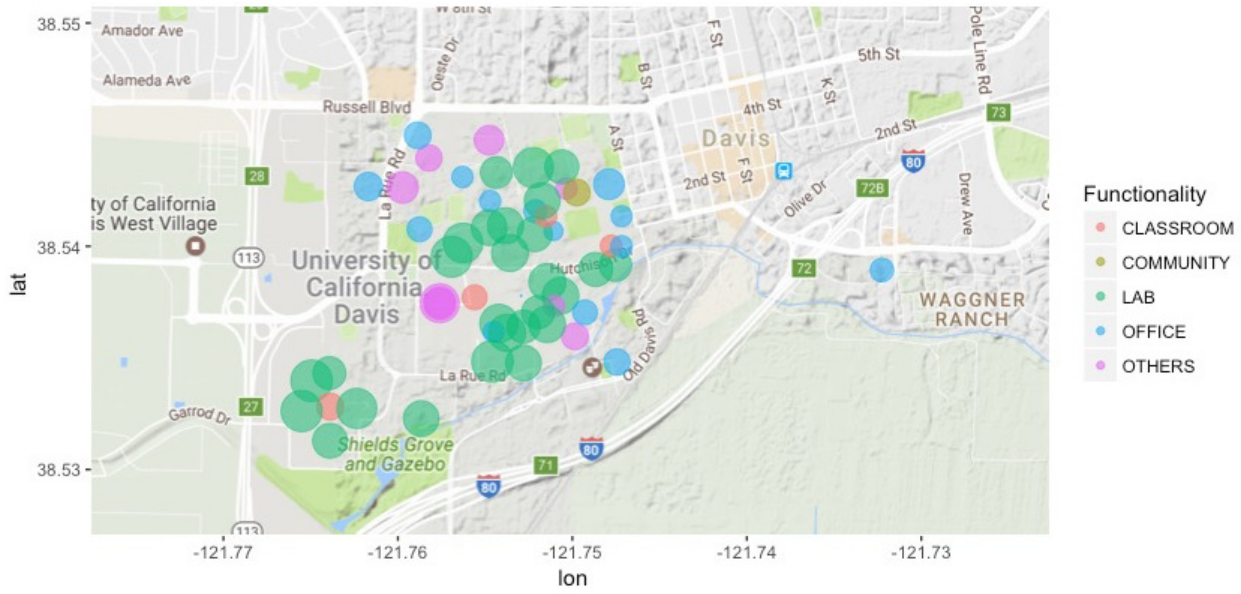


Figure 4.5: Annual energy usage prediction in 2017

A linear regression model was built with "Annual Usage" as the response variable, and "Year of Service", "Square Footage", "Functionality-Lab" as predictor variables, and can be used to predict the annual usage of a certain building in future years. The model is built and selected with AIC criterion on 80% of the data and is validated by the rest 20% data both internally and externally. The results from the diagnostic plots meet the criteria, confirming it to be a good fitting model of the buildings.