Buildings and GHG Prediction

@ Seattle

Analytic Report

by Dave Lam

Oct 2022

Content

- Executive summary and Introduction
- Methodology
- Dataset
- Data Cleaning
- Data Mining
- Data Visualization and pivot table analysis
- Prediction by Polynomial Regression

Executive Summary and Introduction

Summary

According to Seattle Office of Sustainability & Environment, buildings in Seattle are one of the largest and fastest growing sources of climate pollution, more than a third of the city's Green-house Gas (GHG) emissions is from the buildings through energy used*. These emissions pollute the environment and change the climate.

Therefore, in this report, we try to find out buildings in Seattle that they need to be improved in energy use more effective, so as to reduce green-house emission.

*Source from: Seattle's Buildings and Energy

Introduction

This report is to analyze the dataset (2015 Building Energy Benchmarking) from the city providing to the public. From the dataset, we hope to find out:

1) buildings in Seattle that are over the benchmarking in energy use
2) building types in geographical distribution in relation to green-house gas emission
3) prediction of green-house gas emission from a building in respect to its features, such as property type, gross floor area (GFA), energy (electricity, natural gas and steam) use, location, building age etc.

Methodology

In this report, we use Python programming to conduct data analysis, which includes data cleaning, mining and visualization. Also, information* from the Seattle city government provided to the public and technical reference of EnergyStar are adopted.

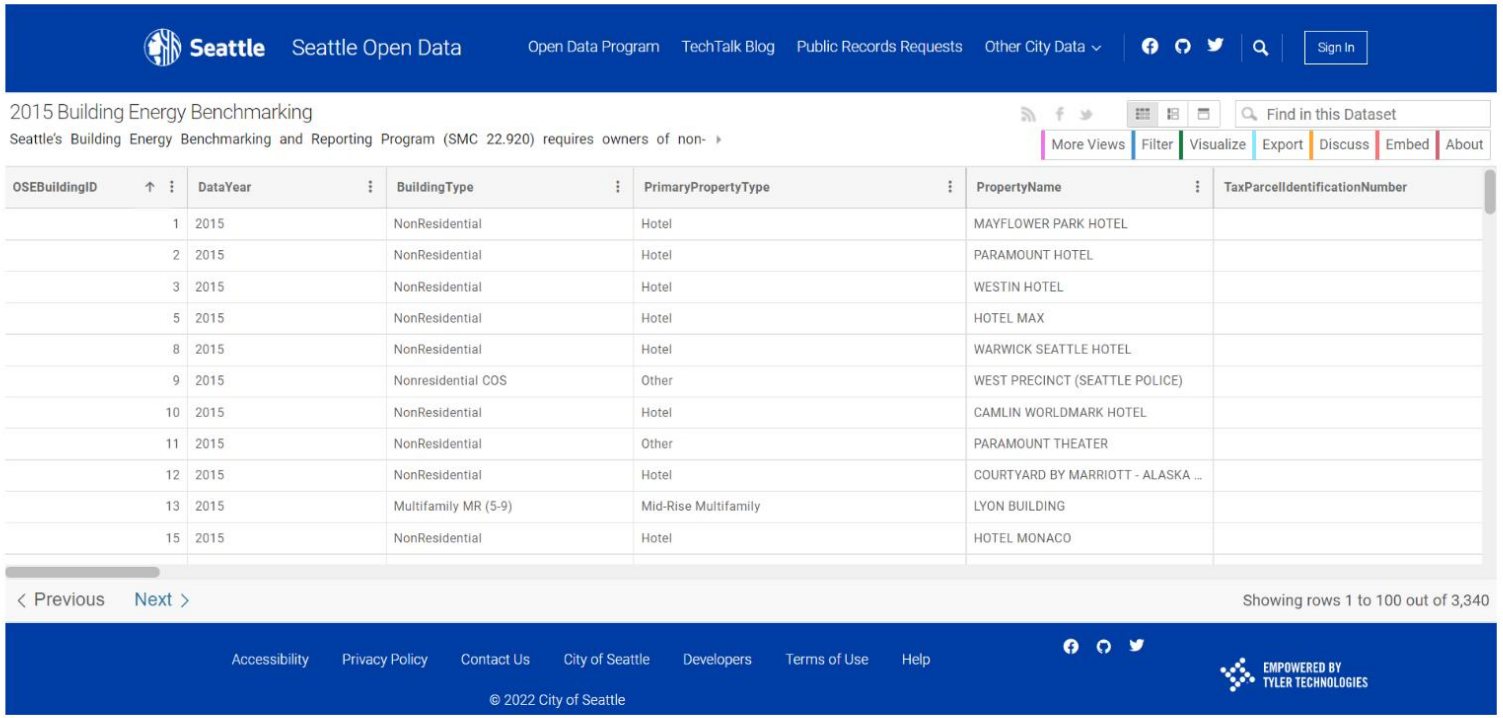
*Source:

Seattle's Buildings and Energy
Energy Star's Technical Reference

Dataset

The dataset* is "2015 Building Energy Benchmarking" from the Seattle Open Data that the data includes 3340 entries and 47 columns*.
*Source:

[Dataset]
[Data's interpretation]



OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber
1	2015	NonResidential	Hotel	MAYFLOWER PARK HOTEL	
2	2015	NonResidential	Hotel	PARAMOUNT HOTEL	
3	2015	NonResidential	Hotel	WESTIN HOTEL	
5	2015	NonResidential	Hotel	HOTEL MAX	
8	2015	NonResidential	Hotel	WARWICK SEATTLE HOTEL	
9	2015	Nonresidential COS	Other	WEST PRECINCT (SEATTLE POLICE)	
10	2015	NonResidential	Hotel	CAMLIN WORLDMARK HOTEL	
11	2015	NonResidential	Other	PARAMOUNT THEATER	
12	2015	NonResidential	Hotel	COURTYARD BY MARRIOTT - ALASKA ...	
13	2015	Multifamily MR (5-9)	Mid-Rise Multifamily	LYON BUILDING	
15	2015	NonResidential	Hotel	HOTEL MONACO	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3340 entries, 0 to 3339
Data columns (total 47 columns):
 #   Column                                            Non-Null Count  Dtype
---  ------                                            --------------  -----
 0   OSEBuildingID                                     3340 non-null   int64
 1   DataYear                                          3340 non-null   int64
 2   BuildingType                                      3340 non-null   object
 3   PrimaryPropertyType                               3340 non-null   object
 4   PropertyName                                      3340 non-null   object
 5   TaxParcelIdentificationNumber                     3338 non-null   object
 6   Location                                          3340 non-null   object
 7   CouncilDistrictCode                               3340 non-null   int64
 8   Neighborhood                                      3340 non-null   object
 9   YearBuilt                                         3340 non-null   int64
 10  NumberofBuildings                                 3340 non-null   int64
 11  NumberofFloors                                    3332 non-null   float64
 12  PropertyGFATotal                                  3340 non-null   int64
 13  PropertyGFAParking                                3340 non-null   int64
 14  PropertyGFABuilding(s)                            3340 non-null   int64
 15  ListOfAllPropertyUseTypes                         3213 non-null   object
 16  LargestPropertyUseType                            3204 non-null   object
 17  LargestPropertyUseTypeGFA                         3204 non-null   float64
 18  SecondLargestPropertyUseType                      1559 non-null   object
 19  SecondLargestPropertyUseTypeGFA                   1559 non-null   float64
 20  ThirdLargestPropertyUseType                       560 non-null    object
 21  ThirdLargestPropertyUseTypeGFA                    560 non-null    float64
 22  YearsENERGYSTARCertified                          110 non-null    object
 23  ENERGYSTARScore                                   2560 non-null   float64
 24  SiteEUI(kBtu/sf)                                  3330 non-null   float64
 25  SiteEUIWN(kBtu/sf)                                3330 non-null   float64
 26  SourceEUI(kBtu/sf)                                3330 non-null   float64
 27  SourceEUIWN(kBtu/sf)                              3330 non-null   float64
 28  SiteEnergyUse(kBtu)                               3330 non-null   float64
 29  SiteEnergyUseWN(kBtu)                             3330 non-null   float64
 30  SteamUse(kBtu)                                    3330 non-null   float64
 31  Electricity(kWh)                                  3330 non-null   float64
 32  Electricity(kBtu)                                 3330 non-null   float64
 33  NaturalGas(therms)                                3330 non-null   float64
 34  NaturalGas(kBtu)                                  3330 non-null   float64
 35  OtherFuelUse(kBtu)                                3330 non-null   float64
 36  GHGEmissions(MetricTonsCO2e)                      3330 non-null   float64
 37  GHGEmissionsIntensity(kgCO2e/ft2)                 3330 non-null   float64
 38  DefaultData                                       3339 non-null   object
 39  Comment                                           13 non-null     object
 40  ComplianceStatus                                  3340 non-null   object
 41  Outlier                                           84 non-null     object
 42  2010 Census Tracts                                224 non-null    float64
 43  Seattle Police Department Micro Community Policing Plan Areas  3338 non-null   float64
 44  City Council Districts                            213 non-null    float64
 45  SPD Beats                                         3338 non-null   float64
 46  Zip Codes                                         3340 non-null   int64
```

Data Cleaning

The following are how to modify data columns on behalf of analysis:

1. separate the items nested in the “Location” column

2. rename the columns to make them coherent
i.e. GHGemissions(MetricTonsCO2e) and GHGEmissionsIntensity(kgCO2e/ft2) have the same value in amount but the different interpretation in unit, so keep one and rename to TotalGHGEmissions

3. check any null and redundant data in the dataset

4. eliminate any data unnecessary and duplicated to the analysis
i.e. OSEBuildingID, DataYear, CouncilDistrictCode, 2010 Census Tracts, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, SPD Beats, and Zip Codes

5. quantify category to numeric feature
i.e. ListOfAllPropertyUseTypes, LargestPropertyUseType, SecondLargestPropertyUseType,
and ThirdLargestPropertyUseTypeGFA to make change and rename as TotalUseTypeNumber

6. combine related columns into one, such as:
a) LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, and ThirdLargestPropertyUseTypeGFA are the breakdown of PropertyGFATotal
b) SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), SiteEnergyUseWN(kBtu), SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu), OtherFuelUse(kBtu) are the components of SiteEnergyUse(kBtu)

7. re-rate the data, such as
a) change the data of PropertyGFAParking and PropertyGFABuilding into ratio
b) Latitude and Longitude data are modified to Haversine_distance

Data Mining

The following are the wayfinding to data mining:

1. Correlation among the data columns
   a) Using Pearson Correlation
   b) Under the threshold of 0.7 to confine the correlation result
   c) Verification of multicollinearity with the VIF (Variance Inflation Factor)

2. Distribution of GHG Emission and Energy Use and the relationship between them

3. Relationship between haversine distance and GHG Emission in relation to Energy Use

4. Relationship between Building Type and GHG Emission in relation to Energy Use

5. Relationship between Building Age and GHG Emission in relation to Energy Use

6. Number of buildings by the Year in which
   a) property was constructed or underwent a complete renovation year of construction
   b) age of a property

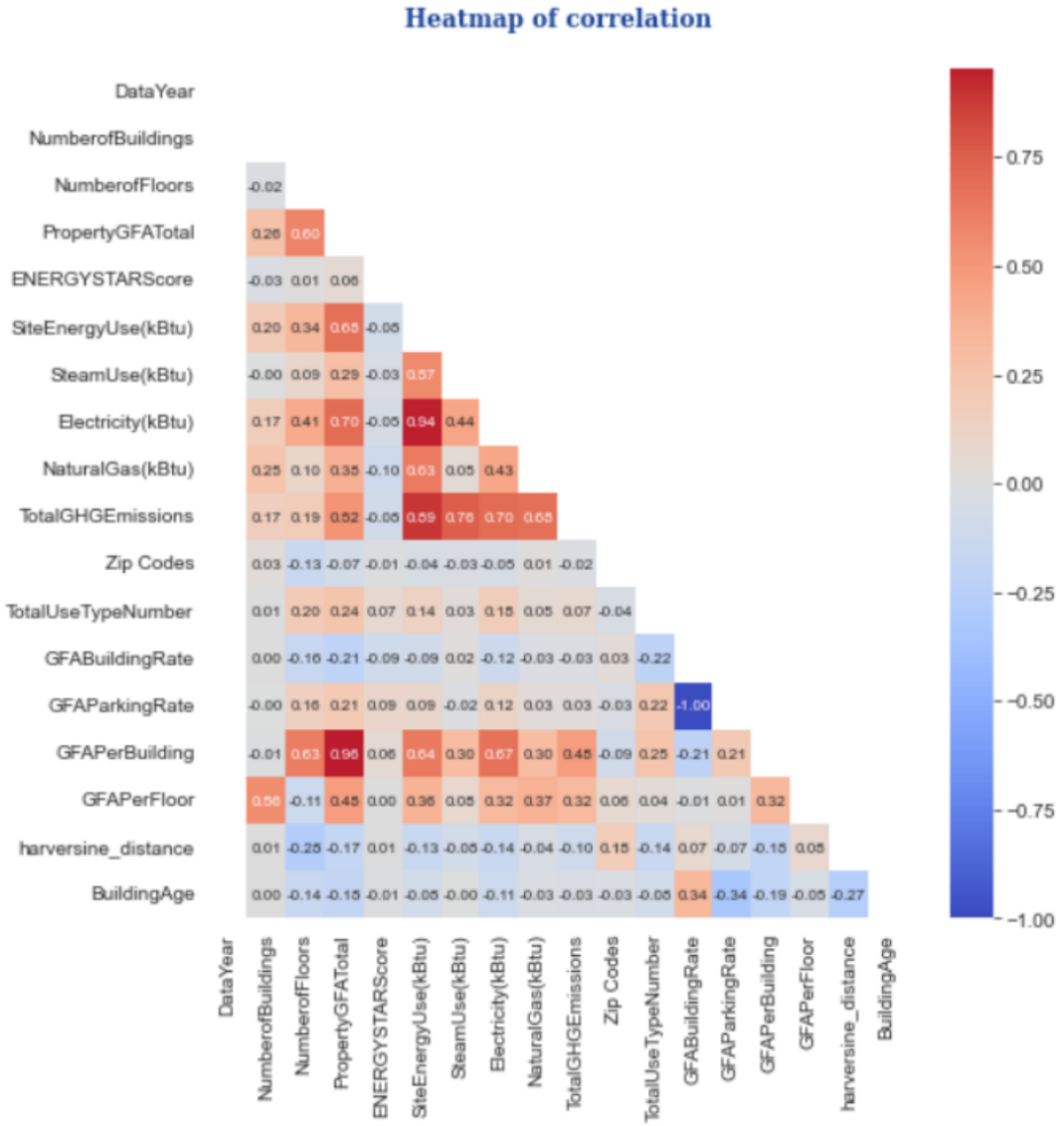
7. Clarify category and numerical features from column data

8. Number of buildings by Primary Property Type

9. Pivot tables

10. Location of the number of buildings
    a) using Folium Map

Visualization



	level_0	level_1	corr_coeff
12	PropertyGFATotal	GFAPerBuilding	0.955969
10	SiteEnergyUse(kBtu)	Electricity(kBtu)	0.941414
8	TotalGHGEmissions	SiteEnergyUse(kBtu)	0.892636
6	TotalGHGEmissions	SteamUse(kBtu)	0.756680
4	Electricity(kBtu)	PropertyGFATotal	0.700599
2	GFABuildingRate	GFAParkingRate	-1.000000
0	BuildingAge	YearBuilt	-1.000000

	feature	VIF
0	PropertyGFATotal	1.349081e+01
1	GFAParkingRate	3.464307e+14
2	GFABuildingRate	3.002400e+15
3	SiteEnergyUse(kBtu)	4.810910e+03
4	Electricity(kBtu)	1.736158e+03
5	SteamUse(kBtu)	6.403370e+01
6	YearBuilt	2.595735e+13
7	BuildingAge	2.065871e+13
8	TotalGHGEmissions	1.505526e+03
9	GFAPerBuilding	1.239428e+01

Analysis:

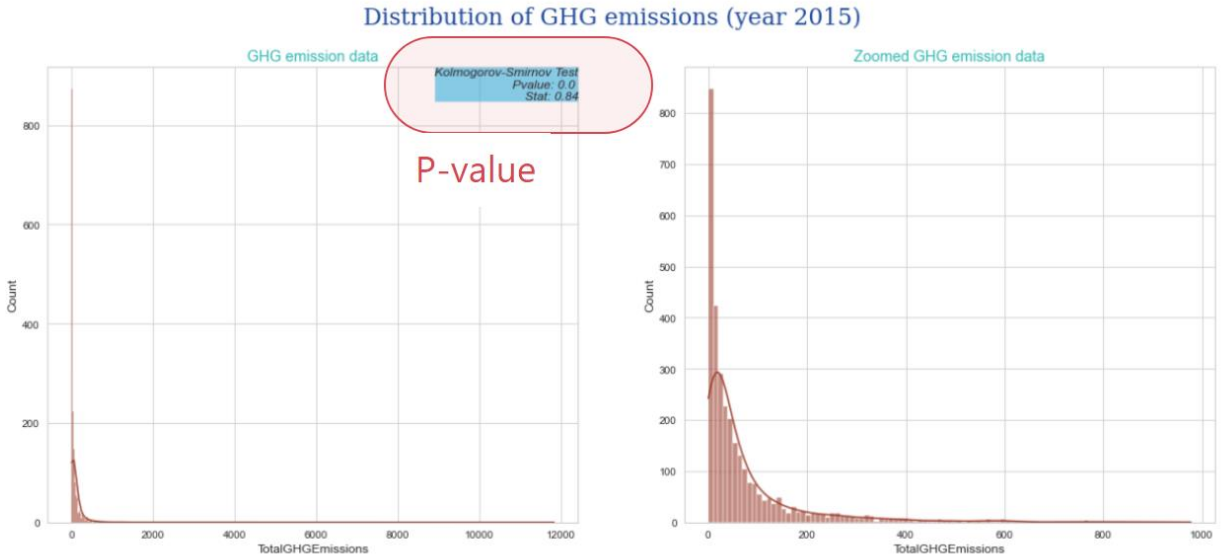
a) Using Pearson Correlation and limiting the result in the threshold of 0.7

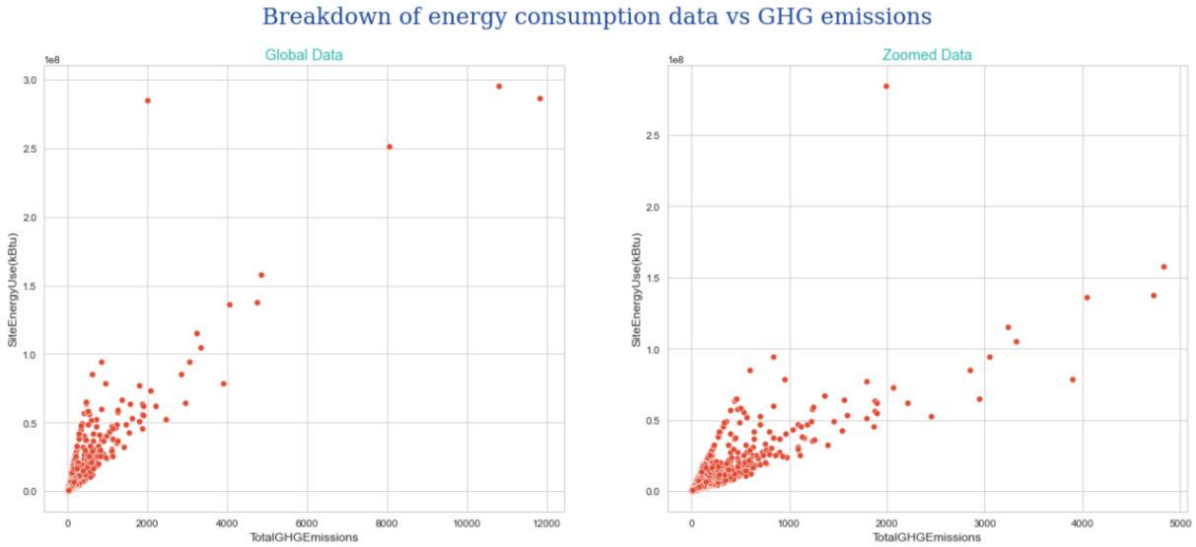
b) Verification of multicollinearity with the VIF (Variance Inflation Factor)

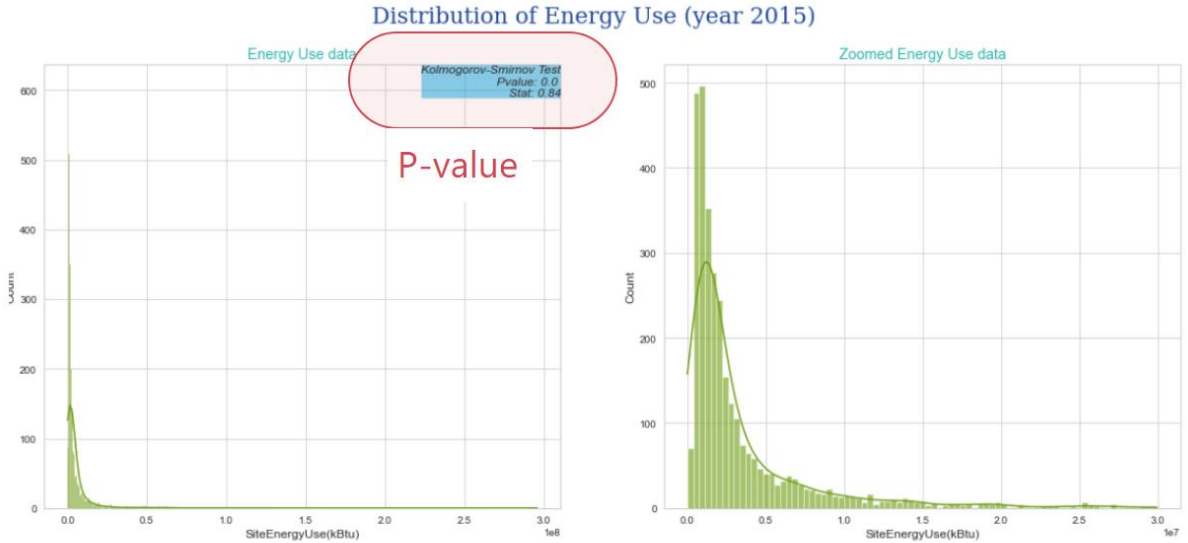
We can see a close relationship between features (GFA building, GFA parking, building age, year of built, energy, and GHG) as seen from the table and graph.

Next, we will find out what features will have impact on GHG.

Visualization





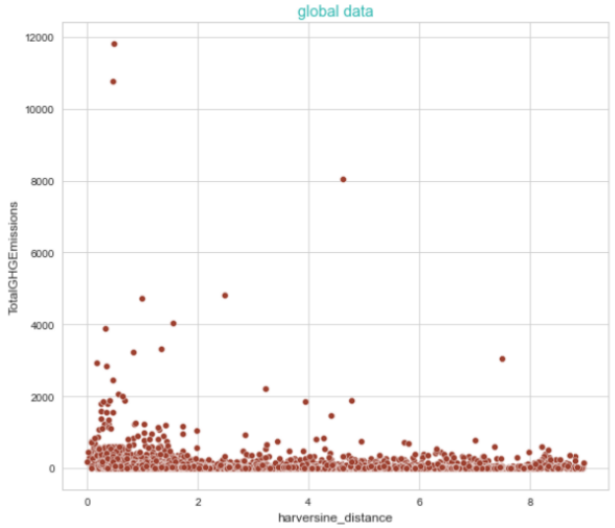


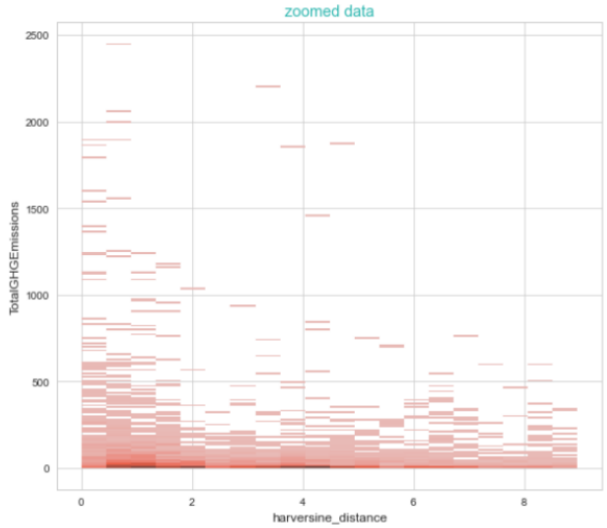
Analysis:

Using Kolmogorov-Smirnov, the p-value is 0. We therefore have **significant evidence to reject the null hypothesis that the variable follows a normal distribution**. The null hypothesis is that **the two dataset values are from the same continuous distribution**. In this case, GHG Emission and Energy Use are correlated.

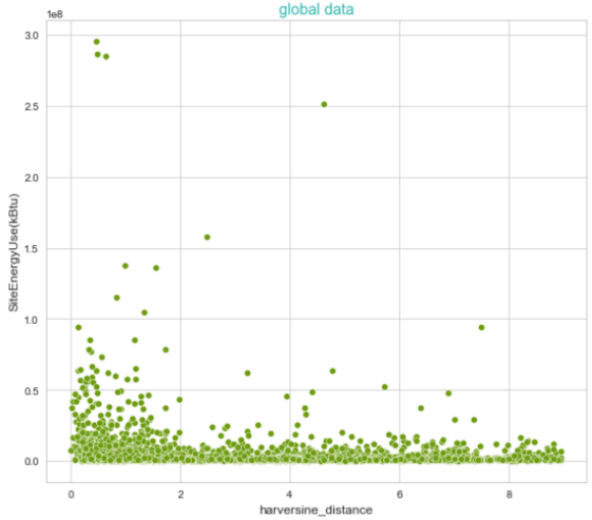
Visualization

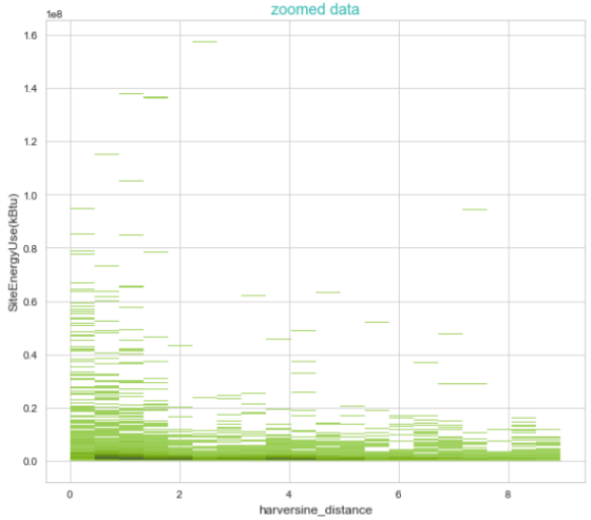
Distribution of GHG emissions data according to geographical coordinates



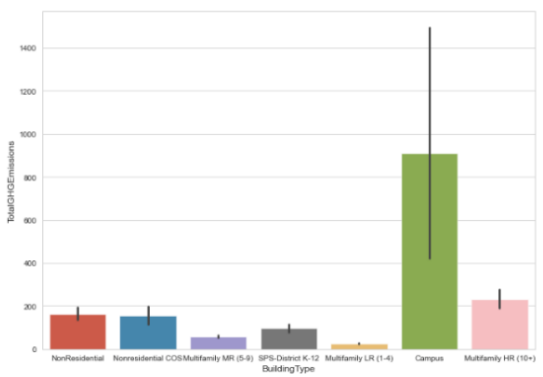


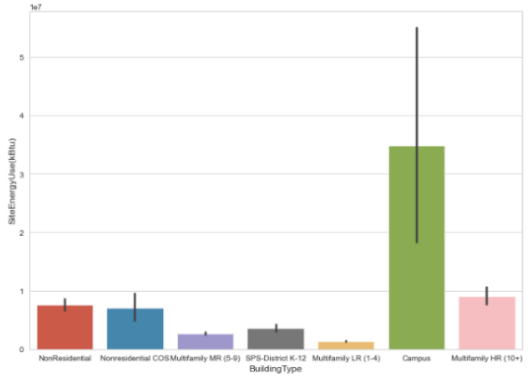
Distribution of Energy consumption data according to geographical coordinates





Breakdown of energy consumption and GHG emissions by Building Type



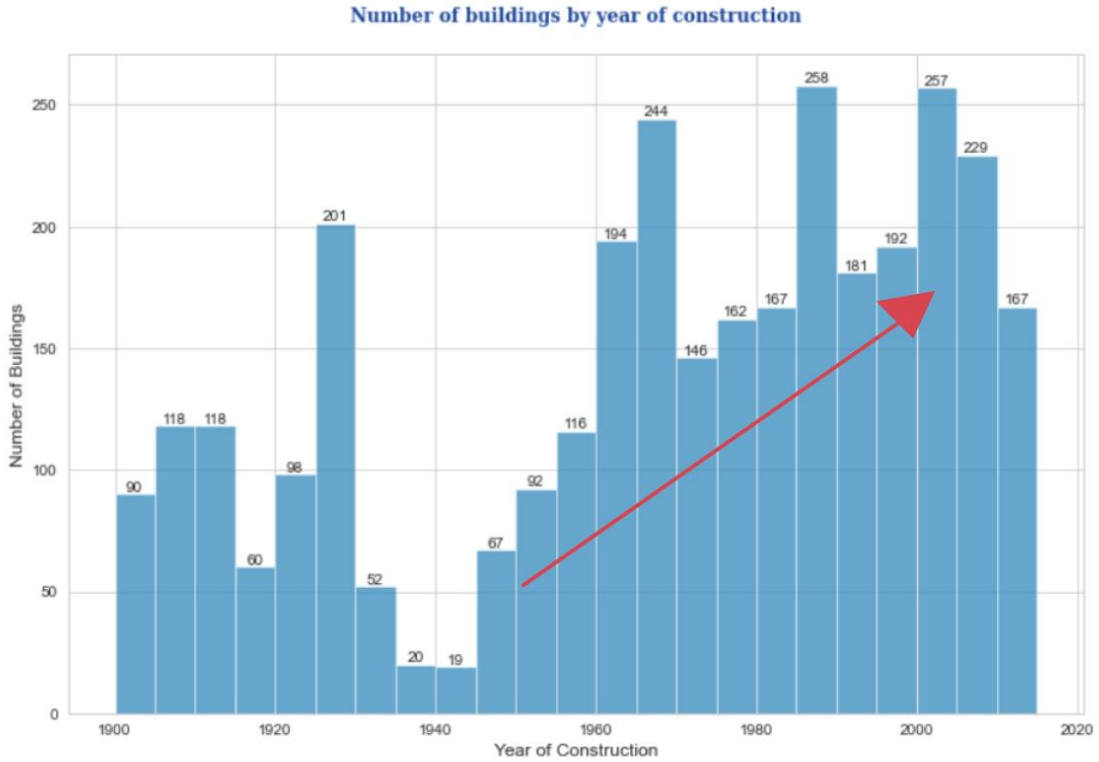


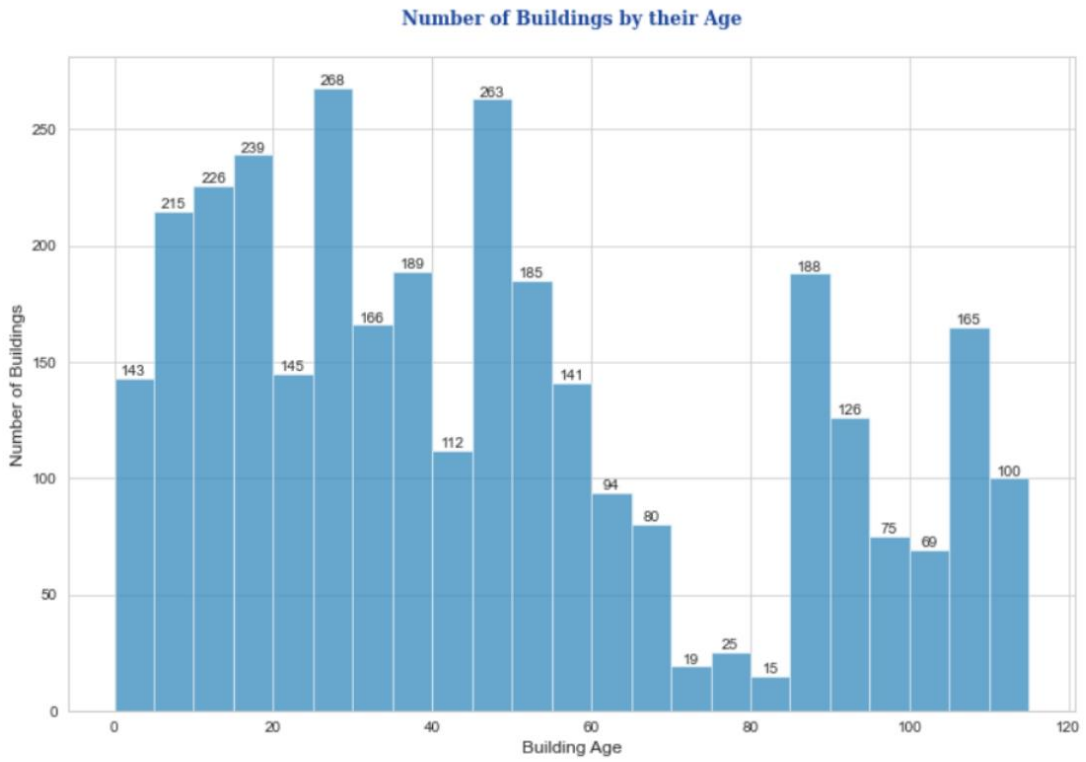
Analysis:

The data, Latitude and Longitude of the buildings have been changed to Haversine distance*. As seen from the graphs, there is no strong evidence to show the relationship between distance and Energy Use, and distance and GHG Emission. Therefore, we would say distance has no significant impact on GHG emission.

*Haversine distance is the distance between the coordinate points of building and the center of Seattle.

Visualization





Analysis:

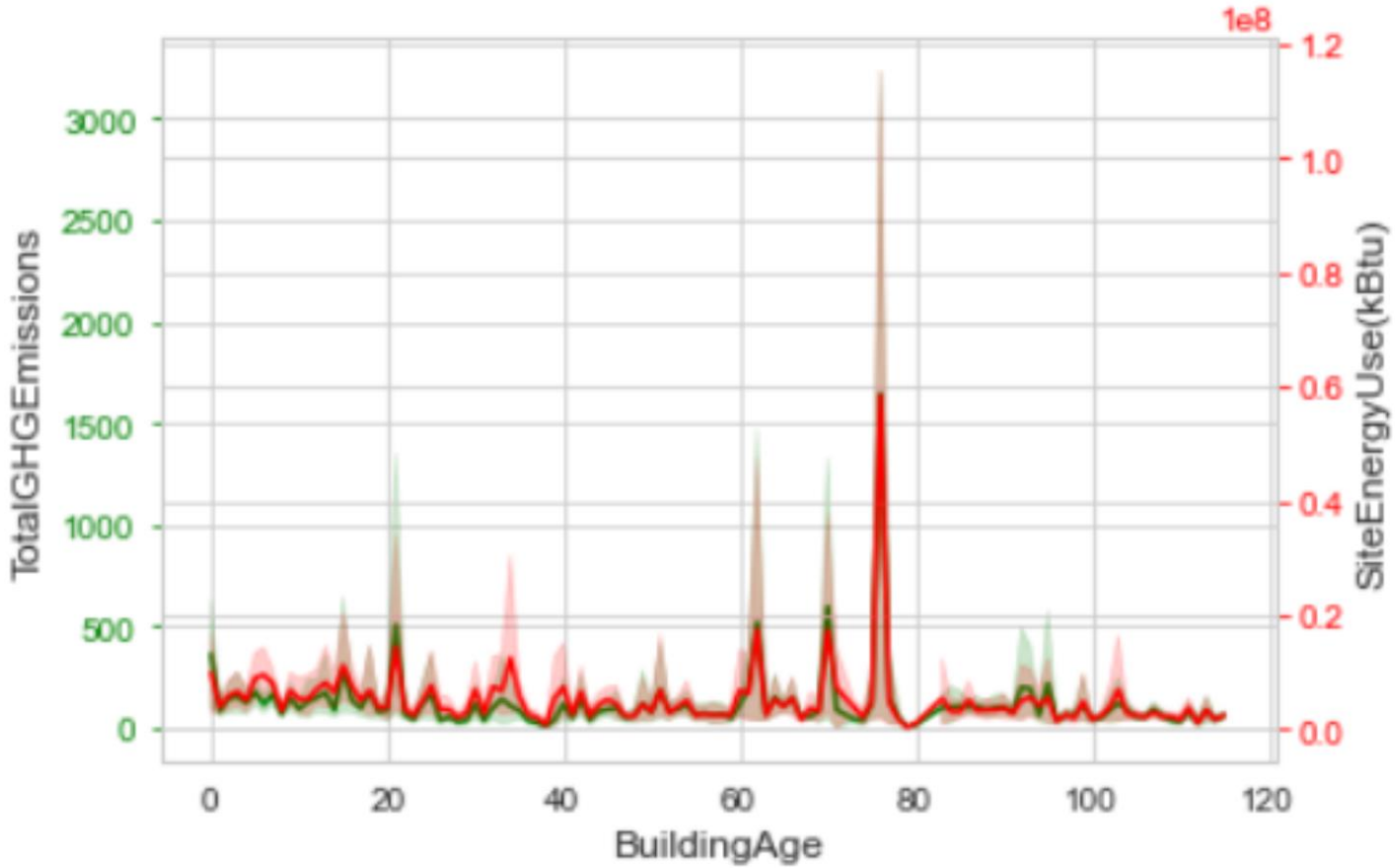
The number of buildings by the Year in which

a) a property was constructed or underwent a complete renovation year of construction

b) building age

We can see the number of buildings from year 1945 to 2000 increasing gradually and the number starts decreasing after year 2000.

Visualization



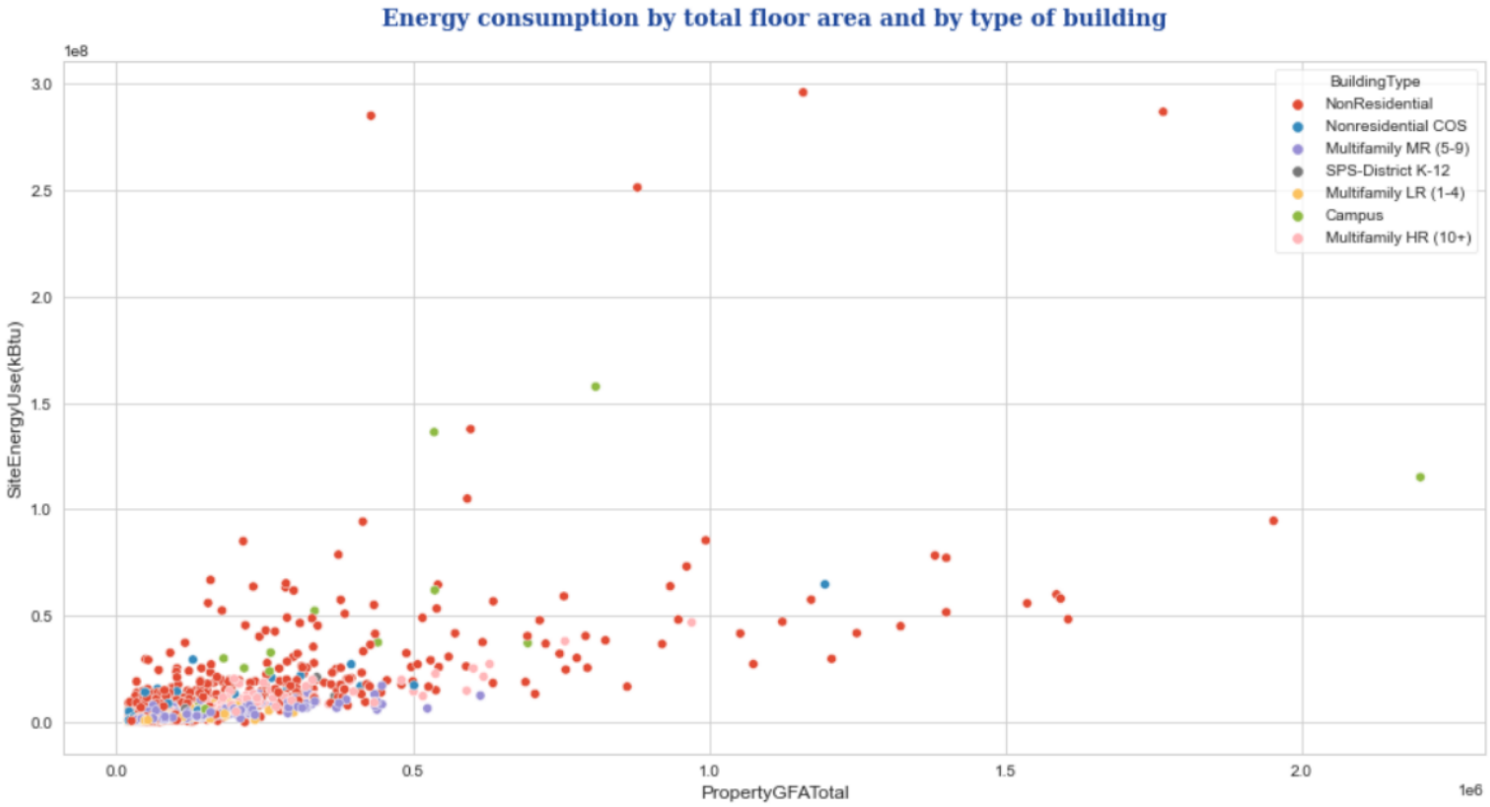
Analysis:

a) Similar patterns between GHG Emission by Building Age and Energy Use by Building Age

b) The buildings in the age group (60 to 80) have higher GHG emission and Energy use

We can see from the graphs:
1) GHG Emission and Energy Use have similar patterns, so they are closely related.
2) Building age is not correlated with GHG Emission and Energy Use.

Visualization



BuildingType	PrimaryPropertyType	PropertyName	Neighborhood	YearBuilt
NonResidential	Hospital	SEATTLE CHILDREN'S HOSPITAL (MAIN CAMPUS)	NORTHEAST	1953
NonResidential	Hospital	HARBORVIEW MEDICAL CENTER	EAST	2000
NonResidential	Other	THE WESTIN BUILDING	DOWNTOWN	1981
NonResidential	Hospital	SWEDISH HOSPITAL MEDICAL CENTER First Hill Campus	EAST	1994
Campus	Other	AMGEN INC. MASTER CAMPUS	MAGNOLIA / QUEEN ANNE	2002

PropertyGFATotal	LargestPropertyUseType	YearsENERGYSTARCertified	ENERGYSTARScore	SiteEnergyUse(kBtu)
879000	Hospital (General Medical & Surgical)	NaN	13.0	251191824.0
1158691	Hospital (General Medical & Surgical)	NaN	30.0	295812640.0
429405	Data Center	NaN	71.0	284867168.0
1765970	Hospital (General Medical & Surgical)	NaN	59.0	286685536.0
808520	Other	NaN	NaN	157606480.0

SteamUse(kBtu)	Electricity(kBtu)	NaturalGas(kBtu)	TotalGHGEmissions
0.0	114748139.0	136448438.0	8046.70
122701720.0	170891586.0	2226424.0	10780.64
0.0	284726322.0	152639.0	1992.96
127869744.0	140448322.0	18373320.0	11824.89
0.0	76742000.0	80867644.0	4829.96

Analysis:

From the scatter plot and the tables below, the top 5 buildings that consume energy the most are:

1) Harborview Medical Center

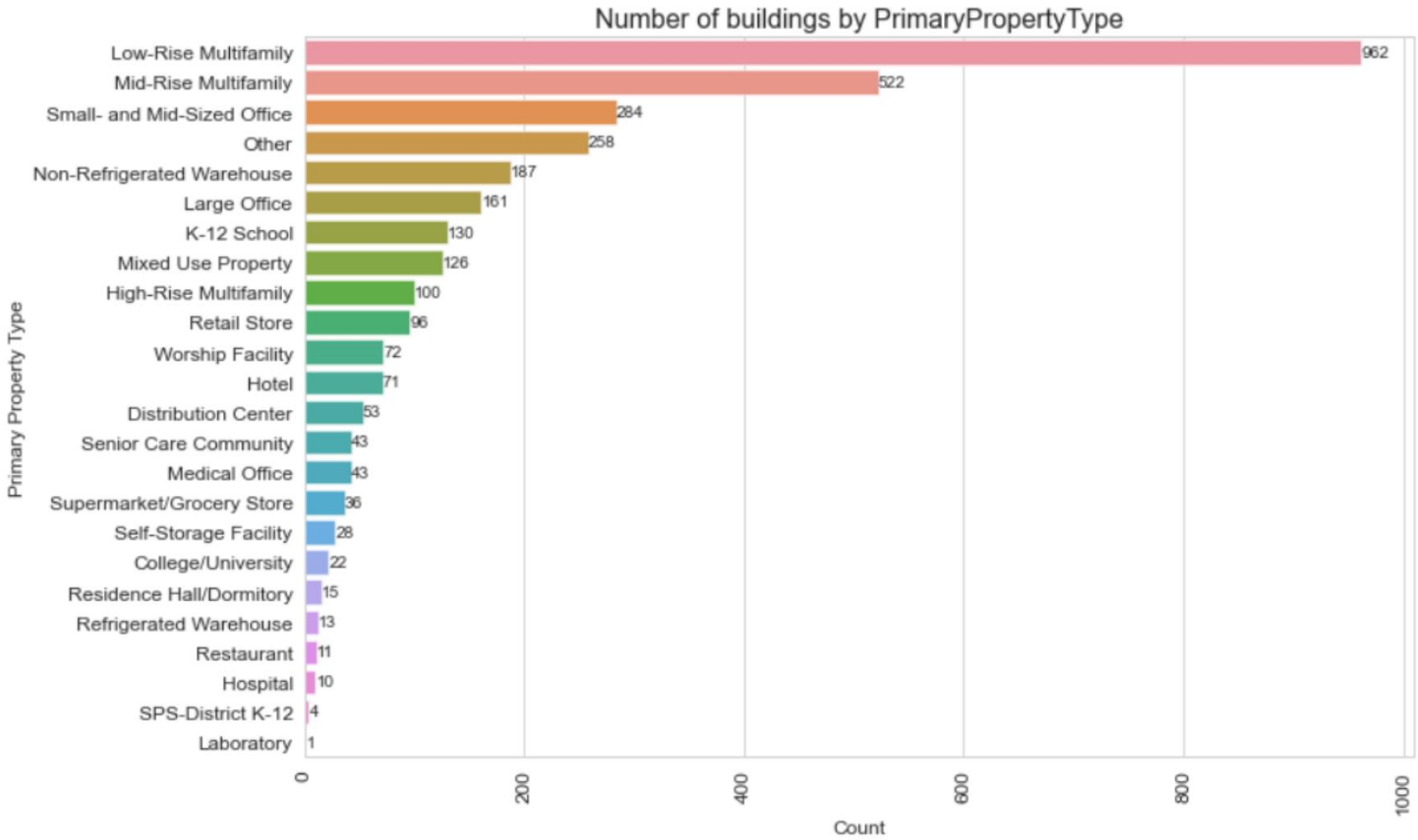
2) Swedish Hospital Medical Center (First Hill Campus)

3) The Westin Building (tele-com hub with data facility)

4) Seattle Children's Hospital (Main Campus)

5) Amgen Inc. Master Campus

Visualization with Pivot Table



	Energy_per_GFA		TotalGHGEmissions
	count	mean	sum
PrimaryPropertyType			
Laboratory	1	247.374774	612.84
Supermarket/Grocery Store	36	235.737882	8191.87
Hospital	10	192.789527	46847.63
Restaurant	11	173.093623	2045.79
Other	258	96.191478	62891.93
Refrigerated Warehouse	13	85.747008	417.72
Senior Care Community	43	76.810149	11574.93
Medical Office	43	76.471661	11463.42
Hotel	71	74.683687	28609.63
Mixed Use Property	126	71.100746	17738.28
College/University	22	62.492728	9792.48
Retail Store	96	61.679761	8138.87
Residence Hall/Dormitory	15	56.249209	1053.47
Small- and Mid-Sized Office	284	53.386318	11170.61
Large Office	161	49.817900	32170.40
High-Rise Multifamily	100	44.258458	21618.73
K-12 School	130	43.283830	11778.43
SPS-District K-12	4	35.458078	403.07
Distribution Center	53	33.902057	2649.22
Worship Facility	72	33.182882	2970.88
Non-Refrigerated Warehouse	187	32.902669	7447.79
Mid-Rise Multifamily	522	32.396518	29753.75
Low-Rise Multifamily	962	31.030194	22963.69
Self-Storage Facility	28	18.802829	627.71

Energy Star

ENERGY STAR® PortfolioManager®

Technical Reference

U.S. National Median Reference Values for All Portfolio Manager Property Types

Broad Category	Primary Function	Further Breakdown (where needed)	Source EUI (kBtu/ft²)	Site EUI (kBtu/ft²)	Reference Data Source - Peer Group Comparison
Banking/Financial Services	Bank Branch *		209.9	88.3	CBECS - Bank/Financial
	Financial Office*		116.4	52.9	CBECS - Office & Bank/Financial
Education	Adult Education		110.4	52.4	CBECS - Education
	College/University		180.6	84.3	CBECS - College/University
	K-12 School*		104.4	48.5	CBECS - Elementary/Middle & High School
	Pre-school/Daycare		131.5	64.8	CBECS - Preschool
	Vocational School		110.4	52.4	CBECS - Education
	Other - Education				
Entertainment/Public Assembly	Convention Center		109.6	56.1	CBECS - Social/Meeting
	Movie Theater		112.0	56.2	CBECS - Public Assembly
	Museum				
	Performing Arts				
	Recreation	Bowling Alley	112.0	50.8	CBECS - Recreation
		Fitness Center/Health Club/Gym			
		Ice/Curling Rink			
		Roller Rink			
		Swimming Pool			
		Other - Recreation			
	Social/Meeting Hall		109.6	56.1	CBECS - Social/Meeting

Recommendation:

The reference table is designed to help comparing property's energy use with the national median (or mid□point) energy use of similar properties.

Therefore, it helps to see any building type (or specific building) is/are over the reference benchmarks.

* Energy Star's Technical Reference

Data Comparison

ENERGY STAR® PortfolioManager® Technical Reference

U.S. National Median Reference Values for All Portfolio Manager Property Types

Broad Category	Primary Function	Further Breakdown (where needed)	Source EUI (kBtu/ft²)	Site EUI (kBtu/ft²)	Reference Data Source - Peer Group Comparison
Banking/Financial Services	Bank Branch *		209.9	88.3	CBECS - Bank/Financial
	Financial Office*		116.4	52.9	CBECS - Office & Bank/Financial
Education	Adult Education		110.4	52.4	CBECS - Education
	College/University		180.6	84.3	CBECS - College/University
	K-12 School*		104.4	48.5	CBECS - Elementary/Middle & High School
	Pre-school/Daycare		131.5	64.8	CBECS - Preschool
	Vocational School		110.4	52.4	CBECS - Education
	Other - Education				
Entertainment/Public Assembly	Convention Center		109.6	56.1	CBECS - Social/Meeting
	Movie Theater		112.0	56.2	CBECS - Public Assembly
	Museum				
	Performing Arts				
	Recreation	Bowling Alley	112.0	50.8	CBECS - Recreation
		Fitness Center/Health Club/Gym			
		Ice/Curling Rink			
		Roller Rink			
		Swimming Pool			
		Other - Recreation			
	Social/Meeting Hall		109.6	56.1	CBECS - Social/Meeting

	Energy_per_GFA		TotalGHGEmissions
	count	mean	sum
PrimaryPropertyType			
Laboratory	1	247.374774	612.84
Supermarket/Grocery Store	36	235.737882	8191.87
Hospital	10	192.789527	46847.63
Restaurant	11	173.093623	2045.79
Other	258	96.191478	62891.93
Refrigerated Warehouse	13	85.747008	417.72
Senior Care Community	43	76.810149	11574.93
Medical Office	43	76.471661	11463.42
Hotel	71	74.683687	28609.63
Mixed Use Property	126	71.100746	17738.28
College/University	22	62.492728	9792.48
Retail Store	96	61.679761	8138.87
Residence Hall/Dormitory	15	56.249209	1053.47
Small- and Mid-Sized Office	284	53.386318	11170.61
Large Office	161	49.817900	32170.40
High-Rise Multifamily	100	44.258458	21618.73
K-12 School	130	43.283830	11778.43
SPS-District K-12	4	35.458078	403.07
Distribution Center	53	33.902057	2649.22
Worship Facility	72	33.182882	2970.88
Non-Refrigerated Warehouse	187	32.902669	7447.79
Mid-Rise Multifamily	522	32.396518	29753.75
Low-Rise Multifamily	962	31.030194	22963.69
Self-Storage Facility	28	18.802829	627.71

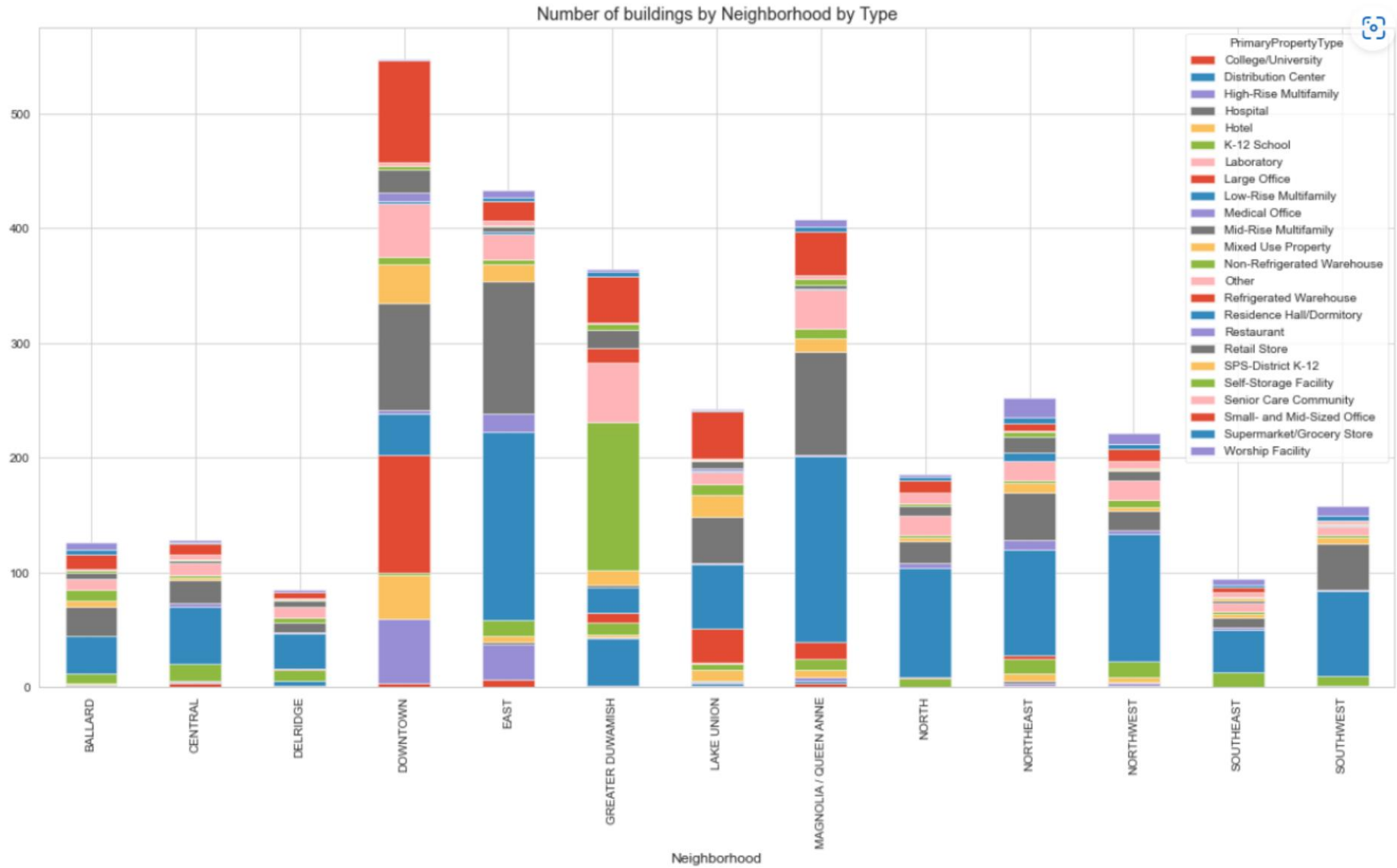
Analysis:

Comparing the data (Energy_per_GFA) shown on the table beside with the benchmarking data (Site EUI (kBtu/ft2) shown the EnergyStar's reference table, we can see the following building types in Seattle are over the benchmarks.

1) Laboratory
2) Supermarket/ Grocery Store
3) Refrigerated Warehouse
4) Medical Office
5) Hotel
6) Mix Use Property
7) Retail Store
8) Small- and Mid-Sized Office
9) Distribution Center
10) Worship Facility
11) Non-Refrigerated Warehouse

*'Other' in the Primary Property Type is not included in this comparison.

Visualization



Pivot Table

PrimaryPropertyType	College/University	Distribution Center	High-Rise Multifamily	Hospital	Hotel	K-12 School	Laboratory	Large Office	Low-Rise Multifamily	Medical Office	Mid-Rise Multifamily	Mixed Use Property
Neighborhood												
DOWNTOWN	4	0	56	0	38	2	0	102	36	4	93	34
EAST	7	0	30	3	5	14	0	0	164	15	116	15
MAGNOLIA / QUEEN ANNE	3	3	3	0	6	10	0	15	161	1	90	12
GREATER DUWAMISH	1	42	1	0	2	10	0	9	22	0	2	13
NORTHEAST	1	0	3	2	6	13	0	3	92	8	42	8
LAKE UNION	1	2	2	1	9	5	1	30	56	1	41	19
NORTHWEST	1	0	3	1	4	14	0	0	111	3	17	3
NORTH	0	0	0	0	0	8	0	1	95	4	19	4
SOUTHWEST	0	0	0	1	0	9	0	0	74	1	40	6
CENTRAL	3	0	2	1	0	14	0	0	50	3	20	3
BALLARD	0	1	0	1	1	9	0	0	33	0	25	6
SOUTHEAST	0	0	0	0	0	13	0	0	37	2	9	3
DELRIDGE	1	5	0	0	0	9	0	1	31	1	8	0

Non-Refrigerated Warehouse	Other	Refrigerated Warehouse	Residence Hall/Dormitory	Restaurant	Retail Store	SPS-District K-12	Self-Storage Facility	Senior Care Community	Small- and Mid-Sized Office	Supermarket/Grocery Store	Worship Facility	Total
6	47	0	2	7	20	0	3	4	88	0	2	548
4	22	0	2	0	5	0	1	4	17	3	6	433
9	33	0	2	0	3	0	5	3	38	4	7	408
129	52	13	0	0	16	0	5	1	40	4	2	364
2	17	0	8	0	13	0	5	1	6	5	17	252
9	11	0	1	2	6	1	0	1	42	1	1	243
6	17	0	0	0	9	1	1	6	11	4	10	222
2	17	0	0	0	8	0	2	10	10	3	3	186
2	7	0	0	1	0	0	1	2	1	5	8	158
2	10	0	0	0	2	0	1	5	9	1	2	128
9	9	0	0	1	5	0	2	1	13	4	6	126
2	7	0	0	0	3	2	1	4	4	2	6	95
5	9	0	0	0	6	0	1	1	5	0	2	85

Analysis: Pivot table (Neigborhood by PrimaryPropertyType)

The table figure refers the previous stacked bar chart (visualization _7), we can see the which types of primary property types in which area is the most:

1) office, high-rise multifamily, hotel, restaurants, and retail store in Downtown

2) non-refrigerated warehouse and distribution center in Greater Duwamish

3) Mid-rise and Low-rise multifamily in East.

Pivot Table

Neighborhood	PrimaryPropertyType	PropertyName	Energy_per_GFA	Electricity(kBtu)	NaturalGas(kBtu)	PropertyGFATotal	SiteEnergyUse(kBtu)
SOUTHWEST	Restaurant	SALTYS RESTAURANT	445.392097	3378201	5707048	20398	9085108
	Supermarket/Grocery Store	ADMIRAL METROPOLITAN MARKET	383.308376	6088718	6116456	31841	12204922
		WEST SEATTLE THRIFTWAY	259.836913	5549794	3208500	33706	8758063
	Other	NW ART AND FRAME/HUSKY ICE CREAM	228.302660	2657078	2406785	22180	5063753
	Supermarket/Grocery Store	QFC	218.858414	5646757	2640334	37864	8286855
NORTHWEST	Other	NORTHWEST HOSPITAL & MEDICAL CENTER (NEW PROFESSIONAL BLDG)	544.945740	20697280	8499046	53575	29195468
		AURORA CHRYSLER PLYMOUTH	290.914760	2213351	4080100	21633	6293359
	Supermarket/Grocery Store	SAFEWAY STORE # 1845 (FORMERLY 3389)	289.398071	7850393	6074900	48117	13924967
	Hospital	NORTHWEST HOSPITAL & MEDICAL CENTER	226.591427	42421724	51758703	415632	94178648
	Supermarket/Grocery Store	QUALITY FOOD CENTER (QFC) (OLD ART'S SHOPPING CENT	218.801314	8216324	5104640	60880	13320624
LAKE UNION	Mixed Use Property	CHANDLERS COVE	400.469875	5703982	8168130	34639	13871876
		FISHER PLAZA - WEST BUILDING	397.414630	84837545	146727	213834	84980760
	Supermarket/Grocery Store	Q F C	333.862052	5689951	3928182	28808	9617898
	Other	ZYMOGENETICS-DENDREON	322.052390	17857273	19420391	115748	37276920
		FHCRC - Weintraub/Hutchinson/Thomas Bldgs Campus	254.206897	69147767	67096519	535947	136241424
SOUTHEAST	Other	RAINIER BEACH COMMUNITY CENTER	291.522250	4261219	9823563	48314	14084606
	Retail Store	EMPIRE CENTER	279.585804	5323739	5268029	37883	10591549
	Supermarket/Grocery Store	SAFEWAY 1965	220.151737	8499577	4941700	61053	13440924
	Other	RAINIER SQUARE PLAZA	206.296907	12415487	8654752	102133	21069722
	Supermarket/Grocery Store	SAARS MARKET (2012)	160.881290	6924747	3926985	67450	10851443

Analysis:

Pivot Table (Neigborhood, PrimaryPropertyType, PropertyName and EnergyperGFA by Electricity, NaturalGas, SiteEnergyUse and PropertyGFATotal)

We can see in Southwest, Saltys restaurant is the building in the most energy use per sqft in 2015 that also show how much the energy use in electricity, natural gas and the total energy the building used. Similarly, we can find out buildings that consume the most energy per sqft from different Neighborhood areas.

Next, we can use the EnergyStar's technical reference to find out which buildings are over the benchmarks.

Pivot Table

	PropertyGFATotal	SiteEnergyUse(kBtu)				TotalGHGEmissions
	sum	count	mean	std	sum	sum
Neighborhood						
DOWNTOWN	96586290	548	9.618702e+06	1.785421e+07	5271048848	98851.15
EAST	35894548	433	5.590880e+06	2.213001e+07	2420850997	75946.83
LAKE UNION	26413616	243	6.691377e+06	1.361614e+07	1626004537	30696.10
MAGNOLIA / QUEEN ANNE	27278671	408	3.242276e+06	9.249954e+06	1322848581	29954.14
GREATER DUWAMISH	27215078	364	3.617007e+06	8.212074e+06	1316590677	23340.07
NORTHEAST	16590028	252	3.878890e+06	1.631367e+07	977480236	23981.97
NORTHWEST	14579321	222	3.441376e+06	7.590375e+06	763985373	17735.05
NORTH	13715761	186	2.778517e+06	3.408668e+06	516804179	9483.70
CENTRAL	8427178	128	3.742174e+06	1.236513e+07	478998216	11903.69
BALLARD	8716058	126	3.397067e+06	6.460336e+06	428030424	9821.59
SOUTHWEST	7841011	158	2.188764e+06	4.240994e+06	345824747	8182.37
SOUTHEAST	6682588	95	3.348249e+06	4.016567e+06	318083652	7454.78
DELRIDGE	6467837	85	3.226518e+06	4.647933e+06	274254017	5581.70

Analysis:

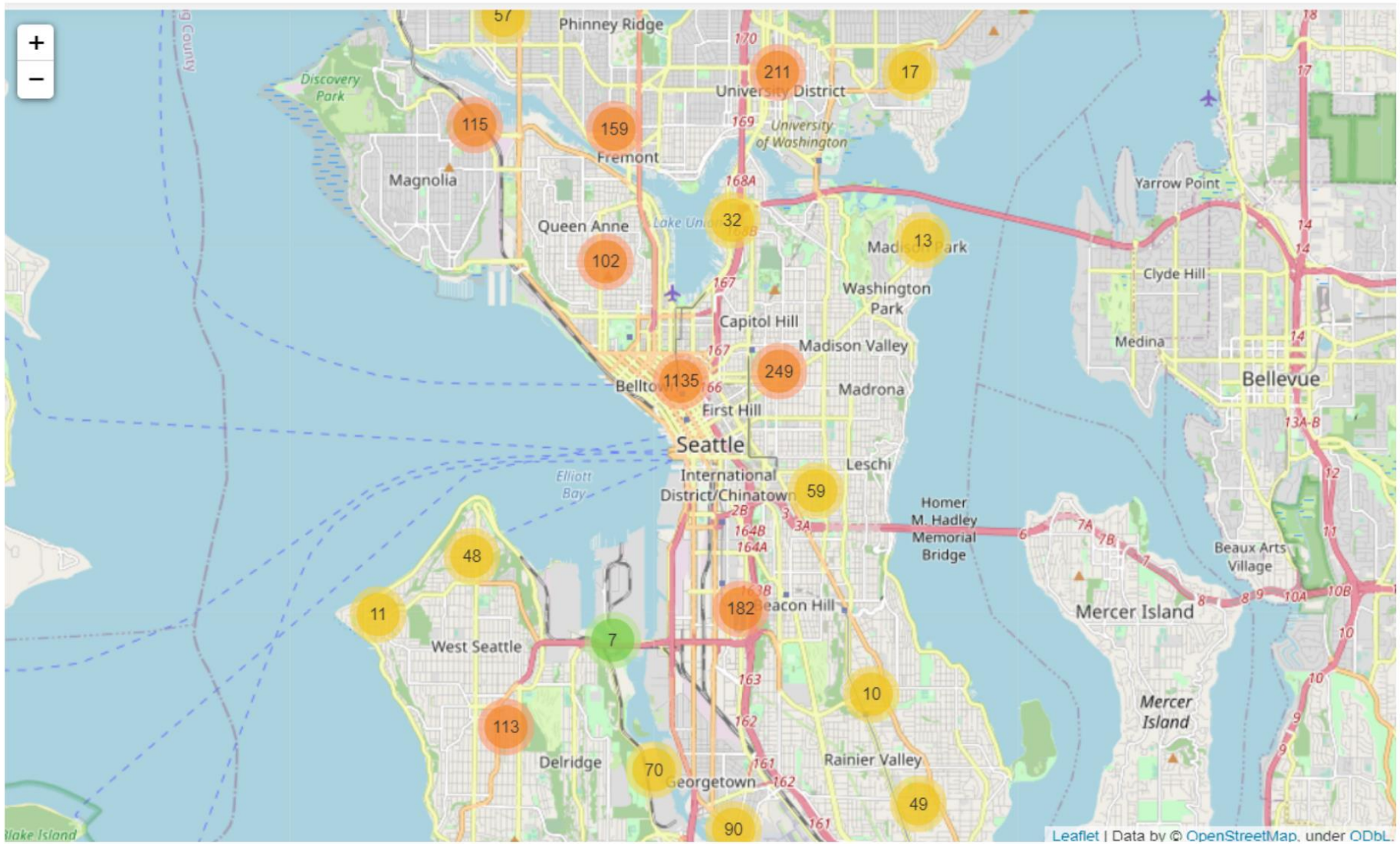
Pivot Table (Neighborhood by GFA Total, Energy Use and GHG Emission)

We can see:

1) The area where energy is used the most is Downtown that also generate GHG emission the most.

2) As comparing the area North with Central, the energy used in North is more than Central but the GHG emission in North is less than Central, we may conclude that buildings in energy use is more effective in North than Central.

Distribution of the number of Buildings in location



Model Development

In this section, we will develop models to attempt prediction in Green House Gas (GHG) emission from a building with its features' relationship, so as to get any significant insight from the features affecting the GHG emission.

Previously, we conducted analysis on some data features and found out 'Energy Use' has a closed relation with GHG emission.

However, this is not the end. We have to testify other features and optimize them to get a better hypothesis. Ordinary Least Squared (OLS) Regression will be used for that in the next stage.

The features, or what we call the independent variables, will be applied for the model training and testing, are:

NumberofBuildings, NumberofFloors, PropertyGFATotal, SiteEnergyUse(kBtu), harversine_distance, BuildingAge, TotalUseTypeNumber

Finally, we will use Polynomial Regression to predict a building the amount of GHG emission under the independent variable(s).

Regression

Process:
1) Import modules

2) Split data to train and test variables

Independent variables:
NumberofBuildings, NumberofFloors, PropertyGFATotal, SiteEnergyUse(kBtu), harversine_distance, BuildingAge, TotalUseTypeNumber

Dependent variable (the prediction/ target variable):
TotalGHGEmissions

3) Ordinary Least Squares (OLS) Regression

4) Polynomial Regression

5) Prediction of the dependent variable

Ordinary Least Squares (OLS) Regression

We develop OLS to describe the relationship between independent quantitative variables and a dependent variable.

OLS Regression Results

Dep. Variable:	TotalGHGEmissions	**R-squared:**	0.805
Model:	OLS	**Adj. R-squared:**	0.804
Method:	Least Squares	**F-statistic:**	1523.
Date:	Fri, 21 Oct 2022	**Prob (F-statistic):**	0.00
Time:	13:15:45	**Log-Likelihood:**	-17023.
No. Observations:	2598	**AIC:**	3.406e+04
Df Residuals:	2590	**BIC:**	3.411e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	47.8731	12.762	3.751	0.000	22.849	72.897
NumberofBuildings	-0.2182	4.176	-0.052	0.958	-8.407	7.970
NumberofFloors	-9.0758	0.877	-10.349	0.000	-10.795	-7.356
PropertyGFATotal	-9.091e-05	3.96e-05	-2.294	0.022	-0.000	-1.32e-05
SiteEnergyUse(kBtu)	2.647e-05	3.23e-07	82.057	0.000	2.58e-05	2.71e-05
harversine_distance	-3.1106	1.602	-1.942	0.052	-6.252	0.031
BuildingAge	0.0790	0.110	0.719	0.472	-0.136	0.294
TotalUseTypeNumber	-8.4935	3.233	-2.627	0.009	-14.834	-2.153

Omnibus:	3999.088	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18187490.848
Skew:	-8.540	**Prob(JB):**	0.00
Kurtosis:	412.539	**Cond. No.**	5.72e+07

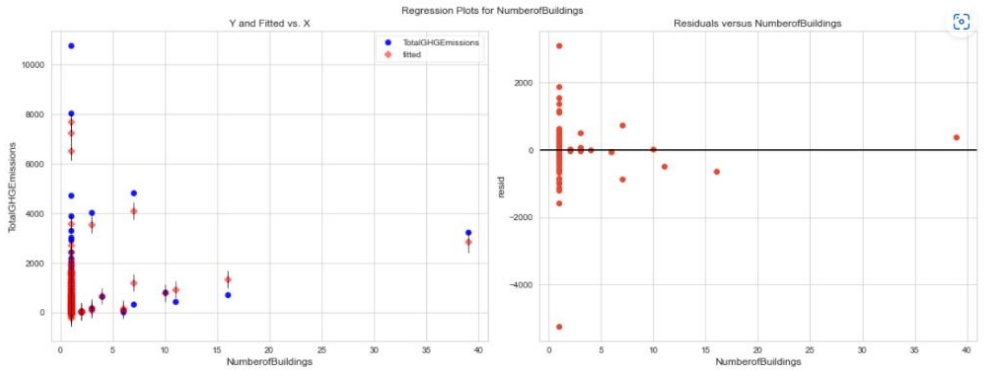
Analysis:

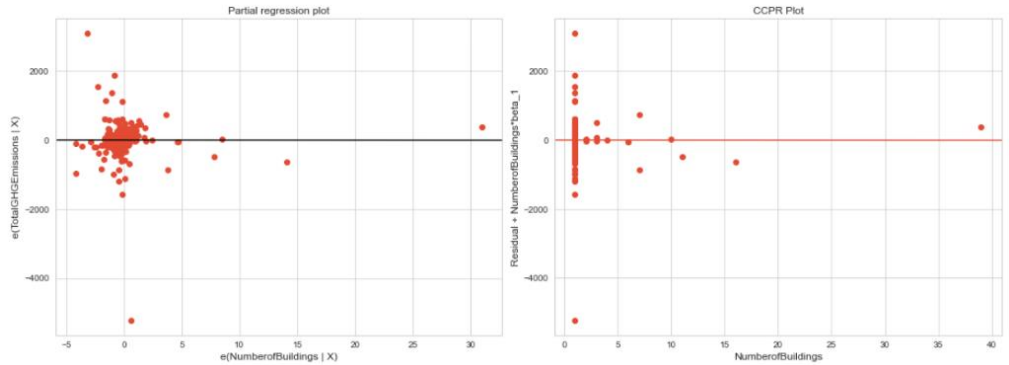
1) t-statistic (t>1.96 or t<-1.96) and P-value <0.05, independent variable is acceptable. In this case, NumberofFloors, PropertyGFATotal, SiteEnergyUse(kBtu), and TotalUseTypeNumber are acceptable.

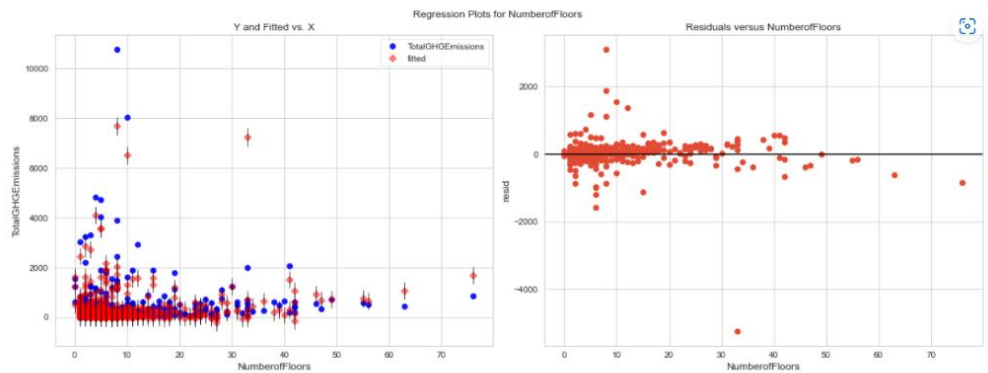
2) R-squared is that % variation in dependent variable is explained due to independent variables. In this case, ~80% variation in GHG emission is explained due to the 7 independent variables

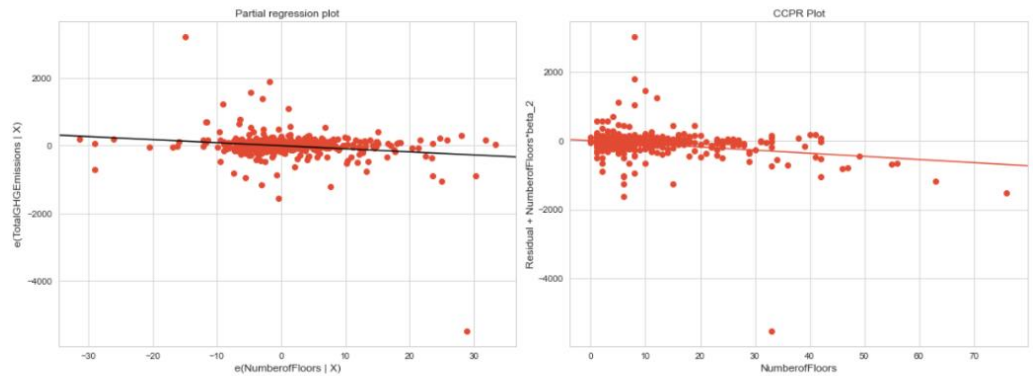
Regression Plots

Now we plot regression and residuals to see any insight from the independent variables. The following are just for reference and may be accessed to the link from appendix for details.









Variance Inflation Factor (VIF)

Now we use VIF for verification of multicollinearity of independent variables, in order to see the degree of variables related to each other. The result of VIF may affect the combination of the independent variables to the end result of dependent variable. For instance, there are 4 variables, A, B, C, and D. A, B, C are independent variables and D is the dependent variable. A, B, C have different degrees of impact on D in different combination, such as ABC, AB, BC, and AC, they will come up different amounts to D.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor as vif

for i in range(len(X_train_cont_features.columns)):
    v=vif(np.matrix(X_train_cont_features),i)
    print("Variance inflation factor for {}: {}".format(X_train_cont_features.columns[i],round(v,2)))
```

```
Variance inflation factor for NumberofBuildings: 2.57
Variance inflation factor for NumberofFloors: 2.86
Variance inflation factor for PropertyGFATotal: 4.05
Variance inflation factor for SiteEnergyUse(kBtu): 2.02
Variance inflation factor for harversine_distance: 1.95
Variance inflation factor for BuildingAge: 2.21
Variance inflation factor for TotalUseTypeNumber: 2.93
```

Analysis:

VIF >5 but less than 10 is a concern for the collinearity. If VIF >10, it indicates a serious collinearity. In this case, it does not constitute a serious problem in collinearity.

Polynomial Regression

Independent variables:
1) Considering t-statistic (t>1.96 or t<-1.96) and P-value <0.05, independent variable is acceptable.

2) According to the OLS Regression table, the independent variables (NumberofFloors, PropertyGFATotal, SiteEnergyUse(kBtu), and TotalUseTypeNumber) are acceptable. However, as having a close look of relationship between NumberofFloors and PropertyGFATotal from the Pearson Correlation (see the table below), we eventually select PropertyGFATotal, SiteEnergyUse(kBtu), TotalUseTypeNumber for the independent variables to the polynomial regression.

*Remark: the more floors a building has, the more GFA it will be.

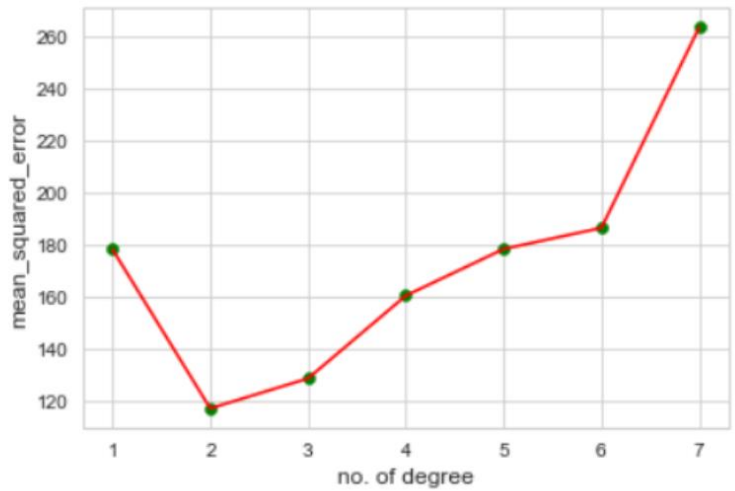
	NumberofBuildings	NumberofFloors	PropertyGFATotal	SiteEnergyUse(kBtu)	harversine_distance	BuildingAge	TotalUseTypeNumber
NumberofBuildings	1.000000	-0.021250	0.304719	0.220195	0.002467	0.001390	0.004999
NumberofFloors	-0.021250	1.000000	0.626292	0.353350	-0.289970	-0.147264	0.172223
PropertyGFATotal	0.304719	0.626292	1.000000	0.659682	-0.162515	-0.166209	0.203721
SiteEnergyUse(kBtu)	0.220195	0.353350	0.659682	1.000000	-0.123735	-0.076039	0.131037
harversine_distance	0.002467	-0.289970	-0.162515	-0.123735	1.000000	-0.268698	-0.127791
BuildingAge	0.001390	-0.147264	-0.166209	-0.076039	-0.268698	1.000000	-0.089680
TotalUseTypeNumber	0.004999	0.172223	0.203721	0.131037	-0.127791	-0.089680	1.000000

Polynomial Regression

Polynomial Degree:
Based on mean squared error (MSE), the number of degrees for the polynomial regression is 2 that will be better.
As seen from the chart below, the lower MSE the better is.

```
df=data_filter
X, y = df[['PropertyGFATotal', 'SiteEnergyUse(kBtu)', 'TotalUseTypeNumber']], df['TotalGHGEmissions']
```

```
# Check accuracy for each degree, the lower the error the better!
number_degrees = [1,2,3,4,5,6,7]
plt_mean_squared_error = []
for degree in number_degrees:

    poly_model = PolynomialFeatures(degree=degree)

    poly_x_values = poly_model.fit_transform(X)
    poly_model.fit(poly_x_values, y)

    regression_model = LinearRegression()
    regression_model.fit(poly_x_values, y)
    y_pred = regression_model.predict(poly_x_values)

    plt_mean_squared_error.append(mean_squared_error(y, y_pred, squared=False))

plt.scatter(number_degrees,plt_mean_squared_error, color="green")
plt.plot(number_degrees,plt_mean_squared_error, color="red")
plt.ylabel("mean_squared_error", size=12)
plt.xlabel("no. of degree", size=12)
```



Polynomial Regression

Comparing Linear Regression with Polynomial Regression by Root Mean Squared Error (RMSE):

In this case, the polynomial regression model performs twice better than the linear regression model:

Linear Regression's RMSE ~ 245.09
Polynomial Regression's RMSE ~ 116.51

Therefore, using Polynomial Regression with degree 2 is better than Linear Regression for the prediction.

```
#Polynomial degree = 2
df=data_filter
X, y = df[['PropertyGFATotal', 'SiteEnergyUse(kBtu)', 'TotalUseTypeNumber']], df['TotalGHGEmissions']
poly = PolynomialFeatures(degree=2, include_bias=False)
poly_features = poly.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(poly_features, y, test_size=0.3, random_state=42)
```

```
poly_reg_model = LinearRegression()
poly_reg_model.fit(X_train, y_train)
```

```
LinearRegression()
```

```
poly_reg_y_predicted = poly_reg_model.predict(X_test)
from sklearn.metrics import mean_squared_error
poly_reg_rmse = np.sqrt(mean_squared_error(y_test, poly_reg_y_predicted))
poly_reg_rmse
```

```
116.51215038292571
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
lin_reg_model = LinearRegression()
lin_reg_model.fit(X_train, y_train)
lin_reg_y_predicted = lin_reg_model.predict(X_test)
lin_reg_rmse = np.sqrt(mean_squared_error(y_test, lin_reg_y_predicted))
lin_reg_rmse
```

```
245.09549330622679
```

Model coefficients of Linear Regression and Polynomial Regression:

```
print("Linear Regression Model coefficient:", lin_reg_model.coef_, "intercept", lin_reg_model.intercept_)
```

```
Linear Regression Model coefficient: [-1.70766573e-04  2.31709123e-05 -9.07080011e+00] intercept 21.545878255421712
```

```
print("Polynomial Regression Model coefficient:", poly_reg_model.coef_)
```

```
Polynomial Regression Model coefficient: [-2.87556122e-04  2.81125054e-05  1.69706264e-01 -1.66218359e-09
  3.82108280e-11  1.00706026e-04 -9.79253337e-14 -2.70958952e-06
  1.17095568e+00]
```

Prediction of GHG emission

Prediction of GHG emission from the building with these values:

```
GFA = input("Enter PropertyGFATotal value: ")
EU = input("Enter SiteEnergyUse(kBtu) value: ")
TN = input("Enter TotalUseTypeNumber value: ")
```

```
Enter PropertyGFATotal value: 378525
Enter SiteEnergyUse(kBtu) value: 25476332
Enter TotalUseTypeNumber value: 3
```

```
poly_reg_model.predict(poly.fit_transform([[GFA,EU,TN]]))
```

```
C:\Users\dlaminus\anaconda3\lib\site-packages\sklearn\base.py:
decimal numbers if dtype='numeric'. This behavior is deprecate
onvert your data to numeric values explicitly instead.
  X = check_array(X, **check_params)
```

```
array([591.80402319])
```

Let's say,

the building GFA total amount is 378,525 sqft,
the total energy use is 25,476,332 kBtu,
the number of building types in the building is 3;

Therefore, the prediction amount of the GHG emission from the building is ~591.80 (MetricTonsCO2e).

Prediction of GHG emission from the building

Conclusion:

There is no perfect prediction, it is just for reference. Based upon the dataset, we came up the 3 independent variables have higher degree of impact on GHG emission. There are many more we can dig out if data available, such as the weather, the habit people have in using energy, the number of people occupied in the building, the number of devices and appliances with EnergyStar, and green design features in the building. Those are the factors that can affect these 3 independent variables. However, that does not mean the more would be better, we have to consider whether they are acceptable under criteria, and the more variables we have, the more complicated the regression will be.

Appendix

For detail of the python programming about this report, please access to the link below:

Report (Jupyter Notebook Viewer)

End