ÉCOLE POLYTECHNIQUE
EXECUTIVE EDUCATION

2018

DATA SCIENCE STARTER PROGRAM

Soutenance : Conduire un projet de sciences de données

Mohamed Bouzid

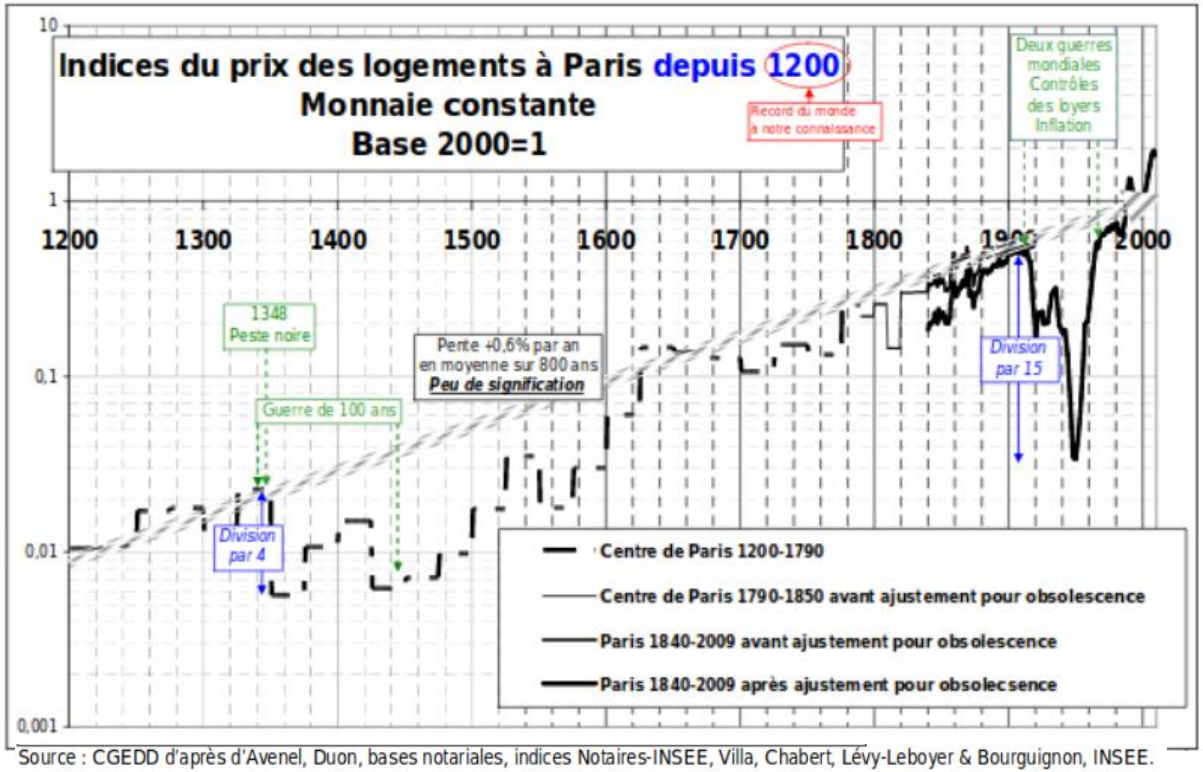
ÉTUDE DU PRIX DE L'IMMOBILIER À PARIS

PROBLÉMATIQUE
ET SOMMAIRE

« Il n'est pas toujours aisé de comprendre où, comment, ni même pourquoi le prix de l'habitat ne cesse d'augmenter dans la capitale. Notre problématique consistera donc à éluder dans la mesure du possible ces questions pour affiner au mieux la connaissance du logement dans la ville. »

- Historique de l'immobilier parisien
- Bases de données contemporaines : Castorus, BIEN, Open Data Paris
- Outils enseignés au DSSP : Numpy, Pandas (Geopandas), régression linéaire
- Cartographies de l'immobilier parisien : marché, ventes réelles, comparatif
- Existe-t-il une logique des prix du marché ?
- Conclusion et perspectives

HISTORIQUE DE L'IMMOBILIER PARISIEN



Source : CGEDD d'après d'Avenel, Duon, bases notariales, indices Notaires-INSEE, Villa, Chabert, Lévy-Leboyer & Bourguignon, INSEE.

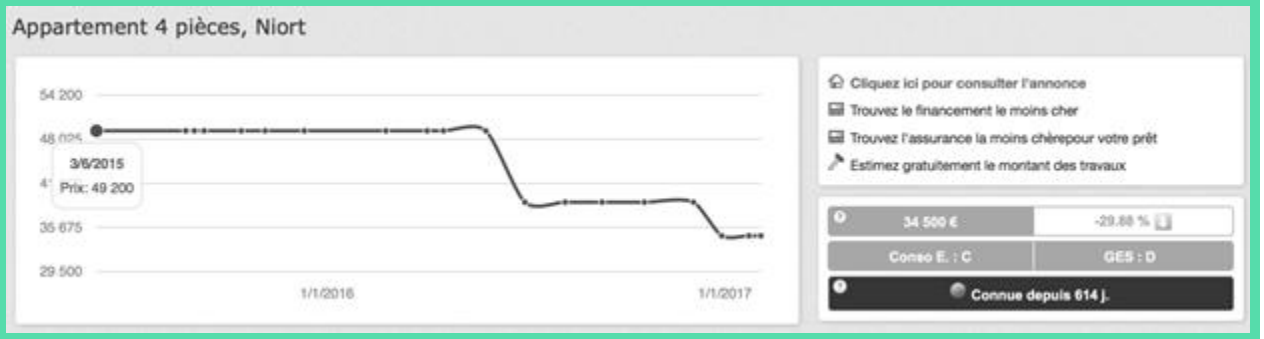
- 1348
  - Épidémie de Peste à Paris
- 1337 - 1453
  - Guerre de Cent Ans
- Fin XV^e - début XVII^e
  - Période de la Renaissance
- 1789
  - 1830
  - 1848
  - Révolutions Françaises
- Fin XIX^e
  - Remboursement de l'indémnité envers l'Allemagne
- 1914 - 1918
  - 1939 - 1945
  - Guerres Mondiales
- 2002
  - 2008
  - Kraks boursiers du XXIe siècle

SITUATION ACTUELLE

- **OBJECTIF DE L'ACHETEUR :**
  - Réaliser la « bonne affaire »
  - Maximiser le retour sur investissement
- **SE RENSEIGNER :**
  - Prix du marché (Castorus)
  - Prix effectif de vente (base notariale BIEN)
- **VISUALISER CES PRIX EN FONCTION DE LA LOCALITÉ**
  - Carte de la ville (Open Data Paris)

CASTORUS

BASE DE DONNÉES SUR LES PRIX DU MARCHÉ



- **DÉFINITION :** Analyse et historique des annonces immobilières professionnelles
- **PRINCIPALES CARACTÉRISTIQUES :**
  - – Informations sur l'annonce (date, modifications du prix, prix moyen/m², etc.),
  - – Communautaire

Quantitatif : 10 000 annonces
Qualitatif : Pas de doublons

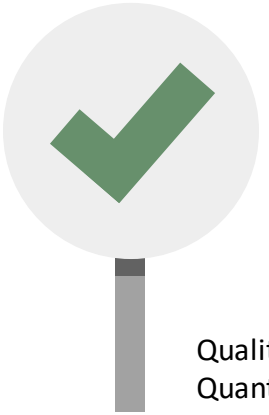
Ne fonctionne pas si l'URL change

BIEN

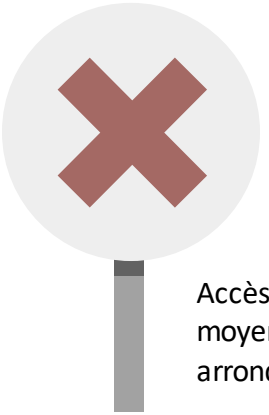
trim	1er	2e	3e	4e	5e	6e	7e
T1 2018	12 060 €	10 880 €	10 990 €	11 920 €	11 390 €	12 240 €	11 760 €
T4 2017	11 560 €	10 890 €	11 170 €	11 540 €	11 100 €	12 360 €	11 530 €
T3 2017	11 380 €	10 590 €	11 020 €	11 840 €	11 210 €	12 020 €	11 980 €
T2 2017	11 060 €	10 080 €	10 310 €	11 990 €	11 010 €	12 090 €	11 690 €
T1 2017	11 410 €	9 830 €	10 490 €	11 630 €	11 040 €	11 730 €	10 720 €
T4 2016	10 250 €	9 630 €	10 260 €	11 310 €	10 720 €	11 240 €	11 130 €
T3 2016	10 860 €	9 920 €	9 740 €	11 550 €	10 510 €	11 250 €	11 060 €
T2 2016	10 630 €	9 230 €	9 960 €	11 300 €	10 390 €	11 330 €	11 230 €
T1 2016	10 390 €	9 670 €	9 610 €	11 330 €	9 910 €	11 370 €	11 070 €
T4 2015	11 030 €	9 290 €	9 550 €	11 130 €	10 300 €	11 160 €	10 860 €
T3 2015	10 500 €	9 560 €	9 820 €	11 520 €	9 940 €	11 560 €	10 810 €

- **DÉFINITION**
  - Base d'informations économiques notariales, depuis 1990

- **PRINCIPALES CARACTÉRISTIQUES**
  - Suivi de l'évolution de la valeur du prix d'un bien immobilier
  - Conseil et recommandations des clients
  - Volume : ~200 000 nouvelles références/an (total actuel : ~ 3 000 000 de références)

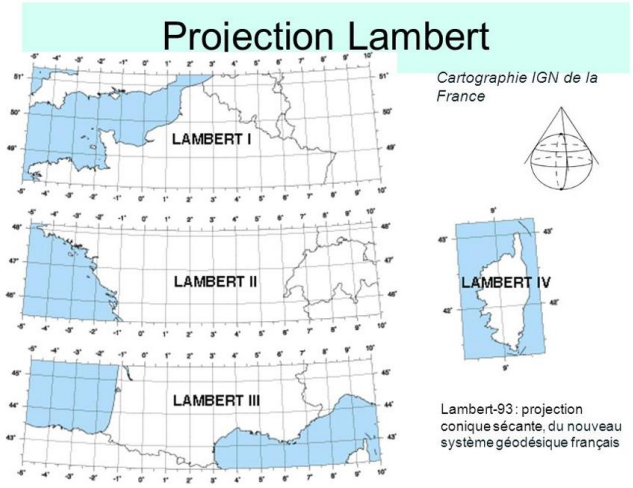


Qualitatif : actes vérifiés
Quantitafif : volume de données



Accès seulement au prix/m² moyen par trimestre et par arrondissement

LA BASE DE DONNÉES DES NOTAIRES



OPEN DATA PARIS

HISTORIQUE

2011 : données utilisées par la mairie en accessibilité libre

USAGE

- Récupération des coordonnées officielles de la Mairie de Paris
  - Projection : Lambert 1
  - Fractionnement : arrondissement « municipal »

Nom	Libellé	Type	O	Valeurs possibles
N_SQ_AR	Identifiant séquentiel de l'arrondissement	N	O	
C_AR	Numéro d'arrondissement	N	O	De 1 à 20
C_ARINSEE	Numéro d'arrondissement INSEE	N	O	De 75101 à 75120
L_AR	Nom de l'arrondissement	C30	O	Ex : 1er Ardt
L_AROFF	Nom officiel de l'arrondissement	C30	O	Ex : Louvre

LES COORDONNÉES DE LA CAPITALE

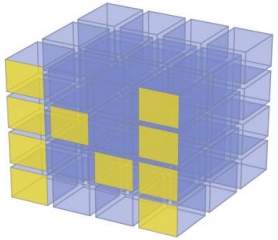
UTIVE
ATION

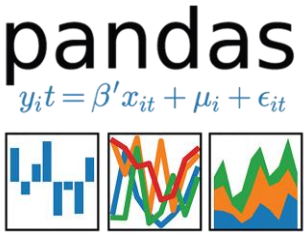
MÉTHODES ET OUTILS ENSEIGNÉS AU DSSP

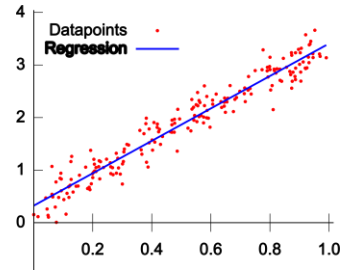
NUMPY

PANDAS (GEOPANDAS)

RÉGRESSION LINÉAIRE







PROBLÉMATIQUE

NUMPY
PANDAS

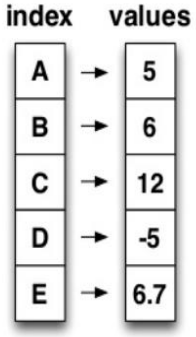
NUMPY

- Opérations d'algèbre linéaire
- Calculs matriciels rapides

PANDAS

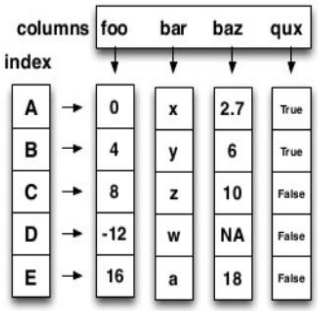
- Librairie de manipulation de tableaux de données
- Fonctionne avec des librairies d'algèbre matricielle (Numpy) et de visualisation (Matplotlib, Seaborn)
- 2 structures : Series & Dataframes
- Indexation des éléments

Series



- Subclass of `numpy.ndarray`
- Data: any type
- Index labels need not be ordered
- Duplicates are possible (but result in reduced functionality)

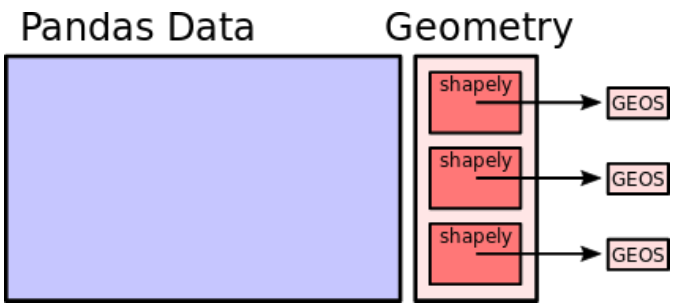
DataFrame

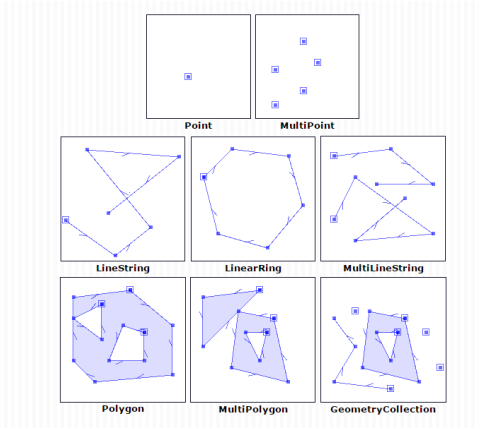


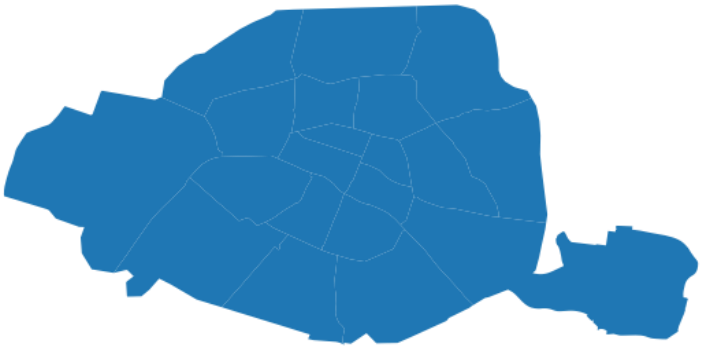
- NumPy array-like
- Each column can have a different type
- Row and column index
- Size mutable: insert and delete columns

GEOPANDAS

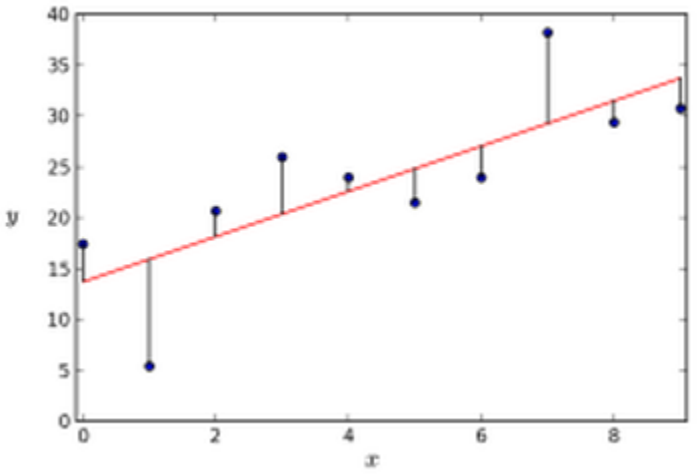
- Projet de 2013 par Kesley Jordahl
- Découle directement de Pandas
- Allège le traitement des données spatiales
- 2 structures (attributs + géométrie) :
  - – Series → GeoSeries
  - – Dataframes → GeoDataframes
- Possibilité de combiner avec d'autres librairies







Cartographier les données Open Data Paris

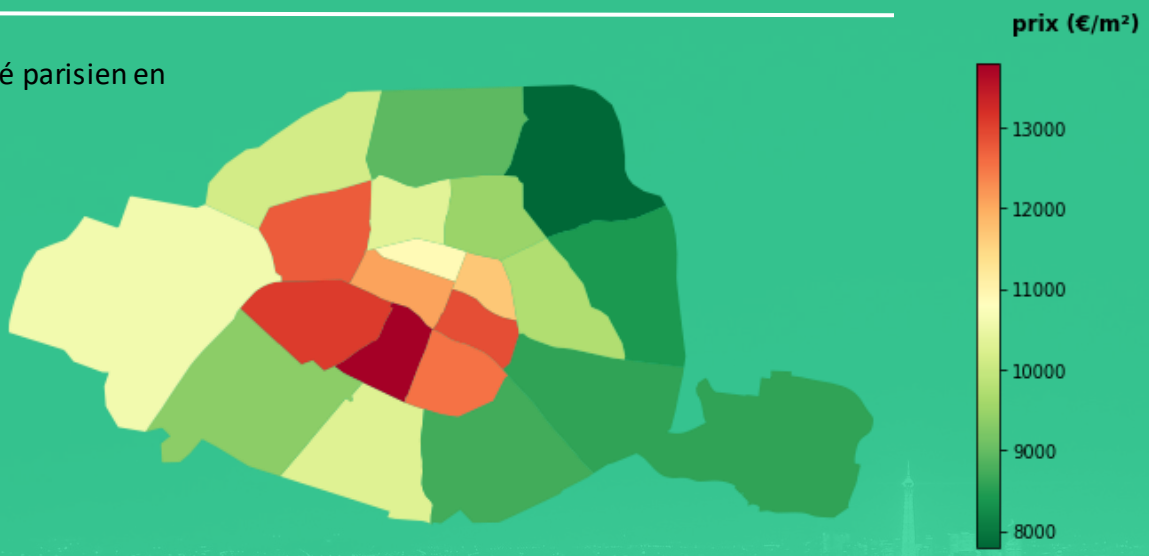


LA RÉGRESSION LINÉAIRE

- Existance d'une relation données observées - projection
- Cette projection = phénomène linéaire
- Droite correspondant « le mieux » aux données sous la forme d'une droite $y_i = \beta_0 + \beta_1 x_i$
- Scipy et Scikit-learn intègrent la régression linéaire

CARTOGRAPHIE DU MARCHÉ IMMOBILIER PARISIEN

Marché parisien en 2018 :



LIBRAIRIES UTILISÉES : Matplotlib, Pandas, Geopandas, Shapely
DONNÉES D'OPEN DATA PARIS : fichier GeoJSON
DONNÉES CASTORUS : fichier CSV (~9300 lignes, 10 colonnes)
PRINCIPE : fusion des deux souces via la fonction *pandas.DataFrame.merge*
NETTOYAGE : conversion des prix en float, moyenne des prix/m² avec *pandas.DataFrame.groupby.mean*, etc.

OBSERVATIONS

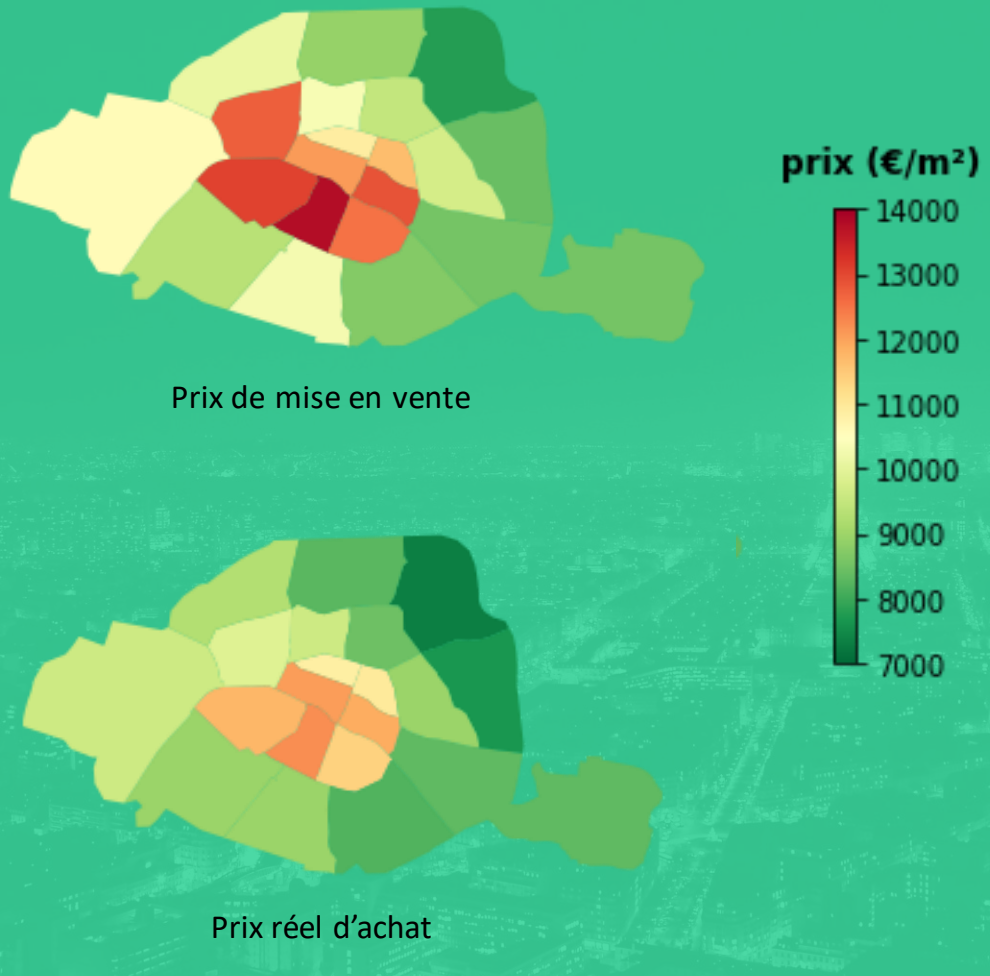
Concentration d'arrondissements

– « chers » au centre

– « moins chers » vers la périphérie

CATÉGORIE	OBSERVATION
Prix moyen	10 613 €/m²
Arrondissement le plus cher	6ème (13 801 €/m²)
Arrondissement le moins cher	19ème (7 804 €/m²)

COMPARATIF DU MARCHÉ IMMOBILIER AVEC LES PRIX DE VENTE RÉELS



Prix de mise en vente

Prix réel d'achat

PRINCIPE : extraction PDF avec *Tabula*
NETTOYAGE : comme précédemment

- + transposition du tableau extrait
- + comparaison avec *matplotlib.pyplot.subplo*

OBSERVATIONS

Concentration d'arrondissements

– « chers » au centre

– « moins chers » vers la périphérie

Prix réels atténués

CATÉGORIE	OBSERVATION
Prix moyen	9 752 €/m²
Arrondissement le plus cher	6ème (12 240 €/m²)
Arrondissement le moins cher	19ème (7 350 €/m²)

COMPARATIF DU MARCHÉ IMMOBILIER AVEC LES PRIX DE VENTE RÉELS

Écart de prix/m² :

OBSERVATIONS

Moyenne des écarts de prix au m² : 7,9 %.

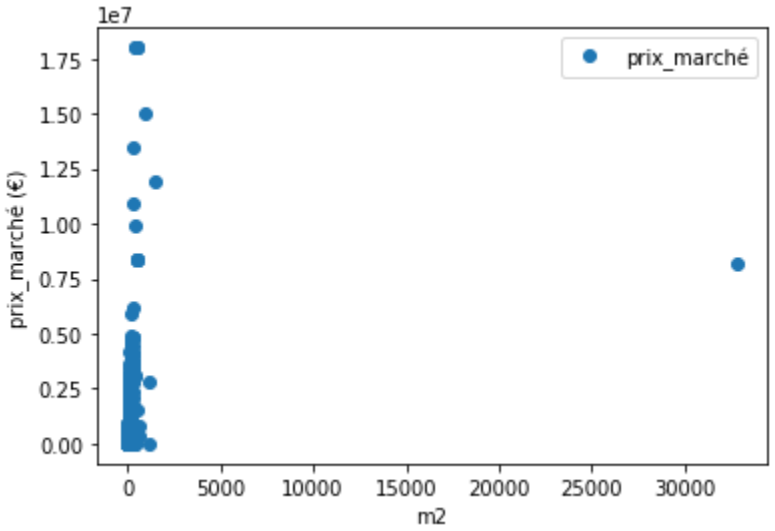
Maximums : 8ème arrondissement (22,3 %), puis 14ème arrondissement (12,7 %).

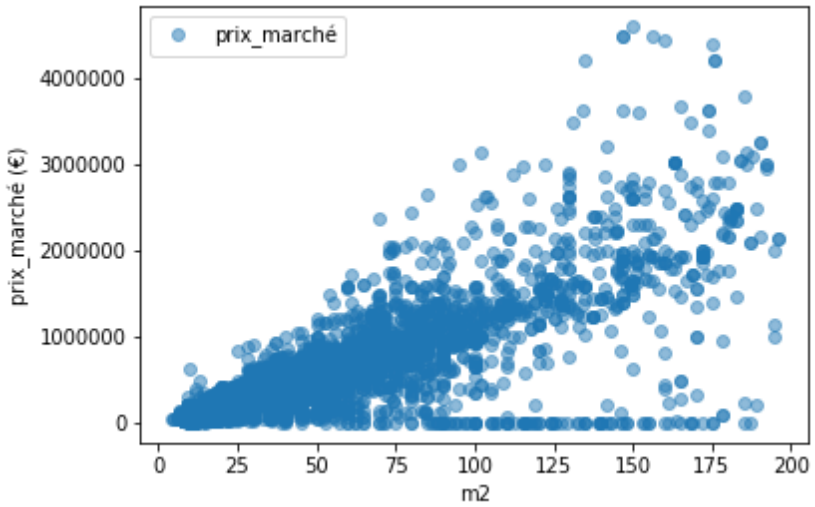
Minimums : 2ème arrondissement (0,2 %), puis 1er arrondissement (0,4 %).

CATÉGORIE	OBSERVATION
Écart moyen	7,9 %
Arrondissement le plus surévalué	8ème (22,3 %)
Arrondissement le plus fidèle à la réalité	2ème (0,2 %)

EXISTE-T-IL UNE LOGIQUE DES PRIX DU MARCHÉ ?

- Seule la base de données Castorus permet de vérifier le lien prix VS m²





OBSERVATIONS

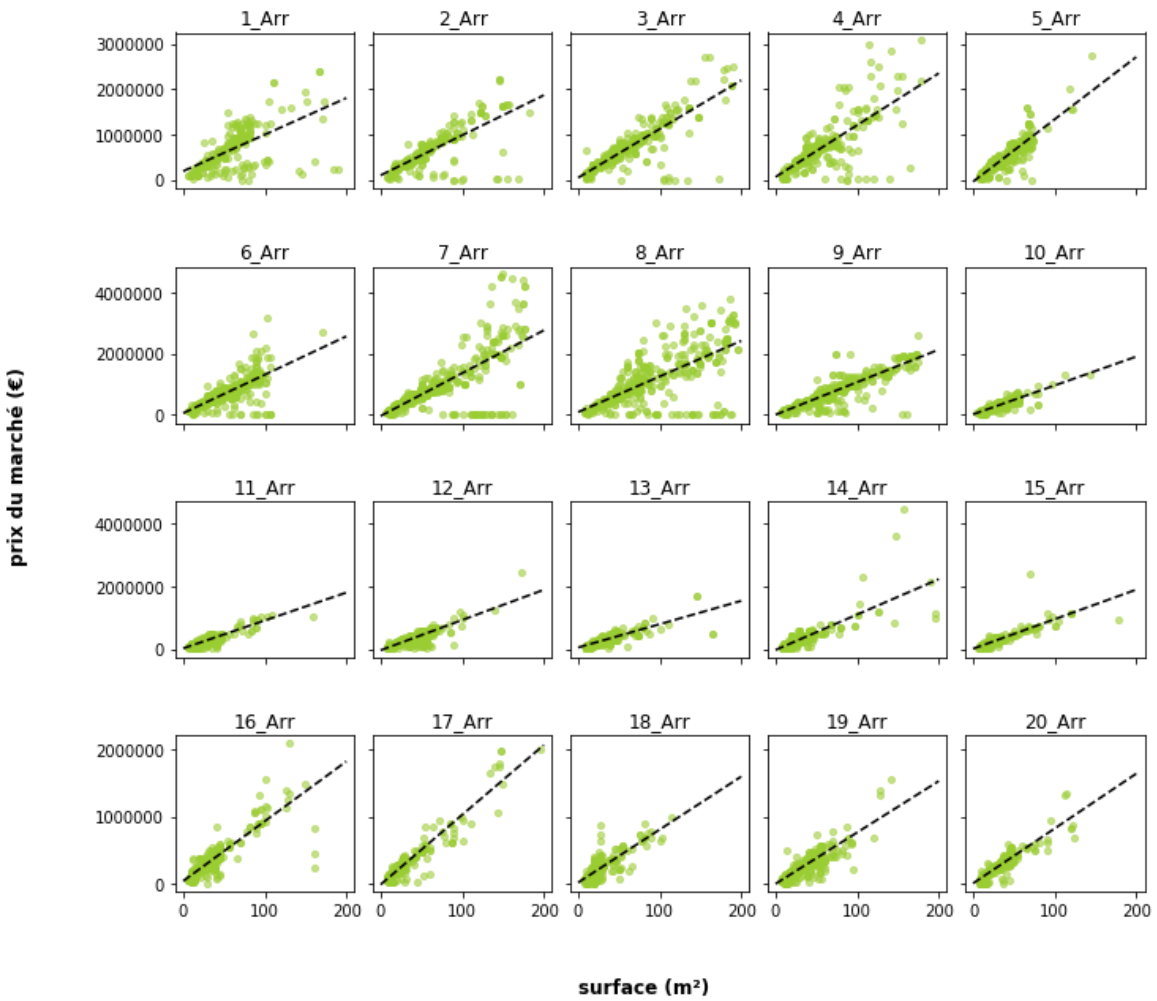
La nuage de points montre quelques informations abbérantes → nécessité de se focaliser sur la modélisation.

Prix des biens lié linéairement à la surface.

Traînée de points parallèle à l'axe des abscisses (indépendants de la surface ?)

Quelques tarifs au dessus de la tendance principale

EXISTE-T-IL UNE LOGIQUE DES PRIX DU MARCHÉ ?



ARR.	ORDONNÉE À L'ORIGINE	PENTE	R^2
17ème	-5718,19	10343,31	**0,9**
20ème	11569,11	8136,71	0,83
10ème	4747,78	9448,02	0,81
9ème	-7709,77	10656,47	0,8
11ème	22087,85	8915,25	0,8
5ème	-42946,41	13783,5	0,78
16ème	45536,78	8872,26	0,75
19ème	4819,64	7620,87	0,74
12ème	-37773,81	9659,69	0,73
3ème	43717,98	10781,62	0,72
15ème	9962,35	9444,03	0,72
18ème	23379,96	7864,79	0,68
13me	49984,82	7456,06	0,65
14ème	-27794,92	11311,81	0,6
6ème	42003,07	12582,88	0,59
4ème	61205,67	11462,57	0,57
2ème	97812,82	8854,07	0,55
7ème	-48637,81	14014,41	0,54
8ème	76959,23	11670,55	0,48
1er	188978,63	8083,06	**0,36**

- **PRIX DES BIENS LIÉ LINÉAIREMENT À LA SURFACE :** projections possibles ! (c.f. tableau)
- **TRAÎNÉE DE POINTS PARALLÈLE À L'ABSCISSE :** quasi inexistant par arrondissement (facteur externe, erreur de saisie?)
- **TARIFS AU DESSUS DE LA TENDANCE :** 7ème et 8ème (surévaluation?), 4ème

CONCLUSION

CARTOGAPHIE

- A l'aide de Geopandas et des bases de données, nous avons pu visualiser :
  - les prix du marché immobilier parisien
  - les prix de vente réelles
- Les arrondissements centraux sont plus chers qu'à la périphérie
- Il existe une surévaluation du prix des logements :
  - de ~8 % en moyenne
  - avec un maximum dans le 8ème (22,3%)

MODÉLISATION

- Nous avons pu établir un modèle de régression linéaire :
  - Le phénomène se distingue indépendament de l'arrondissement
  - La quasi totalité des prix de ventes est dictée par la surface du bien à laquelle s'ajoute un bruit (facteur externe)
- Les calculs de ce modèle permettent de faire des projections
- Attention aux facteurs sociaux ou matériels (travaux, voisinage, mobilier, etc.)

PERSPECTIVE DES TRAVAUX

- LIMITES

- BIEN est en accès libre restreint pour le public (prix/m², par arrondissement et par trimestre).

- 18 000 €/an, difficile à rentabiliser pour cette recherche.

PERSPECTIVES

Obtenir un dataset similaire à Castorus en fonction des quartiers parisiens pour :

- – mieux visualiser la répartition du prix des biens dans un arrondissement

- – affiner les modélisations à l'échelle des quartier.