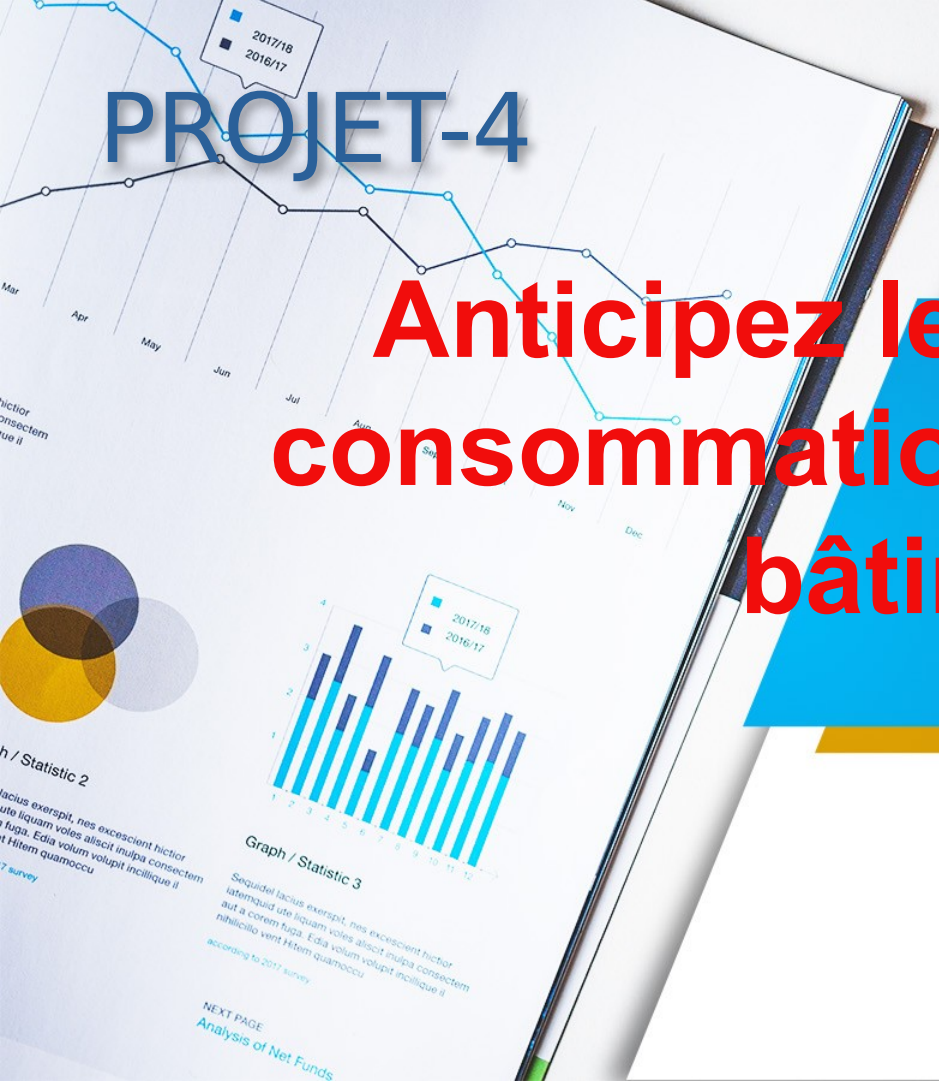


# PROJET-4

## Anticipez les besoins en consommation électrique de bâtiments





# Plan de Présentation

- 1- Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.(5 mn)
- 2- Présentation du cleaning effectué, du feature engineering et de l'exploration.(5mn)
- 3- Présentation des différentes pistes de modélisation effectuées.(10 mn)
- 4- Présentation du modèle final sélectionné ainsi que des améliorations effectuées.(5mn)
- 5- 5 à 10 minutes de questions-réponses

(5mn=750 mots    10mn=1500 mots)

# 1-Problématique

La ville de Seattle veut atteindre son objectif de ville neutre en émissions de carbone en 2050.

Une étude est initiée et porte sur les émissions des bâtiments non destinés à l'habitation.

Des relevés minutieux ont été effectués en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, on va tenter de prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments pour lesquels elles n'ont pas encore été mesurées.

\* Il va falloir aussi évaluer l'intérêt de l'ENERGY STAR Score en modélisant Avec et sans.

\* Nous allons donc entraîner plusieurs algorithmes de régression linéaires. A l'issue de l'apprentissage, nous allons retenir le ou les modèles de régression à performance élevée. Ensuite nous allons procéder à l'optimisation des paramètres en utilisant la méthode du Gridsearch.

The background is a grayscale image of a document. It features a line chart at the top left with two data series: a black line with circular markers and a blue line with circular markers. The x-axis is labeled with months: Jun, Jul, Aug, Sep, Oct. A legend in the top left corner identifies the series as '2017/18' (black square) and '2016/17' (blue square). Below the line chart is a bar chart with five bars, each composed of two segments in light blue and dark blue. A legend to the right of the bar chart shows a light blue square and a dark blue square. At the bottom left, there is a block of Latin text: 'Sequid lacus exera', 'latemquid ute liquan', 'aut a corem fuga. Ege', 'nitalicillo vent fitem', 'according to 2017 survey'. At the bottom right, it says 'NEXT PAGE' and 'Analysis of h'.

# 1-Problématique(suite)

\* Puisque nous devons prédire

-la consommation en énergie 'SiteEnergyUse(kBtu)'

-l'émission en CO2 'TotalGHGEmissions'.

Il va falloir construire deux modèles prédictifs différents.



## 2-Présentation du cleaning effectué, du feature engineering et

de l'exploration

### A-Cleaning

Le Dataset se compose de 2 fichiers : 2015 et 2016.

Cleaning effectué:

1-Renommage des colonnes définissant les mêmes concepts:

Exemple:

'Comment' et 'Comments'

'GHGEmissionsIntensity(kgCO2e/ft2)' et 'GHGEmissionsIntensity'

2-Fusion des datasets 2015 et 2016 après alignement des variables.

3-Suppression de variables non pertinentes.

## B-Feature Engineering

1-Decomposition de la variable 'location'

2-Calcul Pourcentage pour chaque type d' energie

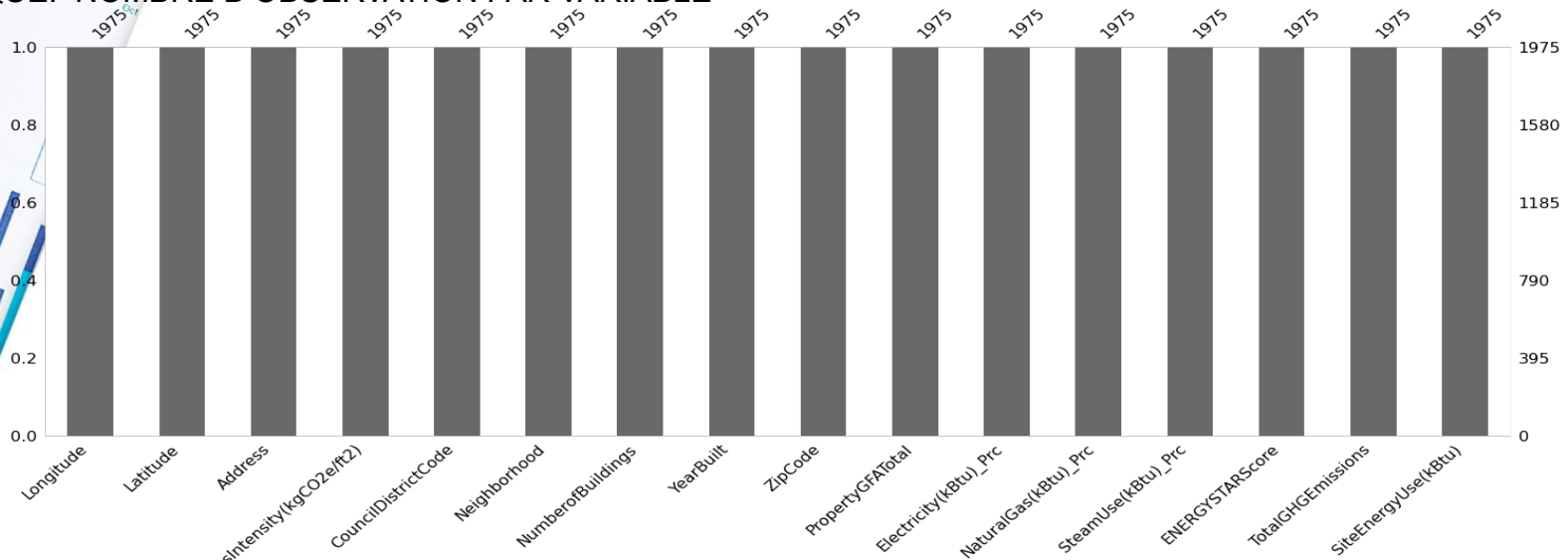
```
df15_16['Electricity(kBtu)_Prc']=df15_16['Electricity(kBtu)'].div(df15_16['AllTypeEnergy'])  
df15_16['NaturalGas(kBtu)_Prc']=df15_16['NaturalGas(kBtu)']/df15_16['AllTypeEnergy']  
df15_16['SteamUse(kBtu)_Prc']=df15_16['SteamUse(kBtu)']/df15_16['AllTypeEnergy'].
```

# C-Exploration

Dimension DataSet final

Nombre de lignes : 1975.  
Nombre de Colonnes: 17.

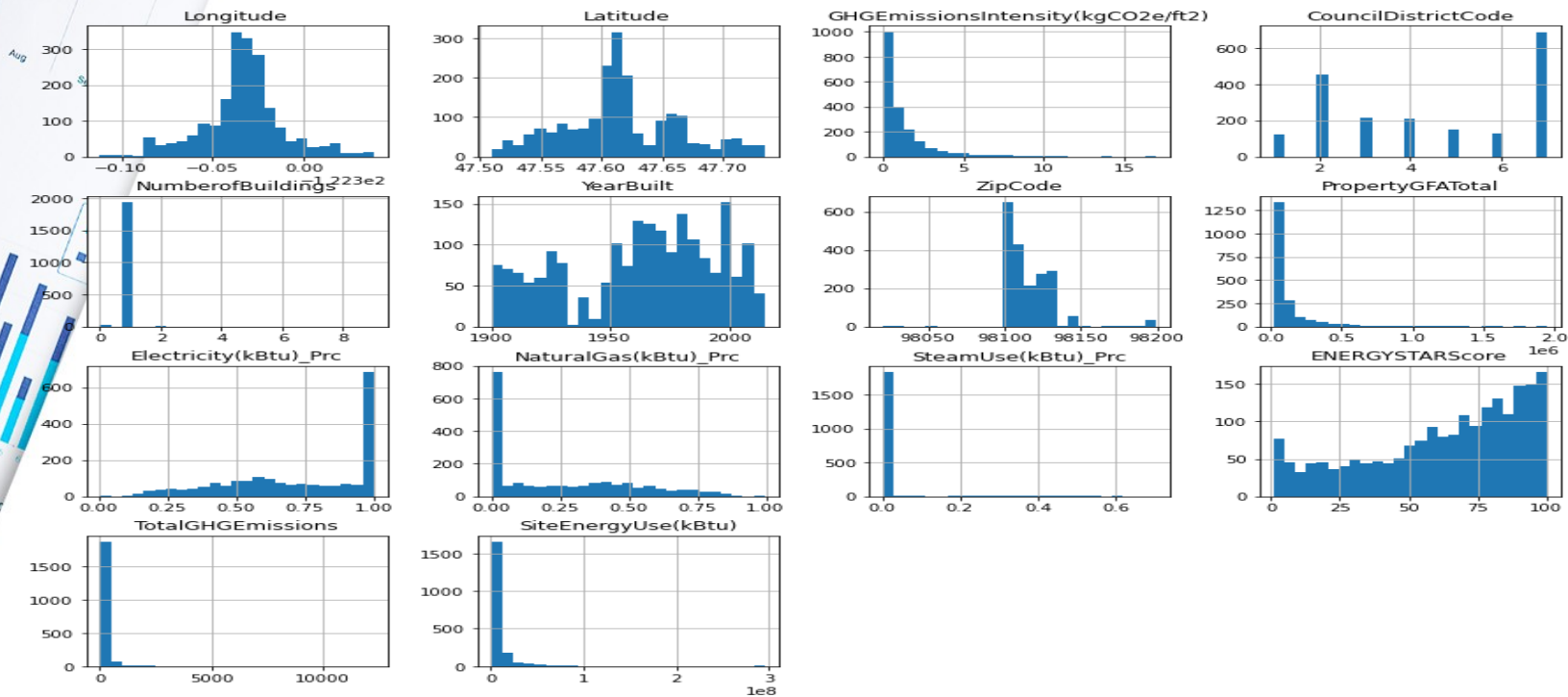
GRAPHIQUE: NOMBRE D'OBSERVATION PAR VARIABLE



# C-Exploration(suite1)

## Analyses Univariées

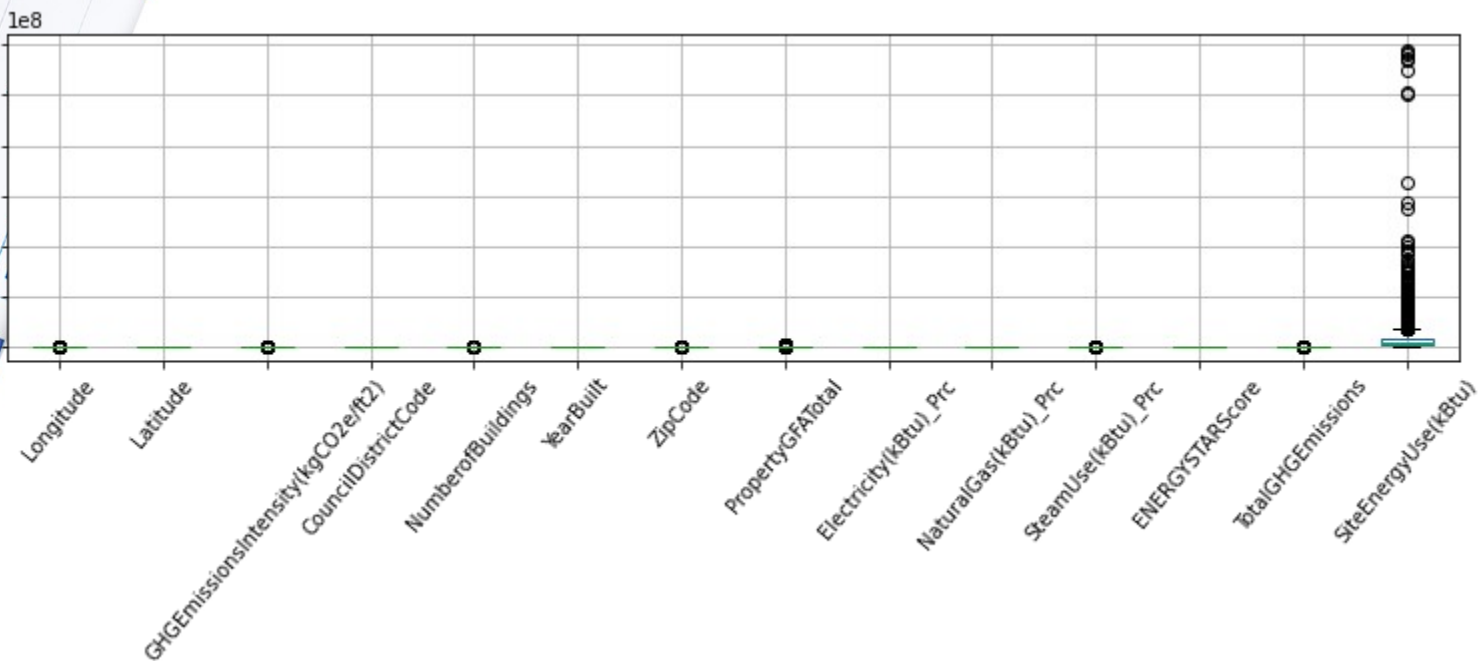
### 1-Variables numeriques : histogramme





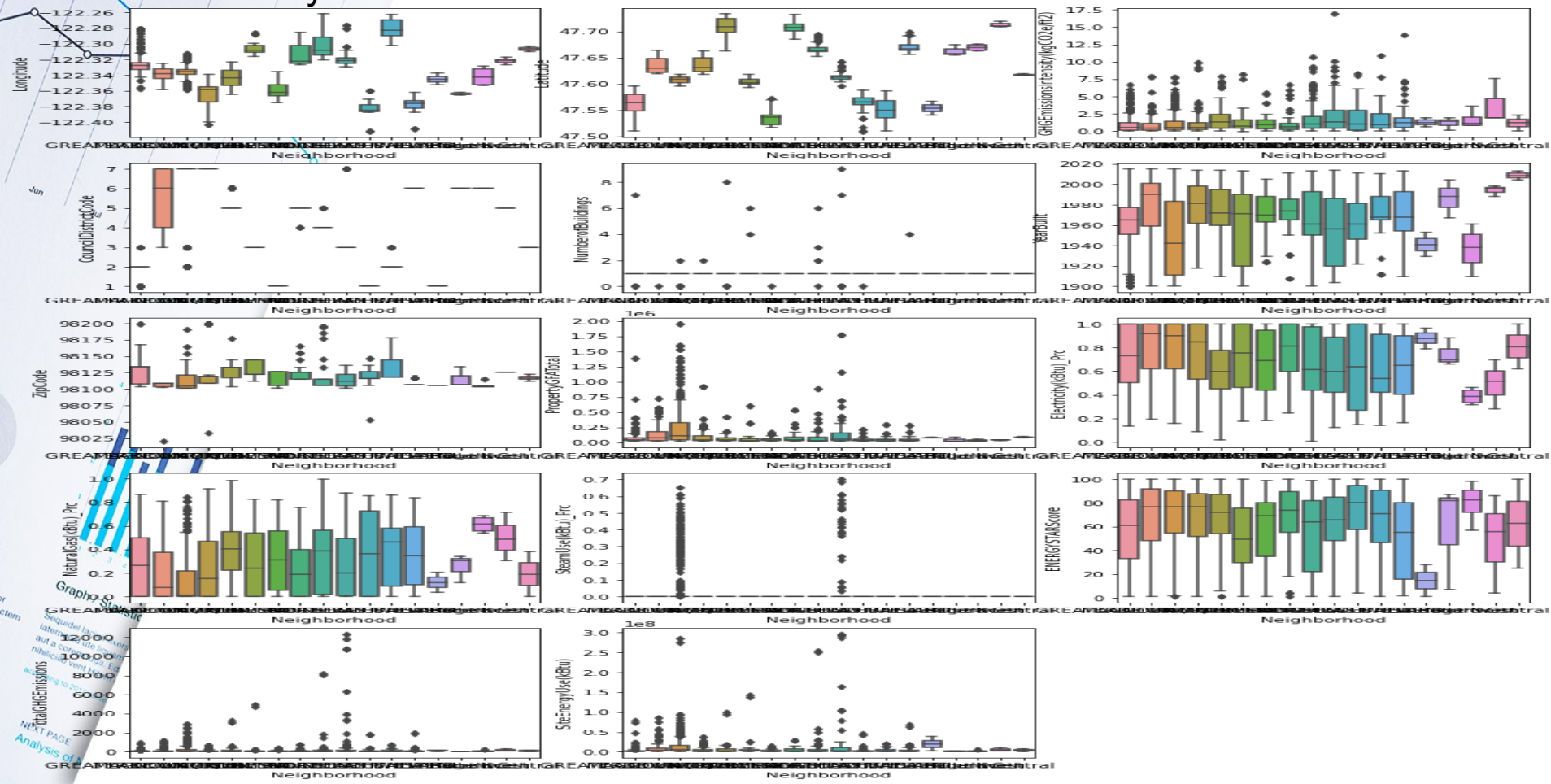
## C-Exploration(suite2)

2-Valeurs numeriques :boxplot



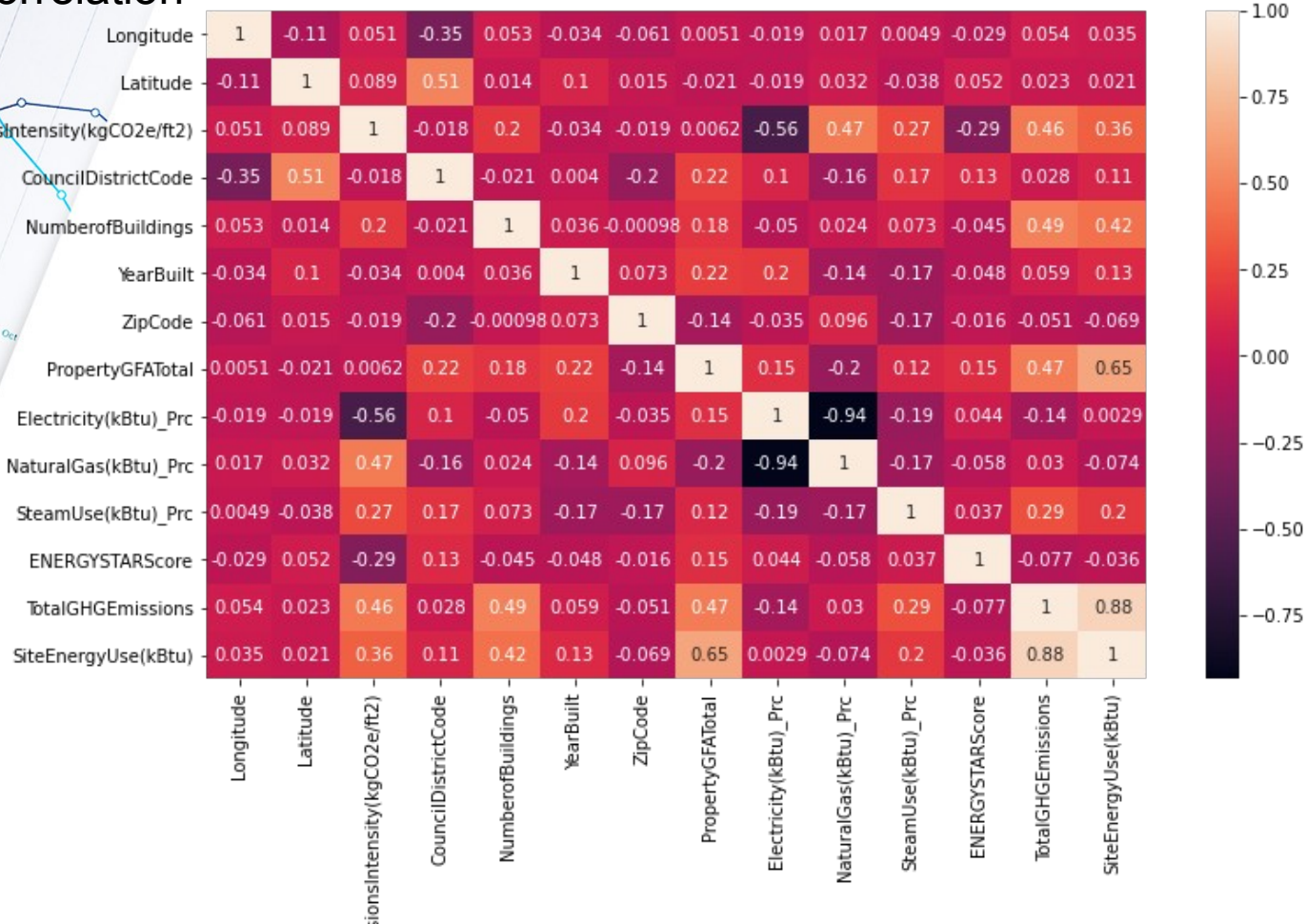
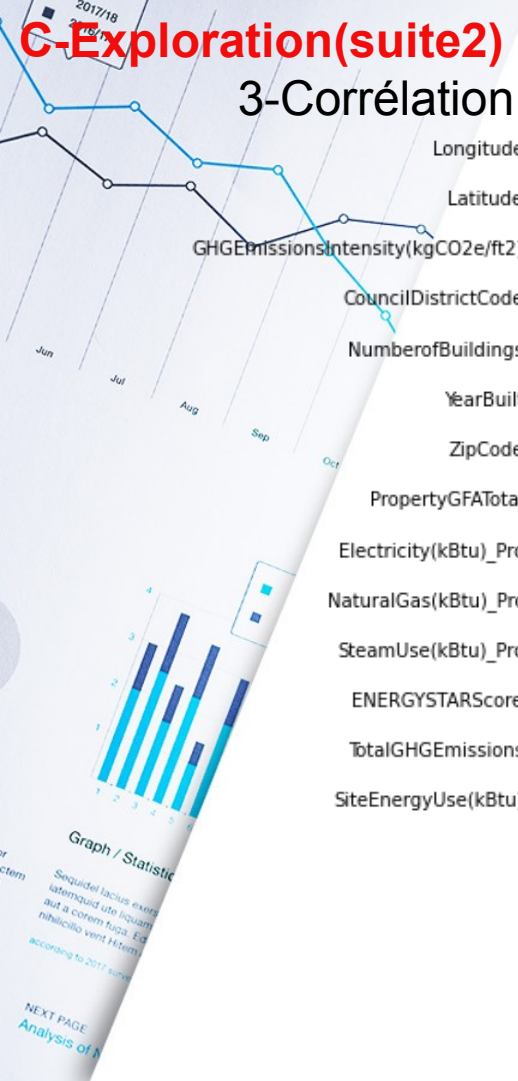
# C-Exploration(suite2)

## 3-Analyse bivariée



# C-Exploration(suite2)

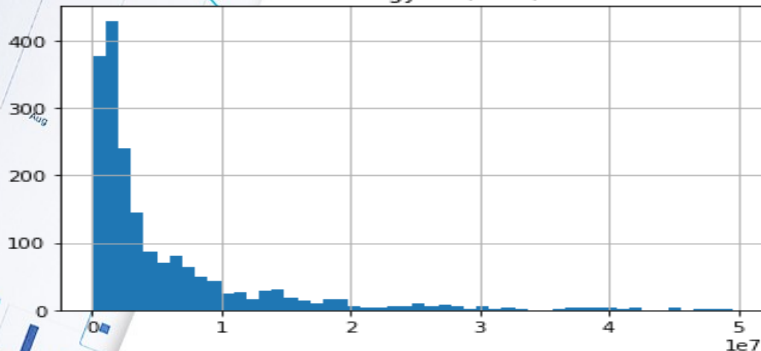
## 3-Corrélation



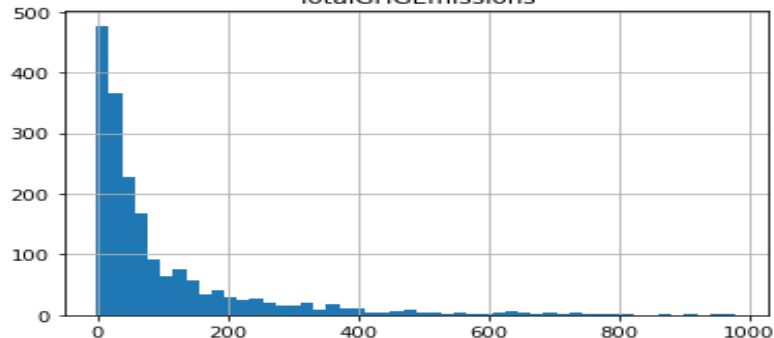
### 3-Modélisation et Amélioration

**A-Variables à prédire : SiteEnergyUse(kBtu) et TotalGHGEmissions**  
Sans transformation et avec transformation logarithmique.

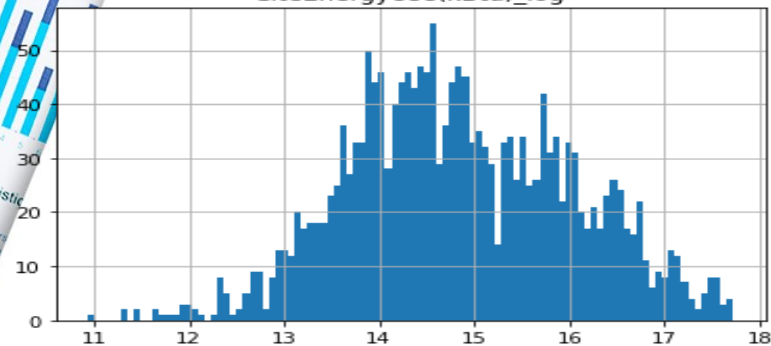
SiteEnergyUse(kBtu)



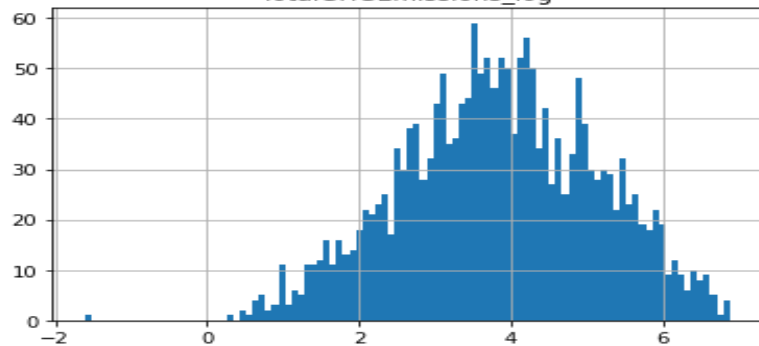
TotalGHGEmissions



SiteEnergyUse(kBtu)\_log



TotalGHGEmissions\_log





### 3-Modélisation et Amélioration(suite1)

#### A-Modèle consommation energie(sans Energy Star Score) :

##### 1-Modelisation par regression lineaire

-Calcul valeurs des métriques

##### 2-Modelisation par plusieurs algorithmes de régression

-Algorithmes

LinearRegression,  
RandomForestRegressor,  
KNeighborsRegressor,  
Ridge,  
Lasso,  
ElasticNet,  
SVR.

-Calcul valeurs des métriques.

##### 3-Optimisation des paramètres du meilleur modèle avec GridsearchCV.

The background features a collage of data-related images. At the top left, a line chart shows two data series for the years 2017/18 and 2016/17, with data points plotted for June, July, and August. Below this, a bar chart displays five bars with varying heights, each composed of two stacked segments in different colors. At the bottom left, there is a snippet of a document titled 'Graph / Statistic' containing some placeholder text and a 'NEXT PAGE' label.

### 3-Modélisation et Amélioration(suite2)

**B-Modèle d'émission CO2(avec et sans Energy Star Score) :**  
**(Même étape que précédemment).**

1-Modélisation par regression linéaire  
-Calcul valeurs des métriques

2-Modélisation par plusieurs algorithmes de regression  
-Algorithmes

LinearRegression,  
RandomForestRegressor,  
KNeighborsRegressor,  
Ridge,  
Lasso,  
ElasticNet,  
SVR.

-Calcul valeurs des métriques.

3-Optimisation des paramètres du meilleur modèle GridSearchCV.

# COMPARAISON METRIQUES

## A-Consommation Energie

### ALGORITHMES

### RMSE

### RMSE(avec log)

- Linear regression 13832782.600
- Random Forest 10798669840950.254

0.8850754448

0.3140493867

-



## COMPARAISON METRIQUES(2)

### B-Emission CO2 sans EnergyStar Score

#### ALGORITHMES

- Linear regression
- Random Forest

#### RMSE

8413779.67037

0.320411

#### RMSE(avec log)

0.8850754448

0.312661947453

-



# COMPARAISON METRIQUES(3)

## C-Emission CO2 avec EnergyStar Score

### ALGORITHMES

### RMSE

### RMSE(avec log)

- Linear regression 8413779.67037
- Random Forest 0.320411
- 

0.8850754448  
0.31557151230

-

- 
- 

### CONCLUSION:

Peu d'influence de EnergyStar Score

# COMPARAISON METRIQUES(4)

## D-Tableau Récapitulatif

Consommation Energie	Algorithmes	<i>RMSE avec log</i>	<i>RMSE sans log</i>
	regression	0.8850754448439829	13832782.600850401
	randomforest	0.318629	0.321654
Emission avec EnergyStarScore	regression	0.8850754448439829	8413779.670372294
	randomforest	0.3207	0.322755
Emission sans EnergyStarScore	regression	0.8850754448439829	8413779.670372294
	randomforest	0.319950	0.320411

# 4-Modèle Final

1-Consommation d' énergie :

**RandomForestRegressor(max\_features=8, n\_estimators=300)**

Score avec log :

**0.3131082842177906**

Score sans log :

**10798669840950.254**

2-Emission CO2 sans EnergyStar Score

**RandomForestRegressor(max\_features=8, n\_estimators=500)**

Score avec log:

**0.31210224291048105**

Score sans log :

**10859908779161.967**