



United States™  
**Census**  
Bureau

Projet Machine Learning  
IA School

---

# Objectif:



Exploration des données et analyse statistique sur Python



Modélisation à l'aide d'algorithme d'apprentissage supervisé



Restitution des résultats en soutenance



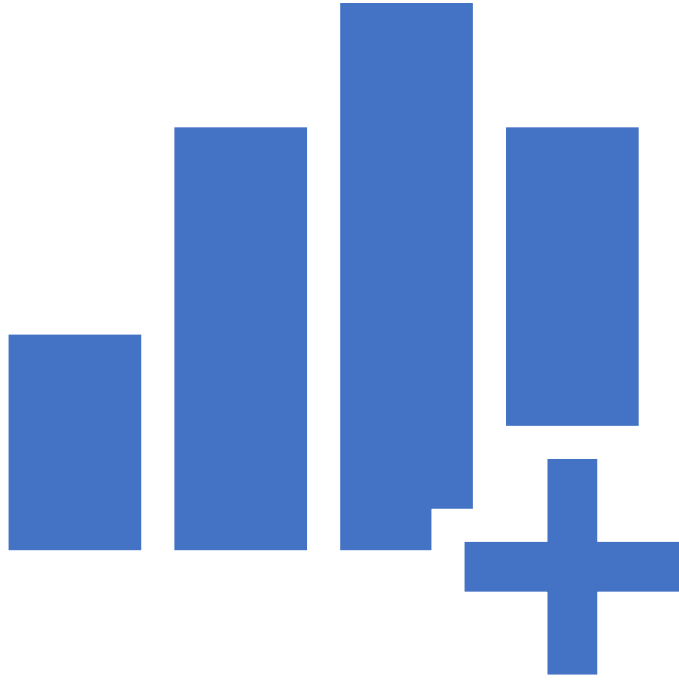
# Groupe

Projet en  
binôme

Durée  
estimée



- 1 mois de preparation: Lundi 28 Juin 2021
- Livrable à mettre sur le Classroom à **8H30 maximum**
- Nomenclature de fichier à respecter  
Prénom\_Nom\_Projet2\_IASCHOOL\_Airbnb.ipynb



# Bloc de compétence à valider

- Constituer un jeu de données exploitable de manière à entraîner un modèle d'apprentissage en utilisant la méthodologie et/ou l'outil approprié en fonction des standards de l'écosystème
- Interpréter les données grâce à des outils de visualisation de données en vue d'expliquer les caractéristiques du jeu de données
- Exploiter un modèle d'apprentissage permettant la classification ou la prédiction d'une variable en fonction des données disponibles et des outils sélectionnés
- Améliorer les performances d'un modèle d'apprentissage à l'aide d'une évaluation de la qualité des données et de la technique de modélisation afin de réduire les biais et les anomalies de résultats
- Concevoir un modèle d'apprentissage efficient en exploitant les méthodes standards d'apprentissage profond pour répondre à une problématique identifiée

## Inside Airbnb

Adding data to the debate

### Paris, Île-de-France, France

See [Paris data visually here](#).

Date Compiled	Country/City	File Name	Description
09 July, 2019	Paris	<a href="#">listings.csv.gz</a>	Detailed Listings data for Paris
09 July, 2019	Paris	<a href="#">calendar.csv.gz</a>	Detailed Calendar Data for listings in Paris
09 July, 2019	Paris	<a href="#">reviews.csv.gz</a>	Detailed Review Data for listings in Paris
09 July, 2019	Paris	<a href="#">listings.csv</a>	Summary information and metrics for listings in Paris (good for visualisations).
09 July, 2019	Paris	<a href="#">reviews.csv</a>	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Paris	<a href="#">neighbourhoods.csv</a>	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Paris	<a href="#">neighbourhoods.geojson</a>	GeoJSON file of neighbourhoods of the city.

[show archived data](#)

# Données Projet 1

<http://insideairbnb.com/get-the-data.html>

Ville de Paris



Data sur Google Classroom

**US census Dataset  
(Classroom)**

Données Projet 2

Pandas



Outils

---



# Les livrables attendus



Un 1 Notebook d'analyses statistiques, une modélisation et une prédiction



Présentation Powerpoint



Soutenance (5 minutes par personne)



# Brief projet 1 Airbnb

- Vous êtes une société immobilière. Vous disposez d'un parc immobilier de 50 appartements et maisons à Amsterdam.

Vous voulez étudier les opportunités d'investissement dans la ville de Paris.

- Pour ce faire, vous disposez d'analyses statistiques réalisées par votre équipe de data scientist sur la ville de paris. (Précédent Projet Data Science)

**Le projet consiste ici à prédire les prix des locations Airbnb sur la ville de Paris à 3, 6 et 12 mois.**

- On téléchargera le fichier `listing.csv.gz` pour faire nos analyses

## Paris, Île-de-France, France

See [Paris data visually here](#).

Date Compiled	Country/City	File Name	Description
09 July, 2019	Paris	<a href="#">listings.csv.gz</a>	Detailed Listings data for Paris
09 July, 2019	Paris	<a href="#">calendar.csv.gz</a>	Detailed Calendar Data for listings in Paris
09 July, 2019	Paris	<a href="#">reviews.csv.gz</a>	Detailed Review Data for listings in Paris
09 July, 2019	Paris	<a href="#">listings.csv</a>	Summary information and metrics for listings in Paris (good for visualisations).
09 July, 2019	Paris	<a href="#">reviews.csv</a>	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Paris	<a href="#">neighbourhoods.csv</a>	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Paris	<a href="#">neighbourhoods.geojson</a>	GeoJSON file of neighbourhoods of the city.

[show](#) archived data

# Les livrables attendus sur le notebook



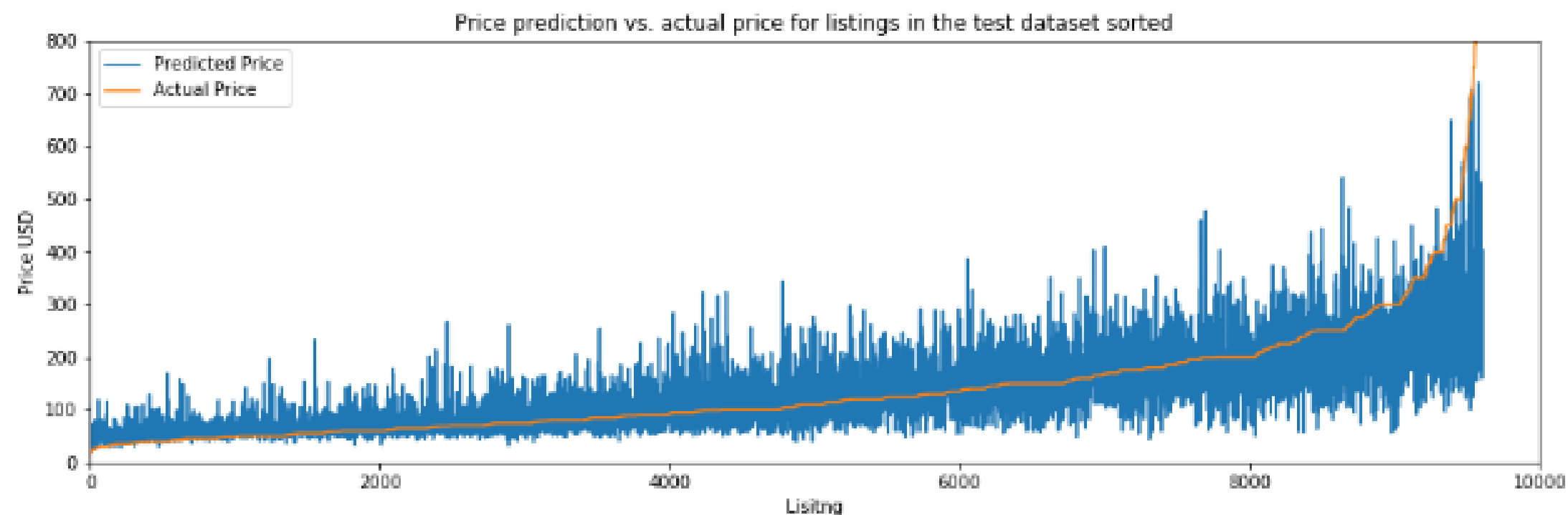
Un 1 Notebook avec au moins 4 analyses statistiques (graphique, correlation etc...)



Une modélisation (Au moins 3 algorithmes de regression à utiliser )



Un Grid Search sur les 2 meilleurs algorithmes issus de la base line



# Brief projet 2 - US Census

---

- Nous chercherons à prédire le niveau de revenus chez un adulte américain s'étant fait recenser.
- Notre variable à prédire est **income**. Elle correspond aux revenus des adultes sur ce recensement. Cette variable est binaire. On commencera par charger les librairies utiles pour notre projet.

\* Le projet consiste ici à prédire les modalités de la variable "Income"





Suite

- Il comprend 15 variables et 48842 lignes. Les variables correspondant aux champs suivants:
- age
- workclass
- fnlwgt
- education
- education.num
- marital.status
- Occupation
- Relationship
- race
- sex
- capital.gain
- capital.loss
- hours.per.week
- native.country
- income

# Les livrables attendus sur le notebook



Un 1 Notebook avec au moins 4 analyses statistiques



Une modélisation (Au moins 5 algorithmes de classification à utiliser)



Un Grid Search sur les 2 meilleurs algorithmes issues de la base line





Et  
surtout...une  
histoire !