Rethinking Baseline Greenhouse Gas Emissions with Data Science

Goal

Can you estimate the energy use intensity (EUI) of these new buildings if they were to have existed back in 2006 using historical data from buildings?

$$\text{EUI} = \frac{\text{annual energy use (KBtu)}}{\text{building square footage (GSF)}}$$

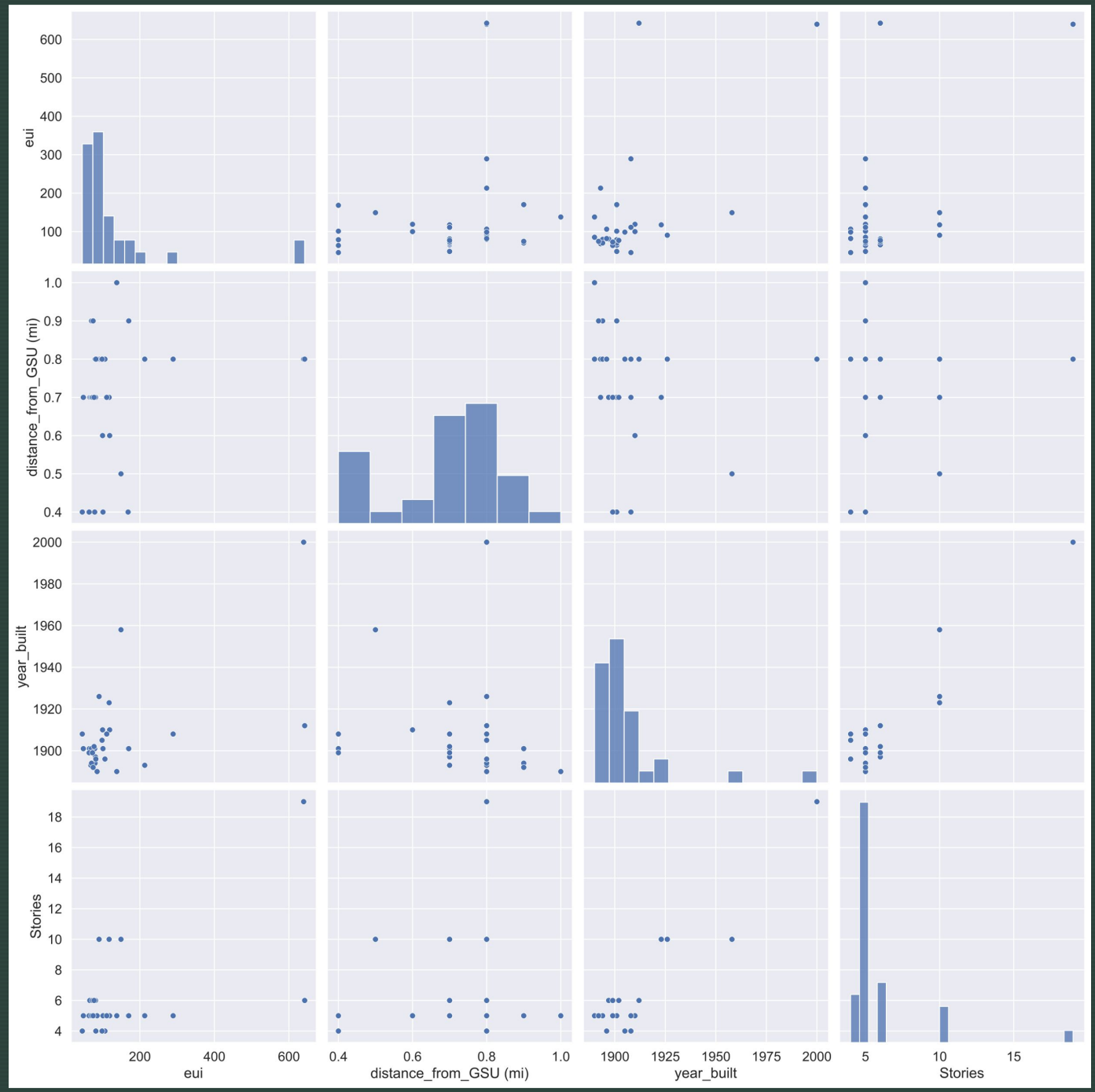
Data Collection

- Unsatisfied with provided data, decided to collect more data on buildings
- Collected data included number of stories, height, distance from GSU, and year built
- Used BU's IMPs to find year built, height, and number of stories
- Used OCR scanner software to turn IMPs into usable excel spreadsheets
- Used Google Maps API to extract latitude and longitude of each building
- Used Haversine formula to turn coordinates into distance from GSU
- Spent time cleaning and merging data into original data sets
- Used pandas, regular expressions, and fuzzy matching algorithms to merge data sets together.

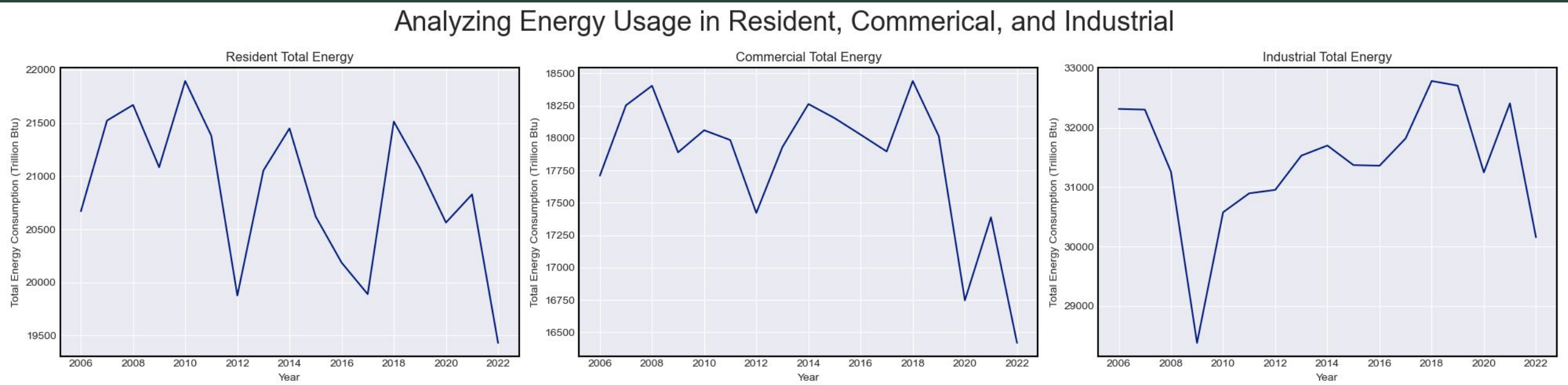
Address	Year Built	Stories	Height	Bldg Area	Land Area	Use
	/ ca 1958					
541 Comm Ave	1894	B + 6	70'	21,196		Coml
565 Comm Ave	1952	B + 2	21' 4"	12,908	16,216	Acad
577-601 Comm Ave, SMG	1996	3L+ 9+ P	166' 2"	481,119	49,686	Acad
582-588 Comm Ave, Sci Ctr		B + 5	60'	102,500	23,395	Pkg
590-596 Comm Ave, Sci Ctr	1983	B + 4	57'	167,000	52,048	Acad
602 Comm Ave , Morse	1907	B + 3		21,919	13,508	Acad
617-621 Comm Ave		B + 4	51' 4"	22,762	9,422	Acad
622-640 Comm Ave, COM	1956	B + 3	47'	84,022	67,232	Acad
631-639 Comm Ave, Sargent Col	1957	B + 7	75'	113,621	19,225	Acad
645-655 Comm Ave		NA	NA	NA	41,574	Pkg
675 Comm Ave, Stone Science	1938	B + 5	62'	54,527		Acad
675-775 Comm Ave, Central Campus Land		NA		NA	561,605	Acad
685 Comm Ave Building	1939	B + 5 + P	79'	141,257		Acad
700 Comm Ave, Warren Towers Pkg	1966	B + 3		251,712	63,472	Pkg
704 Comm Ave	1910	B + 5		31,552	6,480	Acad/Res
710 Comm Ave	1875	B + 4		5,300		Coml
718 Comm Ave, CLA	1910	B + 5	58'	22,068	5,549	Acad
725 Comm Ave, CAS	1948	B + 6 + P	80'	132,261		Acad
730-732 Comm Ave, Eng	1920 ca	B + 3		58,264		Acad/Coml
735 Comm Ave, Marsh Chapel	1949	B + 2		14,964		Acad
736-738 Comm Ave	1965	B + 1		5,840		Coml
742 Comm Ave		B + 2		64,788		Pkg
745-755 Comm Ave, School Theology	1947	B + 6 + P		114,978		Acad
750 Comm Ave	1929	B + 2		44,749	118,483	Acad
756-766 Comm Ave		NA	NA	NA		Pkg
763 Comm Ave, Heating Plant						Acad
765 Comm Ave, Law School	1964	B + 21	232' 5"	167,671	6,480	Acad
767 Comm Ave, Law Library	1964	B + 3		28,616		Acad
771 Comm Ave, Mugar Library	1966	B + 6 + P	114'	218,657		Acad
775 Comm Ave, Student Union	1963	B + 5		202,105		Acad
785 Comm Ave, BU Acad	1931	B + 3	27'	54,767	84,404	Acad
795 Comm Ave		NA		NA	35,474	open
808 Comm Ave, Fuller Building	1928	B + 6 + P	77'	266,029	138,710	Acad
834-846 Commonwealth	1920	B + 2	18'	36,153	35,443	Coml
855 Comm Ave, School of Fine Arts	1919	B + 5	81' 5"	207,318	76,456	Acad
871 Comm Ave, Coll of General Studies	1924	5	43'	95,968	41,520	Acad
881 Comm Ave	1917	B + 7	95'	107,773	26,000	Admin
888 Comm Ave	1924	B + 3		99,352		Admin/Coml

Building Code	Property Type	distance_from_GSU (mi)	year_built	Stories	Height	E (kWh)	G(therms)	O(gallon)	Building Gross Footage
500	Residential	0.8	2000	19	195	13205242	1658548	253894.7	384941.16
506	Residential	0.6	1910	5	44	15476	852	2571.7	4961.74
535	Residential	0.6	1910	5	44	83808	9248	0	4951.86
508	Residential	0.4	1901	5	40	21382	7269	0	8621.75
509	Residential	0.8	1926	10	80	1201216	4408	99773	203525.46
510	Residential	0.7	1895	5	51	18803	0	0	11906.9
720	Residential	0.7	1895	5	51	8663	0	3263.2	6507.28
512	Residential	0.7	1923	10	96	42136.42	13568	0	8736.45
518	Residential	0.7	1923	10	96	1266800	5469	60751.4	113335.15
514	Residential	0.7	1899	6	58	26833	1658	3544.3	11435.95
515	Residential	0.4	1901	5	52	80168	10035	3775.8	10703.48
560	Residential	0.4	1901	5	52	44199	2890	0	7200.27
525	Residential	0.4	1899	4	41	23609	0	0	5128.24
519	Residential	0.7	1899	6	69	60568	20143	0	28553.25
522	Residential	0.5	1958	10	80	1457880	91418	35656.9	127815.2
527	Residential	0.5	1896	5	516	52164	6846	0	8140.83
528	Residential	0.8	1893	5	55	90699	3420	6648.7	7397.32
718	Residential	0.8	1893	5	55	26250	0	3547.6	9622.17
529	Residential	0.5	1901	5	50	22961	0	3343.9	6082.02
736	Residential	0.5	1901	5	40	35873	0	3445.3	7184.56
532	Residential	0.4	1903	4	43	53468	2514	0	4903.89
533	Residential	0.4	1901	5	47	39059	8945	0	6898.98
826	Residential	0.7	1893	5	55	27124	86	4192.6	9997.92
537	Residential	0.4	1901	5	52	38825	3582	0	6546.28
538	Residential	0.4	1901	5	52	44291	4382	0	6820.32
557	Residential	0.4	1901	5	52	31459	6067	0	6933.11
540	Residential	0.4	1908	4	46	26151	0	1700	5085.87
541	Residential	0.4	1901	5	42	34663	1254	3172.9	6780.59
543	Residential	0.4	1901	5	52	28363	1143	1792.6	7128.95
544	Residential	0.4	1901	5	48	38399	305	2837	7057.38

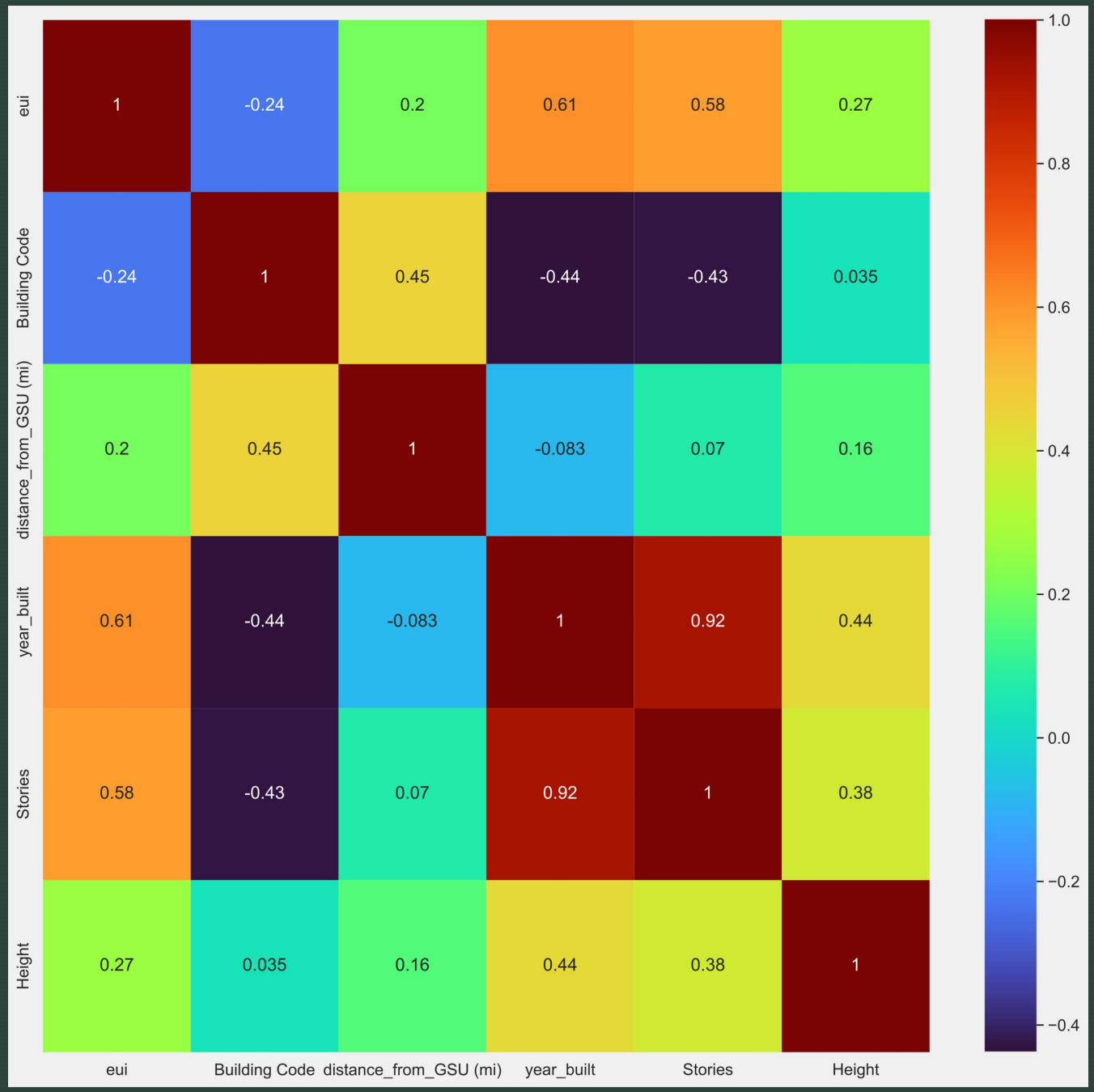
Data Exploration



Looking for Trends in Time Series Data

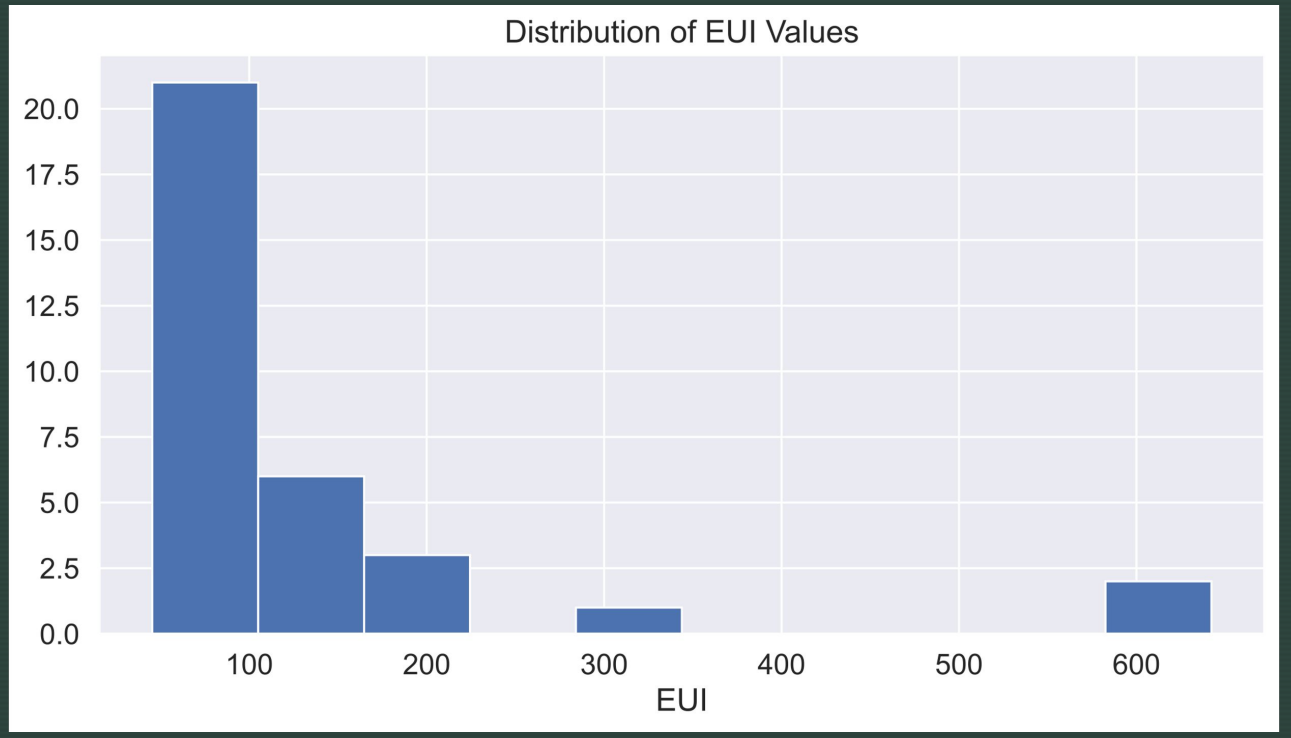


Pearson Correlations of Features with Energy Use Intensity (EUI)



Observations

- Location does not affect energy use intensity (EUI)
- Year built and stories of a building are highly correlated with EUI
- EUI values range from 45 to 650



Challenges

- Small dataset (125 values after processing)
- Limited feature set
- Some of energy, gas, and oil usages provided are 0
  - Assume that data is invalid
- Extrapolating predictions to 2006
- There were not many tall buildings in the dataset, so the model is biased toward shorter buildings.

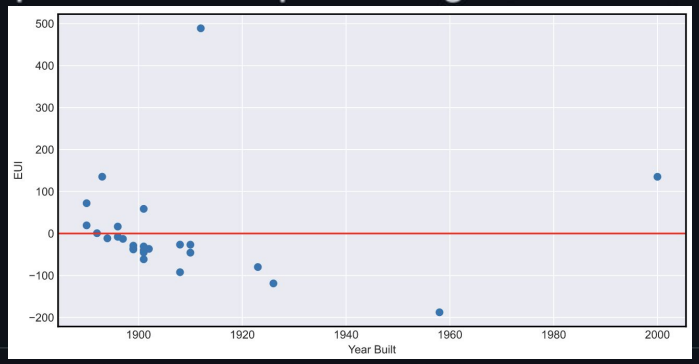
Modeling and Strategies

- Simple Linear Regression
  - Benefits: fast and easy to interpret
  - Disadvantages: often unrealistic of real-world data with noise
- Support Vector Regression
  - Benefits: shown to be very successful in practice
  - Disadvantages: can be computationally intensive and must search for optimal hyperparameters
- K-Nearest Neighbors
  - Benefits: fast, easy to interpret, and no training process
  - Disadvantages: non-parametric algorithm (model grows with dataset size)

Results

- Simple Linear Regression performs well predicting EUI with covariate, year built:

Metric	Value
R^2	0.26
MAE	41.34

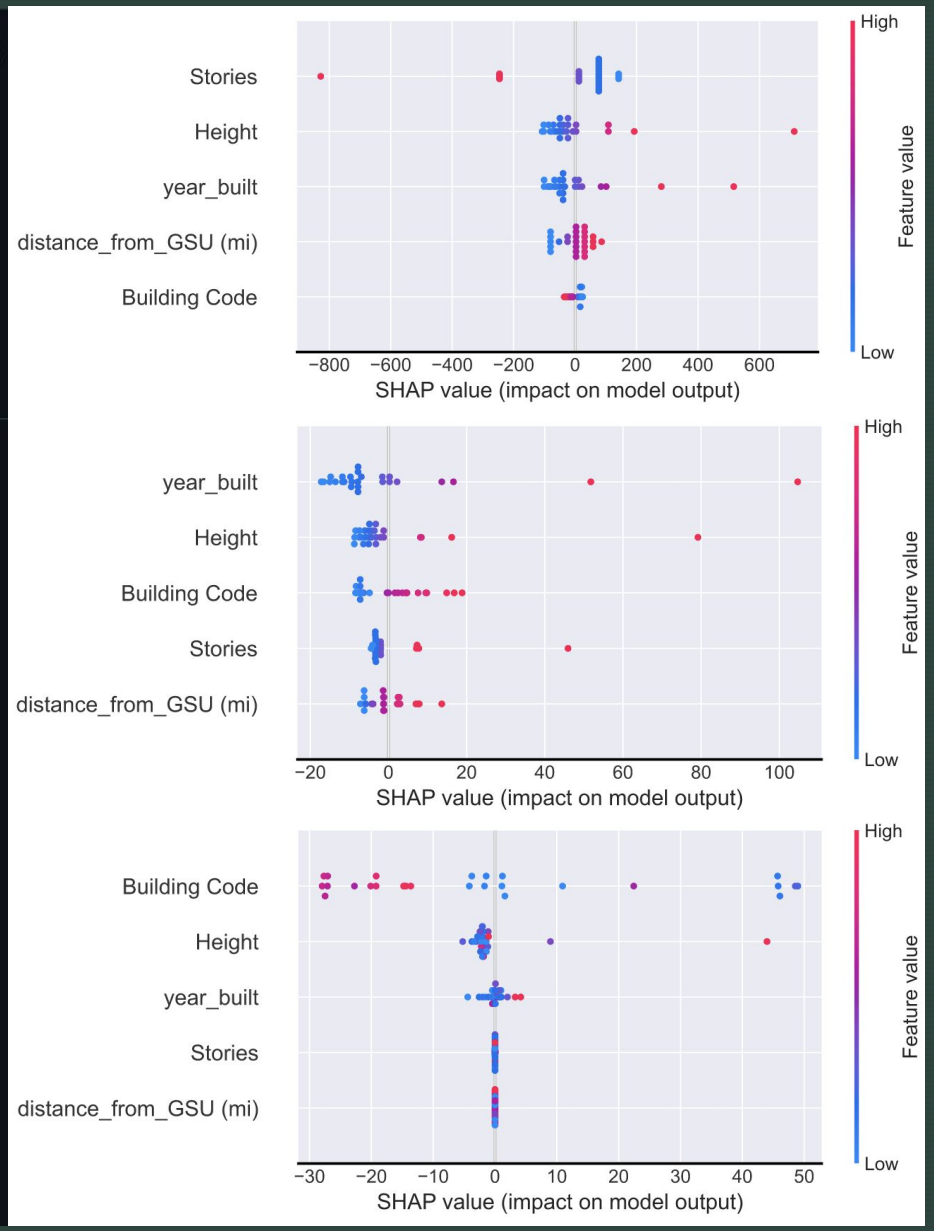


- Support Vector Regression performs well in predicting energy use intensity (EUI):

Metric	Value
R^2	-0.31 (arbitrarily worse than 0)
MAE	49.91

- K-Nearest Neighbors performs best in predicting energy use intensity (EUI):

Metric	Value
R^2	0.44
MAE	42.49



Results using KNN

Code	Address	Predicted_EUI
600	33 harry agganis way, boston, ma	289.7847
918	815 albany street, boston, ma	289.7847
972	peabody hall (210 riverway), boston, ma	182.3993
973	riverway house (160-162 riverway), boston, ma	189.1418
975	campus center and student residence (150 riverway), boston, ma	289.7847
969	37 pilgrim road, boston, ma	182.3993
974	154 riverway, boston, ma	189.1418
608	610 commonwealth avenue, boston, ma	289.7847

Takeaways

- Promising performance for small subset of data and potentially erroneous samples
- With more data and features algorithms such as gradient boosting (XGBoost) and Neural Networks can be explored