INTRODUCTION TO STATISTICAL COMPUTING

ASSIGNMENT 2 - Project Report

Comprehensive Analysis of Chicago Energy Benchmarking Data: Insights into Energy Consumption and GHG Emissions.

Group 9

Astha Bathla - S3999096

Atia Uzma - S4057268

Rashmi Chandrakant Borgave - S4064369

Kawal Kaur - S4060459

Snigdha Mathur - S4017572

Table of Contents

Introduction

This report aims to examine total energy usage in Chicago with the goal of further enhancing efficiency and reducing emissions in the city. It helps in understanding the patterns of energy consumption and the emissions caused, while focusing on natural gas use, electricity use, as well as greenhouse gas emissions across diverse building types and their ages. Our objective is to analyse trends between different types of energy consumptions and how they are affected by building attributes in a number of years in Chicago.

The open sourced dataset utilized in this report is sourced from data.gov: https://catalog.data.gov/dataset/chicago-energy-benchmarking.

Methods

In our analysis of Chicago's energy benchmarking data, we employed statistical methods and visualization tools to assess energy consumption and efficiency using SAS software.

Initial steps involved using PROC SQL for data manipulation and PROC SGPLOT for creating visualizations, focusing on natural gas consumption by property type and energy star scores across different ratings. Time series analysis highlighted energy consumption trends from 2014 to 2022, while regression analysis explored the relationship between building size and electricity usage, confirming the model's relevance through statistical tests.

We also conducted univariate analysis to examine the distribution of greenhouse gas emissions, employing normality tests to confirm the data's non-normal distribution. The Chi-Square test, conducted with PROC FREQ, analyzed the association between building age and energy ratings, while correlation analysis quantified the relationship between natural gas usage and greenhouse gas emissions. TTest was used to perform hypothesis testing on energy star score.

Analysis and Results

1. Top 5 property types by natural gas consumption

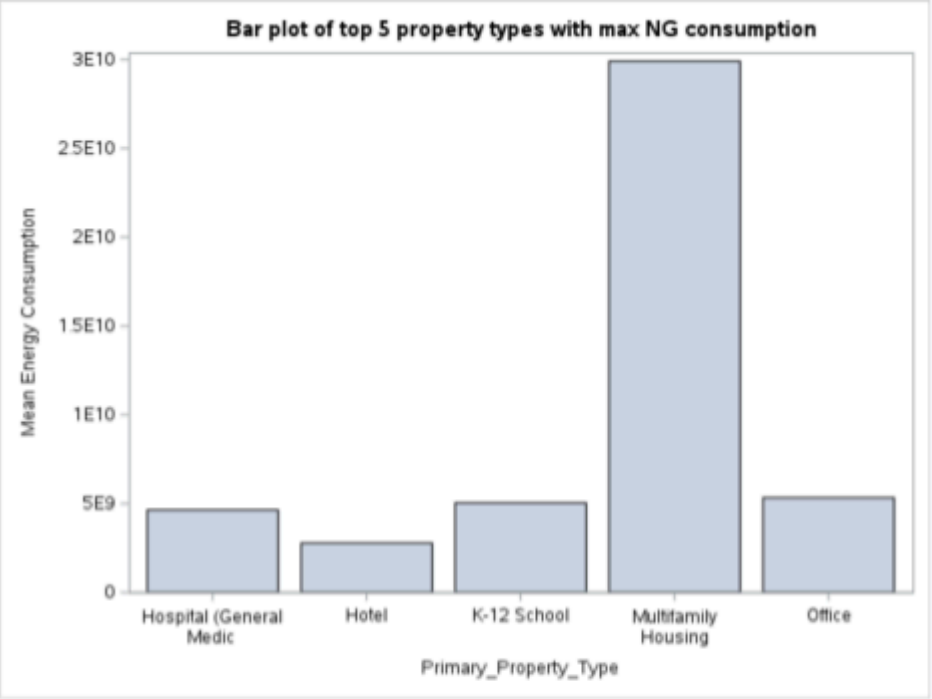


Figure 1: Top Natural Gas Consumption by Property Type Based on Reporting status as 'Submitted Data'

The bar chart illustrates the mean natural gas consumption across various property types based on submitted data. Multifamily housing exhibits the highest natural gas consumption, with usage nearing 30 billion kBtu.

This is significantly higher than the other property types. Hospitals and K-12 schools follow, each with approximately 5 billion kBtu in natural gas consumption. Offices and hotels show the lowest consumption among the top five, with values under 5 billion kBtu. This indicates that multifamily housing units are the largest consumers of natural gas, suggesting a potential area for targeted energy efficiency improvements to reduce overall consumption and associated costs.

2. Total Energy Consumption corresponding to Chicago Energy Rating

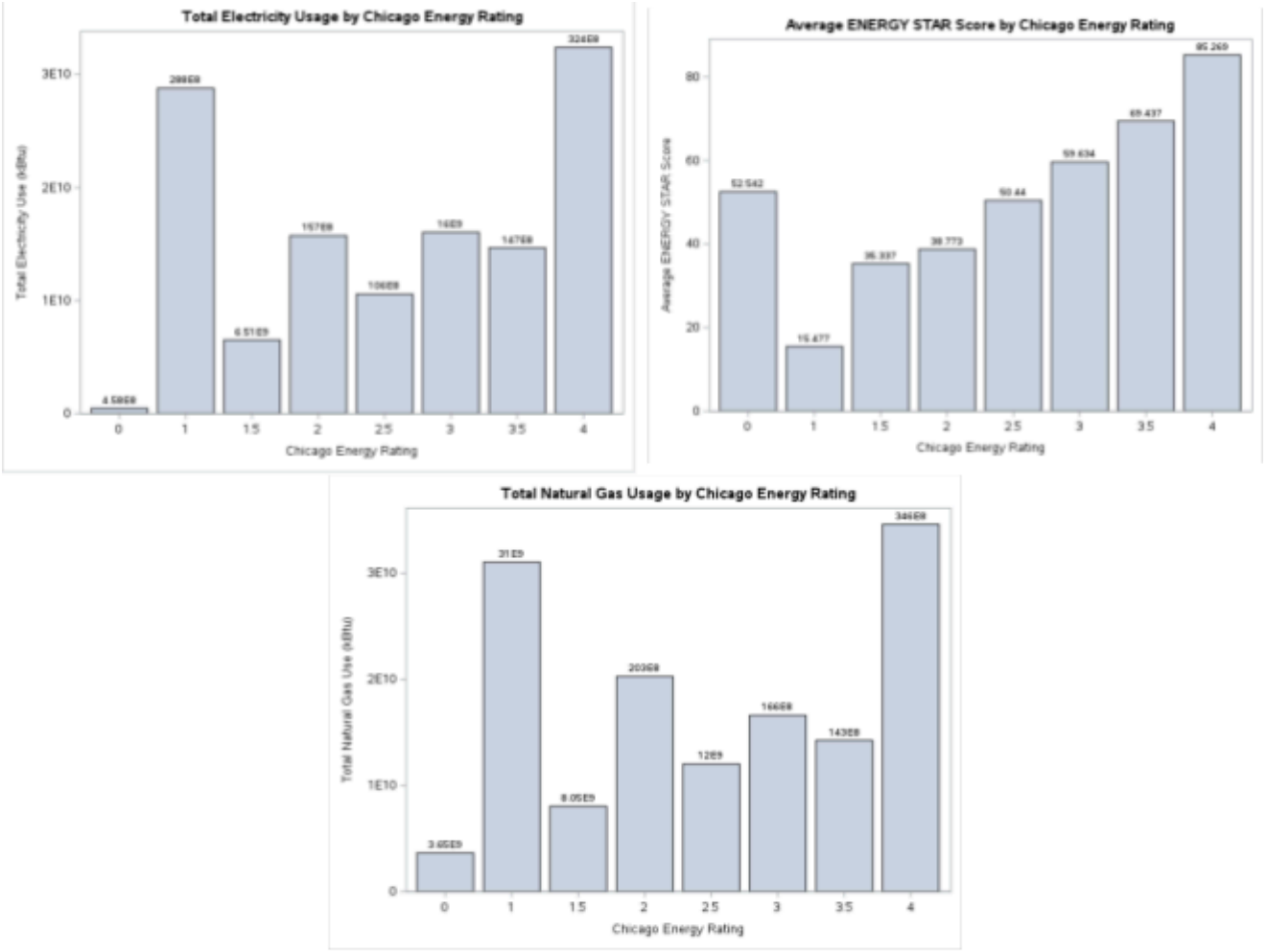


Figure 2: Analysis of buildings of different energy ratings compared in terms of electricity and gas usage, and overall Energy Star scores

Figure 2 consists of 3 graphs depicting total electricity usage, total natural gas usage and average energy star score by Chicago Energy Rating. The graphs on the average energy star score and total natural gas and electricity usage by Chicago Energy Rating illustrate a complex relationship between energy efficiency and consumption.

While the energy star score generally increases with higher energy ratings, indicating improved energy efficiency, the consumption of natural gas and electricity does not follow a linear pattern. This suggests that factors such as building size or operational intensity might significantly impact energy consumption. Notably, despite a sharp increase in energy star scores from rating 1 to 4, energy usage peaks at the highest ratings, underscoring the interplay between achieving high energy efficiency and managing consumption in larger or

more energy-intensive buildings. These insights could guide energy policies and efficiency strategies to optimize both energy usage.

3. Trend of Electricity Consumption over years

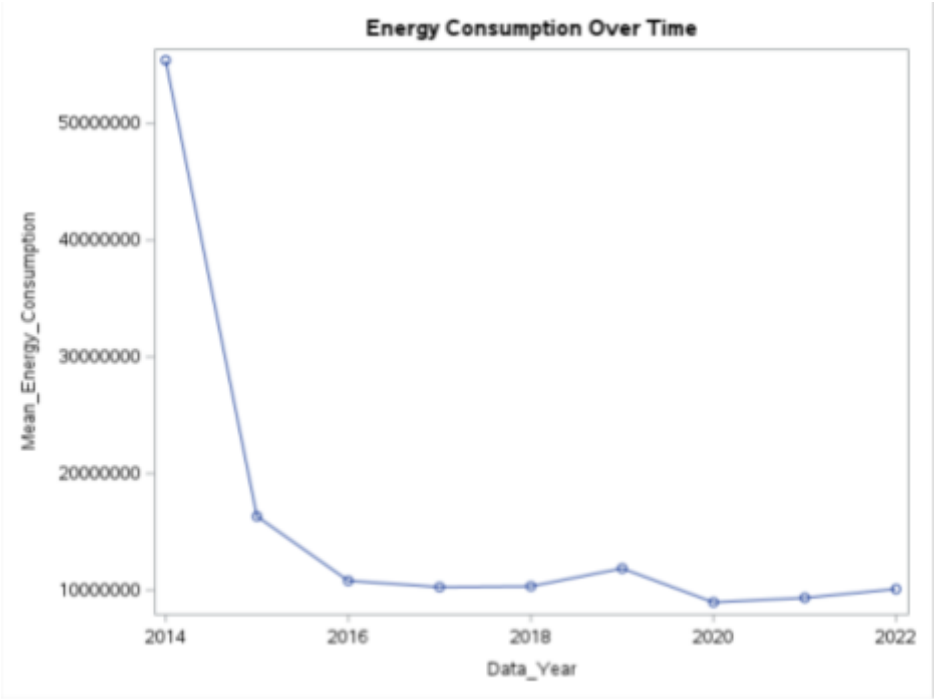


Figure 3: Energy Consumption Over Time

The line graph illustrates mean energy consumption with yearly trends. The line plot, with the x-axis representing years (2014-2022) and the y-axis depicting mean energy consumption , highlights key intervals of change and stability. The analysis result clearly shows that there is a significant drop in mean energy consumption from 2014 to 2015, suggesting substantial improvements in energy efficiency or changes in usage patterns.

From 2015 onwards, the data shows smaller fluctuations, indicating a more stable trend in energy consumption. Minor variations from 2019 to 2022 might be attributed to policy changes, seasonal effects, or other factors affecting energy use. Overall, the trend suggests successful energy-saving measures were implemented around 2014-2015, leading to a more stable and consistent energy management practice in subsequent years.

4. Exploring relationships between Gross Floor area and Electricity Usage

Regression Analysis of Gross Floor area vs Electricity Usage

The REG Procedure
Model: MODEL1
Dependent Variable: Gross_Floor_Area___Buildings__sq

Number of Observations Read	24895
Number of Observations Used	20248
Number of Observations with Missing Values	4647

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.372614E15	2.372614E15	30985.7	<.0001
Error	20246	1.550261E15	76571203748		
Corrected Total	20247	3.922875E15			

Root MSE	276715	R-Square	0.6048
Dependent Mean	273274	Adj R-Sq	0.6048
Coeff Var	101.25903		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	144926	2076.85039	69.78	<.0001
Electricity_Use__kBtu_	1	0.01142	0.00006489	176.03	<.0001

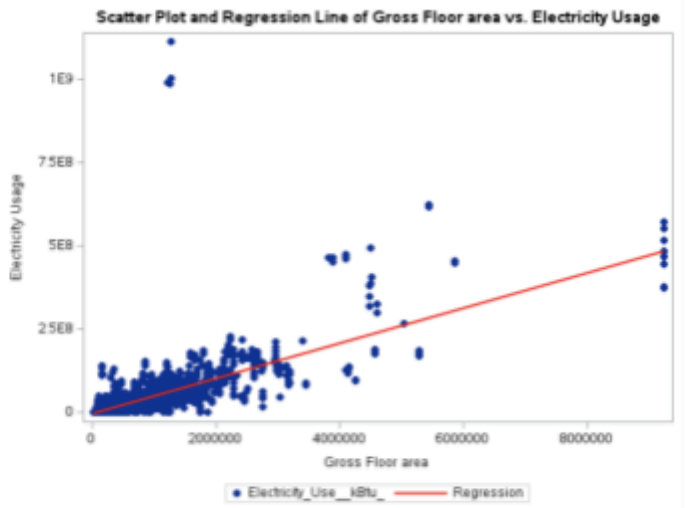


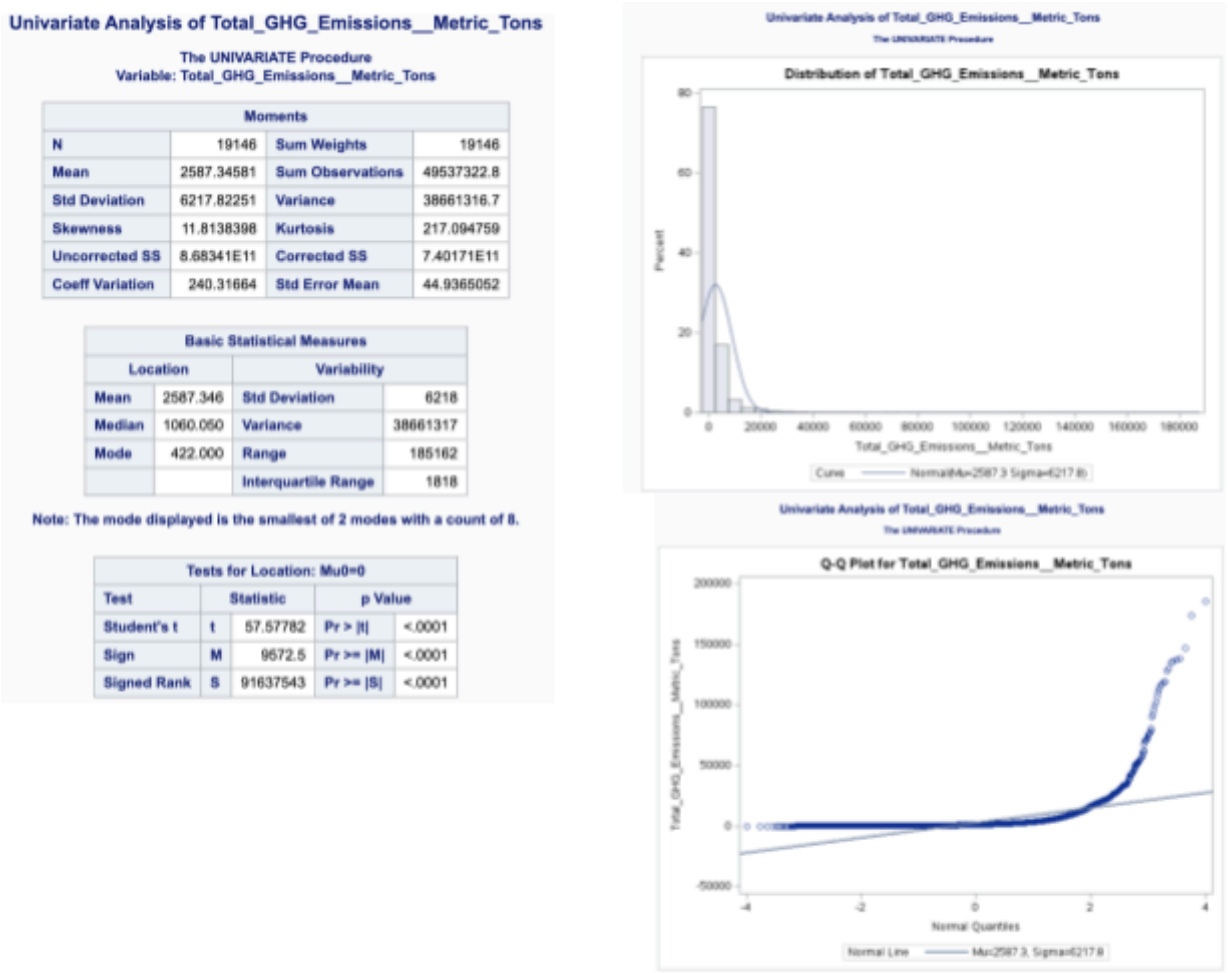
Figure 4: Scatter plot along with regression analysis of Gross Floor area with Electricity usage

The gross floor area is the dependent variable and the electricity is the independent variable. The regression line helps in determining models how changes in the gross floor area of buildings might predict changes in their electricity usage.

Further a scatter plot is created to show the distribution of 2 variables along with the regression line in red colour. The model's F-value is 30,985.7 with a P-value less than 0.0001, indicating the model is statistically significant. This suggests a strong relationship between gross floor area and electricity usage R-Squared: 0.6048, meaning about 60.48% of the variance in the electricity usage of the buildings can be explained by the variance in the building's floor area. The value for model intercept lies at 144,926 with a standard error of 2,076.85. The t-value is 69.78, and the P-value is less than 0.0001, indicating the intercept is significantly different from zero.

There's a significant positive relationship between the building's gross floor area and its electricity consumption. Larger buildings tend to consume more electricity The model is able to show how changes in a building's size affect its electricity use, but there could be other factors that also influence electricity use.

5. Univariate Analysis of GHG Emissions



Univariate Analysis of Total_GHG_Emissions__Metric_Tons

The UNIVARIATE Procedure
Variable: Total_GHG_Emissions__Metric_Tons

Moments			
N	19146	Sum Weights	19146
Mean	2587.34581	Sum Observations	49537322.8
Std Deviation	6217.82251	Variance	38661316.7
Skewness	11.8138398	Kurtosis	217.094759
Uncorrected SS	8.68341E11	Corrected SS	7.40171E11
Coeff Variation	240.31664	Std Error Mean	44.9365052

Basic Statistical Measures			
Location		Variability	
Mean	2587.346	Std Deviation	6218
Median	1060.050	Variance	38661317
Mode	422.000	Range	185162
		Interquartile Range	1818

Note: The mode displayed is the smallest of 2 modes with a count of 8.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	57.57782	Pr > |t|	<.0001
Sign	M	9572.5	Pr >= |M|	<.0001
Signed Rank	S	91637543	Pr >= |S|	<.0001

Figure 5: Analysis of Total GreenHouse Emissions in Metric tons

The univariate analysis of Total GHG Emissions (Metric Tons) provides detailed statistical insights into the distribution of greenhouse gas emissions data. The mean emission is 2,587.35 metric tons and standard deviation is 6,217.82, indicating high variability. The distribution is highly skewed to the right (skewness = 11.8134) and exhibits significant kurtosis (217.0975), suggesting a heavy-tailed distribution with many extreme values.

The histogram confirms the right-skewed distribution, showing that most buildings have lower emissions, while a few have very high emissions. The Q-Q plot illustrates that the data significantly deviates from a normal distribution, particularly in the upper tail, where extreme values are prevalent.

Statistical tests for normality, including the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests, all reject the null hypothesis of normality (p-values < 0.01), confirming GHG emissions are not normal.

6. Relationship between GHG Emissions and Gross floor Area

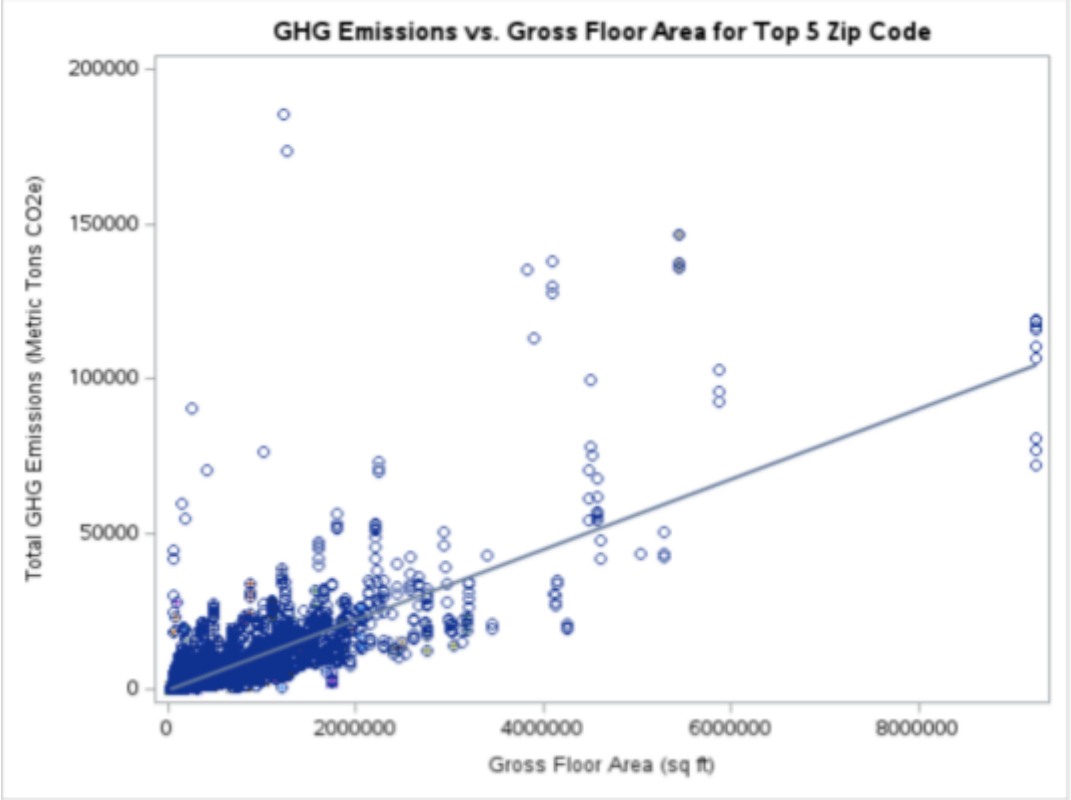


Figure 6: How do building characteristics affect GHG emissions for the top 5 zip codes

The scatter plot illustrates the relationship between Gross Floor Area and Total GHG Emissions for the top five zip codes. The plot reveals a positive correlation, where buildings with larger floor areas tend to produce higher greenhouse gas emissions. This is evidenced by the upward-sloping regression line. Despite the presence of some outliers, the general trend indicates that as the gross floor area increases, total GHG emissions also rise. This relationship suggests that larger buildings have a greater environmental impact in terms of emissions. Therefore, strategies for reducing GHG emissions should particularly target larger buildings to achieve significant reductions in environmental impact.

7. Is the average Energy Star score 59?

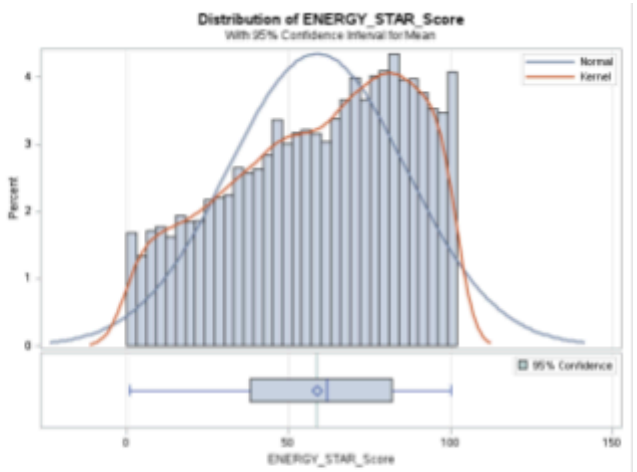
The TTEST Procedure

Variable: ENERGY_STAR_Score

N	Mean	Std Dev	Std Err	Minimum	Maximum
17386	58.7618	27.5262	0.2088	1.0000	100.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
58.7618	58.3526	59.1710	27.5262	27.2399	27.8186

DF	t Value	Pr > |t|
17385	-1.14	0.2538



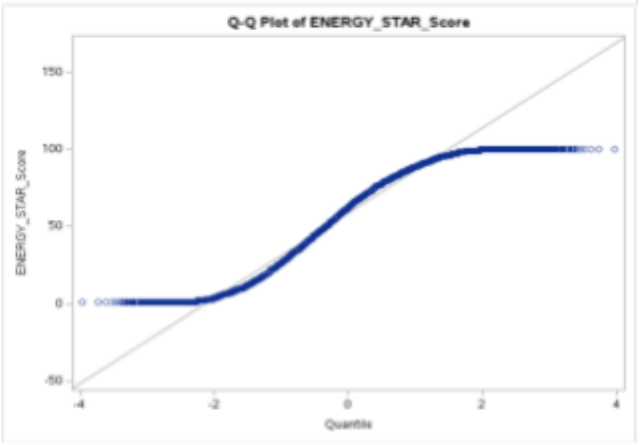


Figure 7: T- test to check for average energy score is equal to 59

The t-test examines whether the sample mean (58.7618) significantly differs from the hypothesized mean (59). The analysis is conducted using the t-test function in SAS. The null hypothesis (H0) assumes that the true mean of the variable ENERGY_STAR_Score is equal 59, while the alternative hypothesis (Ha) assumes that the true mean is not equal to 59

- **Null Hypothesis (H0)**: The mean ENERGY_STAR_Score is 59.
- **Alternative Hypothesis (Ha)**: The mean ENERGY_STAR_Score is not 59.

The result from the test reveals the p-value of 0.2538, which exceeds the alpha level of 0.05, hence we fail to reject the null hypothesis. This means that there is insufficient evidence to

claim that the true mean ENERGY_STAR_Score differs from 59. The detailed understanding of the t-test states that the sample mean ENERGY_STAR_Score is 58.7618, and the 95% confidence interval for the mean, ranging from 58.3526 to 59.1710, includes 59. The t-test statistic is -1.14 with a p-value of 0.2538. Consequently, given the p-value is greater than 0.05, we do not reject the null hypothesis and conclude there is no significant difference between the sample mean and the hypothesized mean of 59.

8. Building Age Groups and their mean Site and Source EUI

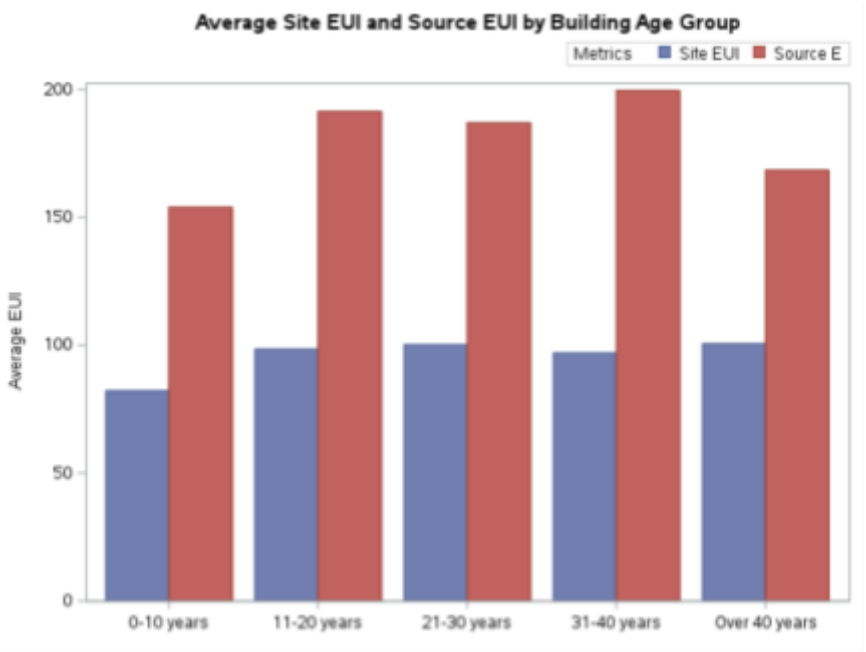
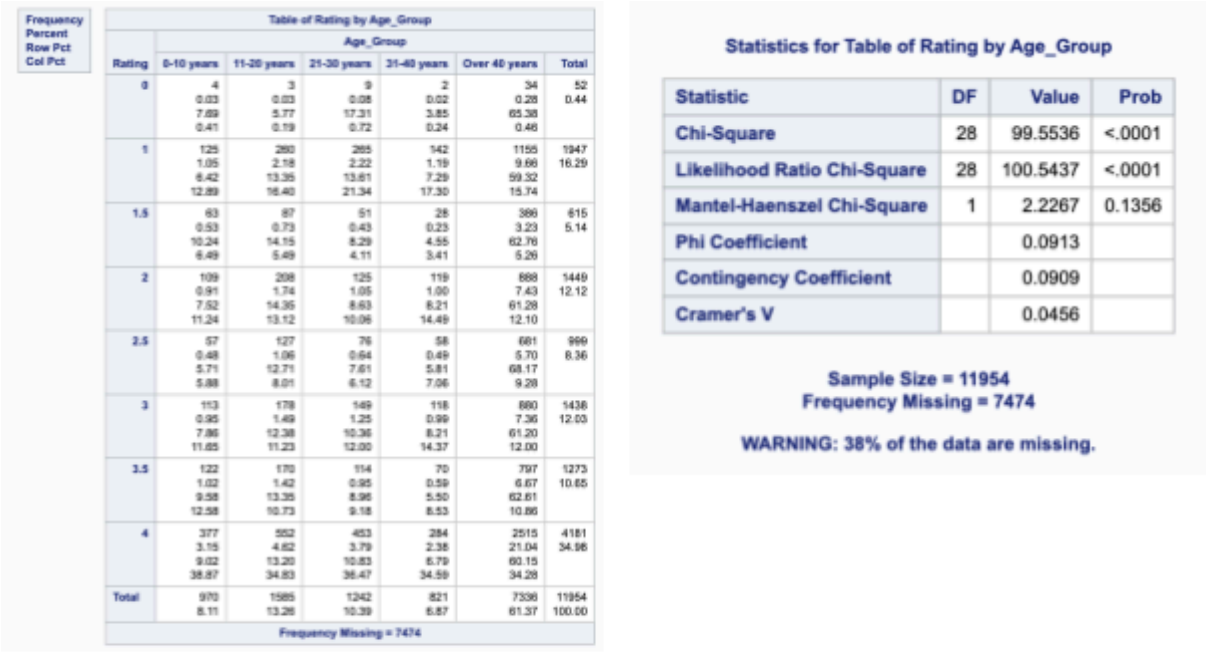


Figure 8: Comparative Analysis of Average Site EUI and Source EUI Across Building Age Groups

The clustered bar chart illustrates the relationship between the age groups (0–10 years, 11–20 years, 21–30 years, 31–40 years, and over 40 years) of buildings and their Site EUI(energy used inside the building) and Source EUI(total energy needed, including generation and distribution). According to the table, buildings that are 31–40 years old and older than 11–20 years old have higher average Source EUIs, while newer structures (0–10 years old) typically have lower Site EUIs, which indicates greater energy efficiency.

Overall, older buildings often have poorer energy efficiency and higher energy consumption, but newer structures are generally more energy-efficient because of contemporary construction standards and technologies.

9. Categorical Association between Chicago Energy Rating and Building age groups



Frequency
Percent
Row Pct
Col Pct

Table of Rating by Age_Group

Rating	Age_Group					
	0-10 years	11-20 years	21-30 years	31-40 years	Over 40 years	Total
0	4 0.03 7.69 0.41	3 0.03 5.77 0.19	9 0.08 17.31 0.72	2 0.02 3.85 0.24	34 0.28 65.38 0.46	52 0.44
1	125 1.05 6.42 12.89	260 2.18 13.35 16.40	265 2.22 13.61 21.34	142 1.19 7.29 17.30	1155 9.66 59.32 15.74	1947 16.29
1.5	63 0.53 10.24 6.49	87 0.73 14.15 5.49	51 0.43 8.29 4.11	28 0.23 4.55 3.41	386 3.23 62.76 5.26	615 5.14
2	109 0.91 7.52 11.24	208 1.74 14.35 13.12	125 1.05 8.63 10.06	119 1.00 8.21 14.49	888 7.43 61.28 12.10	1449 12.12
2.5	57 0.48 5.71 5.88	127 1.06 12.71 8.01	76 0.64 7.61 6.12	58 0.49 5.81 7.06	681 5.70 68.17 9.28	999 8.36
3	113 0.95 7.86 11.65	178 1.49 12.38 11.23	149 1.25 10.36 12.00	118 0.99 8.21 14.37	880 7.36 61.20 12.00	1438 12.03
3.5	122 1.02 9.58 12.58	170 1.42 13.35 10.73	114 0.95 8.96 9.18	70 0.59 5.50 8.53	797 6.67 62.61 10.86	1273 10.65
4	377 3.15 9.02 38.87	552 4.62 13.20 34.83	453 3.79 10.83 36.47	284 2.38 6.79 34.59	2515 21.04 60.15 34.28	4181 34.98
Total	970 8.11	1585 13.26	1242 10.39	821 6.87	7336 61.37	11954 100.00

Frequency Missing = 7474

Statistics for Table of Rating by Age_Group

Statistic	DF	Value	Prob
Chi-Square	28	99.5536	<.0001
Likelihood Ratio Chi-Square	28	100.5437	<.0001
Mantel-Haenszel Chi-Square	1	2.2267	0.1356
Phi Coefficient		0.0913	
Contingency Coefficient		0.0909	
Cramer's V		0.0456	

Sample Size = 11954
Frequency Missing = 7474

WARNING: 38% of the data are missing.

Figure 9: Categorical Association test between age_group and Energy rating

The images show the results of a Chi-Square test of independence aimed at analysing the relationship between Chicago Energy Rating and the Age Group of Buildings. This analysis, performed using the "chisq" function in SAS, involves creating a contingency table between the age_group and rating variables. This is a test of categorical association. The null hypothesis states that the two categories are independent, while the alternative hypothesis asserts that they are dependent.

The Chi-Square test statistic is 99.5536 with 28 degrees of freedom (DF), and a p-value less than 0.0001, indicating a statistically significant association between Rating and Age_Group. The Likelihood Ratio Chi-Square, which is 100.5437, further supports this significance. However, the Mantel-Haenszel Chi-Square test indicates no significant linear trend in the association. Despite the significance, the measures of association suggest that the association's strength is weak. Additionally, a limitation of this analysis is the significant portion of missing data (38%), which could affect the results' reliability.

10. Correlation between Natural Gas and GHG Emissions

Correlation between Natural Gas and GHG Emissions

The CORR Procedure

2 Variables:	Natural_Gas_Use__kBtu_ Total_GHG_Emissions__Metric_Tons

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Natural_Gas_Use__kBtu_	19114	12535707	33710366	2.39608E11	-4774718	1642941217
Total_GHG_Emissions__Metric_Tons	19146	2587	6218	49537323	0	185162

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Natural_Gas_Use__kBtu_	Total_GHG_Emissions__Metric_Tons
Natural_Gas_Use__kBtu_	1.00000 19114	0.67102 <.0001 17883
Total_GHG_Emissions__Metric_Tons	0.67102 <.0001 17883	1.00000 19146

Figure 10: Correlation between Natural Gas and Total GHG Emissions

Using SAS's CORR function, a correlation analysis was carried out between Natural Gas Use and Total GHG Emissions. 19,114 observations of natural gas use and 19,146 observations of total greenhouse gas emissions were included in the analysis. There is a quite high positive correlation between the two variables, as indicated by the 0.67102 Pearson correlation coefficient. This implies that total greenhouse gas emissions tend to rise along with an increase in natural gas use. With a p-value of less than 0.001, the correlation is statistically significant, demonstrating the validity of this association.

Conclusion

Our comprehensive analysis of Chicago's Energy Benchmarking Data uncovers crucial insights into patterns of energy consumption across different property types. Significant observations were noted. Multifamily housing is the largest consumer of natural gas. the relationship between gross floor area and electricity usage showed that larger buildings consume more energy. Further analysis of energy star scores relative to Chicago Energy Ratings highlighted that higher ratings correlated with improved efficiency, though the overall energy consumption patterns remained nonlinear. Historical data from 2014 to 2022 revealed that energy-saving measures implemented around 2014-2015 have effectively stabilized energy consumption over the subsequent years.

Additionally, regression analysis linked larger building sizes to increased electricity usage. A significant correlation was also witnessed between natural gas use and greenhouse gas emissions, emphasizing the environmental impact of current energy practices. This relationship underscores the need for sustainable energy solutions that can mitigate environmental impacts while catering to the demands of larger structures.

Overall, these insights enrich our understanding of the data, illustrating dependency of different energy consumption on building’s attributes. Through detailed statistical testing, such as t-tests and Chi-Square analyses, we witnessed the dynamics of energy ratings. These findings can guide the formulation of targeted energy policies and broader sustainability strategies, aimed at reducing carbon footprints and enhancing urban energy management in Chicago.

References

1. Nur Insani (2024) Module 4 - SAS Summaries [PDF Document, MATH1322], RMIT University, Melbourne.

2. Nur Insani (2024) Module 5 - SAS Outputs [PDF Document, MATH1322], RMIT University, Melbourne.

3. Nur Insani (2024) Module 5 - SAS PROC SQL [PDF Document, MATH1322], RMIT University, Melbourne.

4. Nur Insani (2024) Module 5 - Intermediate PROC SQL [PDF Document, MATH1322], RMIT University, Melbourne.

5. Data.Gov - Open dataset (Feb 17, 2024) Chicago Energy Benchmarking, accessed 25th May 2024, https://catalog.data.gov/dataset/chicago-energy-benchmarking

6. Chicago Data Portal (2024), Chicago Energy Benchmarking, accessed 25th May 2024,https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking/xq83-jr8c/about_data

7. TutorialsPoint (2024) , SAS SQL, accessed on 29th May 2024, https://www.tutorialspoint.com/sas/sas_sql.htm

8. Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group, accessed on 2nd June 2024 https://stats.oarc.ucla.edu/stata/ado/analysis/.

9. SAS Institute Inc. (2020) , The FREQ Procedure, accessed on 30th May 2024, https://documentation.sas.com/doc/en/statug/15.2/statug_freq_details08.htm

10. SAS Institute Inc. (2020) , VALIDVARNAME= SAS System Option, accessed on 30th May 2024, https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/acreldb/n0vnyuzncldjabn1923ug8svx7uh.htm

Appendix

The SAS CODE

```
libname mydata "/home/u63822929/sasuser.v94/Assignment 2/";
options validvarname=v7;

proc import
datafile='/home/u63822929/sasuser.v94/Assignment2//Chicago_Energy_Benchmarking.csv
'
   out=mydata.energy_data dbms=csv replace;
run;

proc print data=mydata.energy_data(obs=10);
run;
```

/* 1. Top Natural Gas Consumption by Property Type Based on Reporting status = Submitted Data */

```
proc sql;
 create table ng_buildingwise_energy_table as
 select Primary_Property_Type, sum(Natural_Gas_Use__kBtu_) as NG_Consumption
 from mydata.energy_data
 where Primary_Property_Type is not null
 and UPPER(reporting_status) = "SUBMITTED DATA"
 group by Primary_Property_Type
 order by NG_Consumption desc;
quit;

proc sgplot data=ng_buildingwise_energy_table(obs=5);
 title 'Bar plot of top 5 property types with max NG consumption';
 vbar Primary_Property_Type / response=NG_Consumption;
```

```
  xaxis label='Primary_Property_Type';
  yaxis label='Mean Energy Consumption';
run;
```

/* 2. Analysis of buildings of different energy ratings compare in terms of electricity and gas usage, and their overall ENERGY STAR scores */

```
proc sql;
    create table energy_by_rating as
    select Chicago_Energy_Rating,
           count(*) as Number_of_Properties,
           sum(Electricity_Use__kBtu_) as Total_Electricity_Use,
           sum(Natural_Gas_Use__kBtu_) as Total_Natural_Gas_Use,
           mean(ENERGY_STAR_Score) as Avg_ENERGY_STAR_Score
    from mydata.energy_data
    where Chicago_Energy_Rating is not null
    group by Chicago_Energy_Rating
    order by Chicago_Energy_Rating;
quit;


proc sgplot data=energy_by_rating;
    vbar Chicago_Energy_Rating / response=Total_Electricity_Use datalabel;
    title "Total Electricity Usage by Chicago Energy Rating";
    xaxis label="Chicago Energy Rating";
    yaxis label="Total Electricity Use (kBtu)";
run;

proc sgplot data=energy_by_rating;
    vbar Chicago_Energy_Rating / response=Total_Natural_Gas_Use datalabel;
    title "Total Natural Gas Usage by Chicago Energy Rating";
```

```
    xaxis label="Chicago Energy Rating";
    yaxis label="Total Natural Gas Use (kBtu)";
run;

proc sgplot data=energy_by_rating;
    vbar Chicago_Energy_Rating / response=Avg_ENERGY_STAR_Score datalabel;
    title "Average ENERGY STAR Score by Chicago Energy Rating";
    xaxis label="Chicago Energy Rating";
    yaxis label="Average ENERGY STAR Score";
run;
```

/* 3. Energy Consumption Over Time */

```
proc sql;
  create table yearwise_mean_energy_table as
  select Data_Year, mean(Electricity_Use__kBtu_) as Mean_Energy_Consumption
  from mydata.energy_data
  where Data_Year is not null
  group by Data_Year;
quit;

proc sgplot data=yearwise_mean_energy_table;
  title 'Energy Consumption Over Time';
  series x=Data_Year y=Mean_Energy_Consumption / markers;
run;
```

/* 4. Scatter plot along with regression analysis of Gross Floor area with Electricity usuage */

```
proc reg data=mydata.energy_data;
    model Gross_Floor_Area___Buildings__sq = Electricity_Use__kBtu_;
```

```
   title "Regression Analysis of Gross Floor area  vs Electricity Usage";
run;


proc sgplot data=mydata.energy_data;
                scatter x=Gross_Floor_Area___Buildings__sq y=Electricity_Use__kBtu_ /
markerattrs=(symbol=CircleFilled);
   reg x=Gross_Floor_Area___Buildings__sq y=Electricity_Use__kBtu_ / lineattrs=(color=red);
   title "Scatter Plot and Regression Line of Gross Floor area vs. Electricity Usage";
   xaxis label="Gross Floor area ";
   yaxis label="Electricity Usage";
run;
```

/* 5. Analysis of Total_GHG_Emissions__Metric_Tons */

```
proc univariate data=mydata.energy_data;
   var Total_GHG_Emissions__Metric_Tons;
   histogram Total_GHG_Emissions__Metric_Tons/ normal;
   qqplot Total_GHG_Emissions__Metric_Tons / normal(mu=est sigma=est);
   title 'Univariate Analysis of Total_GHG_Emissions__Metric_Tons';
run;
```

/* 6. How do building characteristics (e.g., Gross Floor Area) affect GHG emissions for the top 5 zip codes */

```
proc sql;
   create table total_ghg_by_prop_floor_area as
   select Zip_Code,
        Total_GHG_Emissions__Metric_Tons,
        Gross_Floor_Area___Buildings__sq
   from mydata.energy_data
   group by Zip_Code
```

```
   order by Total_GHG_Emissions__Metric_Tons desc;
quit;

proc sgplot data=total_ghg_by_prop_floor_area;
        scatter  x=Gross_Floor_Area___Buildings__sq  y=Total_GHG_Emissions__Metric_Tons  /
group=Zip_Code;
   reg x=Gross_Floor_Area___Buildings__sq y=Total_GHG_Emissions__Metric_Tons;
   xaxis label="Gross Floor Area (sq ft)";
   yaxis label="Total GHG Emissions (Metric Tons CO2e)";
   title "GHG Emissions vs. Gross Floor Area for Top 5 Zip Code";
run;
```

/* 7 Ttest to check if average energy score is equal to 59 */

```
PROC TTEST data= mydata.energy_data h0= 59;
var ENERGY_STAR_Score;
RUN;
```

/* 8. Comparative Analysis of Average Site EUI and Source EUI Across Building Age Groups */

```
proc sql;
 create table energy_data_age_groups as
 select Year_Built,
       (year(today()) - Year_Built) as Building_Age,
       case
         when (year(today()) - Year_Built) <= 10 then '0-10 years'
         when (year(today()) - Year_Built) <= 20 then '11-20 years'
         when (year(today()) - Year_Built) <= 30 then '21-30 years'
         when (year(today()) - Year_Built) <= 40 then '31-40 years'
         else 'Over 40 years'
```

```
        end as Age_Group,
        Site_EUI__kBtu_sq_ft_ as Site_EUI,
        Source_EUI__kBtu_sq_ft_ as Source_EUI,
        Chicago_Energy_Rating as Rating
  from mydata.energy_data
  where Site_EUI__kBtu_sq_ft_ is not null
    and Source_EUI__kBtu_sq_ft_ is not null;
quit;

proc print data=energy_data_age_groups(obs=20);
  var Year_Built Building_Age Age_Group Site_EUI Source_EUI;
  title 'Data with Building Age Groups and Calculated Energy Metrics';
run;

/* Reshape data for clustered bar chart */
data reshaped_energy_data;
  set energy_data_age_groups;
  Metric = 'Site EUI';
  EUI = Site_EUI;
  output;
  Metric = 'Source EUI';
  EUI = Source_EUI;
  output;
  keep Age_Group Metric EUI;
run;
```

/* Clustered bar chart for Average Site EUI and Average Source EUI by Building Age Group */

```
proc sgplot data=reshaped_energy_data;
  title 'Average Site EUI and Source EUI by Building Age Group';
```

```
  vbar Age_Group / response=EUI stat=mean group=Metric groupdisplay=cluster;
  xaxis display=(nolabel) label='Building Age Group';
  yaxis label='Average EUI';
  keylegend / title='Metrics' position=topright;
run;
```

/* 9 Categorical Association test between age_group and Energy rating */

```
proc freq data=energy_data_age_groups;
    tables Rating*Age_Group / chisq;
run;
```

/* 10 Correlation between Natural Gas and Total GHG Emissions */

```
proc corr data=mydata.energy_data;
    var Natural_Gas_Use__kBtu_ Total_GHG_Emissions__Metric_tons;
    title "Correlation between Natural Gas and GHG Emissions";
run;
```