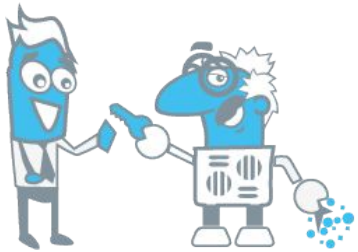
# ANALYSE EXPLORATOIRE DES DONNÉES

## Data preprocessing



HUPI
VALORISER ANALYTIQUEMENT VOTRE DIFFÉRENCE

bai euskarari elkartea
LAN ETA BIZI EUSKARAZ

# INDEX

1 — 2 — 3 — 4

1. Statistiques descriptives
2. Analyse univariée
3. Analyse bivariée
4. Tests de corrélation

# 01 STATISTIQUES DESCRIPTIVES

1

# « Pourquoi explorer les données avant toute modélisation ? »

# 1 Objectif

**Résumer les caractéristiques principales d'une variable ou d'un jeu de données ;**

**Identifier les anomalies ;**

**Visualiser les relations entre les variables ;**

**Orienter le choix des transformations**

2

# « Connaissez-vous des indicateurs permettant d'atteindre ces objectifs ? »

# 2 Indicateurs principaux

**Exemples d'indicateurs (métriques principales) :**

- Moyenne, médiane, mode
- Écart-type, variance
- Min, max

Pour aller plus loin :

- Quartiles (pour identifier les valeurs aberrantes)
- Distribution des données (histogrammes)

# 02 ANALYSE UNIVARIÉE

1

« Quel est l'objectif principale de l'analyse univariée ? »

# 1 Objectif

**Étudier chaque variable individuellement.**

2

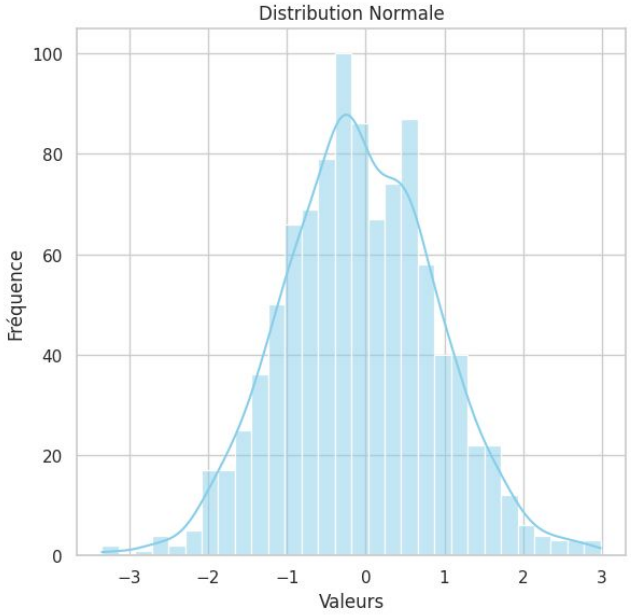
# « Quelle est la meilleure façon de détecter des outliers dans une variable quantitative ? »

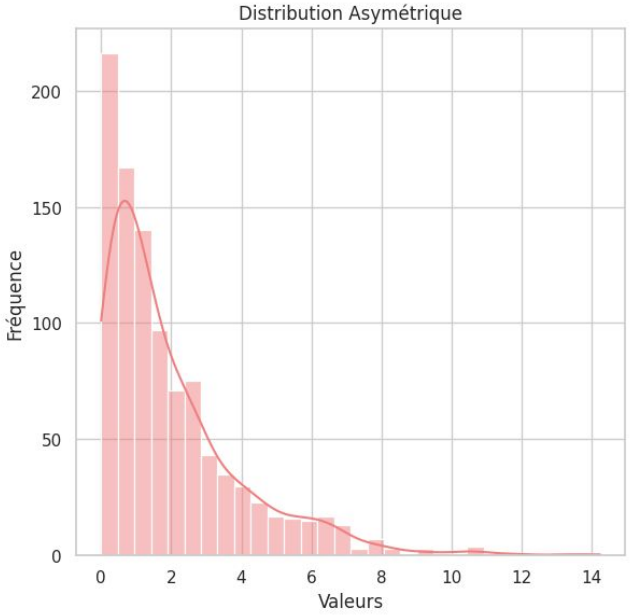
# 2 Méthodes principales

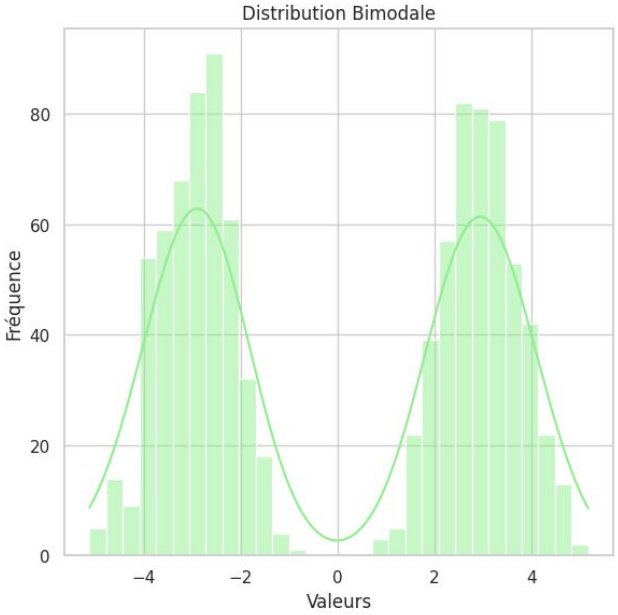
**Outils :**

- Countplots (pour analyser les variables catégorielles)
- Histogrammes
- Boxplots (pour visualiser les valeurs aberrantes)
- Recherche des outliers :
  - Méthode IQR (Écart interquartile)
  - Méthode Z-score

# 3 Types de distributions principaux



# 03 ANALYSE BIVARIÉE

1

# « Quel est l'objectif principale de l'analyse bivariée ? »

# 1 Objectif

**Étudier les relations entre deux variables.**

2

# « Quelles méthodes utilisées ? »

# 2 Méthodes principales

**Exemples d'outils :**

- Scatterplots (pour les variables continues)
- Boxplots (Boxplots pour une variable catégorielle vs quantitative)
- Heatmaps pour visualiser des corrélations

# 04 TESTS DE CORRÉLATION

1

« Qu'est-ce que la corrélation de Spearman ? En quoi est-elle différente de Pearson ? »

# 1 Objectif et définition

**Identifier les relations linéaires ou monotones entre variables.**

- Corrélation de Pearson : Relations linéaires
- Corrélation de Spearman : Relations monotones

**Outil pour la visualisation :** Matrice de corrélation avec des heatmaps.

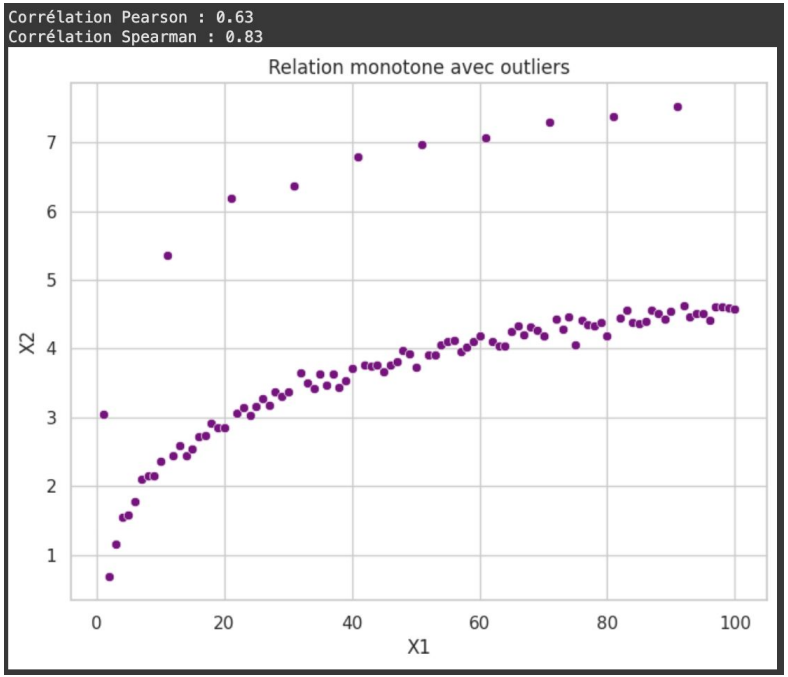
2

« Dans quel cas je peux avoir une corrélation de Spearman plus élevée que de Pearson ? »

# 2 Simulation d'un exemple

**Exemple :**

- Générer des données monotones non linéaires à l'aide d'une relation logarithmique
- Ajouter des outliers
- Calculer les corrélations de Pearson et de Spearman



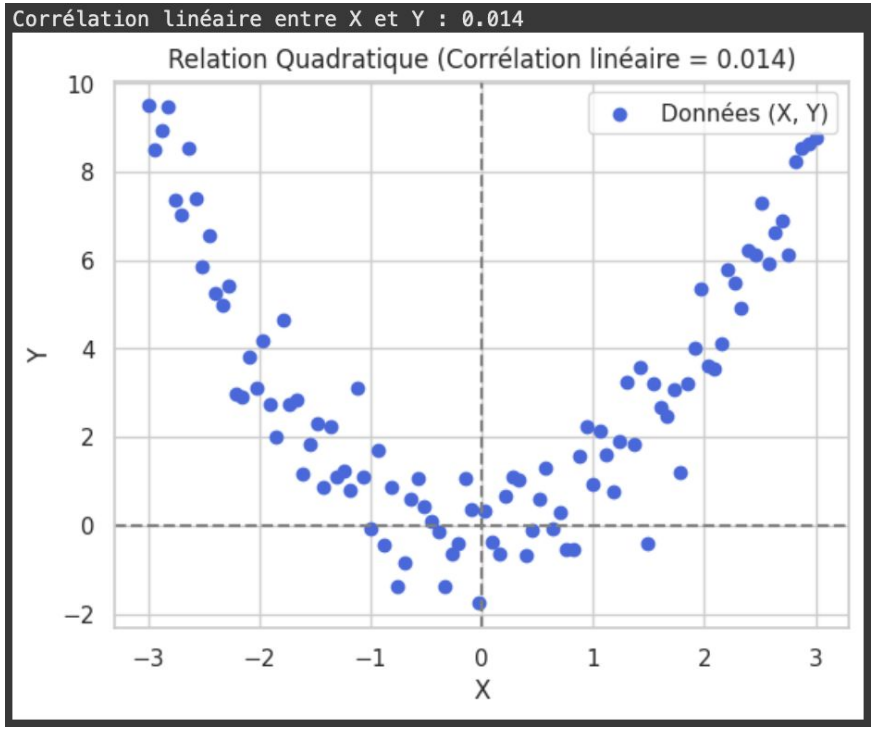
3

« Si deux variables ont une corrélation proche de zéro, cela signifie-t-il qu'elles ne sont pas liées ? Pourquoi ? »

# 3 Simulation d'un exemple

**Exemple :**

- Générer des données avec une relation quadratique avec un bruit gaussien
- Calculer la corrélation linéaire (Pearson)
- Visualisation : interpréter visuellement la relation quadratique existante



# CONCLUSIONS

## Bonnes pratiques :

- Nettoyer les données avant d'analyser
(valeurs manquantes, doublons).
- Toujours explorer les données avant de modéliser.
- Utiliser des visualisations adaptées pour comprendre les relations.

| Objectif | Type de graphique |
|---|---|
| Distribution d'une variable | Histogramme, Boxplot |
| Comparaison de catégories | Barplot, Countplot |
| Relation entre deux variables | Scatterplot, Regression |
| Corrélation entre variables | Heatmap |

- Réflexion sur l’interprétabilité des résultats.
Un graphique doit parler !

# AIDE AU CODAGE

- pandas.read_csv()
- describe()

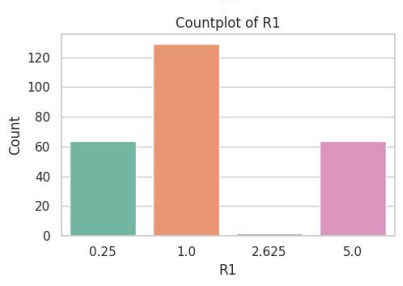
- plt.figure()
- plt.subplots()

- seaborn.countplot()
- seaborn.barplot()
- seaborn.histplot()
- seaborn.scatterplot()
- seaborn.PairGrid()
- seaborn.boxplot()

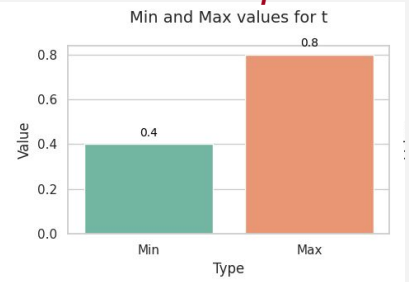
- corr()
- seaborn.heatmap()

## ANALYSE UNIVARIÉE

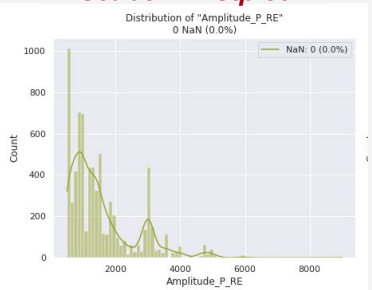
***seaborn.countplot***




***seaborn.barplot***

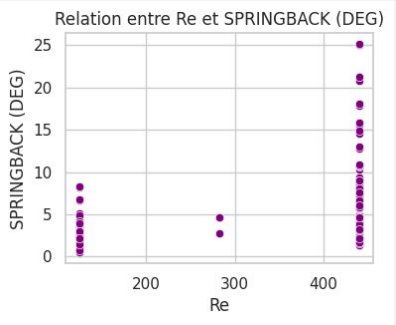



***seaborn.histplot***

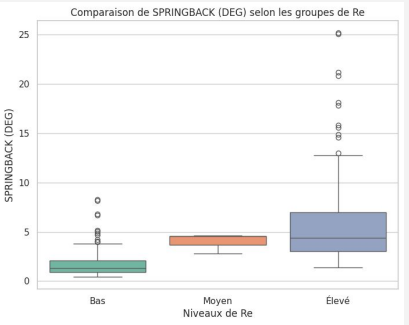



## ANALYSE BIVARIÉE

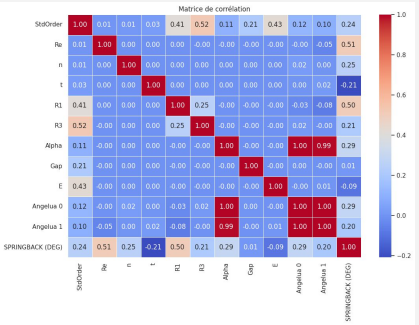
***seaborn.scatterplot***




***seaborn.boxplot***




***seaborn.heatmap***



# AIDE AU CODAGE

- pandas.read_csv()
- describe()

- plt.figure()
- plt.subplots()

- seaborn.countplot()
- seaborn.barplot()
- seaborn.histplot()
- seaborn.scatterplot()
- seaborn.PairGrid()
- seaborn.boxplot()

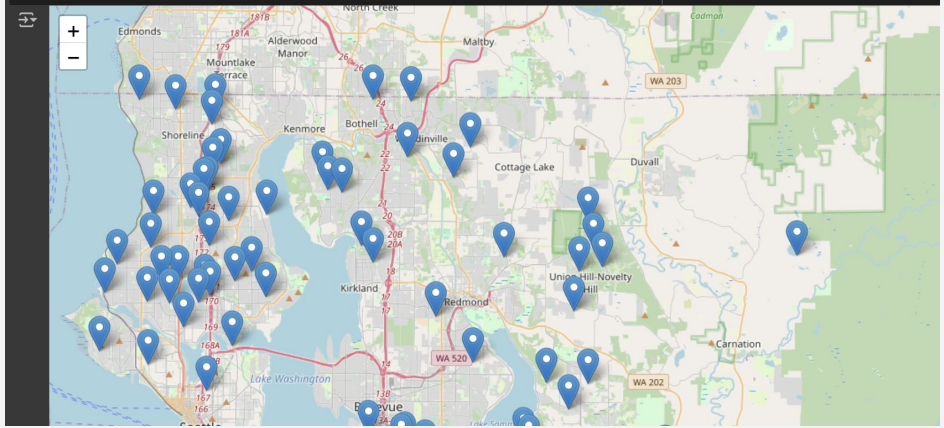
- corr()
- seaborn.heatmap()

## ANALYSE DONNÉES GÉOLOCALISÉES

***folium.Map***

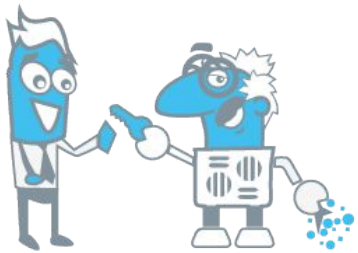
Here I just use it to show the localisation of the houses, but remember that it can be used for several purpose way useful than this one.

```
[ ]   import folium
      map = folium.Map(location=[df['lat'].mean(), df['long'].mean()], zoom_start=7)

      # Select 100 random rows from the DataFrame
      sample_df = df.sample(n=100, random_state=1)

      # Add markers to the map for each row in your DataFrame
      for index, row in sample_df.iterrows():
          folium.Marker([row['lat'], row['long']]).add_to(map)

      # Display the map
      map
```



# FEATURES ENGINEERING

Creating, transforming and selecting features



HUPI
VALORISER ANALYTIQUEMENT VOTRE DIFFÉRENCE

bai euskarari elkartea LAN ETA BIZI EUSKARAZ

# INDEX

1

2

## Introduction

1. Qu'est-ce que l'ingénierie des fonctionnalités ?

## Techniques courantes

1. Traitement des données manquantes
2. Codage des variables catégorielles
3. Mise à l'échelle et normalisation
4. Création de caractéristiques

# 01 INTRODUCTION

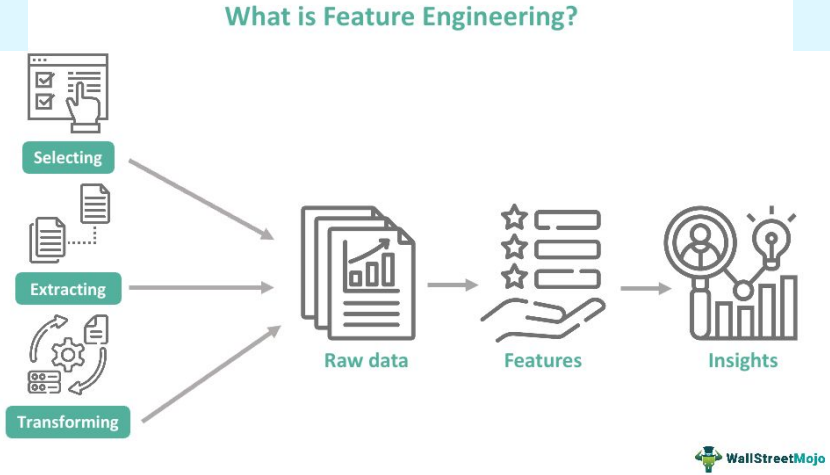
1

# « Qu'est-ce que Feature Engineering ? »

# 1 Importance de la création de caractéristiques

*“Feature engineering et sa **pertinence** dans la science des données »*

*« Importance de l'ingénierie des caractéristiques et son **impact** sur la **performance des modèles** »*




« L'ingénierie des caractéristiques est le processus de création de nouvelles caractéristiques significatives à partir de données existantes afin d'améliorer les modèles d'apprentissage automatique »

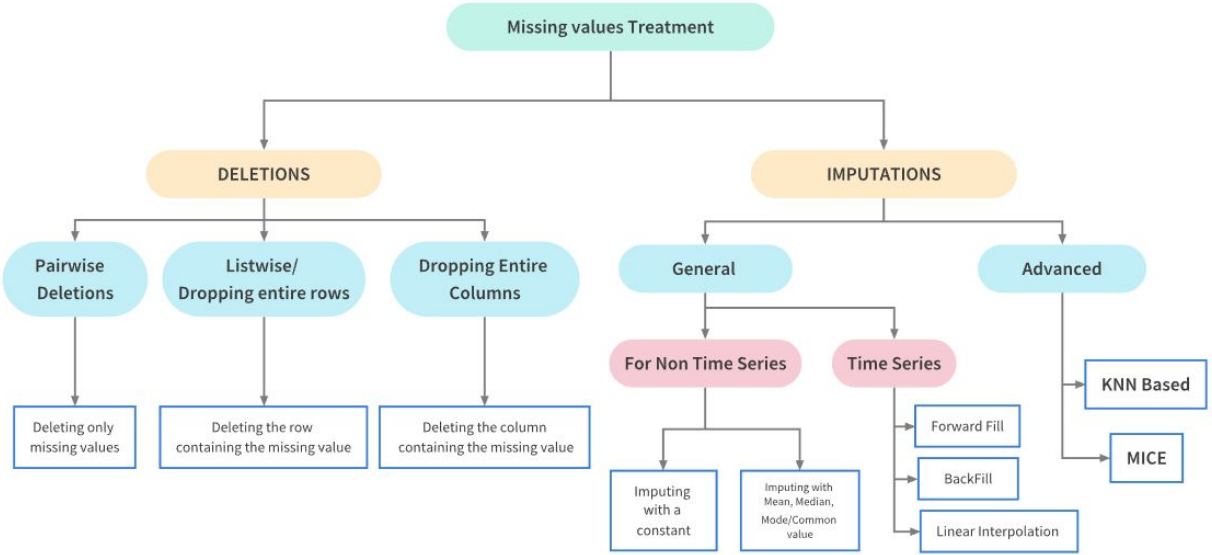
# 02 TECHNIQUES COURANTES

2

# « Qu'est-ce que la Feature Engineering ? »

# 1 Traitement des données manquantes (Handling missing data)

*« Traitement des données manquantes et **techniques** pour résoudre ce problème ».*

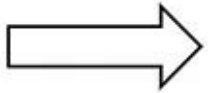


# 2 Encodage de variables catégorielles (Encoding categorical variables)

*« Encodage des variables catégorielles et méthodes pour y parvenir »*

**One-Hot Encoding**

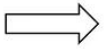
| Places |
|---|
| New York |
| Boston |
| Chicago |
| California |
| New Jersey |



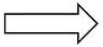
| New York | Boston | Chicago | California | New Jersey |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |

**Label Encoding**

| Places |
|---|
| New York |
| Boston |
| Chicago |
| California |
| New Jersey |



| Places | Map |
|---|---|
| New York | 1 |
| Boston | 2 |
| Chicago | 3 |
| California | 4 |
| New Jersey | 5 |

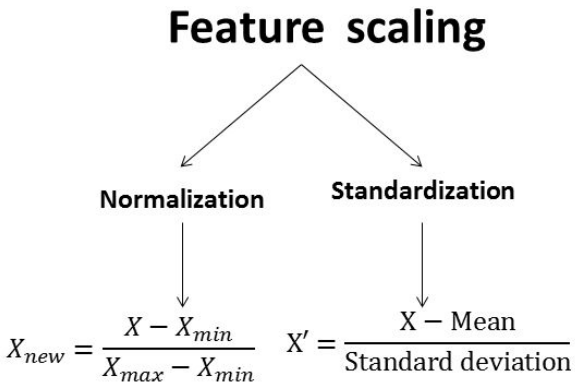


| Places | Encoded |
|---|---|
| New York | 1 |
| Boston | 2 |
| New York | 1 |
| California | 4 |
| Boston | 2 |

- **Dummy Encoding**
- **Label Encoding**
- **Ordinal Encoding**
- **Binary Encoding**
- **Count Encoding**

# 3 Mise à l'échelle et normalisation (Scaling and Normalization)

*« La nécessité d'une mise à l'échelle et d'une normalisation »*




- **“Standardization”**: vous modifiez l'étendue de vos données.
- **“Normalization”** : vous modifiez la forme de la distribution de vos données.

**Avantages de la “Standardization” :**

- Amélioration des performances du modèle : En particulier ceux qui s'appuient sur des mesures de distance ou des gradients pour l'optimisation. Elle garantit que les caractéristiques à grande échelle ne dominent pas le processus d'apprentissage.

**Avantages de la “Normalization” :**

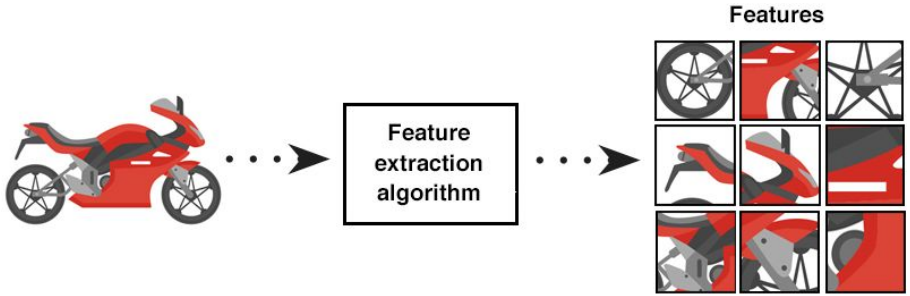
- Influence égale des caractéristiques : Toutes les caractéristiques ont la même influence ou le même poids dans les analyses et les modèles. Cette caractéristique est particulièrement utile lorsque l'on compare des caractéristiques à des échelles différentes, car elle permet d'éviter que l'une d'entre elles ne domine l'analyse.
- Amélioration des comparaisons

**Principaux inconvénients de la “Normalization” et de la “Standardization” :** Perte de l'interprétabilité originale

# 4 Création de caractéristiques (Feature creation)

**Quelques exemples :**

1. En classification d'images :




1. Création de groupes (Kmeans, PCA) : de nouveaux groupes d'individus avec des caractéristiques similaires