## Summary of the Data

The dataset was generated using a building energy simulation tool called 'Honeybee', which is a plugin tool that works via a CAD software called Rhino. The data targets operational energy simulations for buildings located in Los Angeles, California, providing a clear overview of pure operational performance. The pseudo designs are generated using a Monte Carlo approach with random building attributes.

Some of the key attributes of the data that we filtered are:

**PREDICTORS**

| Name | Metric | Data Type |
|---|---|---|
| Orientation | Degrees | Quantitative |
| nonMassWallR | m^2-K/W | Quantitative |
| MassWallR | m^2-K/W | Quantitative |
| RoofR | m^2-K/W | Quantitative |
| ExteriorFloorR | m^2-K/W | Quantitative |
| WWRNorth | R-value | Quantitative |
| WWRWest | R-value | Quantitative |
| WWREast | R-value | Quantitative |
| WWRSouth | R-value | Quantitative |
| SHGC | ratio(percentage) | Quantitative |
| WindowR | R-value | Quantitative |
| numFloor | count(integer) | Quantitative |
| AspectRatio | ratio(percentage) | Quantitative |
| VolumetoFacadeRatio | ratio(percentage) | Quantitative |
| Equipment (0-5) | One-hot-encoded | Categorical |
| Program(0-9) | One-hot-encoded | Categorical |
| WallType(0-3) | One-hot-encoded | Categorical |

**RESPONSE**

| Name | Metric | Data Type |
|---|---|---|
| Operational Energy(OE) | kWh/m^2 | Quantitative |

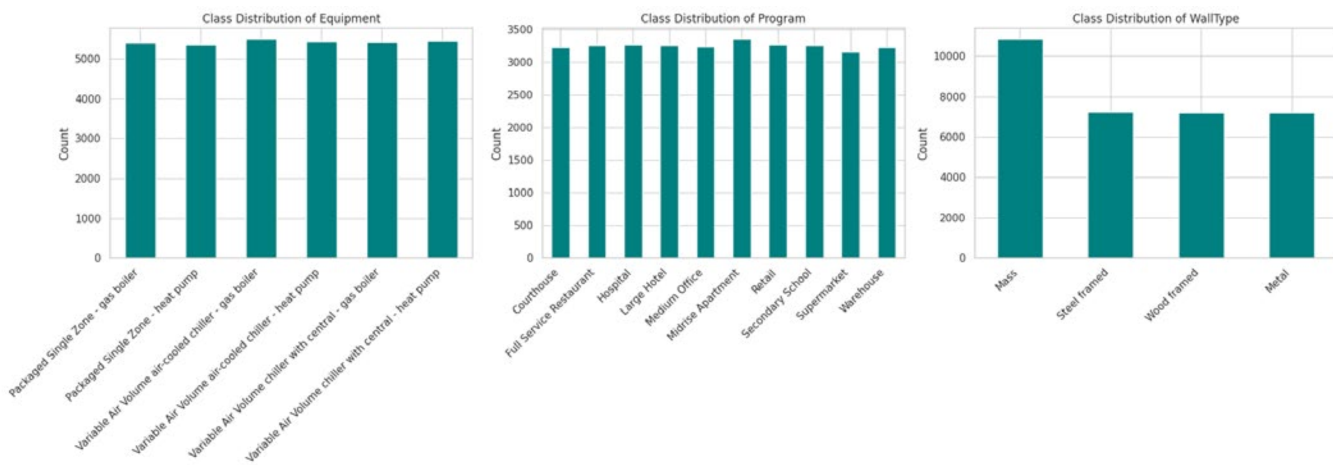
*Please refer to the attached ipynb notebook for more details regarding summary of features and purpose of the project (written in detail)*

## Deeper Understanding

In order to understand our data, we sought for primarily patterns, trends, and relationships of variables, in addition to confirming that the data is good enough to work with and identify inconsistencies or any erroneous/troublesome values.

---

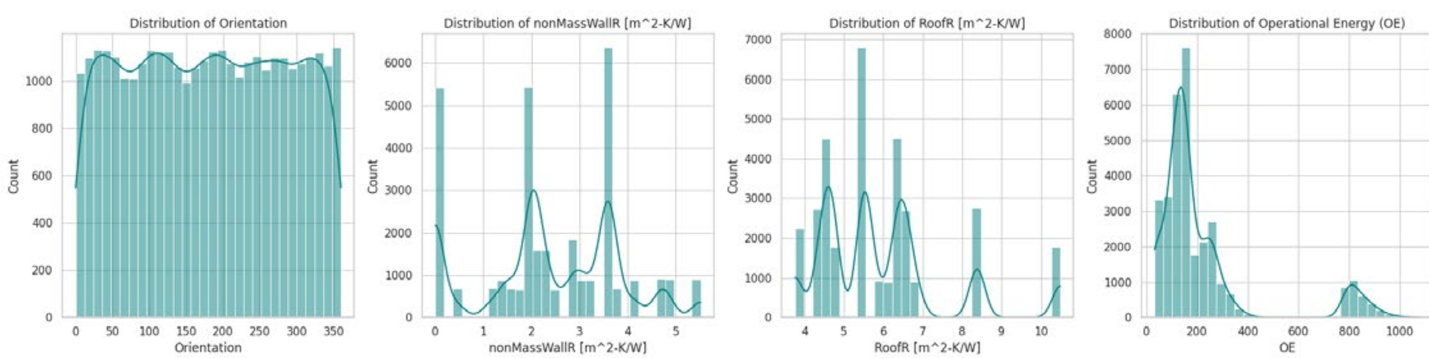
### Data Imbalances




All programs, wall types and equipment are well balanced in terms of data counts, confirming that there are no imbalances.

---

### Distributions

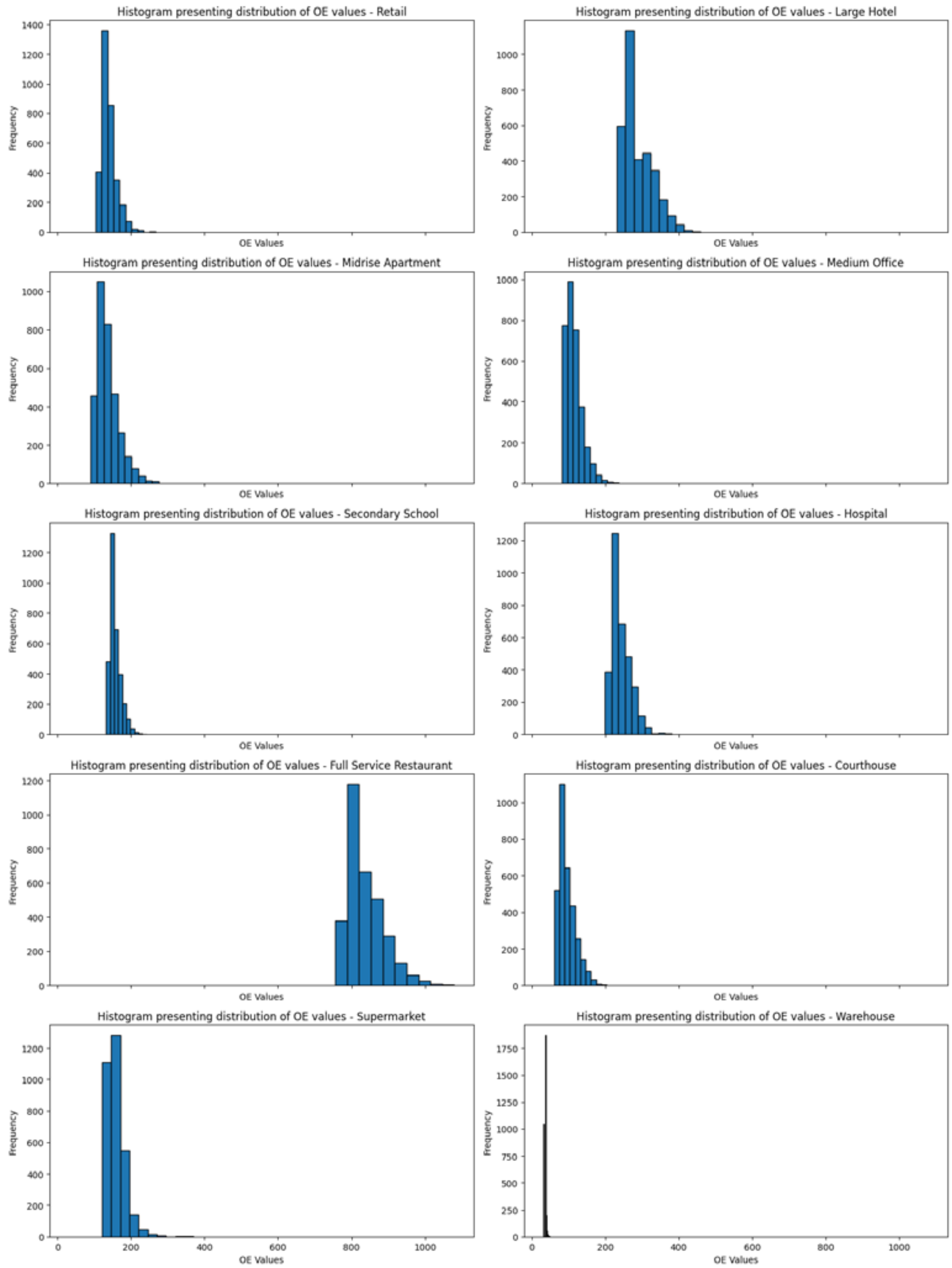



Select few predictors plotted to observe distributions.

The summary statistics for the predictors reveal a variety of scales and distributions. For instance, the orientation seems uniformly distributed between 0 and 360 degrees, non-mass wall R-values range between 0 and 5.5, and roof R-values range between 3.76 and 10.49. The response variable, operational energy (OE), has a mean of approximately 221 with a wide standard deviation of about 217, indicating significant variation in energy consumption across the dataset.

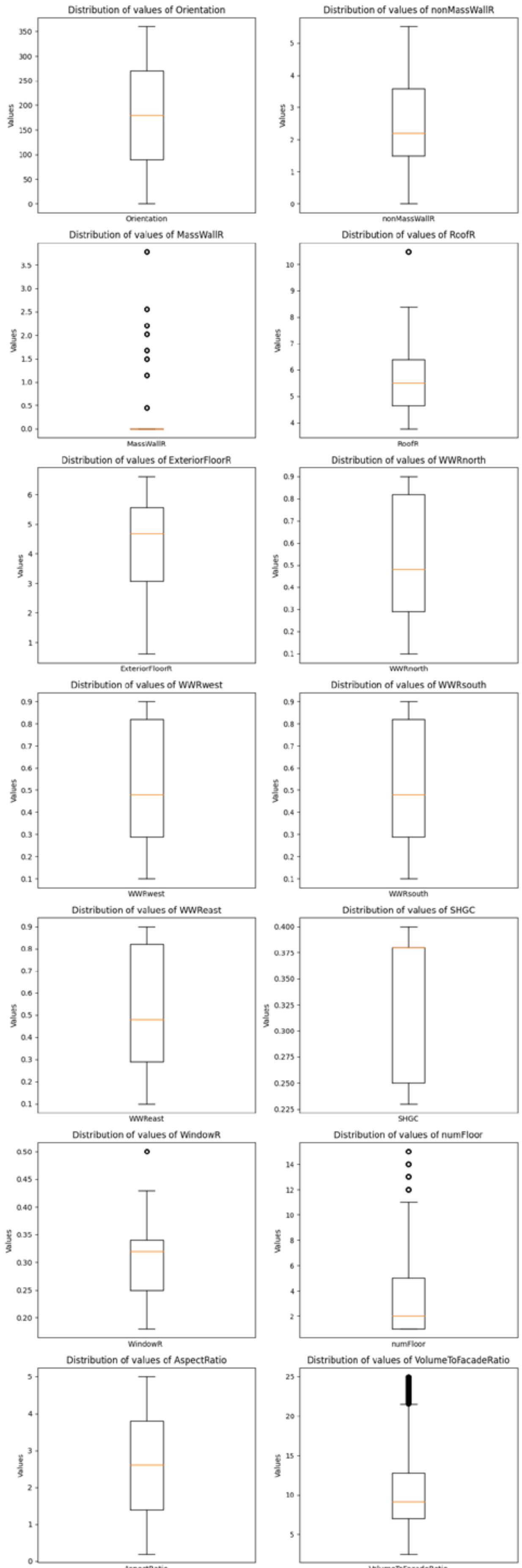
The histograms for selected predictors and the operational energy suggest the following:

The Orientation histogram is approximately uniform, suggesting no particular orientation bias in the dataset.
nonMassWallR is non-uniform, with varying frequency of values
RoofR shows a slight right skew.
The operational energy distribution is right-skewed, with a few values significantly higher than the rest, which could be potential outliers or simply represent high-energy-consuming buildings.

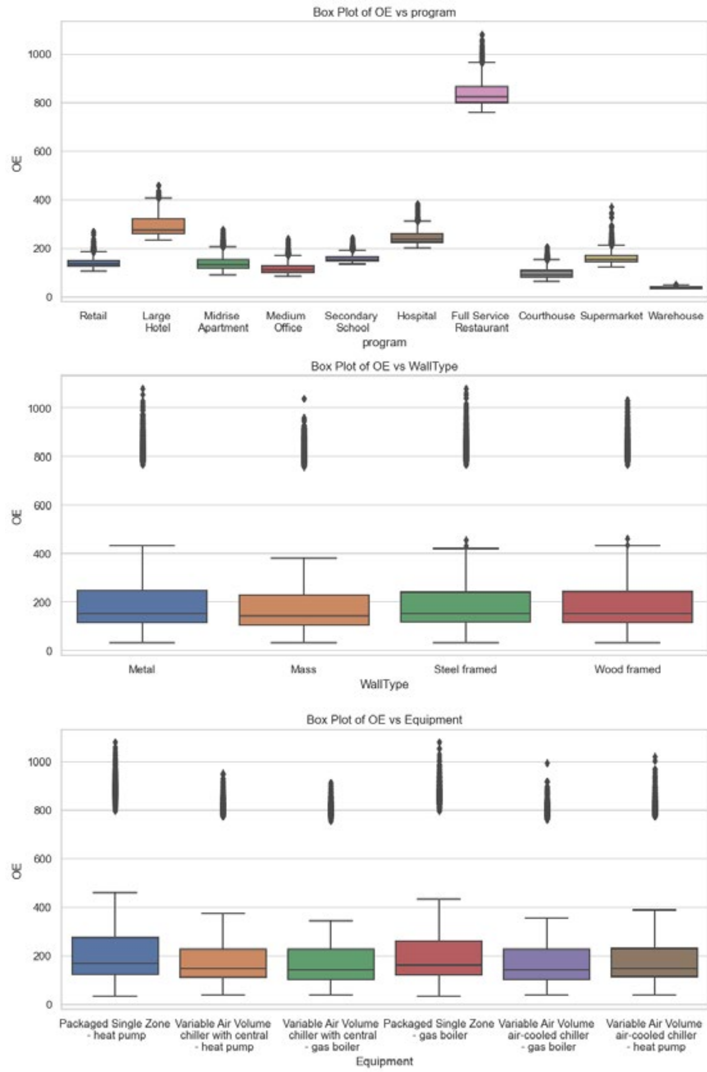



Distribution of OE per programs(labeled)

The distributions of Operational energy consumption per program is very helpful to determine how buildings of different types differ in energy usage and at what frequency. In the plots above, Warehouse types for instance use very low evergy as they often have low numbers of embedded mechanical systems and equipment loads. In contrast, a building like a restaurant tend to operate actively throughout 24 hrs, some operating beyond 10 - 12 hours per day, with a lot of loads from cooking equipment, lighting and ventilation systems. The plot displays high fidelity with common assumptions of building usages.




Distribution of values per Predictor

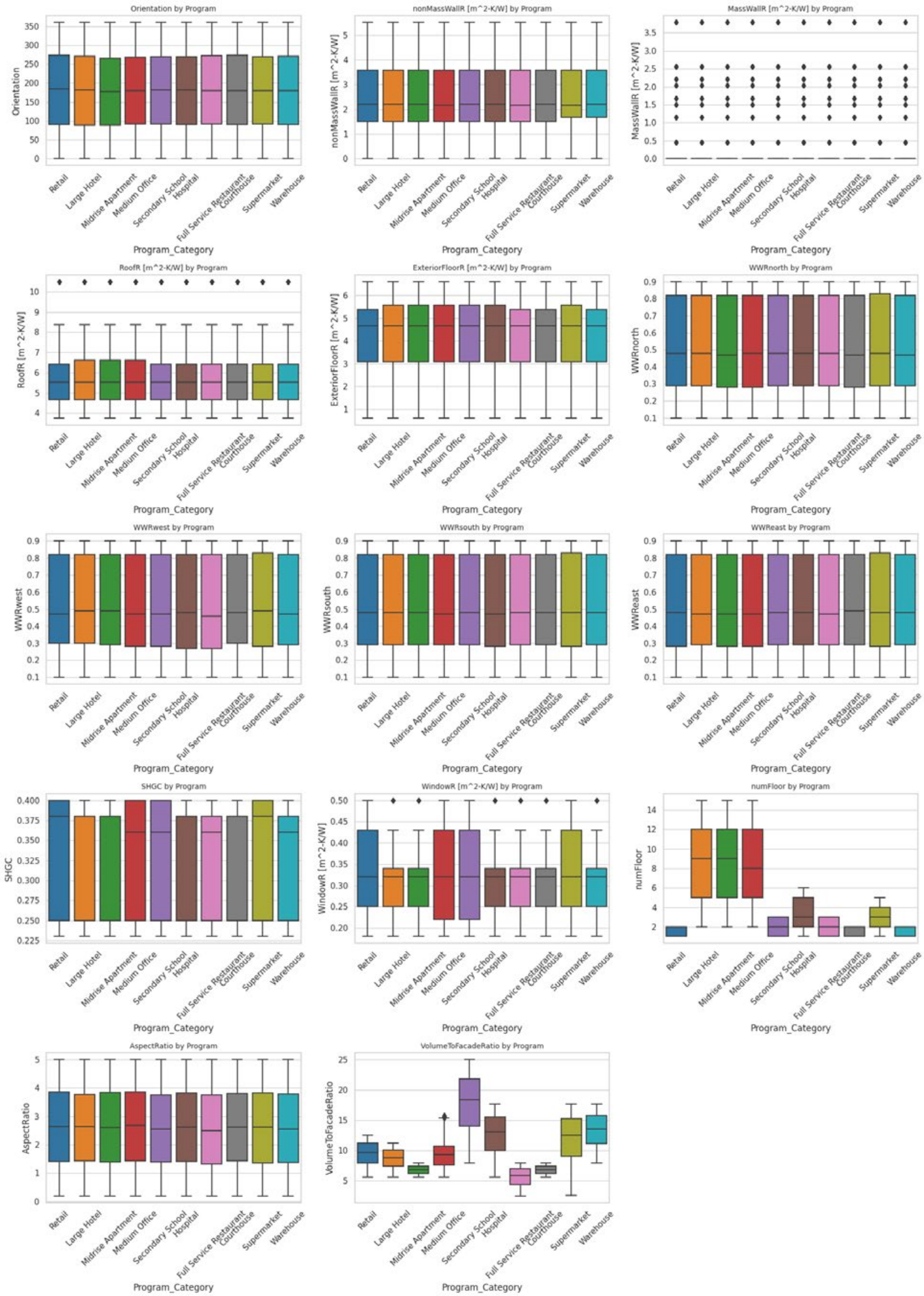



OE vs Categorical Variables

In this section of research, we are seeking for outliers and trying to understand where these may come from. In the left plots for distribution values per predictor we can easily find outliers of values per predictor, particularly Window R values, mass wall R values, number of floors and volume to facade ratio having the most noticeable outliers. To further understand why, we plotted the OE vs the categorical variables. The plot above contradicts our initial belief that perhaps the outliers came from full service restaurants which used the most OE, and instead we find outliers in all building types, which means the outliers are nearly well distributed across all types, being general outliers rather than a particular type's outliers.

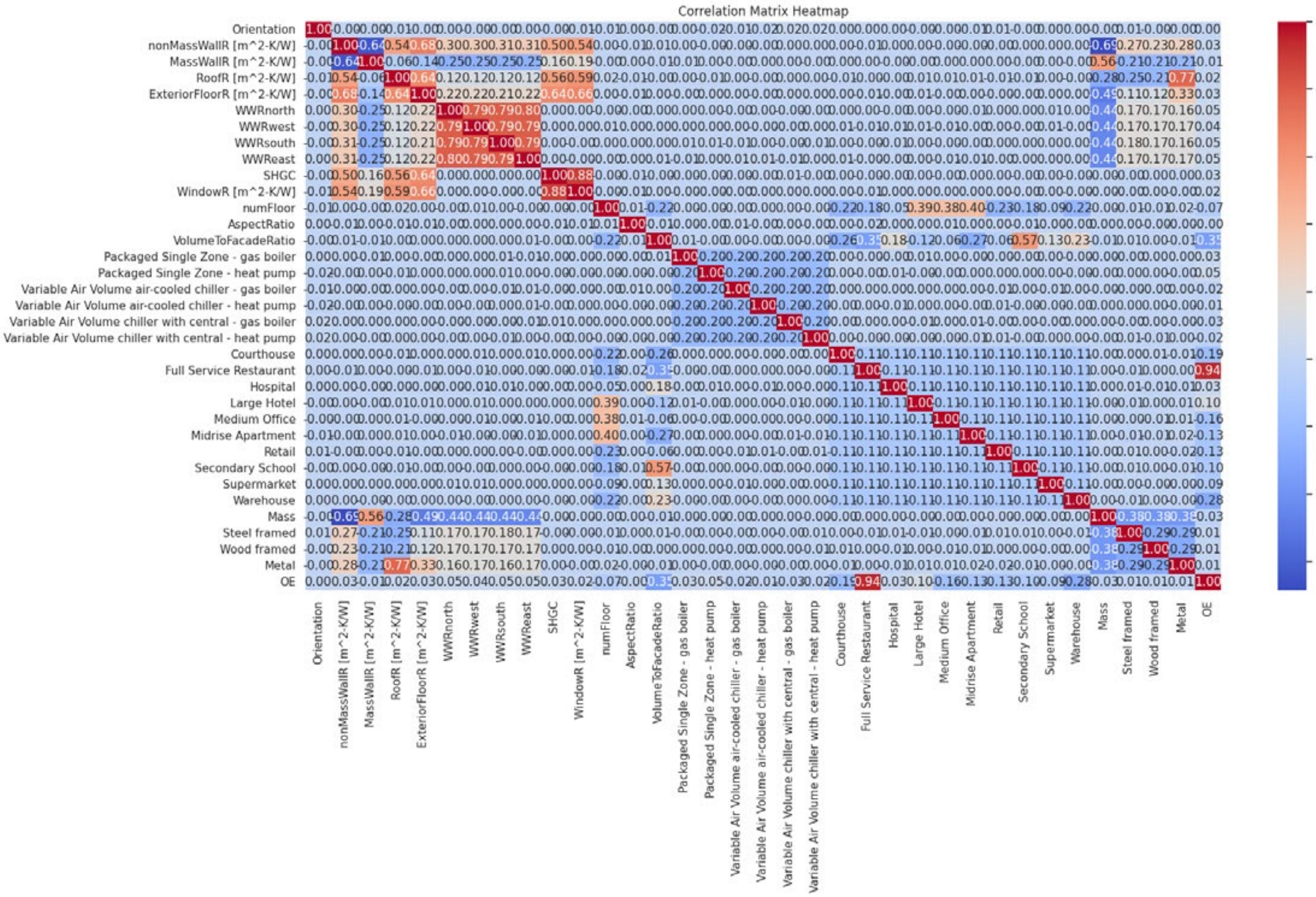
In the next page, we analyze this even further by plotting each individual predictor vs building programs, which further strengthen our findings so far.

Refer to plot in pg5




Predictors vs Programs

## Trends and Relationships



A correlation matrix is made and visualized in a heat map using seaborn to show relationships and correlation coefficients between variables. In this process we are trying to identify which variables have strong relationship with each other.

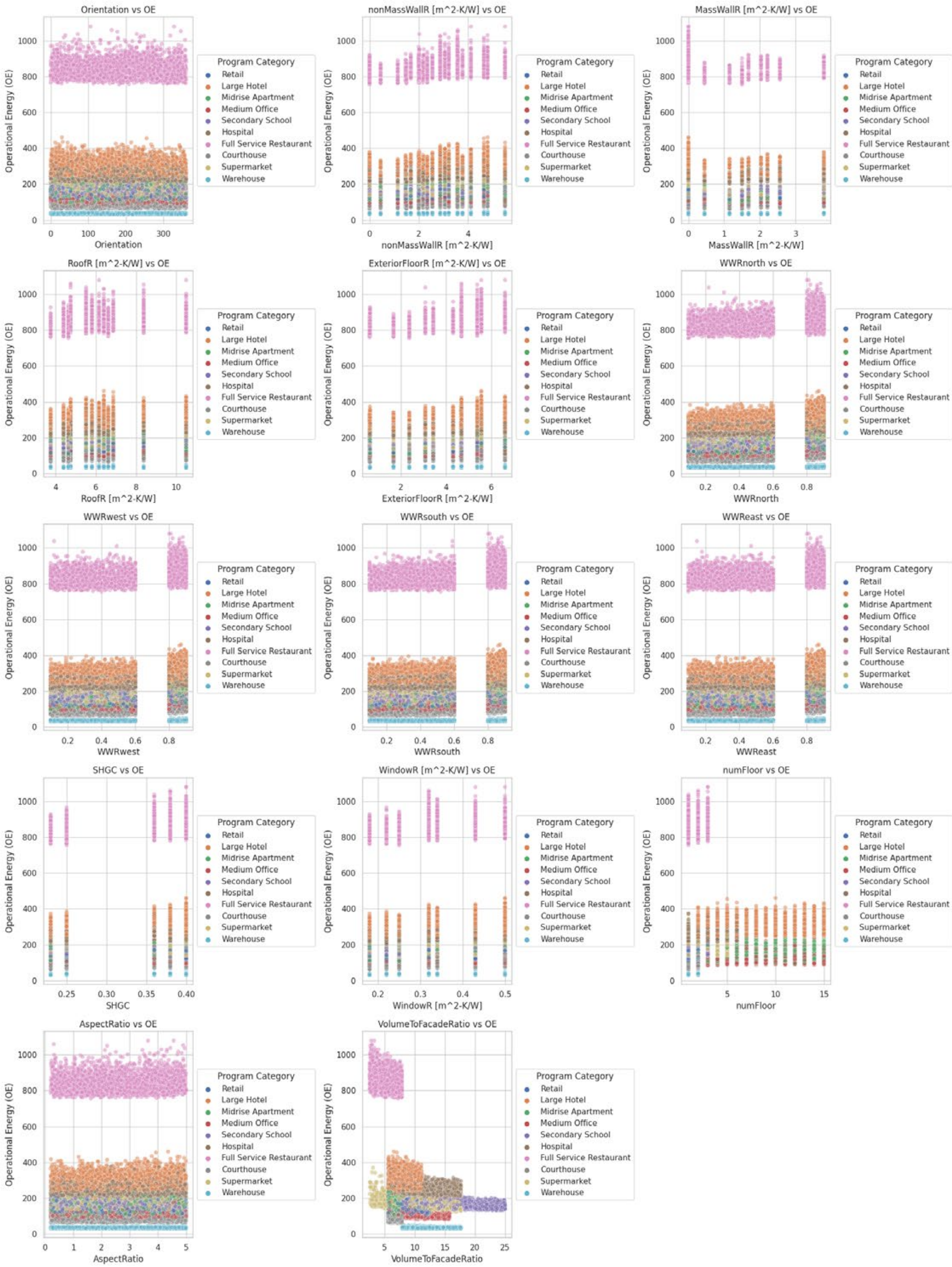
In the visualization, most interesting aspect is the VolumeToFacadeRatio having a negative correlation with Operational Energy (OE) of -0.35, suggesting that buildings with high volume to facade ratio tend to have lower operational energy consumptions, holding all variables constant. This is interesting in the context of building technology, as one may often assume that larger buildings would consume more energy, but as we are holding all other variables constant, we need to look into how the other variables influence the output.

As we move to analyzing relationships further by observing trends when plotting each predictor vs OE, we can see a few important aspects to remember and carefully consider: (plot in page 7)

1. Some predictors such as window to wall ratios per orientation have gaps between 0.6 - 0.8 consistently across all 4, this is because the simulations take the situation concerning the glass storefront as well.

2. All predictors, when observed plotted against the response variable, have non-linear relationships, this is crucial to observe in order to determine what model to choose to implement for predictions. For milestone 2, we roughly started with a Random Forest Regression model, which this proves why the model we chose initially could be potentially effective.

3. Across all plots we have a trend of restaurants having the highest energy consumption when considering a predictor holding all others constant.


Orientation vs OE
Operational Energy (OE)
Orientation
nonMassWallR [m^2-K/W] vs OE
nonMassWallR [m^2-K/W]
MassWallR [m^2-K/W] vs OE
MassWallR [m^2-K/W]
RoofR [m^2-K/W] vs OE
RoofR [m^2-K/W]
ExteriorFloorR [m^2-K/W] vs OE
ExteriorFloorR [m^2-K/W]
WWRnorth vs OE
WWRnorth
WWRwest vs OE
WWRwest
WWRsouth vs OE
WWRsouth
WWReast vs OE
WWReast
SHGC vs OE
SHGC
WindowR [m^2-K/W] vs OE
WindowR [m^2-K/W]
numFloor vs OE
numFloor
AspectRatio vs OE
AspectRatio
VolumeToFacadeRatio vs OE
VolumeToFacadeRatio
Program Category
Retail
Large Hotel
Midrise Apartment
Medium Office
Secondary School
Hospital
Full Service Restaurant
Courthouse
Supermarket
Warehouse

## Project Questions & Plans

How can we make an effective predictive model based on the fact that the data has non-linear relationships, relatively high dimensionality, and correlations between predictors?

How do we address and handle gaps? are they critical in determining the accuracy and effectiveness of our model?

Is there a safe range of resulting Operational Energy consumption that we can use to roughly determine before proceeding to validation strategies that our model is yielding accurate predictions?

Our baseline model has been implemented already in Milestone 2 (Random Forest Regression).
We plan to test and compare other models such as:

- Decision Trees
- Polynomial regression
- Lasso|ridge regression
- Gradient Boosting (pending discussion)

The decision for such models comes from the data analysis. We have carefully observed the non-linear relationships and also a relatively high level of complex interactions between variables. Decision Trees and Random Forest regression are models that handle non-linearity effectively without explicitly defining it, while polynomial regression is effective in handling non-linearity between independent and dependent variables by adding polynomial terms to the model. A lasso and ridge regression will also be a good model to test, again because it is effective in handling non-linearity as an extension to polynomial regression.

***For more information on the work in progress please refer to the github link below, where all our notebooks and data can be found:

https://github.com/snwnkang/CS109A_Final/tree/main