



Office central pour la répression du faux monnayage

Détection de faux billets

Pour rappel : Office central pour la répression du faux monnayage

❖ Création : 11/09/2011

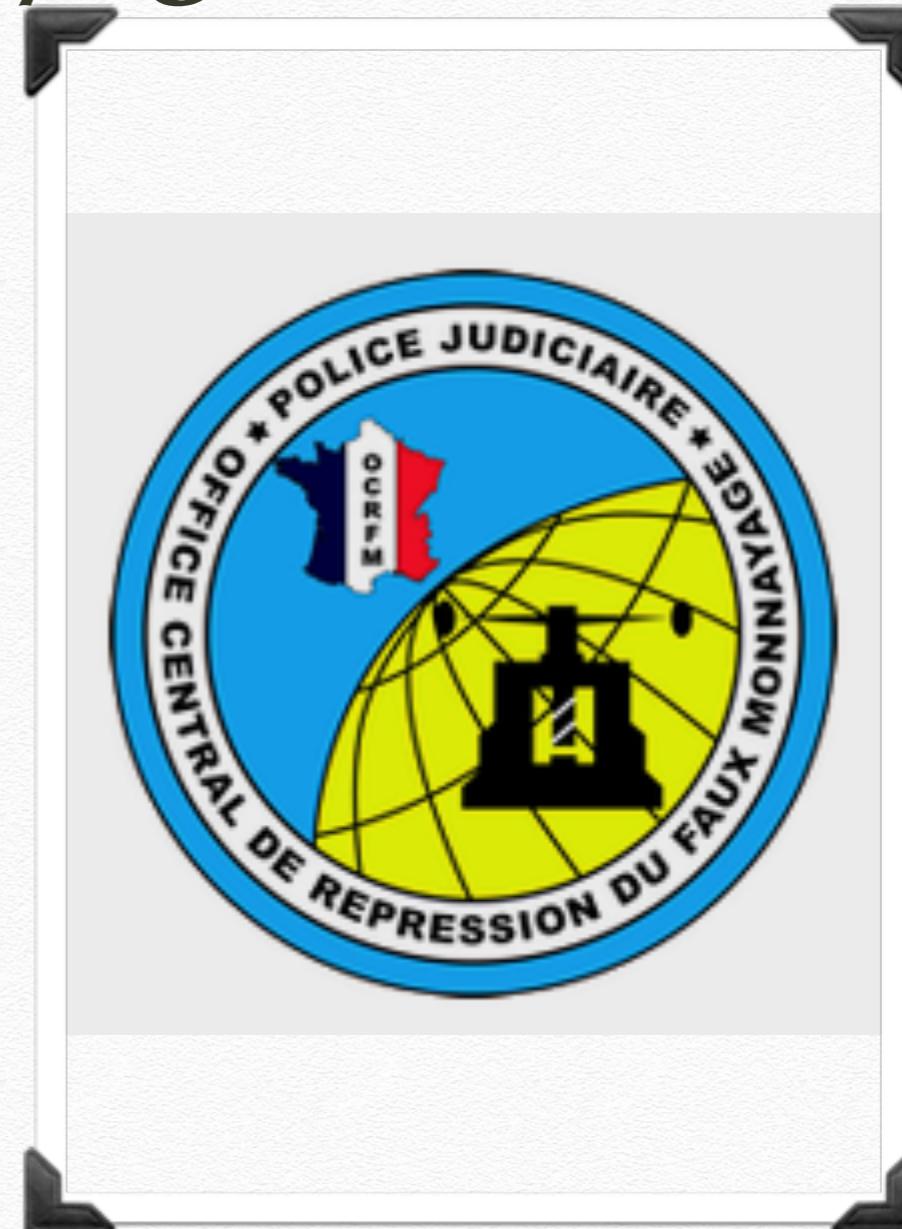
❖ Mission : Centraliser les renseignements pouvant faciliter :

- les recherches
- la prévention
- la répression

... sur le faux monnayage.

❖ Moyens (applications informatiques):

- Le répertoire automatisé pour l'analyse des contrefaçons de l'euro (RAPACE)
- Le fichier national du faux monnayage (FNFM)



Objectif : Créer un algorithme de détection de faux billets

- ❖ Etude du jeu de données d'entraînement
 - Présentation du jeu de donnée & analyses univariée et bivariée.
 - Analyse multivariée avec l'ACP.
- ❖ Classification des billets
 - Classification (KMeans) intuitive des billets
 - Prédiction du type de billet (Modèle de Régression logistique)

Etude du jeu de données d'entraînement

Présentation du jeu de données

- ❖ Données sur 170 billets
- ❖ Les caractères des billets :
 - Is_genuine : indique si un billet est vrai ou faux
 - Length : longueur du billet en mm
 - Height_left : hauteur mesurée à gauche du billet (en mm)
 - Height_right : hauteur mesurée à droite du billet (en mm)
 - Margin_low : La marge entre le bord inférieur du billet et l'image de celui-ci (en mm)
 - Margin_up : La marge entre le bord supérieur du billet et l'image de celui-ci (en mm)
 - Diagonal : la diagonal du billet (en mm)

Présentation du jeu de données

Vérification de la présence de valeurs manquantes

```
sys.path.append('..../project_5_prod_market_study/code')
```

```
import my_functions_revue as mfct
```

```
mfct.verif_presence_nan_in_df(data, 'data')
```

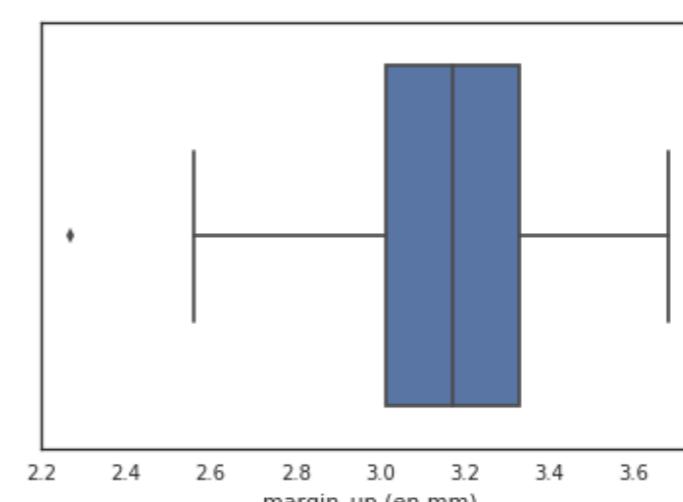
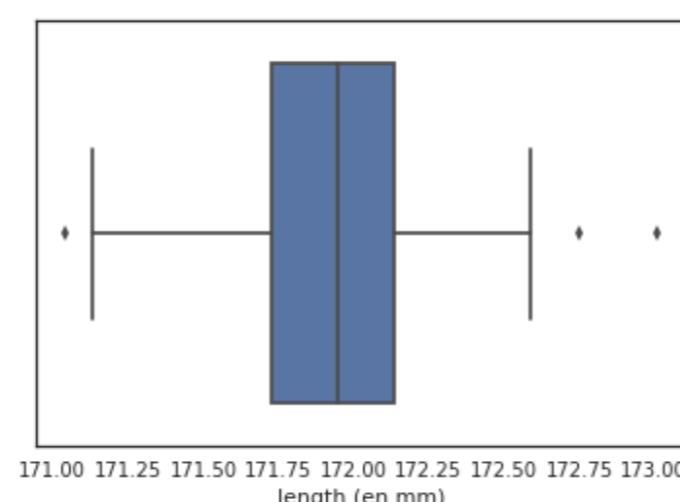
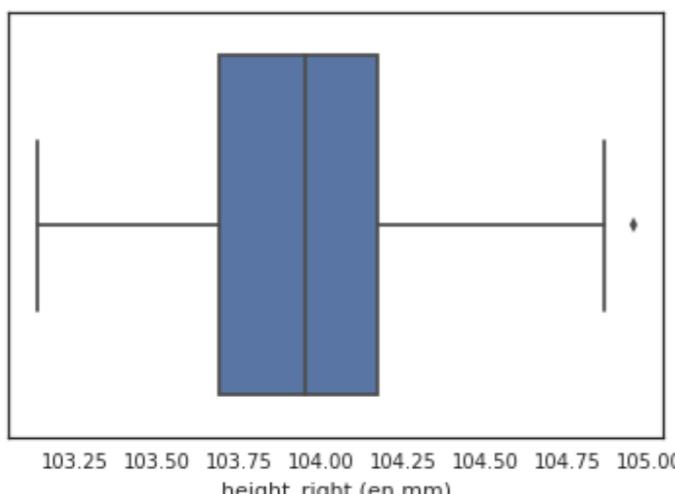
Il n'y a pas de valeur manquante dans data

Vérification de la présence de doublons

```
mfct.verif_doublon(data, 'data')
```

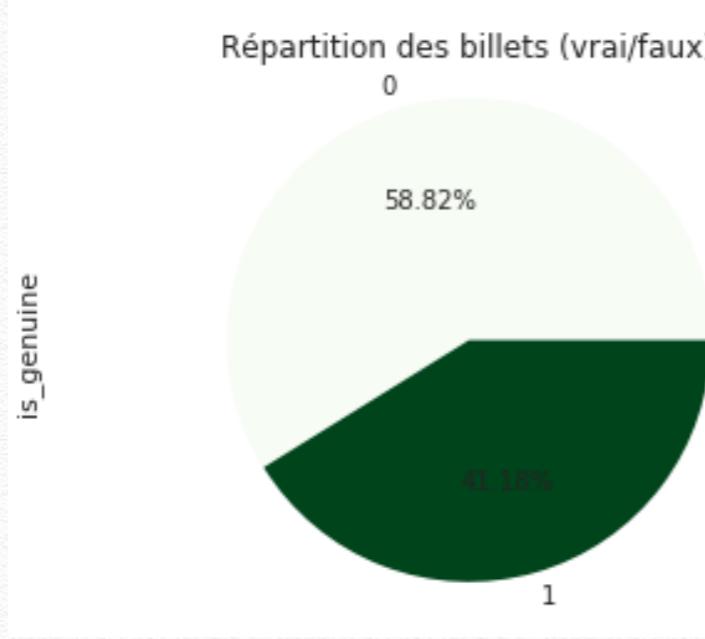
Absence de doublon, il n'y a pas de retraitement à faire pour data

Vérification de valeurs aberrantes : Présence de quelques unes

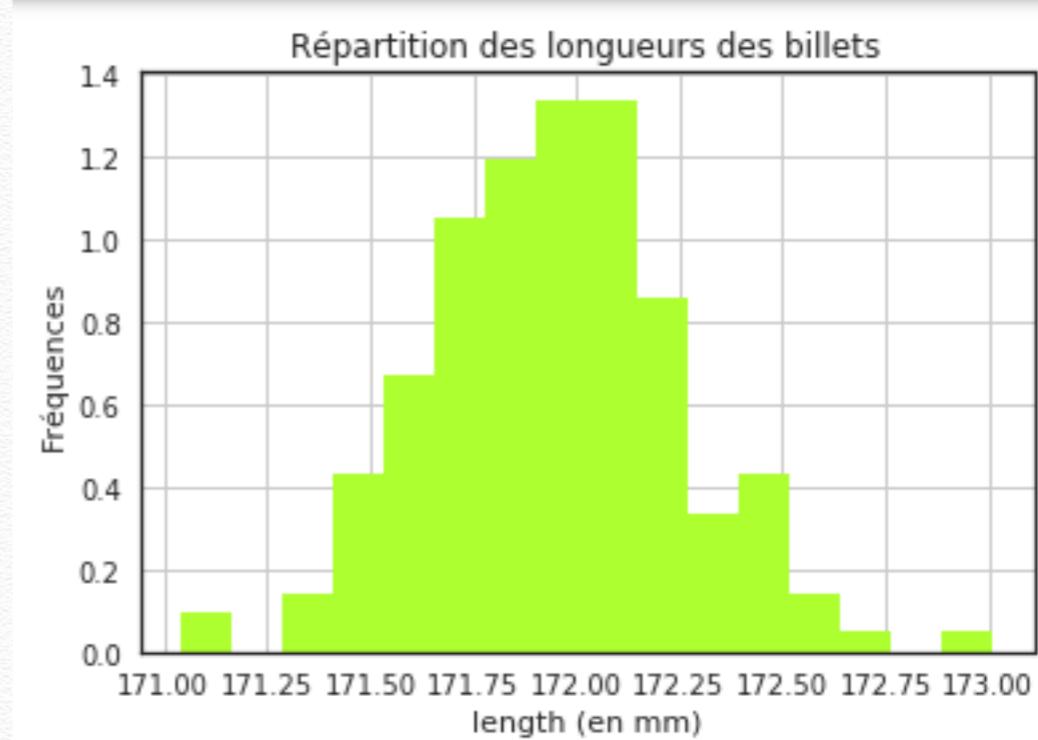


Analyses univariées

Type vrai/faux



Longueur



Variable is_genuine :

- Mode = 0 True

is_genuine	n	f
0	True	100 0.59
1	False	70 0.41

Variable length :

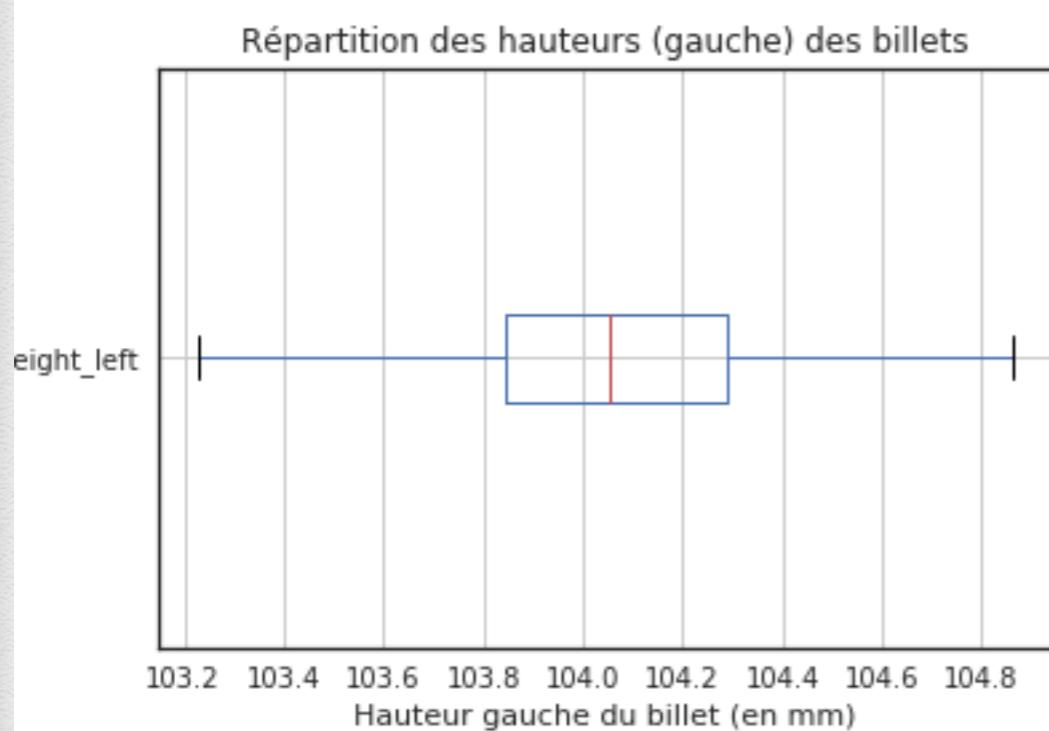
- Moyenne = 171.94
- Médiane = 171.945

Variable length :

- Variance = 0.093
- Ecart-type = 0.304

Analyses univariées

Hauteur gauche



La mediane est 104.055, Q1 est égal à 103.842 et Q3 est égal à 104.288

L'écart inter-quartile est égal à 0.44 et les bornes sont respectivement de 103.23 à 104.86

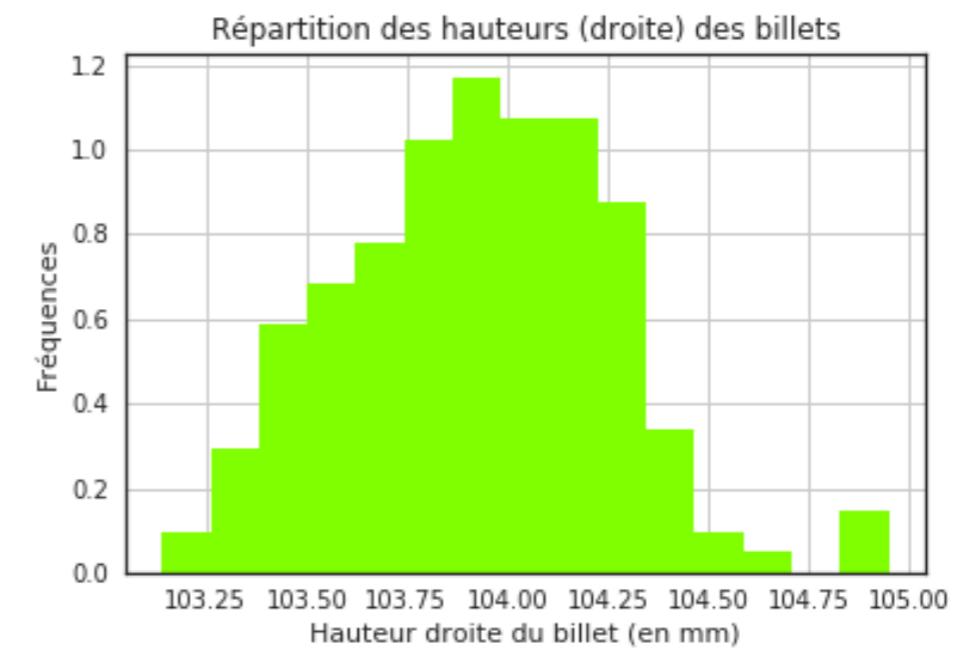
Variable height_left :

- Moyenne = 104.066
- Médiane = 104.055

Variable height_left :

- Variance = 0.088
- Ecart-type = 0.297

Hauteur droite



Variable height_right :

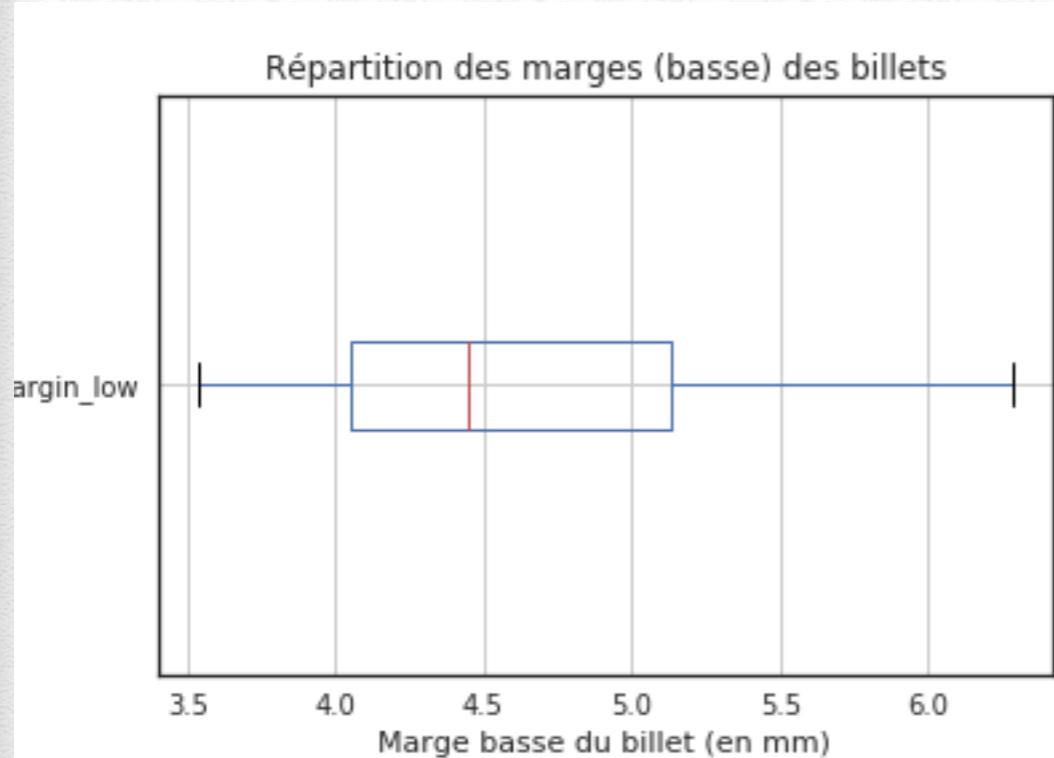
- Moyenne = 103.928
- Médiane = 103.95

Variable height_right :

- Variance = 0.109
- Ecart-type = 0.330

Analyses univariées

Marge basse



La médiane est 4.45, Q1 est égal à 4.05 et Q3 est égal à 5.13

L'écart inter-quartile est égal à 1.077 et les bornes sont respectivement de 3.54 à 6.28

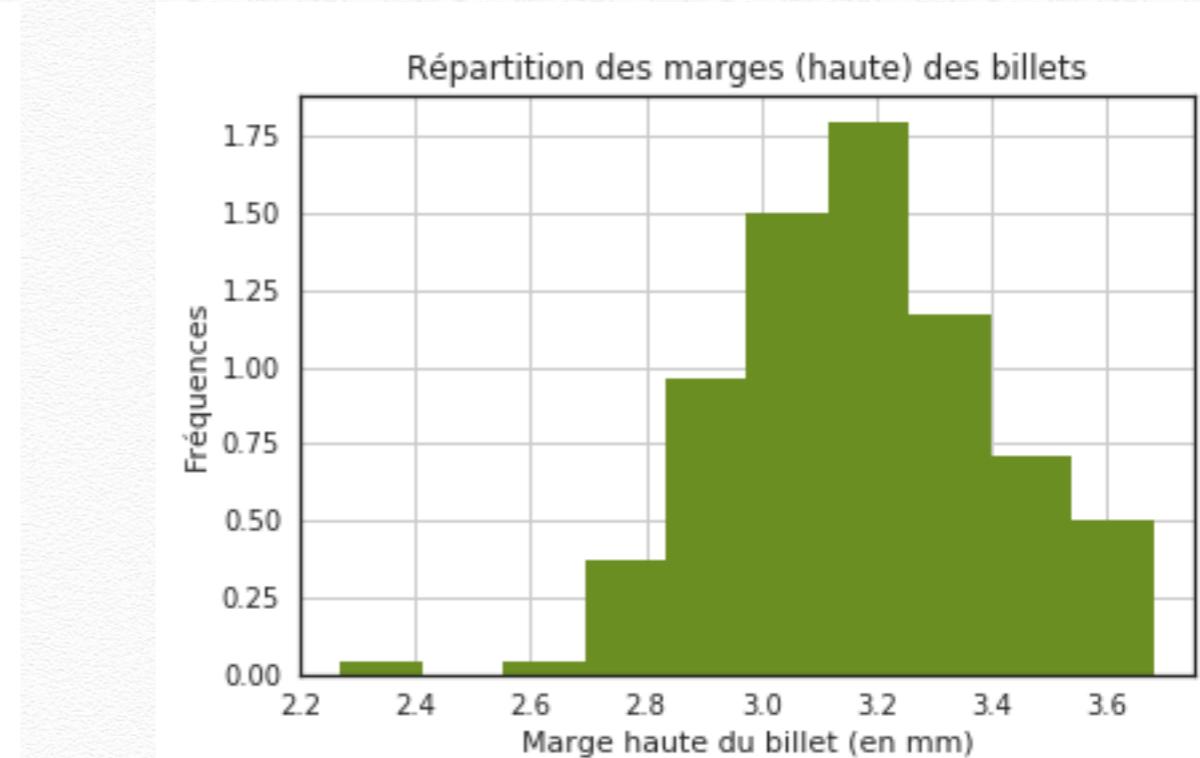
Variable margin_low :

- Moyenne = 4.612
- Médiane = 4.45

Variable margin_low :

- Variance = 0.490
- Ecart-type = 0.700

Marge haute



Variable margin_up :

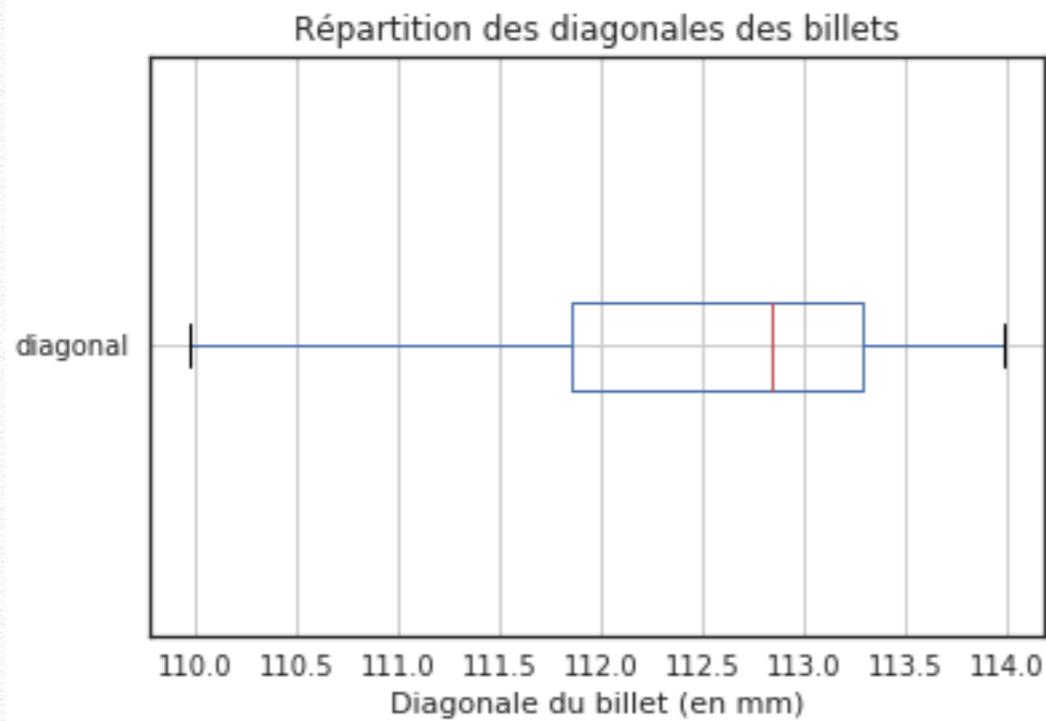
- Moyenne = 3.17
- Médiane = 3.17

Variable margin_up :

- Variance = 0.056
- Ecart-type = 0.235

Analyses univariées

Diagonale



La médiane est 112.845, Q1 est égal à 111.855 et Q3 est égal à 113.288

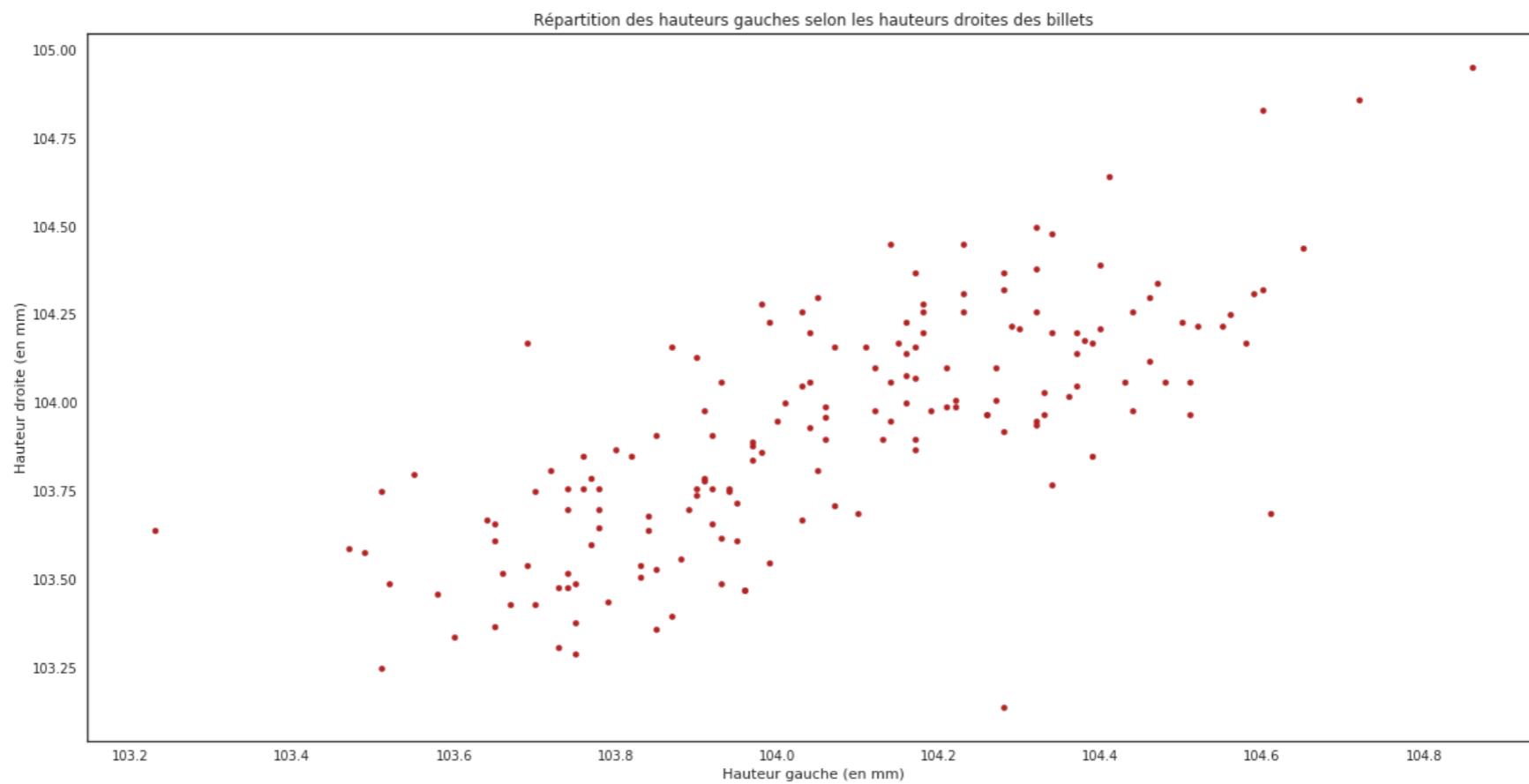
L'écart inter-quartile est égal à 1.433 et les bornes sont respectivement de 109.97 à 113.98
Variable diagonal :

- Moyenne = 112.570
- Médiane = 112.845

Variable diagonal :

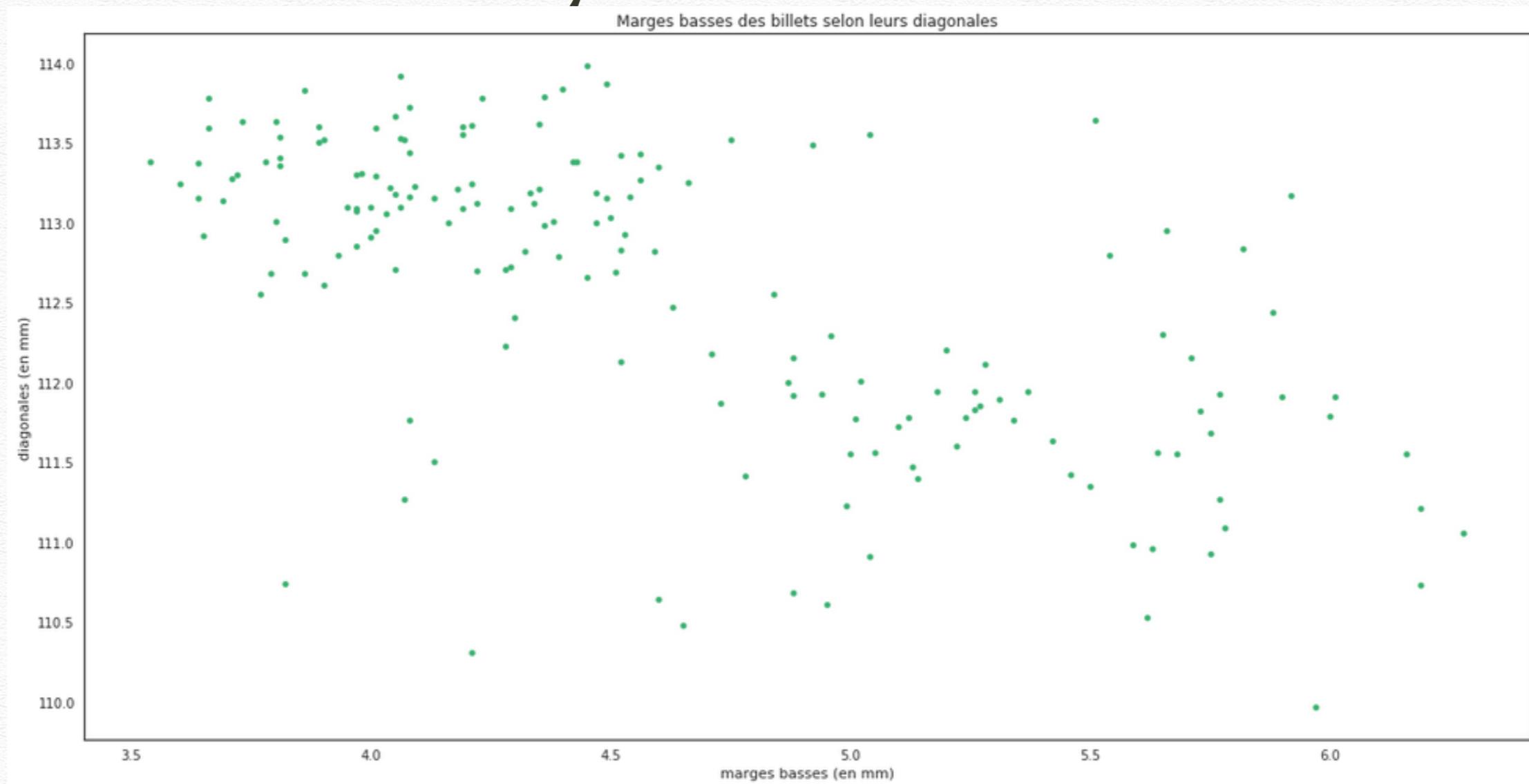
- Variance = 0.850
- Ecart-type = 0.922

Analyses bivariées



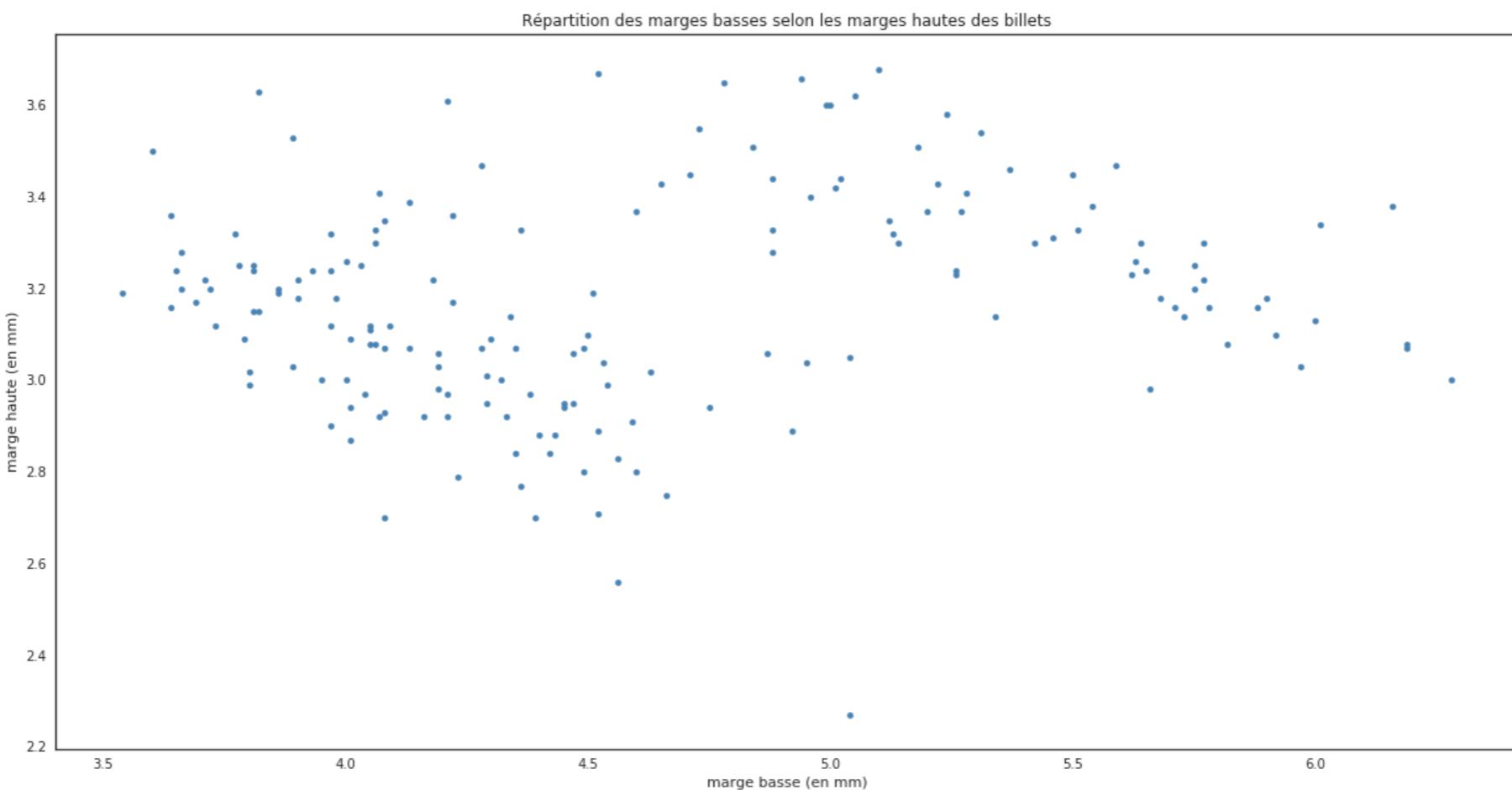
Le coefficient de corrélation (Pearson) est égal à 0.73
Les variables sont corrélées positivement car 0.73 est supérieur à 0,60

Analyses bivariées



Le coefficient de corrélation (Pearson) est égal à -0.64
Les variables sont négativement corrélées car -0.64 est inférieur à -0,60

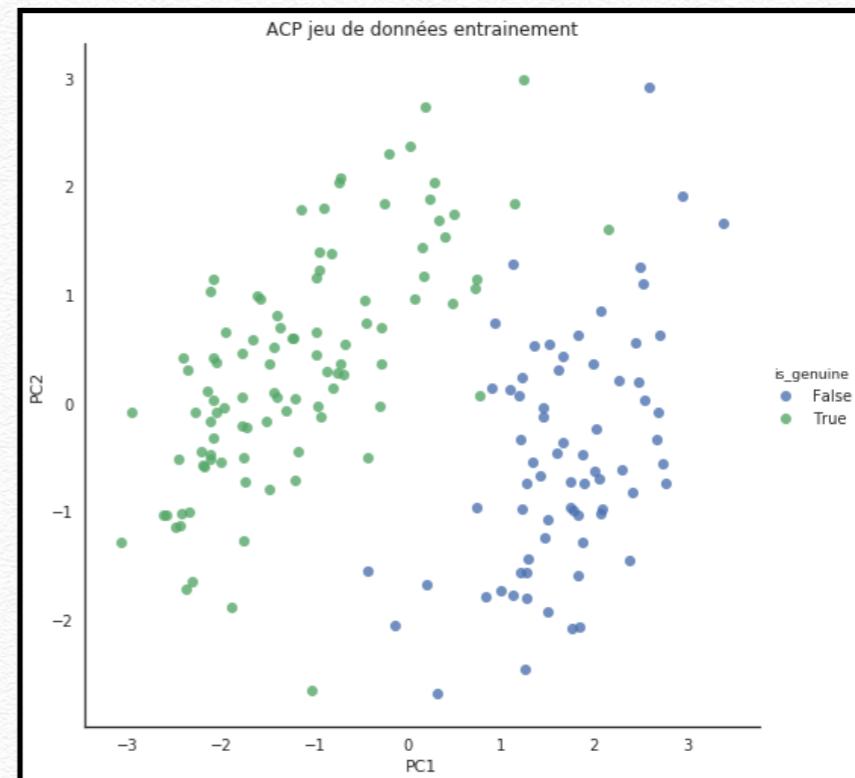
Analyses bivariées



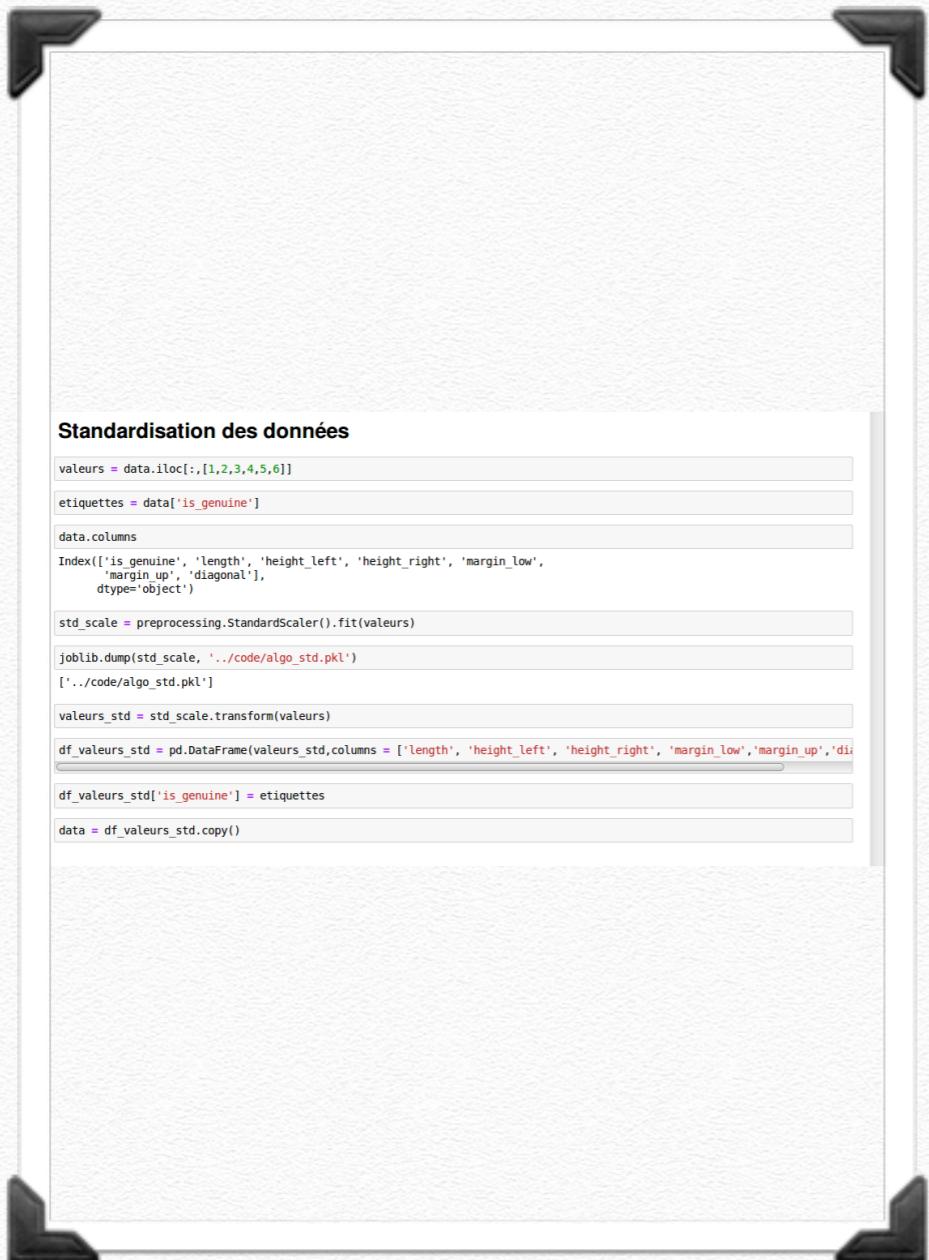
Le coefficient de corrélation (Pearson) est égal à 0.17
Les variables ne sont pas positivement corrélées car 0.17 est inférieur à 0,40

Analyse multivariée : méthode de l'analyse par composante principale.

- ❖ Objectif : Résumer l'information contenu dans les sept caractères contenus dans notre jeu de données à l'aide de deux axes principaux.



Analyse par composante principale (ACP) : Au préalable



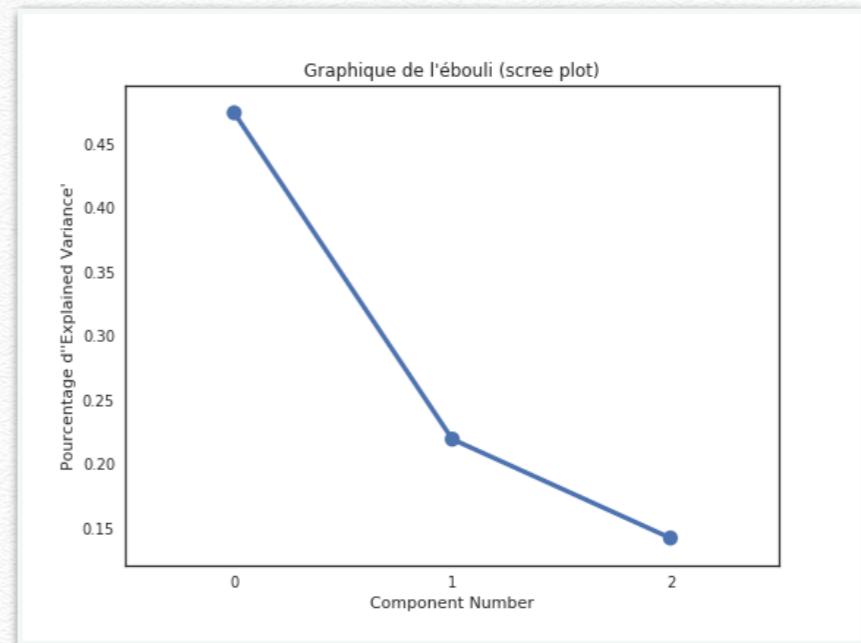
The screenshot shows a Jupyter Notebook cell with the title "Standardisation des données". The code within the cell is as follows:

```
valeurs = data.iloc[:,[1,2,3,4,5,6]]  
etiquettes = data['is_genuine']  
  
data.columns  
Index(['is_genuine', 'length', 'height_left', 'height_right', 'margin_low',  
       'margin_up', 'diagonal'],  
      dtype='object')  
  
std_scale = preprocessing.StandardScaler().fit(valeurs)  
  
joblib.dump(std_scale, '../code/algo_std.pkl')  
['../code/algo_std.pkl']  
  
valeurs_std = std_scale.transform(valeurs)  
  
df_valeurs_std = pd.DataFrame(valeurs_std,columns = ['length', 'height_left', 'height_right', 'margin_low','margin_up','diagonal'])  
  
df_valeurs_std['is_genuine'] = etiquettes  
  
data = df_valeurs_std.copy()
```

- ❖ Standardisation des données (centrer-réduire) : Evite qu'une variable prenne le dessus sur une autre compte tenu de son ordre de grandeur.
- ❖ Pour cela, on centre notre variable (moyenne=0), et on la réduit (écart type =1).

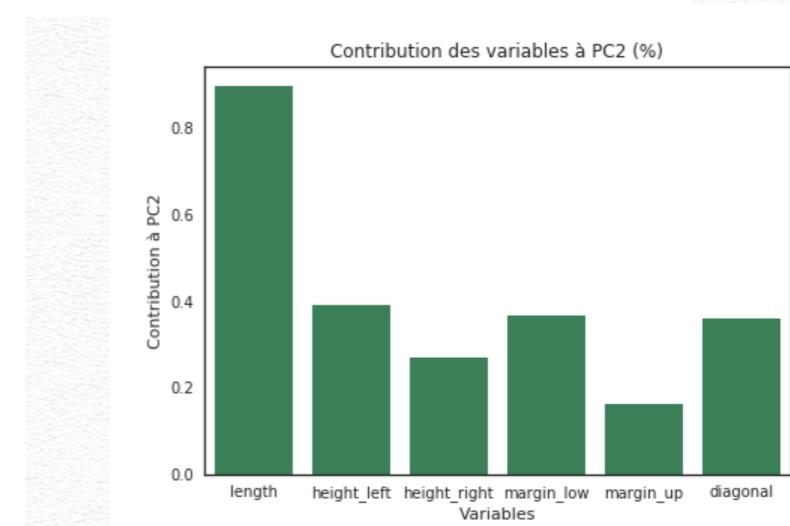
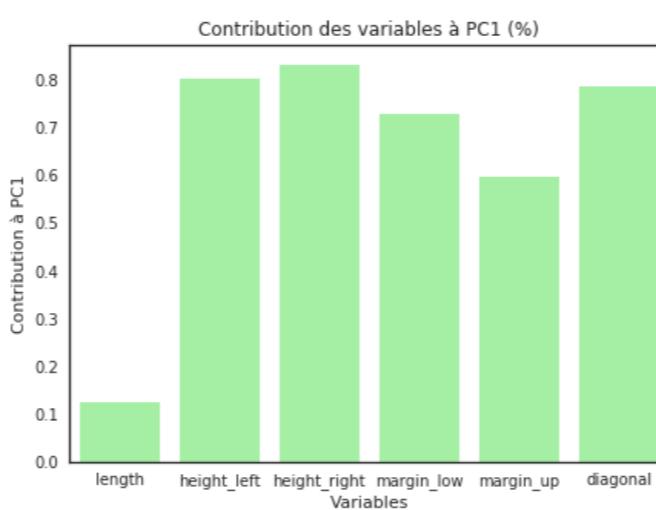
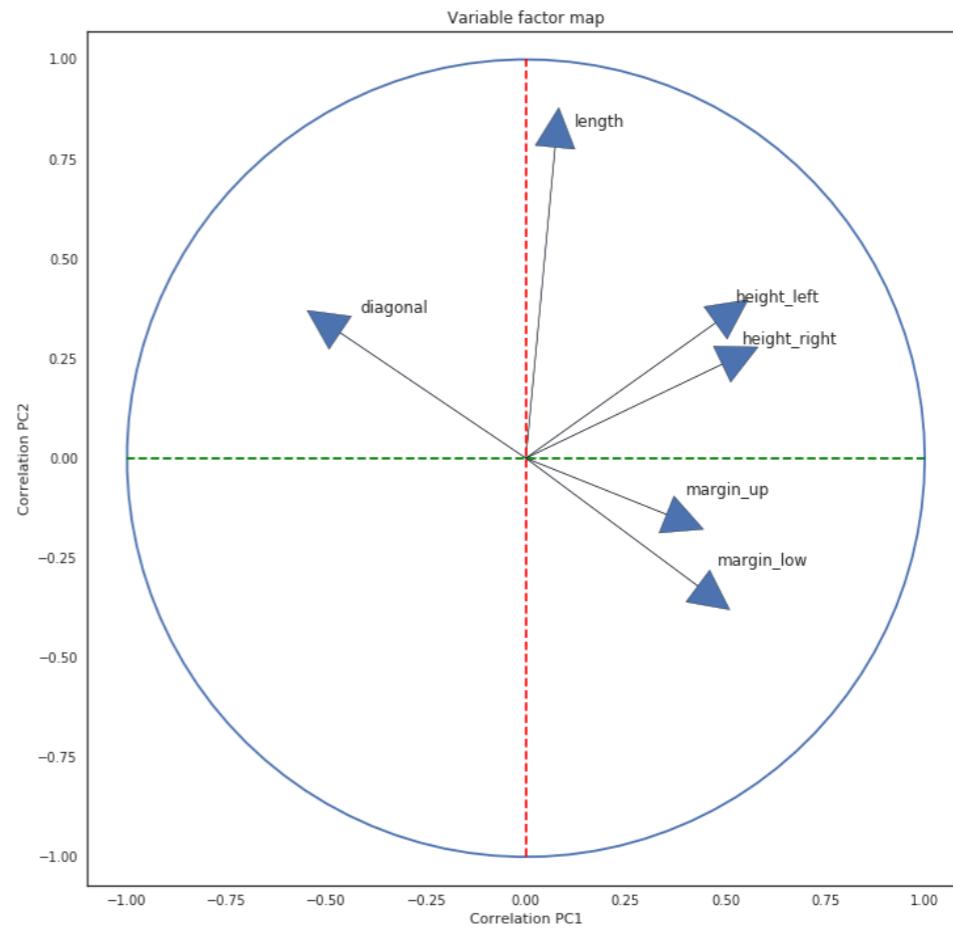
ACP : Choix du nombre de composantes principales (axes)

- ❖ Objectif : Sélectionner les axes qui maximisent la variance (information résumée ou inertie)
- ❖ Cette information est données par la matrice des valeurs propres
- ❖ Ici, les deux premiers axes résument 84% de l'information. Nous les sélectionnerons.



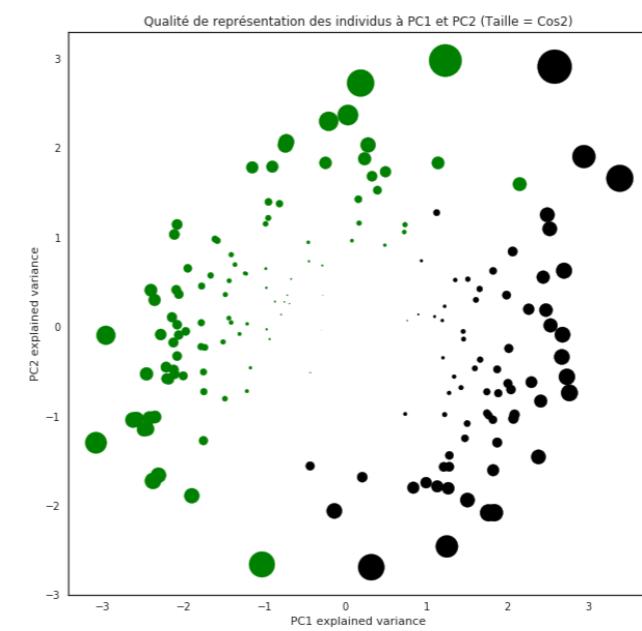
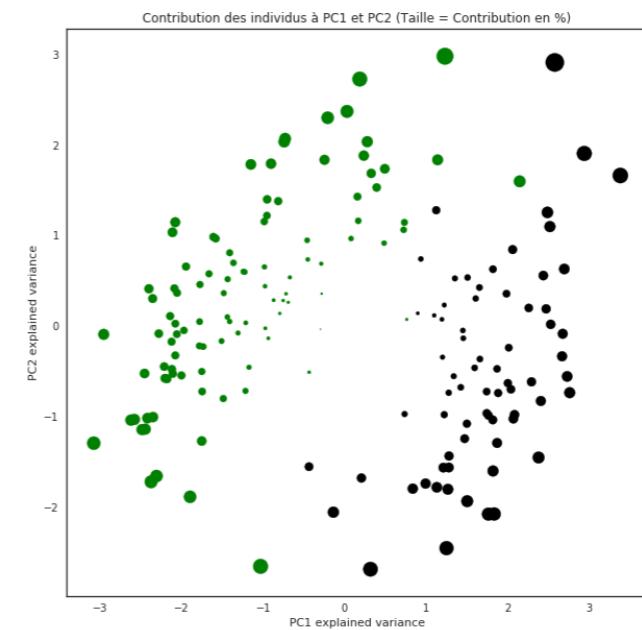
ACP : Cercle des corrélations

- ❖ Objectif : Montrer les corrélations entre les variables et les corrélations entre variables et composantes principales.
- ❖ Si les flèches sont proches entre elles, les variables sont positivement corrélées.
- ❖ Si les flèches négatives corrélées sont regroupées sur le quadrant opposé.
- ❖ Plus une flèche est proche du cercle plus elle est représentée par la composante principale.



ACP : Contribution et qualité de représentation des individus.

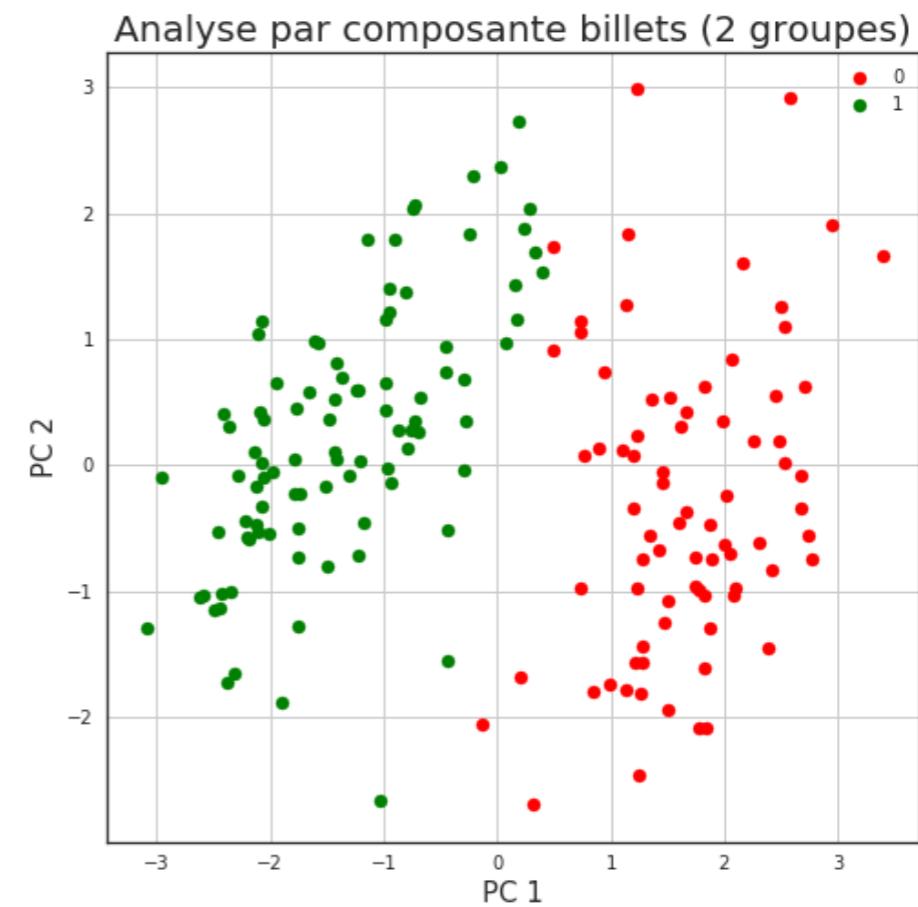
- ❖ En vert, les vrais billets et en noir les faux billets.
- ❖ Plus les points sont gros plus les individus contribuent aux deux composantes principales et sont mieux représentés.



Classification des billets

Classement des billets selon un algorithme non supervisé : Kmeans

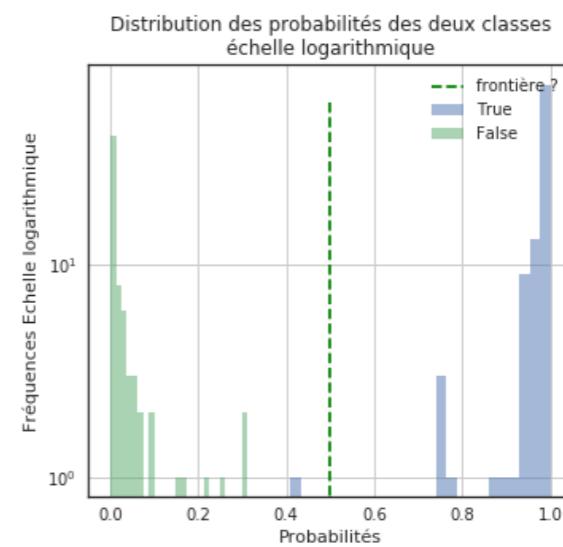
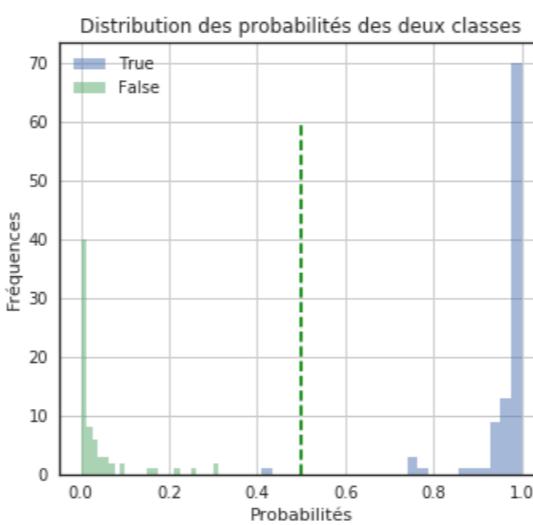
- ❖ Placement de points au hasard dans un plan, puis regroupement des individus autour de ces points.
- ❖ Calcul des centroïdes des groupes formés et regroupement des individus autour de ces centroïdes.
- ❖ Répétition de l'opération jusqu'à ce que les centroïdes ne bougent plus.



cluster ID	0	1	Total
is_genuine			
False	69.0	1.0	70.0
True	7.0	93.0	100.0
total	76.0	94.0	170.0

Construction d'un modèle de prédiction via un modèle de régression logistique.

- ❖ La régression logistique est une méthode permettant de modéliser une variable binaire en fonction de variables explicatives.
- ❖ Objectif : prédire si un billet est faux ou non.



	prédict False	prédict True
vrai False	70	0
vrai True	1	99

Test de l'algorithme de
classification ...

Conclusion

- L'analyse univariée des variables révèle la présences d'outliers.
- La standardisation des données limite l'impact des différents ordres de grandeur.
- L'ACP résume correctement l'information via un premier axe qui compile l'ensemble des variables (sauf « longueur ») et un second axe représentant la longueur des billets
- De manière non supervisée, on arrive à dégager deux classes se rapprochant de notre cible à prédire.
- Le modèle de régression logistique détecte facilement les billets.