AIRBNB PRICE PREDICTION IN NEW YORK

OUR TEAM

FRASNI OUBEY

BOUSSAA RANIA

MARAH ANASS

2007
NAISSANCE D'AIRBNB

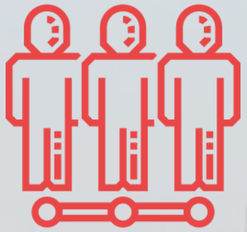
60 MILLIONS
DE RÉSERVATIONS PAR AN

airbnb

81K
VILLES PROPOSÉES

1 MILLIARD
CHIFFRE D'AFFAIRES EN 2018

8,7 MILLIONS



70 MILLIONS

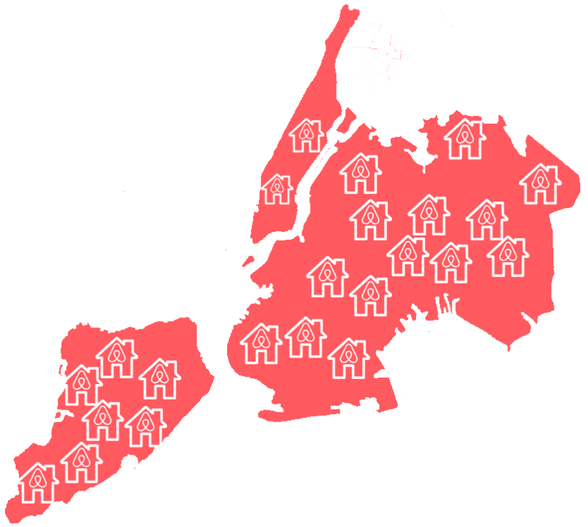
NEW YORK CITY

4%

70 MILLIARDS $

PROBLÉMATIQUE

SATISFACTION DES VOYAGEURS ANNUEL



DISPARITÉ DES LOGEMENTS À NEW YORK

OPTIMISATION DU CHIFFRE D'AFFAIRES

MORE PROFIT

TOOLS

jupyter

tableau SOFTWARE

Visual Studio Code

python™

matplotlib

pandas

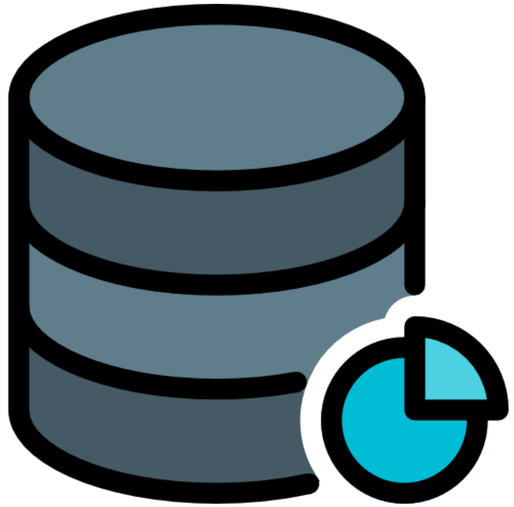
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

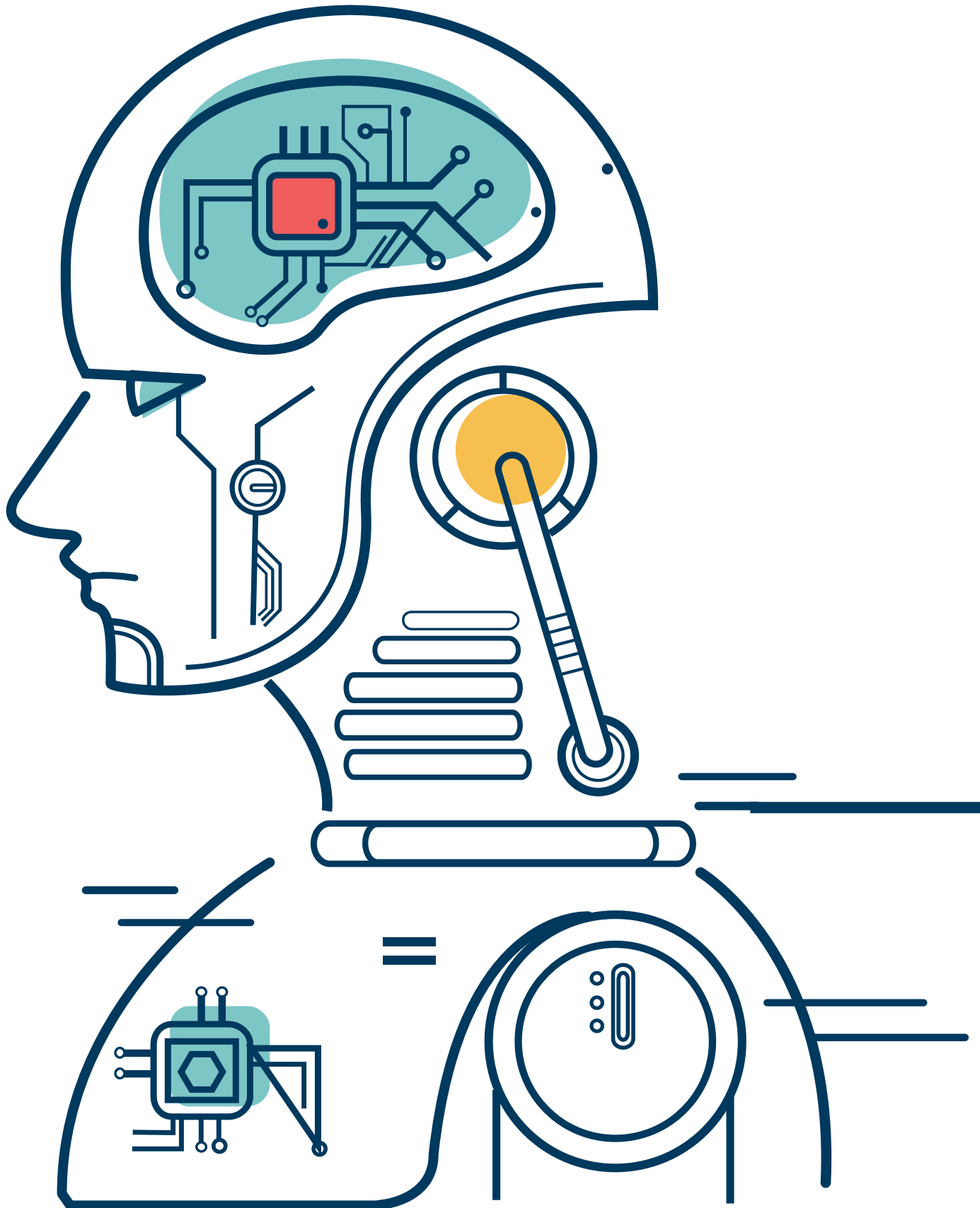
dmlc XGBoost

scikit learn

+

NEW YORK CITY AIRBNB OPEN DATA 2019



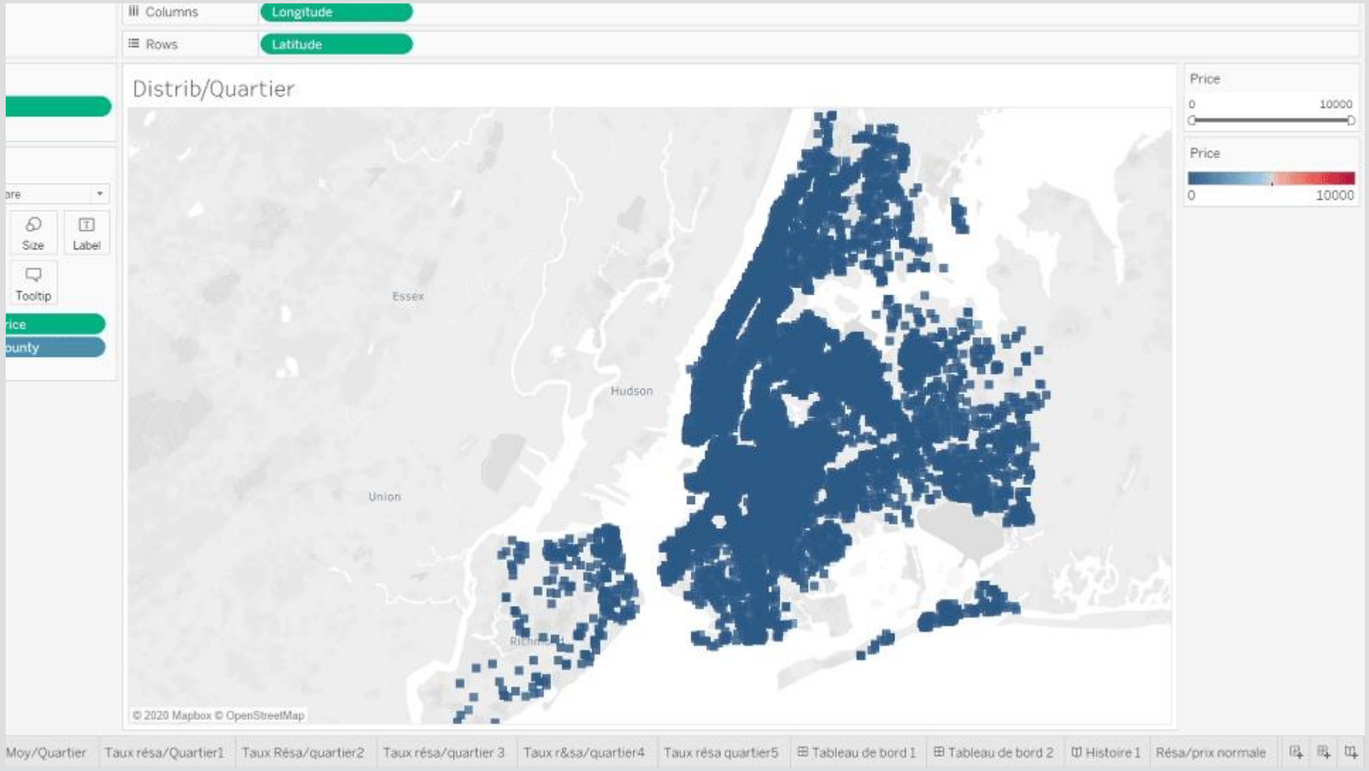


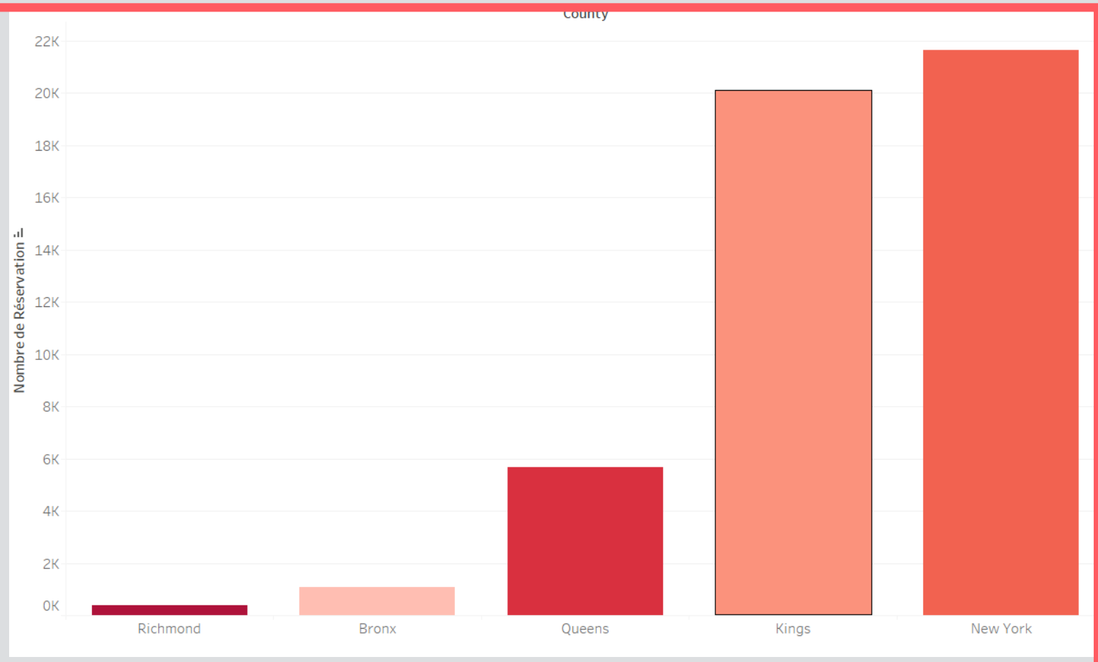
DATA ANALYSIS

EN 2019

MANHATTAN MONOPOLISE LA GAMME LUXE

PLUS QUE 5000$/RÉSERVATION



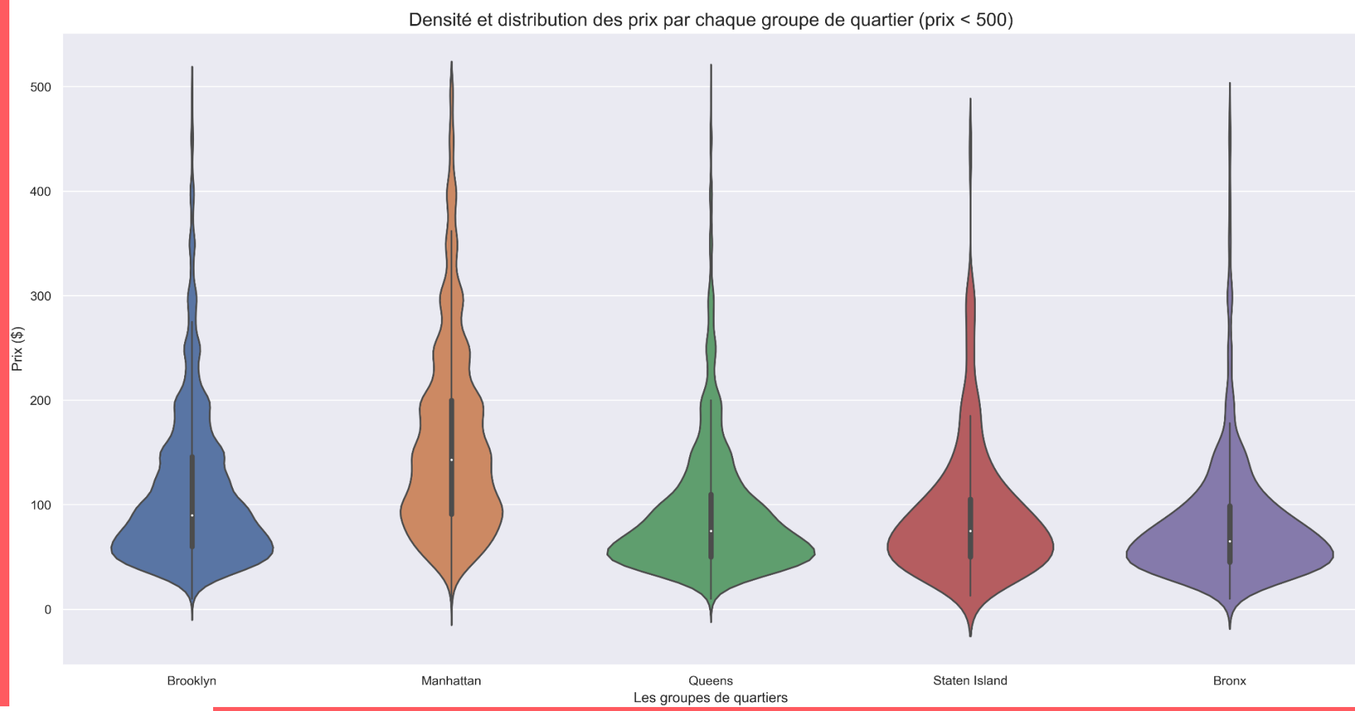


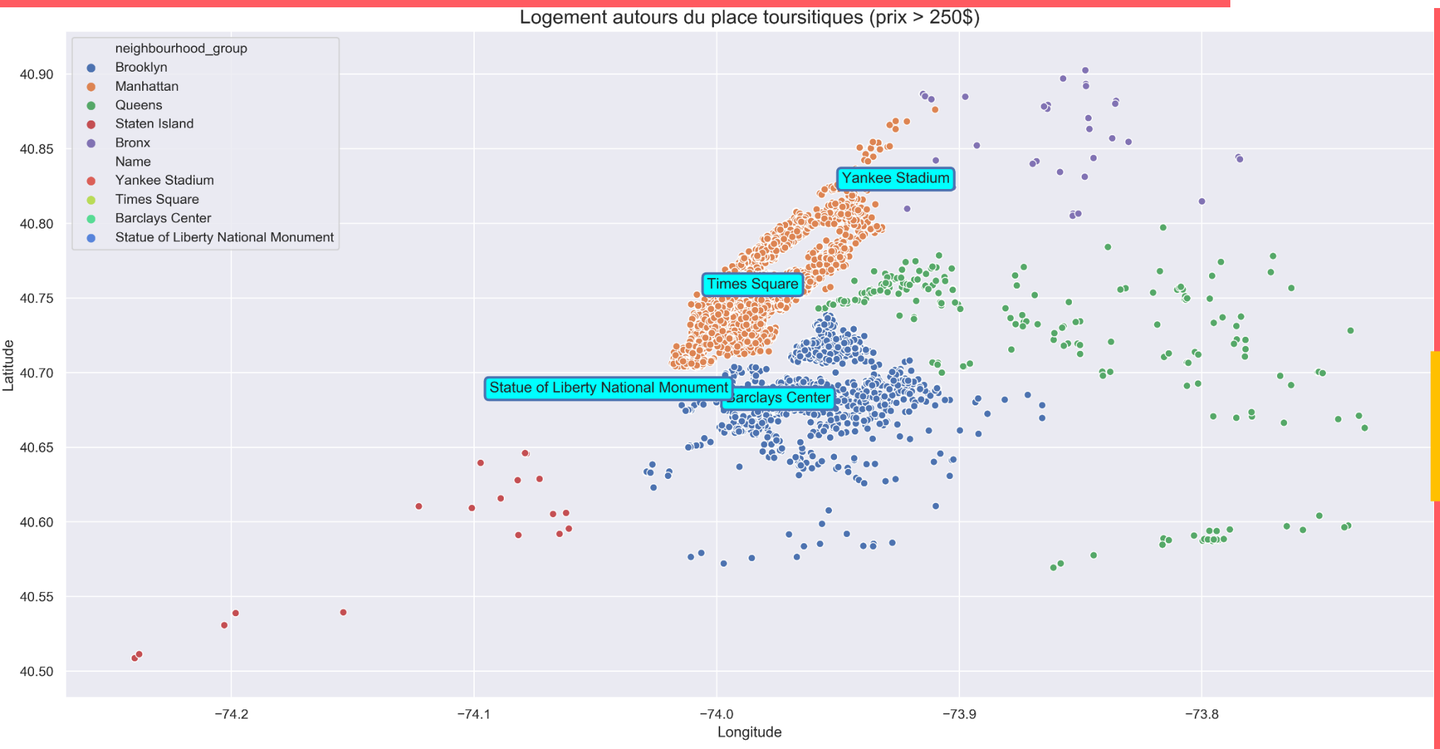
MANHATTAN ET BROOKLYN OCCUPENT LE MARCHÉ DU NEW YORK

+85,4%

EN 2019

LES PRIX SE VARIENT ENTRE 50 ET 150 $ DANS TOUS LES GRANDS QUARTIERS



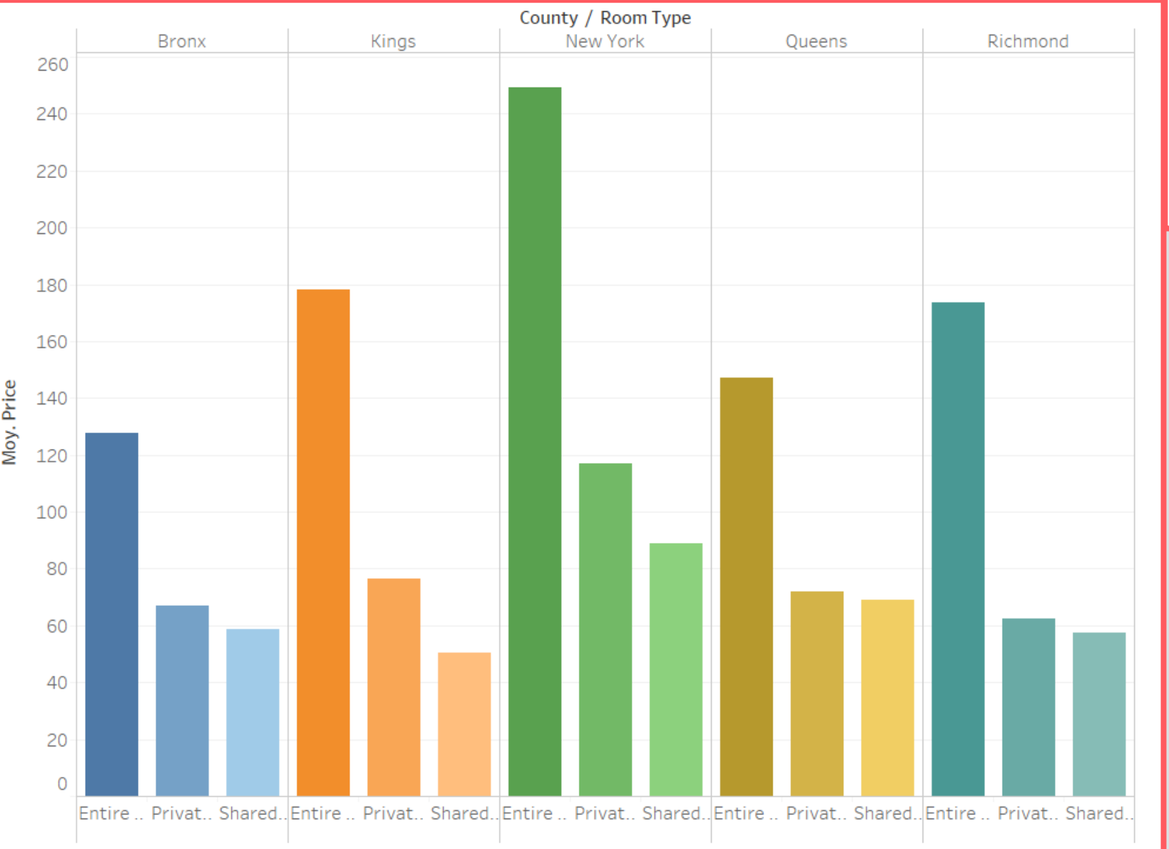


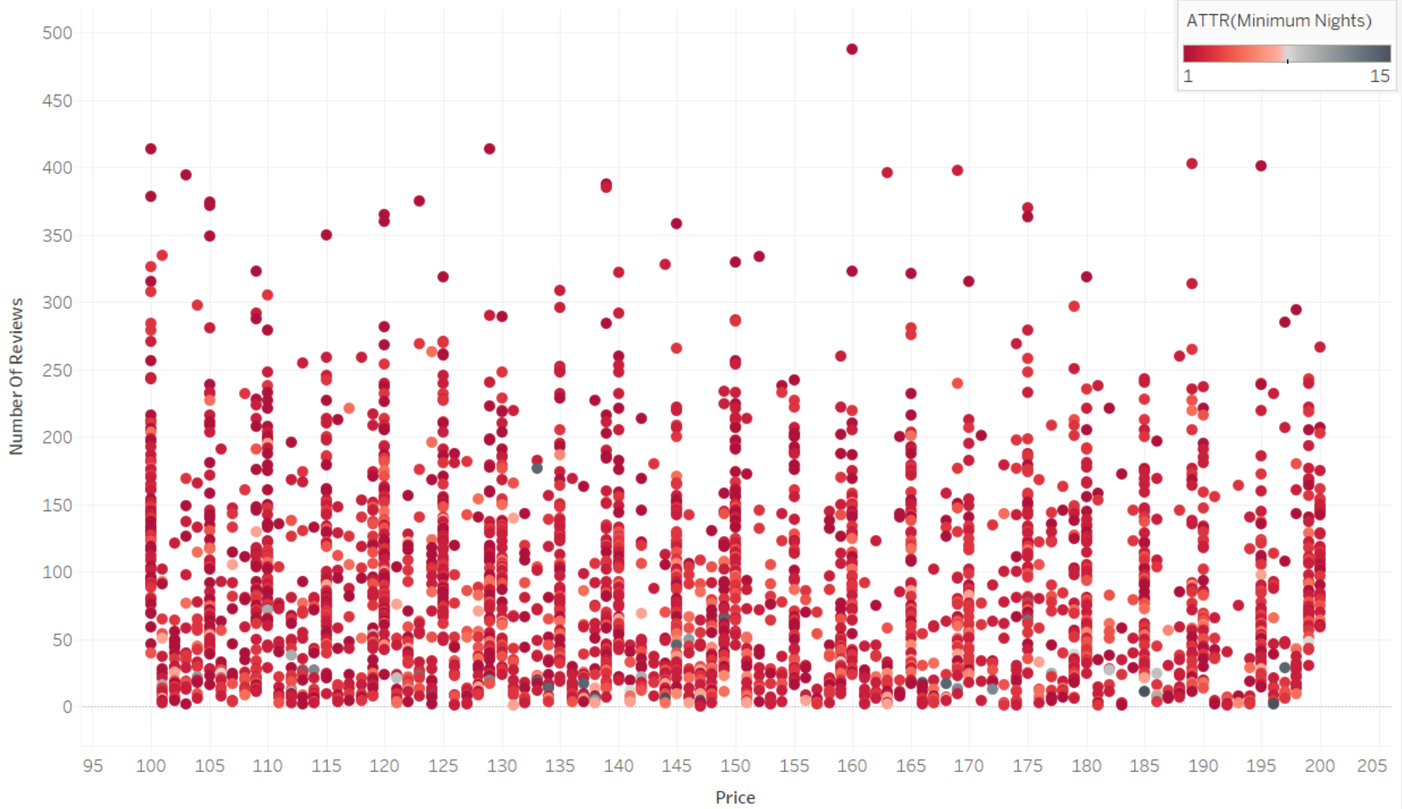
CROISSANCE DU PRIX DÉPEND DES PLACES TOURISTIQUES LES PLUS VISITÉES

EN 2019

“ENTIRE HOME/APT” EST LE TYPE DE LOGEMENTS PRÉFÉRÉS POUR LES CLIENTS

52%

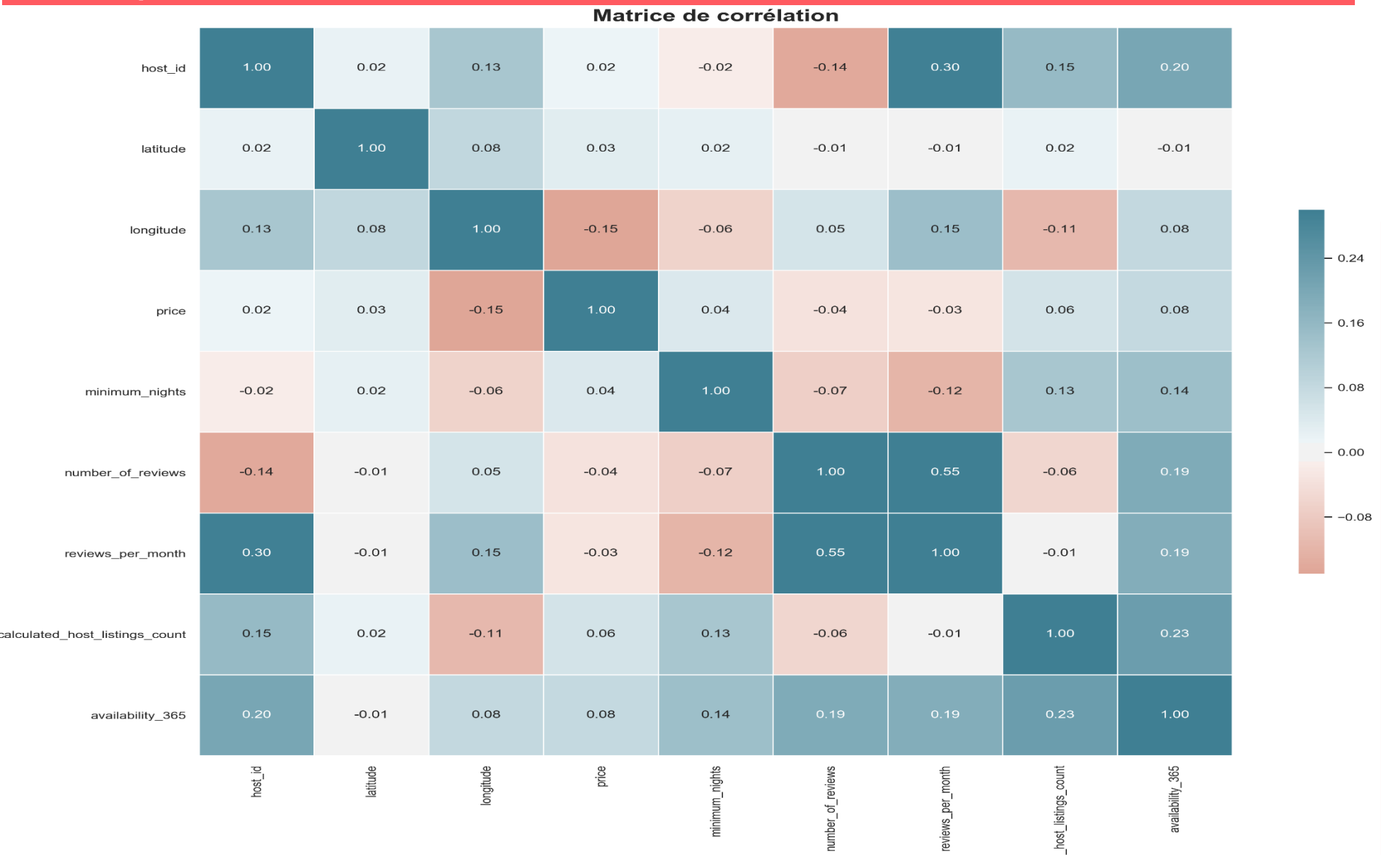


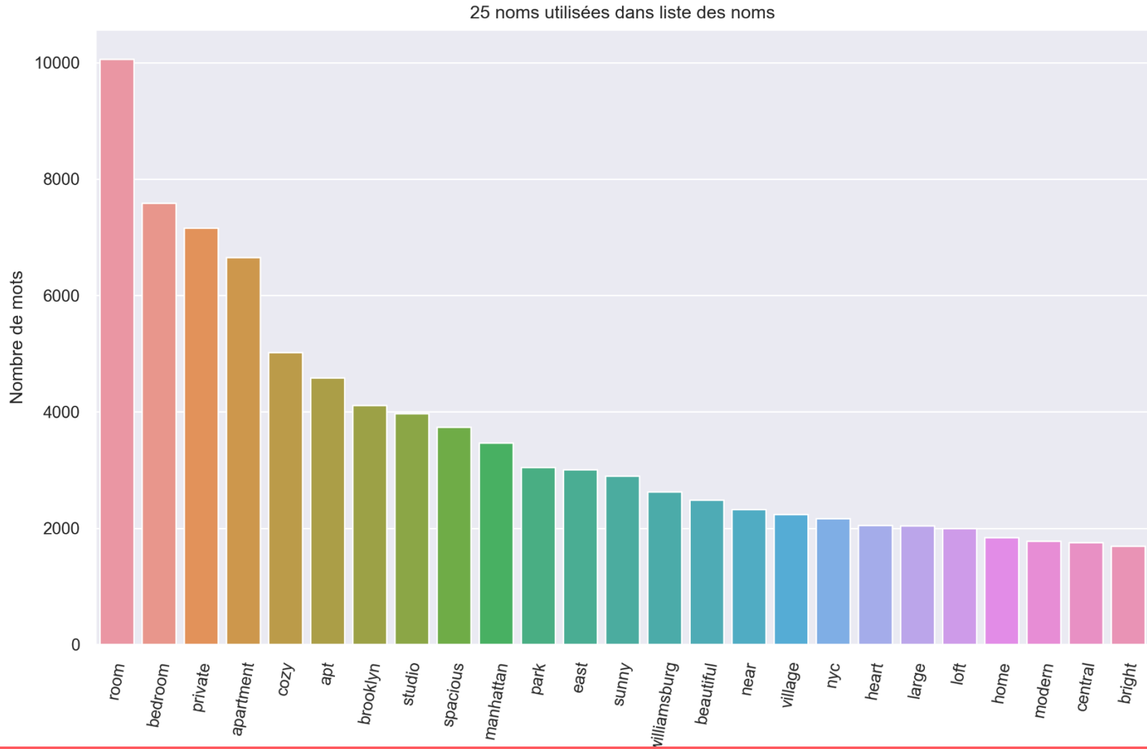


LES LOGEMENTS À COURTE RESERVATION (ENTRE 1 ET 5 NUITS) ET LES PRIX (ENTRE 100 ET 250) ONT UN NOMBRE D’AVIS PLUS IMPORTANT

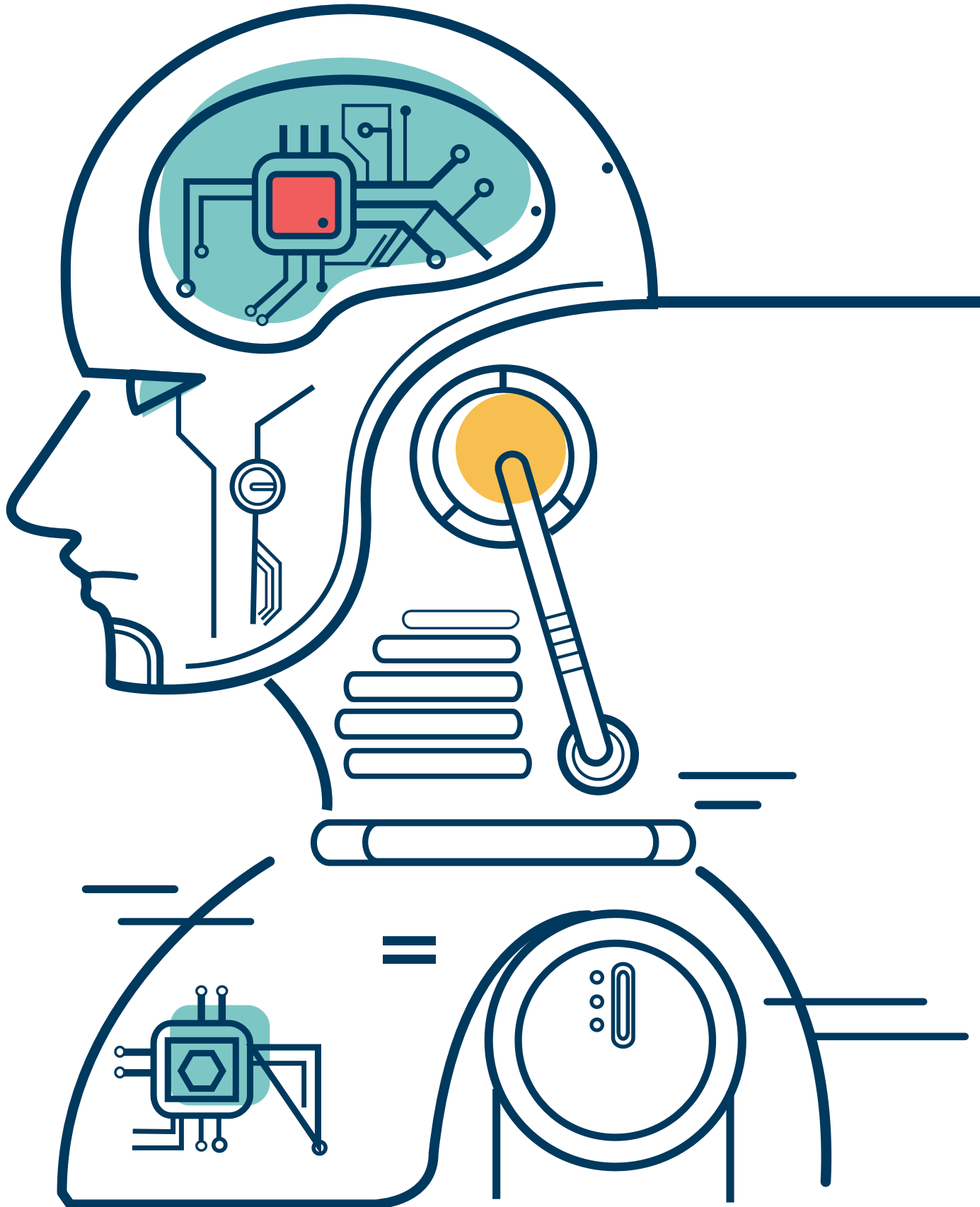
EN 2019

LA LONGITUDE EST ANTICORRÉLÉE (15%) AVEC LE PRIX, BRONX ET QUEENS SONT LES MOINS CHÈRES





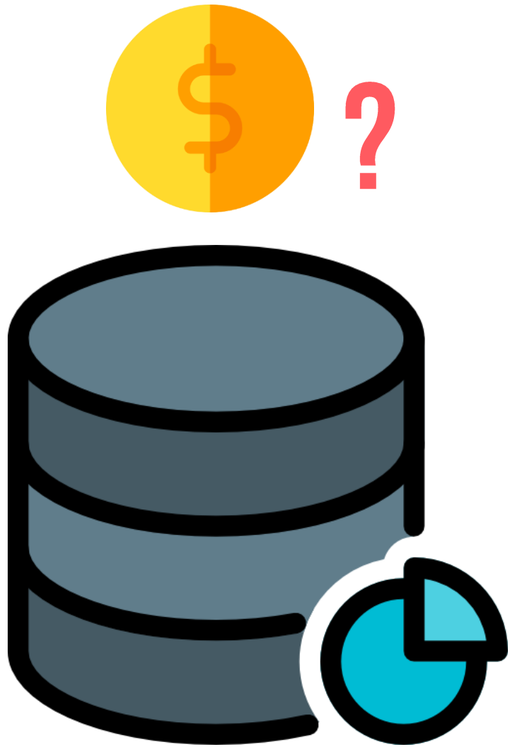
L'UTILISATION DES MOTS TECHNIQUES AUGMENTE LES RÉSERVATIONS



MACHINE LEARNING

MODELING

1- PREPROCESSING
2- METRICS
3- LINEAR REGRESSION
4- XGBOOST
5- DECISION TREE REGRESSION



1- PREPROCESSING

- DEALING WITH MISSING VALUES
- TRANSFORMING TEXT TO NUMERIC
- GENERATING NEW DATA
- SPLITING TRAIN/TEST

Id	Quartier	Latitude	Longitude	Type_chambre	Prix	Minimum_nuit	Avis_par_mois	Nombre_list_hote	Depart_Brooklyn	Depart_Manhattan	Depart_Queens	Depart_Staten Island	Depart_Bronx	Disponabilitie_365_scale	Nomb
2539	109	40.64749	-73.97237	2	149	1	0.21	6	0	1	0	0	0	2.77	
2595	128	40.75362	-73.98377	1	225	1	0.38	2	0	0	1	0	0	2.70	
3647	95	40.80902	-73.94190	2	150	3	0.00	1	0	0	1	0	0	2.77	
3831	42	40.68514	-73.95976	1	89	1	4.64	1	0	1	0	0	0	1.47	
5022	62	40.79851	-73.94399	1	80	10	0.10	1	0	0	1	0	0	0.00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
36484665	14	40.67853	-73.94995	2	70	2	0.00	2	0	1	0	0	0	0.07	
36485057	29	40.70184	-73.93317	2	40	4	0.00	2	0	1	0	0	0	0.27	
36485431	95	40.81475	-73.94867	1	115	10	0.00	1	0	0	1	0	0	0.21	
36485609	96	40.75751	-73.99112	3	55	1	0.00	6	0	0	1	0	0	0.02	
36487245	96	40.76404	-73.98933	2	90	7	0.00	1	0	0	1	0	0	0.17	

2- METRICS

RACINE D'ERREUR QUADRATIQUE MOYENNE RMSE

Plus la valeur rmse est basse, plus la valeur prédite est correcte

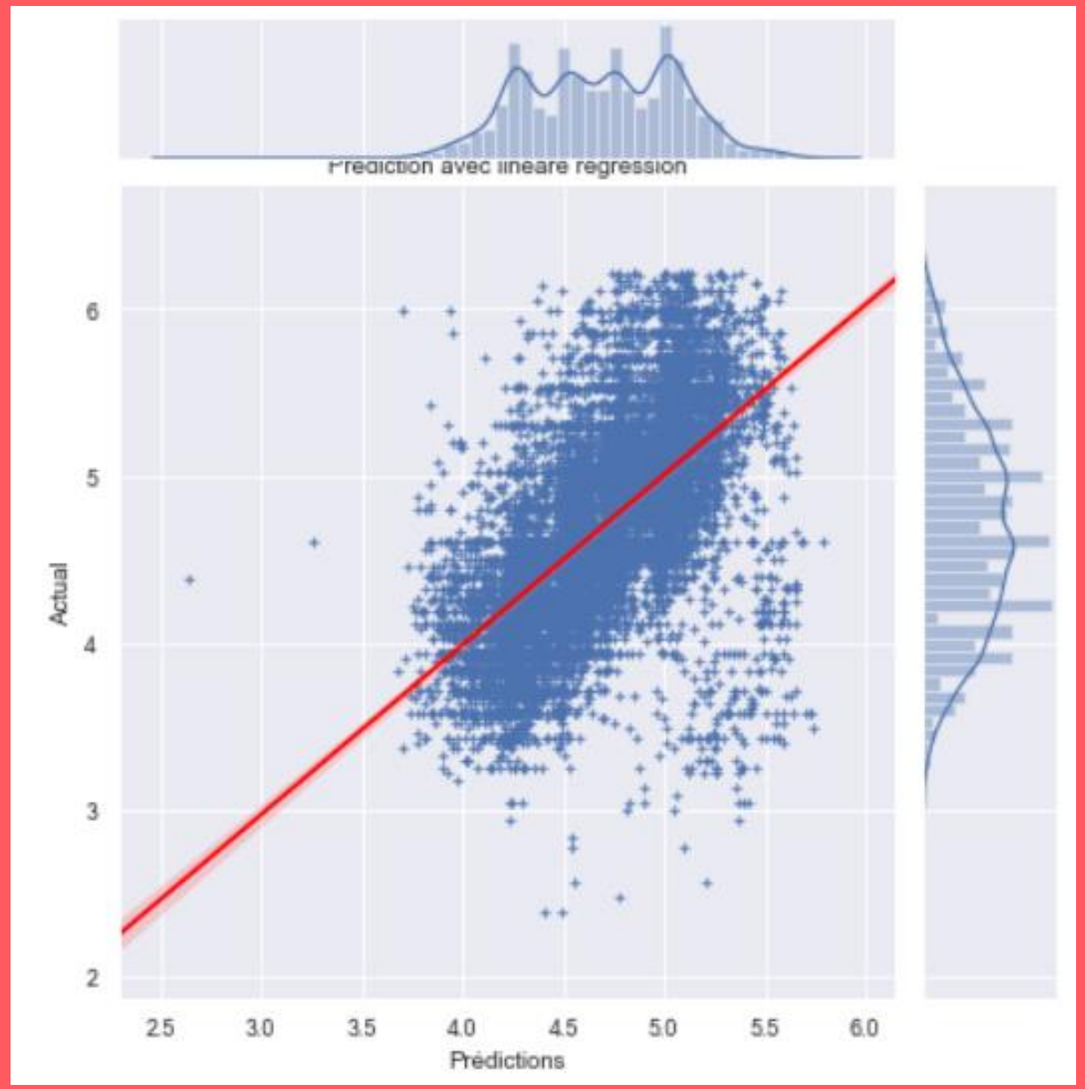
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

R^2 SCORE

Plus le score est proche de 1, plus la valeur prédite est égale à la valeur initial

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

3- LINEAR REGRESSION



RMSE: 227,44

R^2 SCORE: 0,05

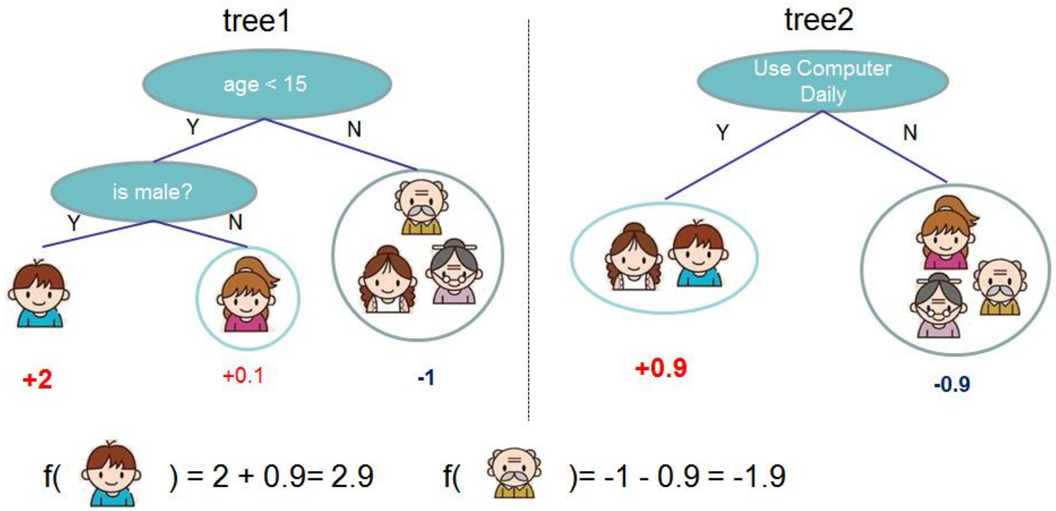
POUR PRIX < 500$

RMSE: 73,04

R^2 SCORE: 0,26

4- XGBOOST

XGBOOST EST PLUS RAPIDE PAR RAPPORT AUX AUTRES IMPLÉMENTATIONS DE GRADIENT BOOST



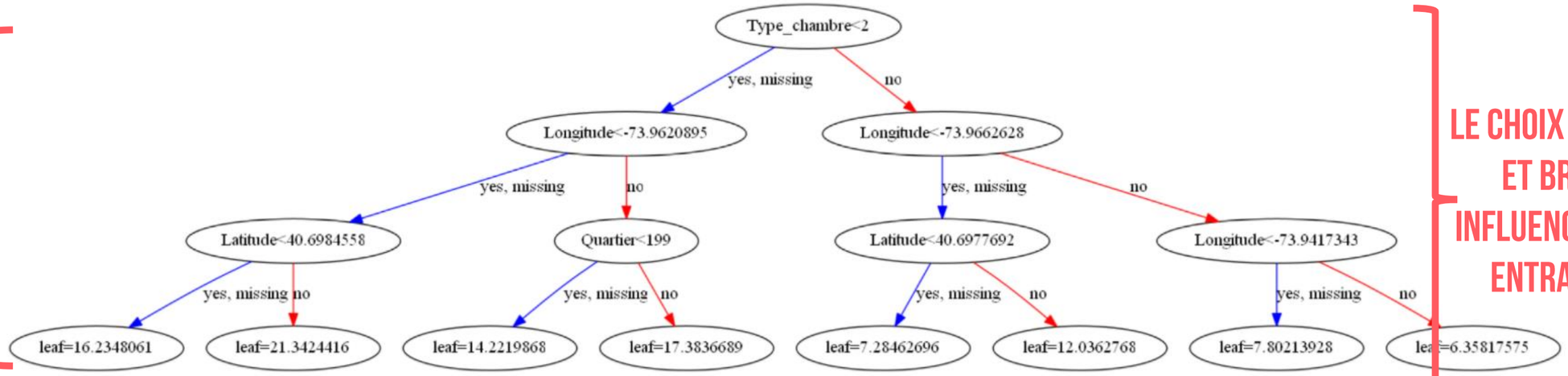
"As the winner of an increasing amount of Kaggle competitions, XGBoost showed us again to be a great all-round algorithm worth having in your toolbox."

| *Dato Winners' Interview*, Mad Professors

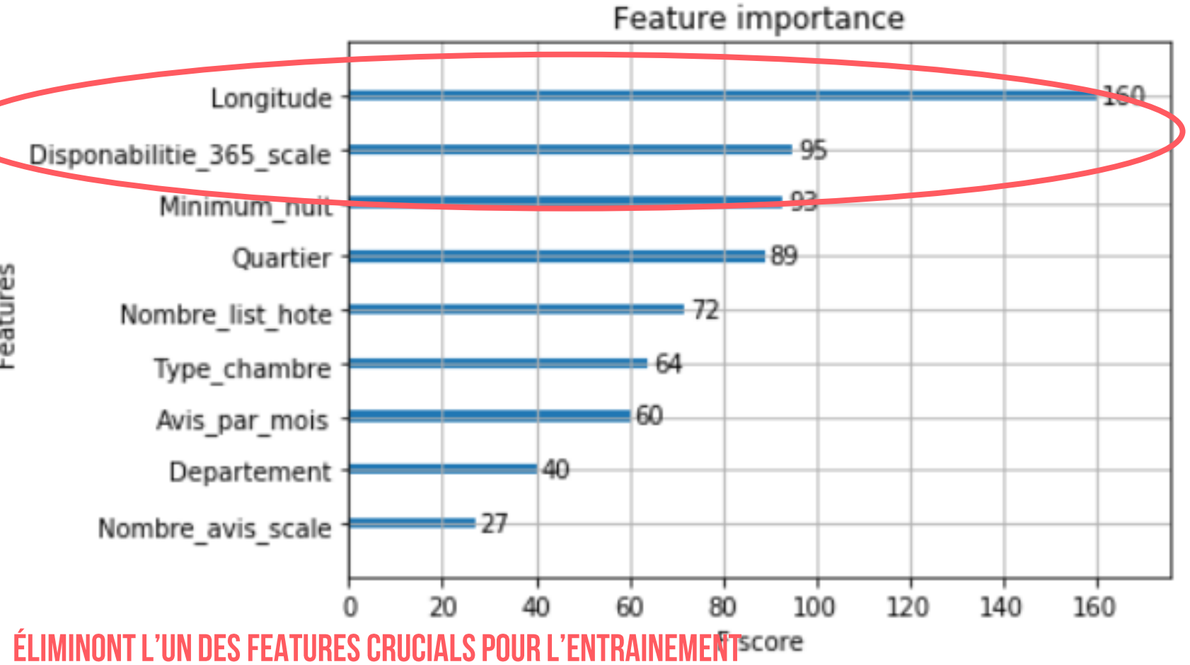
4- XGBOOST

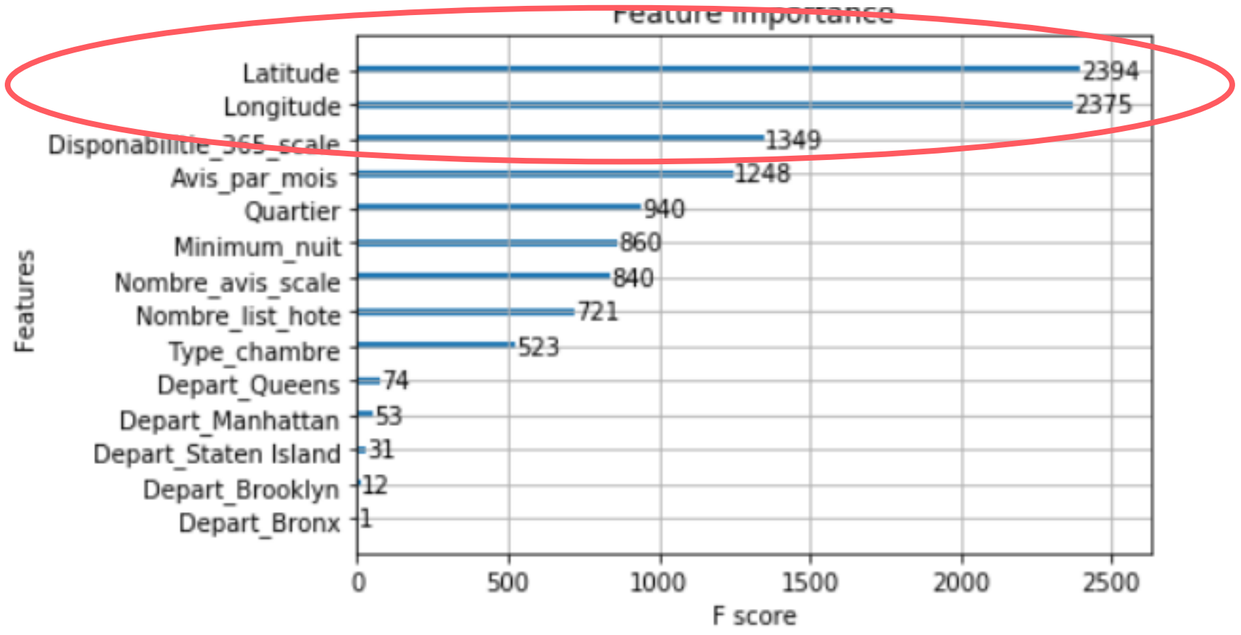
IMPLÉMENTATION

PLUS LA PROFONDEUR EST IMPORTANTE MEILLEURE SERONT LES RÉSULTATS



LE CHOIX DES ARBRES ET BRANCHES INFLUENCENT LE BON ENTRAINEMENT





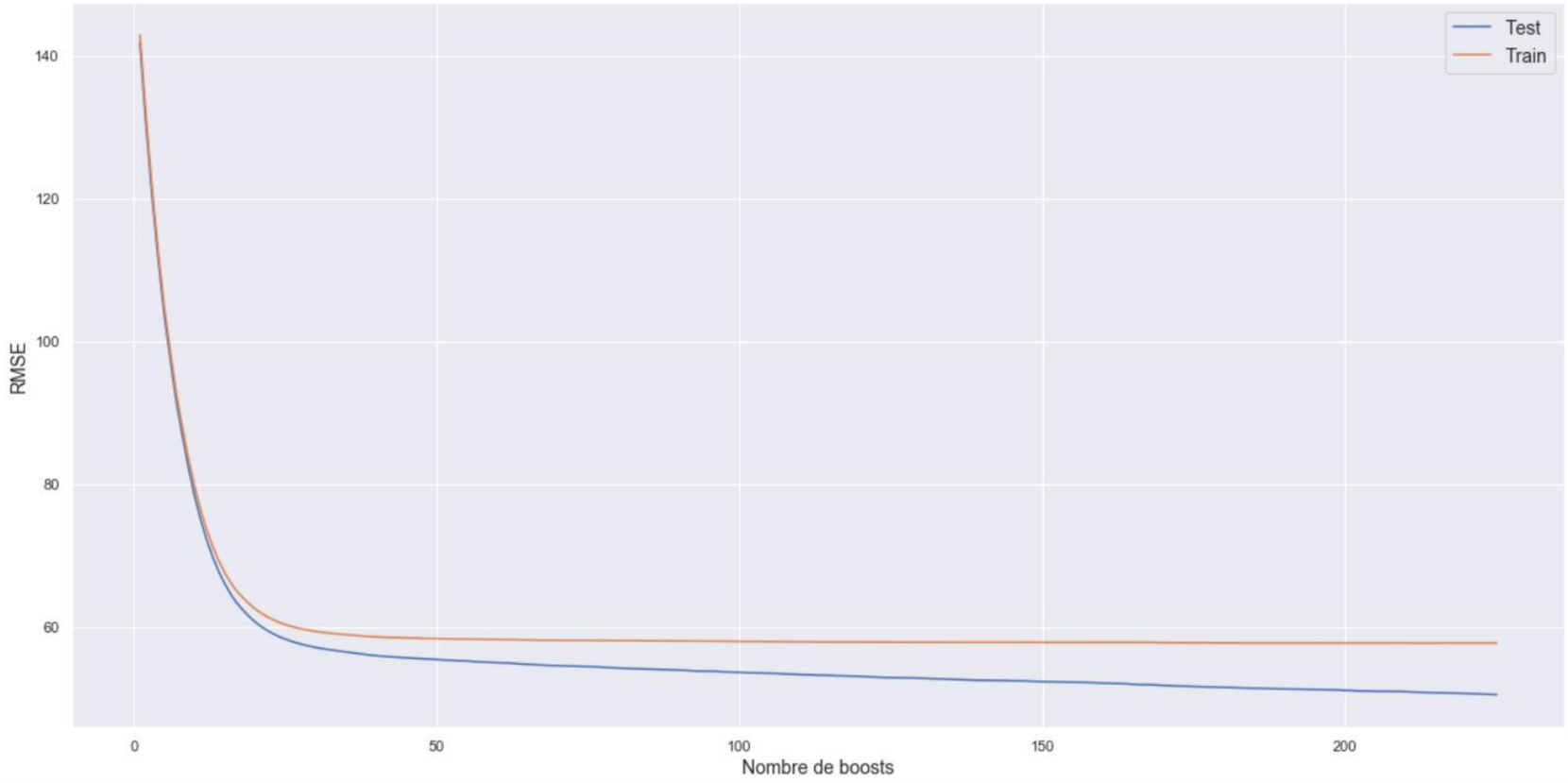
ÉLIMINONT L'UN DES FEATURES CRUCIALS POUR L'ENTRAINEMENT

4- XGBOOST

RÉSULTAT

AVEC HYPERPARAMÉTRATION, ON A CONFIGURE LE MODÈLE POUR QU'IL PUISSE TROUVER LE MEILLEURE ENTRAINEMENT (LEARNING)

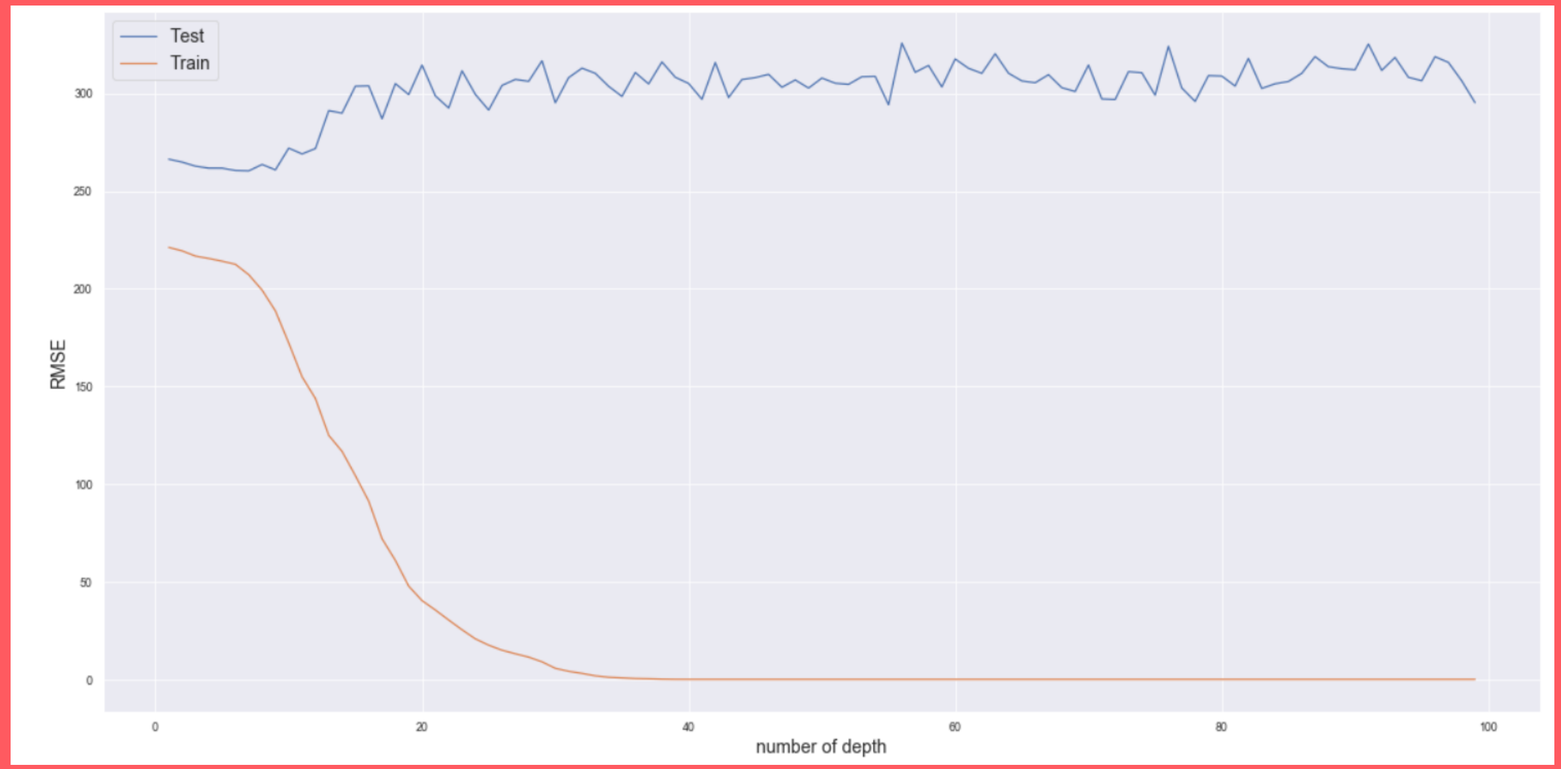
300 ITERATIONS (BOOSTS) LE RMSE DIMINUE JUSQU'À SE STABILISER



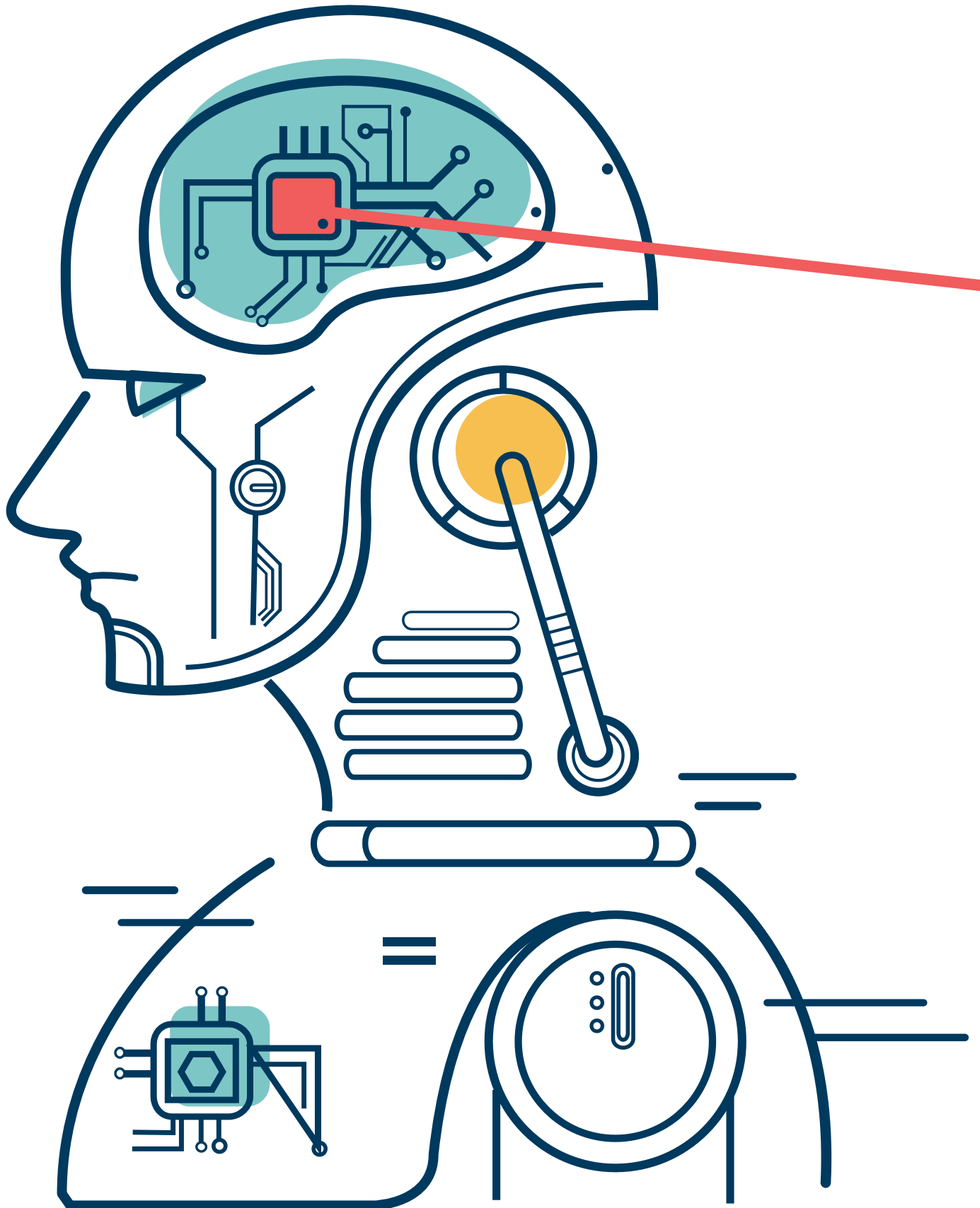
5- DECISION TREE REGRESSION

EN UTILISANT UN AUTRE MODÈLE DU GRADIANT TREE BOOST, ON VOIT QUE LES RÉSULTATS NE SONT PAS BONS

- STABILITÉ EN PHASE TEST SANS DÉGRADATION
- DIMINUTION EN PHASE ENTRAINTEMENT DANS LE RMSE



Mauvais modèle utilisé dans notre cas

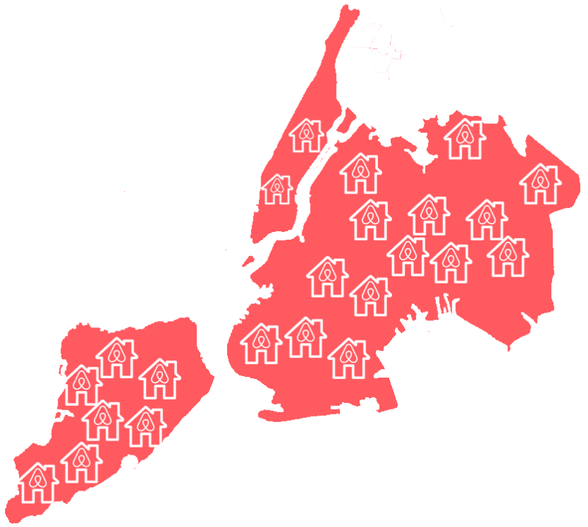


RECOMMANDATION

SUSTAINABILITY

PLUS LES AVIS SONT NOMBREUX SUR LES PRIX LES PLUS BAS, PLUS LES VOYAGEURS RÉSERVENT

LA DATE DE RÉSERVATION AMÉLIORE LA PRÉDICTION POUR UNE MEILLEURE OPTIMISATION



PLUS DE DISPARITÉ DES LOGEMENTS, MEILLEURE PRÉDICTION DES PRIX

TO THE NEXT AIRBNB 2020