Projet 6

Détectez les faux billets

Millet Teïlo

460 000

fausses coupures en euros ont été retirées de la circulation en 2020 (dont 220 000 au second semestre), soit une baisse de 17,7 % par rapport à 2019.

Sommaire

5 étapes

1. Analyse des données
2. Analyse en Composantes Principales
3. Algorithme de classification
4. Regression Logistique
5. Algorithme de prédiction

Les données

Données fournis

	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	171.67	103.74	103.70	4.01	2.87	113.29
2	True	171.83	103.76	103.76	4.40	2.88	113.84
3	True	171.80	103.78	103.65	3.73	3.12	113.63
4	True	172.05	103.70	103.75	5.04	2.27	113.55
...	...	...	...	...	...	...	...
165	False	172.11	104.23	104.45	5.24	3.58	111.78
166	False	173.01	104.59	104.31	5.04	3.05	110.91
167	False	172.47	104.27	104.10	4.88	3.33	110.68
168	False	171.82	103.97	103.88	4.73	3.55	111.87
169	False	171.96	104.00	103.95	5.63	3.26	110.96

170 rows × 7 columns

1. Véritable billet
2. Diagonale
3. Hauteur gauche
4. Hauteur droite
5. Marge basse
6. Marge haute
7. Longueur

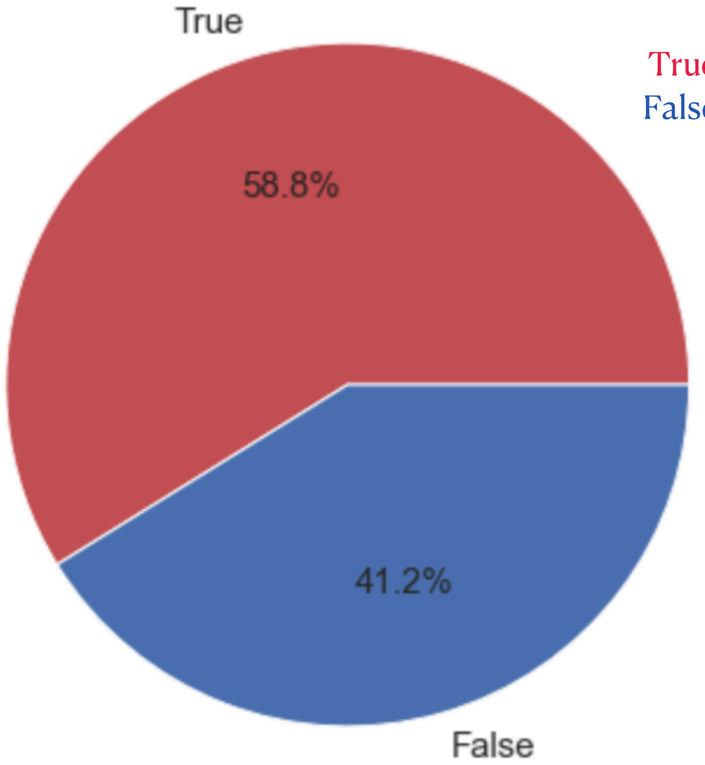
Analyse

Vérification et nettoyage des données

Durant cette étape nous vérifions si les données fournis contiennent des valeurs non renseignées ou négatives

On en profite pour voir la répartition des vrais

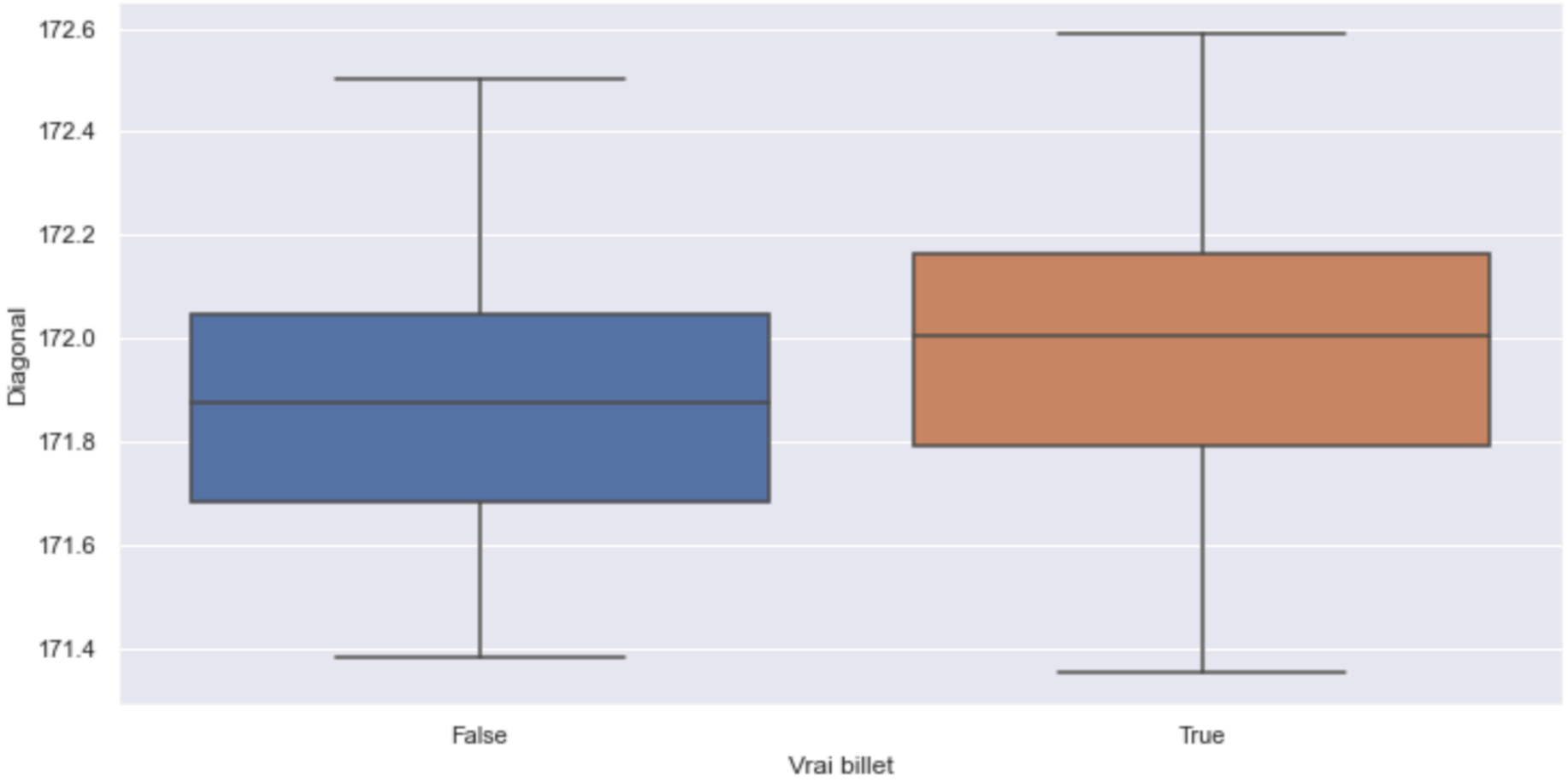
et faux billets:

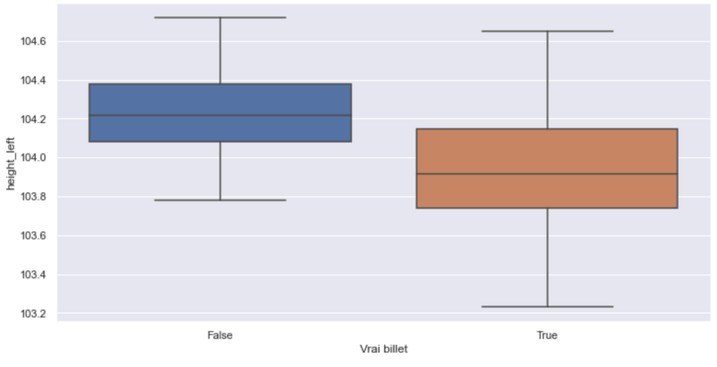


True = Vrai,
False = Faux

Analyse des variables

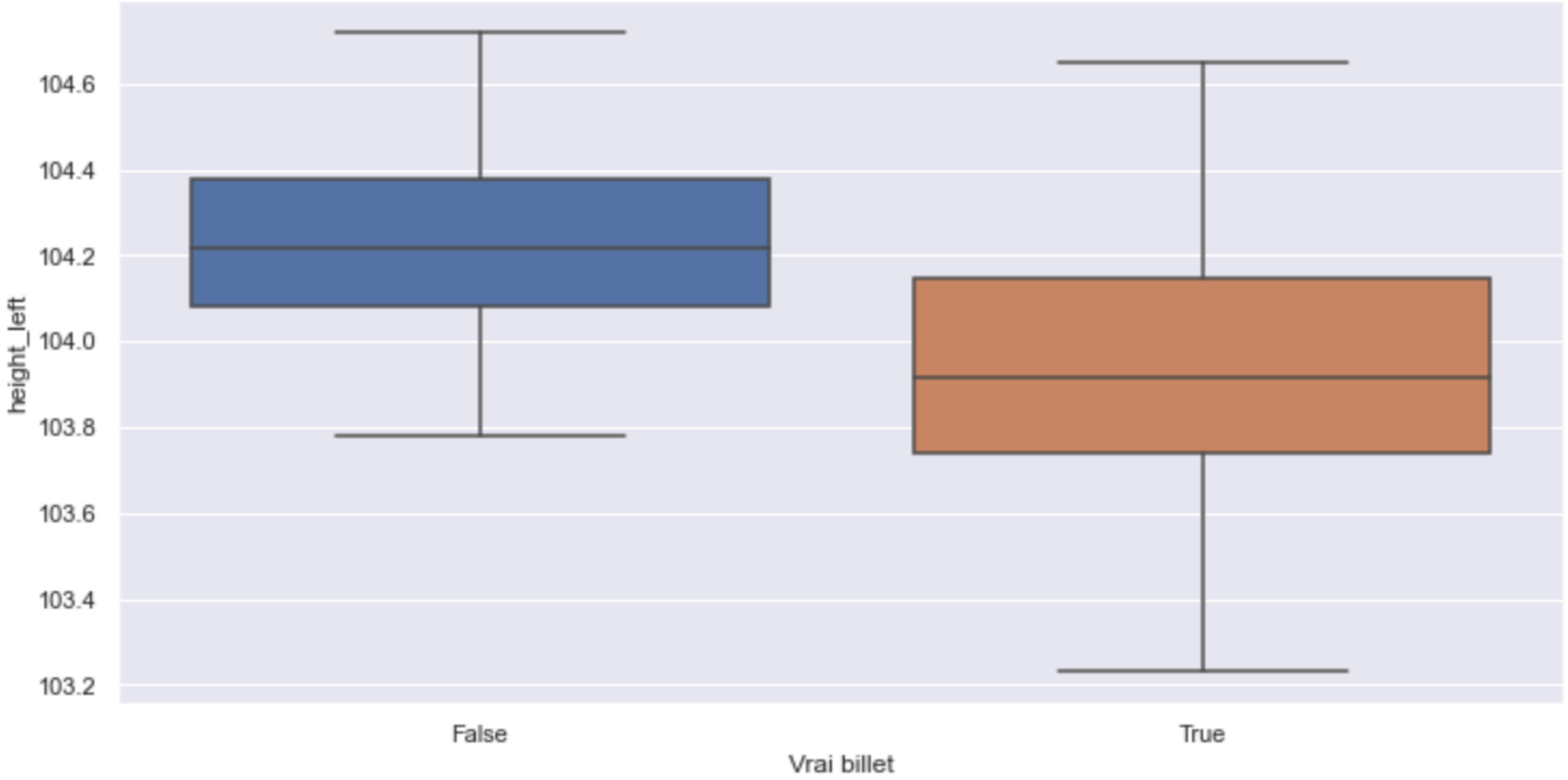
Diagonale

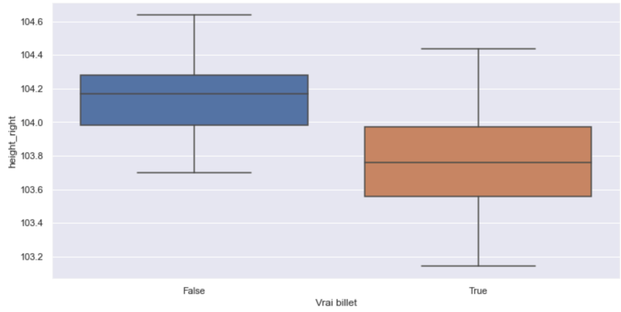




Analyse des variables

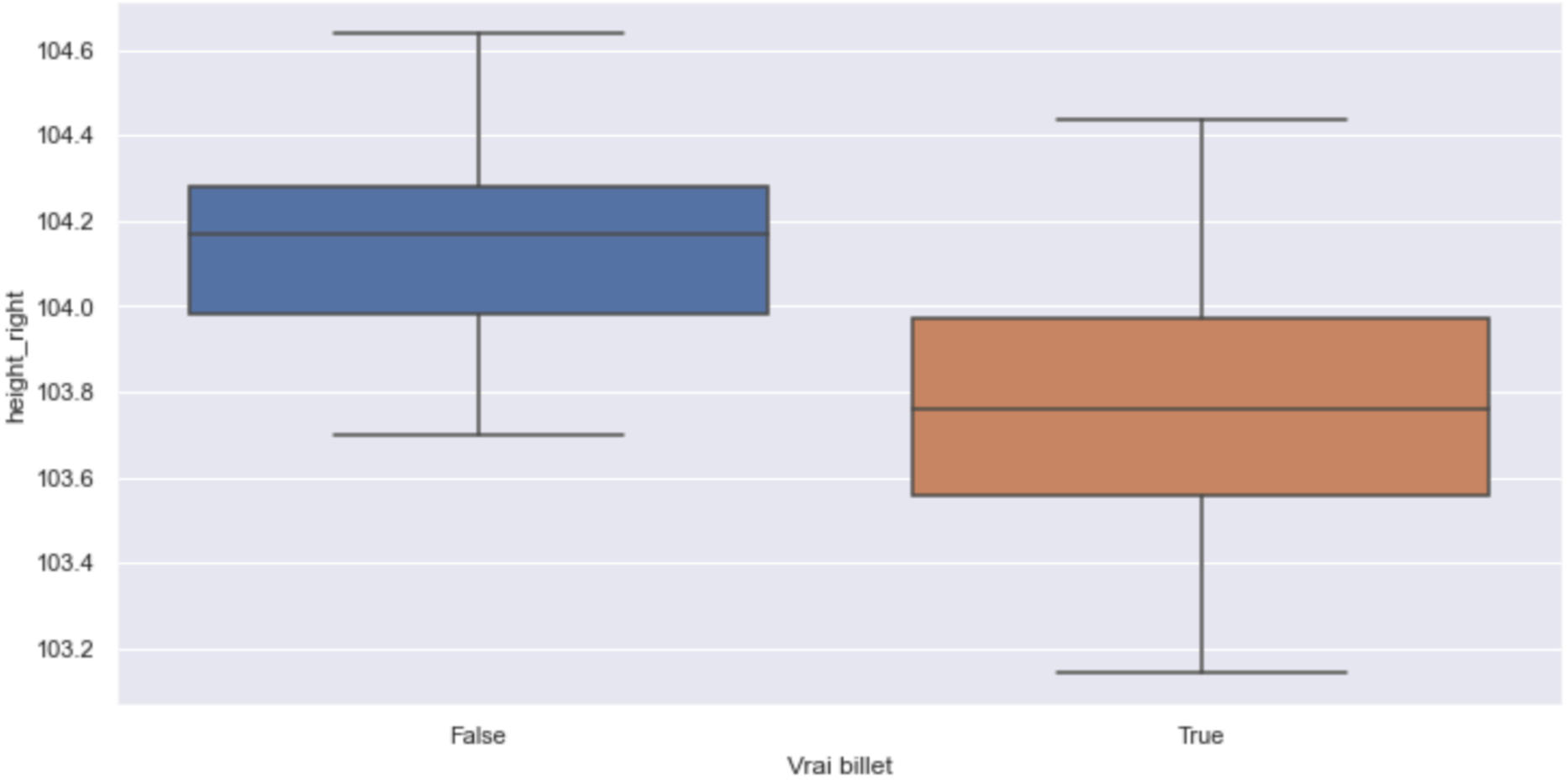
Hauteur Gauche

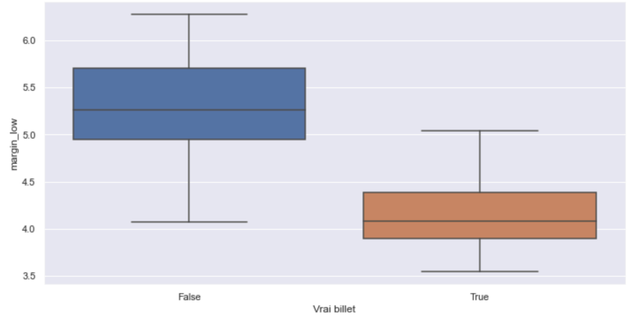




Analyse des variables

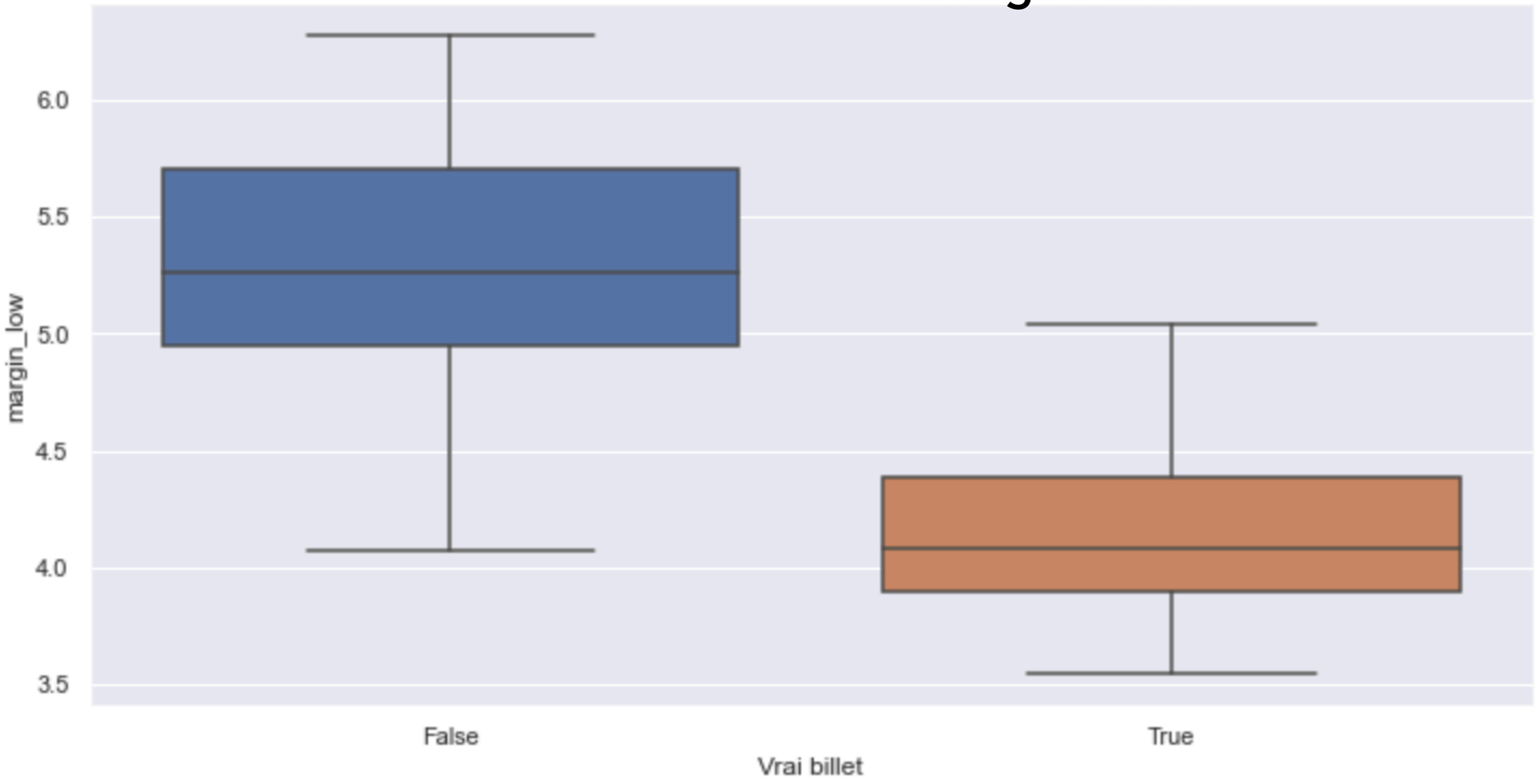
Hauteur Droite

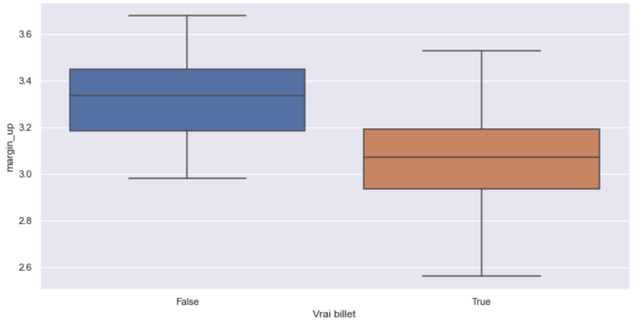




Analyse des variables

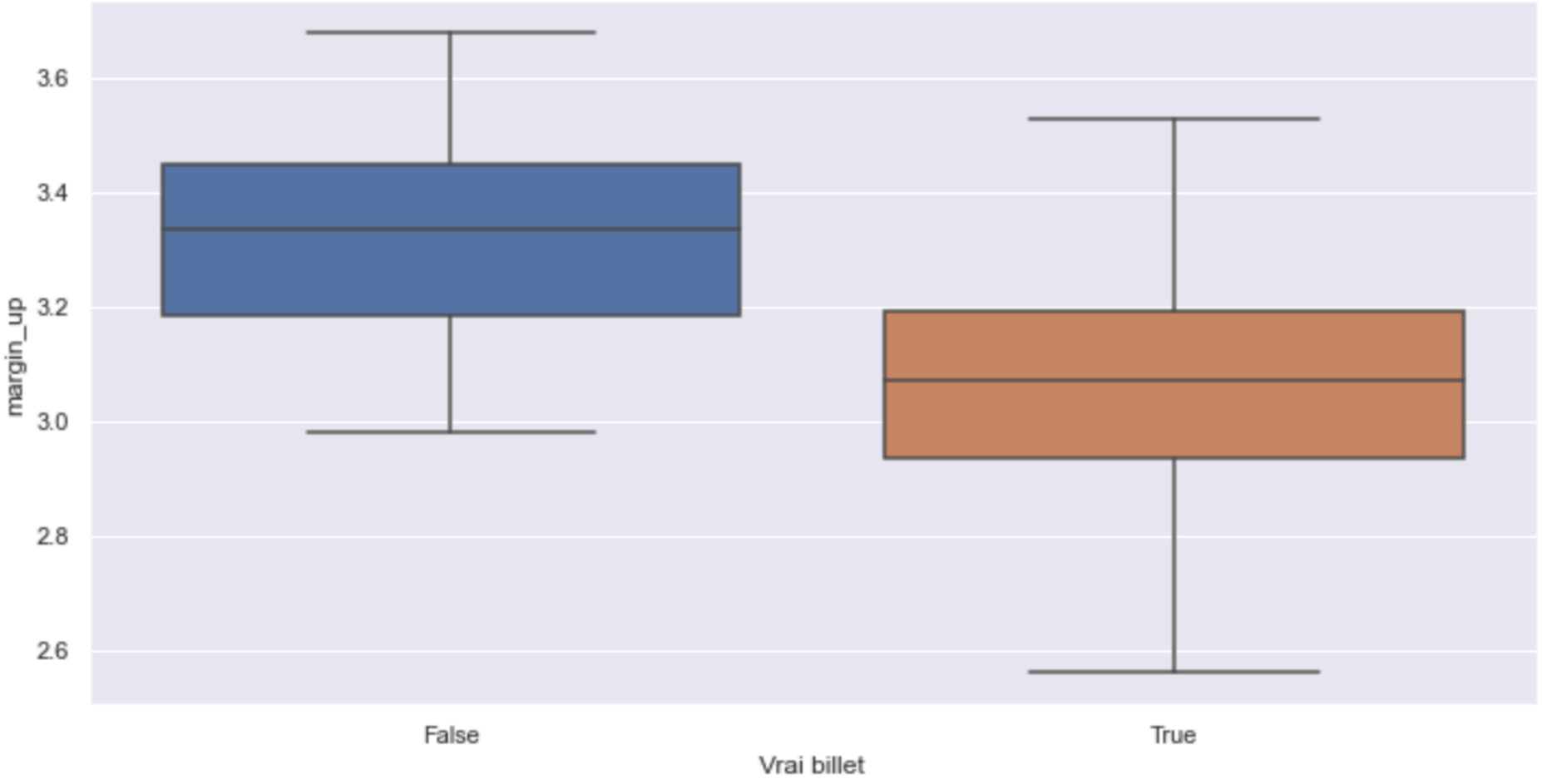
Marge Basse

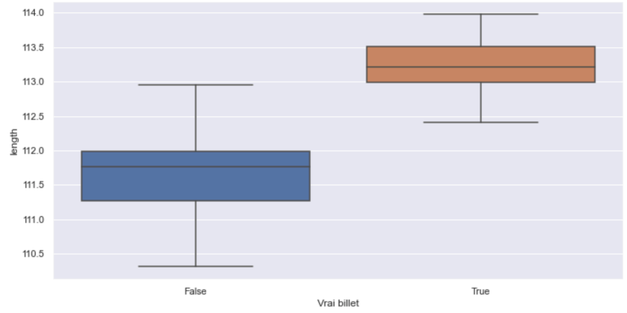




Analyse des variables

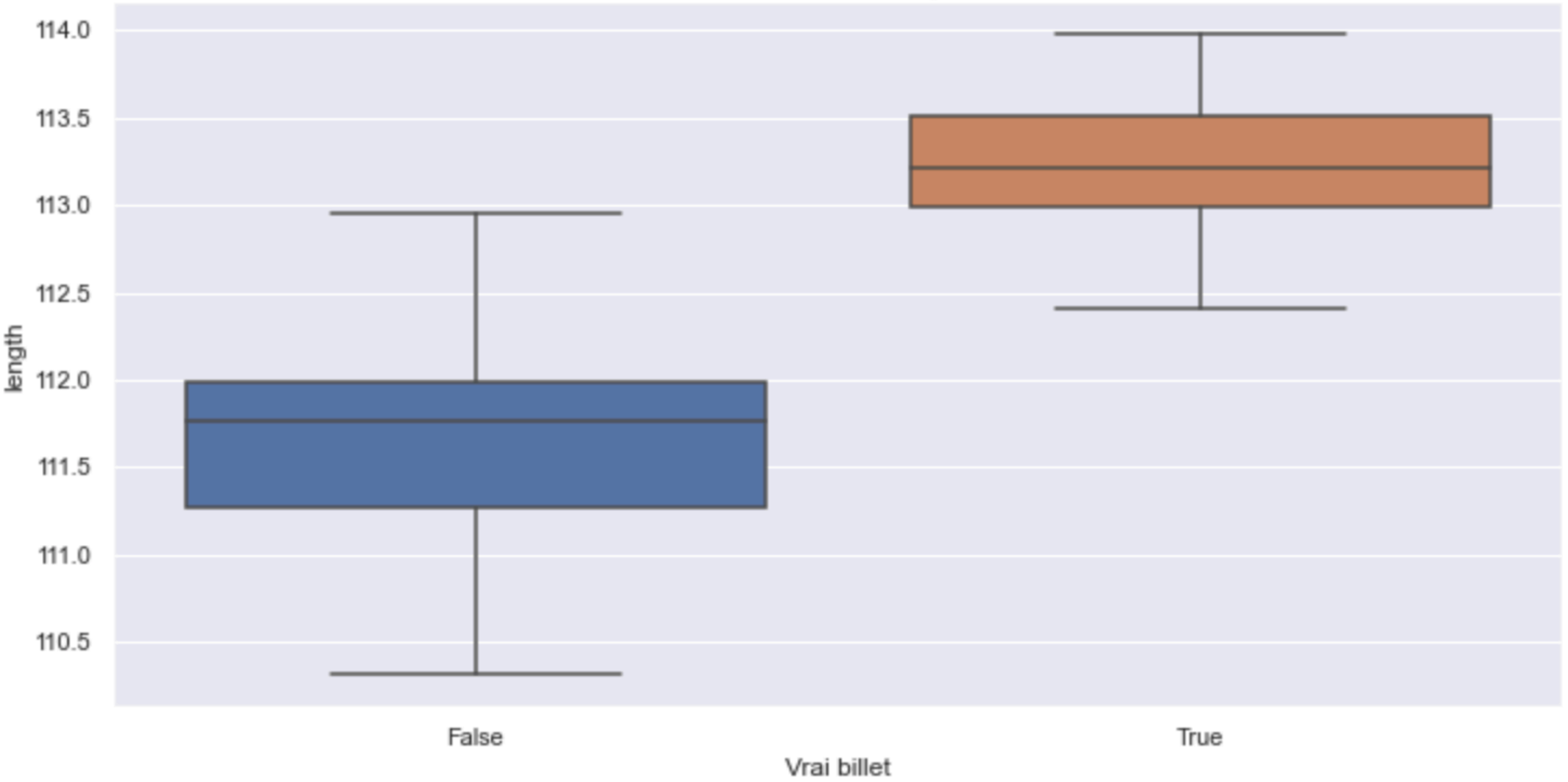
Marge Haute



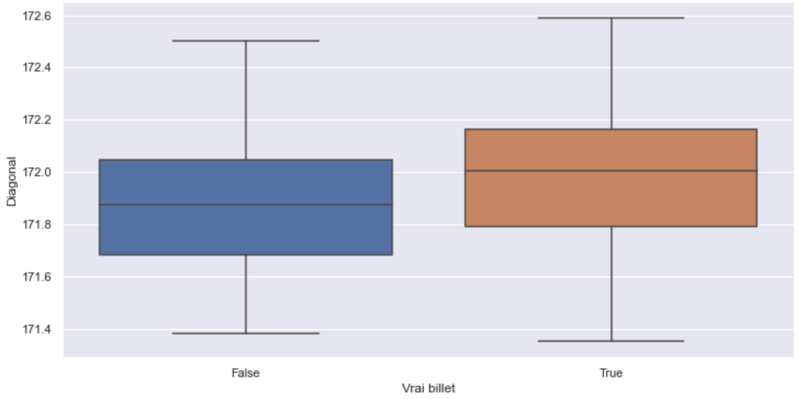


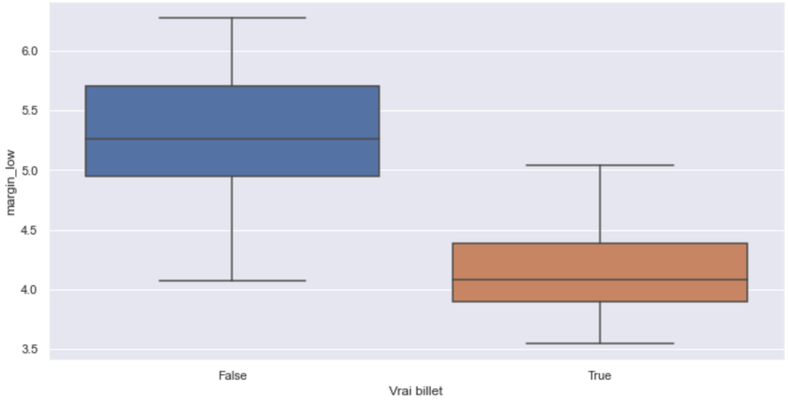
Analyse des variables

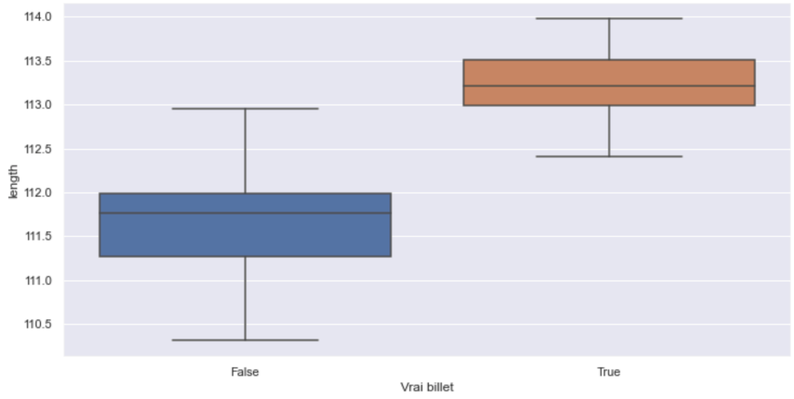
Longueur

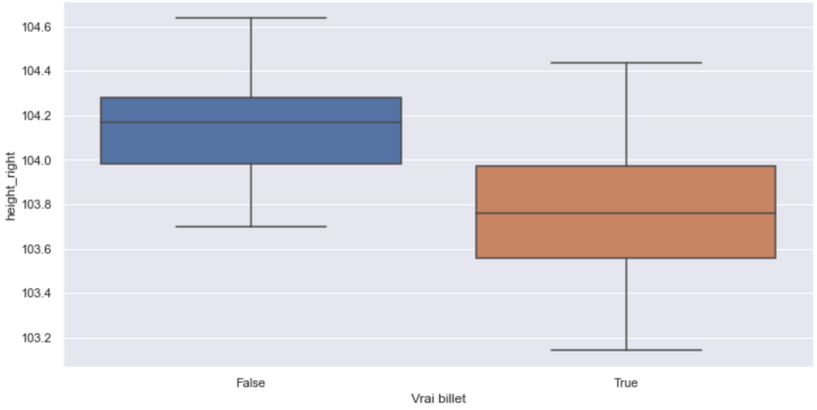


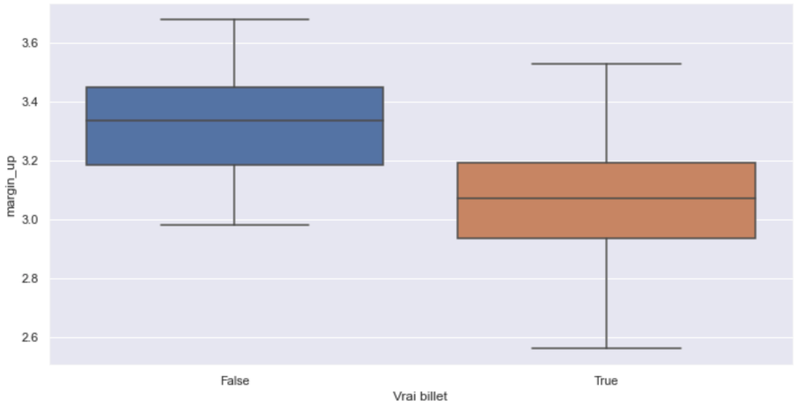
Analyse des variables

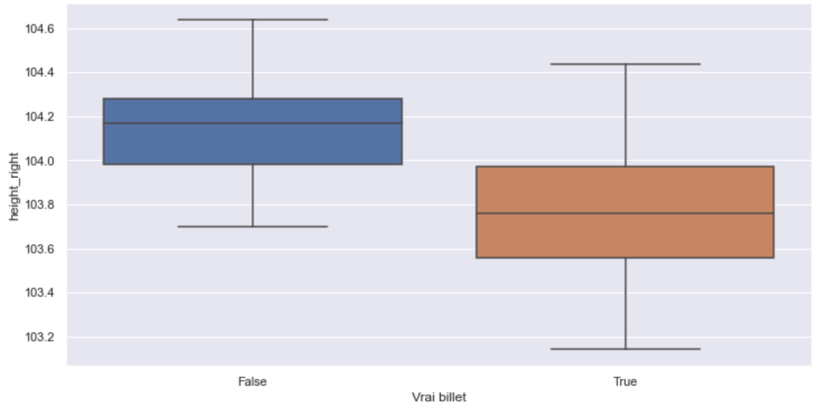




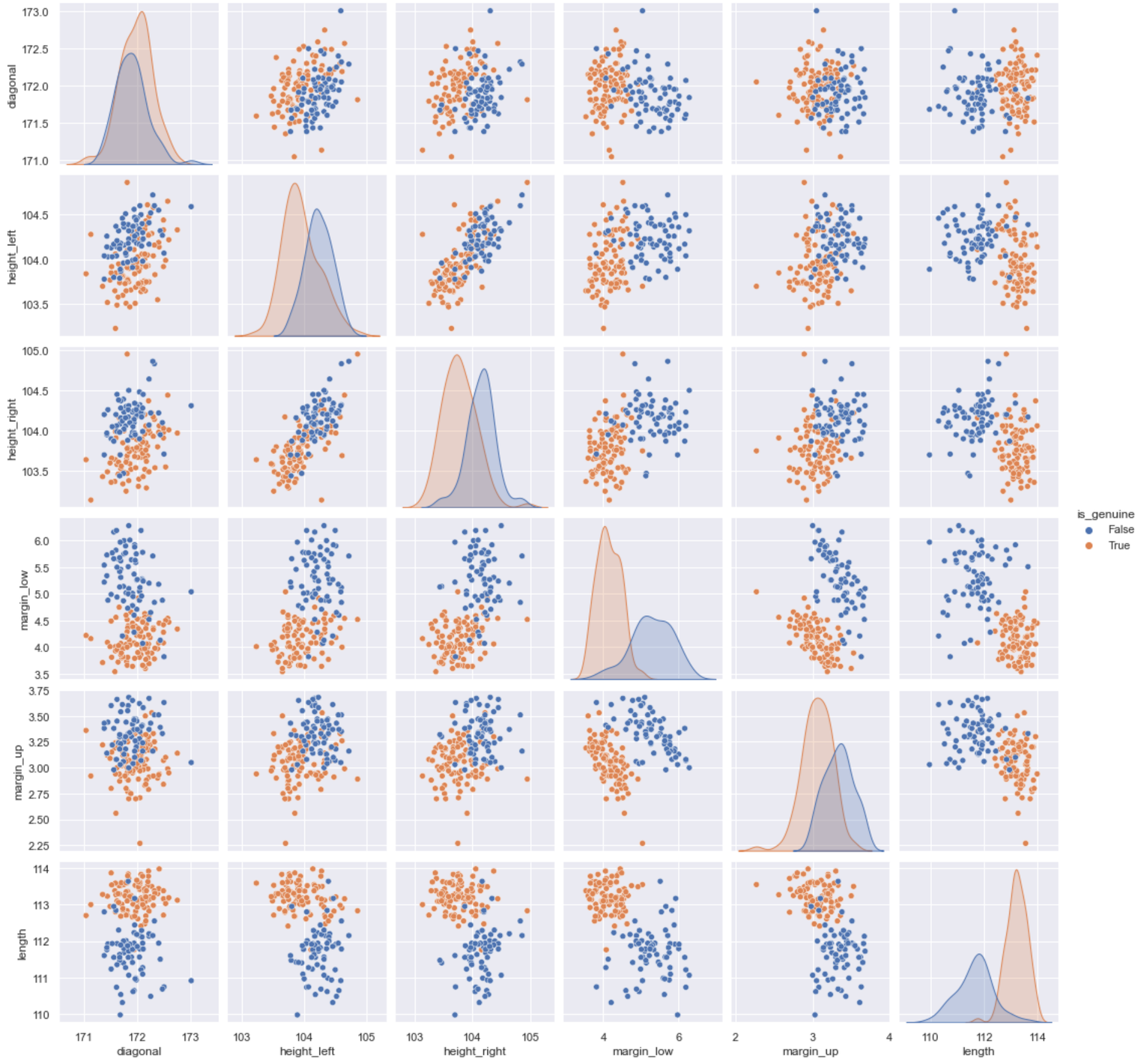








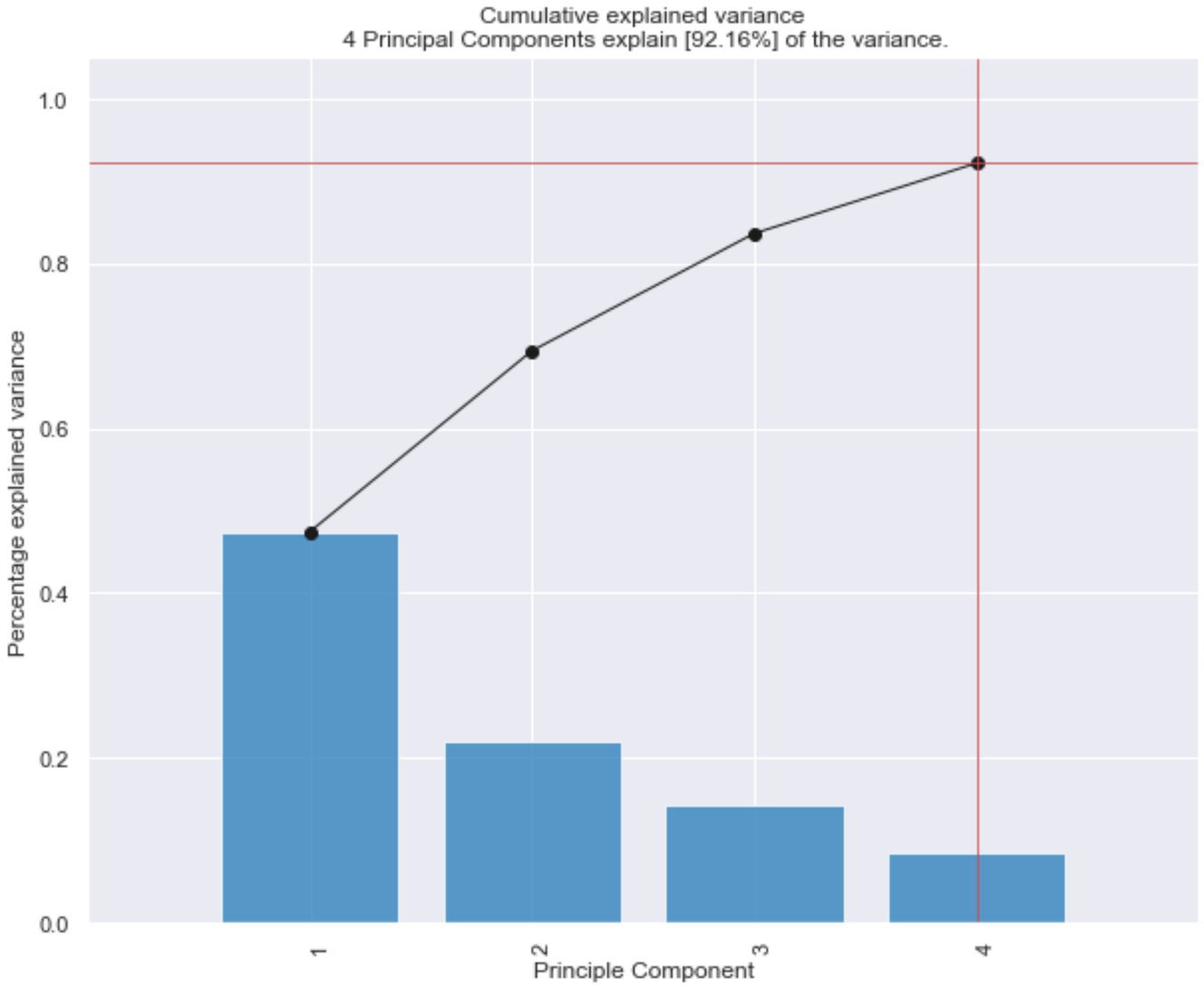
Les faux billets ont tendances a avoir des valeurs plus grande que les vrais billets, excepté pour la longueur qui est bien plus élevé chez les vrais billets


diagonal
height_left
height_right
margin_low
margin_up
length
is_genuine
False
True

Analyse en Composantes Principales

ACP

Analyses de l'éboulis des valeurs propres

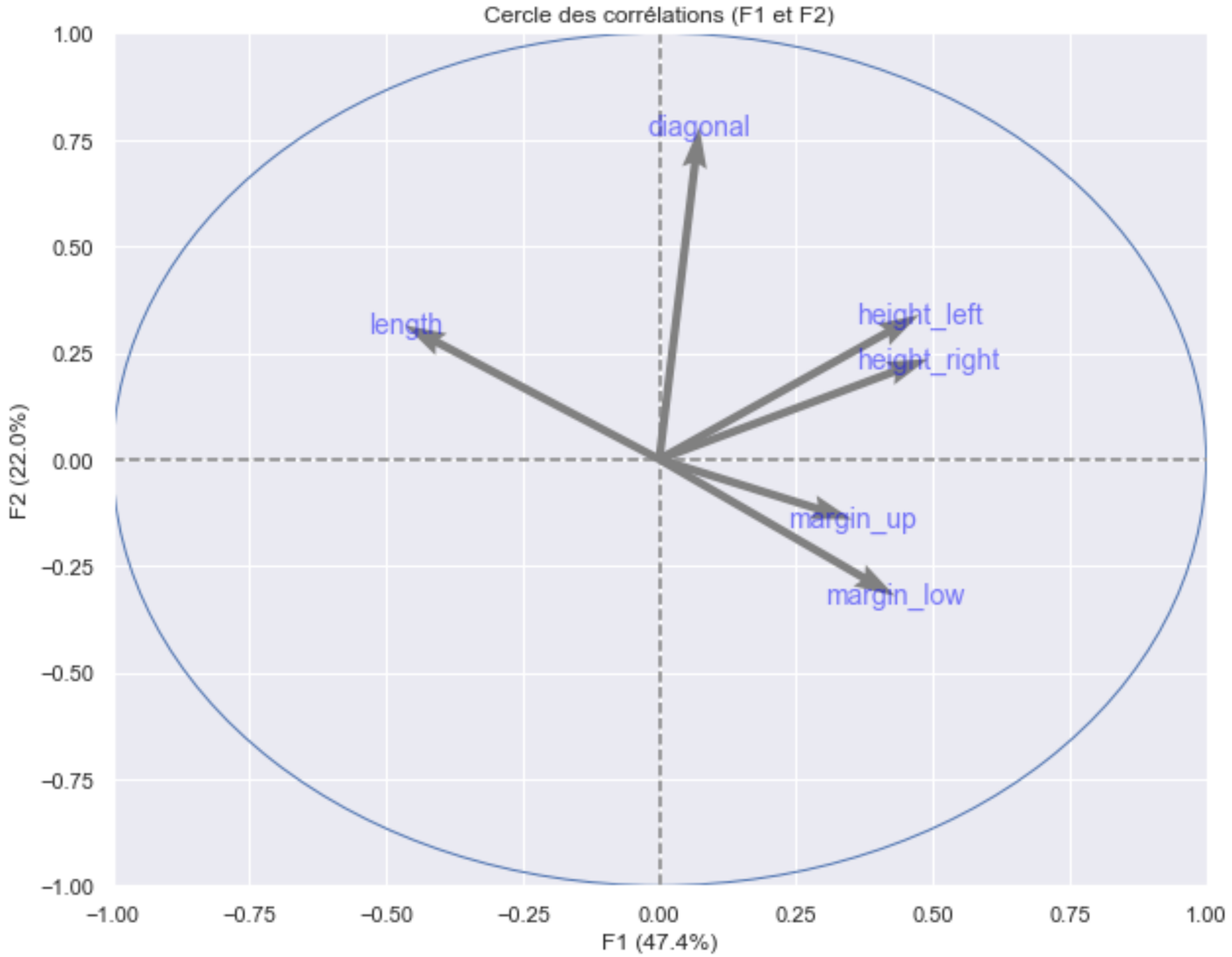


4 composantes explique 92.16% de la variance

L'analyse en Composantes Principales portera sur 4 variables.

ACP

Cercle de corrélations



L'axe F1 rassemble 47% de l'information et l'axe F2, 22%.

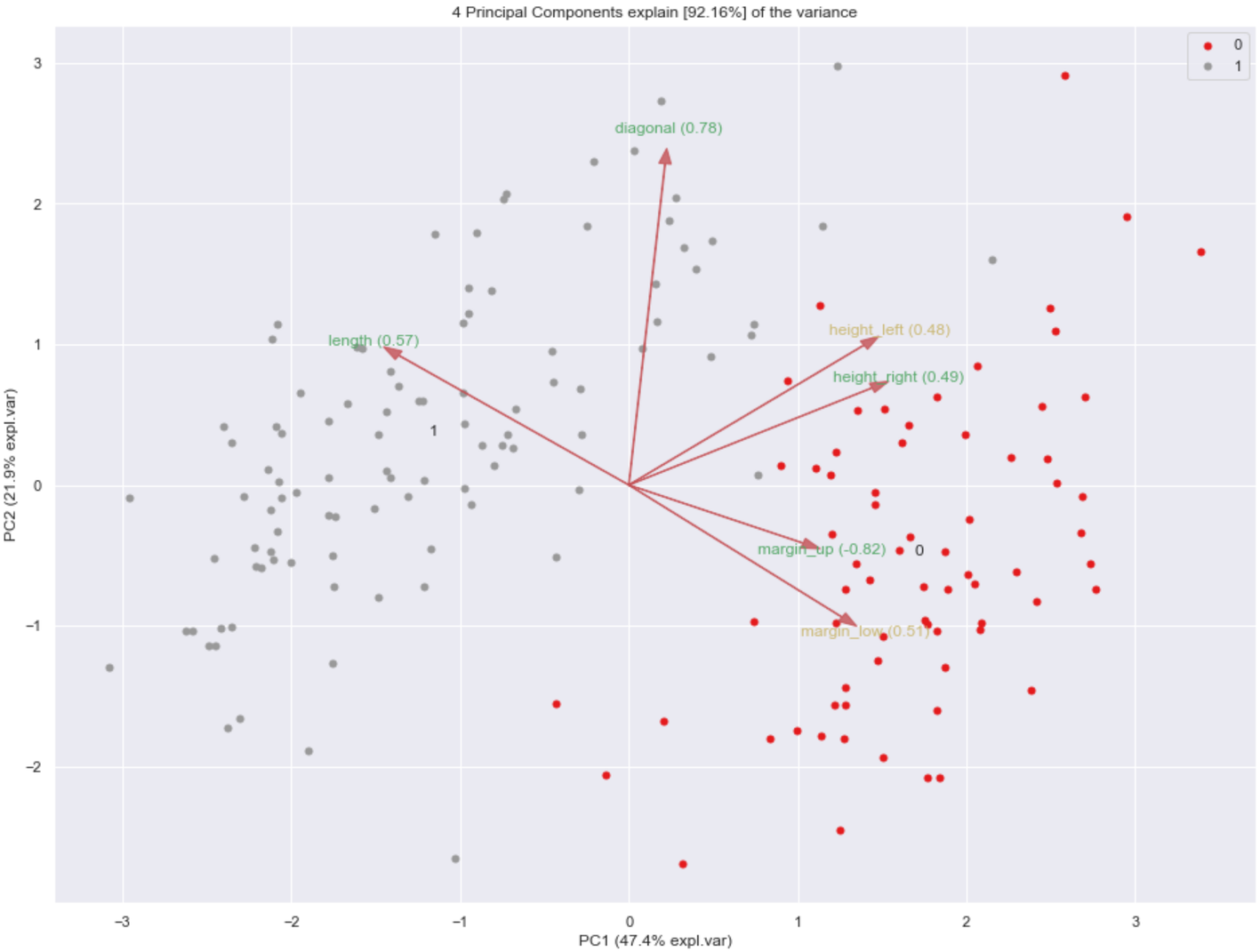
La variable 'diagonal' est très corrélé à la deuxième dimension (75%).

En revanche, la variable 'length' semble fortement anti-corrélé à la premiere dimension(-48%).

C'est à dire que les individus qui ont une faible valeur sur le plan F1, ont une forte 'length'.

ACP

Représentation des individus sur le plan factoriel



2 groupes se séparent.

Il semble que les individus du groupe ‘1’ sont fortement corrélés aux variables ‘length’ et ‘diagonal’.

Les individus du groupe ‘2’ semble eux, plus corrélés aux autres variables.

Cela confirme notre précédente analyse.

ACP

Quelle résultat ?

L'Analyse en Composantes Principales, complete nos précédentes analyses, à savoir:

- 2 groupes:
  - Celui des billets avec des valeur de marge et de hauteur des billets élevé.
  - Celui des billets avec de plus grande valeur quand à la longueur et la diagonale des billets.

Cette analyse confirme principalement qu'un cluster de billets à une longueur plus élevé que l'autres. Ainsi les individus du deuxième clusters ont une valeur de marge et de hauteur plus élevé.

Algorithme de classification

k-Means Clustering

Nous allons utiliser k-Means comme algorithme de classification.

- standardise les données afin d'éviter des erreurs du au différence entre les variables,
- sélectionne le nombre de groupe que l'on souhaite obtenir, ici 2 car il existe seulement 2 possibilités: le billet est un vrai ou un faux.

k-Means sélectionne 2 points afin de déterminer les clusters,

- il va mesurer la distance entre les différents points et les 2 clusters pour determiner à quels clusters les points appartiennent.
- il sélectionnera les clusters qui ont la somme de variance inter clusters minimale.

kMeans Clustering

Choix des valeurs

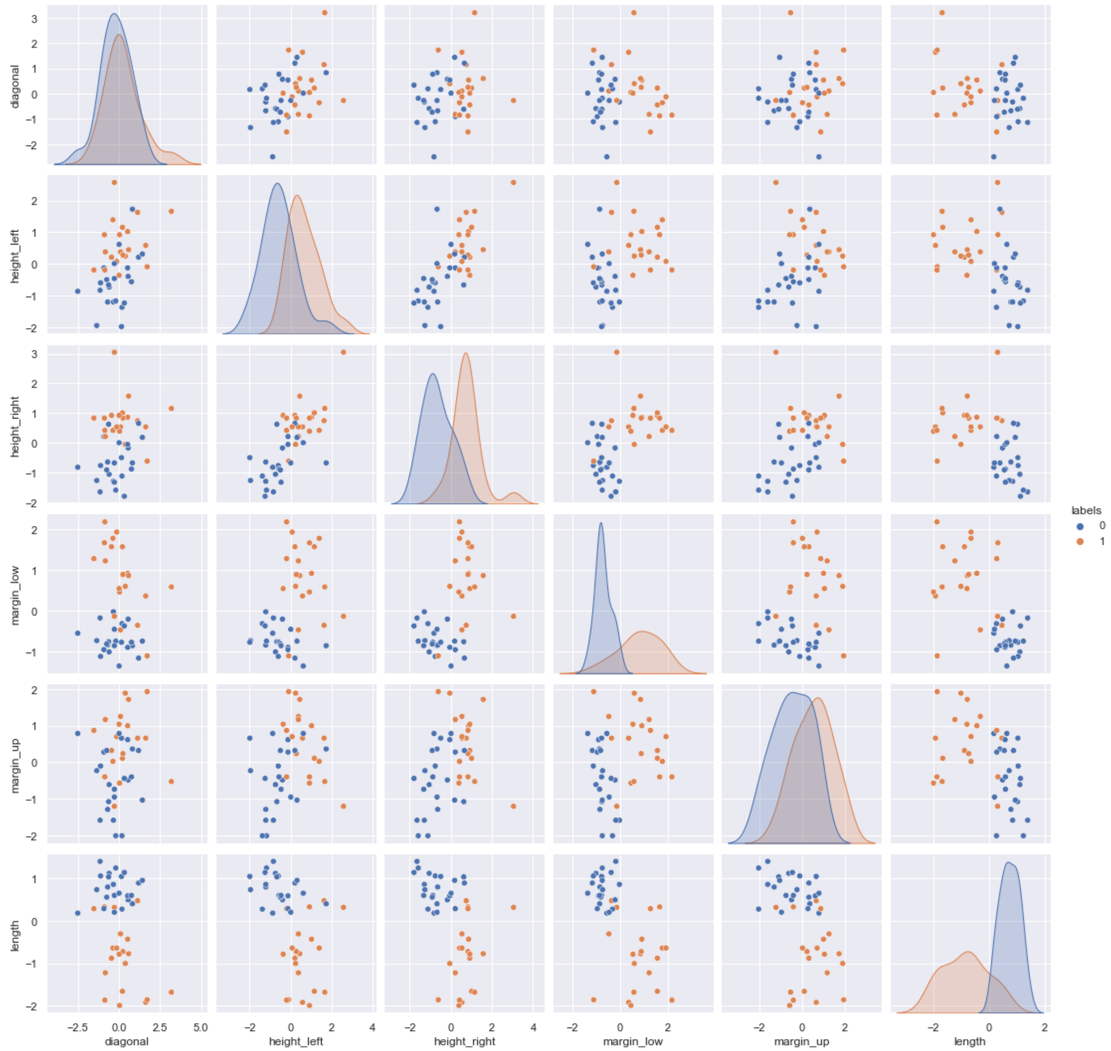
Creation de 4 dataframes différents:

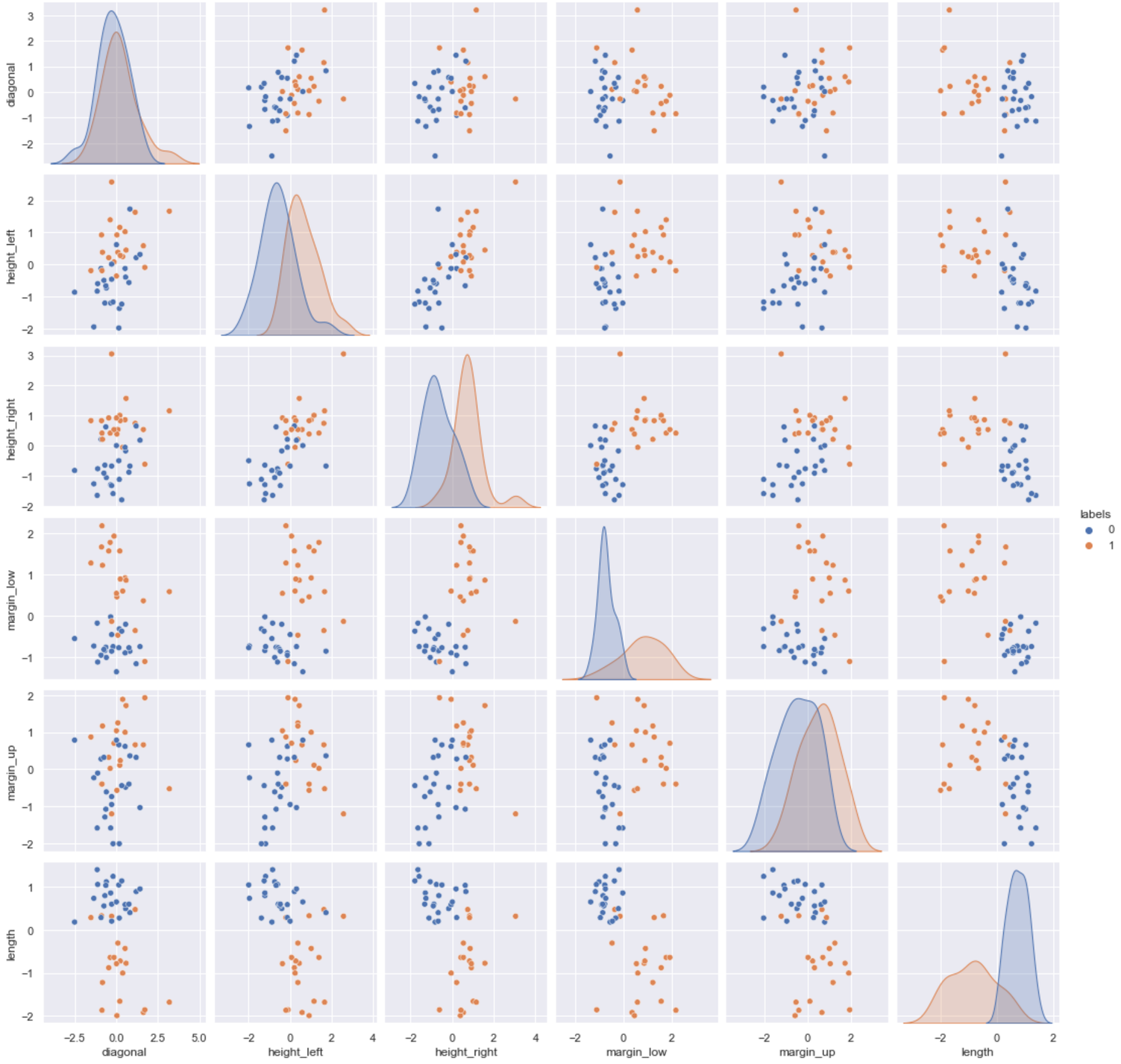
1. x_train : 75% valeurs (sauf valeurs ‘vrai ‘ et ‘faux’)
2. x_test : les valeurs restantes (25%)
3. y_train: valeurs ‘vrai’ et ‘faux’ de x_train
4. y_test: valeurs ‘vrai’ et ‘faux’ de x_test

Comparer la performance de l’algorithme de classification et l’algorithme de prediction.

kMeans Clustering

Modélisation des données de test




diagonal
height_left
height_right
margin_low
margin_up
length
labels
0
1

kMeans Clustering

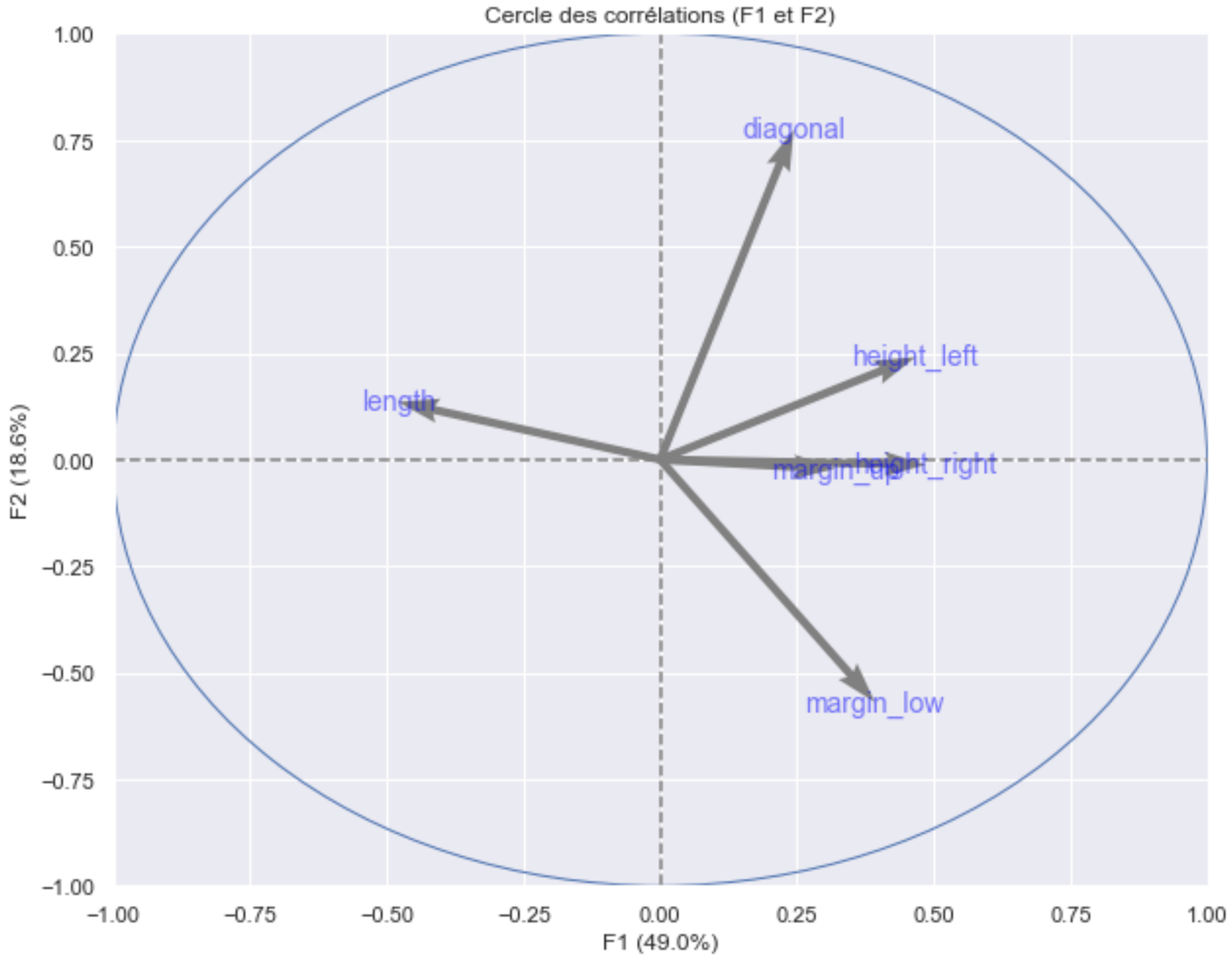
Matrice de corrélation

On remarque :

- La longueur est corrélé négativement à toute les autres variables.
- La diagonale est corrélé négativement à la marge basse.
- Les autres valeurs sont corrélé positivement entre elle.

kMeans Clustering

Analyse en Composantes Principales

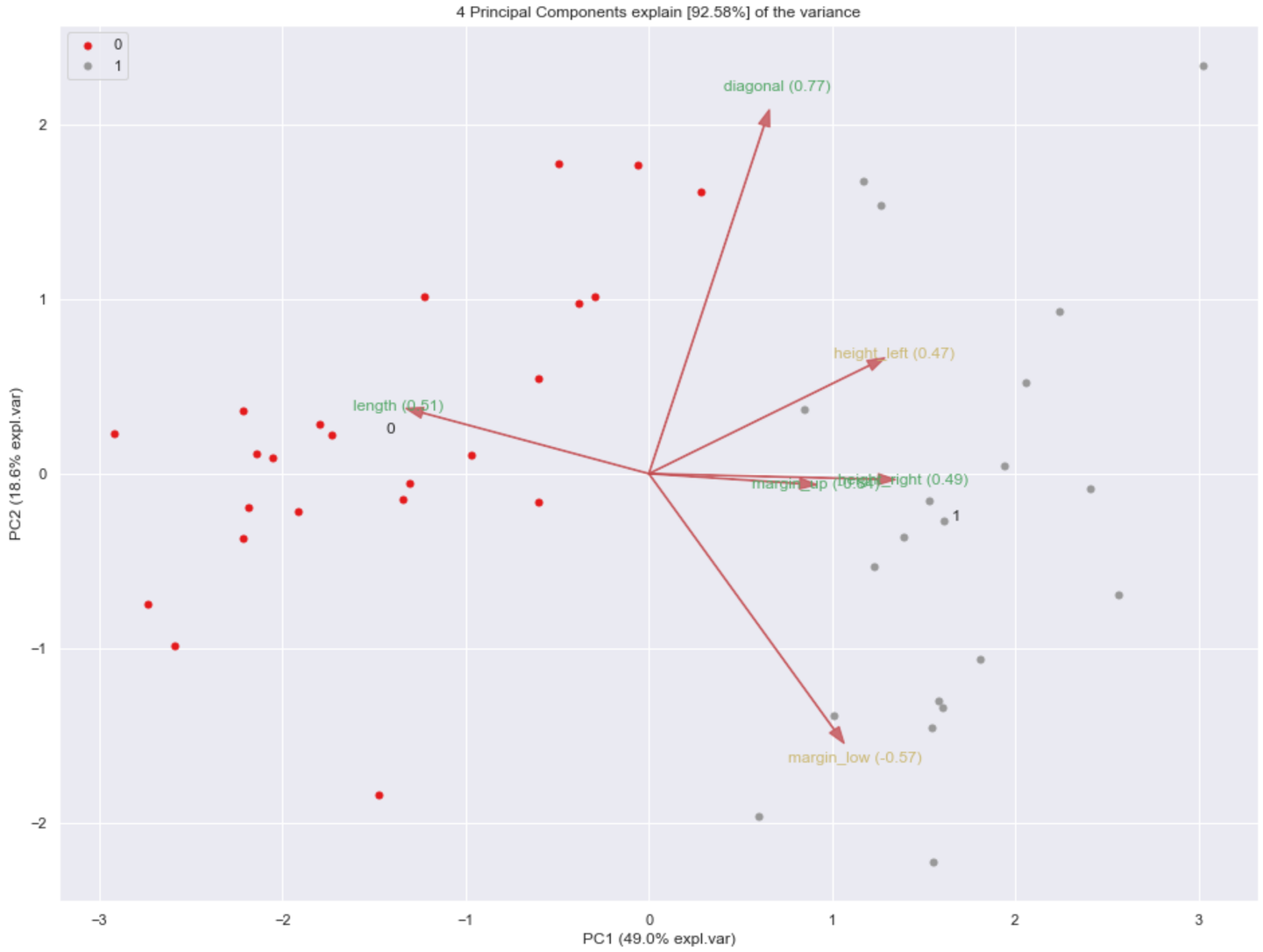


On remarque :

- La variable longueur est corrélé négativement a la marge haute et la hauteur droite.
- La diagonale est corrélé au plan F2 (75%).
- Les variables, excepté la longueur, sont fortement corrélés au plan F1.

kMeans Clustering

Analyse en Composantes Principales



On remarque :

- Le cluster 0 est fortement corrélé à la longueur.
- Le cluster 1 est corrélé au 1e plan et ses variables.

kMeans Clustering

Analyse de l'algorithme de classification

Pour analyser le résultat de kMeans, nous utilisons le Rand Index.

Celui-ci va mesurer les similarités entre les clusters calculés et les véritables valeurs.

```
Rand Index is: 0.9091915836101883
Adjusted Rand Index is: 0.8183945257891252
```

Le Rand Index ajusté (0.82) obtenu nous permet de confirmer que le model contient des similarités avec les réels clusters mais les résultats ne sont pas parfait. (18% de marge d'erreur)

Regression Logistique

Regression Logistique

Choix des valeurs

Nous effectuons la Regression Logistique sur le meme jeux de données qu'avec kMeans.

Ainsi le model est entrainé avec x_train et y_train. Puis nous l'essayons sur x_test et y_test.

De la meme manière que kMeans, nous comparons les clusters et les valeurs réel.

```
Le score de l'algorithme est : 0.9767441860465116
```

Regression Logistique

Verification des données

	coef	std err	t	P>|t|	[0.025	0.975]
diagonal	-0.1369	0.068	-2.018	0.051	-0.274	0.001
height_left	0.1356	0.147	0.921	0.363	-0.163	0.434
height_right	-0.0874	0.131	-0.668	0.508	-0.352	0.177
margin_low	-0.3628	0.056	-6.534	0.000	-0.475	-0.250
margin_up	-0.4967	0.137	-3.629	0.001	-0.774	-0.219
length	0.1985	0.036	5.560	0.000	0.126	0.271

En regardant la p-valeur (<0.05), on voit que les marges ainsi que la longueur on un impact sur la véracité des billets.

Regression Logistique

Prediction sur les données

	precision	recall	f1-score	support
0	1.00	0.99	0.99	70
1	0.99	1.00	1.00	100
accuracy			0.99	170
macro avg	1.00	0.99	0.99	170
weighted avg	0.99	0.99	0.99	170

Quand le model est testé sur le total des données, celui-ci obtient un score de 0.99, ce qui est très performant.

Algorithme de prédiction

Merci