# PROJECT 2
# DATA SCIENCE
# ANTICIPATE THE ELECTRICITY CONSUMPTION NEEDS OF BUILDINGS

Seattle

STUDENT : SYLVAIN CARLEVATO

# INTRODUCTION

- In 2050, the city of Seattle has set itself the goal of achieving a level of carbon neutrality.

- In order not to bear new additional costs, like the readings of the campaigns carried out in 2015 and 2016, the city needs predictions for its buildings not intended for housing.

# PLAN

- Introduction
- Presentation of The Problem
- Data Analysis
- Modeling Track & Results of the Selected Models
- Conclusion

# PROBLEM

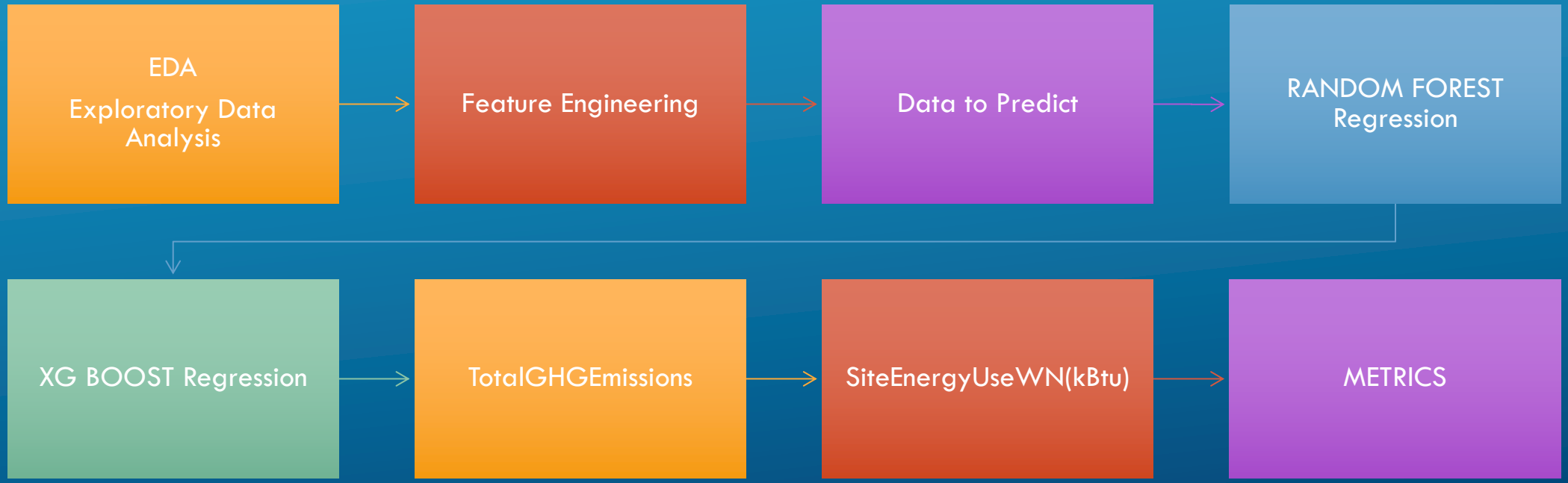
# PRESENTATION OF THE PROBLEM

- Consumption Data available for Buildings in the City of Seattle for the Year 2016.
- Significant Cost of obtaining the Readings to be collected.
- Goals :

  Predictions of Carbon Dioxide Emissions and Total Energy.

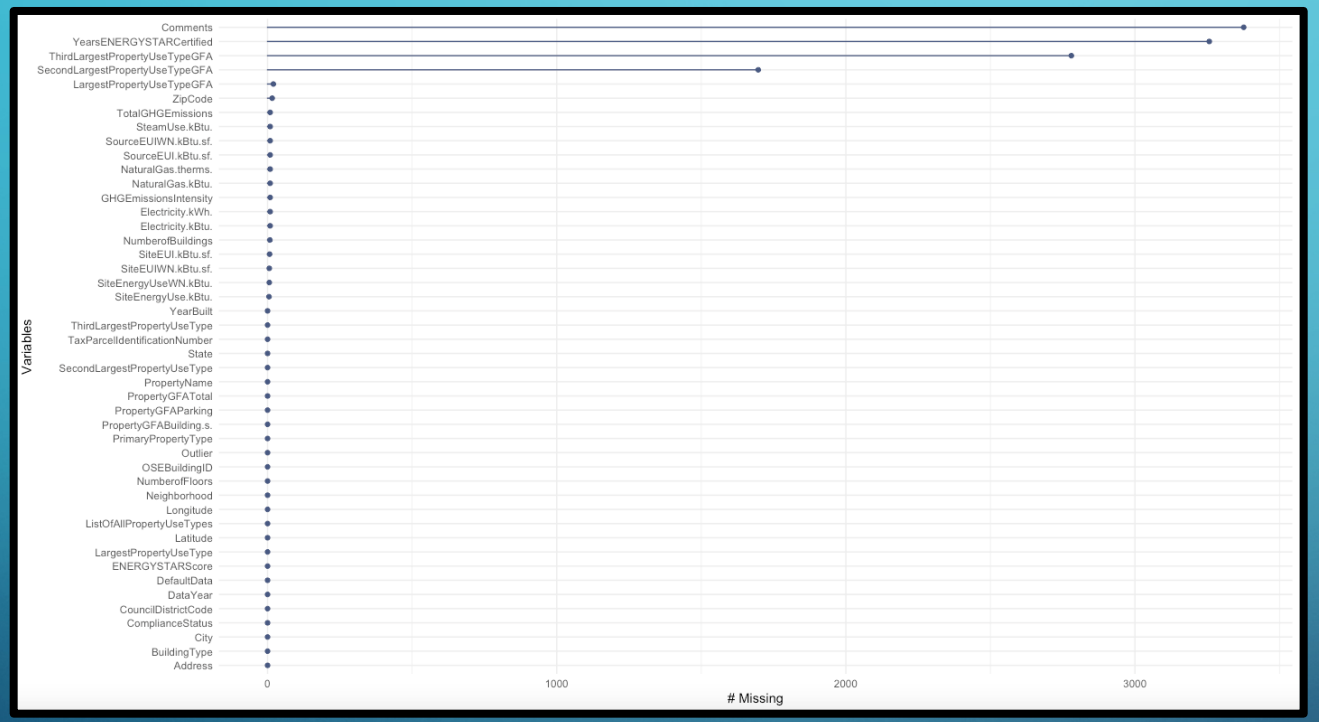
  Consumption in kBtu without Recovery of Annual Readings.
- Implementation of a reusable Prediction Model.

# INTERPRETATION OF THE PROBLEM




DATA ANALYSIS

# NUMBER OF NANS



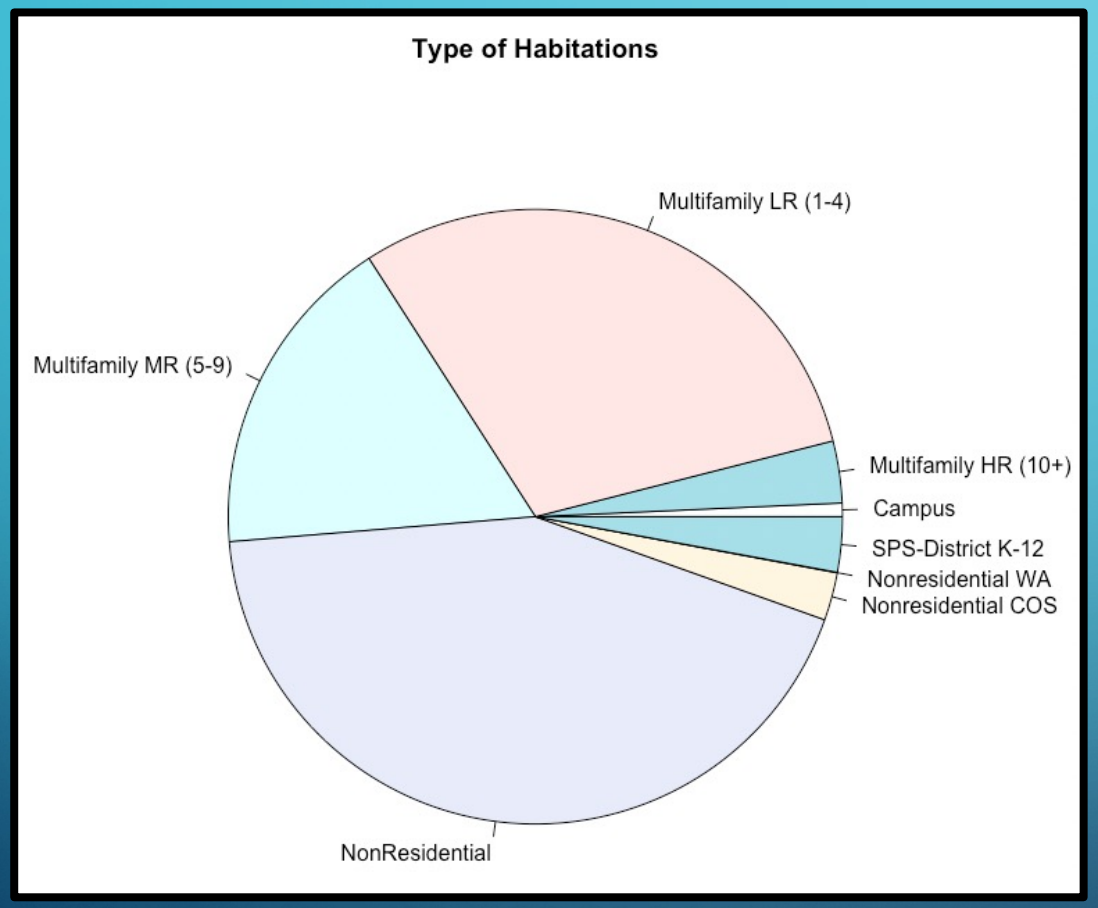

- The dataset contains 11259 NaNs.
- Missing Values are in «Comments», «YearsENERGYSTARCertified», «ThirdLargestPropertyUseTypeGFA» and «SecondLargestPropertyUseTypeGFA» Variables.

# DESCRIPTIVE STATISTICS OF THE TARGETS

| TotalGHGEmissions | |
|---|---|
| Min. | -0.800 |
| 1st Qu. | 9.495 |
| Median | 33.920 |
| Mean | 119.724 |
| 3rd Qu. | 93.940 |
| Max. | 16870.980 |
| NA's | 9 |

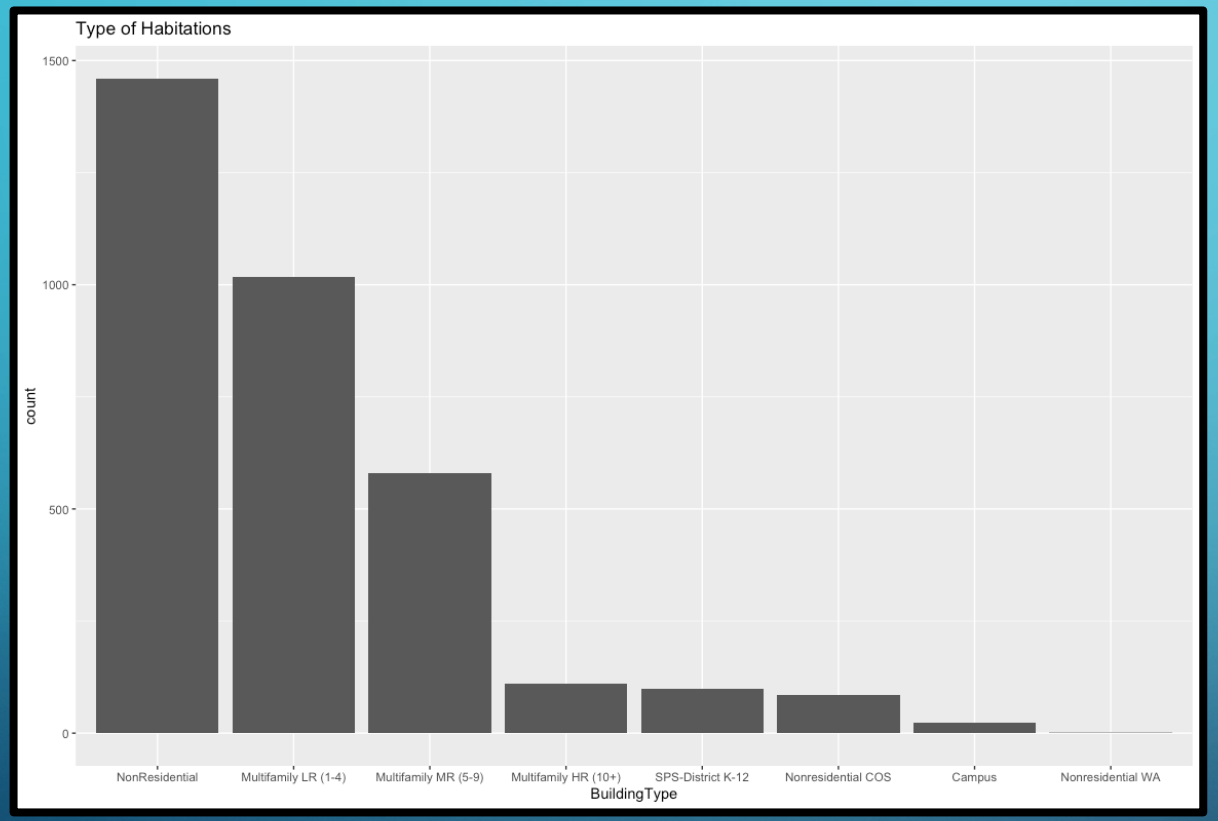
| SiteEnergyUseWN.kBtu. | |
|---|---|
| Min. | 0 |
| 1st Qu. | 970182 |
| Median | 1904452 |
| Mean | 5276726 |
| 3rd Qu. | 4381429 |
| Max. | 471613856 |
| NA's | 6 |

# SELECTION OF THE TYPE OF DWELLINGS



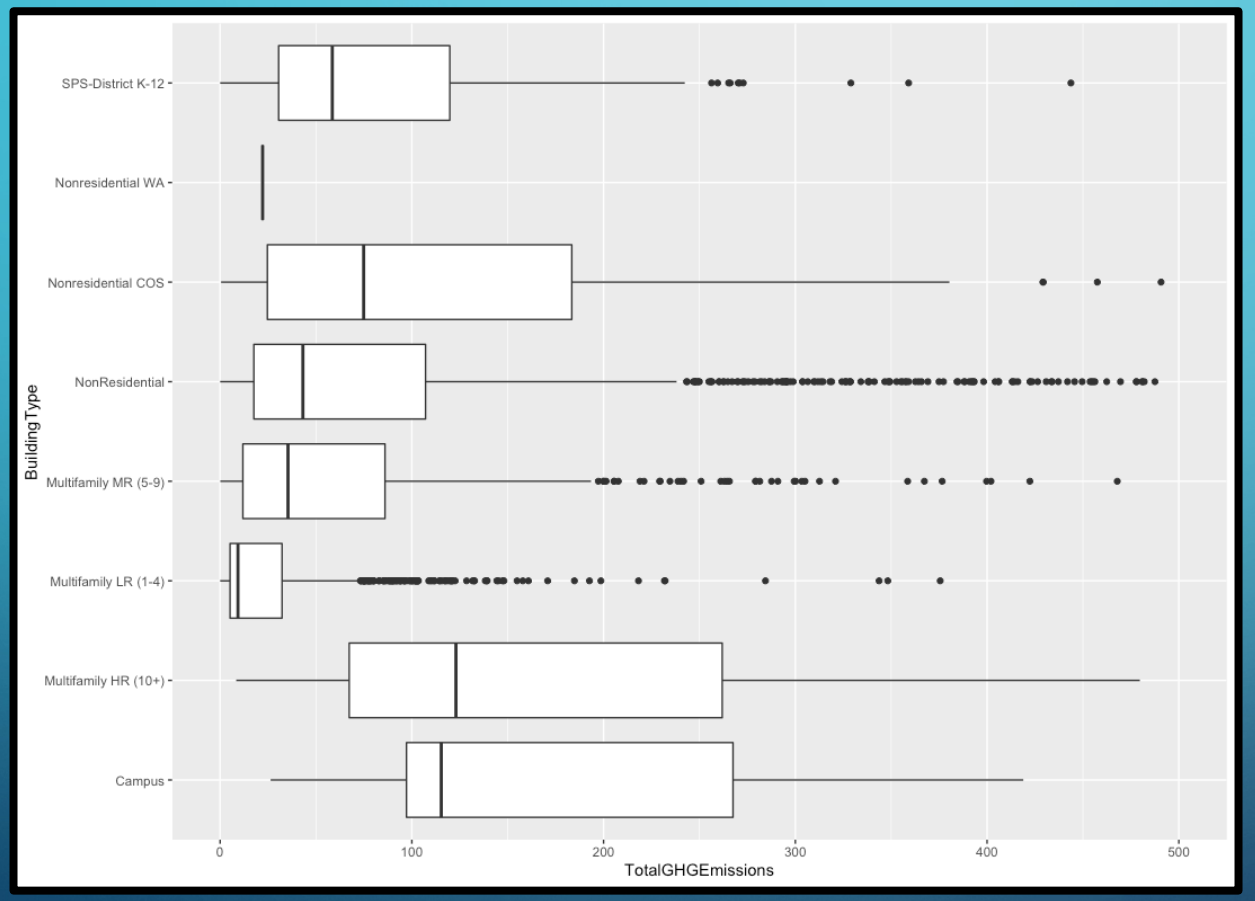

- « Non Residential » and « Multifamily LR (1-4) » are the most important types of Dwellings.

# UNIVARIATE ANALYSIS



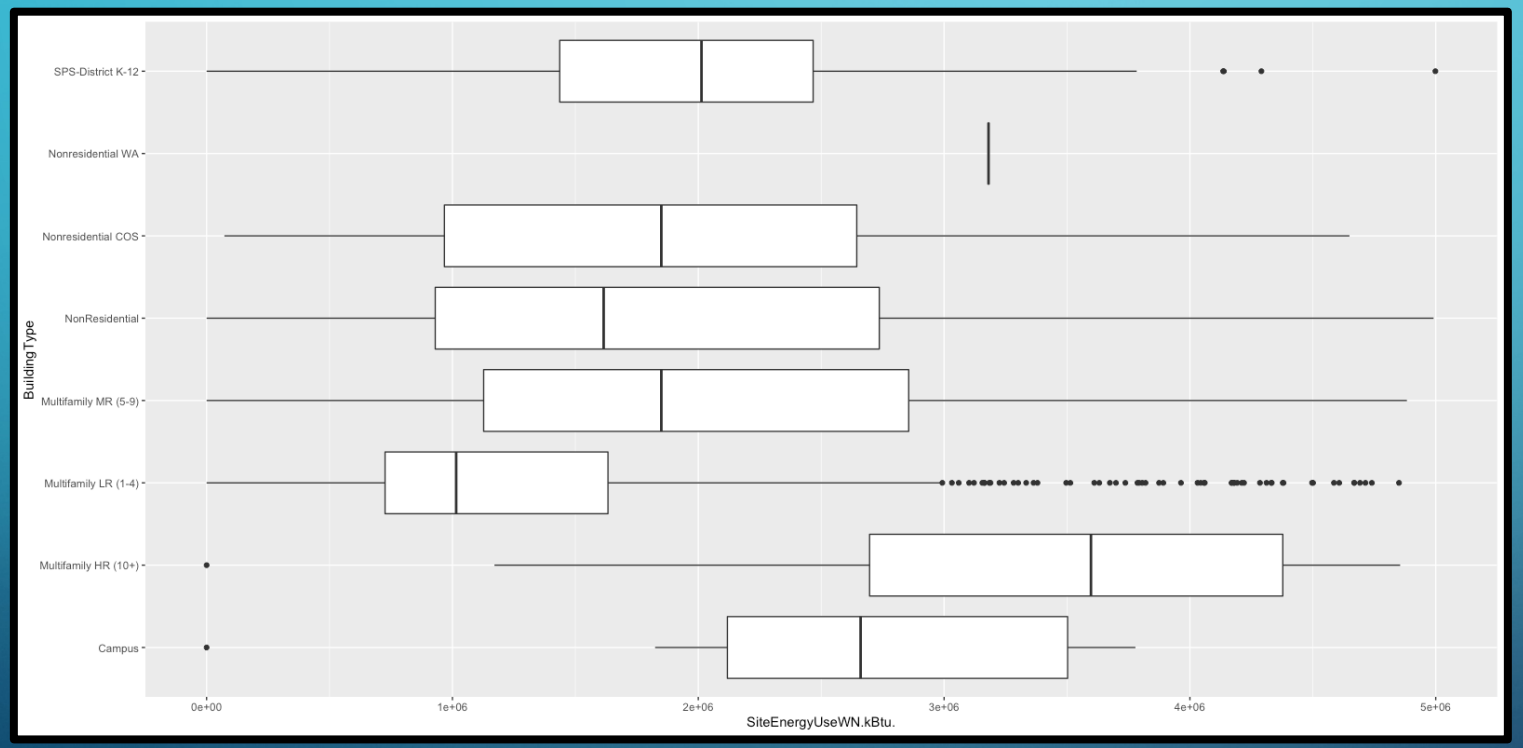

- The Most Important Type Of Habitations is «Non Residential» and «Multifamily LR (1-4)». For my study, I am interested in «Non Residential» Type.

# BIVARIATE ANALYSIS : TOTALGHGEMISSIONS



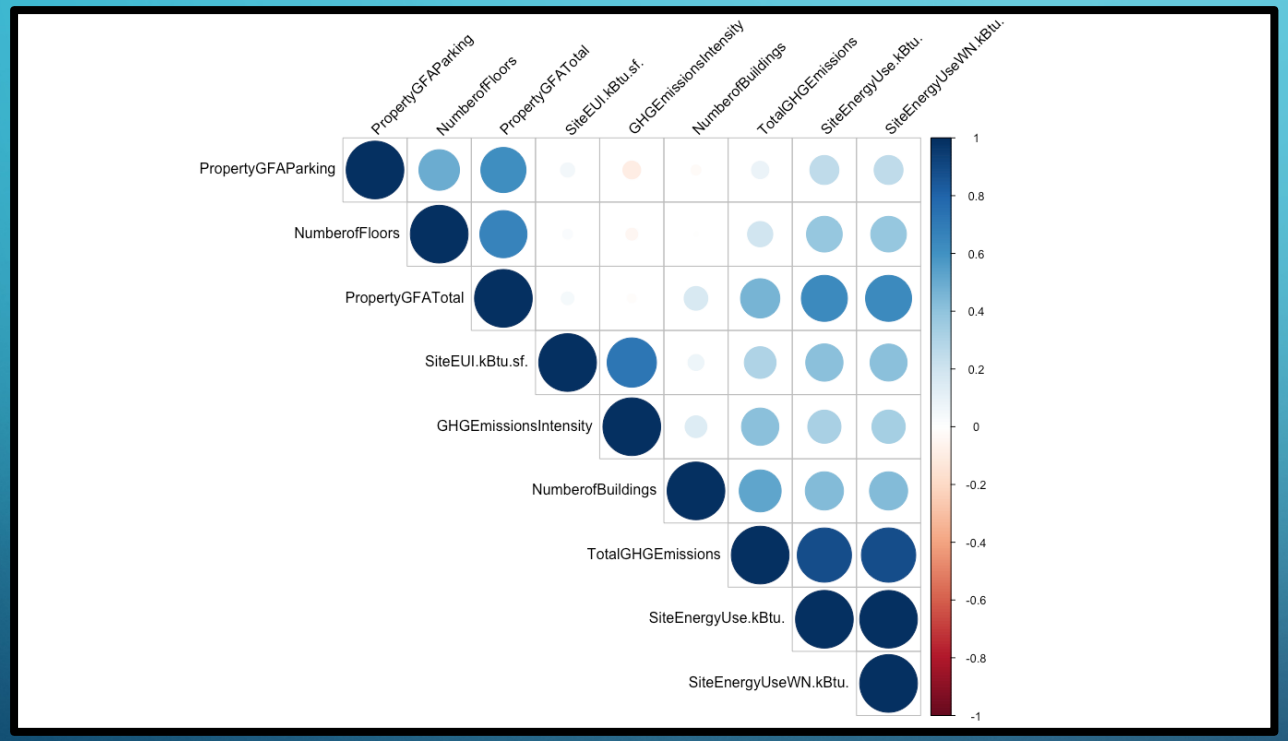

- Variable «Multifamily HR(10+)» has the most important spread of values for the TotalGHGEmissions.

# BIVARIATE ANALYSIS : ENERGY




- Variable «NonResidential» has the most important spread of values for the SiteEnergyUseWN.kBtu.

# MULTIVARIATE ANALYSIS : CORRELATION MATRIX




- Variables like «Number Of Buildings» or «Number Of Floors» have a real impact on Variables like «SiteEnergyUse.kBtu.», «SiteEnergyUseWN.kBtu.», «TotalGHGEmissions».


MODELING TRACK

# XG BOOST REGRESSION

# XGBOOST REGRESSION

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.

Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XGBoost can be used directly for regression predictive modeling.

XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling.

Evaluate an XGBoost regression model using the best practice technique of repeated k-fold cross-validation.

Fit a final model and use it to make a prediction on new data.

# RESULTS : XG BOOST REGRESSION : TARGETS 1 & 2

| Metrics | TARGET 1 : TotalGHGEmissions | TARGET 2 : SiteEnergyUseWN.kBtu. |
|---|---|---|
| MAE | 37 | 596555.5 |
| MSE | 83872 | 5.146327e+12 |
| RMSE | 289 | 2268552 |
| R_squared | 0.898775 | 0.9823269 |

# RANDOM FOREST REGRESSION

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression.

Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

# RESULTS : RANDOM FOREST REGRESSION : TARGETS 1 & 2

| Metrics | TARGET 1 : TotalGHGEmissions | TARGET 2 : SiteEnergyUseWN.kBtu. |
|---|---|---|
| MAE | 66.04764 | 3458829 |
| MSE | 58440.86 | 3.187042e+14 |
| RMSE | 241.7454 | 17852288 |
| R_squared | 0.872856 | 0.6728722 |

# CONCLUSION

- Use of two regression models.
- Efficient results with XGBOOST and with Random Forest for Target SiteEnergyUseWN(kBtu) and for Target TotalGHGEmissions.