ENERGY DATA REPORT

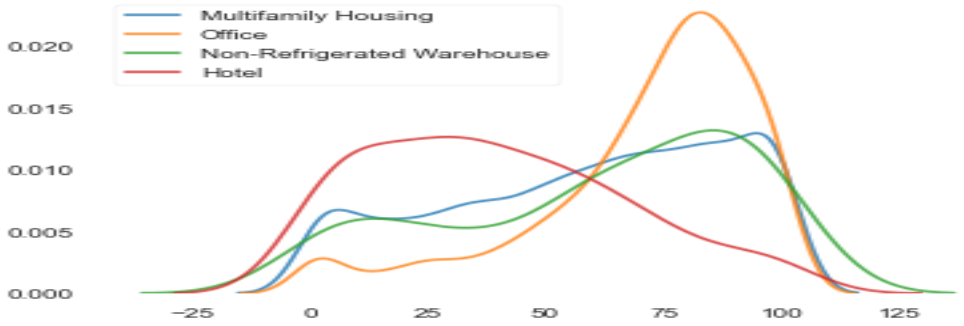
1.Data cleaning

- In the given data, the data dtypes are not correct for the respective data features, so here we assigned type of the data correctly.
- Removing extra spaces in the values of some features
- Observed some missing values and removing them

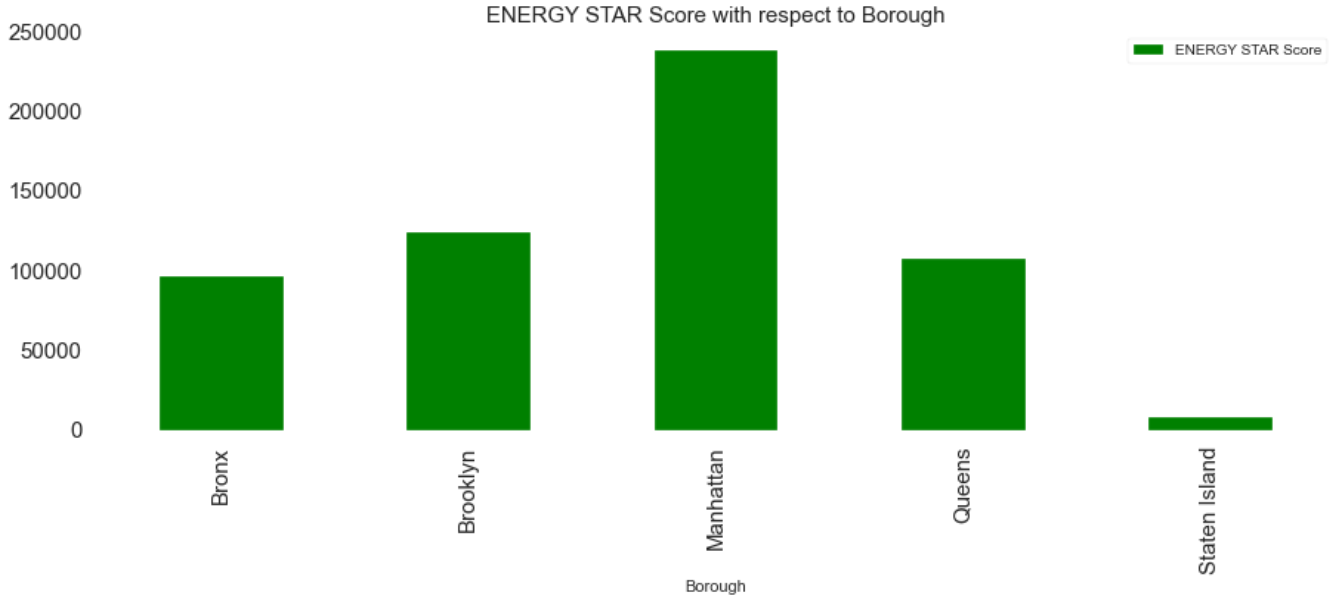
2.EDA

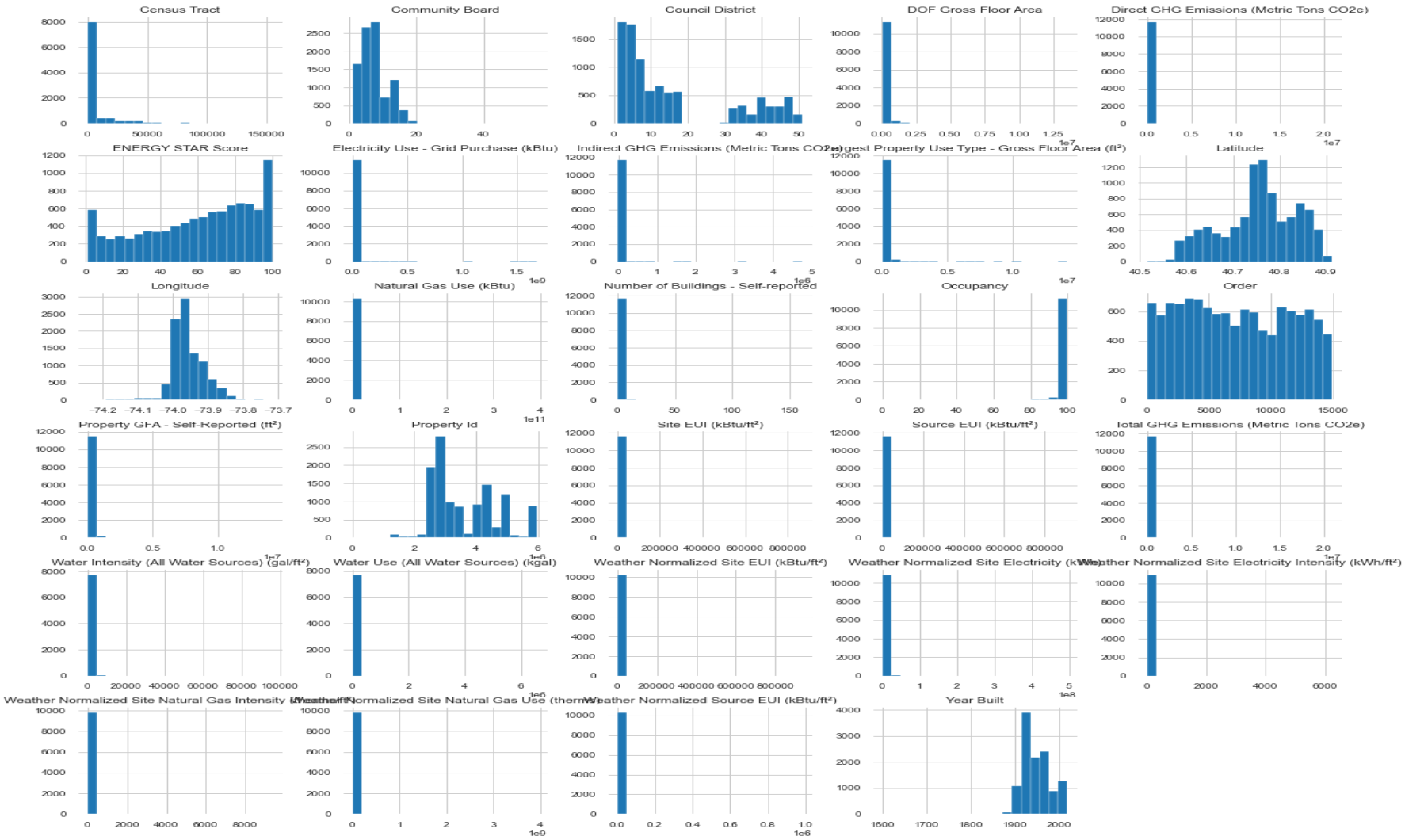
- We found the correlation between all the features with target value. Here our target is 'Energy star score'.
- Performing outlier analysis on our target values and removing them.
- Observed some of the important features with respect to energy score from the data obtained after outlier's removal.
    - Site EUI (kBtu/ft²) 0.723864
    - Weather Normalized Site EUI (kBtu/ft²) 0.713993
    - Weather Normalized Source EUI (kBtu/ft²) 0.645542
    - Source EUI (kBtu/ft²) 0.641037
- Here are the plots for categorical and Numerical data

ENERGY STAR Score with respect to Largest Property Use Type
ENERGY STAR Score
400000
300000
200000
100000
0
Adult Education
Ambulatory Surgical Center
Automobile Dealership
Bank Branch
College/University
Convenience Store without Gas Station
Courthouse
Data Center
Distribution Center
Enclosed Mall
Financial Office
Food Service
Hospital (General Medical & Surgical)
Hotel
K-12 School
Laboratory
Library
Mailing Center/Post Office
Manufacturing/Industrial Plant
Medical Office
Movie Theater
Multifamily Housing
Museum
Non-Refrigerated Warehouse
Office
Other
Other - Education
Other - Entertainment/Public Assembly
Other - Lodging/Residential
Other - Mall
Other - Public Services
Other - Recreation
Other - Services
Other - Specialty Hospital
Other - Technology/Science
Outpatient Rehabilitation/Physical Therapy
Parking
Performing Arts
Pre-school/Daycare
Refrigerated Warehouse
Repair Services (Vehicle, Shoe, Locksmith, etc.)
Residence Hall/Dormitory
Residential Care Facility
Restaurant
Retail Store
Self-Storage Facility
Senior Care Community
Social/Meeting Hall
Strip Mall
Supermarket/Grocery Store
Urgent Care/Clinic/Other Outpatient
Wholesale Club/Supercenter
Worship Facility
Largest Property Use Type



Top 4 building types with high energy score





Numerical data histogram plot

3.Feature Engineering and Selection

- We one hot encoded the categorical columns which are important from the data
- Also we added log transformations and square root values for the numerical data to avoid some non-linear relationship which we have seen in the correlation plot with respect to target values.

4.Model Building

- Here we removed null values from the data where there are no energy score values.
- We replaced the nan values with median of the column in data.
- We performed the data on many machine learning models which are Linear regression, KNN, SVM, random forest, lightgbm, xgboost, gradient boost
- Out of these, lightgbm performance is good when compared with remaining models.

Here is the table of models,

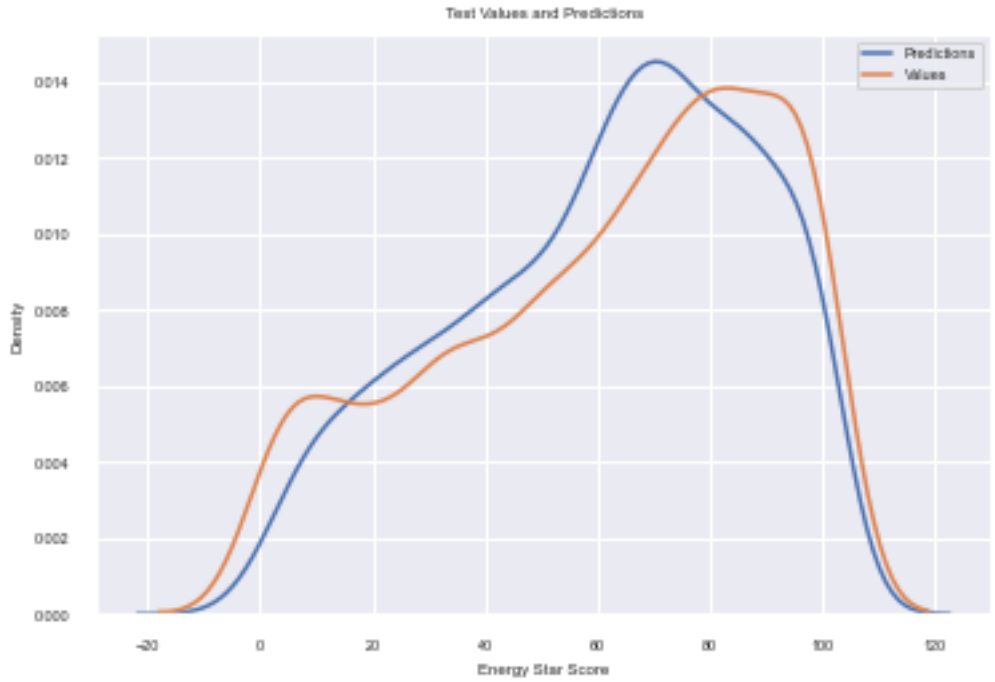
	model_name	MAPE_Result
0	lr	12.325714
1	svm	10.065244
2	randomforest	8.962814
3	knn	13.045509
4	gradboost	9.087056
5	ltb	8.882567
6	xgb	8.975074

5.Hyper parameter tuning

- From the models results, Lightgbmregressor model performance is good. So we optimize the model with best hyperparameters using Randomsearchcv
- After performing hyperparameter tuning on the model, we got lightgbm with these parameters improved the error.
- The best model score is,

model_name	MAPE_Result
lr	12.325714
svm	10.065244
randomforest	8.962814
knn	13.045509
gradboost	9.087056
ltb	8.882567
xgb	8.975074
best_lgb	8.819985

```
LGBMRegressor(bagging_fraction=0.9, bagging_freq=100, boosting_type='gbdt',
              class_weight=None, colsample_bytree=1.0, feature_fraction=0.5,
              importance_type='split', learning_rate=0.02, max_depth=-1,
              min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
              n_estimators=800, n_jobs=-1, num_leaves=10, objective=None,
              random_state=None, reg_alpha=0.0, reg_lambda=0.0, silent=True,
              subsample=1.0, subsample_for_bin=200000, subsample_freq=0)
```



Density plot with lightgbm best model

6.Model Interpretation

- Here we extracted top 80 as important features out of the 108 features from the training data.
- The model result obtained is little improved with 0.02 difference with the best model mape result that is 8.79.

Conclusion:

- From the data cleaning and EDA, we observed there is a missing data.
- Also there is no energy score for around 1500 observations in the data
- Only few building types that is nearly 10 are having high energy score
- Most of the numeric data values are skewed right, and this is due to presence of outliers in the data and we removed outliers which are in only top feature relation with our target.
- We applied log and sqrt transformations to the data to get linear relationship as they are negatively correlated with the target.
- Some more the hyper parameter tuning of model and that model optimization is required and should ex tract and understand the importance of the features in the data.