DÉTECTEZ DES FAUX BILLETS

Sebastien LIM

Introduction

L'ONCFM contribue à la lutte contre le faux-monnayage grâce à l'analyse de données et à l'identification des tendances et des modèles pertinents. Elle a pour objectif de mettre en place des méthodes d'identification des contrefaçons des billets en euros
Nous sommes charges de mettre en place un algorithme qui soit capable de différencier automatiquement les vrais des faux billets.

Ainsi, il faudrait construire un algorithme qui, à partir des caractéristiques géométriques d'un billet, serait capable de définir si ce dernier est un vrai ou un faux billet.

Le jeu de données

*1500 non-null Valeurs

*6 Dimensions géométriques

*37 valeurs manquantes (margin low)

Dimensions géométriques

length : la longueur du billet (en mm)

height_left : la hauteur du billet (mesurée sur le côté gauche, en mm)

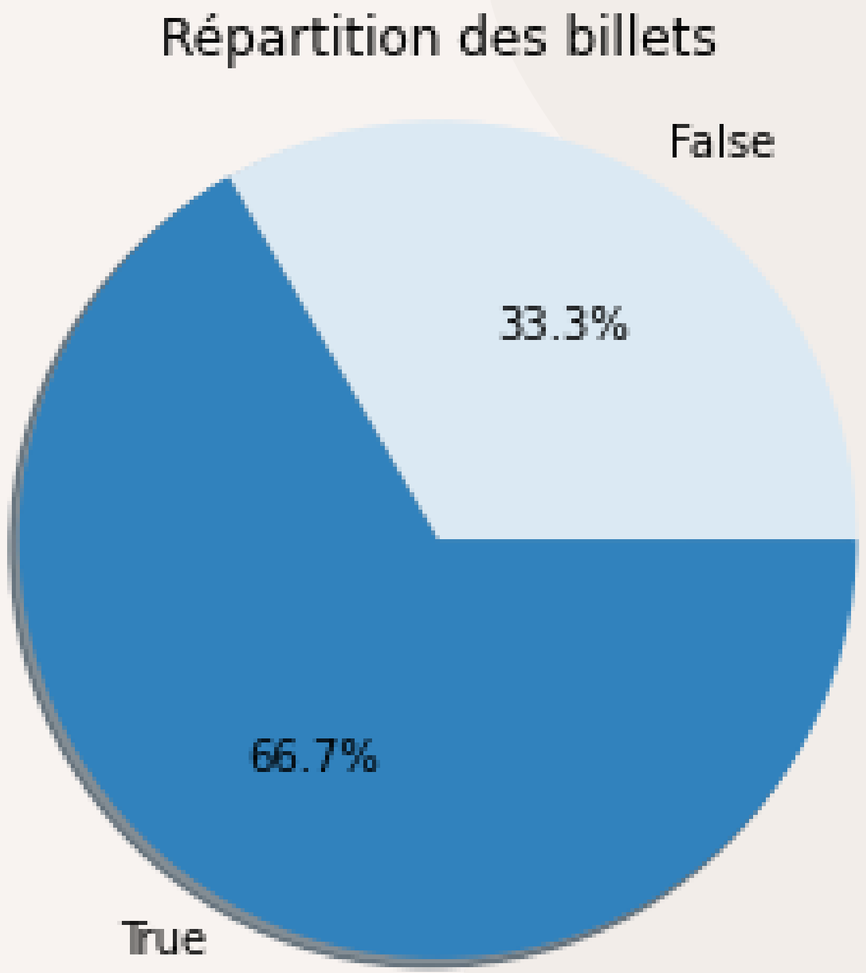
height_right : la hauteur du billet (mesurée sur le côté droit, en mm)

margin_up : la marge entre le bord supérieur du billet et l'image de celui-ci (en mm) ;26

margin_low : la marge entre le bord inférieur du billet et l'image de celui-ci (en mm)

diagonal : la diagonale du billet (en mm)

La répartition des billets



500 faux billets
1000 vrais billets

Régression linéaire multiple

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.477
Model:	OLS	Adj. R-squared:	0.476
Method:	Least Squares	F-statistic:	266.1
Date:	Mon, 24 Apr 2023	Prob (F-statistic):	2.60e-202
Time:	09:11:59	Log-Likelihood:	-1001.3
No. Observations:	1463	AIC:	2015.
Df Residuals:	1457	BIC:	2046.
Df Model:	5		
Covariance Type:	nonrobust		

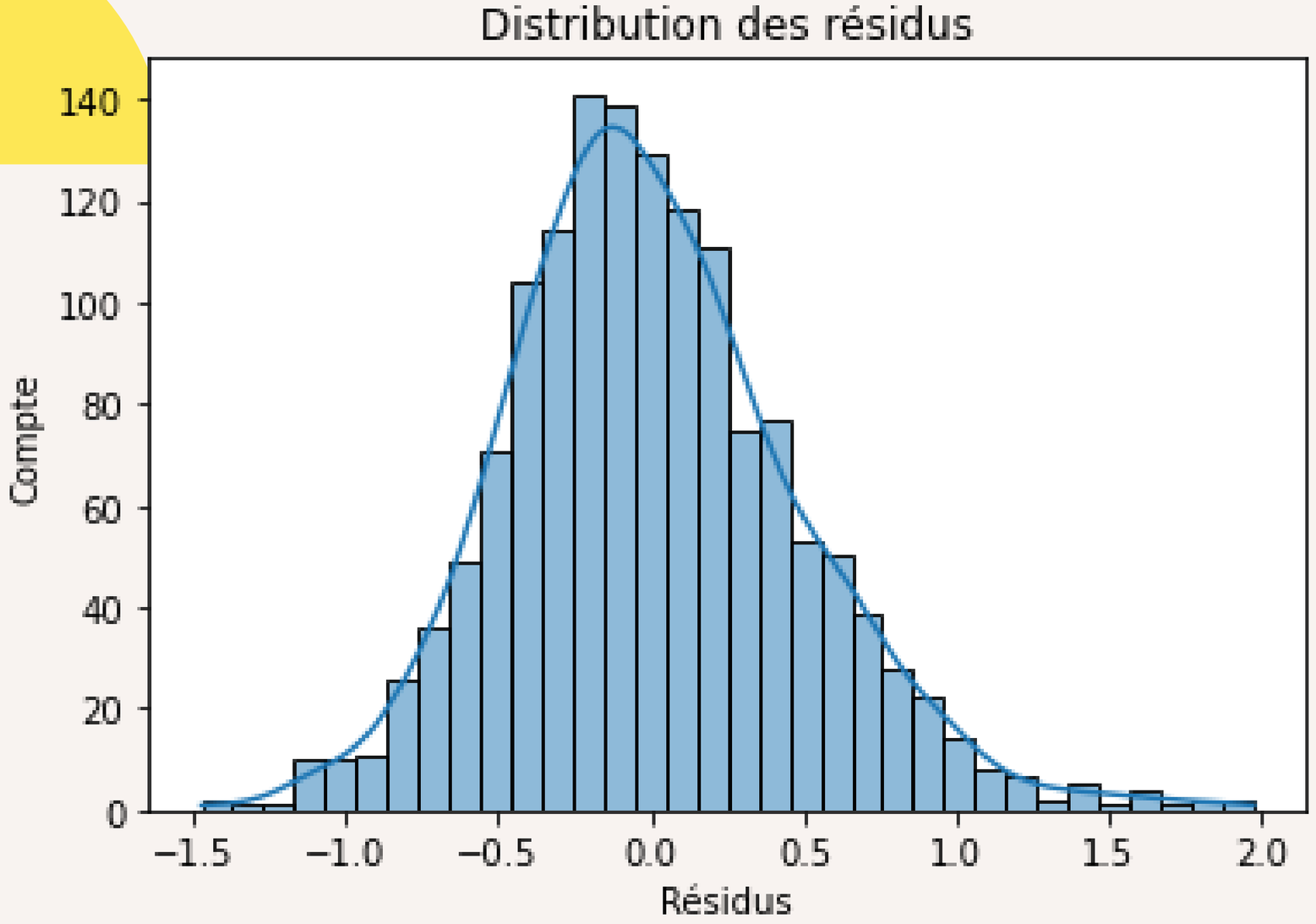
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.9948	9.656	2.382	0.017	4.055	41.935
margin_up	0.2562	0.064	3.980	0.000	0.130	0.382
height_right	0.2571	0.043	5.978	0.000	0.173	0.342
height_left	0.1841	0.045	4.113	0.000	0.096	0.272
diagonal	-0.1111	0.041	-2.680	0.007	-0.192	-0.030
length	-0.4091	0.018	-22.627	0.000	-0.445	-0.374

Omnibus:	73.627	Durbin-Watson:	1.893
Prob(Omnibus):	0.000	Jarque-Bera (JB):	95.862
Skew:	0.482	Prob(JB):	1.53e-21
Kurtosis:	3.801	Cond. No.	1.94e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

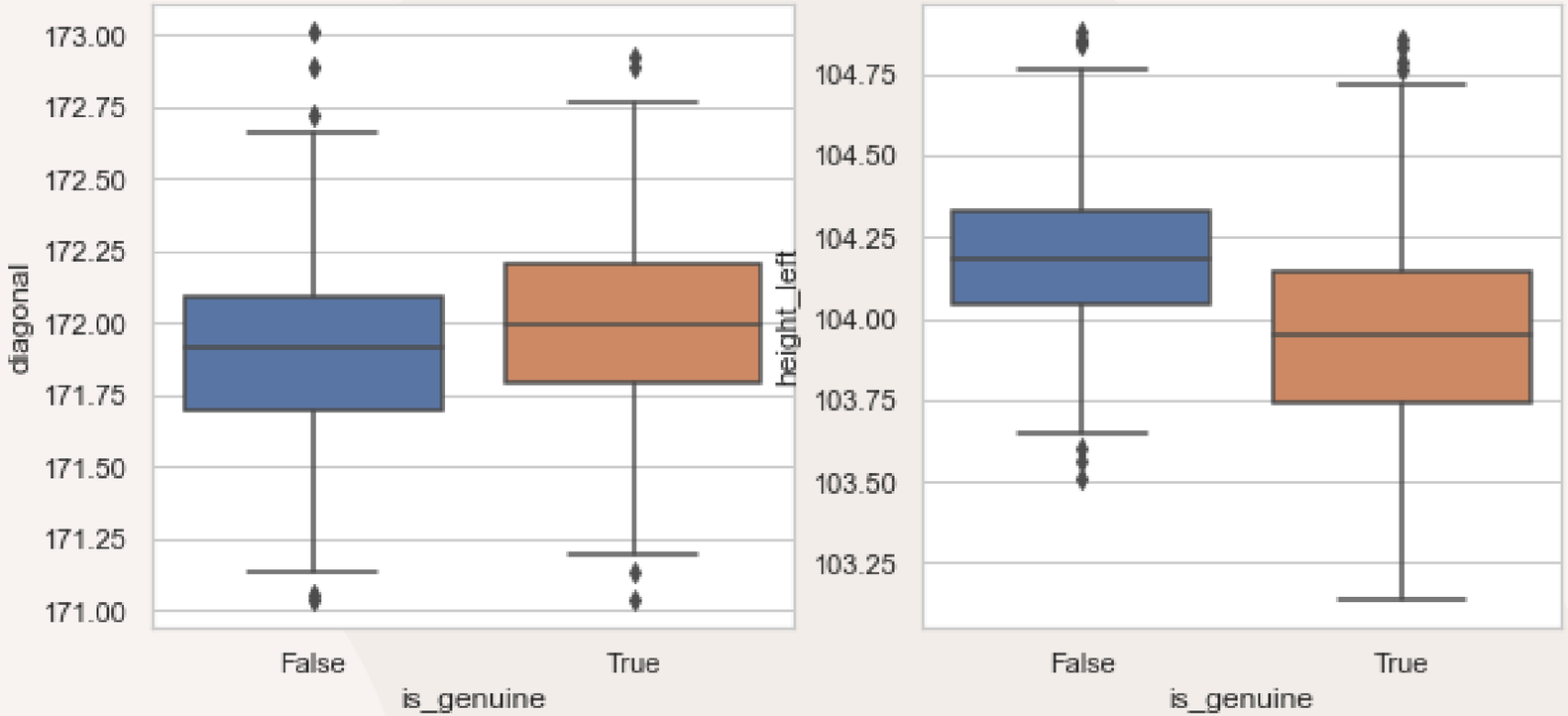
Grâce à la régression linéaire multiple, les 37 données manquantes pour la colonne Margin low ont pu être remplacées.

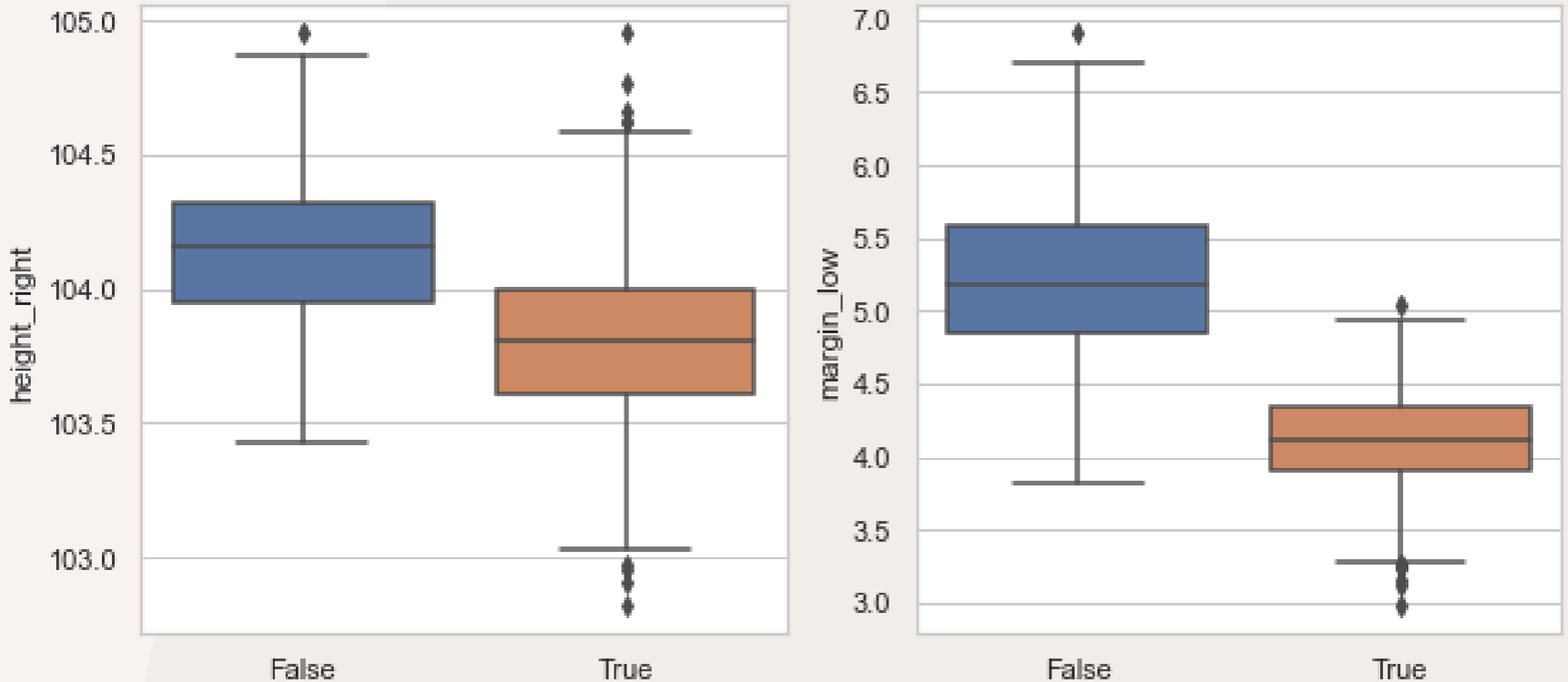
Distribution des résidus



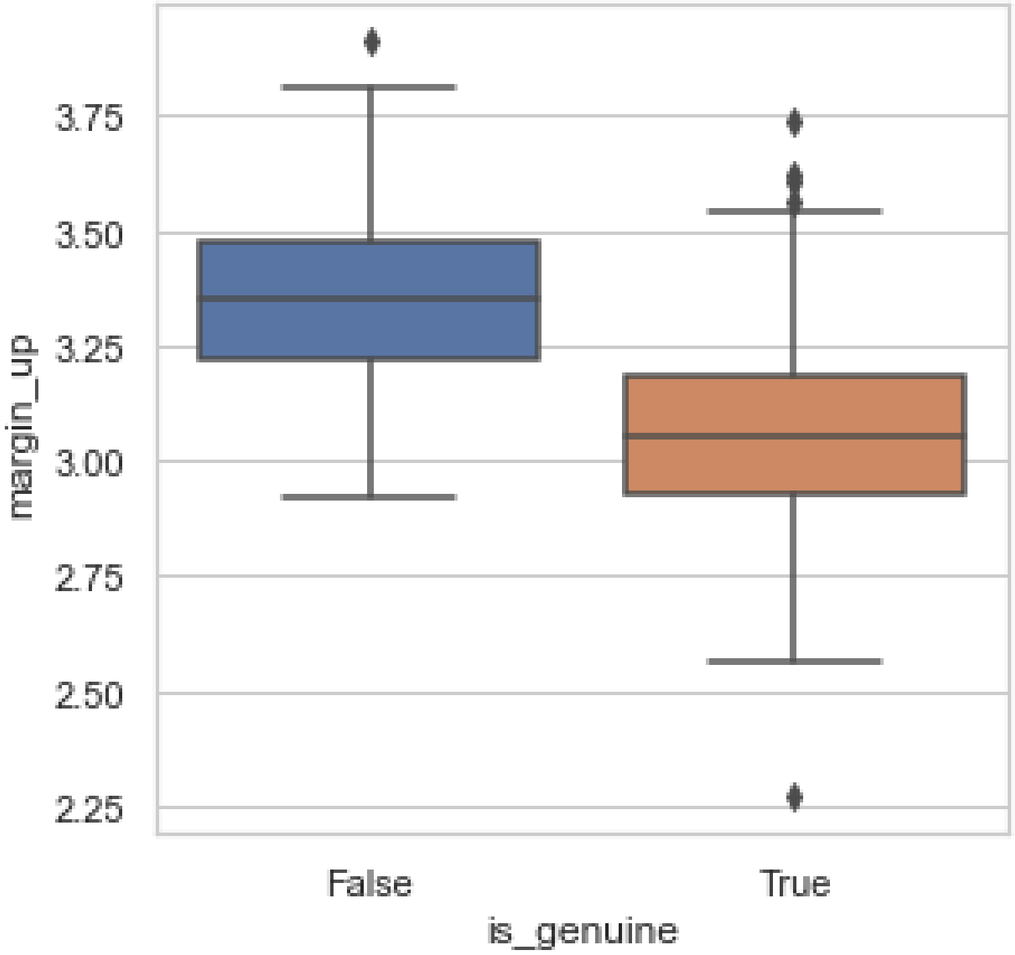
La p-value du test d'Aderson-Darling vaut 0.0

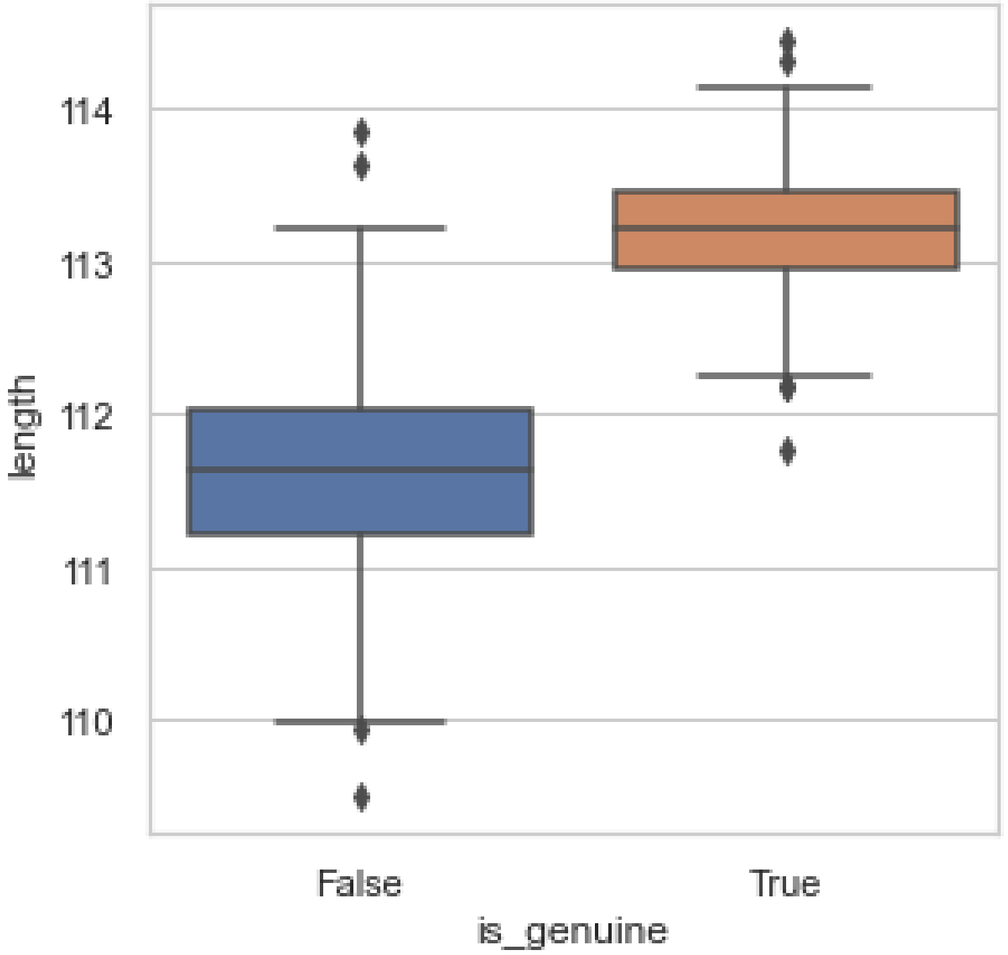
La répartition des dimensions des billets



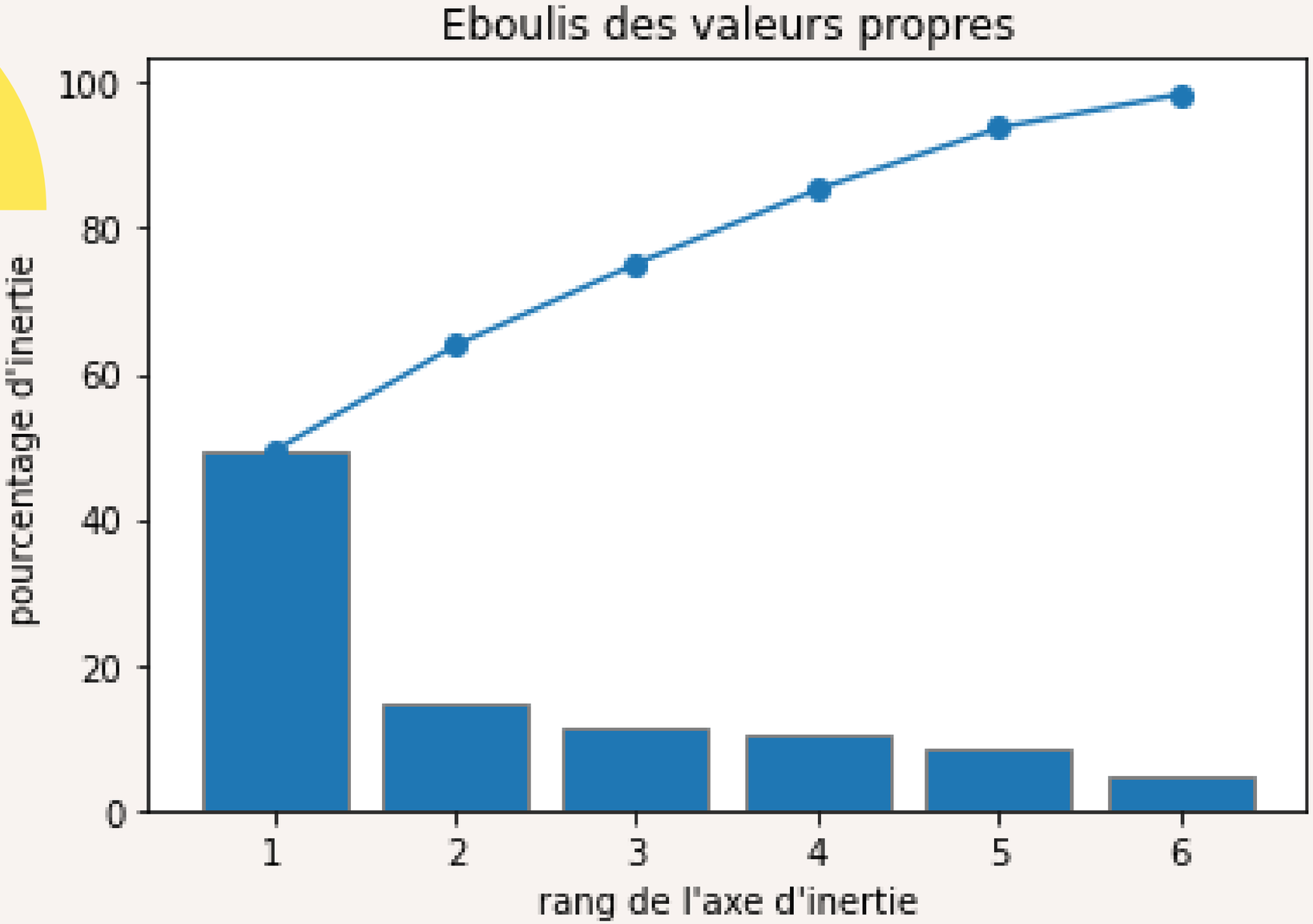


La répartition des dimensions des billets

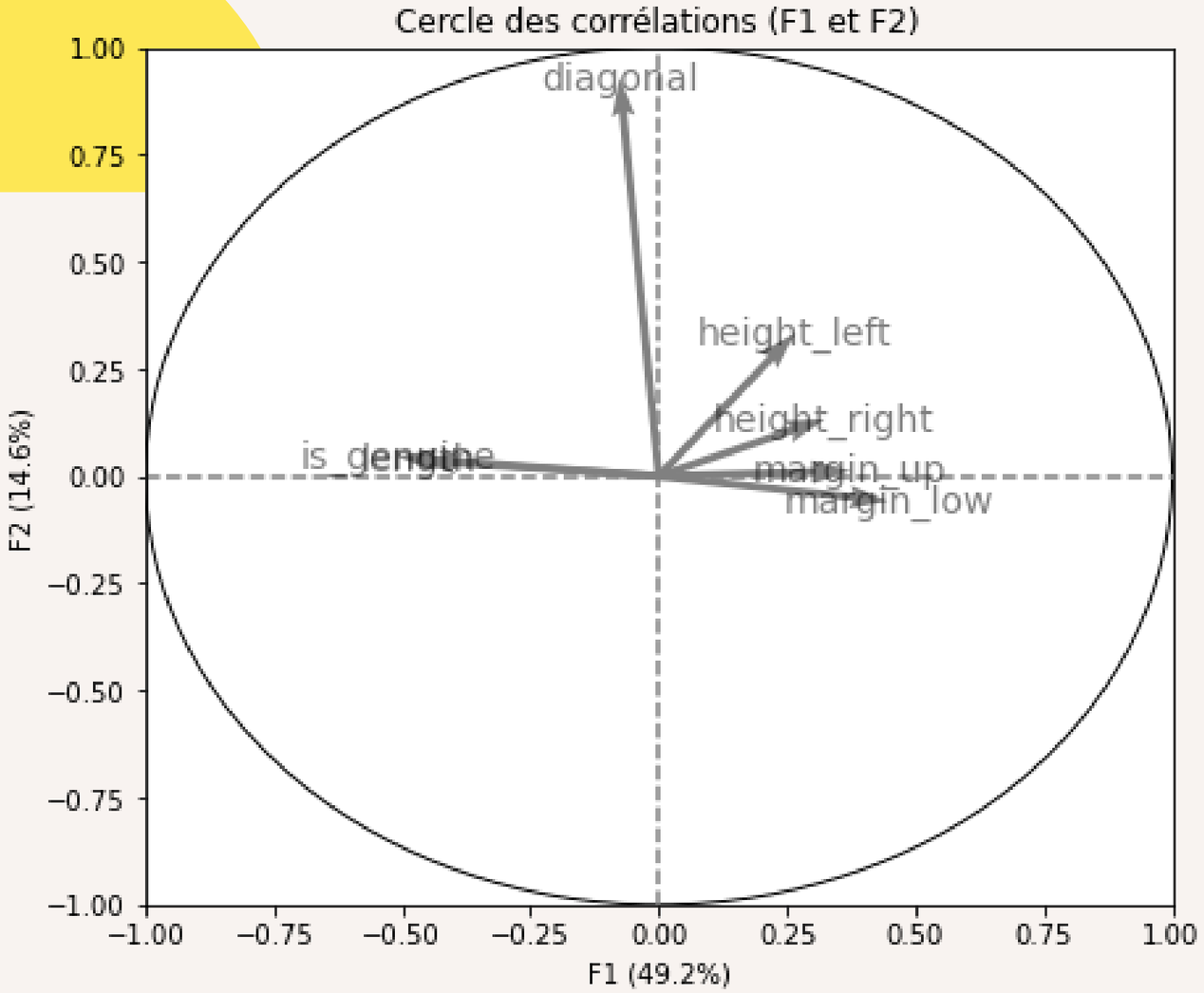




ACP


Eboulis des valeurs propres
pourcentage d'inertie
100
80
60
40
20
0
1
2
3
4
5
6
rang de l'axe d'inertie

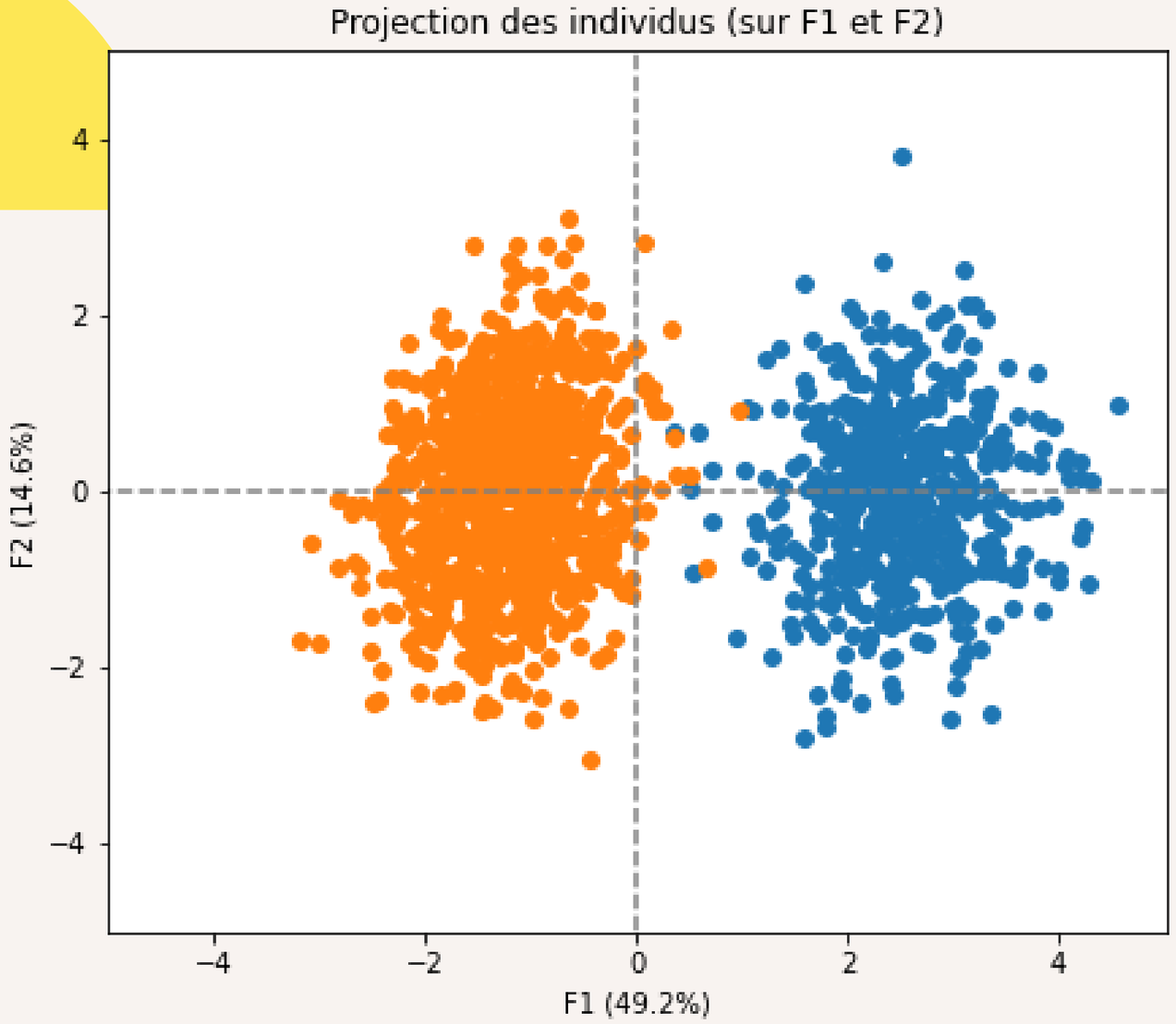
Cercle des corrélations



- Avec le cercle des corrélations F1-F2 (la projection de la flèche sur F1 correspond au coefficient de corrélation), on peut trouver des variables qui sont bien corrélées aux composantes principales:

Les variables les plus corrélées positivement à F1 sont 'height' et 'margin'
Les variables les plus corrélées négativement à F1 est 'length'
Les variables les plus corrélées positivement à F2 est 'diagonal'

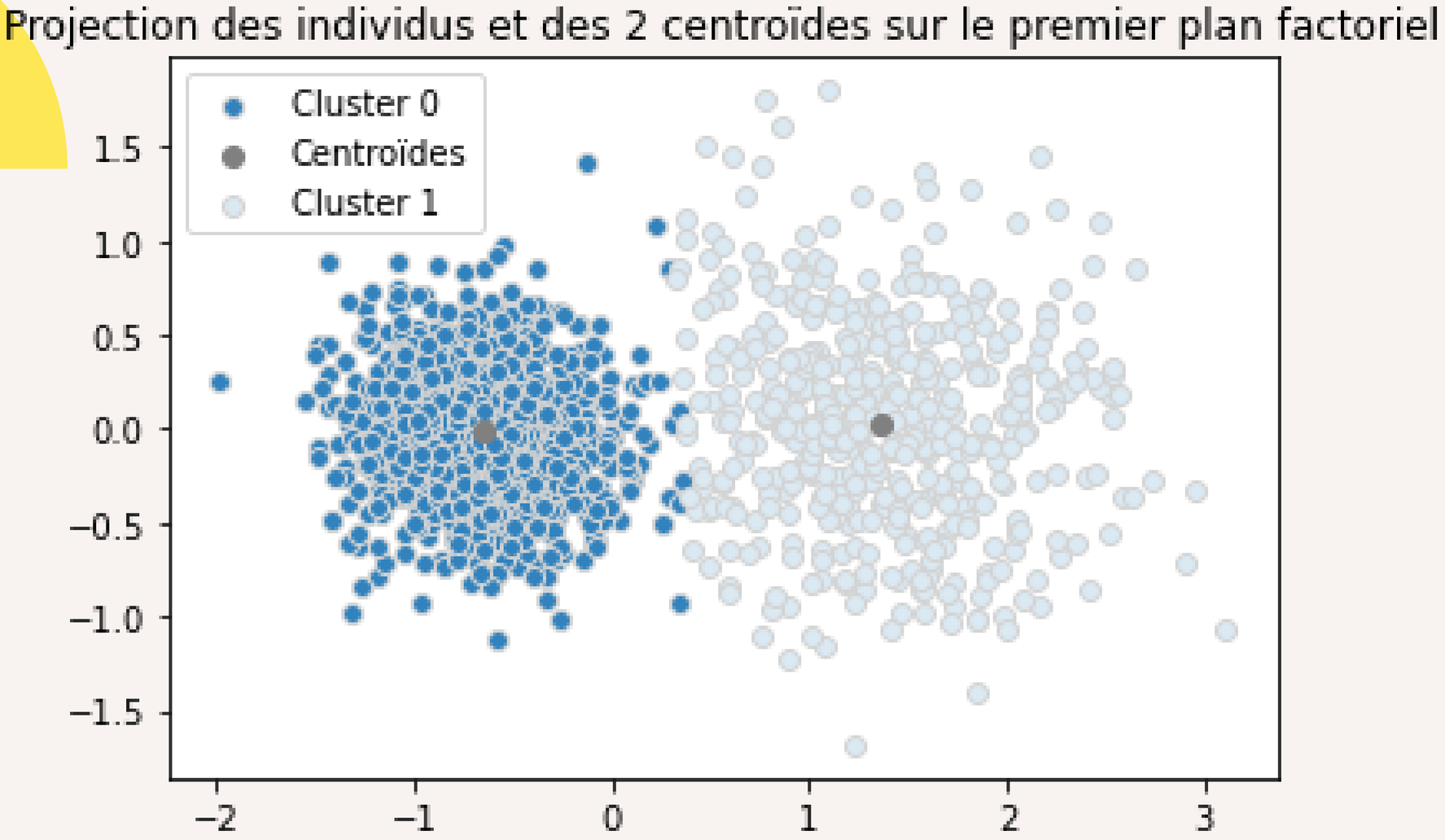
Projection des individus



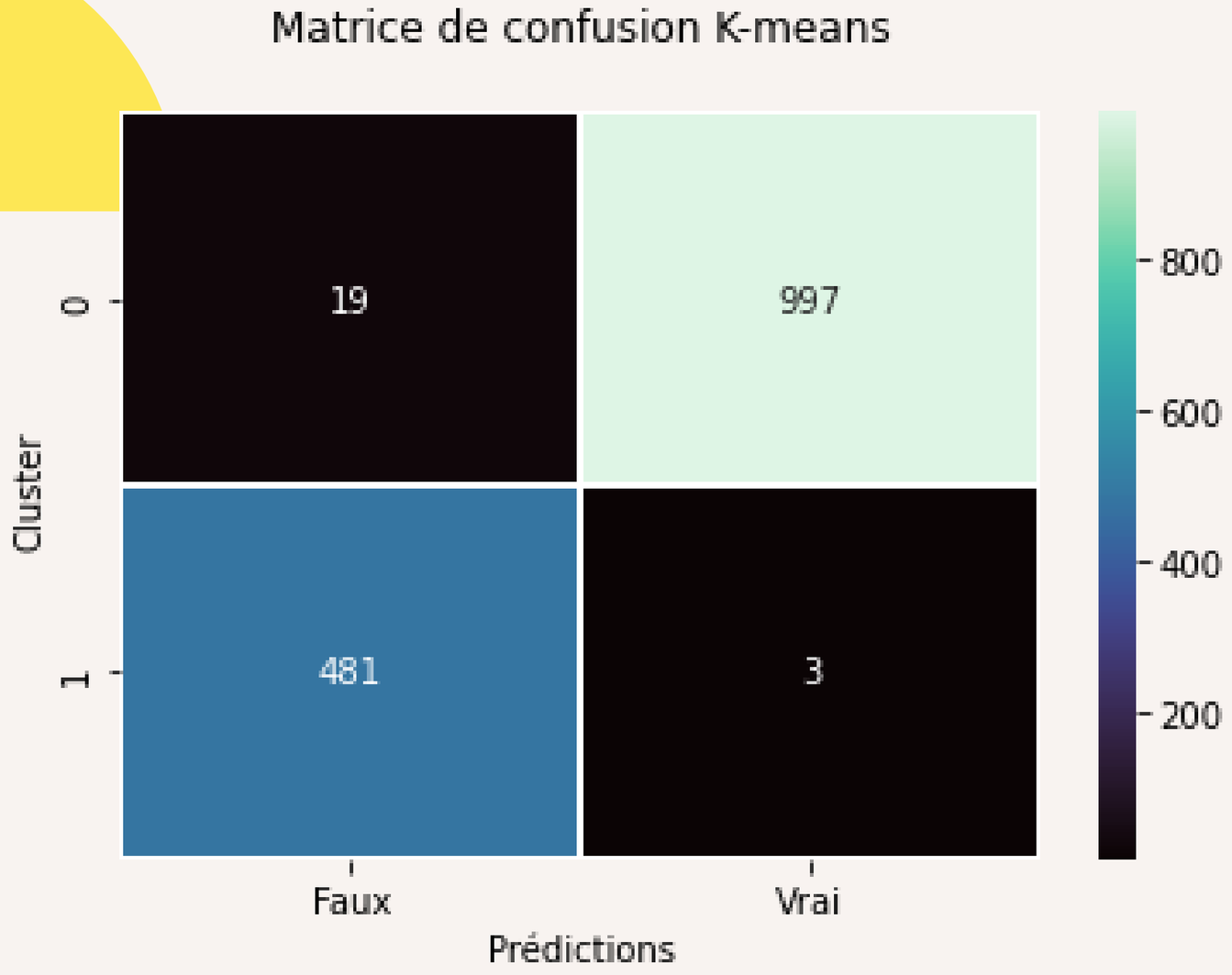
On observe 2 clusters distinctes

Orange:Vrai
Bleu:Faux

K-Means

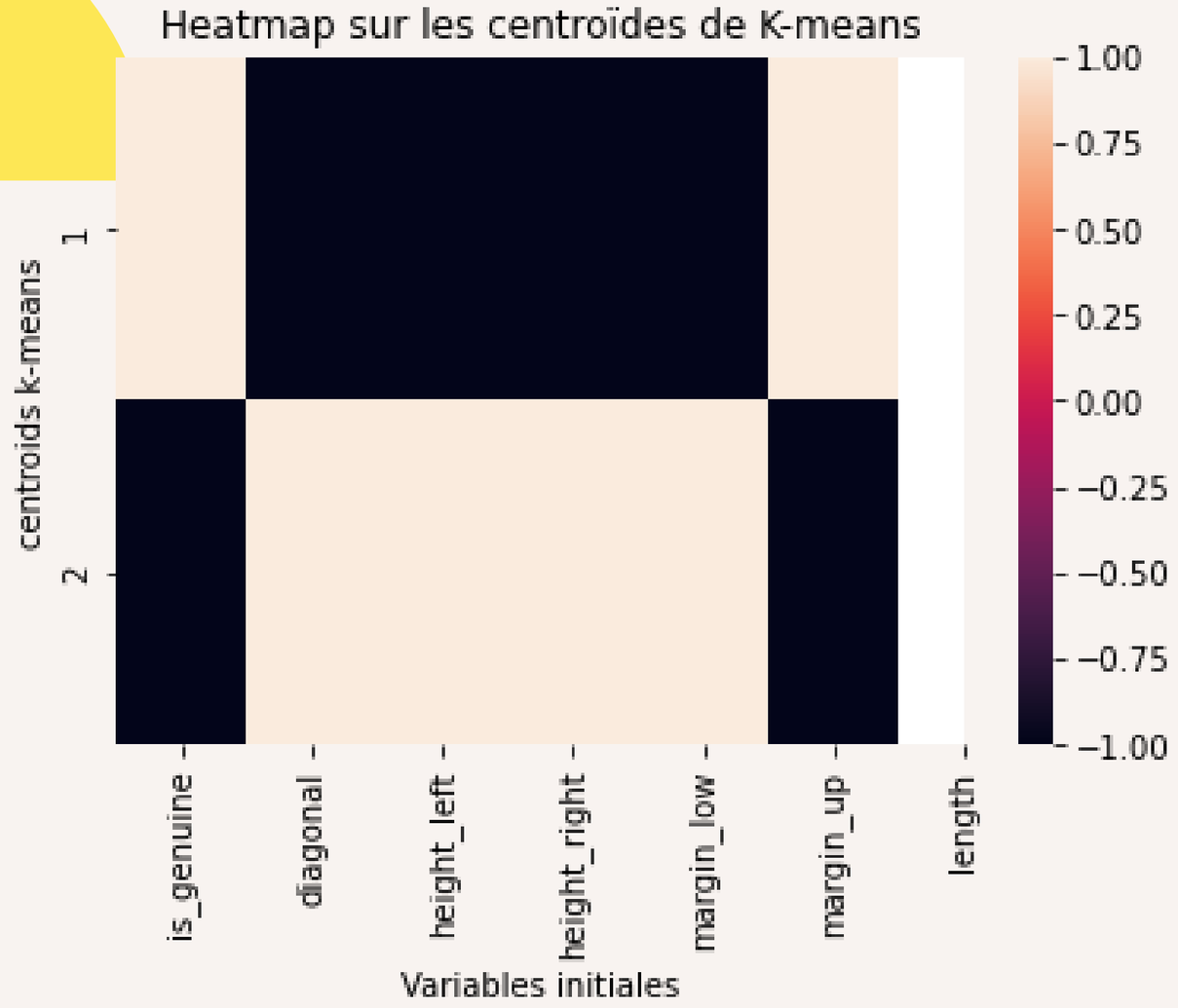

Projection des individus et des 2 centroïdes sur le premier plan factoriel
Cluster 0
Centroïdes
Cluster 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2
3

Matrice de confusion K-means



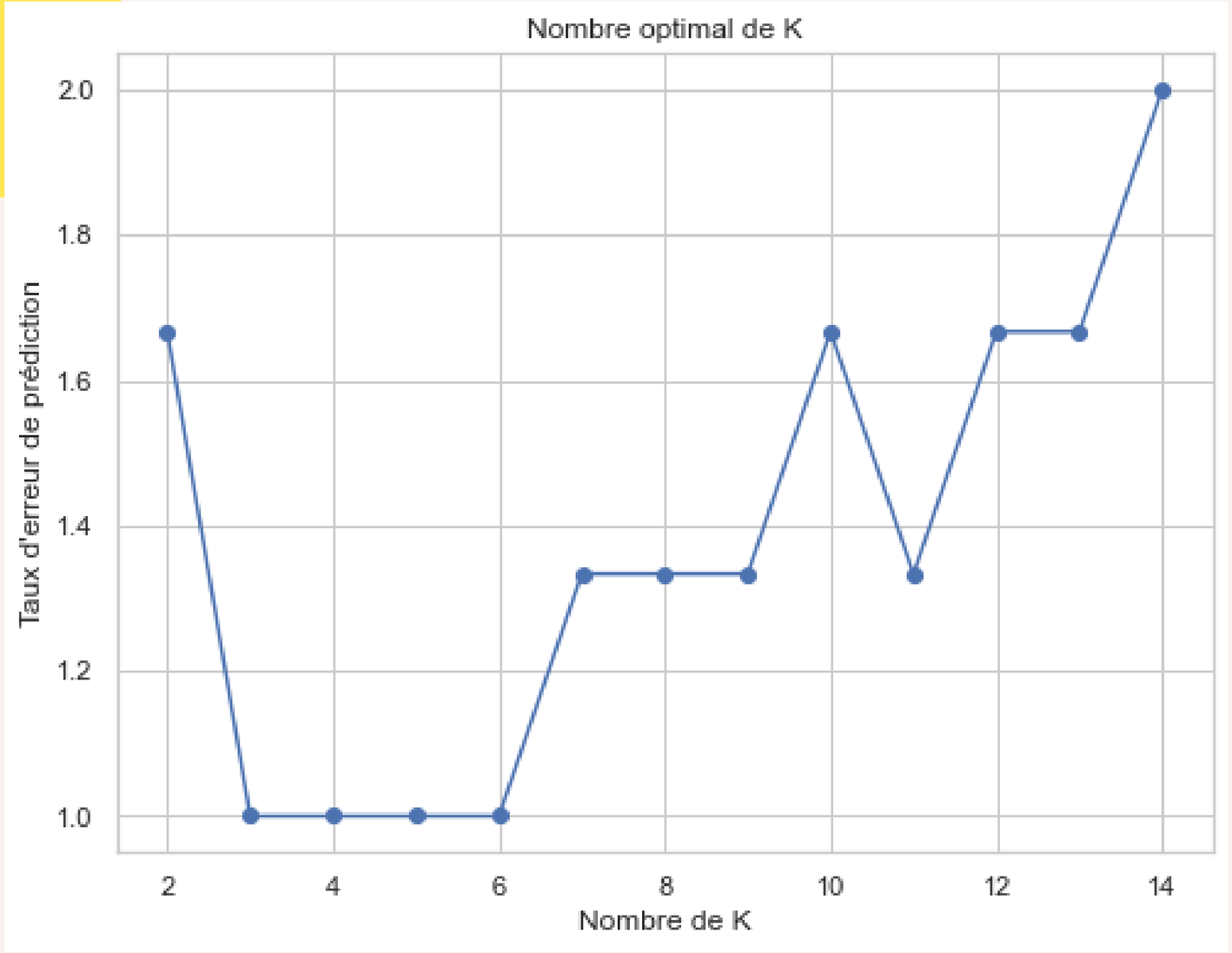
Vrais positifs : 997
Vrais négatifs : 481
Faux positifs : 3
Faux négatifs : 19

Heatmap sur les centroïdes de K-means



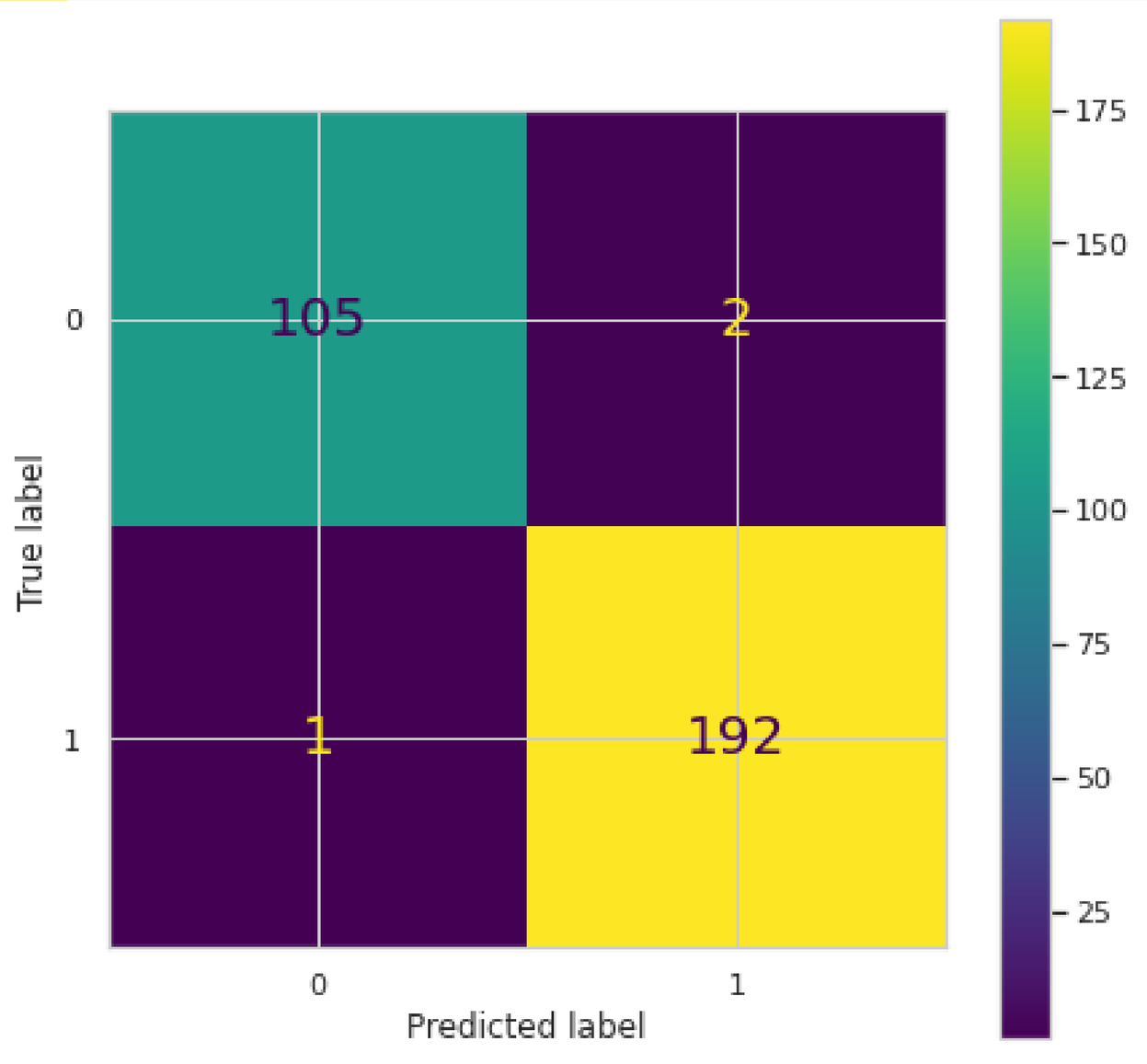
Longueur mis à part, on remarque un très fort contraste entre les centroïdes des clusters pour chaque donnée.

KNN


Nombre optimal de K
Taux d'erreur de prédiction
2.0
1.8
1.6
1.4
1.2
1.0
2
4
6
8
10
12
14
Nombre de K

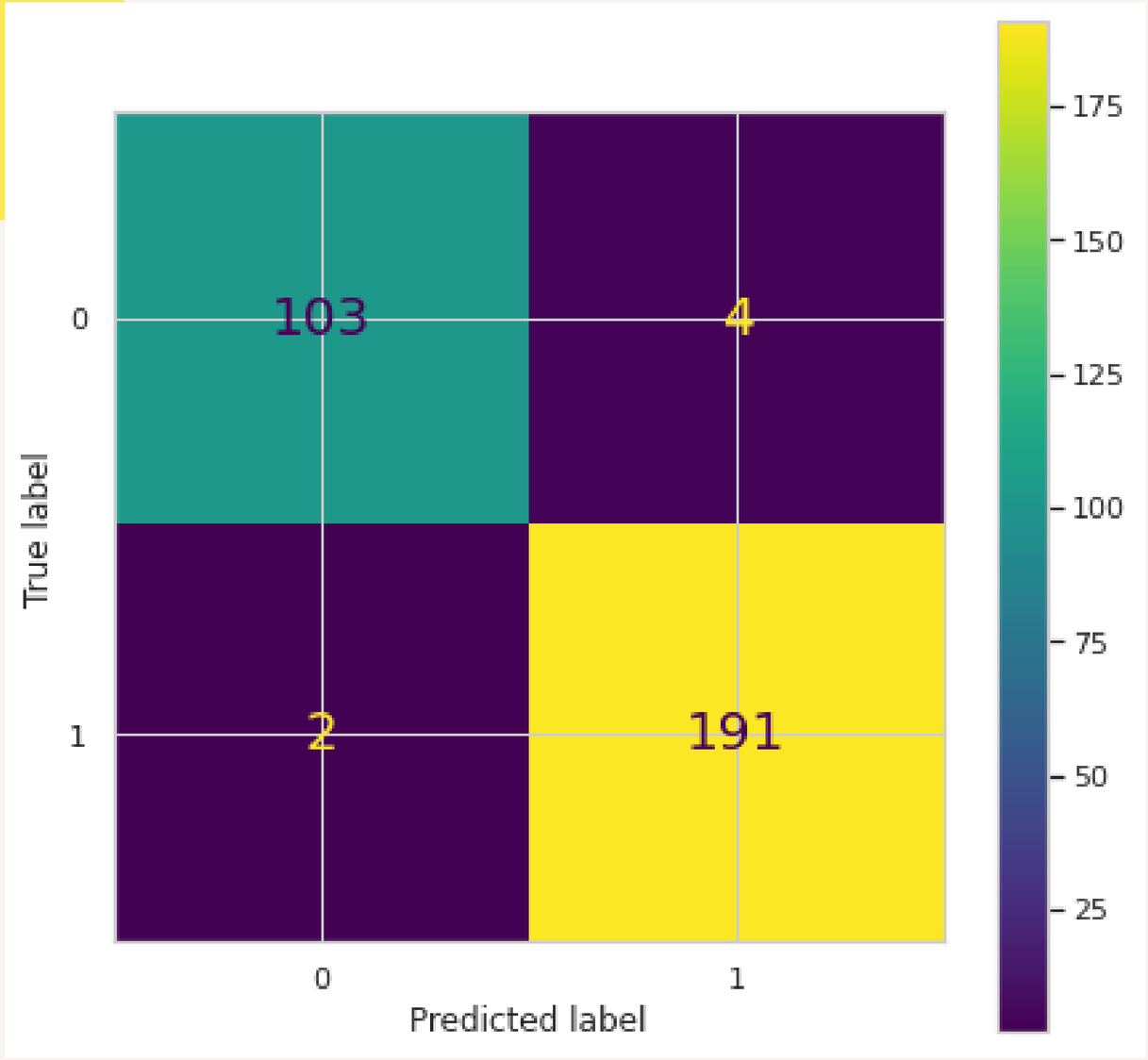
K=3

Matrice de confusion (KNN)



Vrais positifs : 192
Vrais négatifs : 105
Faux positifs : 2
Faux négatifs : 1

Matrice de confusion (Regression logistique)



Vrais positifs : 191
Vrais négatifs : 103
Faux positifs : 4
Faux négatifs : 2

Donc nous avons 2 vrai billet détecté comme un faux, et 4 faux billets détectés comme des vrais billets.

Test de l’algorithme

```
def verif_billet_rl(csv):
    billet_test= pd.read_csv(csv)
    billet_value=billet_test.drop('id', axis=1)
    y_pred = model_logit.predict(billet_value)
    proba_true = model_logit.predict_proba(billet_value)[:, 1]
    billet_test['Prediction'] = y_pred
    billet_test['Probability_is_true'] = proba_true.round(3)
    billets_predict_rl = billet_test[['id','Prediction','Probability_is_true']].set_index("id")
    return billets_predict_rl
```

Thank You