Predictive Analysis of Urban Power Consumption in Los Angeles

Data Science and Analytics

GA SLC
2021-2022
PID: 205101

Table of Contents

Introduction

One of the United Nation’s seventeen goals is to achieve sustainable cities and communities. A significant obstacle hindering this goal from realization is rising levels of electricity usage and the pollutants they produce. With California alone accounting for 10% of the United States’s energy use, this analysis seeks to hone in on California flagship metropolis, Los Angeles, and investigate energy use at a hyper-localized level (EIA, 2018). Using publicly accessible data published by the City of Los Angeles, the analysis will iterate through the steps of data cleanup, exploratory data analysis, and the use of a random forest classifier to predict whether a building in Los Angeles is eligible for Energy Star Certification. With this information, officials within the Los Angeles Department of Water and Power (LADWP) can receive a holistic perspective on the factors that go into rendering an energy efficient building.

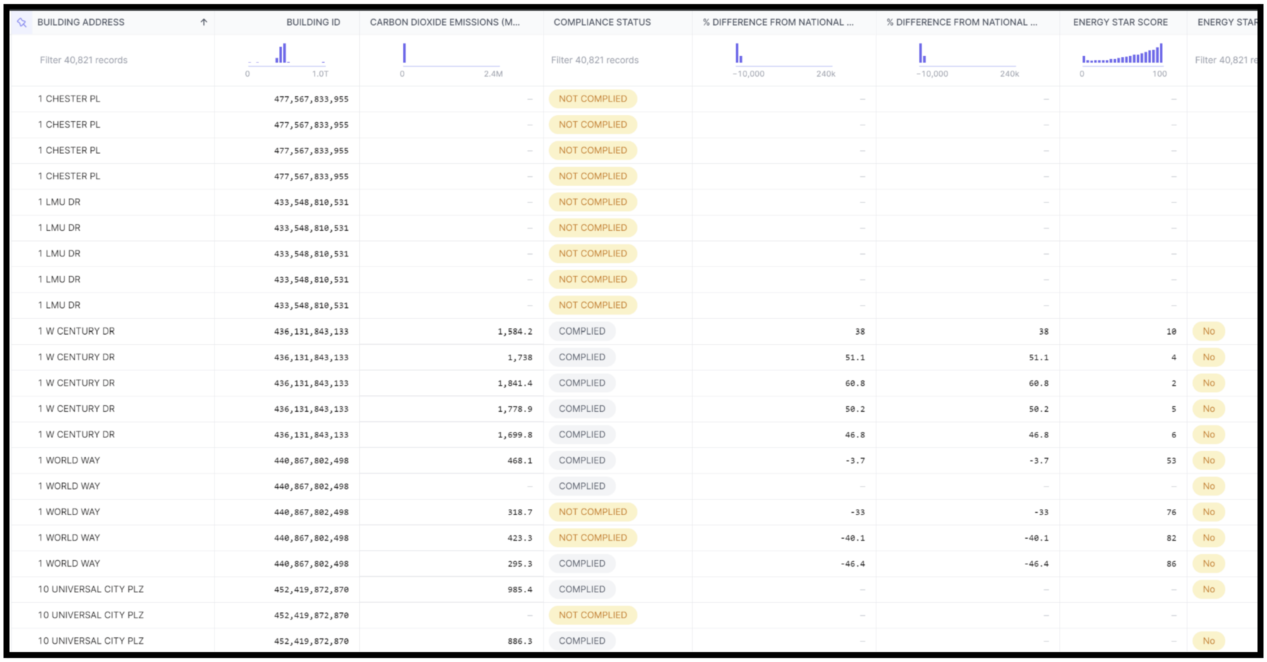
Purpose

Although non-renewable sources of energy are becoming decreasingly sustainable, metropolises around the United States continue to rely on fossil fuels to power their electricity usage. As a result, prolonged electricity use leads to increased demand for fossil fuels, lessening the probability that clean energy can rise to prominence. This issue finds itself growing in Los Angeles, a city desperately working to assimilate clean energy in totality by 2035 (Werner, 2021) Studies estimate that 38% of Los Angeles buildings would need to be fitted with solar panels to meet this ambitious goal—six times the current rate. However, city officials in Los Angeles contend that a switch to clean energy in Los Angeles would set a landmark precedent for other American cities to transition to clean energy as well. Thus, by analyzing Los Angeles buildings' energy consumption, the LADWP can receive critical information on how specific factors (carbon dioxide emissions, source energy usage, water consumption, etc) the tendency of a building to be energy efficient.

Methods

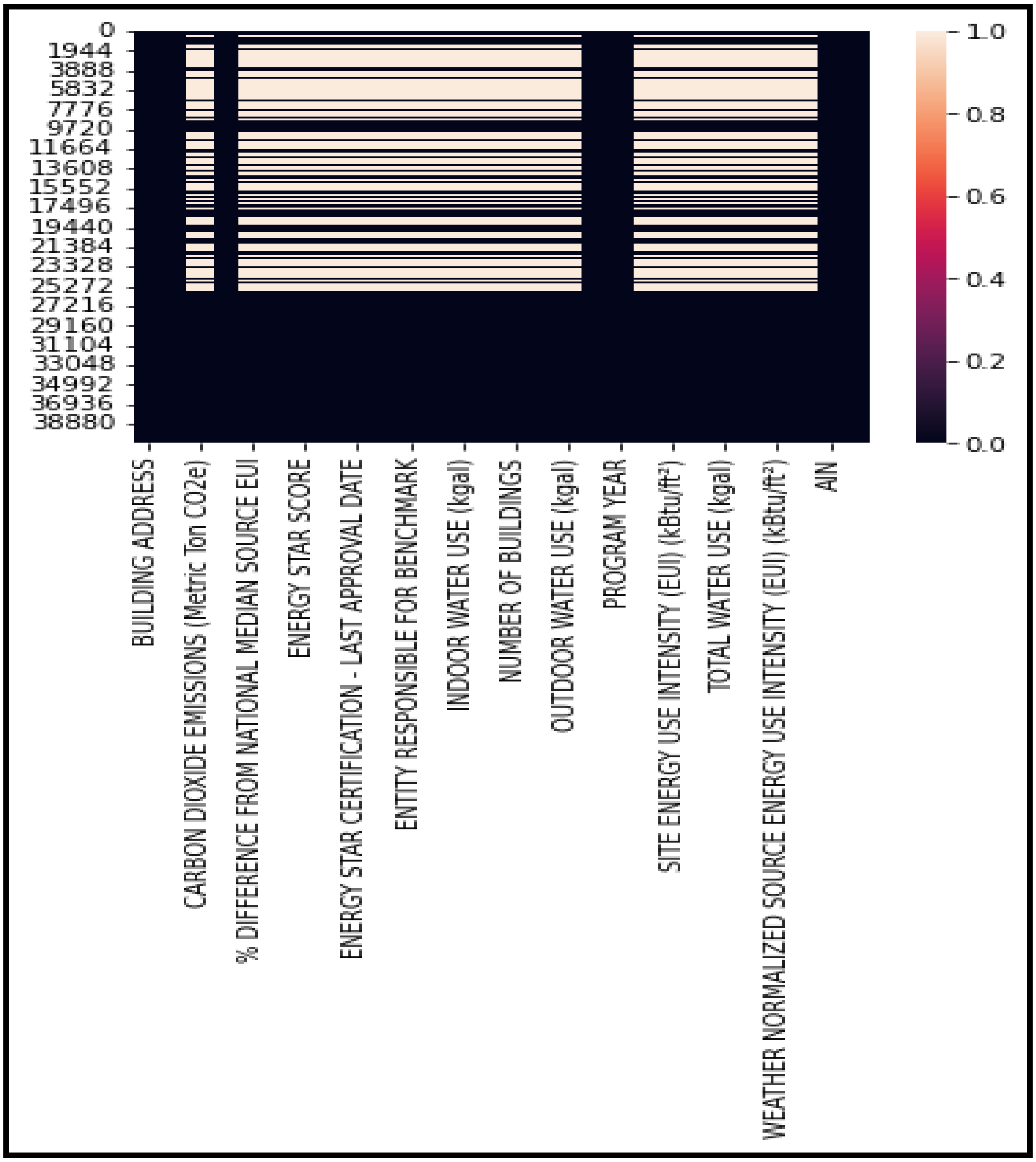
1. Los Angeles OpenData: We utilized Los Angeles's open-source database to sift through datasets concerning citywide energy consumption. By conditioning datasets for relevance and the amount of redacted data, we utilized datasets that were recent (within the past five years) and also accounted for many different variables that may create outliers within the data.
2. Folium: Folium is a graphing module for Python that can leverage GeoJSON mappings of Los Angeles's layout to construct comprehensive maps. These maps pinpoint each building in the dataset onto a visualization of Los Angeles.
3. Decision Trees/Random Forests: Decision Trees are classification algorithms that subdivide datasets into *n* decisions to ultimately predict a binary result. Random Forests accomplish the same goal, but are better suited for datasets with missing or artificially imputed data, a common archetype among publicly available datasets.
4. Scatter Plots: Scatter plots do a good job of demonstrating data of two numeric values and with a lot of variance in data points. It makes it easy to read and understand what the data is showing and the correlation between the two factors.
5. Line Plots: Line plots were used in this analysis to compare numerical factors.
6. Distribution Plots: A distribution plot illustrates how data within one factor is distributed across a specific range. These graphs are critical for spotting potential outliers and disproportionate data.

Results



BUILDING ADDRESS	BUILDING ID	CARBON DIOXIDE EMISSIONS (M...	COMPLIANCE STATUS	% DIFFERENCE FROM NATIONAL ...	% DIFFERENCE FROM NATIONAL ...	ENERGY STAR SCORE	ENERGY STAR
Filter 40,821 records	0 1.0T	0 2.4M	Filter 40,821 records	−10,000 240k	−10,000 240k	0 100	Filter 40,821 re
1 CHESTER PL	477,567,833,955	–	NOT COMPLIED	–	–	–	
1 CHESTER PL	477,567,833,955	–	NOT COMPLIED	–	–	–	
1 CHESTER PL	477,567,833,955	–	NOT COMPLIED	–	–	–	
1 CHESTER PL	477,567,833,955	–	NOT COMPLIED	–	–	–	
1 LMU DR	433,548,810,531	–	NOT COMPLIED	–	–	–	
1 LMU DR	433,548,810,531	–	NOT COMPLIED	–	–	–	
1 LMU DR	433,548,810,531	–	NOT COMPLIED	–	–	–	
1 LMU DR	433,548,810,531	–	NOT COMPLIED	–	–	–	
1 LMU DR	433,548,810,531	–	NOT COMPLIED	–	–	–	
1 W CENTURY DR	436,131,843,133	1,584.2	COMPLIED	38	38	10	No
1 W CENTURY DR	436,131,843,133	1,738	COMPLIED	51.1	51.1	4	No
1 W CENTURY DR	436,131,843,133	1,841.4	COMPLIED	60.8	60.8	2	No
1 W CENTURY DR	436,131,843,133	1,778.9	COMPLIED	50.2	50.2	5	No
1 W CENTURY DR	436,131,843,133	1,699.8	COMPLIED	46.8	46.8	6	No
1 WORLD WAY	440,867,802,498	468.1	COMPLIED	-3.7	-3.7	53	No
1 WORLD WAY	440,867,802,498	–	COMPLIED	–	–	–	No
1 WORLD WAY	440,867,802,498	318.7	NOT COMPLIED	-33	-33	76	No
1 WORLD WAY	440,867,802,498	423.3	NOT COMPLIED	-40.1	-40.1	82	No
1 WORLD WAY	440,867,802,498	295.3	COMPLIED	-46.4	-46.4	86	No
10 UNIVERSAL CITY PLZ	452,419,872,870	985.4	COMPLIED	–	–	–	No
10 UNIVERSAL CITY PLZ	452,419,872,870	–	NOT COMPLIED	–	–	–	
10 UNIVERSAL CITY PLZ	452,419,872,870	886.3	COMPLIED	–	–	–	No

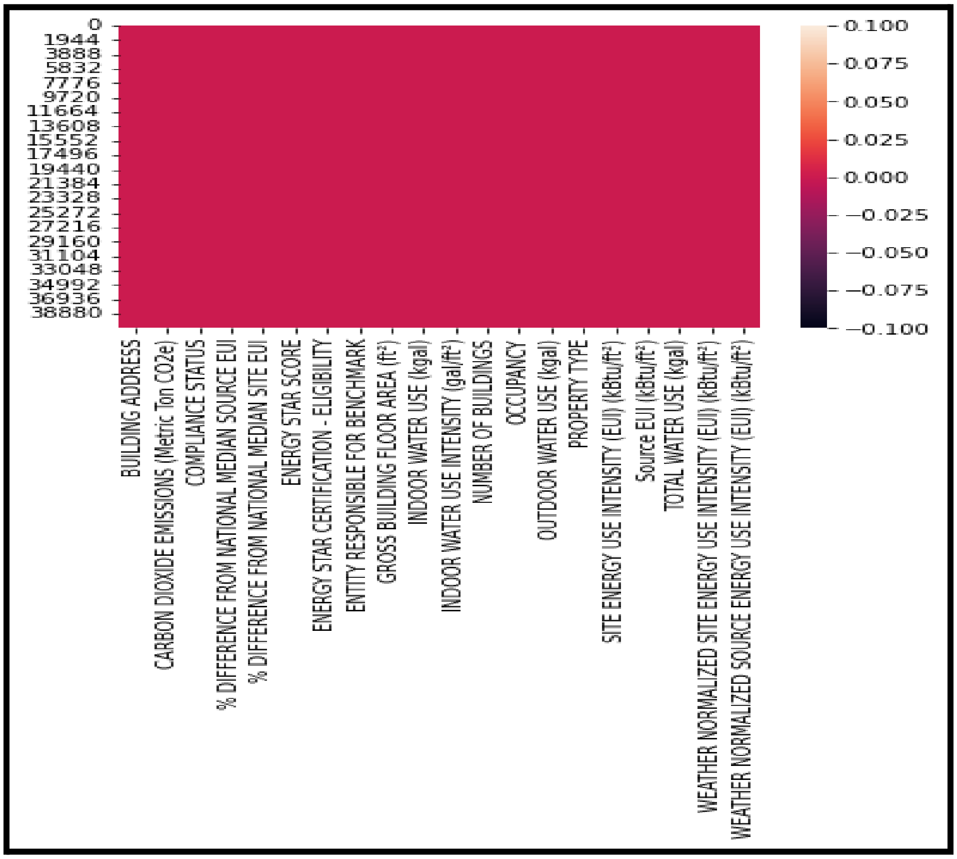
We started our data analysis with a dataset that documents over 80 thousand Los Angeles buildings and characteristics (the columns) of their energy consumption. Particularly notable characteristics include source/site energy consumption, energy star score, water use, carbon dioxide use, and weather-normalized energy consumption. These factors are critical to extracting valuable insights into Los Angeles's holistic energy consumption trend. By leveraging this data, the trends can be analyzed to predict an individual building's energy star certification status, a binary value of "Yes" or "No."



When working with public datasets, oftentimes significant amounts of data are missing. The heatmap above depicts where the missing data lies in the dataset, with the energy factors on the x-axis. As depicted, certain factors have notable amounts of missing data within the dataset that cannot be fixed through normalization or imputation because of limited amounts of existing data. As a result, these columns were dropped using the code below.

```
# Dropping Useless Columns
df = df.drop('POSTAL CODE', axis=1)
df = df.drop('ENERGY STAR CERTIFICATION - LAST APPROVAL DATE', axis = 1)
df = df.drop('LADBS Building Category', axis=1 )
df = df.drop('YEAR BUILT',axis=1 )
df = df.drop('ENERGY STAR CERTIFICATION - YEAR(S) CERTIFIED', axis=1)
df = df.drop('AIN', axis=1)
df = df.drop('PROGRAM YEAR', axis=1)
df = df.drop('BUILDING ID', axis=1)
```

Then, for the rest of the missing data, a normalization algorithm was used to impute data from the row indexed before the current cell. The result was a heatmap illustrating zero discrepancies in available data.



Since, the vast majority of missing data were assigned the “NaN” data type, making the process of filtering this data out of the dataset fairly simple. However, a significant portion of missing data were assigned the string “Not Available.” These cells would render processing the numerical data problematic because algorithms would break attempting to parse over non-numerical values. As a result, we developed the analysis’s first data engineering algorithm to remove all rows that contain a cell with the value “Not Available.”

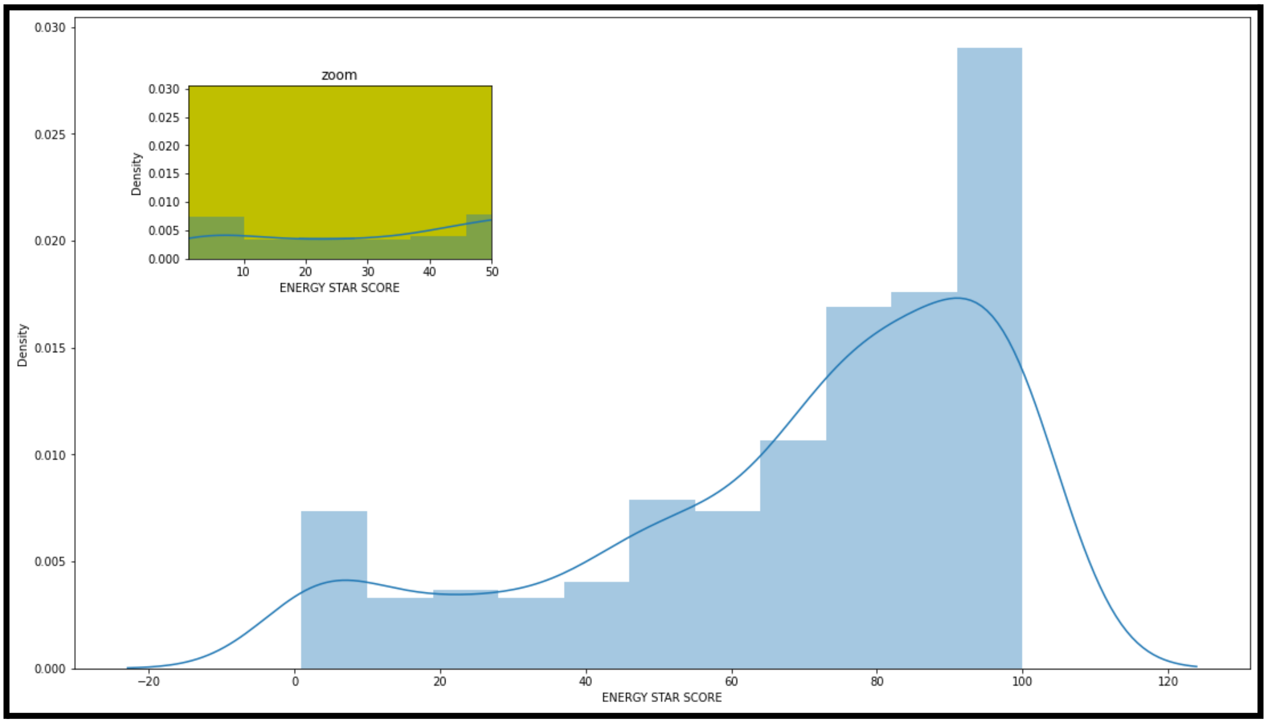
Data Engineering Algorithm #1 (requires Repl.it)

1. First, we initialized a double for-loop to iterate through every row and the contents of every row.
2. For each individual row, if the value “Not Available” occurred in one of its cells, we added 1 to a variable that tracks the number of “Not Available” instances in that row. For every iteration through the row, whether or not the current value was “Not Available,” we added 1 to a variable named “count” that tracked the number of cells in the row.
3. If the for-loop reached the end of the row (the 17th cell) and the variable tracking the number of “Not Available” was 0, this row was appended into a new dataframe that carried only clean rows.
4. The new dataframe with clean rows would henceforth be used for analysis.

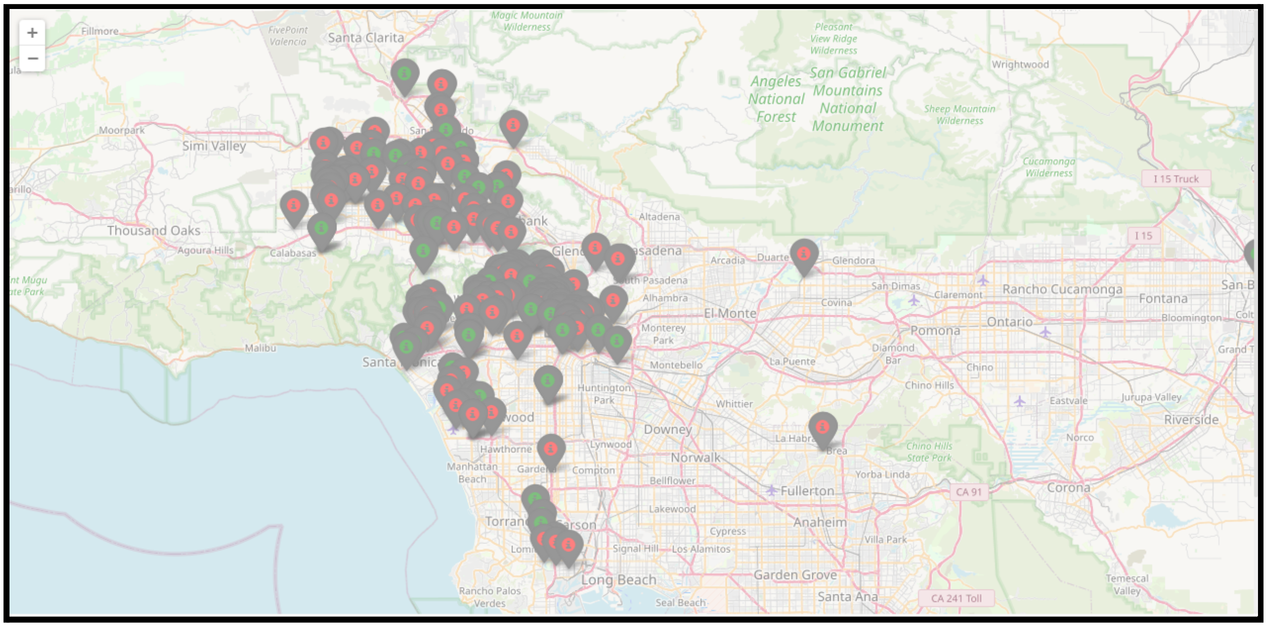
Data Engineering Algorithm #2 (requires Repl.it)

As typical of public datasets, some columns contained numerical data that were actually strings. The implication of this situation is that no processing algorithms would be able to process these seemingly numerical columns because to the algorithm, the data is a string and therefore incompatible for numerical operations. Thus, we developed this algorithm to find columns with numerical data disguised as strings and convert the data into its implied numerical data type.

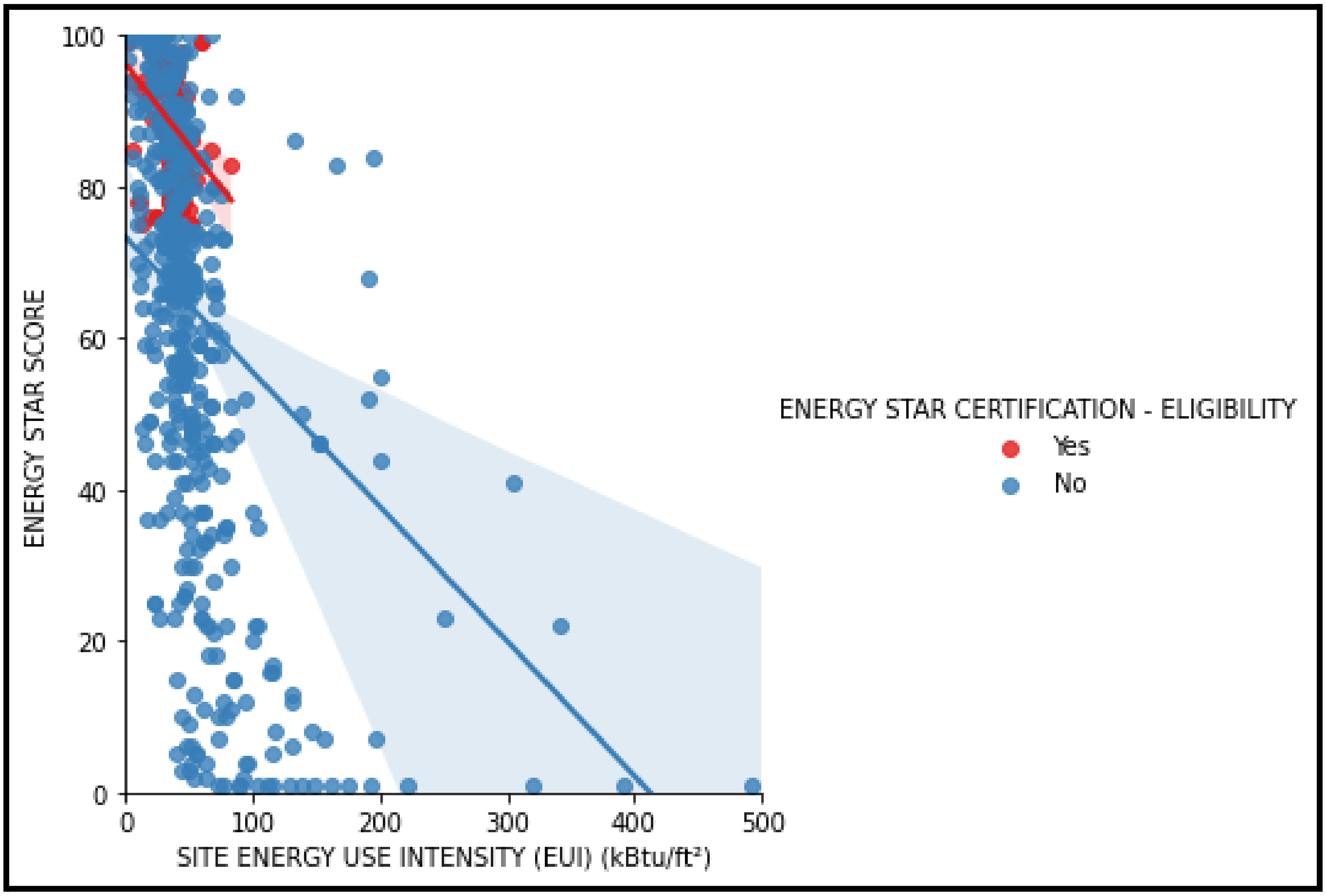
Ending the preliminary data engineering phase of the analysis, we extracted a random sample of 1000 rows from the cleaned dataset. Since the parent dataset contains upwards of 80 thousands rows, computational efficiency would be drastically damaged by attempting to iteratively process or visualize this data. Thus, to ensure fairness and limit bias, we extracted ample data to analyze using a random sampling algorithm.



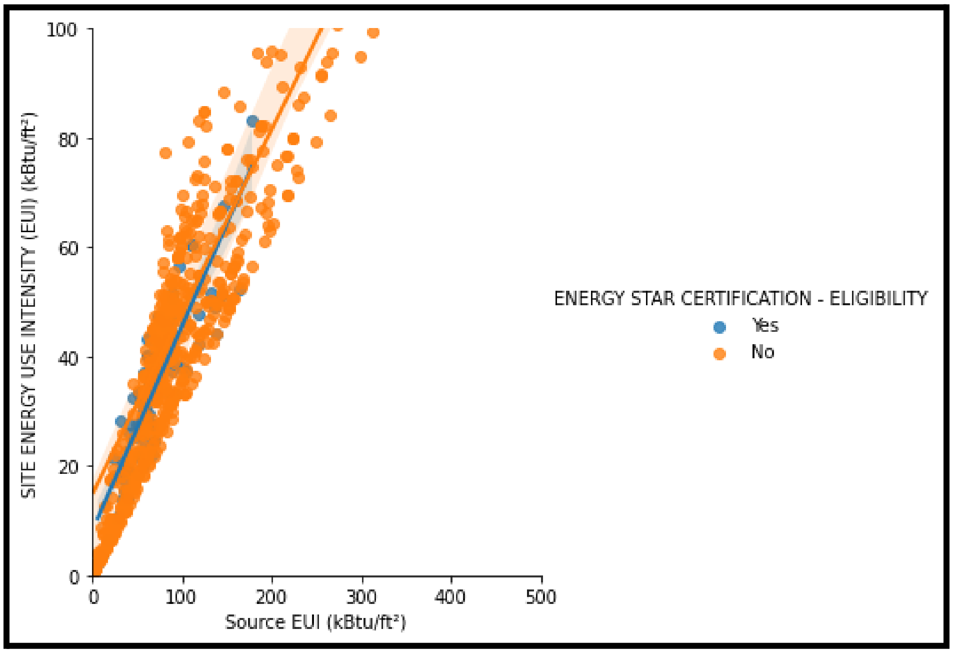
Starting off the exploratory data analysis (EDA) phase, we plotted the distribution of energy star scores attributed to the buildings represented in the dataset. An energy star score is a quantified rating on how energy efficient and high energy performing a building is (EnergyStar, 2014). As depicted, most Los Angeles buildings within this dataset have strong energy star scores, ranging primarily between 80 and 100 (on a 1-100 scale). However, the lower hump in the [0,20] range indicates significant amounts of data that have critically low energy star scores. This insight demonstrates that Los Angeles, although close to reaching holistically high energy star scores, still has remaining stains of abnormally high energy use within the city.



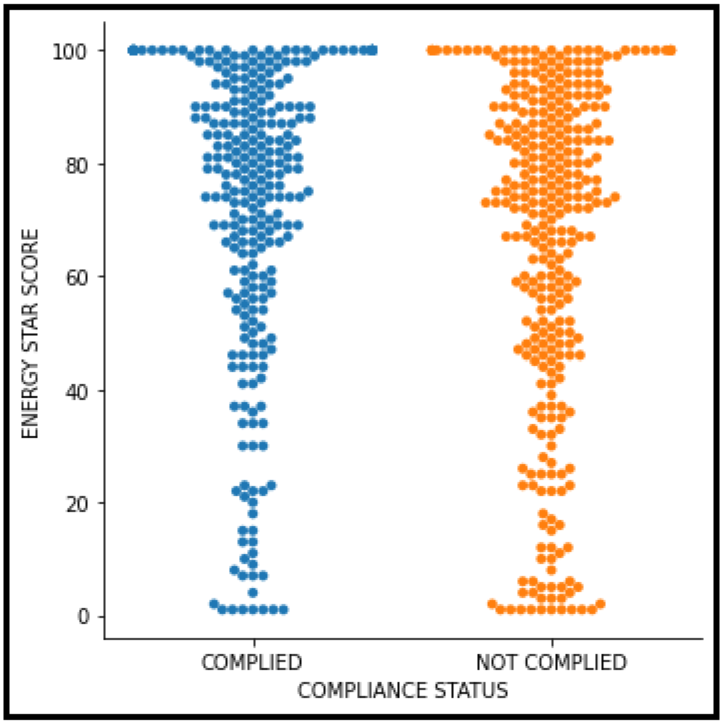
Next, we created a geographical map of all the buildings in this Los Angeles dataset and flagged them with a color that indicates their Energy Star score. Any point with a green icon is a building with an Energy Star score above 75. According to Energy Star, a score of 75 or higher indicates that that building is a top efficient energy user and is eligible for Energy Star certification (EnergyStar, 2014). Any point with a red icon is a building with an Energy Star score less than 75, illustrating they are also not yet eligible for Energy Star certification. The map shows a significant number of buildings with a low Energy Star score, indicating that these buildings must improve in their efficient energy use in order to help the city of Los Angeles reach its goal in 2035.



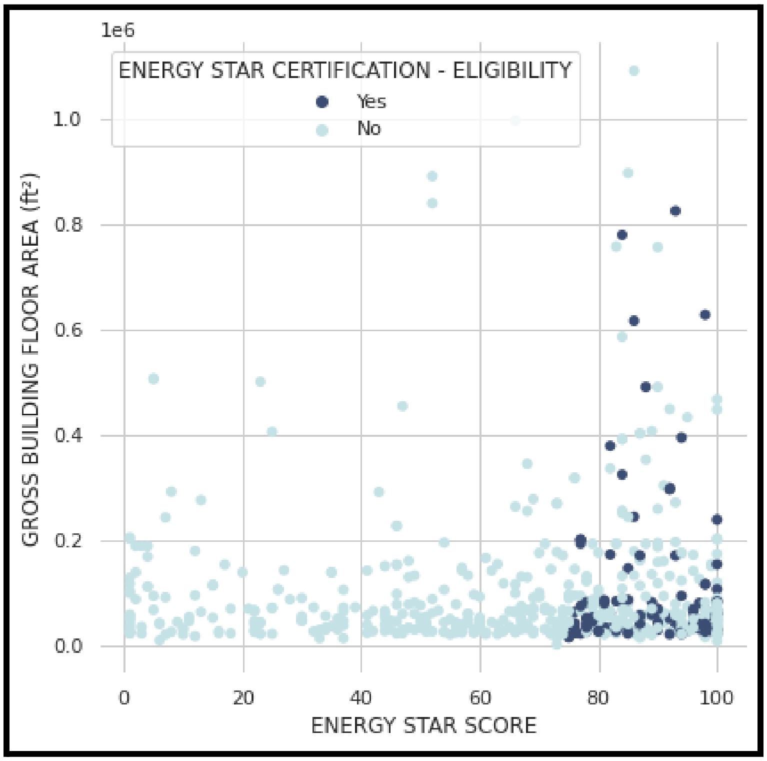
Afterwards, we plotted Site Energy Use Intensity (EUI) versus the Energy Star Score with the Energy Star Certification Eligibility hue applied to the data points. As illustrated, the vast majority of buildings are ineligible for Energy Star Certification, indicating a paradox in Los Angeles energy management when a significant number of buildings have high Energy Star Scores but cannot qualify for certification. Additionally, the ineligible buildings have significantly varying Energy Star Scores, indicating that Energy Star Scores are not the lone determinant for Energy Star Certification and that a more holistic approach to energy usage is applied when deciding certification. As for Site Energy Use Intensity, all buildings that have energy use intensities of over 150 kBtu/ft^2 are not eligible for Energy Star Certification, confirming that Site EUI is influential in deeming Energy Star Certification.



This line/scatter plot hybrid depicts Source EUI versus Site EUI. The strong linear trend confirms that the higher Source EUIs produce higher Site EUIs. Thus, by the transitive property we can likewise consider Source EUI an influential metric on Energy Star Certification.

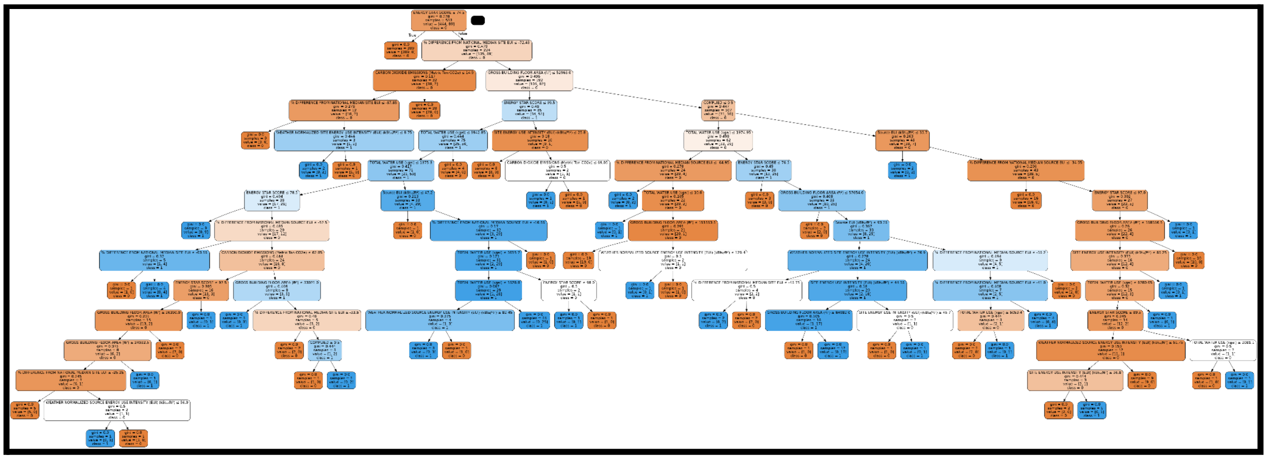


This plot illustrates the swarm distribution of energy star scores under two conditions: whether the building complied with energy use examination or not. The similarity of these distributions indicates that compliance and non-compliance are weak metrics for determining energy star certification because of the impact energy star score leverages on certification.



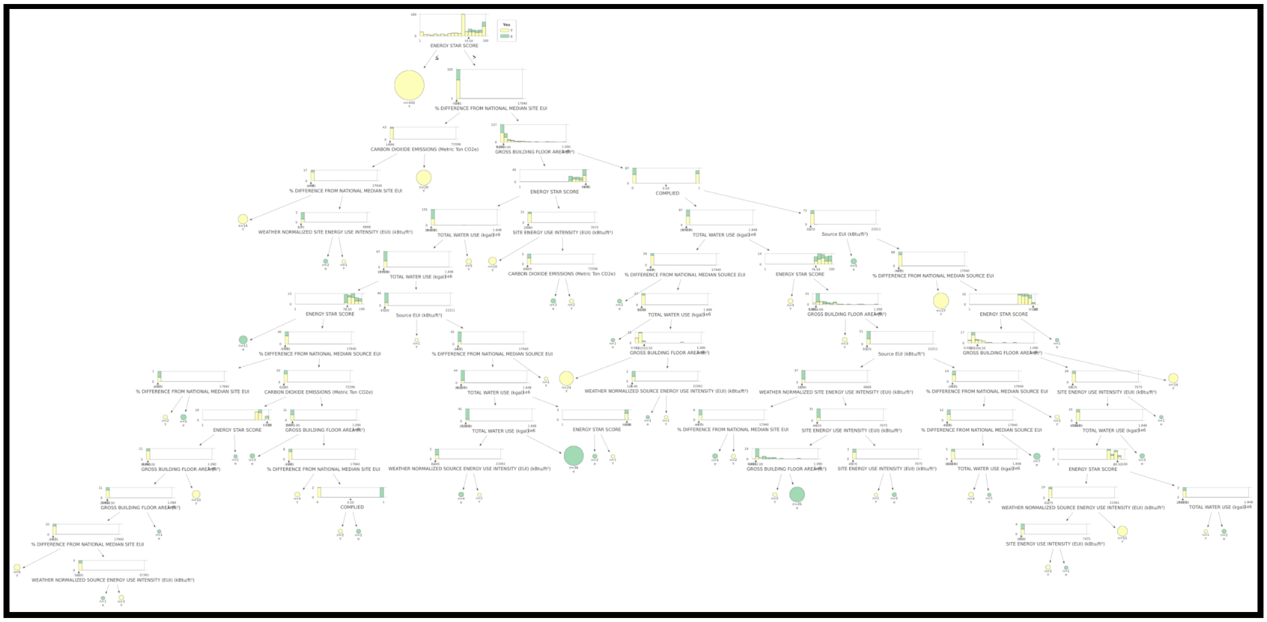
This scatterplot is a visual of a building's Energy Star score and its total floor area, or essentially its size. It also has a color hue that indicates which buildings are and are not Energy Star certification eligibility. This graph shows that the size of the building has a very low effect on the ability to be eligible for certification because of how spread out the data is. Additionally, not all buildings with a score above 75 are automatically granted eligibility, meaning that there are other factors that weigh in on determining eligibility. It also further proves the point that eligibility is not solely determined by Energy Star score because of how spread out the buildings are above a 75 score. This graph helps to confirm that Gross Building Floor Area (ft^2) does not affect the certificate eligibility of a certain building.

Classification Algorithm Code (requires Repl.it)



Decision Tree Render (requires Google Drive)

We chose a decision tree because a decision tree works well with datasets with significant amounts of missing data, which is a commonality in public datasets such as this one. A decision tree classification also gives a more clear understanding and we can see what factors play roles in determining the eligibility of a certain building and its specifications. This decision tree is essentially a step-by-step guide that shows, based on trends analyzed within the dataset, what a building needs to do in order to be eligible for the Energy Star certification. This decision tree rendered an accuracy of 95%, demonstrating that if the LADWP were to input a building's energy efficiency specifications, the decision tree model would have a 95% chance of predicting the eligibility for Energy Star certification.


Yes
Y
ENERGY STAR SCORE
≤
>
n=442
Y
% DIFFERENCE FROM NATIONAL MEDIAN SITE EUI
CARBON DIOXIDE EMISSIONS (Metric Ton CO2e)
GROSS BUILDING FLOOR AREA (ft²)
n=26
Y
ENERGY STAR SCORE
COMPLIED
TOTAL WATER USE (kgal)
SITE ENERGY USE INTENSITY (EUI) (kBtu/ft²)
TOTAL WATER USE (kgal)
Source EUI (kBtu/ft²)
WEATHER NORMALIZED SITE ENERGY USE INTENSITY (EUI) (kBtu/ft²)
% DIFFERENCE FROM NATIONAL MEDIAN SOURCE EUI
WEATHER NORMALIZED SOURCE ENERGY USE INTENSITY (EUI) (kBtu/ft²)

Here is an alternate visualization of the decision tree.

Conclusion

The decision tree further illustrates how Los Angeles must acknowledge their biggest deficits in energy efficiency and find ways to improve at these points. Energy Star certification is important to reaching the goal of 100% clean energy use, and the decision tree shows what factors and what elements of the building would most need to be improved with new techniques, new resources, and new methods of using energy efficiently. Our analysis can be used in several different ways and applications. Most importantly, the LADWP can use these results and conclusions to focus their improvement efforts on specific factors and pieces to the puzzle of energy efficiency. Additionally, since Energy Use Intensity and compliance were both influential to the eligibility, the LADWP can ensure that all of their buildings have a normal EUI for both Site and Source and can enforce compliance to the regulations, which would help toward the city's goal even more. Our analysis can be beneficial to the city of Los Angeles' goals and their efforts toward pursuing this lofty goal they have declared.

Next Steps

Los Angeles' DWP can take a number of steps toward achieving their set goal. In order to protect the environment around Los Angeles, relieve a big cost for most buildings, and to increase the abundance of these nonrenewable natural energy resources. In cooperation with the EPA, LADWP can work on mass improving the city's buildings and their energy efficiency in order to convert all of these buildings into Energy Star certified buildings. Additionally, if more data on other quantified factors about building's energy usage was available, further machine learning and data analysis could be done to find other factors that play into certification and a high Energy Star score, perhaps an expansion pertaining geography and finding parts of Los Angeles with struggling energy efficiencies or certain common problems. There could also be an implementation of more data that determines trends within non large-scale building energy use, such as smaller neighborhoods and suburbs or finding high energy users and abnormally low energy efficiencies to further push for the 100% clean energy goal.

References

EIA. (2018, February 18). U.S. Energy Information Administration - EIA - independent statistics and analysis. California - State Energy Profile Overview - U.S. Energy Information Administration (EIA). Retrieved February 5, 2022, from https://www.eia.gov/state/?sid=CA

Hoffman, J. S. (2014, April 10). How the 1-100 energy star score is calculated. ENERGY STAR Buildings and Plants | ENERGY STAR. Retrieved February 5, 2022, from https://www.energystar.gov/buildings/benchmark/understand_metrics/how_score_calculated

Hoffman, J. S. (2021, August 16). What is energy efficiency? About ENERGY STAR | ENERGY STAR. Retrieved February 5, 2022, from https://www.energystar.gov/about/about_energy_efficiency

Werner, E. (2021, October 28). Los Angeles is aiming to be first major carbon-free U.S. city, but obstacles loom. The Washington Post. Retrieved February 5, 2022, from https://www.washingtonpost.com/climate-solutions/2021/10/27/los-angeles-2035-climate-goal/