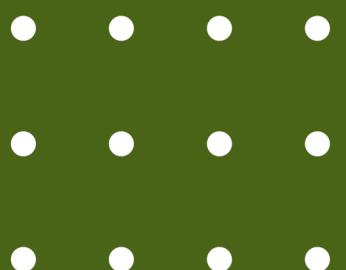


Détection de faux billets



Point abordé pendant la présentation

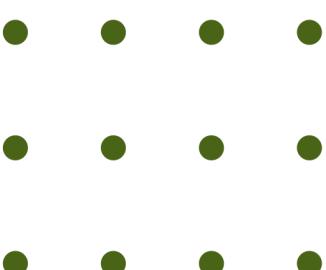


- Contexte du projet
- Observation des données
- Analyse descriptive des données
 - Répartitions des billets
 - Distribution des variables
 - Pairplot sur la variable vrai/faux billet
 - Variance entre les billets vrais/faux par rapport aux autres variables
 - Corrélation entre les différentes variables
- Traitement des valeurs manquantes
 - Vérification par test
 - Test de la normalité des résidus
 - Test de l'homoscédasticité des résidus
 - Test de colinéarité
 - Insertion des données
- Méthode de prédiction
 - Kmeans
 - ACP
 - Éboulis des valeurs propres
 - Cercle des corrélations
 - Projection sur le plan factoriel
 - Métrique de l'inertie
 - Matrice de confusion
 - Régression logistique
 - Matrice de confusion
 - Choix du modèle

Contexte du projet

Mettre en place une modélisation qui permet d'identifier automatiquement les vrais des faux billets, à partir des dimensions du billet et des éléments qui le composent.

Un fichiers est mis à disposition qui contient les informations sur les dimensions des billets.



Observation des données

| | is_genuine | diagonal | height_left | height_right | margin_low | margin_up | length |
|------|------------|----------|-------------|--------------|------------|-----------|--------|
| 0 | True | 171.81 | 104.86 | 104.95 | 4.52 | 2.89 | 112.83 |
| 1 | True | 171.46 | 103.36 | 103.66 | 3.77 | 2.99 | 113.09 |
| 2 | True | 172.69 | 104.48 | 103.50 | 4.40 | 2.94 | 113.16 |
| 3 | True | 171.36 | 103.91 | 103.94 | 3.62 | 3.01 | 113.51 |
| 4 | True | 171.73 | 104.28 | 103.46 | 4.04 | 3.48 | 112.54 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1495 | False | 171.75 | 104.38 | 104.17 | 4.42 | 3.09 | 111.28 |
| 1496 | False | 172.19 | 104.63 | 104.44 | 5.27 | 3.37 | 110.97 |
| 1497 | False | 171.80 | 104.01 | 104.12 | 5.51 | 3.36 | 111.95 |
| 1498 | False | 172.06 | 104.28 | 104.06 | 5.17 | 3.46 | 112.25 |
| 1499 | False | 171.47 | 104.15 | 103.82 | 4.63 | 3.37 | 112.07 |

1500 rows × 7 columns

Il n'y a pas de problème avec les valeurs numériques.

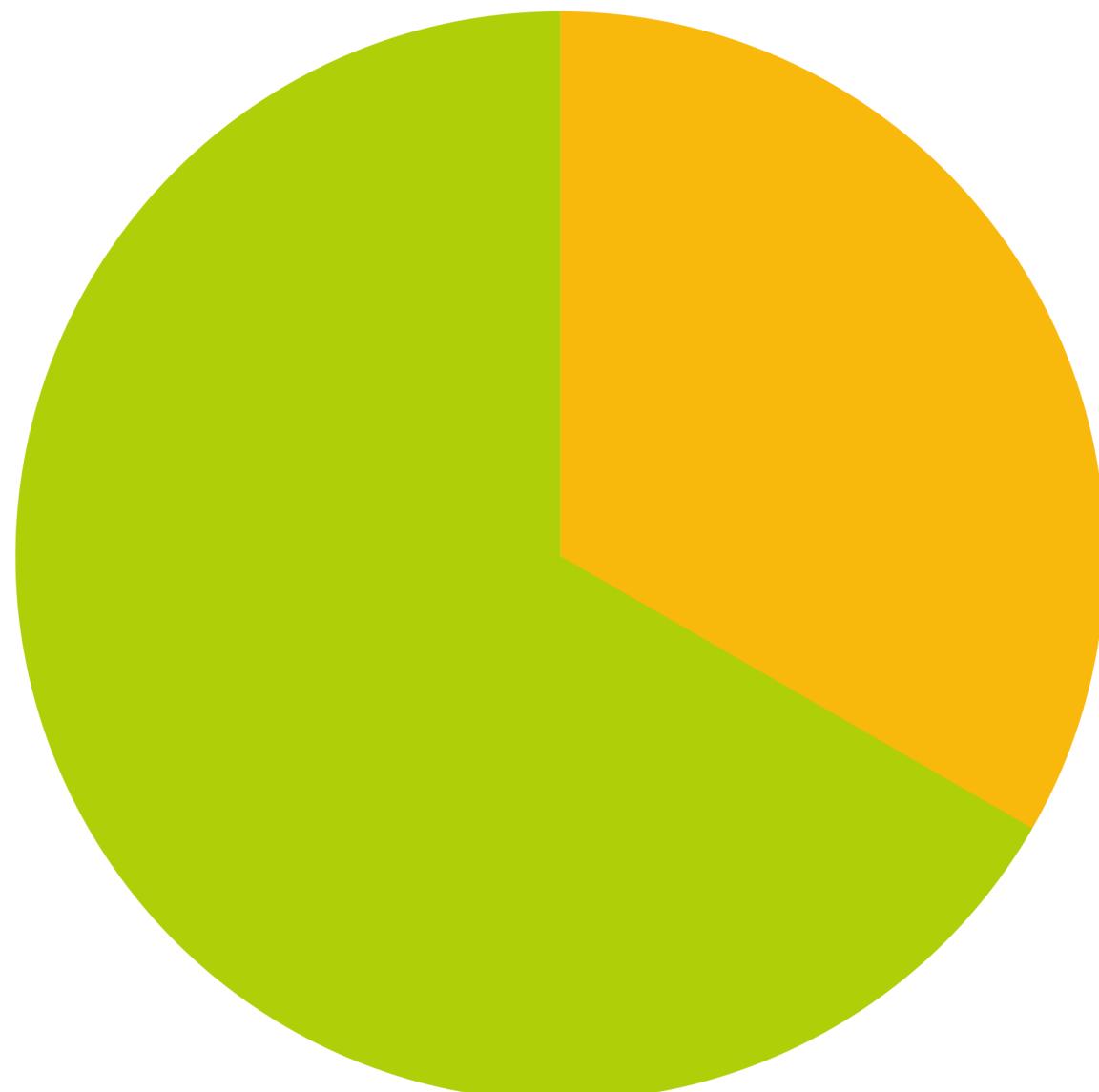
La colonne is_genuine comporte bien les informations 'True' et 'False'.

Il manque des données pour la colonne margin_up (37 données manquantes), elles vont être traiter plus tard.

Répartitions des billets

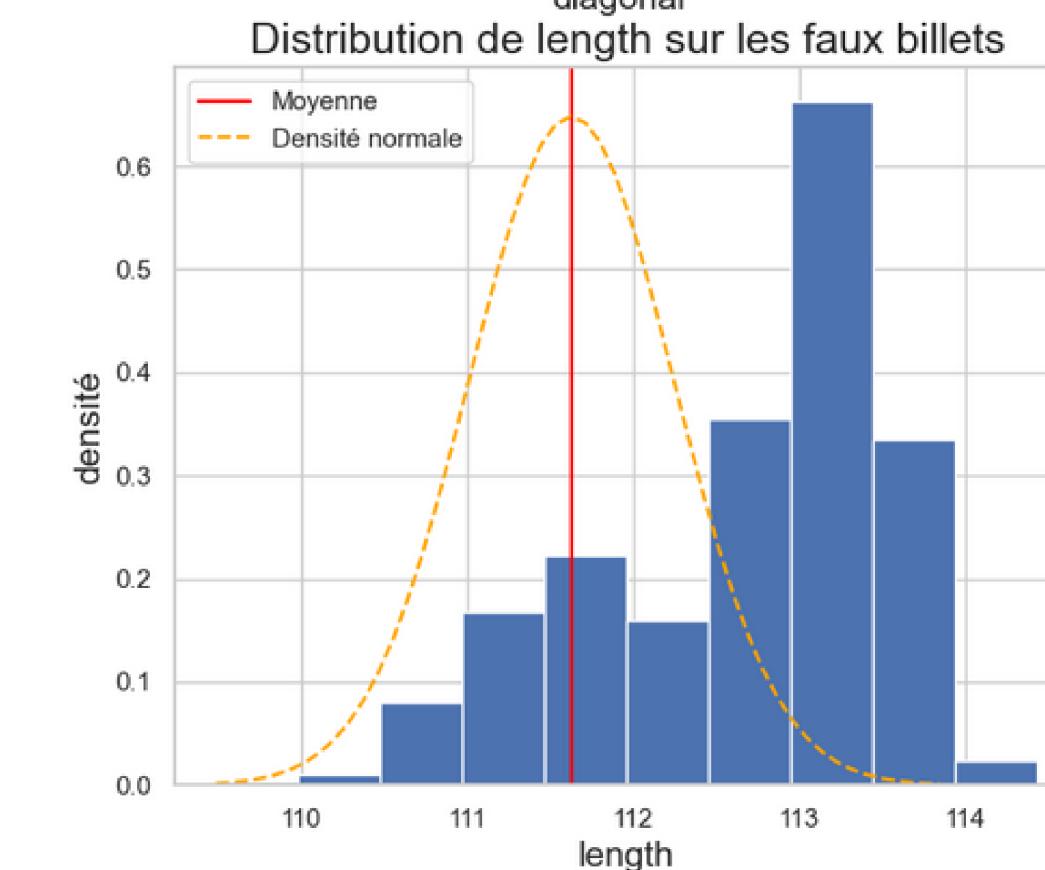
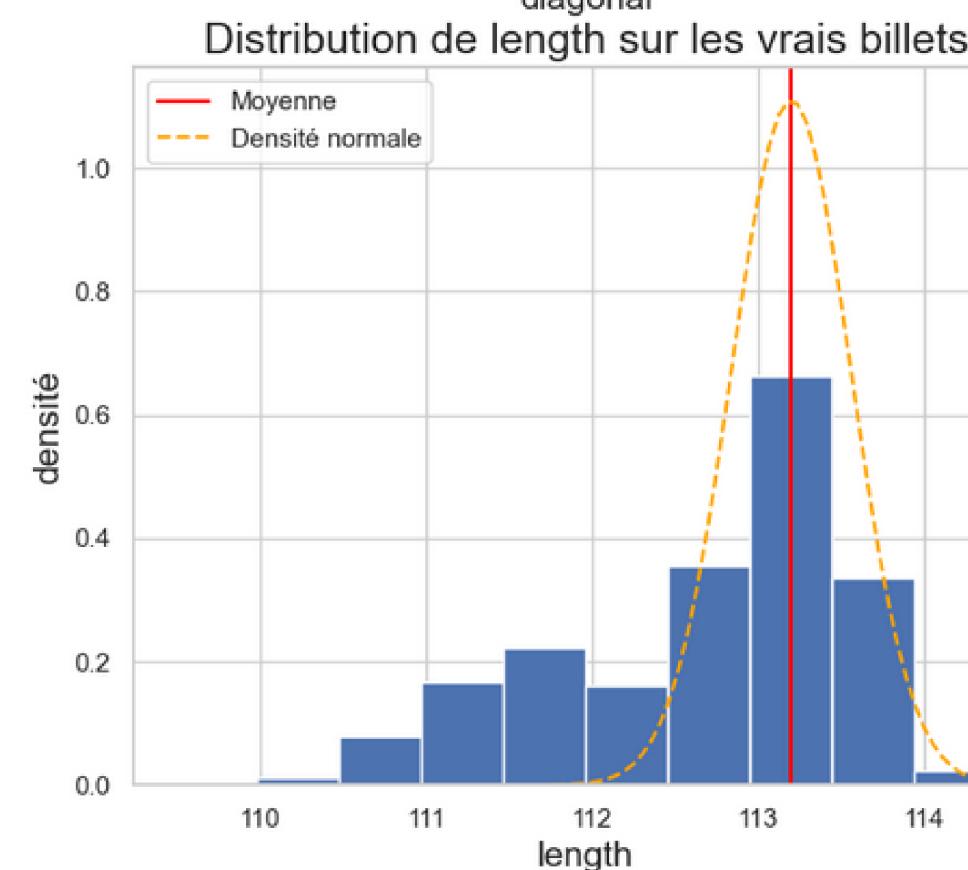
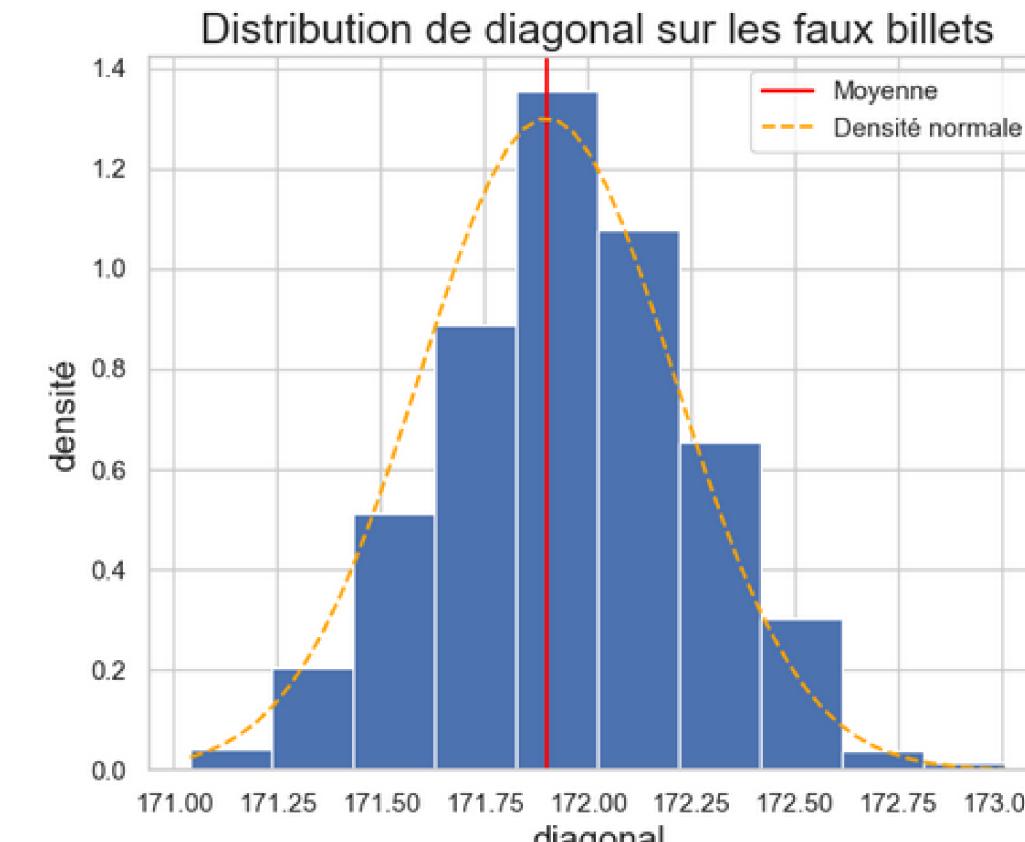
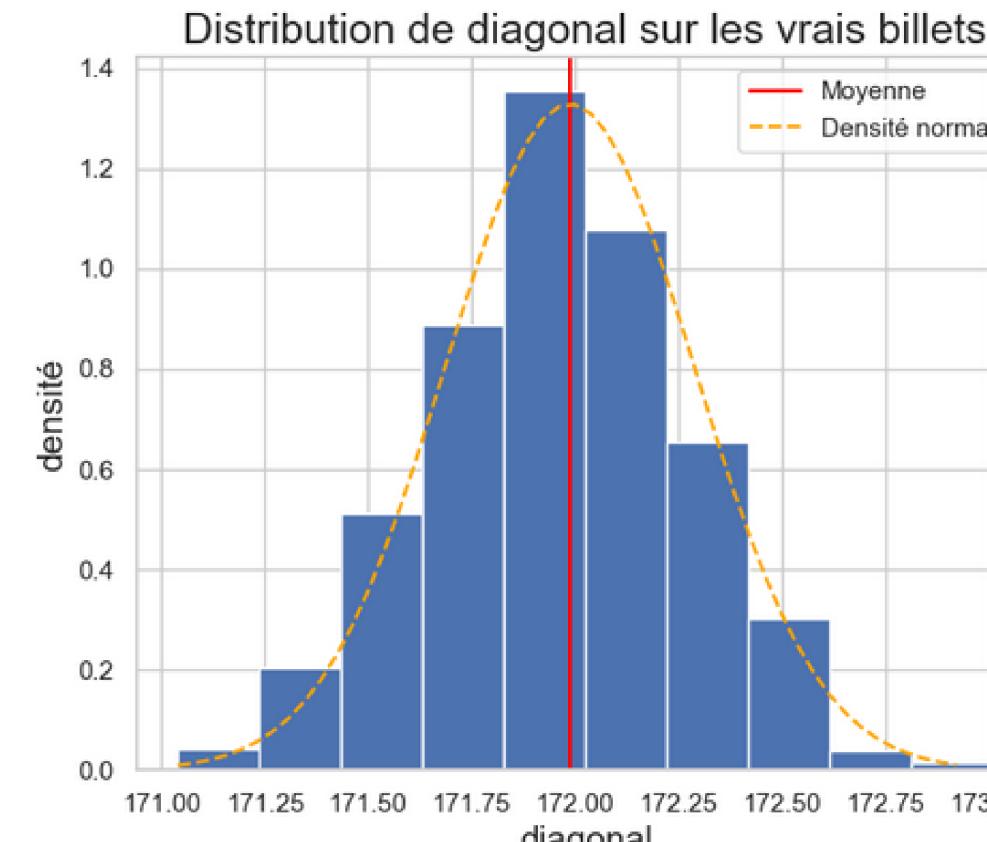
Il y a 1 500 billets au total, 1 000 représentent les vrais billets (66.7 %) et 500 billets sont des faux (33.3%).

Vrais billets
66.7%



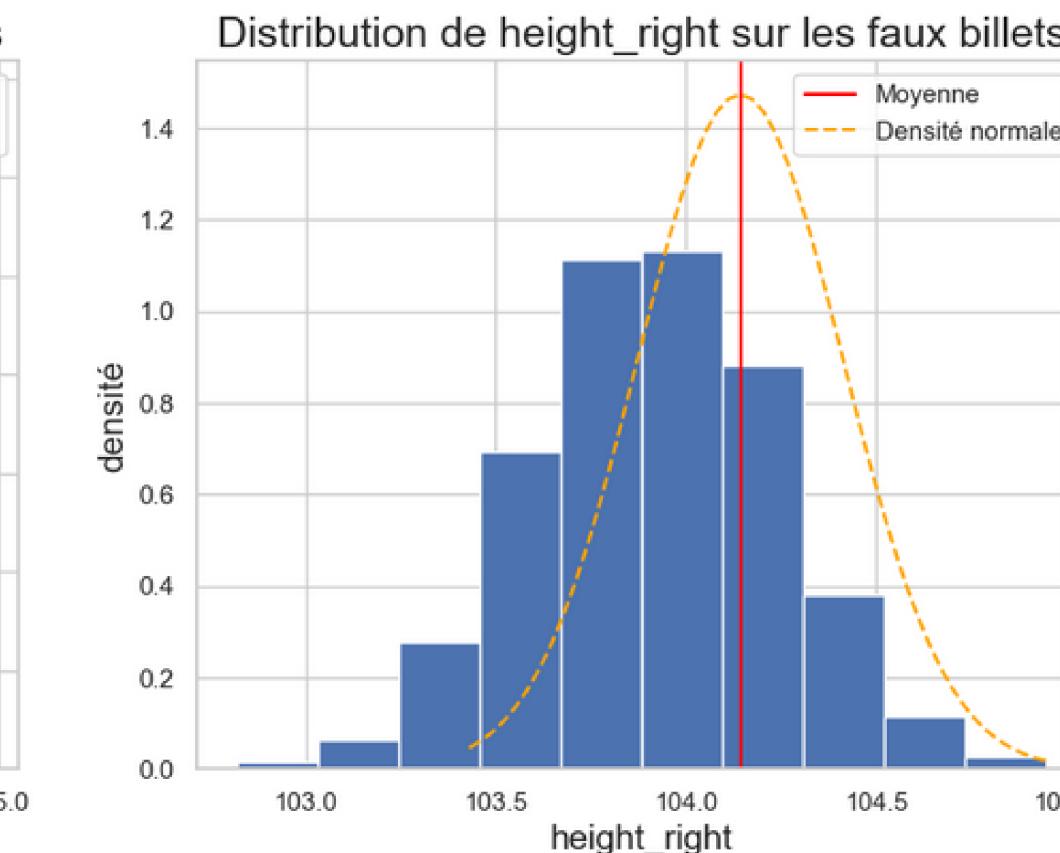
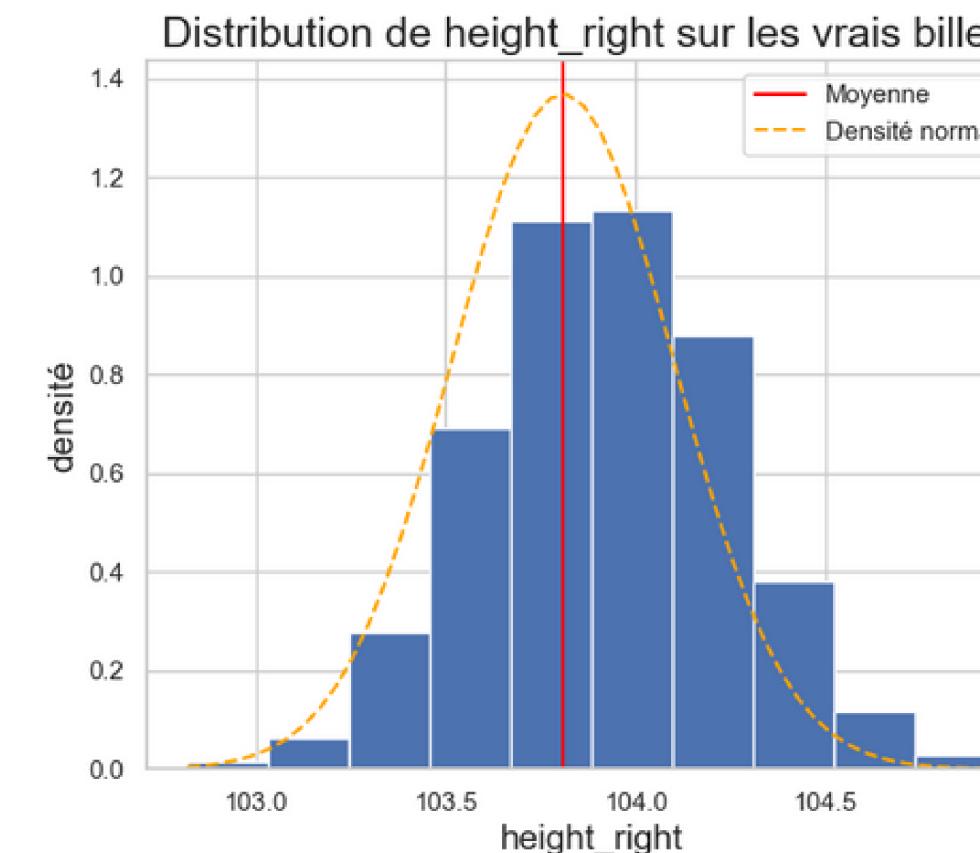
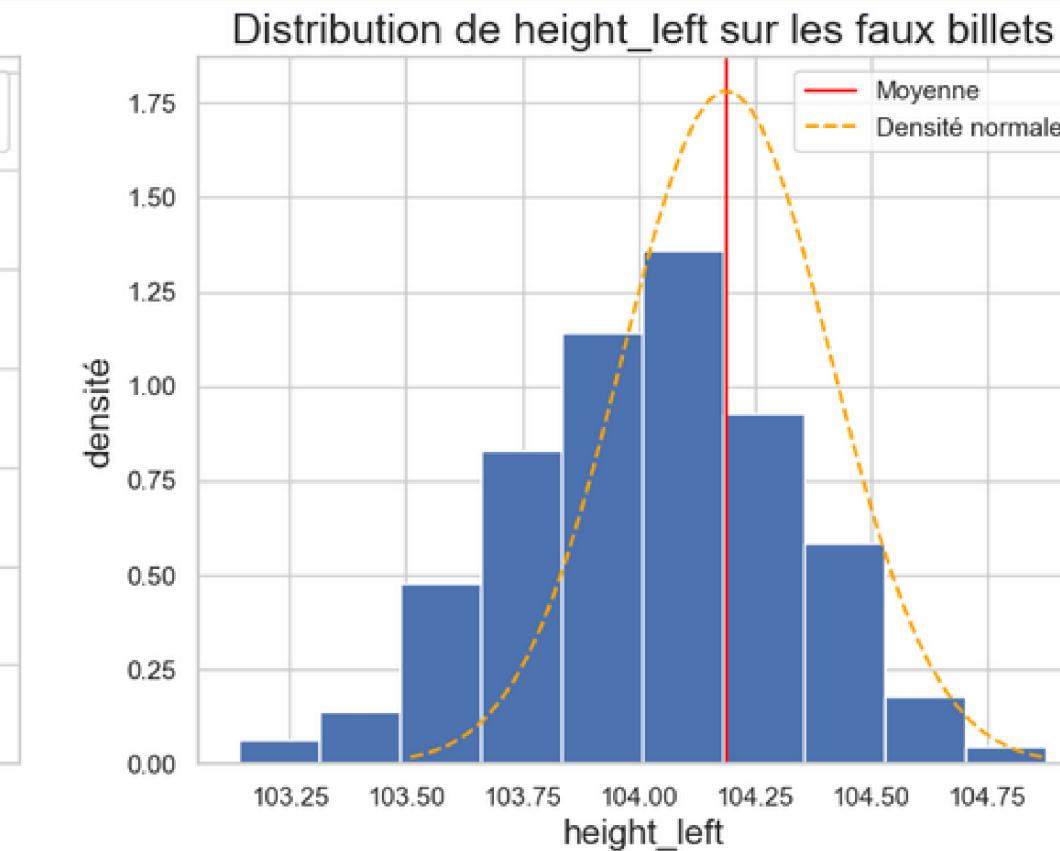
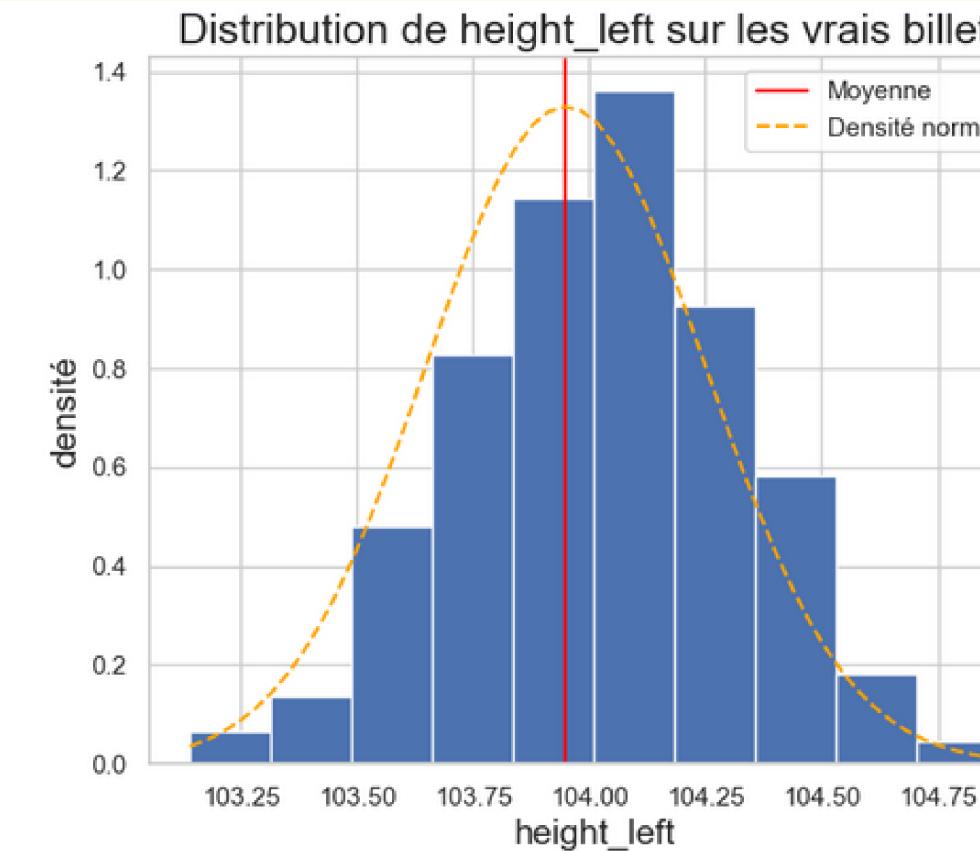
Faux billets
33.3%

Distribution des variables



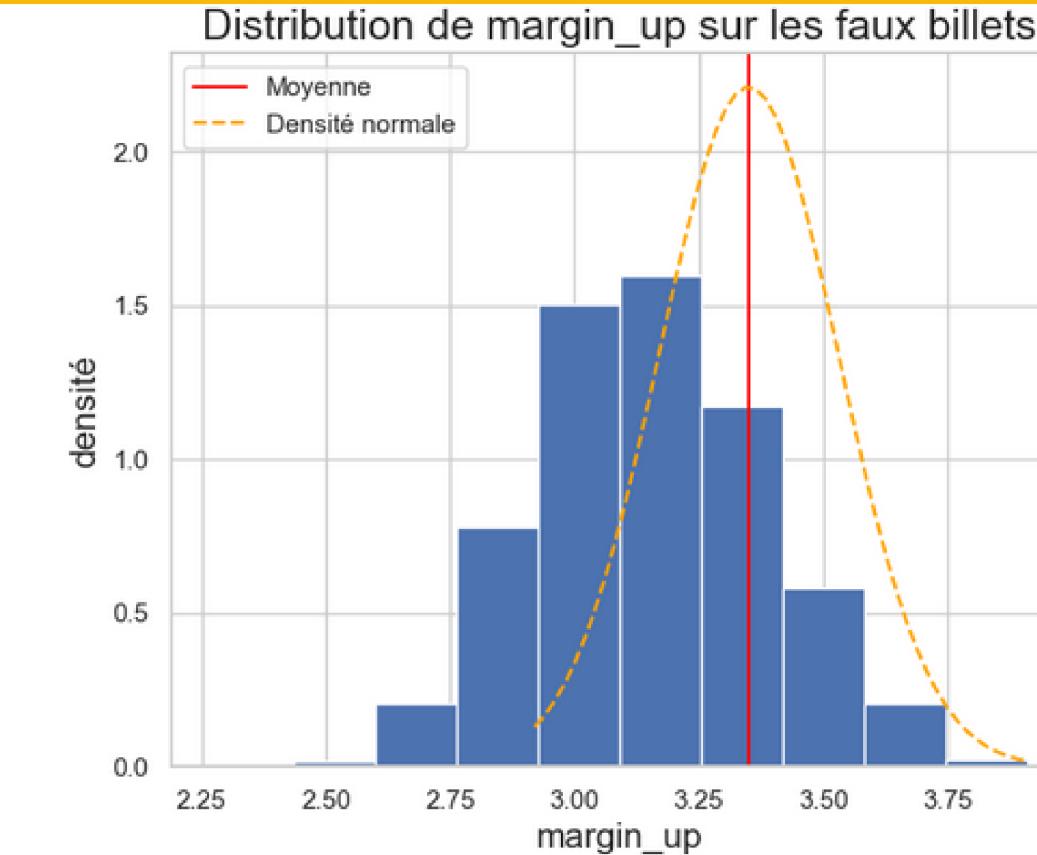
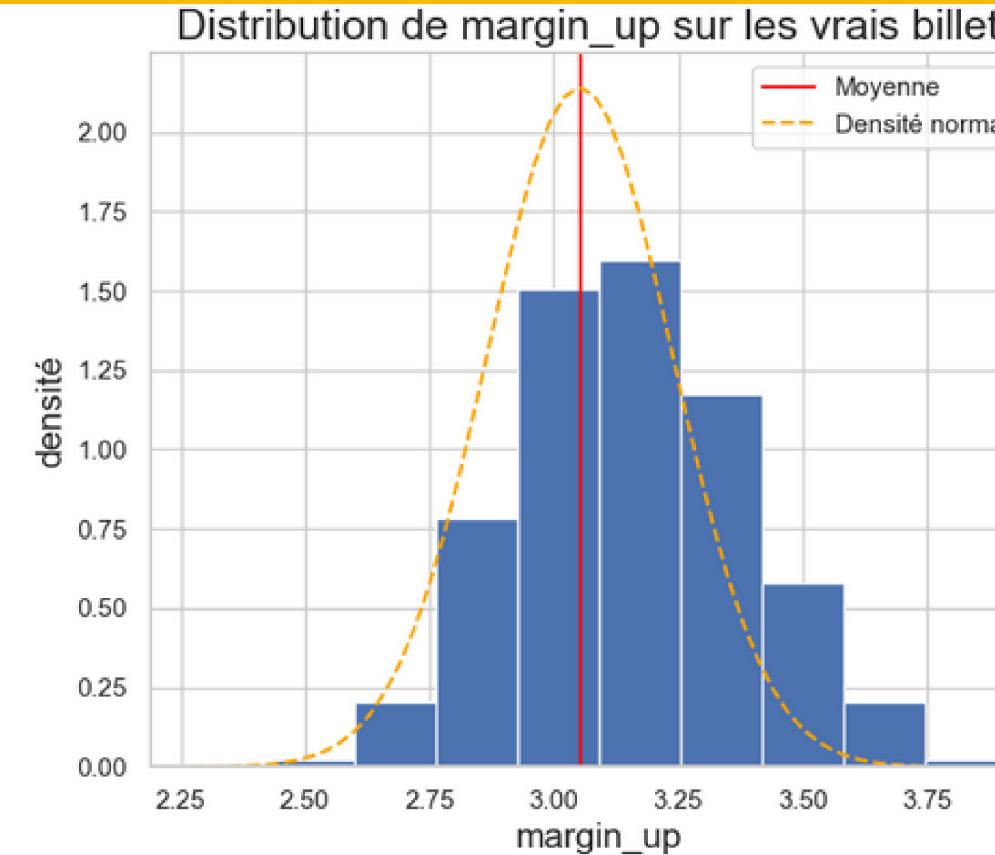
Cette distribution ne
suit pas la loi
gaussienne

Distribution des variables

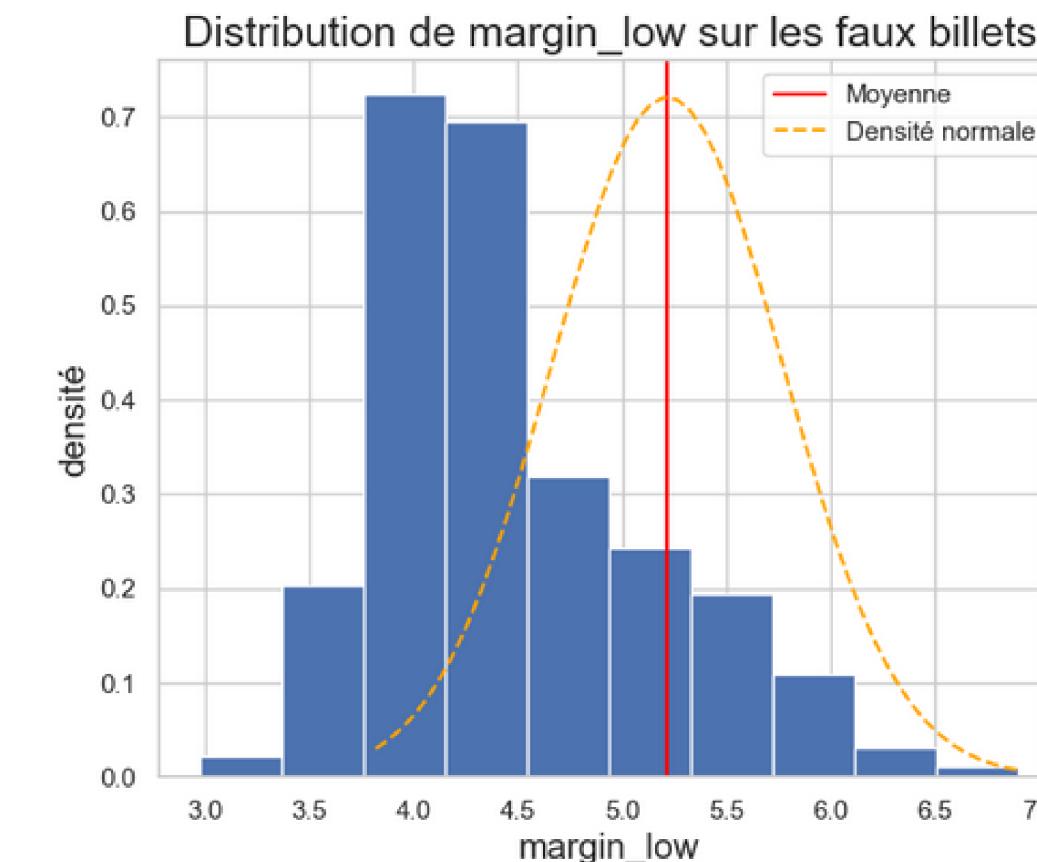
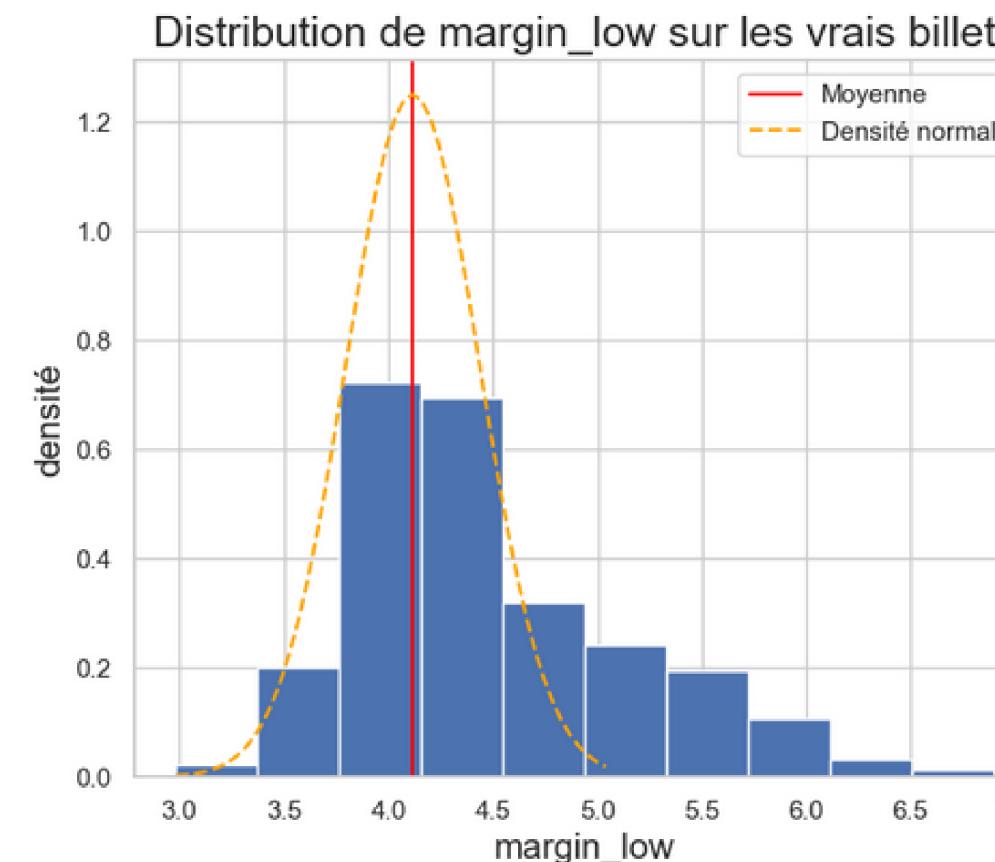


Densité pas normale

Distribution des variables

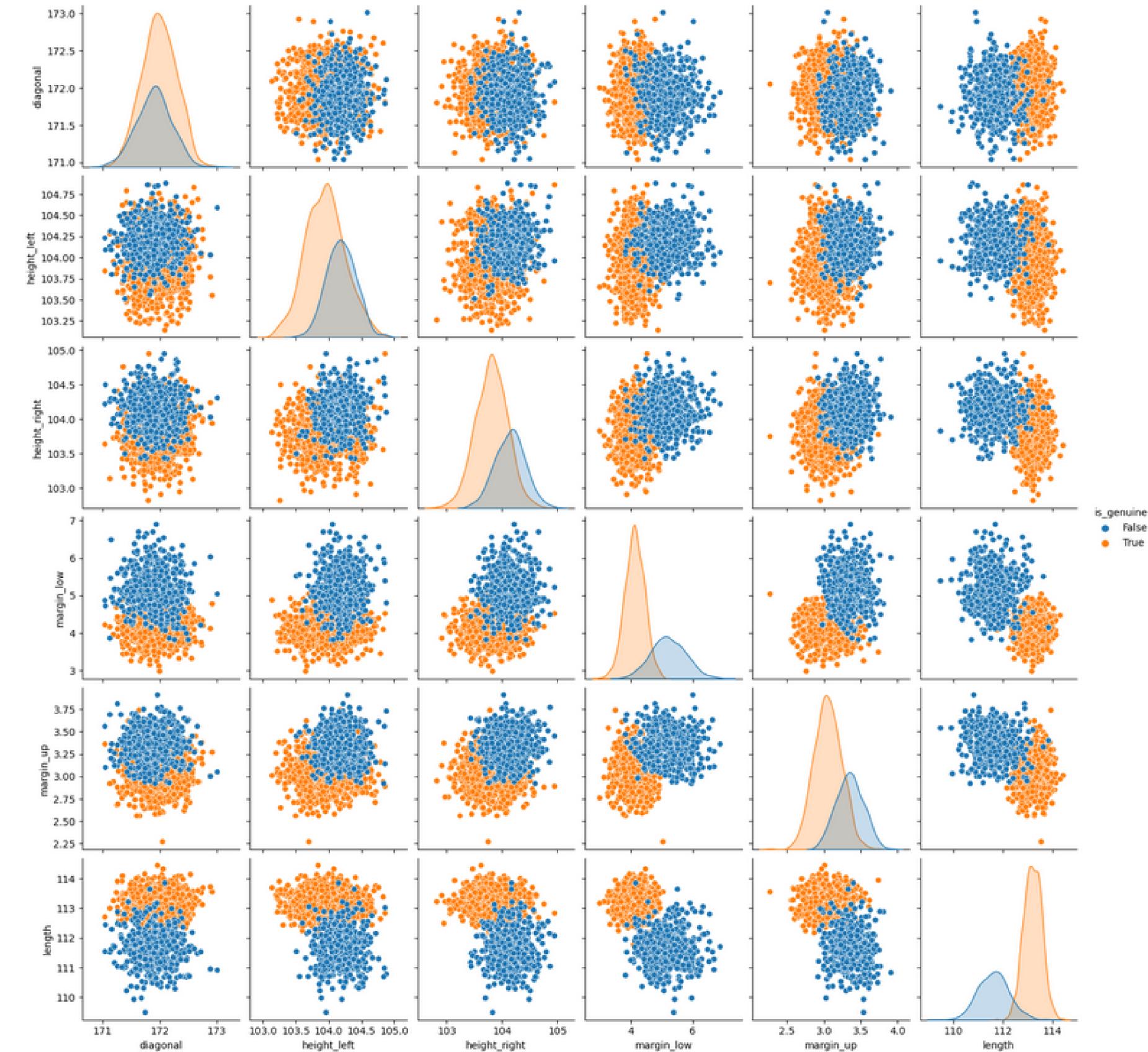


Densité pas normale



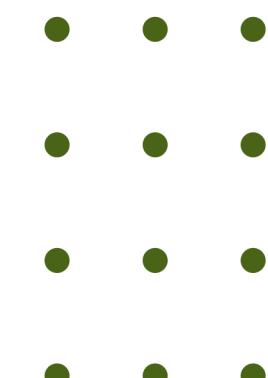
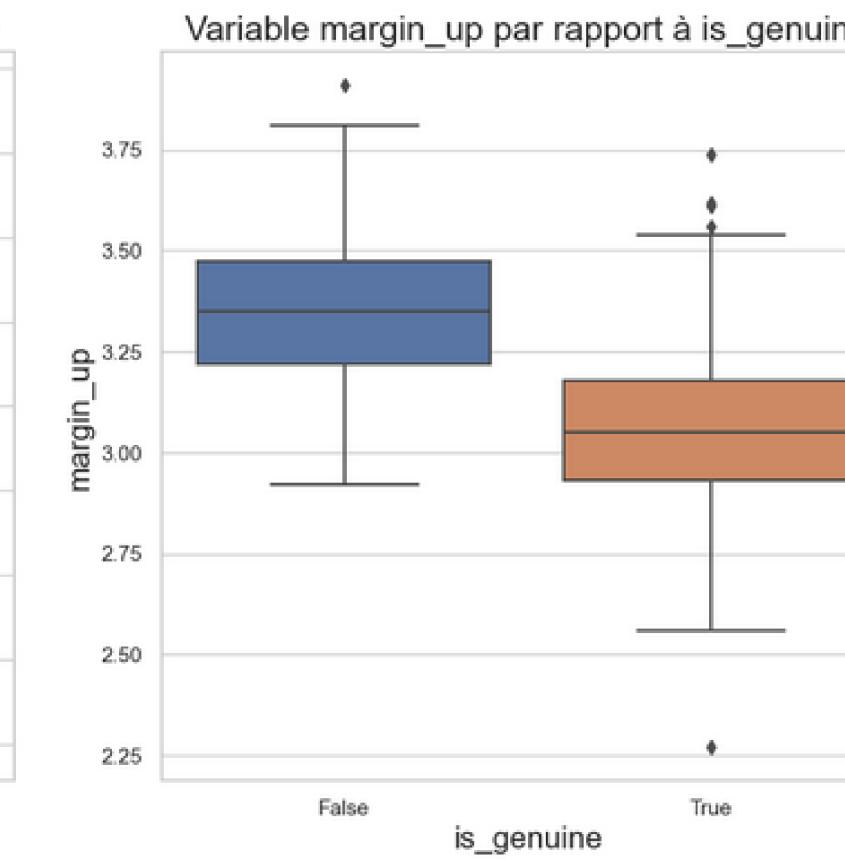
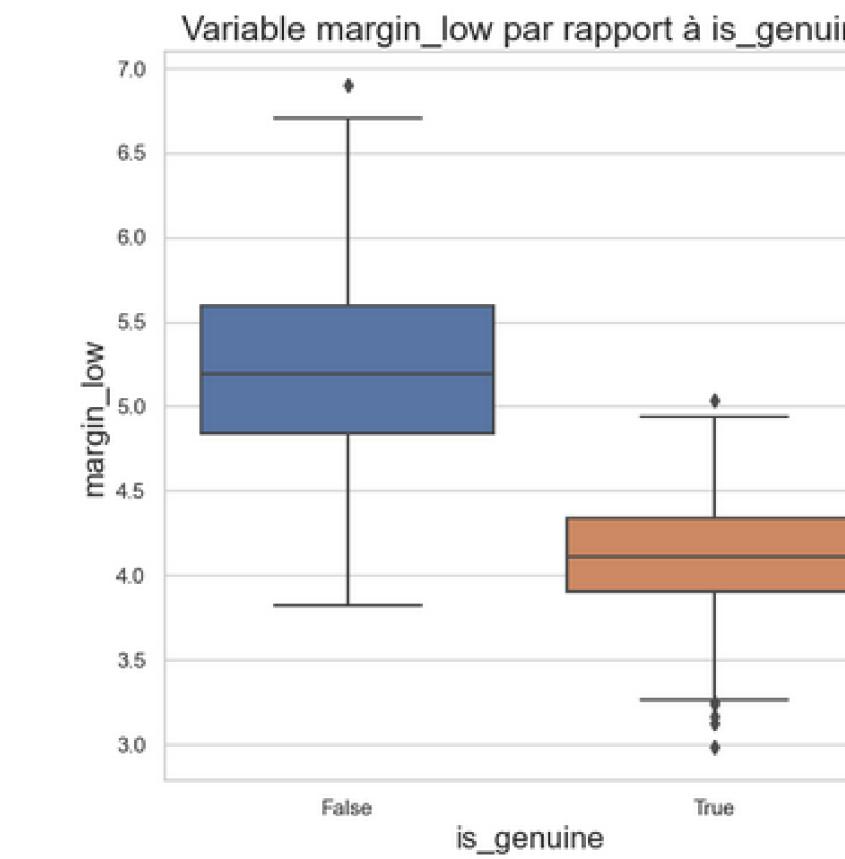
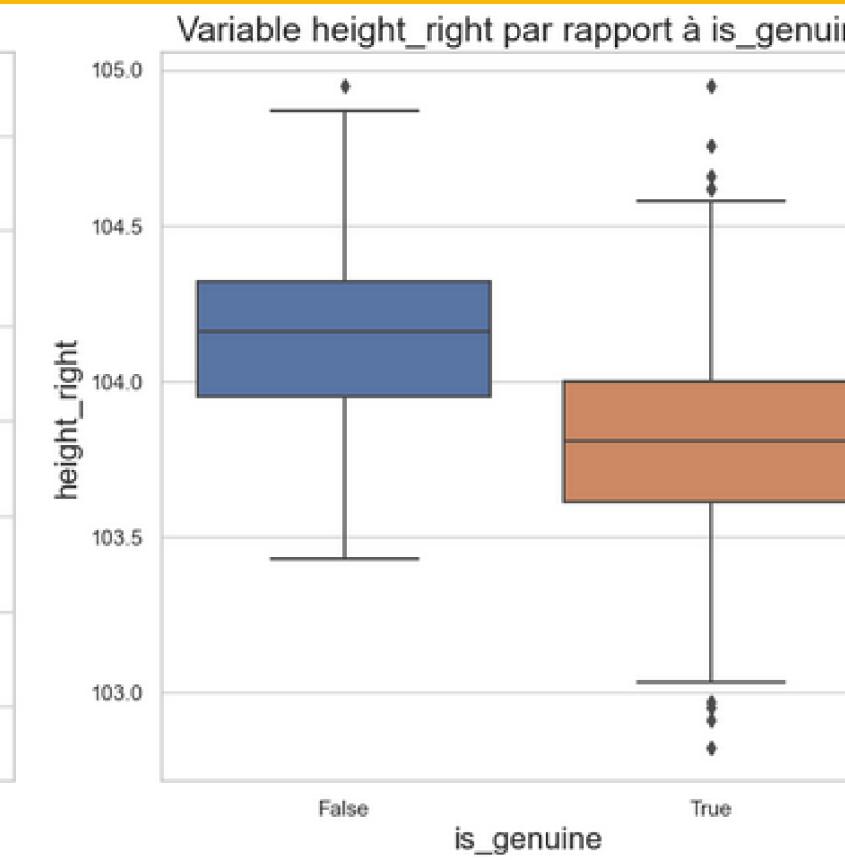
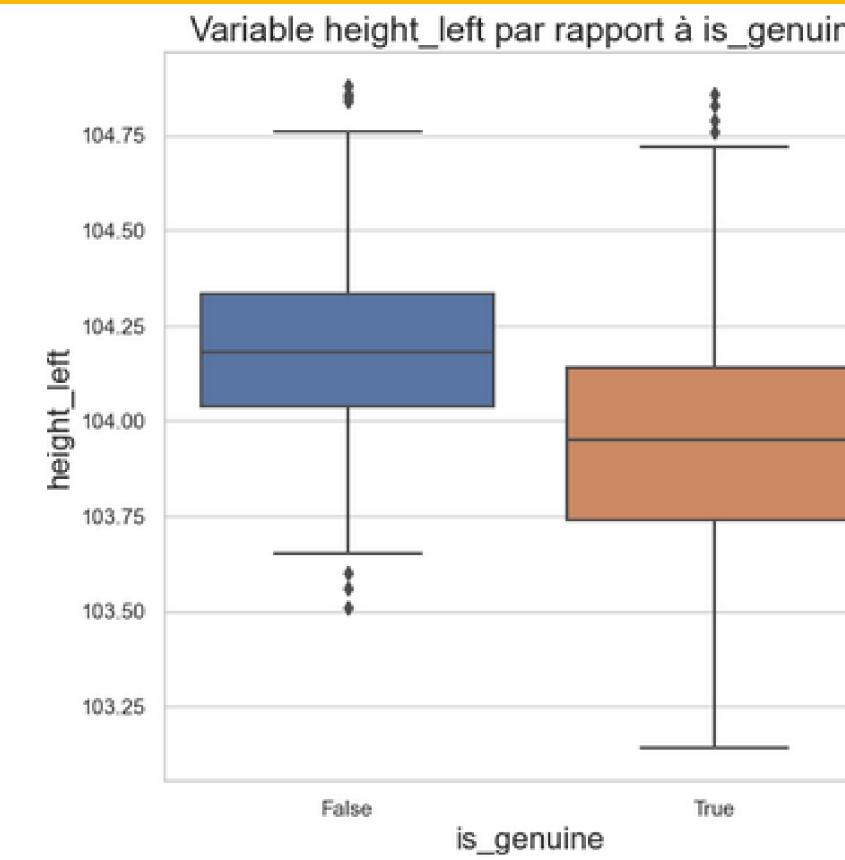
Cette distribution ne
suit pas la loi
gaussienne

Pairplot sur la variable vrai/faux billet

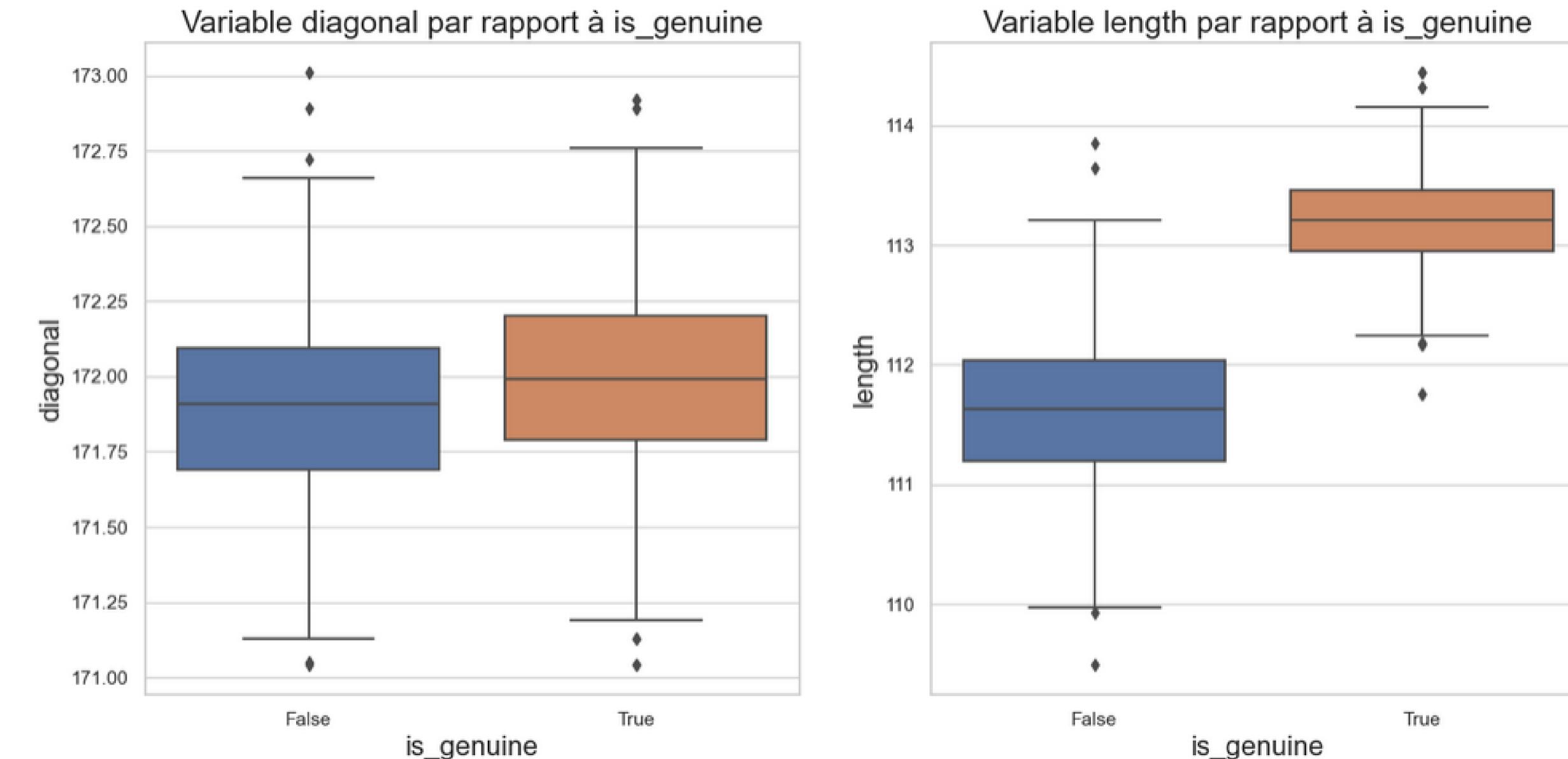


Avec ce pairplot, on remarque bien la séparation entre les vrais et les faux billets pour chaque variable.

Variance entre les billets vrais/faux par rapport aux autres variables

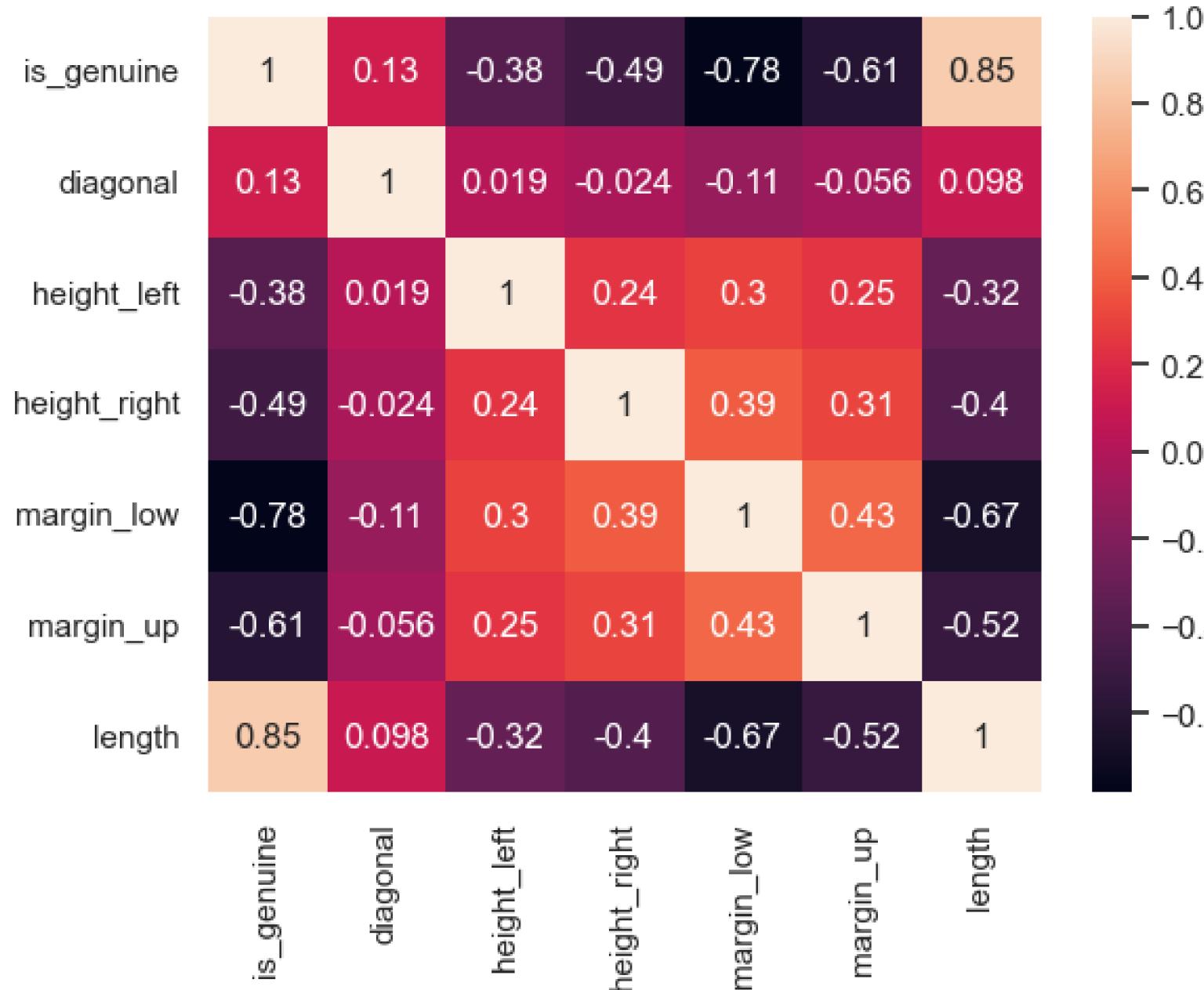


Variance entre les billets vrais/faux par rapport aux autres variables



Il y a une différence de variances entre les vrais et les faux billets pour chaque variable même si la variable "diagonal" est moins importante.

Corrélation entre les différentes variables



La variable is_genuine a une forte corrélation avec length et margin_low

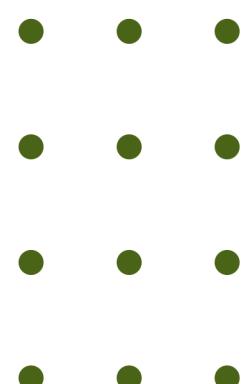
Traitemen~~t~~ment des valeurs manquantes



Pour compléter les valeurs manquantes, on va utiliser une régression linéaire.

Pour faire cette régression linéaire on va d'abord faire :

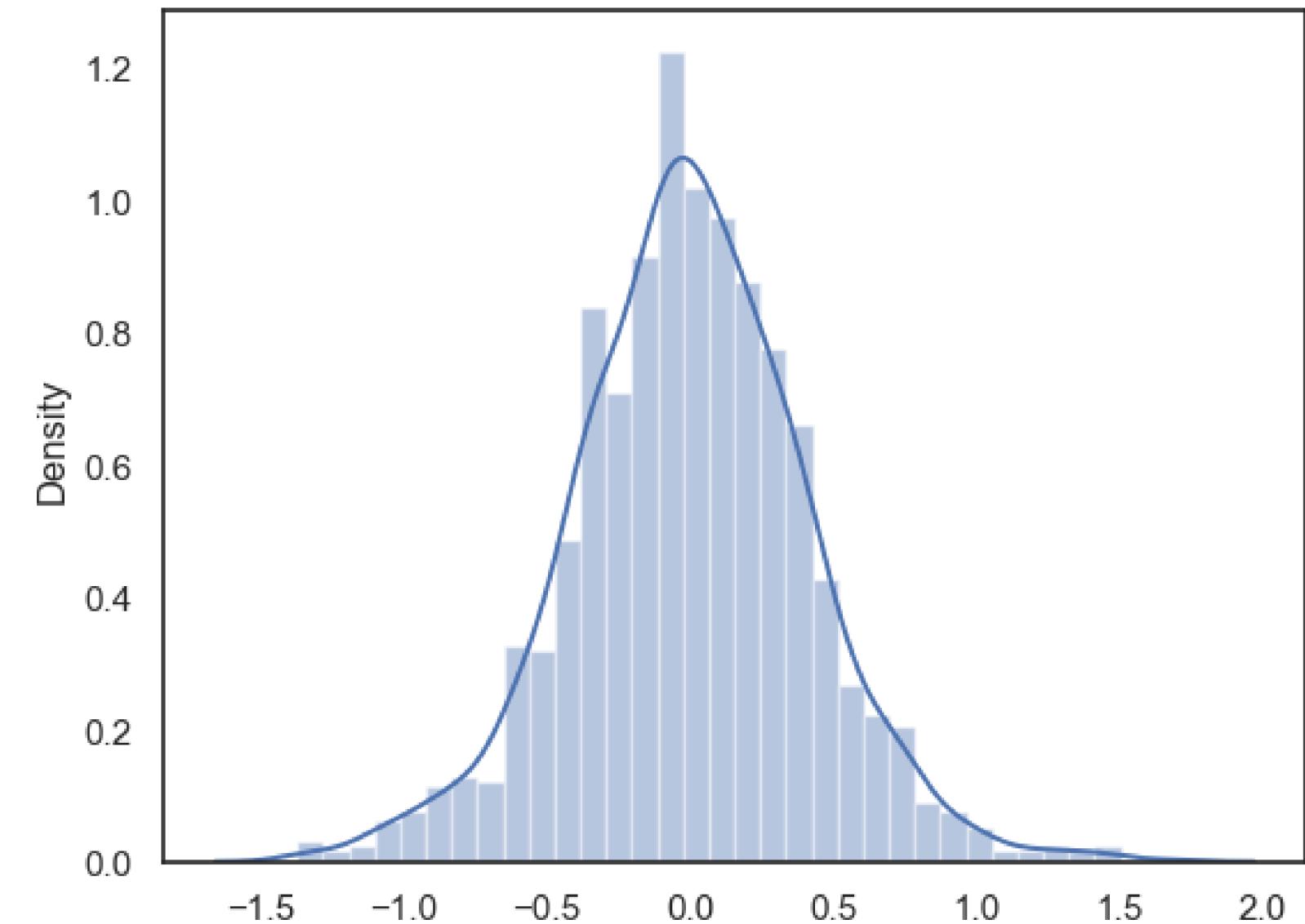
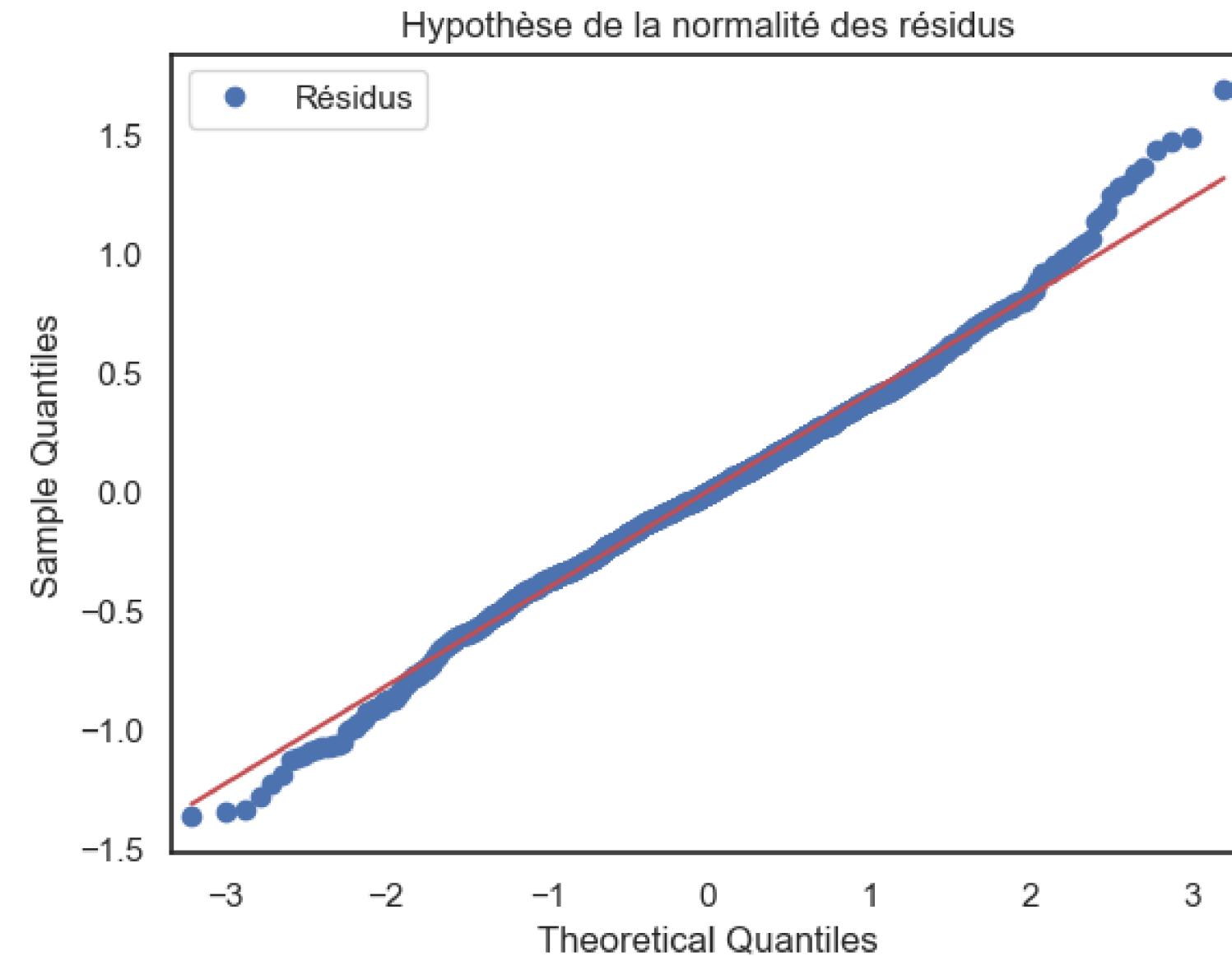
- Utiliser un dataset sans les valeurs manquantes.
- Ensuite trouver les valeurs significatives, ici `is_genuine` et `margin_up`.
- Faire une vérification par test pour valider l'utilisation de la régression linéaire, plusieurs test vont être fait :
 - Normalité : Les erreurs résiduelles doivent être distribuées normalement. Cela signifie que les résidus doivent suivre une distribution normale avec une moyenne de zéro.
 - Homoscédasticité : L'homoscédasticité signifie que la variance des erreurs résiduelles est constante à tous les niveaux de la variable prédictive.
 - Colinéarité : Cette hypothèse concerne la relation entre les variables prédictives (ou indépendantes) dans notre modèle de régression. Elle stipule qu'il ne devrait pas y avoir de forte corrélation linéaire entre les variables indépendantes.



Valeurs significatives

```
OLS Regression Results
=====
Dep. Variable: margin_low R-squared:      0.617
Model:           OLS   Adj. R-squared:    0.616
Method:          Least Squares F-statistic:     1174.
Date:            Wed, 13 Dec 2023 Prob (F-statistic): 1.24e-304
Time:             16:28:50   Log-Likelihood:   -774.73
No. Observations: 1463   AIC:                 1555.
Df Residuals:    1460   BIC:                 1571.
Df Model:        2
Covariance Type: nonrobust
=====
            coef    std err       t   P>|t|    [0.025    0.975]
-----+
Intercept    5.9263    0.198    30.003    0.000    5.539    6.314
is_genuine[T.True] -1.1632    0.029   -40.477    0.000   -1.220   -1.107
margin_up    -0.2119    0.059    -3.612    0.000   -0.327   -0.097
=====
Omnibus:        22.365   Durbin-Watson:    2.041
Prob(Omnibus): 0.000    Jarque-Bera (JB): 39.106
Skew:           0.057    Prob(JB):      3.22e-09
Kurtosis:       3.793    Cond. No.       65.0
=====
```

Test de la normalité des résidus



Si l'on veut tester la normalité des résidus, on peut faire également un test de Shapiro-Wilk.

Les résidus ne suivent pas une distribution normale (hypothèse rejetée).
P-value : 6.20942773821298e-06

L'observation des résidus montre qu'il n'est pas très différent d'une distribution symétrique et le fait que l'échantillon soit de taille suffisante permet de dire que les résultats obtenus par le modèle linéaire gaussien ne sont pas absurdes, même si le résidu n'est pas considéré comme étant gaussien.

Test de l'homoscédasticité des résidus et test de colinéarité



Test d'homoscédasticité

Test de Levene:

Statistique de test : 41.335456318169726

P-value : 7.359885063990044e-33

Le test de Levene indique une statistique de test de 41.33 et une P-value faible (7.35e-33). Cela indique un rejet de l'hypothèse d'homoscédasticité dans notre modèle de régression linéaire.

Test de colinéarité

Variance Inflation Factor (VIF):

| Variable | VIF |
|--------------|----------|
| 0 is_genuine | 1.593885 |
| 1 margin_up | 1.593885 |

Les valeurs VIF pour les variables explicatives is_genuine et margin_up sont toutes deux de 1.59, il n'y a donc pas de problème de colinéarité.

Insertion des données



Après avoir fait la prédiction, on ajoute ces valeurs dans le dataframe principal.

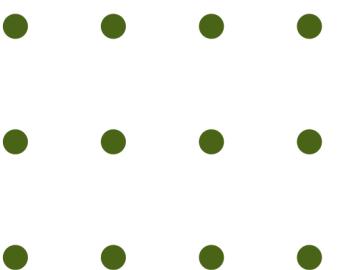
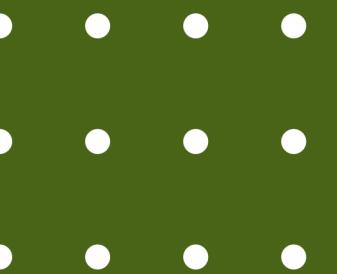
```
billets.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   is_genuine    1500 non-null   bool    
 1   diagonal     1500 non-null   float64 
 2   height_left  1500 non-null   float64 
 3   height_right 1500 non-null   float64 
 4   margin_low   1500 non-null   float64 
 5   margin_up    1500 non-null   float64 
 6   length       1500 non-null   float64 
dtypes: bool(1), float64(6)
memory usage: 71.9 KB
```

```
billets.isnull().sum()
is_genuine      0
diagonal        0
height_left     0
height_right    0
margin_low      0
margin_up       0
length          0
dtype: int64
```

Méthode de prédition

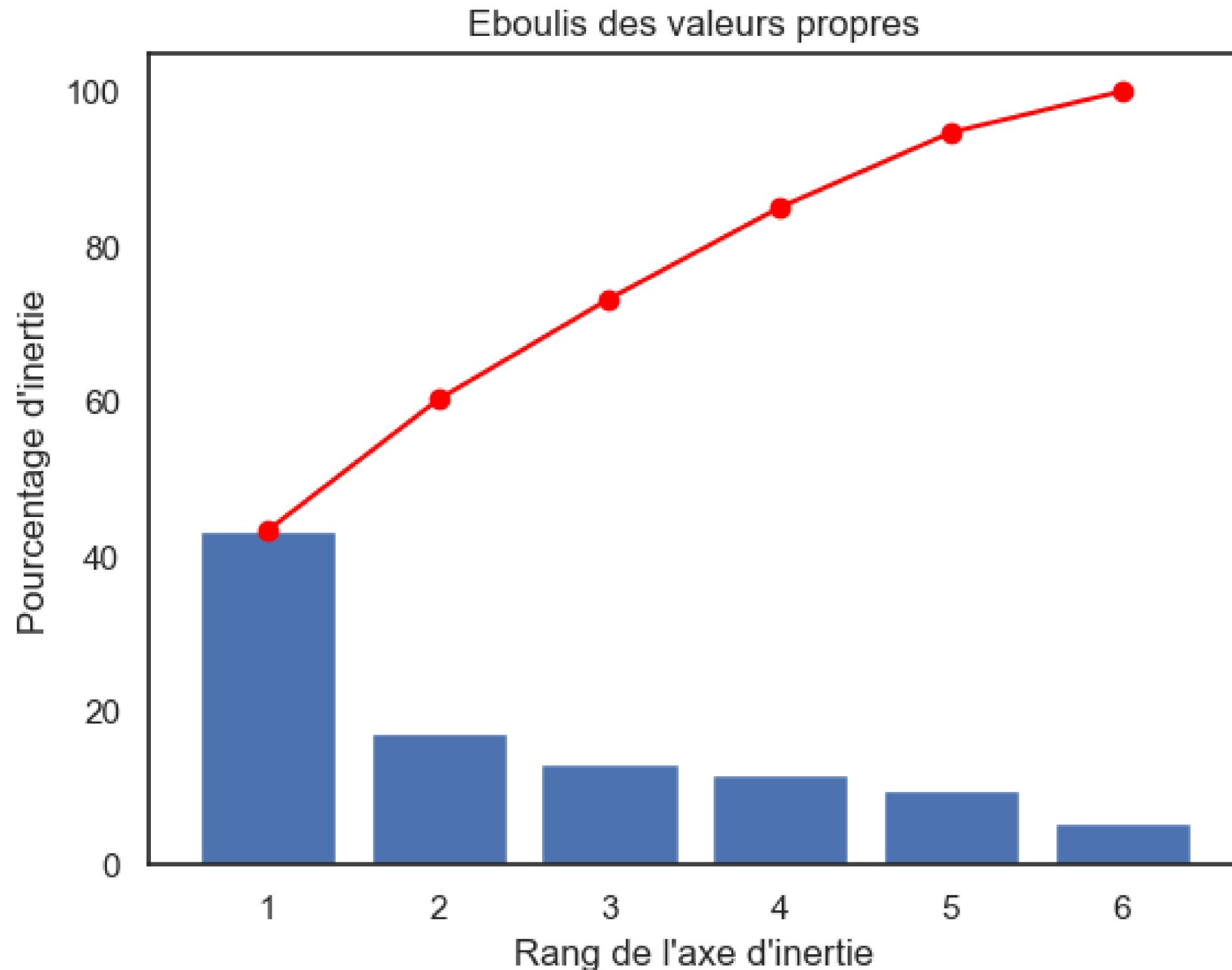
Deux méthodes de prédition vont être mis en concurrence :

- un kmeans, duquel seront utilisés les centroïdes pour réaliser la prédition.
- une régression logistique.



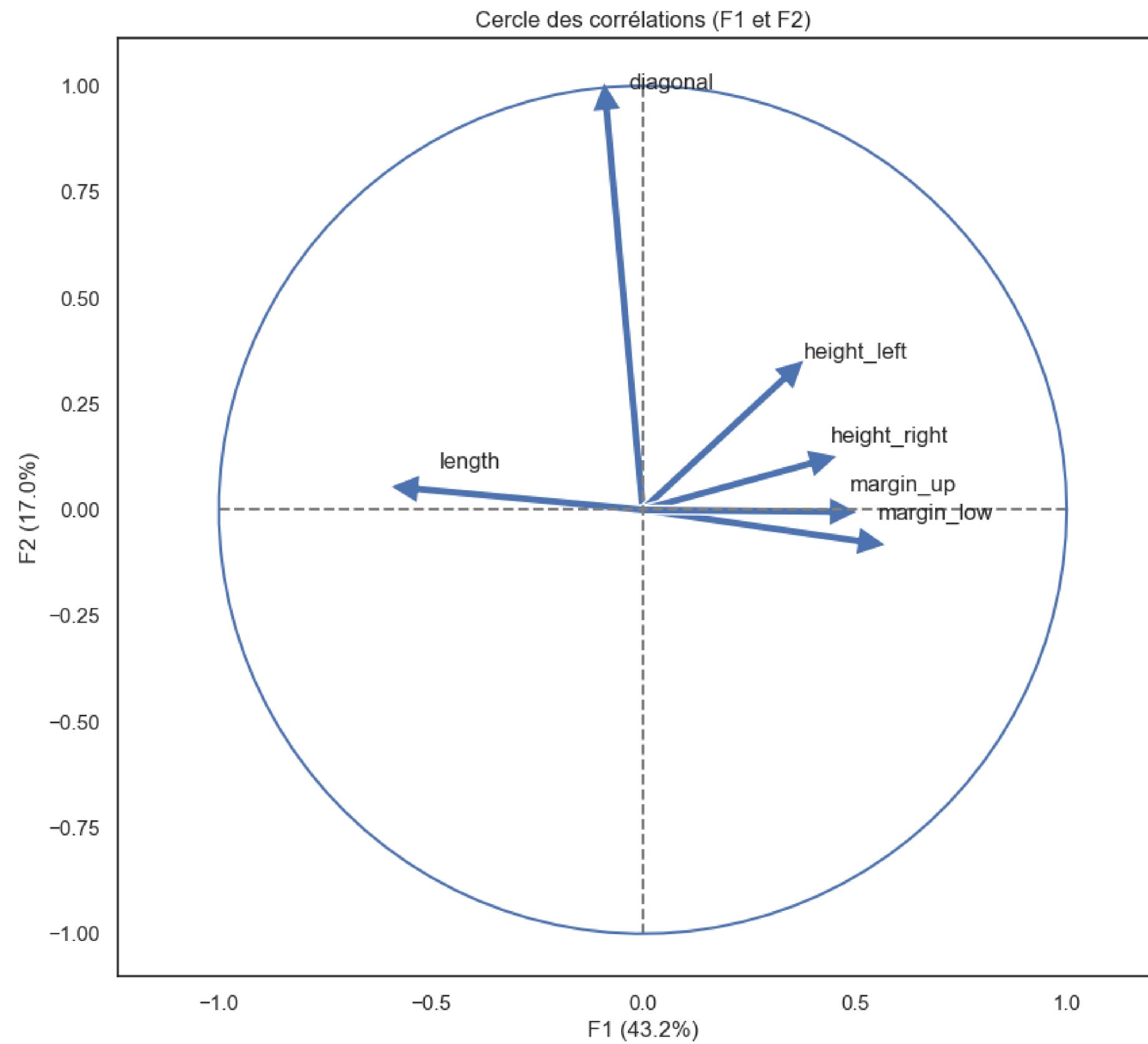
ACP - Éboulis des valeurs propres

• • •
• • •
• • •



Les 2 premières composantes captent 62,16 % ($43.2 + 16.96$) de la variance.

ACP - Cercle des corrélations

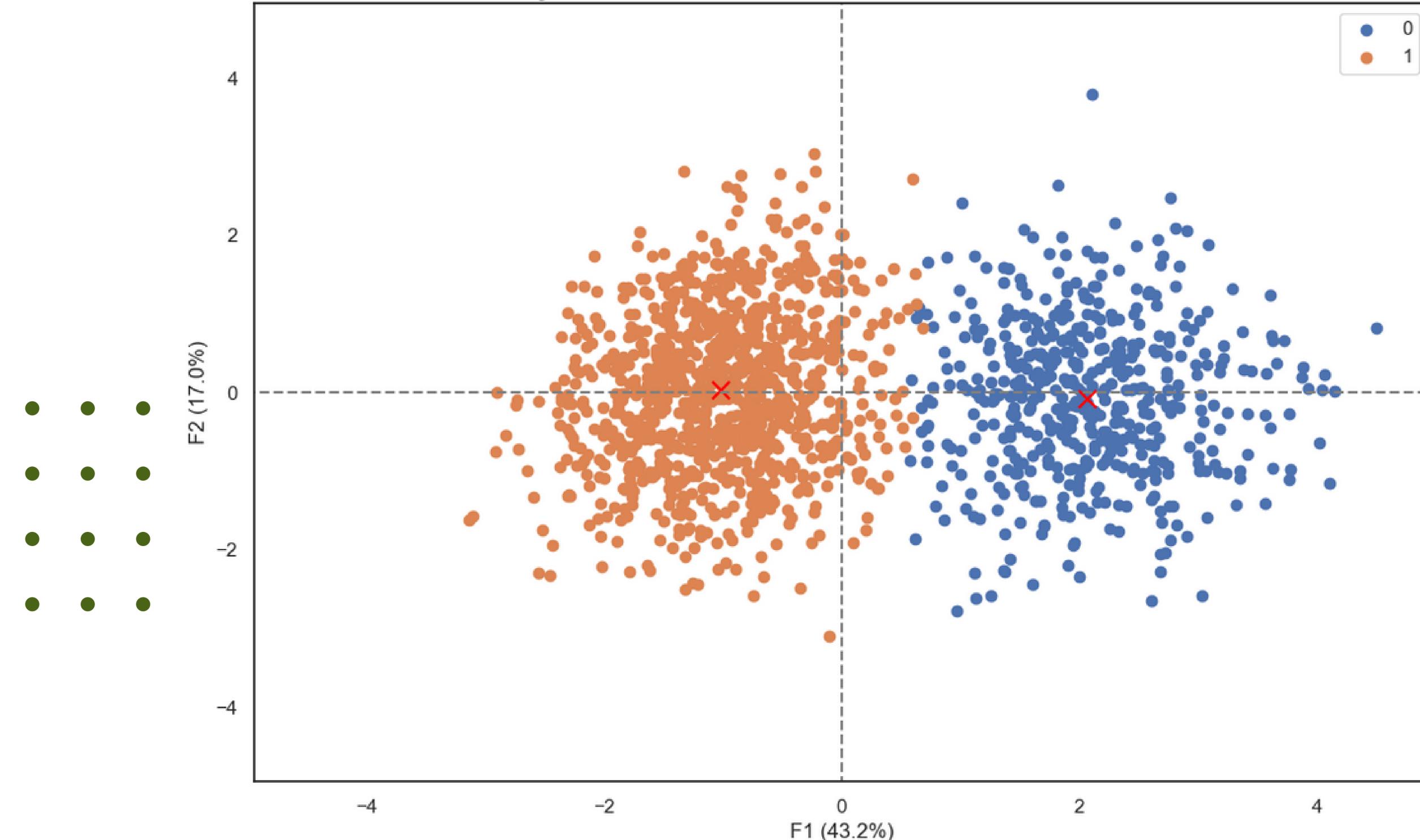


En F1, les variables diagonal et length sont décorrélées par rapport aux autres variables. Pour la composante F2, ce sont les variables margin_low et margin_up qui sont décorrélées par rapport au reste des variables.

ACP - Projection sur le plan factoriel

A yellow background featuring a repeating pattern of white circular dots. The dots are arranged in a grid-like fashion, with horizontal and vertical spacing between them.

Projection des 1500 individus sur F1 et F2

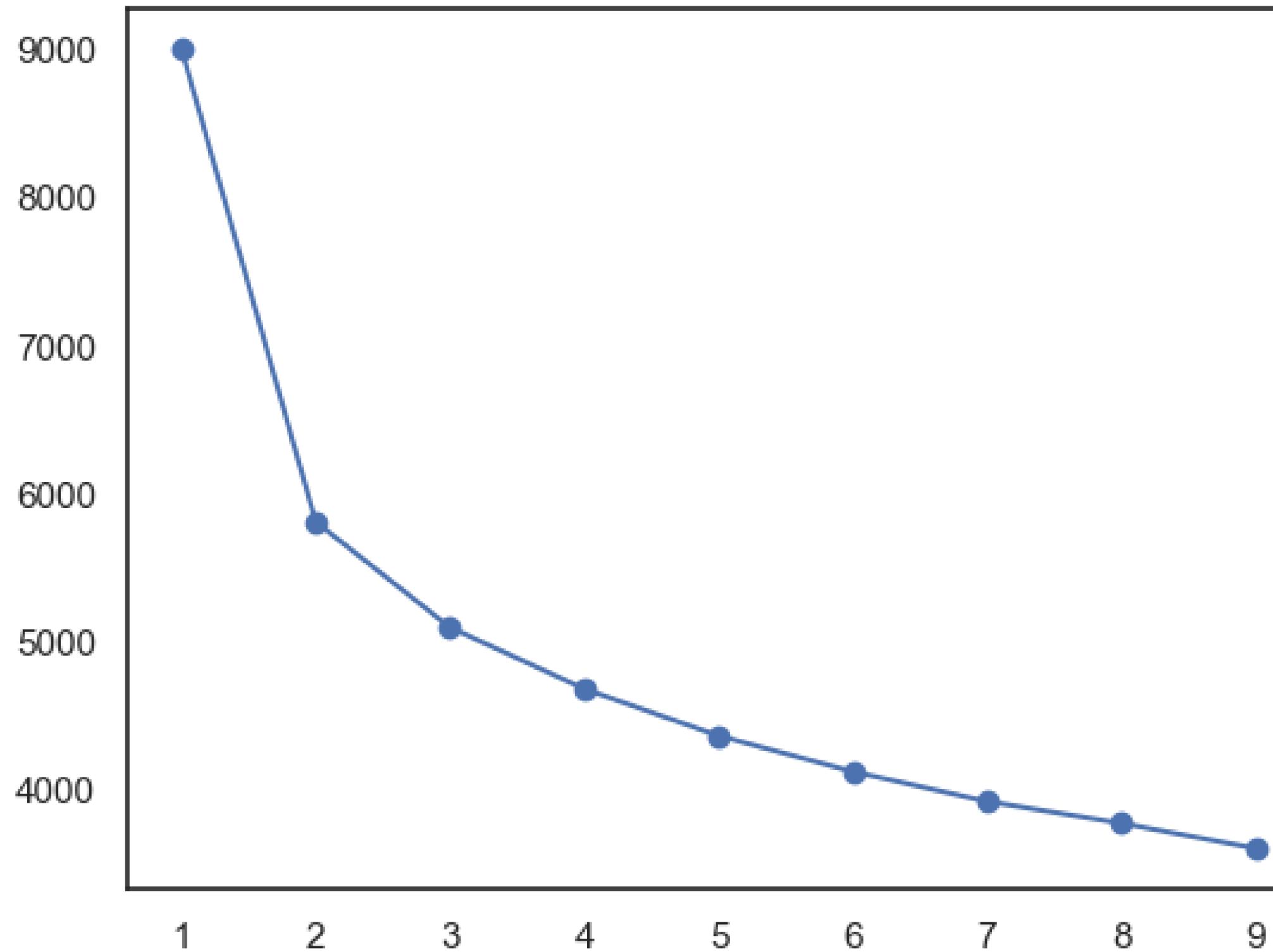


La projection nous permet de mieux représenter les groupes pour permettre une sélection.

Les croix rouges montrent les centroïdes des clusters.

Kmeans - Métrique de l'inertie

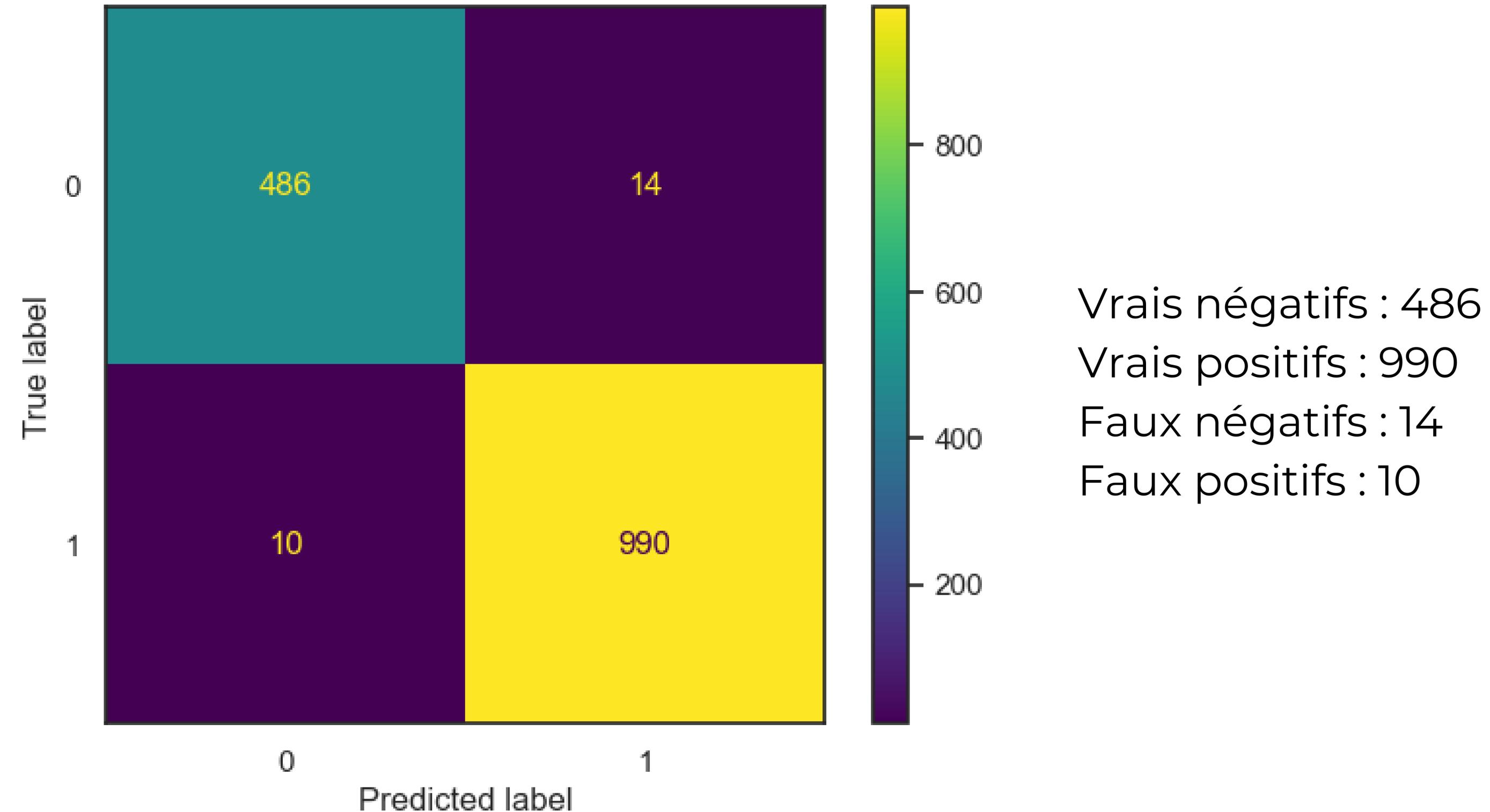
• • •
• • •
• • •



Le choix de 2 clusters est la meilleure solution comme on veut une séparation binaire (vrai / Faux).

Kmeans - Matrice de confusion

⋮ ⋮ ⋮ ⋮
⋮ ⋮ ⋮ ⋮



Régression logistique

A yellow square with white polka dots.

Pour réaliser la régression logistique, il faut préparer les données. La variable `is_genuine` va être transformée en variable numérique (Vrai = 1 et Faux = 0).

Pour faire la prédiction, on va utiliser les variables descriptives significatives suivantes :

- length
 - margin_low
 - margin_up
 - height_right

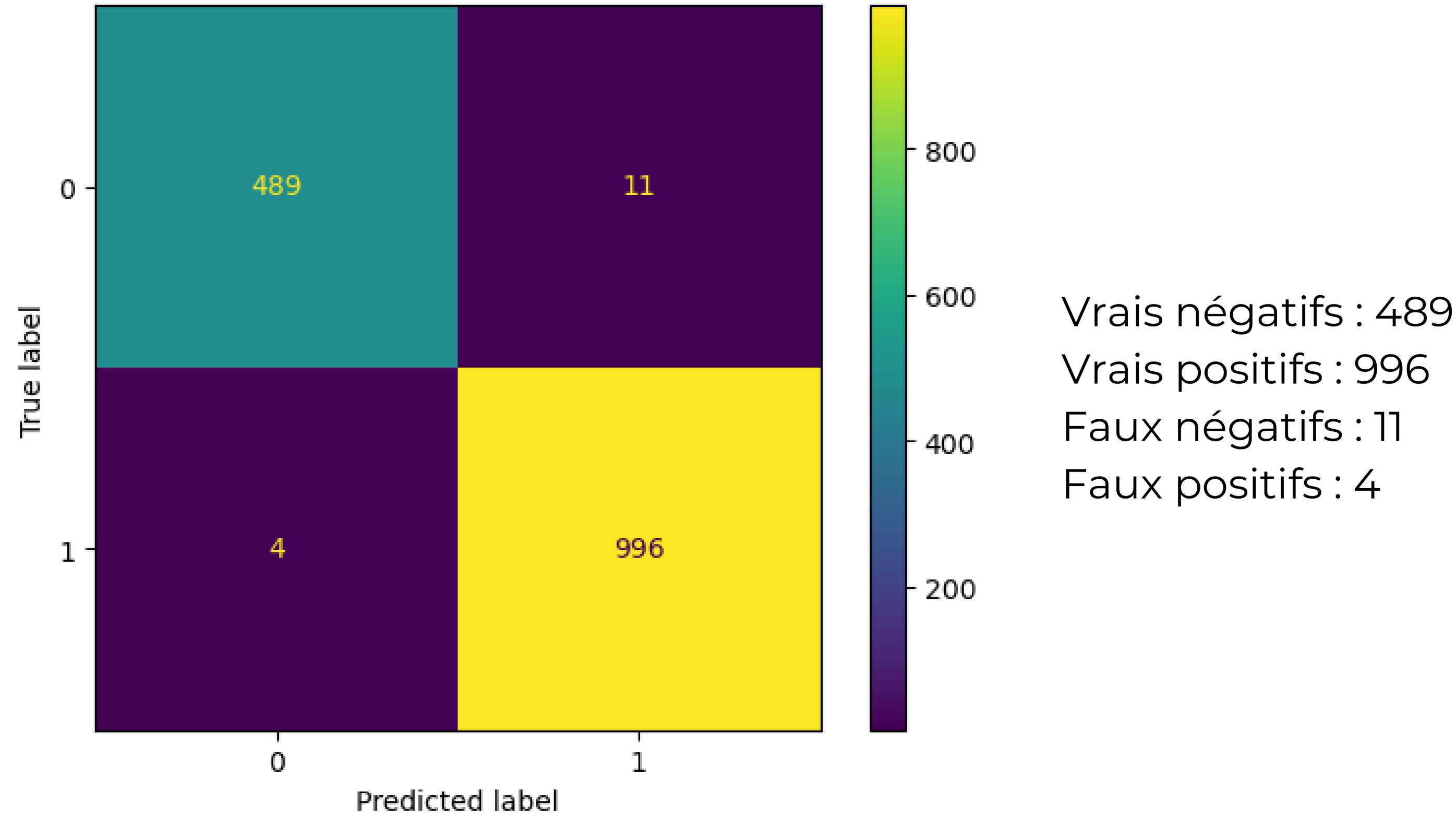
```

Logit Regression Results
=====
Dep. Variable:      is_genuine    No. Observations:      1500
Model:              Logit         Df Residuals:        1495
Method:             MLE          Df Model:            4
Date:              Wed, 13 Dec 2023 Pseudo R-squ.:     0.9579
Time:                16:29:52   Log-Likelihood:   -40.175
converged:          True         LL-Null:           -954.77
Covariance Type:   nonrobust   LLR p-value:       0.000
=====

```

Régression - Matrice de confusion

⋮ ⋮ ⋮ ⋮
⋮ ⋮ ⋮ ⋮



Choix du modèle

• • •
• • •
• • •

Kmeans

Vrais négatifs : 486
Vrais positifs : 990
Faux négatifs : 14
Faux positifs : 10

• • •
• • •
• • •
• • •

Régression logistique

Vrais négatifs : 489
Vrais positifs : 996
Faux négatifs : 11
Faux positifs : 4

On remarque que la méthode de prédiction avec régression logistique est plus concluante qu'avec un Kmeans.