Analyse de données avec Python

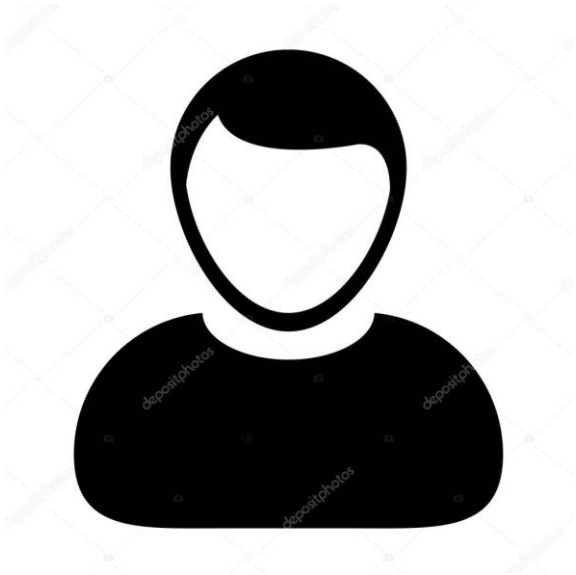
Pourquoi analyser les données

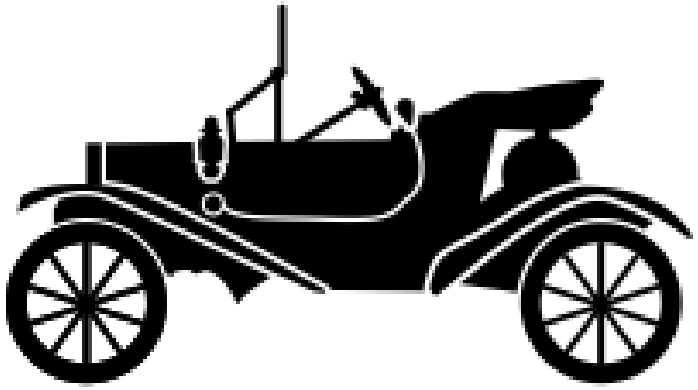
- Les données sont partout.

- L’analyse de données ou la data Science permet d’utiliser la donnée pour résoudre des problèmes.

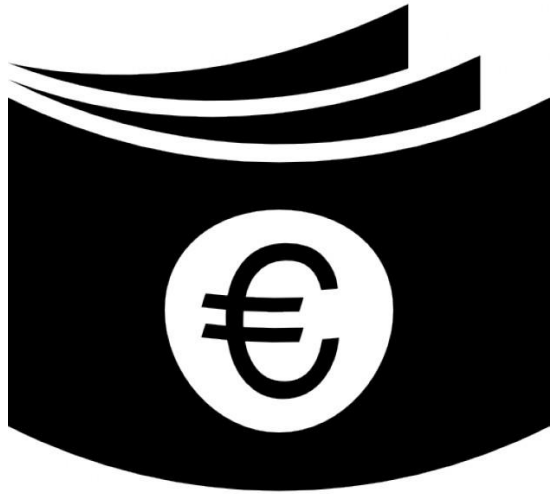
- L’analyse de données joue un rôle important dans :
  - La découverte de patterns cachés,
  - La réponse à certaines question,
  - La prédiction d’événement futurs

Scénario





?



Estimer le prix d’une voiture

- A-t-on accès à des données sur les prix des autres voiture et leurs caractéristiques ?
- Quelles caractéristiques peuvent influencer le prix ?
  - Couleur ? Marque ? Puissance ? Autre chose ?

Bernardo

Données car prices

https://archive.ics.uci.edu/ml/machine-learning-databases/autos/

192	bmw	essence	std	4	berline	t_ar	av	101.20	176.80	64.80	54.30	2395	ohc	4	108	mpfi	3.50	2.80	8.80	101	5800	23	29	16925
188	bmw	essence	std	2	berline	t_ar	av	101.20	176.80	64.80	54.30	2710	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	21	28	20970
188	bmw	essence	std	4	berline	t_ar	av	101.20	176.80	64.80	54.30	2765	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	21	28	21105
?	bmw	essence	std	4	berline	t_ar	av	103.50	189.00	66.90	55.70	3055	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	20	25	24565
?	bmw	essence	std	4	berline	t_ar	av	103.50	189.00	66.90	55.70	3230	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	16	22	30760
?	bmw	essence	std	2	berline	t_ar	av	103.50	193.80	67.90	53.70	3380	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	16	22	41315
?	bmw	essence	std	4	berline	t_ar	av	110.00	197.00	70.90	56.30	3505	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	15	20	36880
121	chevrolet	essence	std	2	coupe	t_av	av	88.40	141.10	60.30	53.20	1488	l	3	61	2bbl	2.91	3.03	9.50	48	5100	47	53	5151
98	chevrolet	essence	std	2	coupe	t_av	av	94.50	155.90	63.60	52.00	1874	ohc	4	90	2bbl	3.03	3.11	9.60	70	5400	38	43	6295
81	chevrolet	essence	std	4	berline	t_av	av	94.50	158.80	63.60	52.00	1909	ohc	4	90	2bbl	3.03	3.11	9.60	70	5400	38	43	6575
118	dodge	essence	std	2	coupe	t_av	av	93.70	157.30	63.80	50.80	1876	ohc	4	90	2bbl	2.97	3.23	9.41	68	5500	37	41	5572
118	dodge	essence	std	2	coupe	t_av	av	93.70	157.30	63.80	50.80	1876	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6377
118	dodge	essence	turbo	2	coupe	t_av	av	93.70	157.30	63.80	50.80	2128	ohc	4	98	mpfi	3.03	3.39	7.60	102	5500	24	30	7957
148	dodge	essence	std	4	coupe	t_av	av	93.70	157.30	63.80	50.60	1967	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6229
148	dodge	essence	std	4	berline	t_av	av	93.70	157.30	63.80	50.60	1989	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6692
148	dodge	essence	std	4	berline	t_av	av	93.70	157.30	63.80	50.60	1989	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	7609
148	dodge	essence	turbo	?	berline	t_av	av	93.70	157.30	63.80	50.60	2191	ohc	4	98	mpfi	3.03	3.39	7.60	102	5500	24	30	8558
110	dodge	essence	std	4	monospace	t_av	av	103.30	174.60	64.60	59.80	2535	ohc	4	122	2bbl	3.34	3.46	8.50	88	5000	24	30	8921
145	dodge	essence	turbo	2	coupe	t_av	av	95.90	173.20	66.30	50.20	2811	ohc	4	156	mfi	3.60	3.90	7.00	145	5000	19	24	12964
137	honda	essence	std	2	coupe	t_av	av	86.60	144.60	63.90	50.80	1713	ohc	4	92	1bbl	2.91	3.41	9.60	58	4800	49	54	6479
137	honda	essence	std	2	coupe	t_av	av	86.60	144.60	63.90	50.80	1819	ohc	4	92	1bbl	2.91	3.41	9.20	76	6000	31	38	6855
101	honda	essence	std	2	coupe	t_av	av	93.70	150.00	64.00	52.60	1837	ohc	4	79	1bbl	2.91	3.07	10.10	60	5500	38	42	5399

Comprendre les données

Le jeu de données : Voitures d'occasion

192	bmw	essence	std	4	berline	t_ar	av	101.20	176.80	64.80	54.30	2395	ohc	4	108	mpfi	3.50	2.80	8.80	101	5800	23	29	16925
188	bmw	essence	std	2	berline	t_ar	av	101.20	176.80	64.80	54.30	2710	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	21	28	20970
188	bmw	essence	std	4	berline	t_ar	av	101.20	176.80	64.80	54.30	2765	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	21	28	21105
?	bmw	essence	std	4	berline	t_ar	av	103.50	189.00	66.90	55.70	3055	ohc	6	164	mpfi	3.31	3.19	9.00	121	4250	20	25	24565
?	bmw	essence	std	4	berline	t_ar	av	103.50	189.00	66.90	55.70	3230	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	16	22	30760
?	bmw	essence	std	2	berline	t_ar	av	103.50	193.80	67.90	53.70	3380	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	16	22	41315
?	bmw	essence	std	4	berline	t_ar	av	110.00	197.00	70.90	56.30	3505	ohc	6	209	mpfi	3.62	3.39	8.00	182	5400	15	20	36880
121	chevrolet	essence	std	2	coupe	t_av	av	88.40	141.10	60.30	53.20	1488	l	3	61	2bbl	2.91	3.03	9.50	48	5100	47	53	5151
98	chevrolet	essence	std	2	coupe	t_av	av	94.50	155.90	63.60	52.00	1874	ohc	4	90	2bbl	3.03	3.11	9.60	70	5400	38	43	6295
81	chevrolet	essence	std	4	berline	t_av	av	94.50	158.80	63.60	52.00	1909	ohc	4	90	2bbl	3.03	3.11	9.60	70	5400	38	43	6575
118	dodge	essence	std	2	coupe	t_av	av	93.70	157.30	63.80	50.80	1876	ohc	4	90	2bbl	2.97	3.23	9.41	68	5500	37	41	5572
118	dodge	essence	std	2	coupe	t_av	av	93.70	157.30	63.80	50.80	1876	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6377
118	dodge	essence	turbo	2	coupe	t_av	av	93.70	157.30	63.80	50.80	2128	ohc	4	98	mpfi	3.03	3.39	7.60	102	5500	24	30	7957
148	dodge	essence	std	4	coupe	t_av	av	93.70	157.30	63.80	50.60	1967	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6229
148	dodge	essence	std	4	berline	t_av	av	93.70	157.30	63.80	50.60	1989	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	6692
148	dodge	essence	std	4	berline	t_av	av	93.70	157.30	63.80	50.60	1989	ohc	4	90	2bbl	2.97	3.23	9.40	68	5500	31	38	7609
148	dodge	essence	turbo	?	berline	t_av	av	93.70	157.30	63.80	50.60	2191	ohc	4	98	mpfi	3.03	3.39	7.60	102	5500	24	30	8558
110	dodge	essence	std	4	monospace	t_av	av	103.30	174.60	64.60	59.80	2535	ohc	4	122	2bbl	3.34	3.46	8.50	88	5000	24	30	8921
145	dodge	essence	turbo	2	coupe	t_av	av	95.90	173.20	66.30	50.20	2811	ohc	4	156	mfi	3.60	3.90	7.00	145	5000	19	24	12964
137	honda	essence	std	2	coupe	t_av	av	86.60	144.60	63.90	50.80	1713	ohc	4	92	1bbl	2.91	3.41	9.60	58	4800	49	54	6479
137	honda	essence	std	2	coupe	t_av	av	86.60	144.60	63.90	50.80	1819	ohc	4	92	1bbl	2.91	3.41	9.20	76	6000	31	38	6855
101	honda	essence	std	2	coupe	t_av	av	93.70	150.00	64.00	52.60	1837	ohc	4	79	1bbl	2.91	3.07	10.10	60	5500	38	42	5399

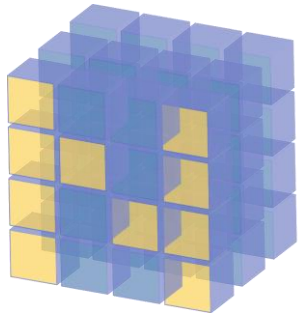
Description des caractéristiques.

- 205 lignes, 26 colonnes

Symboling	-3, -2, -1, 1, 2, 3	**poids**	Continu de 1488 à 4066
Pertes-normalisées	Continu de 65 à 256	Type_moteur	Dohc, dohcv, …
Marque	Alfa romeo …	Nb_cylindres	2, 3, 4, 5,6,8,12
Carburant	Essence, diesel	Taille_moteur	Continu de 61 à 326
Aspiration	Std, turbo	Alimentation	1bbl, 2bbl, 4bbl, idi, mfi
Nb_portes	4, 2	Alesage	Continu de 2.54 à 3.94
Carrosserie	Berline, coupe, …	Course	Continu de 2.07 à 4.17
Motricité	4wd, t_av ou t_ar	Tx_compression	Continu de 7 à 23
Place_moteur	Avant, arriere	Puissance	Continu de 48 à 288
Roues	Continu de 86.6 à 120.9	Reg_pointe	Continu de 4150 à 6600
Longueur	Continu de 141.1 à 208.1	Conso_urb	Continu de 13 à 49
Largeur	Continu de 60.3 à 72.3	Conso_extra_urb	Continu de 16 à 54
hauteur	Continu de 47.8 à 59.8	prix	Continu de 5118 à 45400

Les librairies scientifiques de Python

Pandas

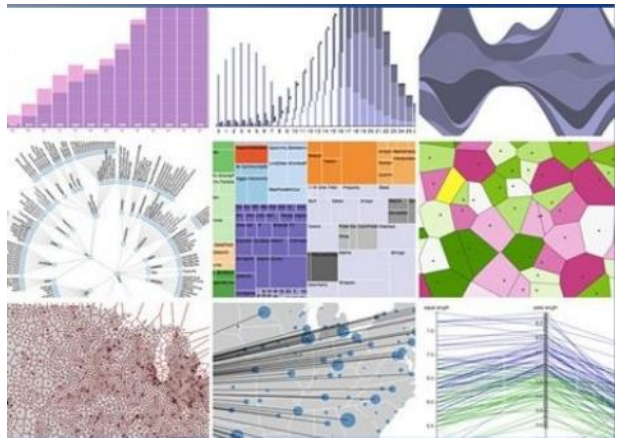


Numpy

Scipy

Les librairies graphiques de Python

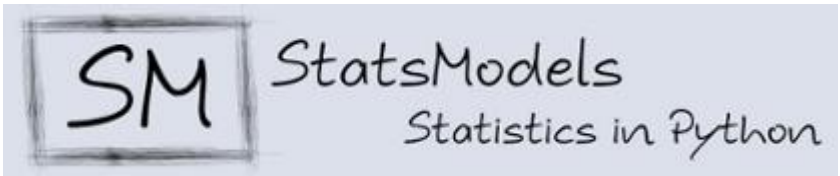
Matplotlib



Seaborn

Les librairies algorithmique de Python

Scikit-learn



Statsmodels

Objectifs

- Comprendre le jeu de données
- Nettoyer les données en utilisant les méthodes pandas
- Analyser l’influence des caractéristiques sur la cible
- Tester les différentes correlations
- Découper le jeu de données (20/80) en effectuant un échantillonnage judicieux
- Entrainer un modèle de régression linéaire et l’évaluer.