



PRÉSENTATION DE PROJET

NEW CAR

PIERRE-ALEXIS LEBAIR



LA PLATEFORME

NEW CAR



SOMMAIRE



- INTRODUCTION
- EXPLORATION DES DONNÉES
- CORRÉLATION LINÉAIRE
- RÉGRESSION LINÉAIRE
- CONCLUSION



NEW CAR

INTRODUCTION



Xxx €



Contexte

Pouvoir prédire le prix d'une voiture en fonction de caractéristiques données

11. Vous souhaitez acheter un véhicule ayant moins de 7 ans, pas plus de 100 000 km, avec une boîte manuelle, **quel devrait-être son prix ?**

Enjeux

- Charger les données
- Analyser les données
- Identifier des tendances
- Réaliser une régression linéaire
- Analyser les résultats



NEW CAR

EXPLORATION DES DONNÉES

Affichage du dataframe

Affichage des données du dataframe pour une première analyse

	A	B	C	D	E	F	G	H	I
1	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brezza	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
13	ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
14	ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
15	ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0
16	dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual	0

Analyse statistique

Identification de potentielles erreurs
Premières remarques sur la qualité du dataset

	Year	Selling_Price	Present_Price	Kms_Driven	Owner
count	301.000000	301.000000	301.000000	301.000000	301.000000
mean	2013.627907	4.661296	7.628472	36947.205980	0.043189
std	2.891554	5.082812	8.644115	38886.883882	0.247915
min	2003.000000	0.100000	0.320000	500.000000	0.000000
25%	2012.000000	0.900000	1.200000	15000.000000	0.000000
50%	2014.000000	3.600000	6.400000	32000.000000	0.000000
75%	2016.000000	6.000000	9.900000	48767.000000	0.000000
max	2018.000000	35.000000	92.600000	500000.000000	3.000000

NEW CAR

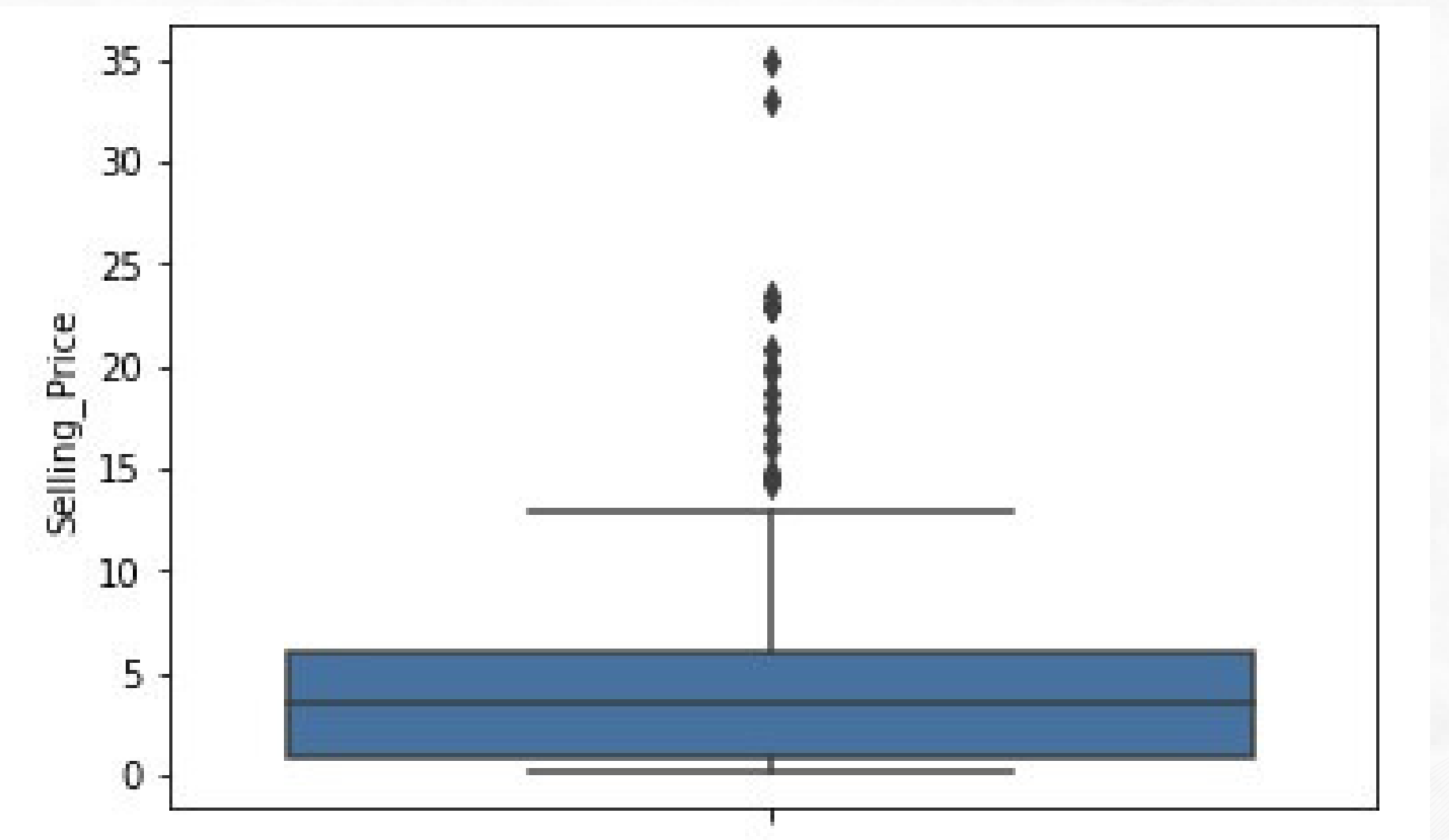
ÉXPLORATION DES DONNÉES

Affichage du dataframe

Affichage des données du dataframe
pour une première analyse

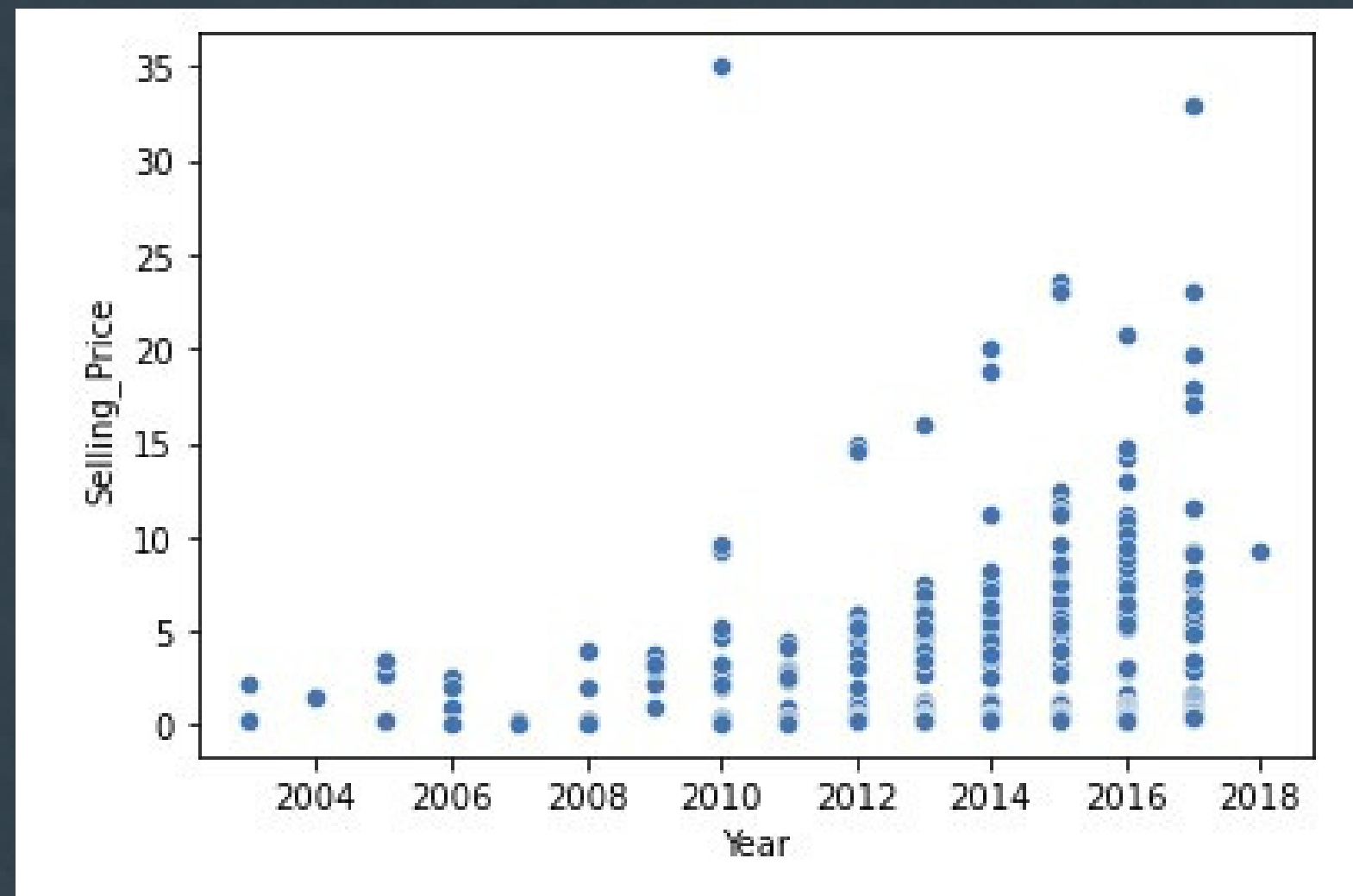
Analyse statistique

Identification de potentielles erreurs
Premières remarques sur la qualité du
dataset



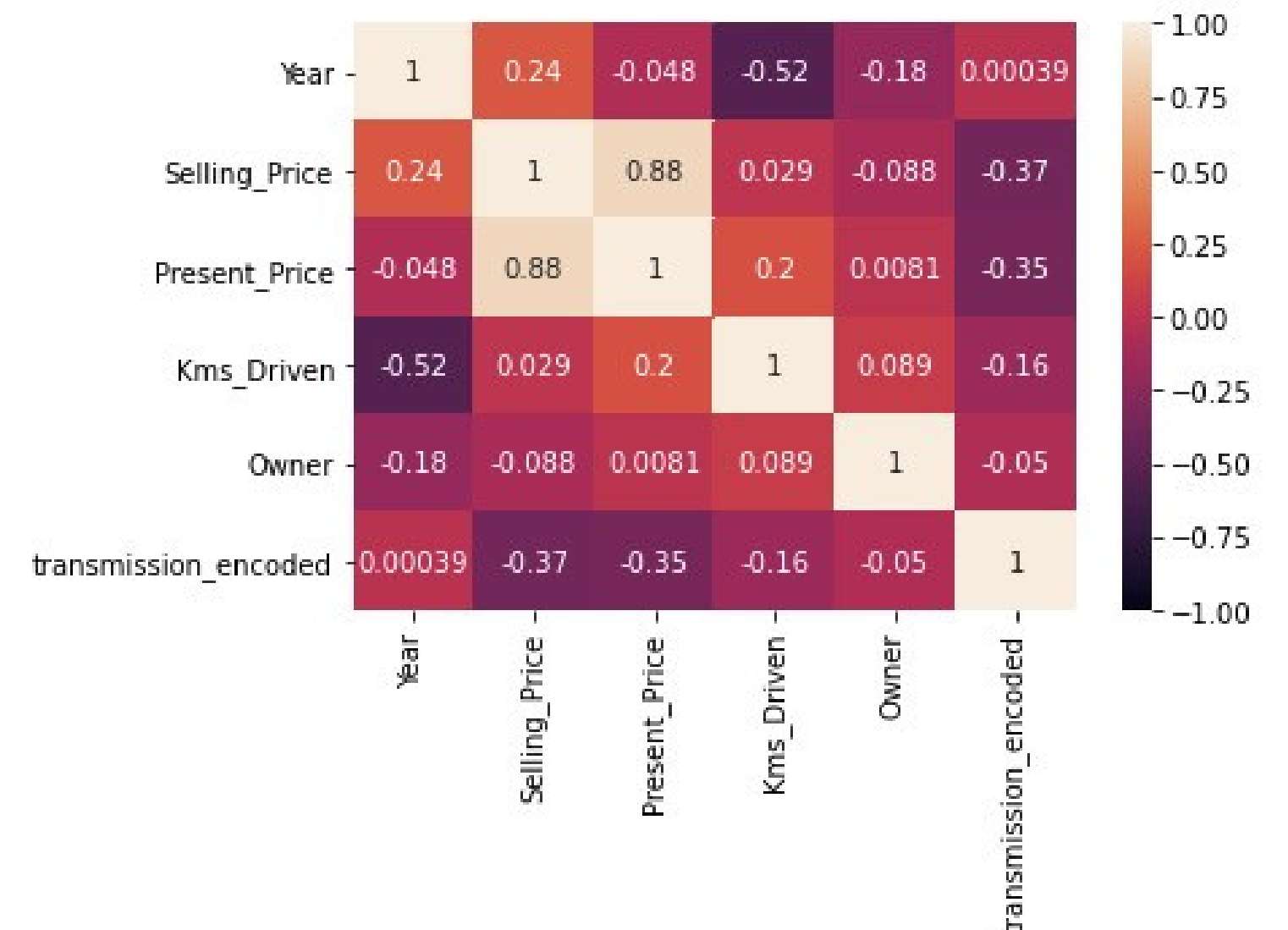
NEW CAR

CORRÉLATION LINÉAIRE



Identification de
tendances

Corrélation des Features

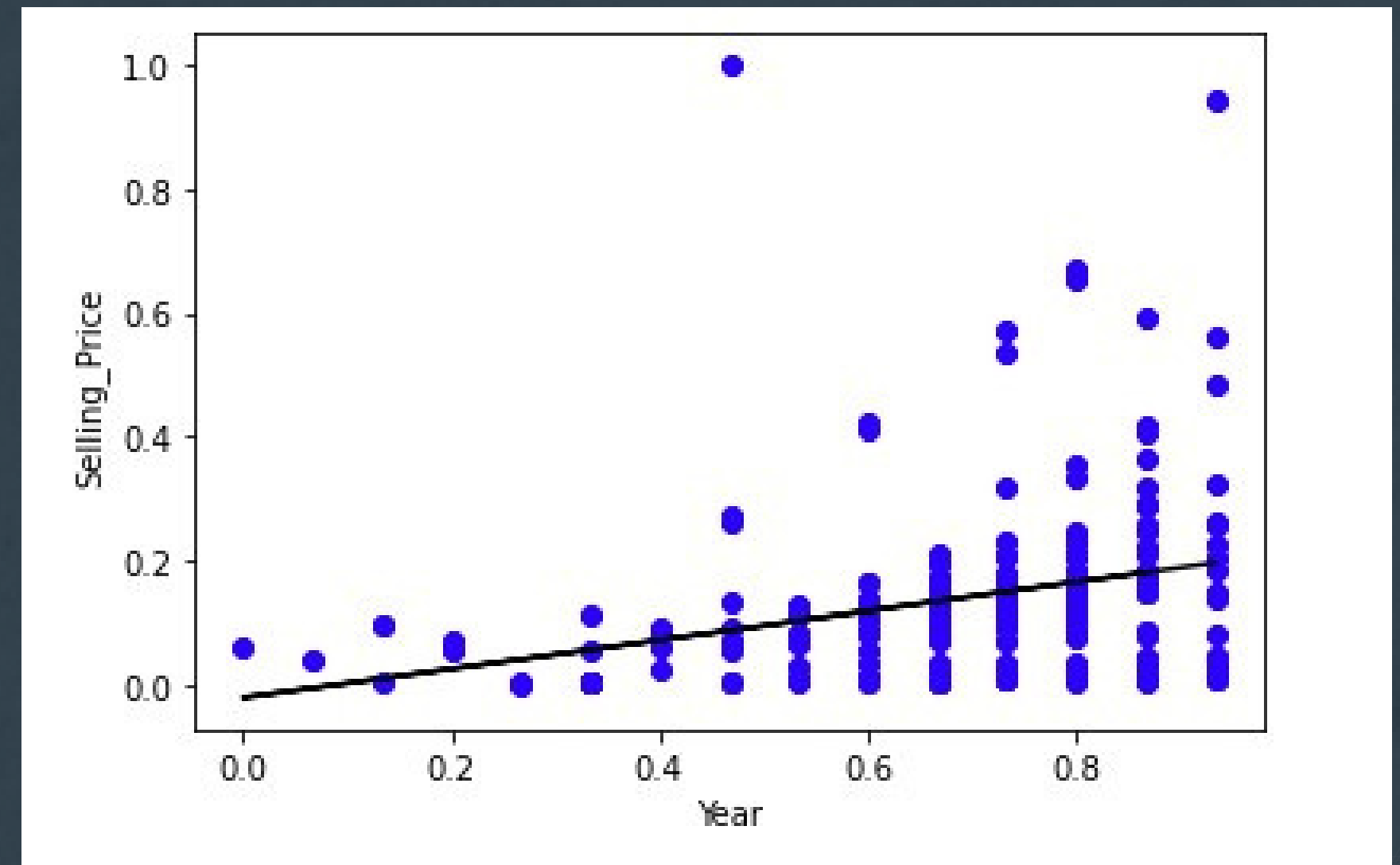


NEW CAR

CORRÉLATION LINÉAIRE

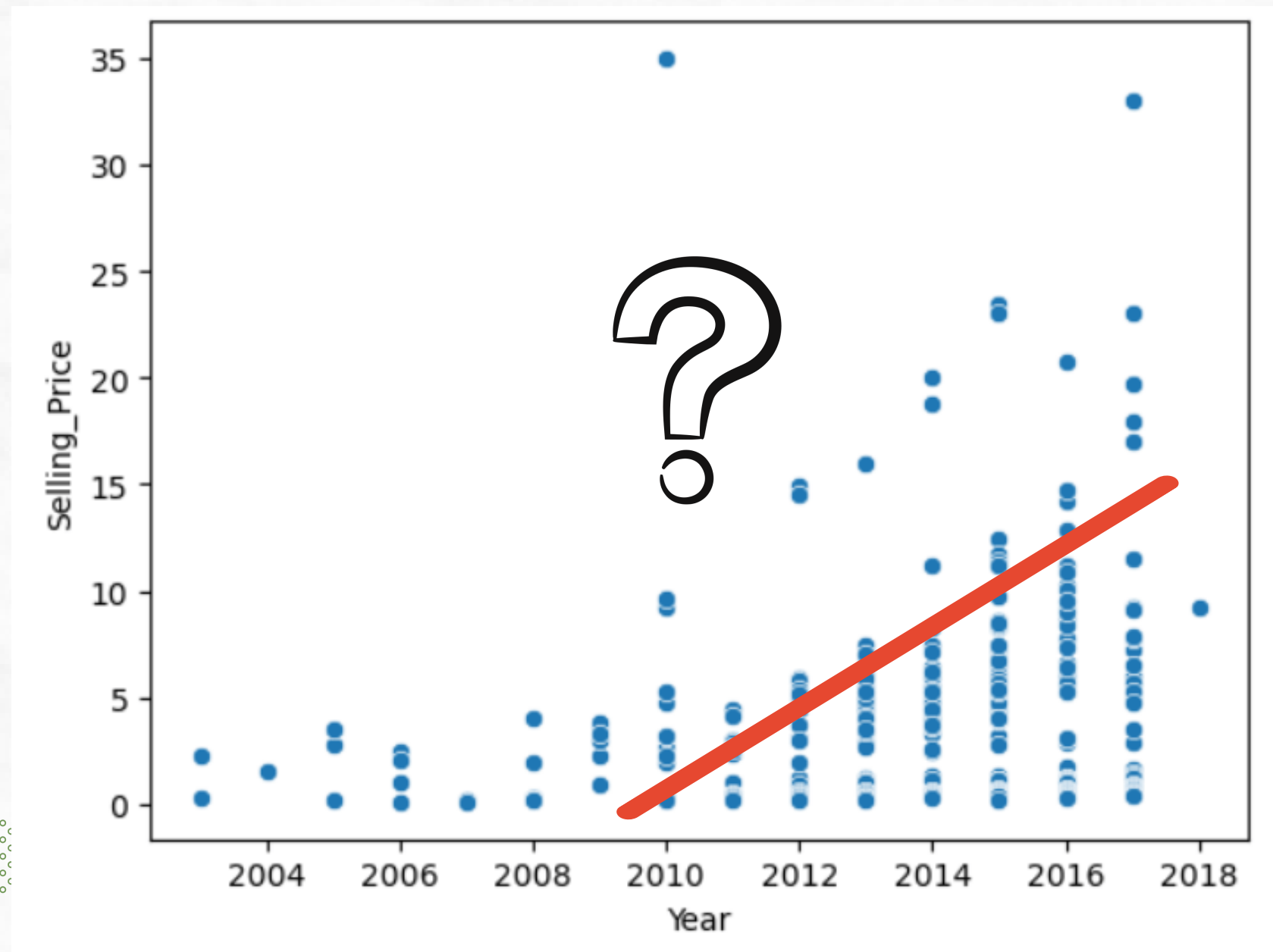
Modèle de Pearson

L'un des modèle les plus utilisé et efficace pour calculer un coefficient de corrélation



NEW CAR

RÉGRESSION LINÉAIRE

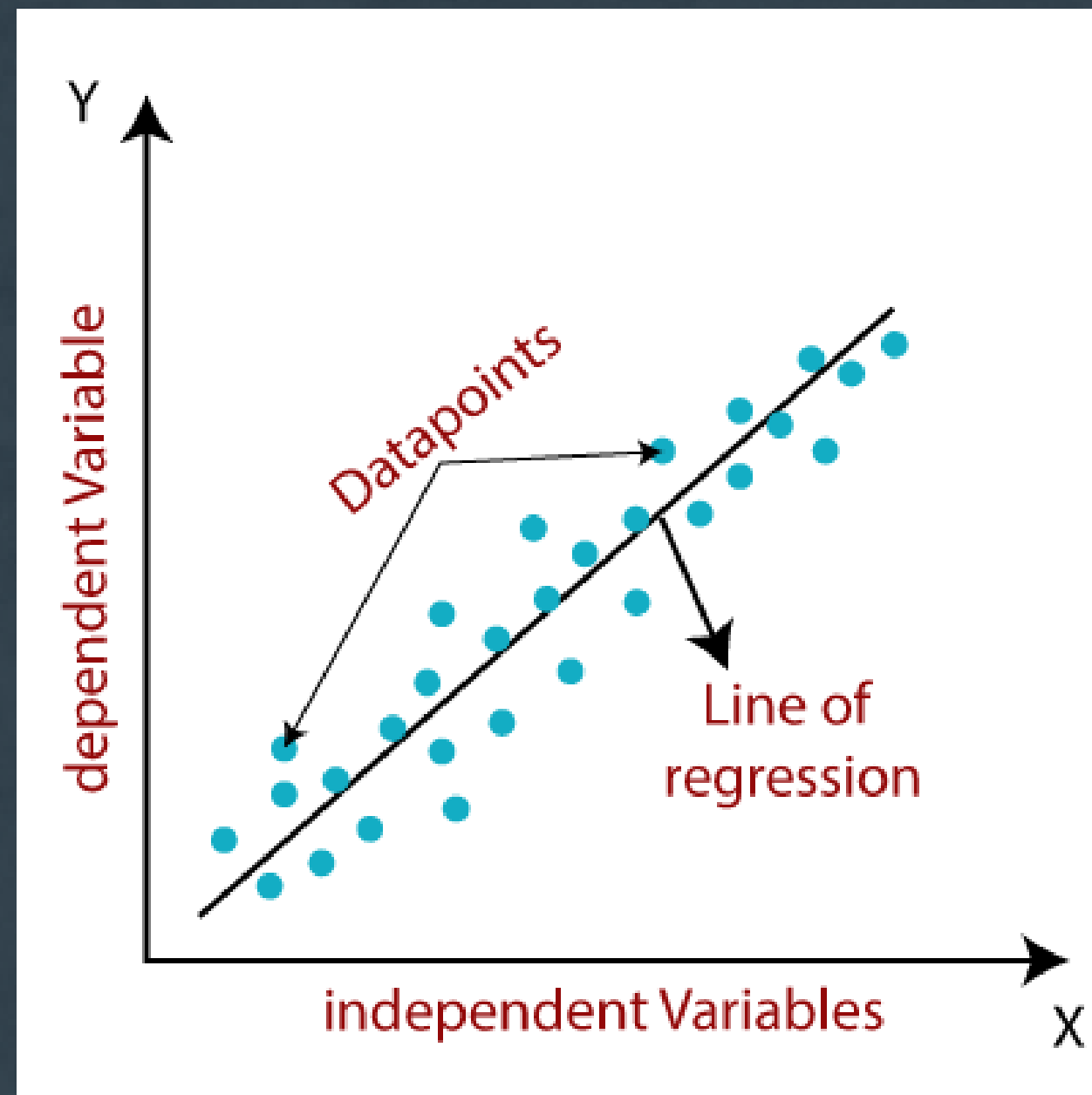


Tendances linéaires ?

Une tendance linéaire peut indiquer qu'une régression linéaire peut être utilisée à des fins de prédiction

NEW CAR

RÉGRESSION LINÉAIRE ?



Qu'est-ce que c'est ?

C'est une droite définie par une équation du type $\beta_0 + \beta_1 X + \varepsilon$ avec :

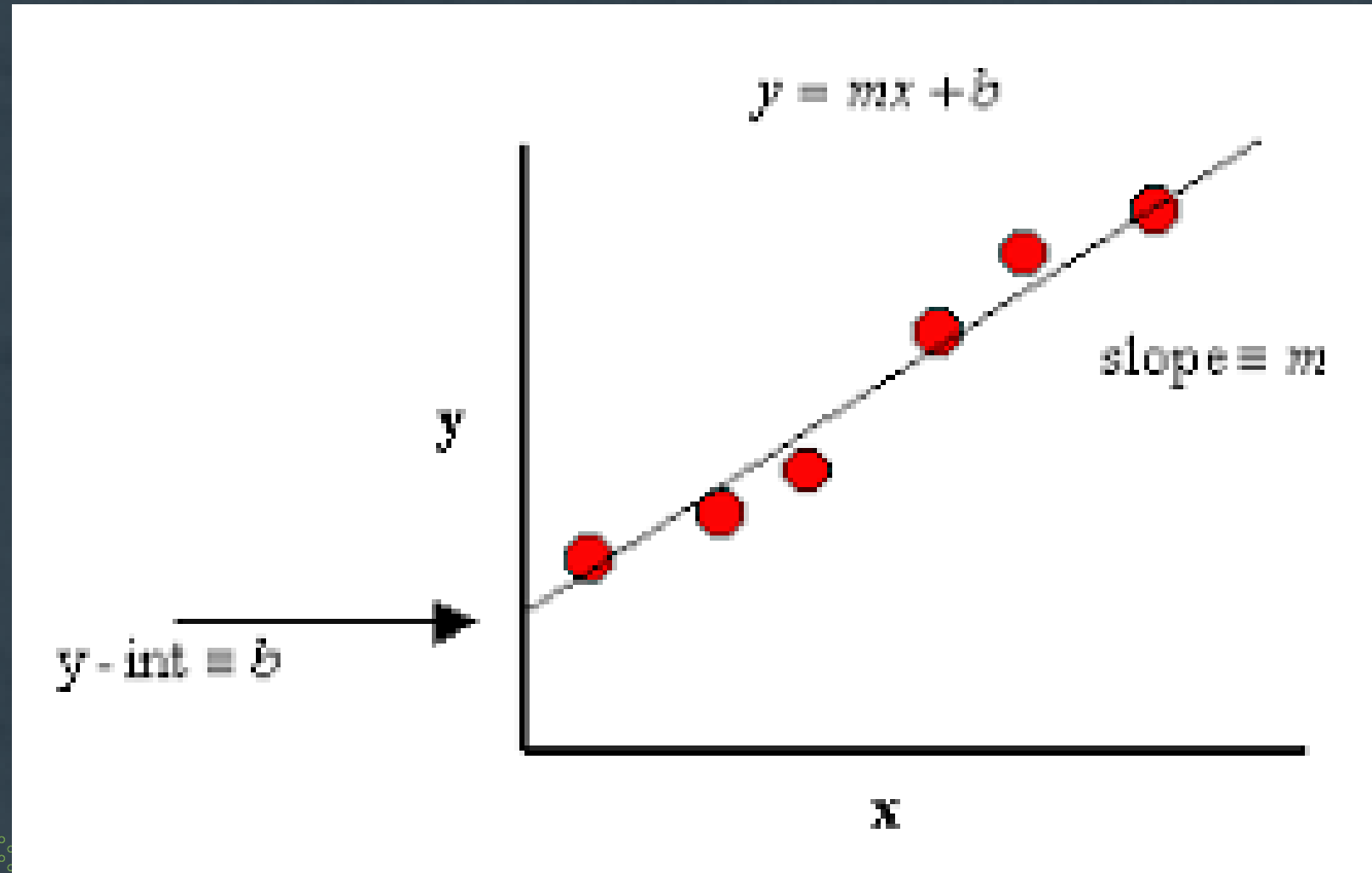
- β_0 -> l'intercept
- β_1 -> le slope
- ε -> l'erreur

A quoi ça sert ?

- Etablir une relation entre deux variables du dataset
- Utiliser l'équation de la droite ainsi obtenue pour faire de la prédiction

NEW CAR

REGRESSION LINÉAIRE ?



Comment ça marche ?

- Tests de coefficients β_0 et β_1
- Calculer l'erreur entre prédiction et les données d'entraînement à l'aide d'une fonction de coût
- Réajuster les coefficients β_0 et β_1 avec un algorithme de minimisation

NEW CAR

NORMALISATION ET ENCODAGE

Normalisation

Réduire l'impact de trop grands écarts dans les valeurs de certaines features

	Year	Selling_Price	Present_Price	Kms_Driven	Owner
count	301.000000	301.000000	301.000000	301.000000	301.000000
mean	2013.627907	4.661296	7.628472	36947.205980	0.043189
std	2.891554	5.082812	8.644115	38886.883882	0.247915
min	2003.000000	0.100000	0.320000	500.000000	0.000000
25%	2012.000000	0.900000	1.200000	15000.000000	0.000000
50%	2014.000000	3.600000	6.400000	32000.000000	0.000000
75%	2016.000000	6.000000	9.900000	48767.000000	0.000000
max	2018.000000	35.000000	92.600000	500000.000000	3.000000

Encodage

Encoder certaines features qualitatives ayant potentiellement un impact sur la target

```
      Seller_Type  Transmission  Owner
0         Dealer         Manual      0
1         Dealer         Manual      0
2         Dealer         Manual      0
3         Dealer         Manual      0
4         Dealer         Manual      0
..          ...             ...     ...
296        Dealer         Manual      0
297        Dealer         Manual      0
298        Dealer         Manual      0
299        Dealer         Manual      0
300        Dealer         Manual      0
```

[301 rows x 9 columns]

NEW CAR

ANALYSE DES RESULTATS



UNIVARIEE

R SCORE = 1.9%

- Score de régression indiquant une faible relation entre feature et target



- Données trop dispersées

MULTIVARIEE

R SCORE = 16,1%

- Score de régression indiquant toujours une faible relation features/target
- Résultats inexploitable pour faire des prédictions

NEW CAR

METRICS DE QUALITÉ

	Year	Selling_Price	Present_Price	Kms_Driven	Owner	transmission_encoded
count	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000
mean	0.708527	0.130696	0.079199	0.072967	0.043189	0.867110
std	0.192770	0.145639	0.093673	0.077852	0.247915	0.340021
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.600000	0.022923	0.009536	0.029029	0.000000	1.000000
50%	0.733333	0.100287	0.065886	0.063063	0.000000	1.000000
75%	0.866667	0.169054	0.103814	0.096631	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	3.000000	1.000000

16%

R^2

Score de détermination

0.09

MAE

Erreur absolue moyenne

0.13

RMSE

Erreur quadratique
moyenne



NEW CAR



CONCLUSIONS



RESULTATS NE PERMETTANT PAS DE PRÉDICTION PRÉCISE

Les indicateurs de qualité de notre régression indiquent que les résultats sont peu susceptibles de générer une prédiction précise pour le problème posé à la fin du sujet

Plusieurs facteurs peuvent en être la cause :

- Données trop générales
- Facteurs externes non pris en compte
- Jeu de donnée trop petit
- ...



NEW CAR

MERCI

POUR VOTRE ATTENTION

