Electricity Consumption in Chicago

The electric power market is evolving rapidly, constantly reshaping our long-lasting expectations for future electric consumption pattern. In this study, we set out to obtain a fresh perspective on the current electricity consumption situation by focusing on Chicago, an industrial metropolitan with an advanced city-wide electric system. We want to know what variables correlate with electricity consumption and how they're correlated. We hypothesize that season, building type have influence over electricity consumption. We firstly investigate the differences in electricity consumption across seasons and among different building types using EDA and ANOVA. After discovering that both variables affect electricity consumption. We then utilize BIC plot to uncover significant factors that affect electricity consumption. Lastly, we conduct linear regression modeling using R and Stata in effort to find the most comprehensive model that helps us gain insight on electricity consumption development in populous cities around the world.

Part 1: Introduction

The electric power landscape is evolving at a rapid pace, reshaping our longstanding expectations for our energy future. There is no doubt that nowadays, individuals, and commercial companies become increasingly reliant on electric power. The purpose of this study is to obtain a comprehensive and fresh understanding on the current electricity consumption situation in Chicago and to construct a reliable model to predict electricity consumption based on our data from the Chicago government website. In this research, we will focus on the questions: "what is the relationship between electricity consumption versus season? Are there hidden factors that affect the aforementioned relationship?" We hypothesize that season and building type exert influence on electricity consumption. Electricity consumption is the highest for commercial building during summer. In addition to season and building type, we also hypothesize that factors such as occupied units, building age, building area and etc. will also affect the aforementioned relationship.

Part 2: Exploratory Data Analysis

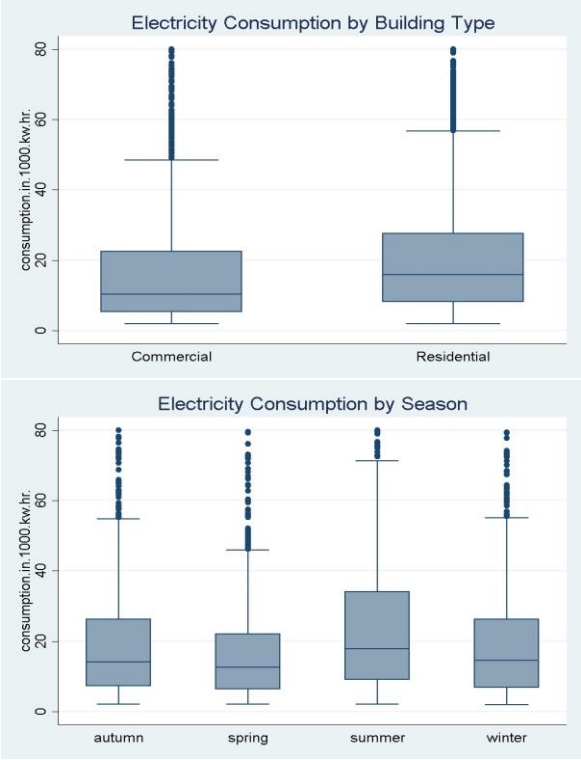


Figure 1

Our original dataset contains 60,000 observations with some influential outliers that prevent us from obtaining meaningful results. Therefore, we decided to trim our data by removing the top 5% and the bottom 5% of the original data to obtain a comprehensive yet reasonable dataset. After the trimming process, we discovered that with 50,000 observations, almost every variable of the study becomes significant. Hence, we decided to utilize random sampling to extract 5,000 observations from the trimmed dataset. By integrating the trimming method and randomized sampling process, we obtained a concise dataset while preserving the integrity of the original dataset. To answer the question posed and to test our hypothesis, we would like to do some exploratory data analysis first as we find it important to have a broad perspective of our data. (You can refer to Appendix 2 and Appendix 3 for description of variables and numerical summary to gain more insights into the findings of this EDA segment.)

To begin our exploratory analysis, we made boxplots of electricity consumption individually by season and by building type to observe their distributions (Figure 1). We observed that mean values of electricity consumption across seasons fluctuate around 19 thousand kWh while the interquartile varies noticeably across seasons. Specifically, summer has the largest mean value, which contradicts to our original hypothesis. (All four seasons' largest outliers seem to align on the same horizon, which was believed to be caused by the trimming process.) From the graphics, the means and the distributions of electricity consumption look distinctive for each season, suggesting that season has an effect on the electricity consumption. As for building type, it seems from the graph that there is still a difference between the mean electricity consumption of residential and commercial building types with commercial building types having lower mean

value. We believe that such case may be a result of the correlation between building type and electricity consumption.

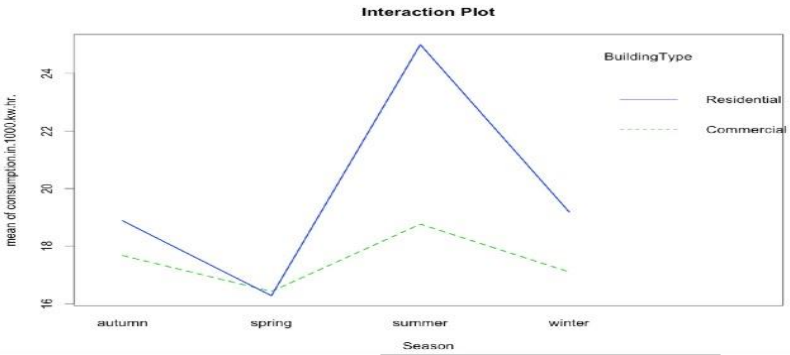


Figure 2: interaction plot

After exploring the aforementioned pairwise data, we decided to test if the conclusions drawn from EDA are consistent for building type throughout season and vice versa by making an interaction plot (Figure 2). The lines in the plot are not paralleled, especially in spring where they crossed each other. Besides, the difference between the two lines in summer is noticeably larger than that of other seasons'. Therefore, we think that there is an interaction between building type and season. We are aware that EDA needed statistical reasoning to verify the conclusions drawn from it. Therefore, we will apply ANOVA test in the next section to see if season and building type have effects on electricity consumption from a statistical perspective.

Part 3: Modeling and inferential Analysis

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
building.type	1	4.944e+09	4.944e+09	20.837	5.12e-06	***
season	3	3.539e+10	1.180e+10	49.716	< 2e-16	***
building.type:season	3	4.815e+09	1.605e+09	6.764	0.00015	***
Residuals	4992	1.185e+12	2.373e+08			

Figure 3: ANOVA analysis

In Figure 3, the p-value of building type and season shows that they are statistically significant variables. In addition, the p-value of the interaction term is 0.00015, suggesting that the interaction between building type and season. Consequently, we concluded that building type, season and their interaction all have effect on electricity consumption.

While our initial research question focused on the effect of building type and season on electricity consumption, we realized that there may be some other variables that affect electricity consumption in addition to these two variables. In order to create a comprehensive and accurate model to predict electricity consumption, we decided observe more variables and check their statistical significance by running multiple linear regression models on them. From the correlation matrix (Appendix 4), we observed correlations between electricity consumption versus population, total units, and occupied units. Therefore, we first built a full model with aforementioned variables (namely total square feet, population, total units, building age and occupied units) and thier corresponding interaction terms. To avoid over-fitting, we selected our model using "backward" stepwise regression and "bic" to produce our initial model. From the BIC plot (Appendix 5), our model achieves smallest value at the number of 4 variables. Therefore we kept the most significant 4 variables. However, after checking for collinearity using vif values, we found total units and occupied units are highly collinear with each other. Consequently, variable occupied units is eliminated.

Now, all the predictors are statistically significant. From the final model (Figure 4), we can tell that residential buildings tend to consume more electricity than commercial buildings, due to the

positive coefficient. The coefficient of summer is positive and largest, suggesting that electricity consumption is the highest in summer. In contrast, consumption is the lowest in spring. Besides building type and season, our model shows that total square feet, total units and building age also have significant effect on electricity consumption. Total units and building age are negative correlated with electricity consumption, while total square feet is positive. Holding other variables constant, 1 unit increase in Total sqft leads to 0.944 units increase in electricity consumption; 1 unit increase in Total Units corresponds with 15.81 units decrease in electricity consumption, and 1 unit increase in building age leads to 53.43 units decrease in electricity consumption.

Electricity Consumption = 8844 + 0.944Total sqft - 15.81Total Units - 53.43Building Age + Factor(Season) + Factor(Building Type) + Factor(Building Type)*Factor(Season)

Coefficients	0.944	-15.81	-53.43	-566.7	4316	1551	1339	-2476	2510	-1917
Predictor	sqft	units	age	spr	sum	wntr	res	spr:res	sum:res	wntr:res

Figure 4: Final model summary

For our final model, all the VIFs are around 1 (Appendix 7), which means that there is no correlation among the predictors. As for checking linear regression assumptions (Appendix 8), we found that most of points on the Normal QQ-plot fall approximately on a straight line indicating no deviation from normality, and fitted vs residual plot shows no serious deviation from constant variance. Though, r^2 value is not very high (0.5088), this is the best model we attempted with respect to the variables in our dataset.

4 Conclusion and implications

From our study, we realize that the electricity consumption is correlated with various variables, both categorical and numerical, in a more complicated way then we have imagined. After analyzing and diagnosing our comprehensive statistical model, we can draw the following conclusions: Firstly, both season and building type have a significant influence on electricity consumption in Chicago. Specifically, the electricity consumption is the highest for summer and residential building type. Secondly, Area of the building, building age, and total units also affect electricity consumption. The relationship between electricity consumption versus building area is positive while the relationship between electricity consumption versus total units or building age is negative.

Since some conclusions of this study are contrary to our original intuition, we hope our study can provide some insights on how to design future electric power grid and transmit electric power in an efficient and adequate way for industrial and populous cities like Chicago. From our standpoint, both the season and characteristics of buildings (total area, total units, building type and building age) should be taken into account while designing electric power grid or planning electric power transmission. For our future work, we want to look at electricity consumption in big cities outside the U.S. to obtain a global view of the electric power market and also to compare cities with similar explanatory variables but different electricity consumption levels in order to find out more hidden variables so that our model can be furthermore modified to accommodate cities with different characteristics.

Appendix:

1. Data Sources:

- Chicago 2010 Energy Consumption

2. Description of the Variables :

- Electricity Consumption (response variable): denotes the amount of electricity consumed per season, measured in 1,000 kilowatt-hour.
- Building Type: this is a categorical data, residential and commercial. In this study, we use number 0 to denote residential building type and number 1 to denote commercial building type.
- Season: this is a categorical data, autumn, spring, summer and winter.
- Total.sqft: denotes the entire area of the corresponding building measured in terms of square feet.
- Population: denotes the number of people residing in the corresponding building with 1 unit equals 1 person for this variable.
- Total.units: denotes for the number of apartments/condos/houses for the residential building type and the number of architectural units specified for the commercial building type.
- Occupied.units: denotes the number of units that was leased to tenants or owned by owners.
- Building age: demotes the age of the corresponding building measured in terms of year.

3. Numerical Summaries:

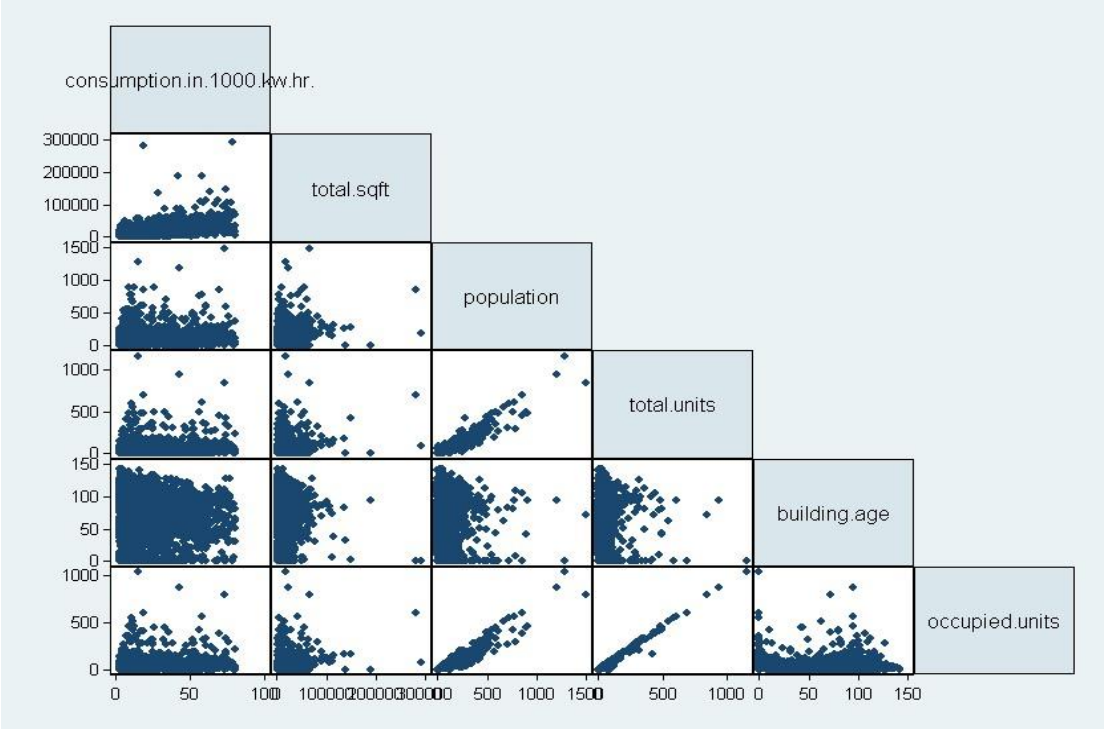
Building Type	mean	median
Residential	19.860	15.930
Commercial	17.460	10.400

Season	mean	median
spring	16.310	12.580
summer	23.680	17.850
fall	18.610	14.120
winter	18.740	14.560

min	median	mean	max	sd
2.02	14.58	19.33	80	15.68

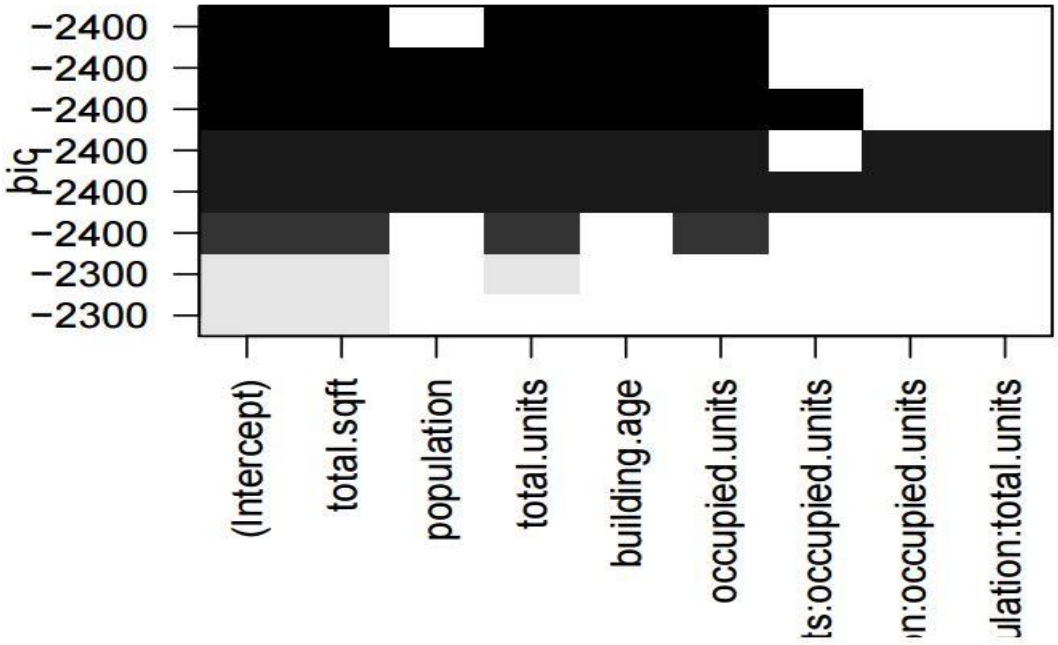
3: summary of electricity consumption (unit: 1,000 kWh)

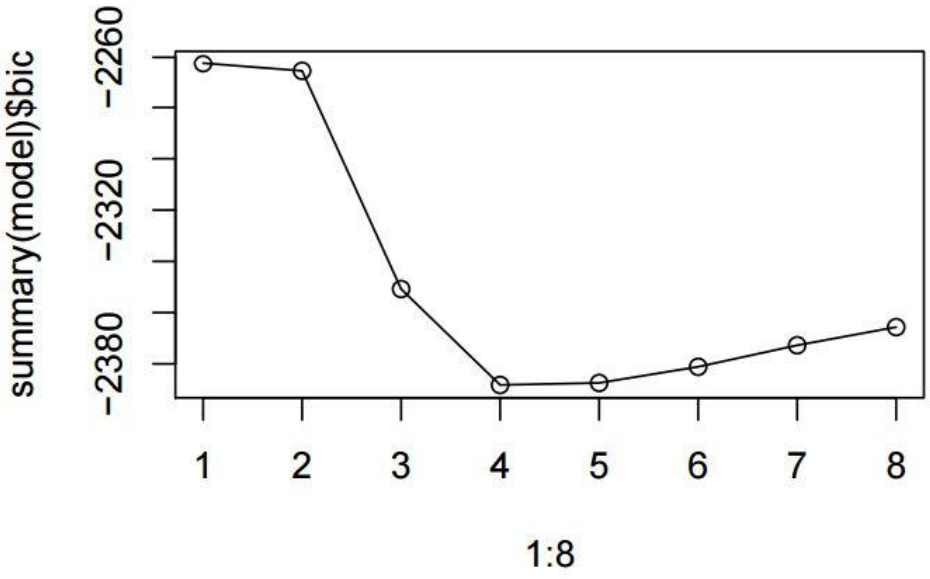
4. Correlation plot:



4: correlation matrix

5. BIC plot





6. Complete linear regression model summary:

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      8.844e+03  7.385e+02  11.976  < 2e-16 ***
total.sqft                                       9.444e-01  1.401e-02  67.414  < 2e-16 ***
building.age                                    -5.343e+01  4.830e+00 -11.062  < 2e-16 ***
total.units                                     -1.581e+01  3.313e+00  -4.770 1.89e-06 ***
factor(season)spring                            -5.667e+02  9.234e+02  -0.614   0.5395
factor(season)summer                             4.316e+03  9.465e+02   4.561 5.23e-06 ***
factor(season)winter                             1.551e+03  9.423e+02   1.646   0.0999 .
factor(building.type)Residential                 1.339e+03  7.612e+02   1.759   0.0787 .
factor(season)spring:factor(building.type)Residential -2.476e+03  1.050e+03  -2.357   0.0184 *
factor(season)summer:factor(building.type)Residential  2.510e+03  1.068e+03   2.350   0.0188 *
factor(season)winter:factor(building.type)Residential -1.917e+03  1.064e+03  -1.802   0.0716 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10930 on 4974 degrees of freedom
Multiple R-squared:  0.5088,    Adjusted R-squared:  0.5079
F-statistic: 515.3 on 10 and 4974 DF,  p-value: < 2.2e-16
```

7. Collinearity analysis:

```
##                          GVIF Df GVIF^(1/(2*Df))
## total.sqft           1.075483  1        1.037055
## building.age         1.083809  1        1.041061
## total.units          1.077713  1        1.038130
## factor(season)       1.004501  3        1.000749
## factor(building.type) 1.107167  1        1.052220
```

8. Checking linear regression model assumptions: