OPENCLASSROOMS

Mission 6

Détectez des faux billets

Mentor : Claire Della Vedova

Laurent Gourdon – Parcours Data Analyst

Agenda

Analyse de l'échantillon
- A propos du dataset
- Analyse univarée
- Analyses bivariées

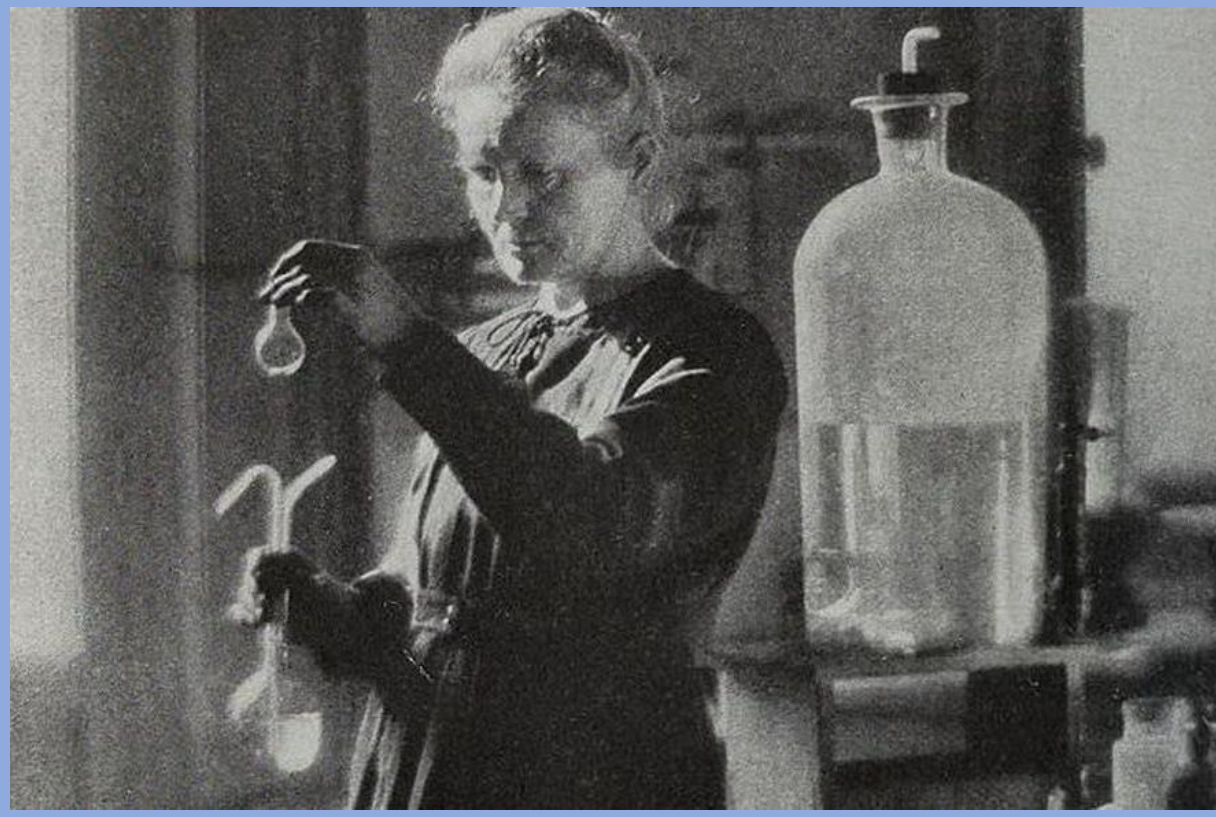
ACP
- Eboulis des valeurs propres et variances
- Contribution et qualité des variables
- Cercle des corrélations
- Projection des individus

K-means
- Projection des individus
- Confusion matrix

Régression logistique
- Multicolinéarité
- Confusion matrix
- Construcion d'un modèle de prédiciton

Mission 0

Analyse de l'échantillon



- A propos du dataset
- Analyse univarée
- Analyse bivariées

A propos du dataset

- L'échantillon a 170 observations et 7 variables
- La variable is_genuine (bool) nous indique vrais et faux billets
- Les autres sont **quantitatives** et caratérisent chaque individu (**mm**)
- Les mesures sont sur des échelles différentes (cf margin vs diagonal, length, height)
- Les écart-types de lengh et margin_low sont élevés : forte dispersion des individus

Dataset info

Number of variables	7
Number of observations	170
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	8.2 KiB
Average record size in memory	49.5 B

Variables types

Numeric	6
Categorical	0
Boolean	1
Date	0
URL	0
Text (Unique)	0
Rejected	0
Unsupported	0

Dataset overview

note_id	is_genuine	length	diagonal	height_left	height_right	margin_low	margin_up
0	True	112.83	171.81	104.86	104.95	4.52	2.89

Dataset describe

	length	diagonal	height_left	height_right	margin_low	margin_up
std	0.92	0.31	0.30	0.33	0.70	0.24
min	109.97	171.04	103.23	103.14	3.54	2.27
25%	111.85	171.73	103.84	103.69	4.05	3.01
50%	112.84	171.94	104.06	103.95	4.45	3.17
75%	113.29	172.14	104.29	104.17	5.13	3.33
max	113.98	173.01	104.86	104.95	6.28	3.68

Analyse univariée

True or False

- sur 170 billets, 100 sont vrai et 70 sont faux (+/- 60%-40%)

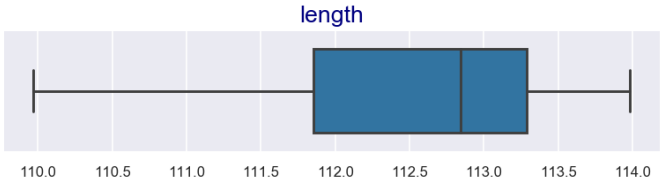
Variables numériques

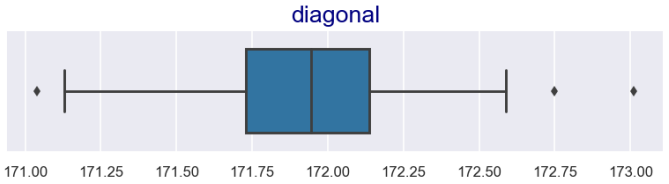
- diagonal, heigh_right et margin_up ont 5 outliers
- length a un skewness à gauche, médiane haute. Les données sont étalées (std=92)
- Heigh_left est plutôt centrée (std=30)
- margin_low a un skewness à droite , médiane haute. Les données sont étalées (std=70)

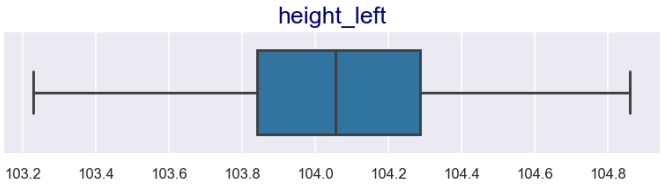
Is_genuine (boolean)

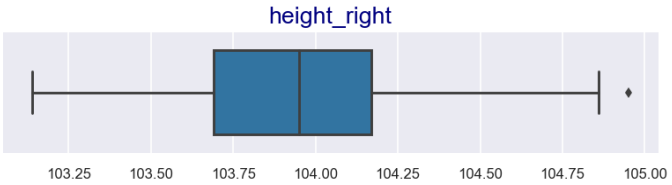
Value	Count	Frequency (%)
True	100	58.8%
False	70	41.2%

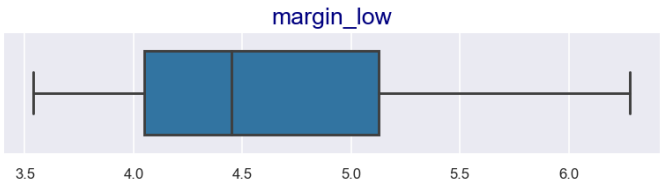
Variables numériques

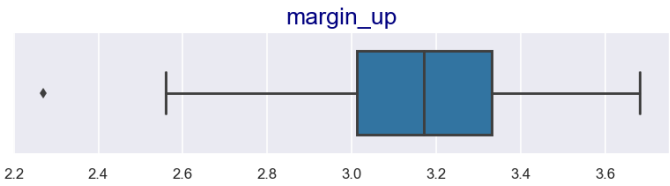








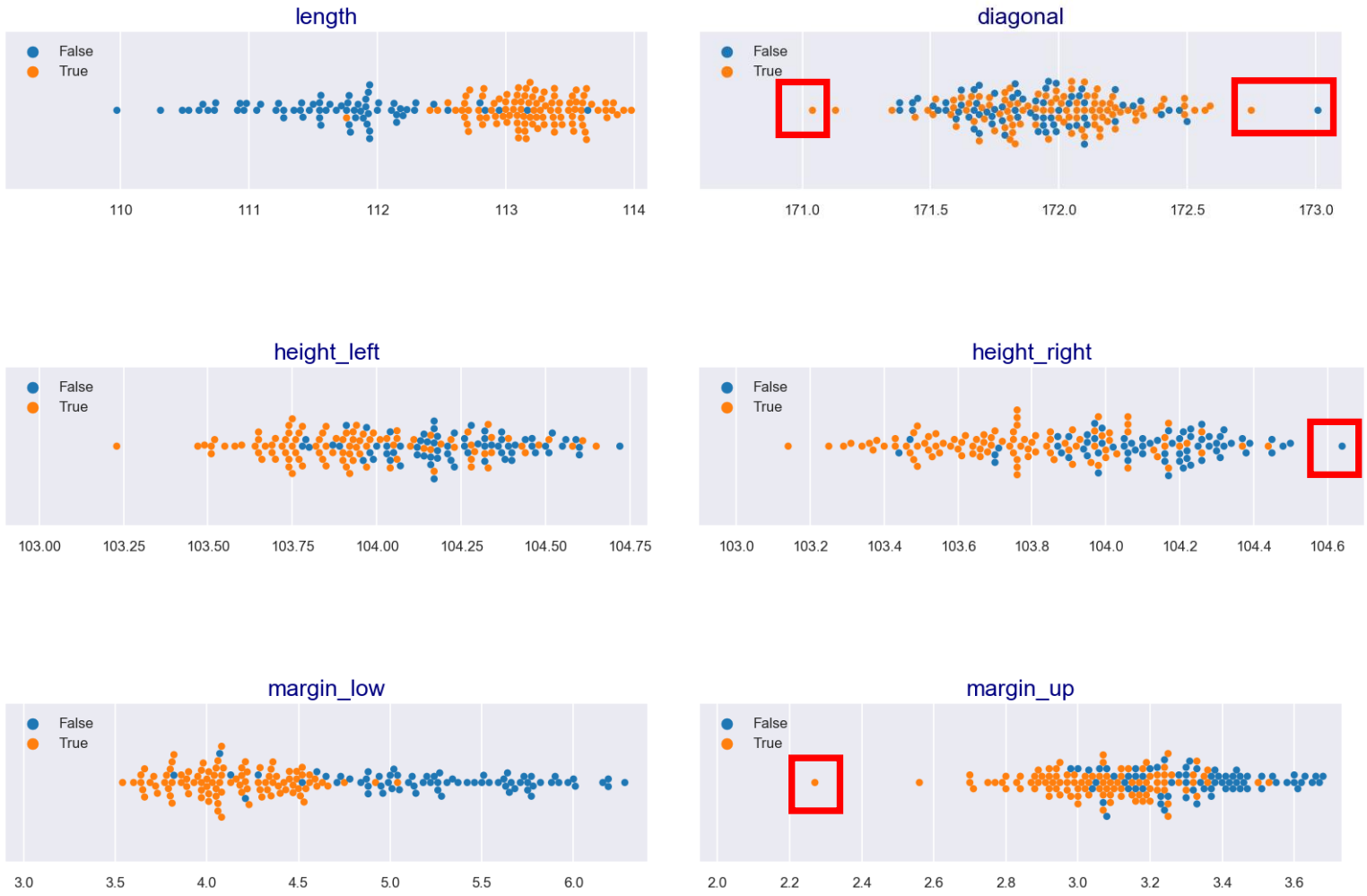




Analyse bivariée

- Pour lenght, la distribution des individus montre que les faux billets sont plutôt moins larges que les vrais billets
- Pour margin_low, la distribution des individus montre que les faux billets tendent à avoir une marge inférieure plus grande que les vrais billets
- Pour les autres variables, on peut difficilement séparer vrai et faux en fonction de la dimension au vu des graphiques

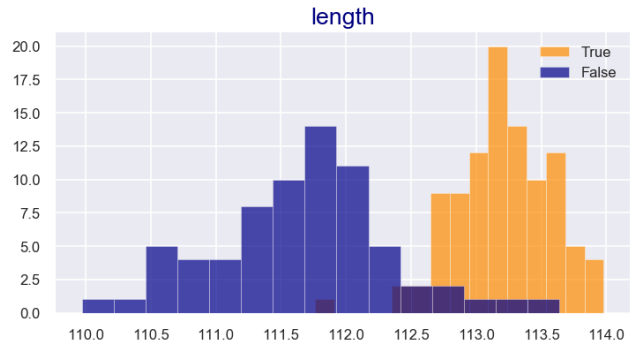
Dispersion de *True* et *False* par mesure (mm) : individus

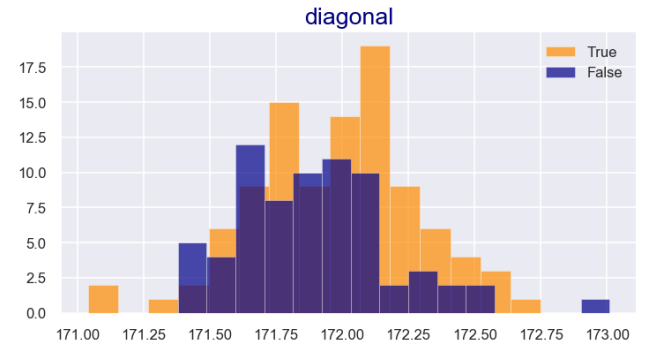


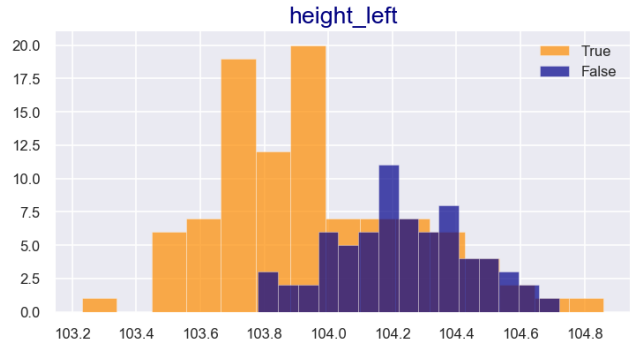
Analyse bivariée

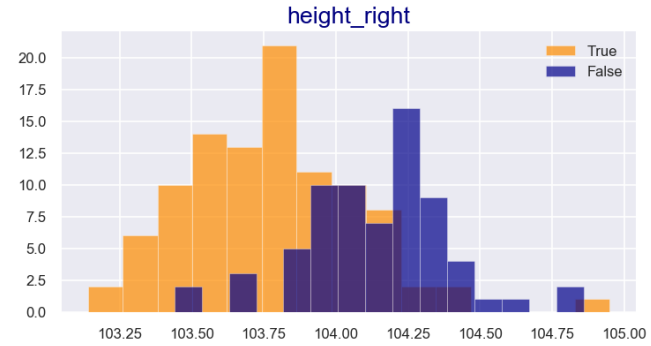
- Pour **length** et **margin_low**, nous confirmons une dichotomie relative entre *True* et *False* selon la mesure des billets
- Pour **length** les vrais billets ont tendance à avoir une mesure élevée
- Pour **height** et **margin** les **billets** *True* tendent à avoir une mesure basse
- Pour **margin**, les *faux billets* tendent à avoir une mesure élevée

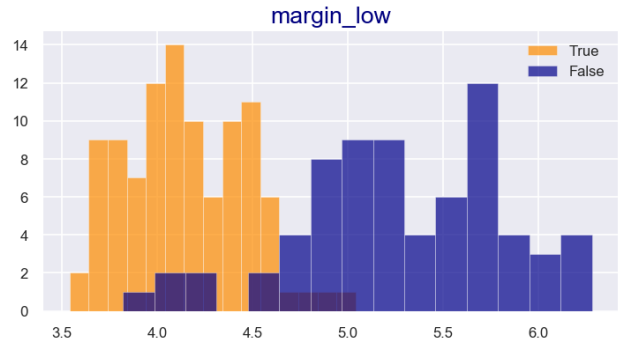
Dispersion de *True* et *False* par mesure (mm) : fréquence

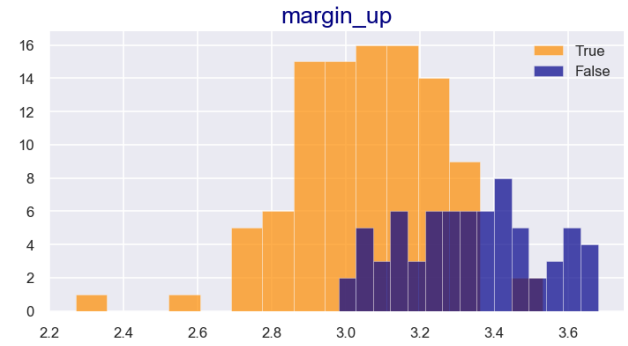












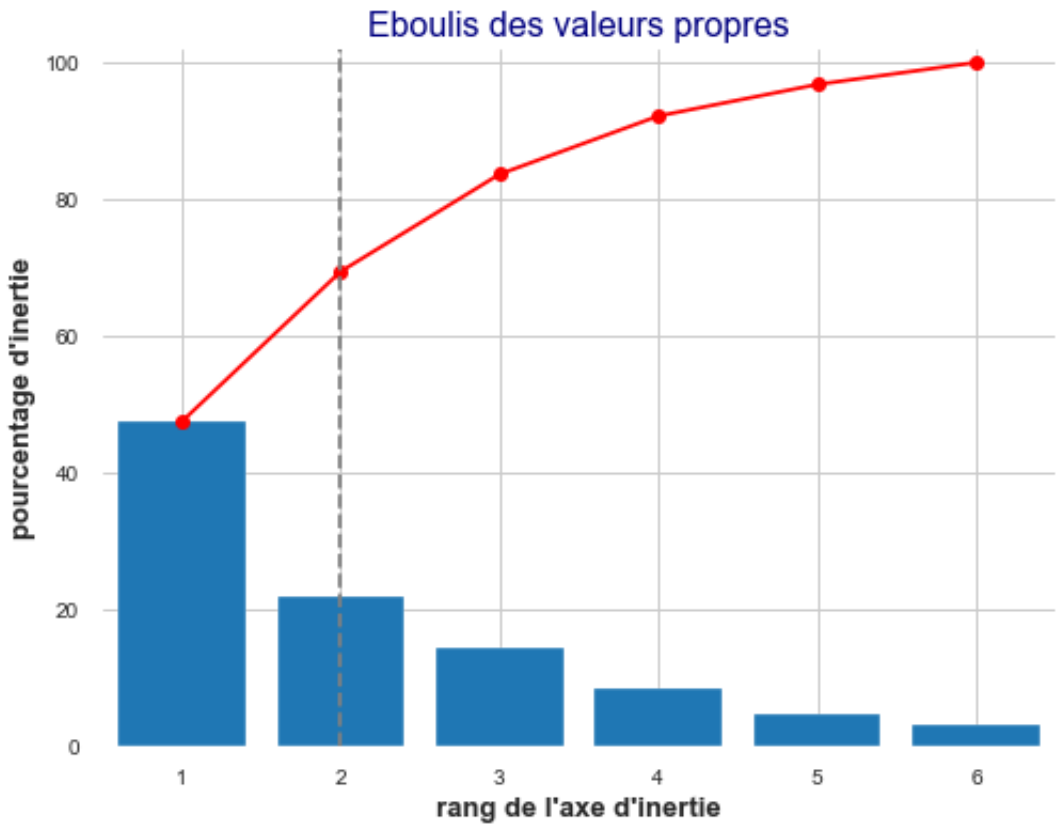
Mission 1

Analyse en composantes principales

- Eboulis des valeurs propres et variances
- Contribution et qualité des variables
- Cercle des corrélations
- Projection des individus

Eboulis des valeurs propres

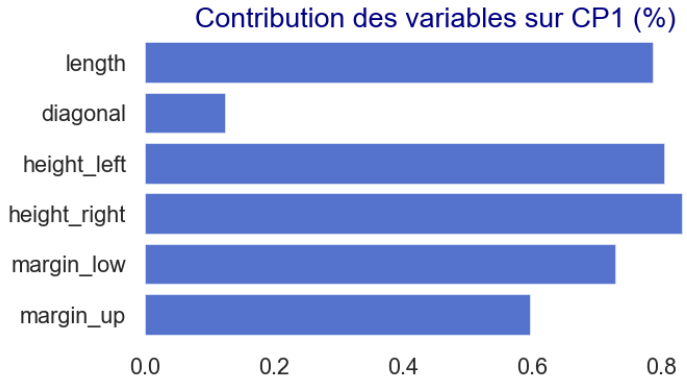
- Le coude apparaît dès la deuxième composante.
- Le premier plan factoriel CP1-CP2 représente 69,4 % variance cumulée
- Il est légitime de se poser la question de travailler avec une représentation en 3D, en ajoutant CP3 (83,6 % variance cumulée)
- Nous nous en tenons aux 2 dimensions offrant la meilleure variance cumulée

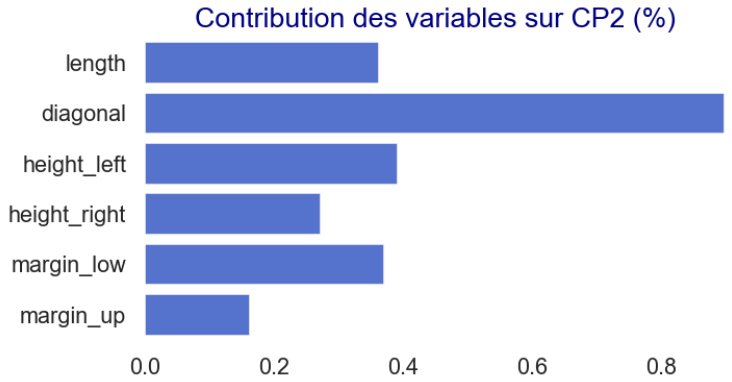


	CP1	CP2	CP3	CP4	CP5	CP6
Variance expliquée	47.45	21.96	14.23	8.53	4.61	3.22
Variance cumulée	47.45	69.41	83.64	92.17	96.78	100.00

Contribution et qualité des variables

- Les variables **heigh_left**, **heigh_right**, **length** et **margin_low** sont très biens représentées sur l'axe CP1
- La variable **diagonal** est de loin la mieux représentée sur l'axe CP2 (90%)
- Cependant, **length**, **height_left** et **margin_low** contribuent à hauteur de 36-39% sur CP2
- Qualité de représentation des variables COS^2 confirme les observations



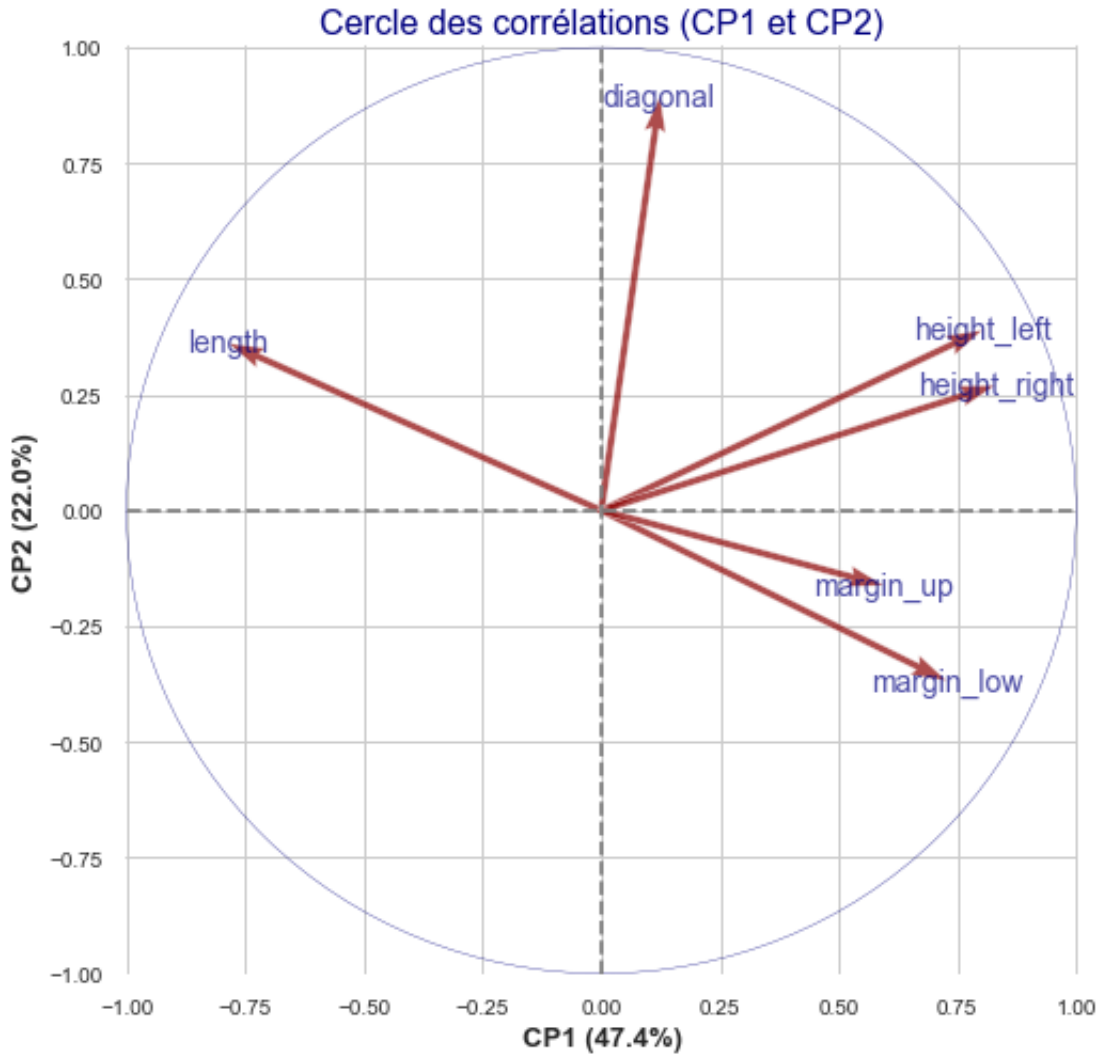


Qualité de représentation des variables (CP1-CP2)

	COS2_1	COS2_2
length	61.66	13.03
diagonal	1.53	80.08
height_left	64.37	15.16
height_right	68.86	7.31
margin_low	52.89	13.54
margin_up	35.38	2.62

Cercle des corrélations : contribution des variables

- Les variables heigh_left, heigh_right et margin_low sont positivement très biens représentées sur l'axe CP1
- La variable length est négativement très bien représentée sur l'axe CP1.
- margin_low est anticorrélée à length
- La variable diagonal est la mieux représentée sur l'axe CP2
- diagonal n'est pas corrélée à margin_up et margin_low, et faiblement à length

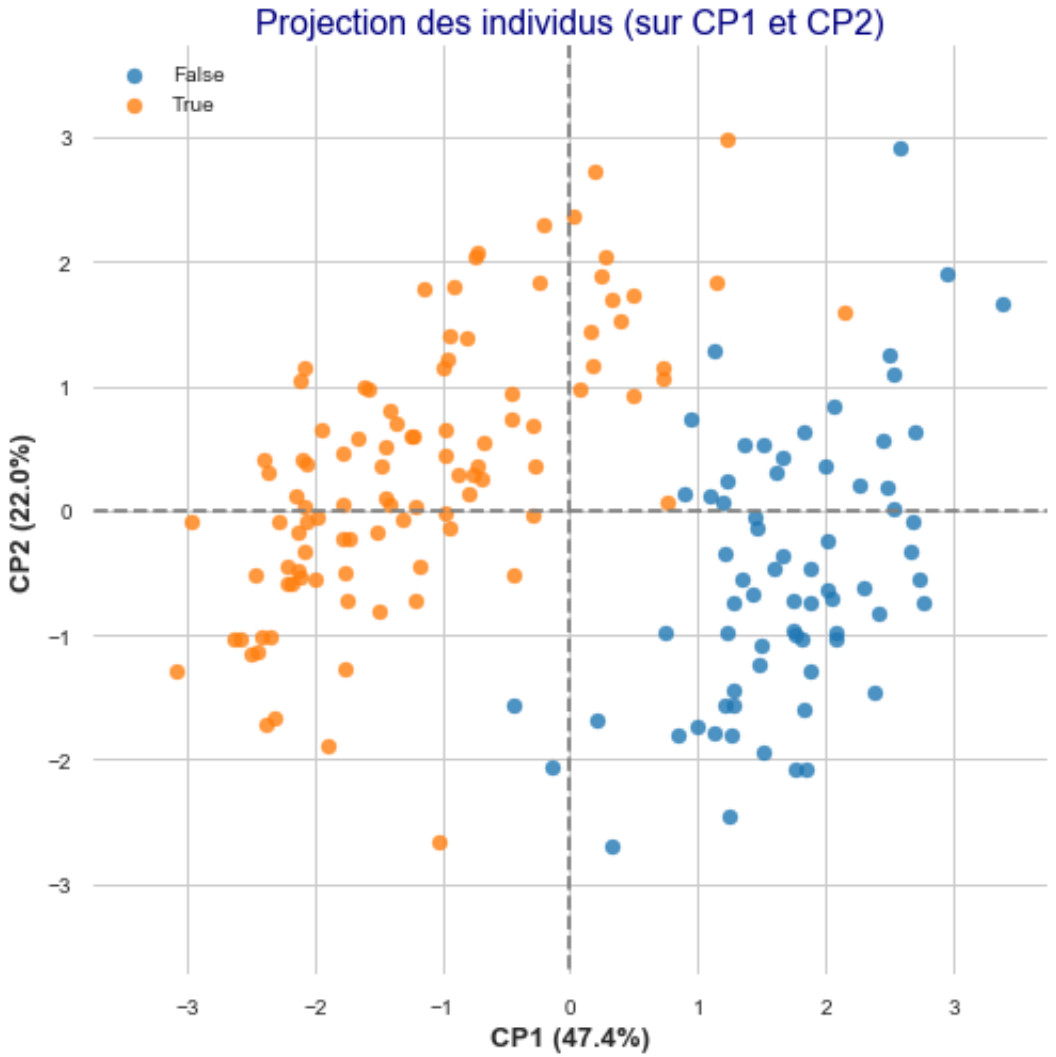


Projection des individus : premier plan factoriel CP1-CP2

Sur le premier plan on peut distinguer 3 zones :

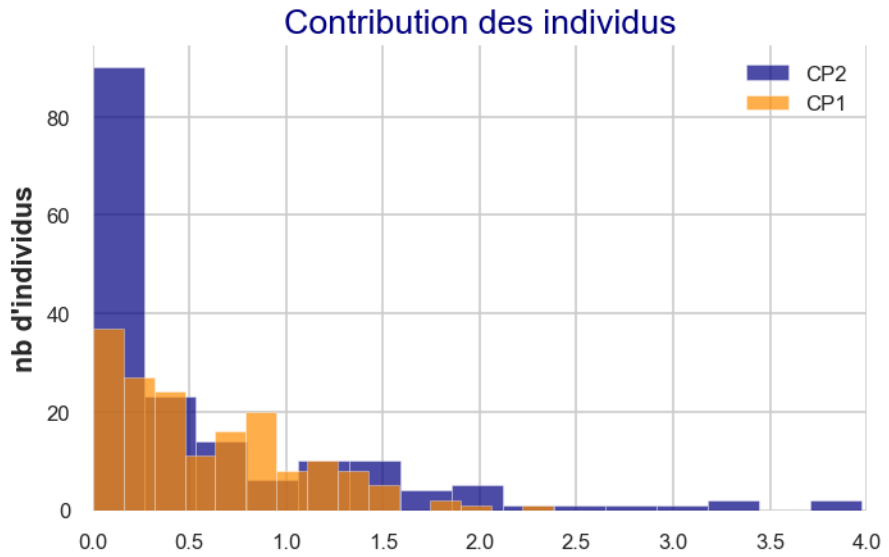
- Sous -0,6 il est presque certain d'avoir des vrais billets
- Au-dessus de 1,0, la probabilité est élevée d'avoir de faux billets
- Entre -0,6 et **1,0** cela peut être un vrai ou un faux billet, à moins de tenir compte de la distribution sur le 2e plan

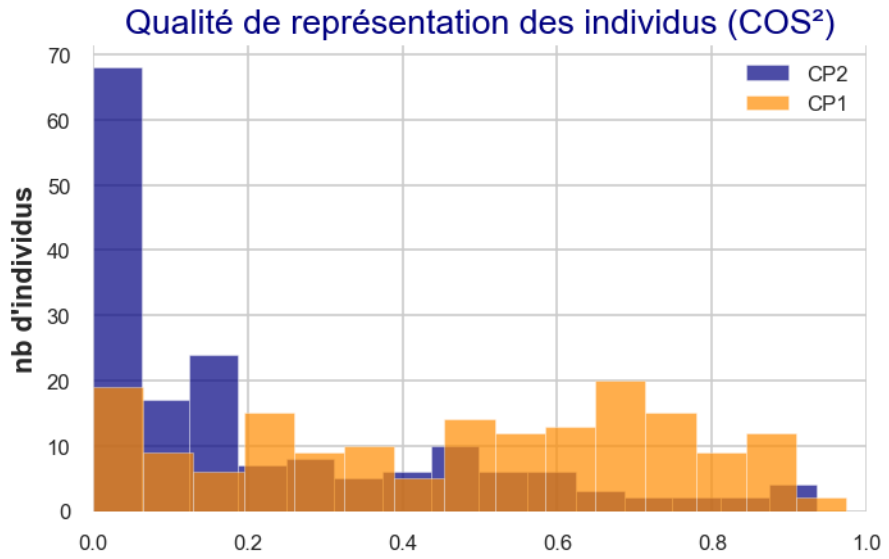
En effet, les vrais billets sur cette zone sont plutôt vrais au-dessus de -0,6 sur CP2 et faux en-dessous



Projection des individus
contribution et
qualité de représentation

- La qualité de représentation des individus a été requise par le commanditaire.
- Cela ne nous semble pas, contrairement à la qualité de représentation des variables, un critère d'analyse pertinent pour objectif de prédiction de vrais/faux billets.
- Ce qui compte pour notre modèle, ce ne sont pas tant les caractéristiques individuelles de chaque billet de l'échantillon que les caractéristiques des mesures utilisées pour déterminer l'authenticité ou non d'un billet.





Mission 2

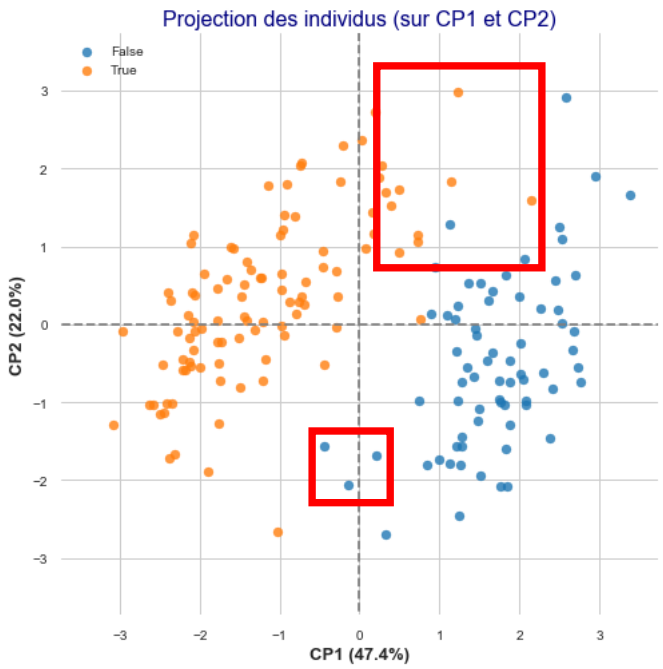
K-means

- Projection des individus
- Confusion matrix

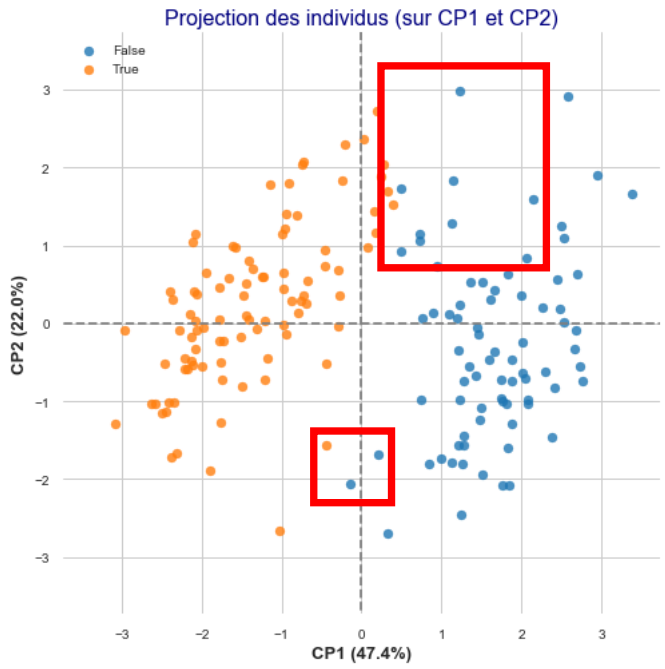
K-means
projection des individus

- L'apprentissage non-supervisé K-means (avec K=2 clusters) est une méthode qui fournit rapidement une solution, pas nécessairement optimale par son appartenance à une classe de problèmes donnés déjà identifiés (heuristique)
- Notre modèle, projeté sur les plans factoriels 1 et 2 propose une solution de classement proche de celle fourni par le statut réel des individus (is_genuine) sur les composantes 1 et 2

ACP (centré-réduit)

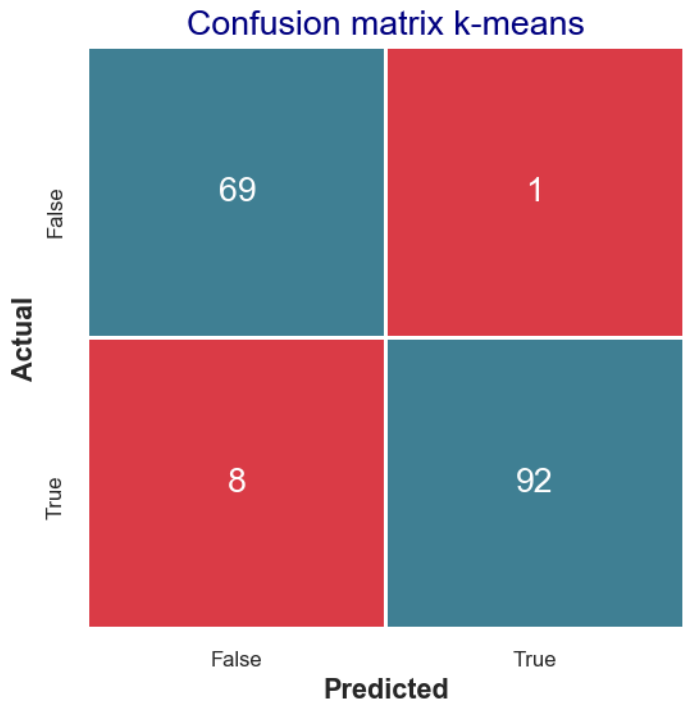


k-means (centré-réduit)



K-means
confusion matrix

- La matrice de confusion nous montre que sur 100 vrais billets, 92 ont été identifies comme vrais et 8 comme faux par l'algorithme
- Sur 70 faux billets, il en classe 69 comme faux et un comme vrai
- On pourrait en déduire que les faux billets sont plutôt mieux identifiés que les vrais par cette méthode



Marge d'erreur constatée (%)

- Faux billets 1,4%
- Vrais billets 8,0%
- Ensemble 6,5%

Mission 3

Régression logistique

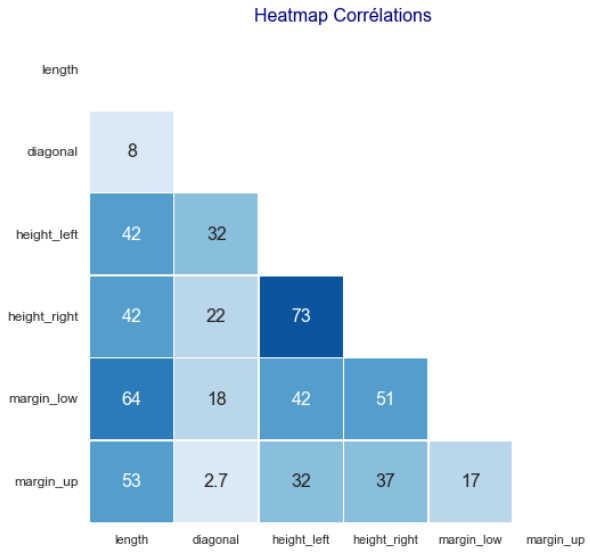
- Multicolinéarité
- Confusion matrix
- Construcion d’un modèle de prédiciton

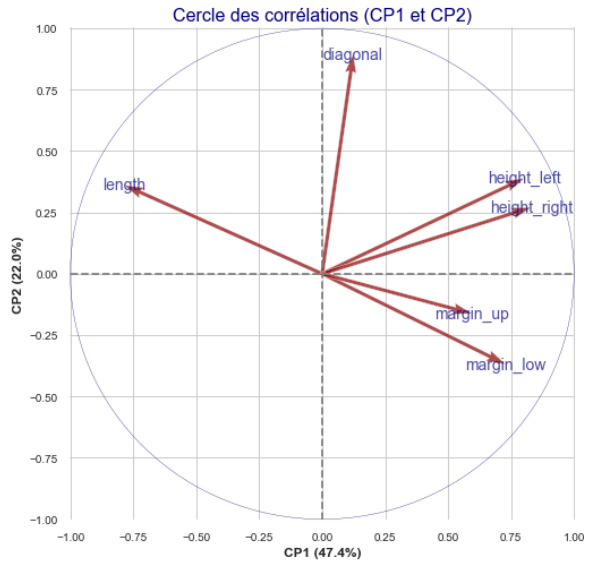
Corrélation & Colinéarité

La régression logistique avec statsmodels ne converge pas avec 6 puis 5 variables

- Forte corrélation entre heigh_left et heigh_right
- Forte corrélation entre length et margin_low
- heigh_left et margin_low ont un fort Variance Inflation Factor indiquant une multi-colinéarité élevée
- Nous réalisons notre modèle de régression logistique, qui converge, avec diagonal, heigh_right, margin_up et lenght

Corrélation





Calcul VIF pour 6 et 4 variables

Six variables	VIF
diagonal	6,63
height_left	12,21
height_right	2,34
margin_low	7,95
margin_up	2,70
length	2,87

Quatre variables	VIF
diagonal	1,14
height_right	1,16
margin_up	1,08
length	1,04

Régression logistique

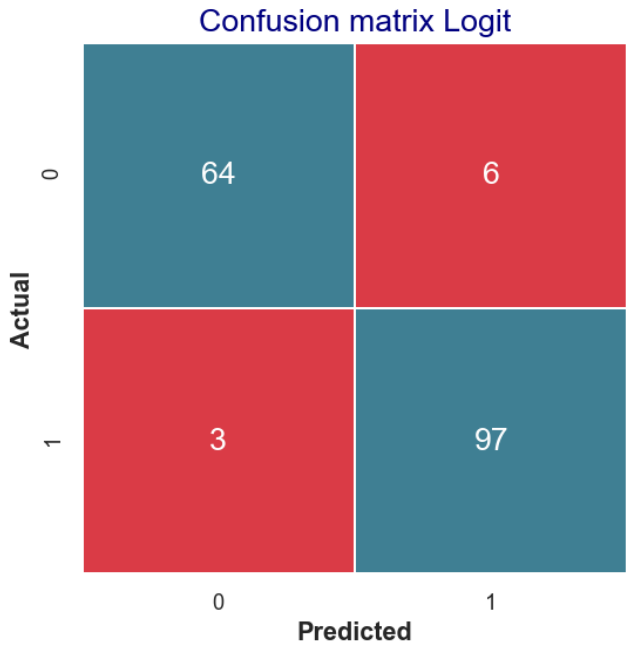
Les résultats de Logit sous statsmodels confirment que le modèle est pertinent

- Log-Likelihood Ratio p-value est quasi-nulle => Notre modèle est significatif
- Les ratio P>z (p-values) des variables sont significatifs
- Nous réalisons un test d'entraînement sur l'échantillon

Logit Regression Results

Dep. Variable:	is_genuine	**No. Observations:**	170
Model:	Logit	**Df Residuals:**	165
Method:	MLE	**Df Model:**	4
Date:	Mon, 16 Sep 2019	**Pseudo R-squ.:**	0.8147
Time:	12:09:55	**Log-Likelihood:**	-21.347
converged:	True	**LL-Null:**	-115.17
Covariance Type:	nonrobust	**LLR p-value:**	1.692e-39

	coef	std err	z	P>|z|
const	-402.2483	258.483	-1.556	0.120
x1	3.4597	0.728	4.753	0.000
x2	2.9821	1.376	2.167	0.030
x3	-4.6289	1.537	-3.011	0.003
x4	-5.6137	2.695	-2.083	0.037



Marge d'erreur constatée (%)

- Faux billets 9,4%
- Vrais billets 3,0%
- Ensemble 5,3%

Train test

- Set d'entraînement : 119 individus
- Set de validation : 51 individus
- Variance expliquée : 0.84
- Score Log_loss : 0.17

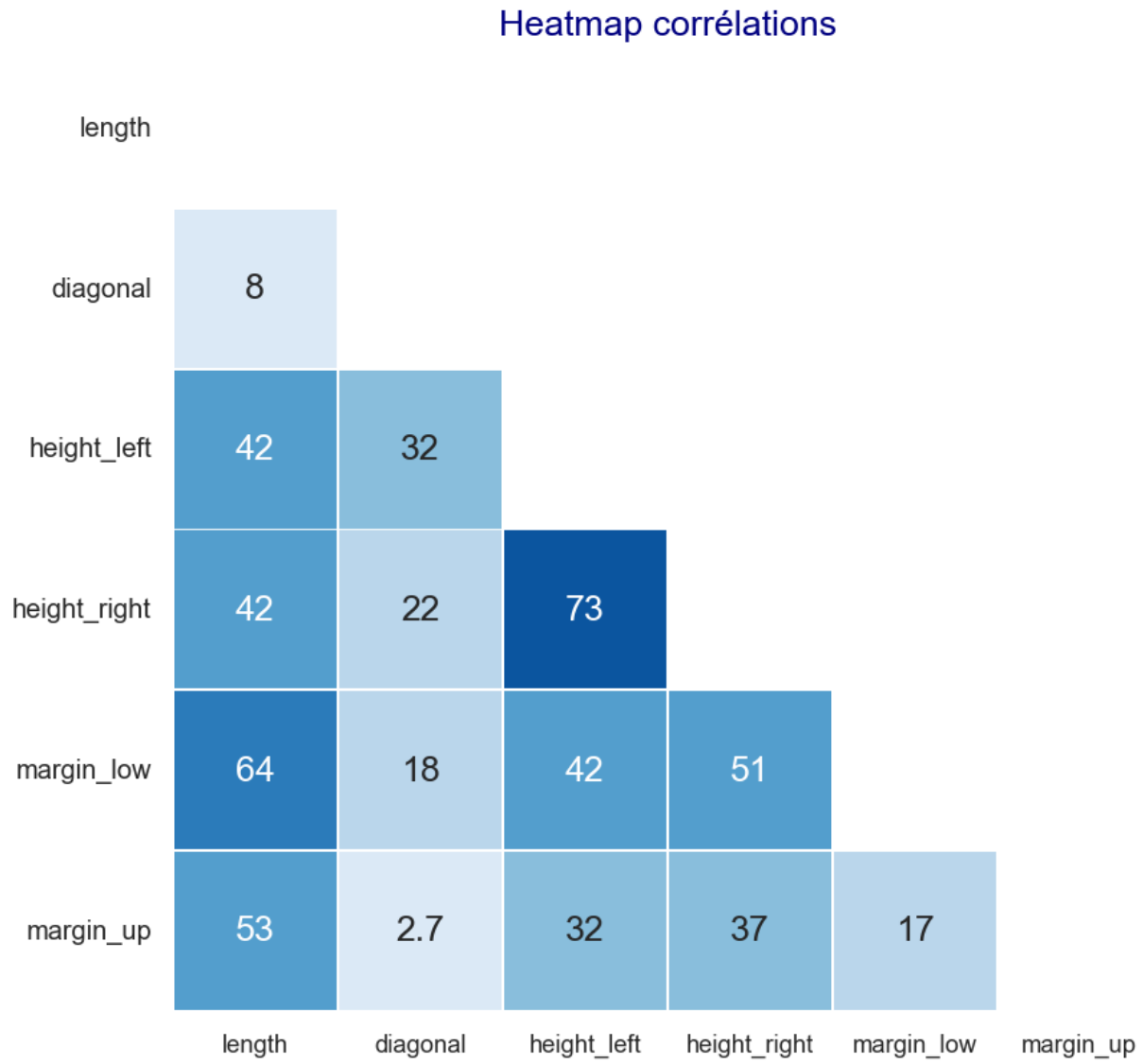
Des questions ?

lgourdon@gmail.com (+34 659 34 05 15)

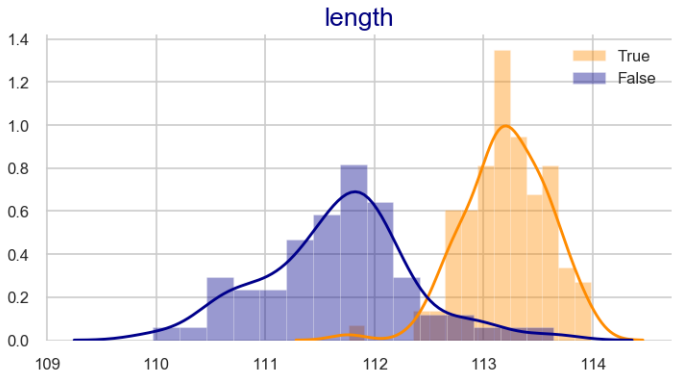
Conclusion

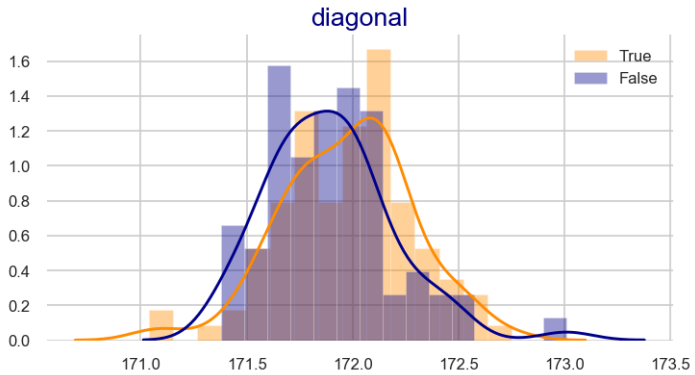
Pour aller plus loin

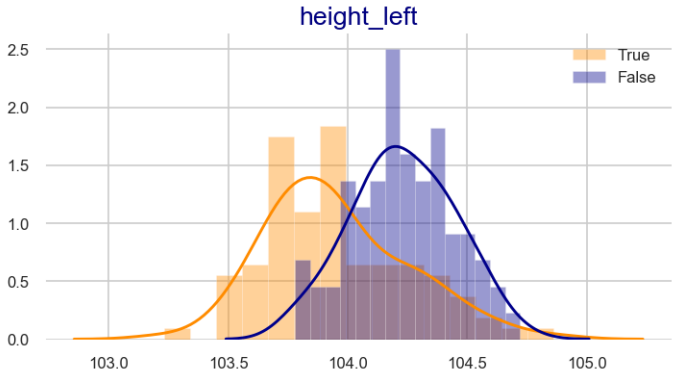
Analyses bivariées corrélations

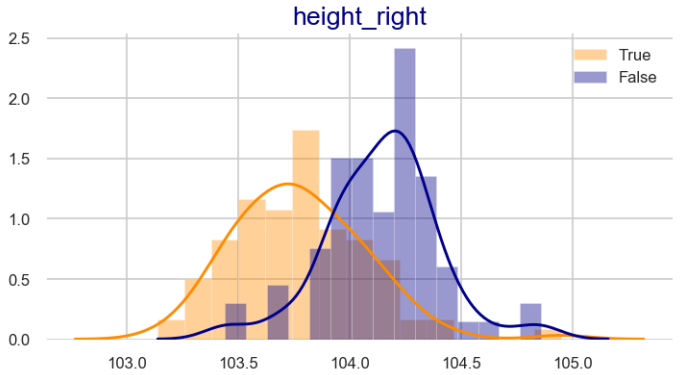


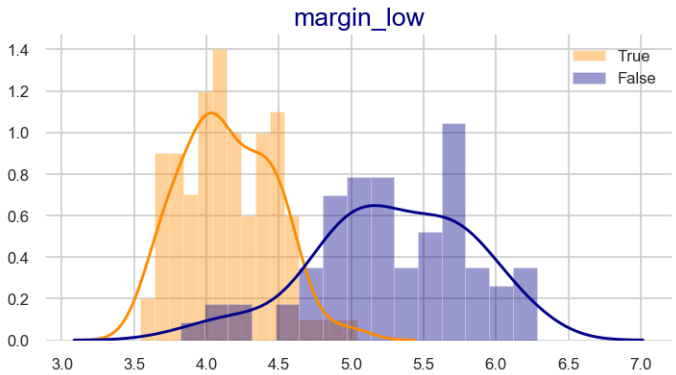
Analyses bivariées True or False

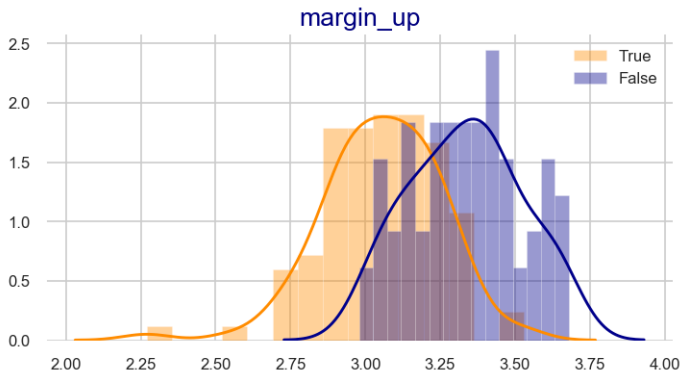












Analyses bivariées dispersion