Big Data Analytics

Data Mining

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...

Data Mining

Attention aux différents biais de vos données !

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...
- trouver de fausses variables explicatives

Data Mining

Meilleures données $>$ Meilleurs modèles
(trash-in, trash-out)

Préparation des données

- valeurs manquantes
- préprocessing (texte, image)
- standardisation
- transformation

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante
  - moyenne de la colonne

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante
  - moyenne de la colonne
  - prédiction d'un autre modèle

Préparation des données — préprocessing

- tokenizer, POS-tagger le texte (`https://spacy.io/`)
- utiliser un réseau de neurones préentraîné sur les images (`https://keras.io/applications/`)
- appliquer une transformée de fourier sur le son
- ...

Préparation des données — standardisation

Beaucoup de modèles travaillent mieux avec des données normales et sont plus efficaces autour de $[-5, 5]$:

- centrer sur la moyenne puis diviser par l'écart-type
- transformation de Box-Cox en cas d'asymétrie
- transformations spécifiques en fonction de la distribution

Préparation des données — transformation

Quand un modèle n'accepte pas de données catégorielles :

- label encoding si ordinal
- one-hot encoding sinon

Préparation des données — label encoding

Si les données sont ordinales :

Ordinal :

Température
Froid
Froid
Tiède
Chaud
Tiède

Label encoding :

Température
1
1
2
3
2

Préparation des données — one-hot encoding

Remplacer une feature par n features avec n le nombre de catégories.

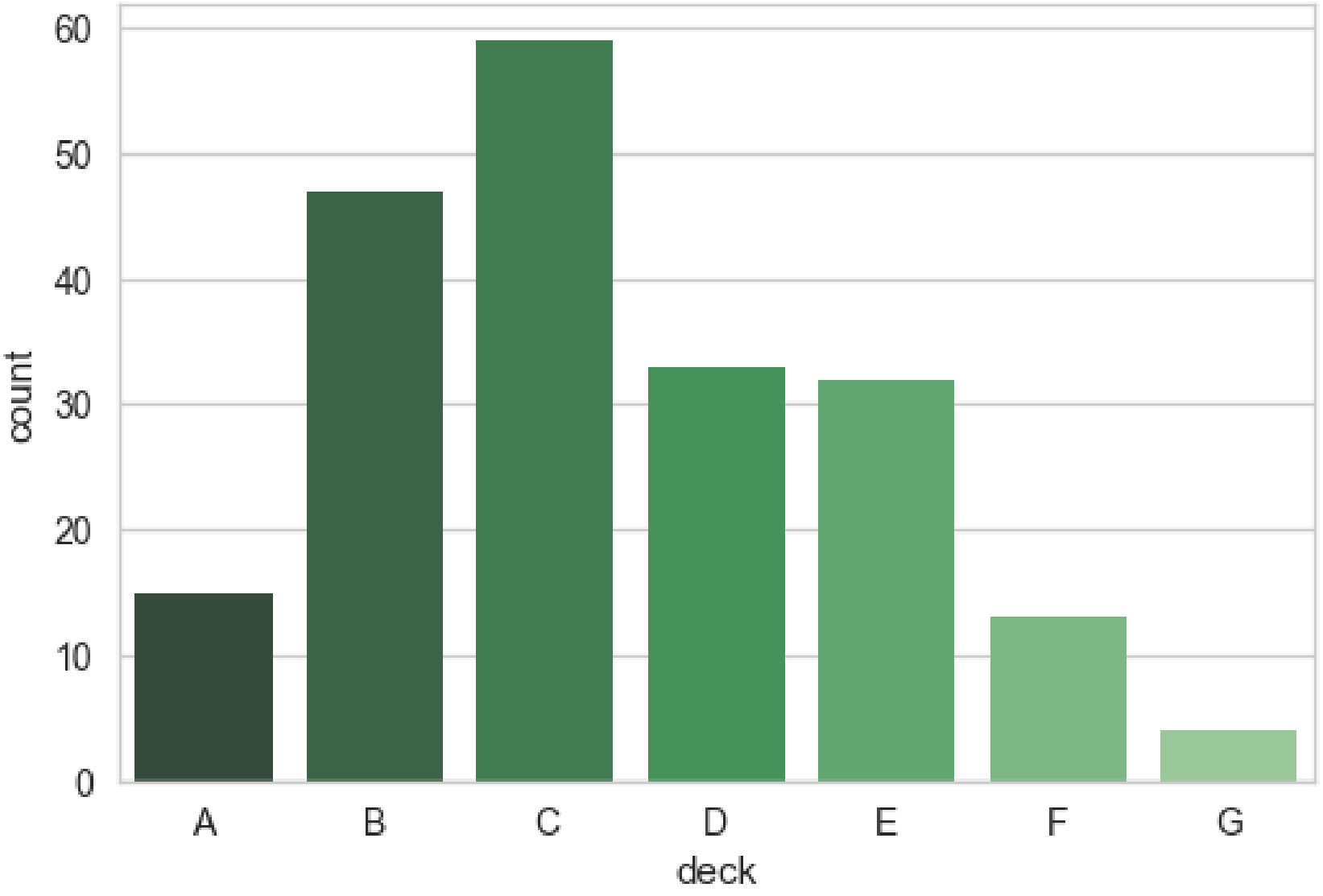
Catégoriel :

Couleur
Rouge
Rouge
Jaune
Vert
Jaune

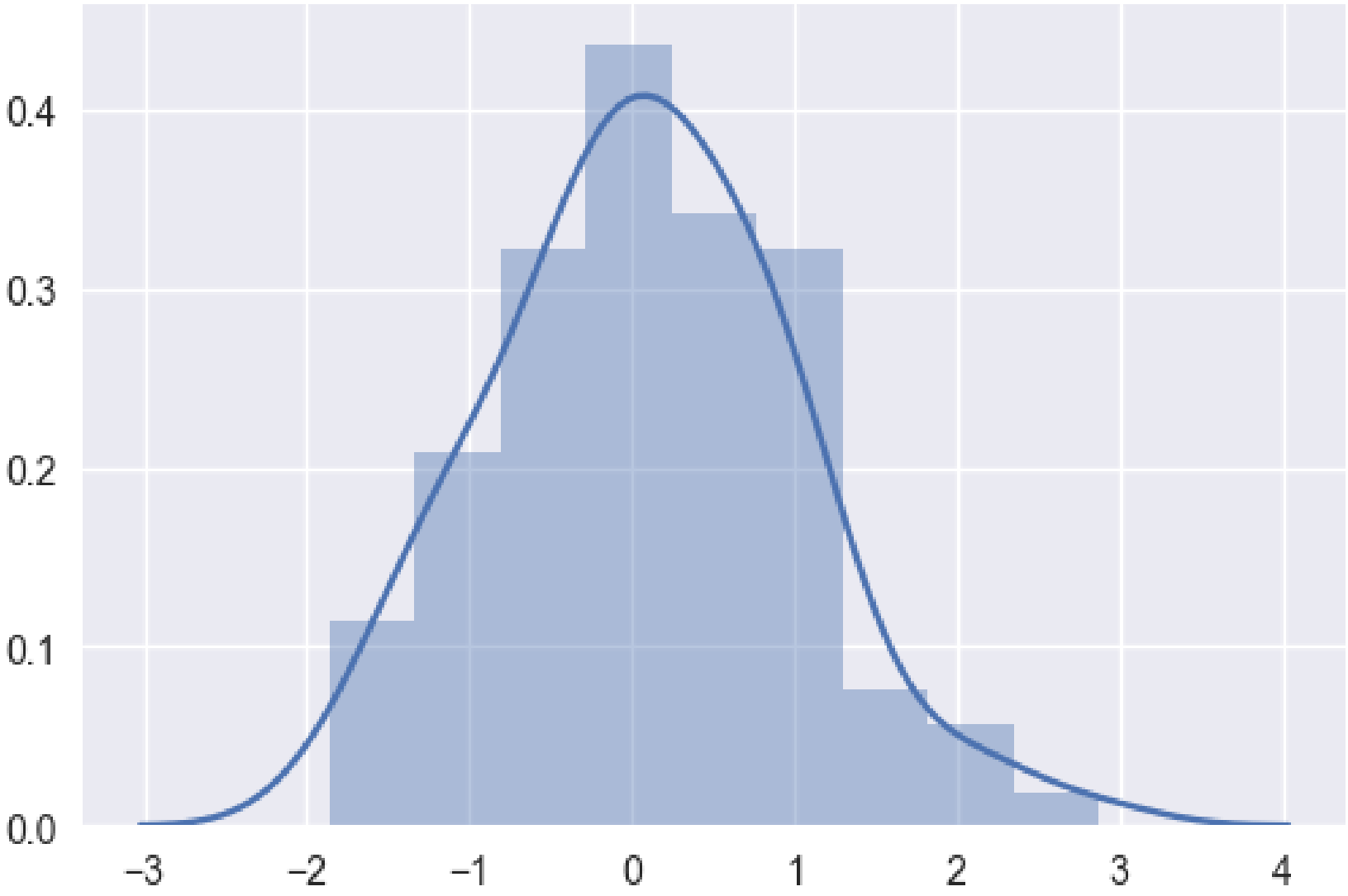
One-hot :

Rouge	Jaune	Vert
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

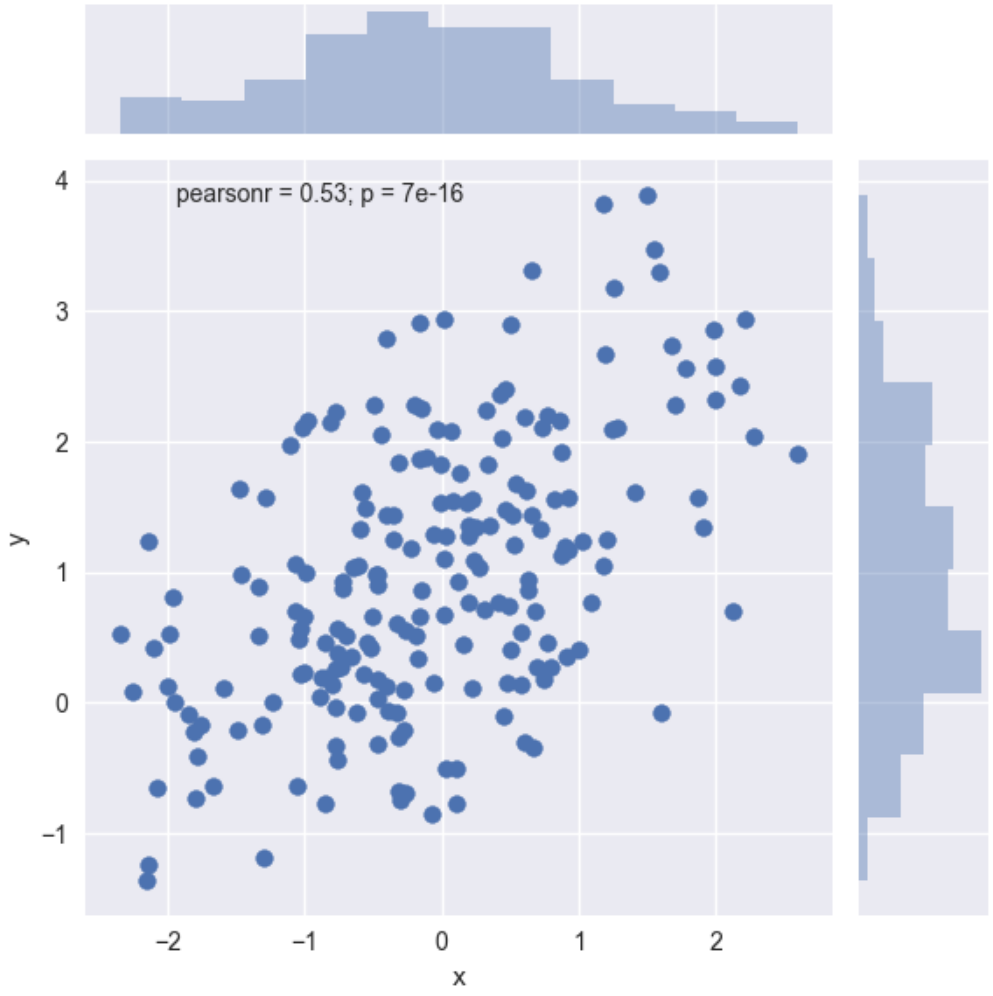
Outils — count plot



Outils — dist plot



Outils — scatter plot



Outils — correlation matrix