



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 & analyse 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')
```

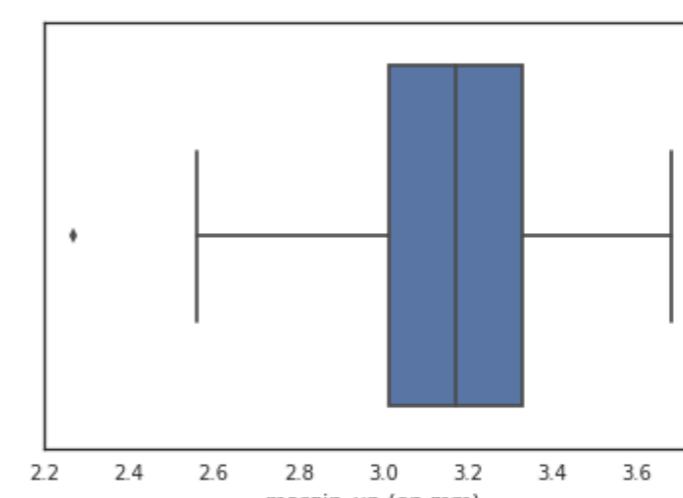
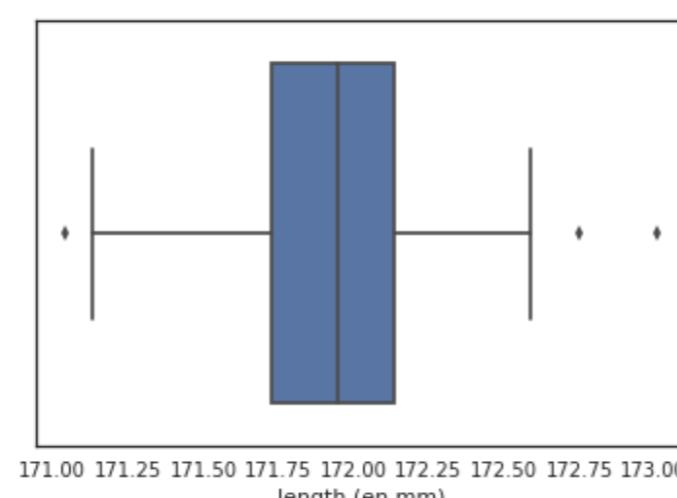
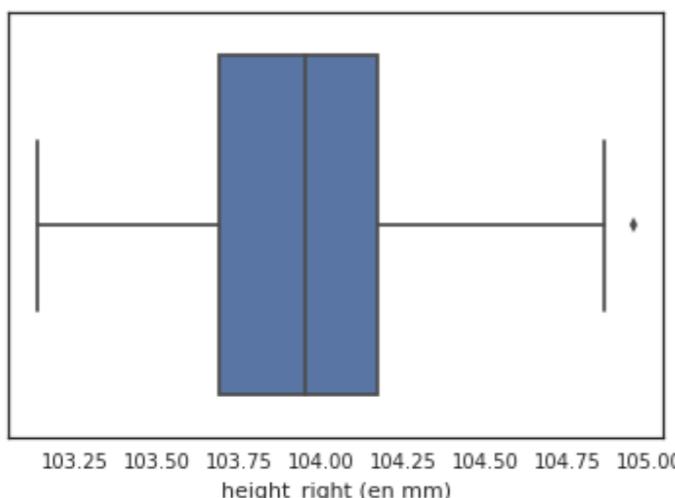
Il n'y a pas de valeur manquante dans data

Vérification de la présence de doublons

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

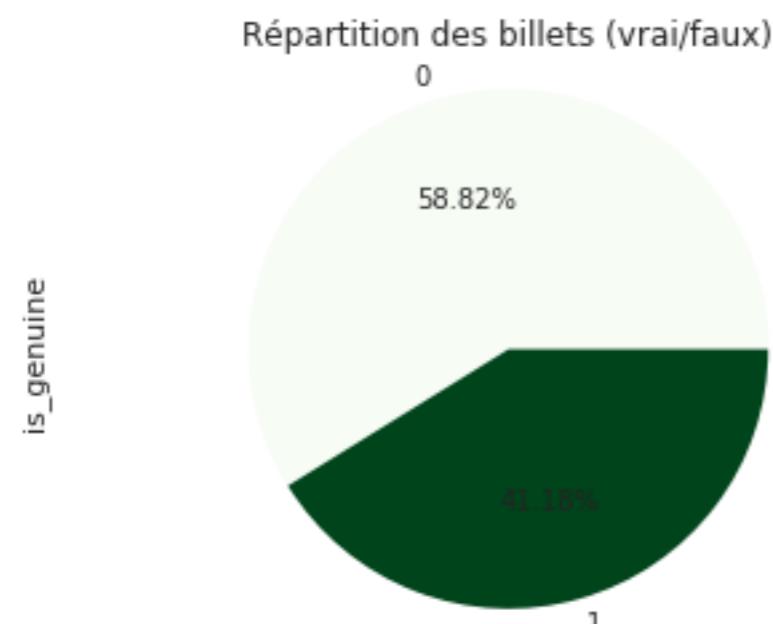
Absence de doublon, il n'y a pas de retraitement à faire pour data

Vérification de valeurs aberrantes : Présence de quelques une



Analyses univariées

Is_genuine



Sauvegarder l'image ? (y/n) :n

Pas de sauvegarde

Variable is_genuine :

- Moyenne = 0.5882352941176471
- Médiane = 1.0
- Mode = 0 True

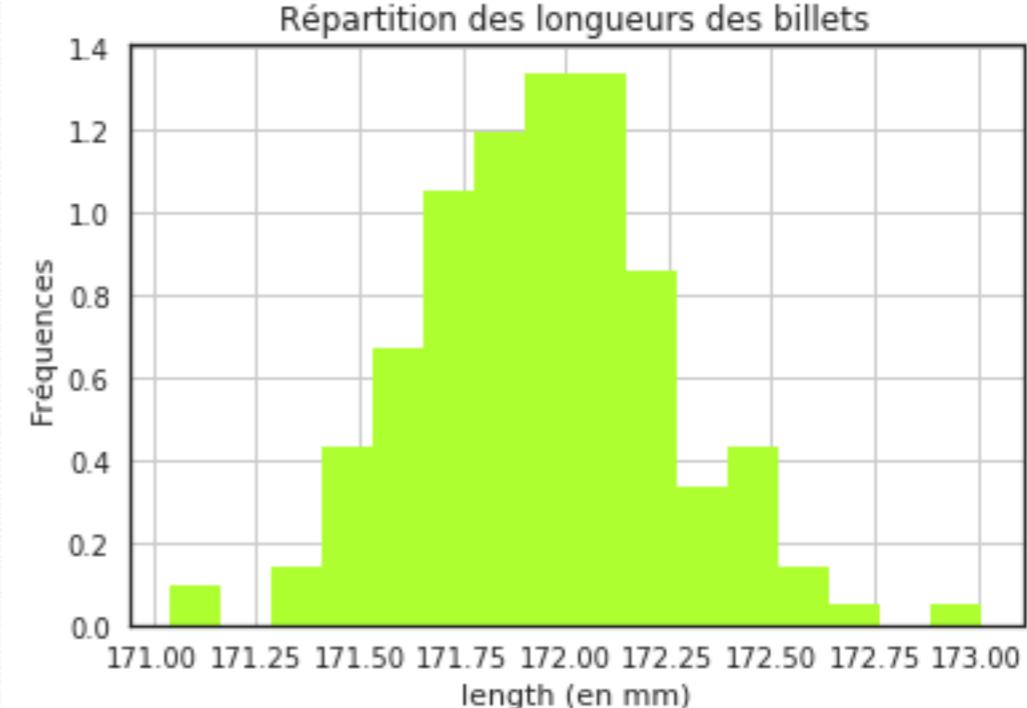
dtype: bool

Variable is_genuine :

- Variance = 0.24221453287197228
- Ecart-type = 0.4921529567847503

is_genuine	n	f
0	True	100 0.588235
1	False	70 0.411765

Length



Sauvegarder l'image ? (y/n) :n

Pas de sauvegarde

Variable length :

- Moyenne = 171.94058823529411
- Médiane = 171.945
- Mode = 0 172.1

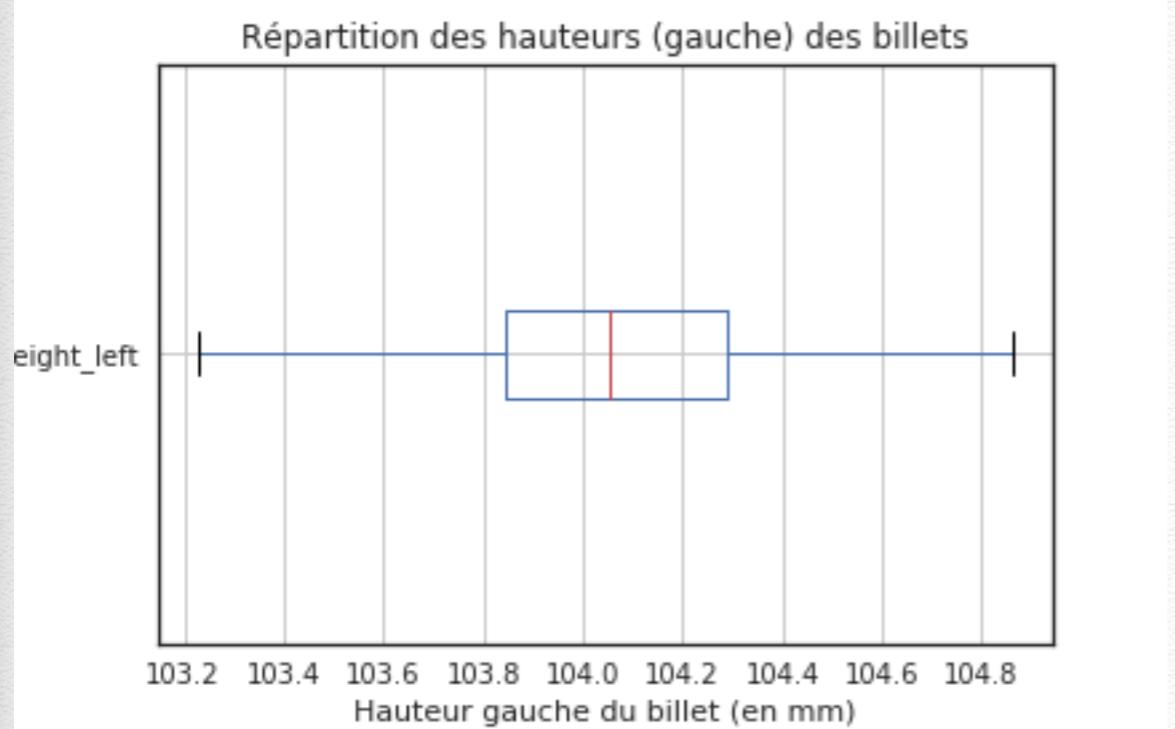
dtype: float64

Variable length :

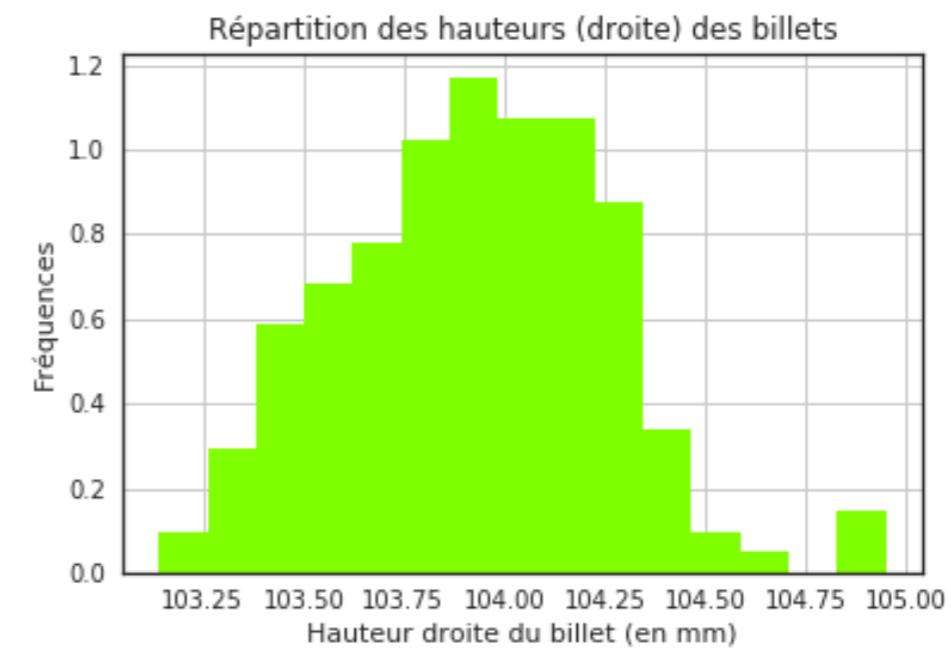
- Variance = 0.09294435986159195
- Ecart-type = 0.30486777439013124

Analyses univariées

Height_left



Height_right



Sauvegarder l'image ? (y/n) :n

Pas de sauvegarde

Variable height_right :

- Moyenne = 103.92811764705881
- Médiane = 103.95
- Mode = 0 103.76

1 104.06

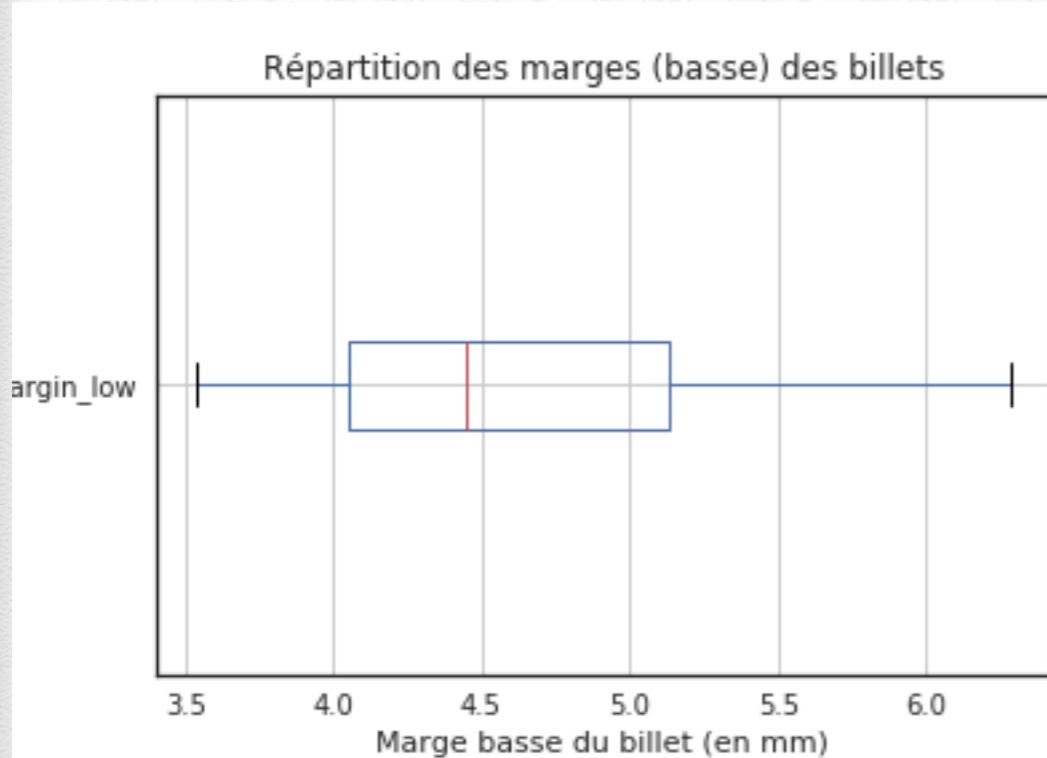
dtype: float64

Variable height_right :

- Variance = 0.10890351557093417
- Ecart-type = 0.330005326579639

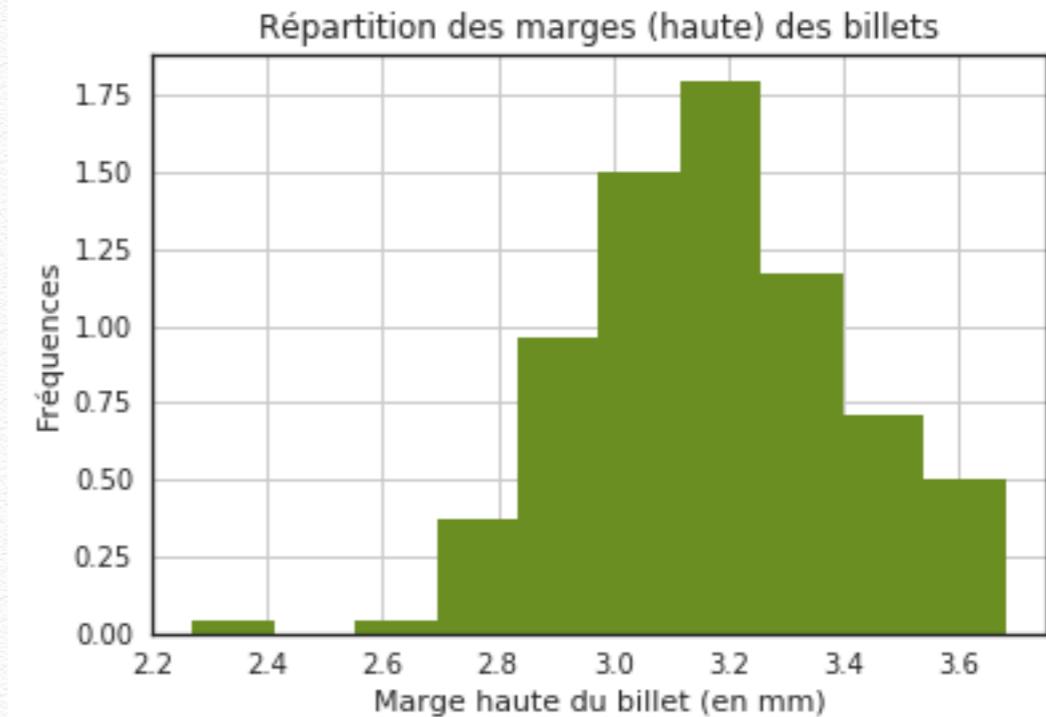
Analyses univariées

Margin_low



```
Sauvegarder l'image ? (y/n) :n
Pas de sauvegarde
La mediane est 4.45, Q1 est égal à 4.05 et Q3 est égal à 5.1275
L'écart inter-quartile est égal à 1.0775000000000006 et les bornes sont respectivement de 3.54 à 6.28
Variable margin_low :
- Moyenne = 4.612117647058823
- Médiane = 4.45
- Mode = 0      3.97
1    4.08
dtype: float64
Variable margin_low :
- Variance = 0.4900484567474049
- Ecart-type = 0.700034611067687
```

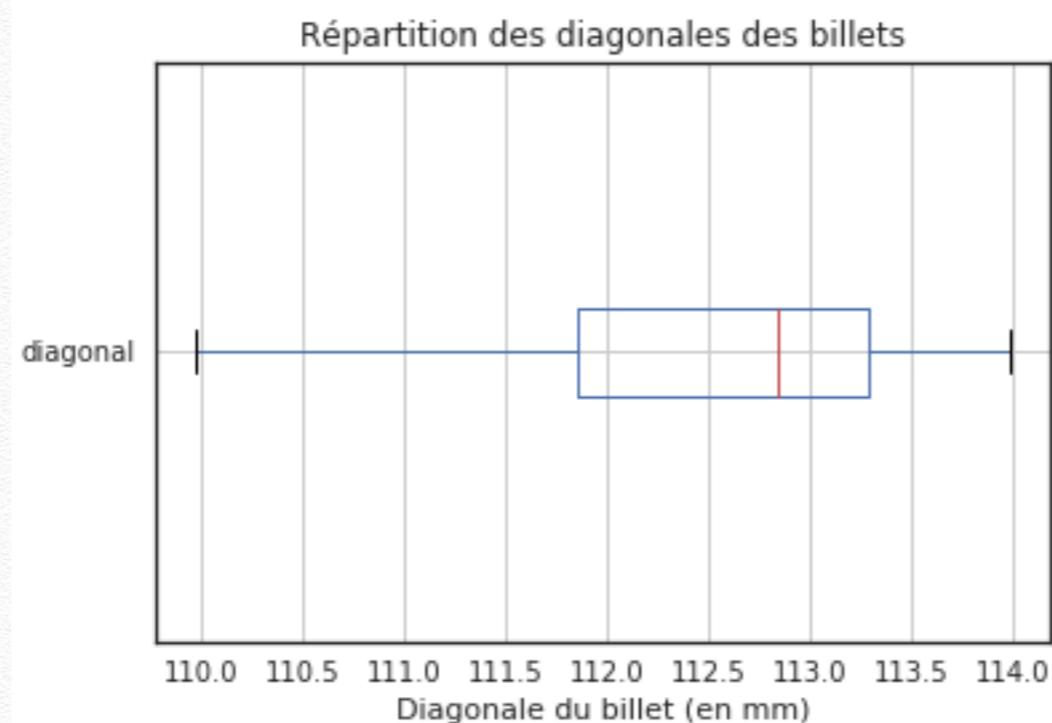
Margin_up



```
Sauvegarder l'image ? (y/n) :n
Pas de sauvegarde
Variable margin_up :
- Moyenne = 3.1704117647058827
- Médiane = 3.17
- Mode = 0      3.07
1    3.24
dtype: float64
Variable margin_up :
- Variance = 0.05553806574394464
- Ecart-type = 0.2356651559818393
```

Analyses univariées

Diagonal



Sauvegarder l'image ? (y/n) :n

Pas de sauvegarde

La mediane est 112.845, Q1 est égal à 111.8549999999999 et Q3 est égal à 113.2875

L'écart inter-quartile est égal à 1.4325000000000045 et les bornes sont respectivement de 109.97 à 113.98

Variable diagonal :

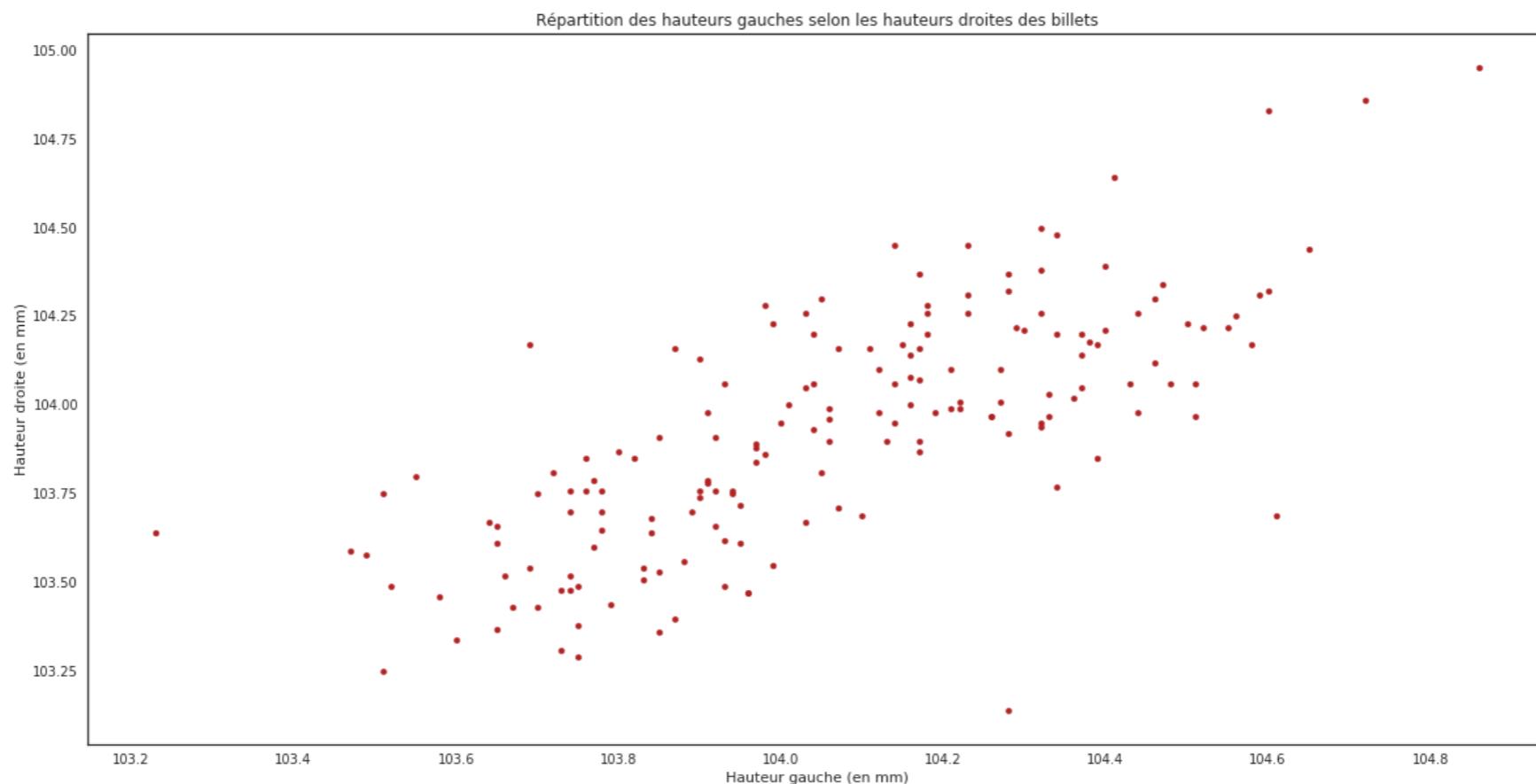
- Moyenne = 112.5704117647059
- Médiane = 112.845
- Mode = 0 113.38

dtype: float64

Variable diagonal :

- Variance = 0.8495768892733561
- Ecart-type = 0.9217249531575871

Analyses bivariées



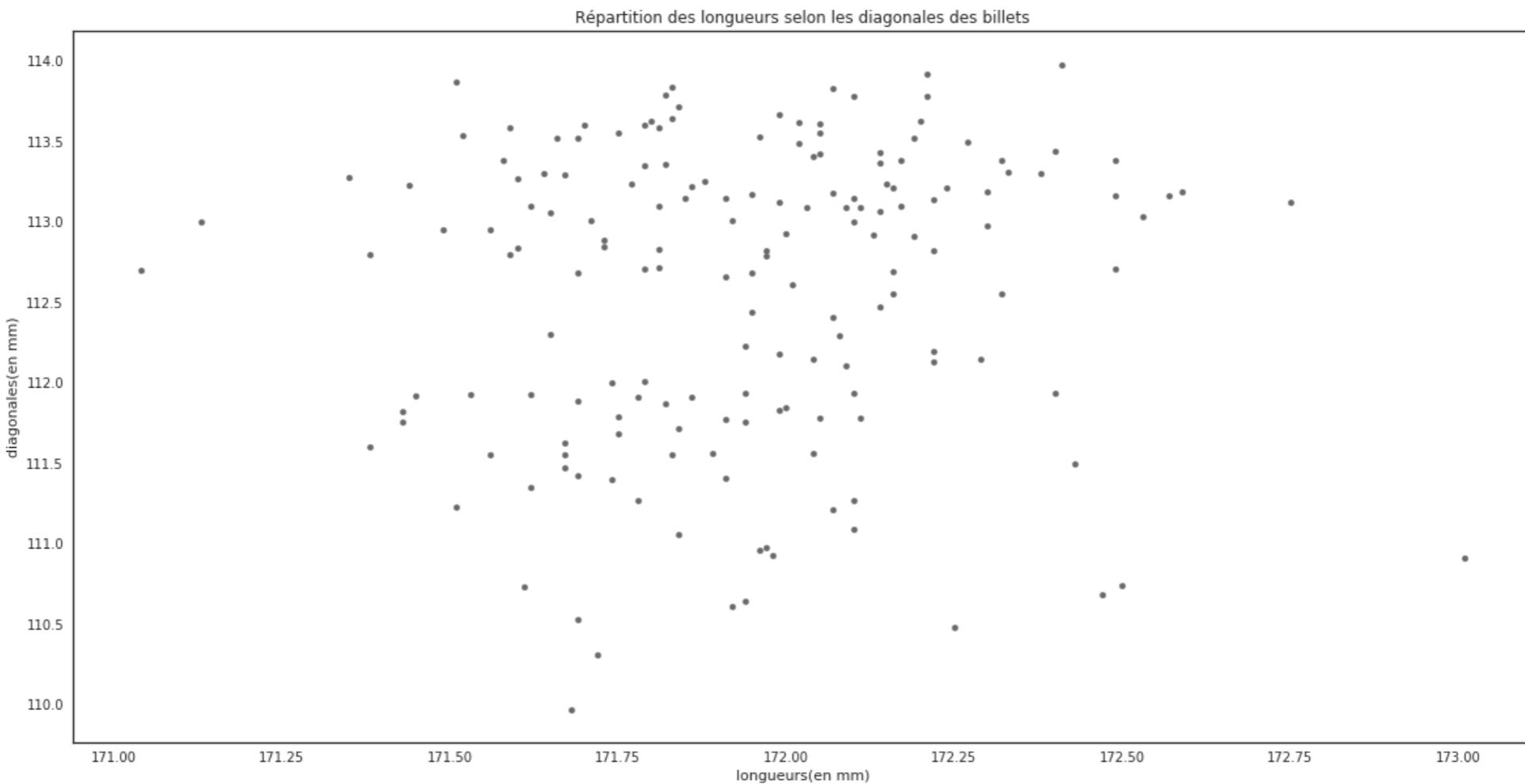
Sauvegarder l'image ? (y/n) :y

Indiquer le chemin du dossier.../presentation/images

Le coefficient de corrélation (Pearson) est égal à 0.73

Les variables sont corrélées

Analyses bivariées



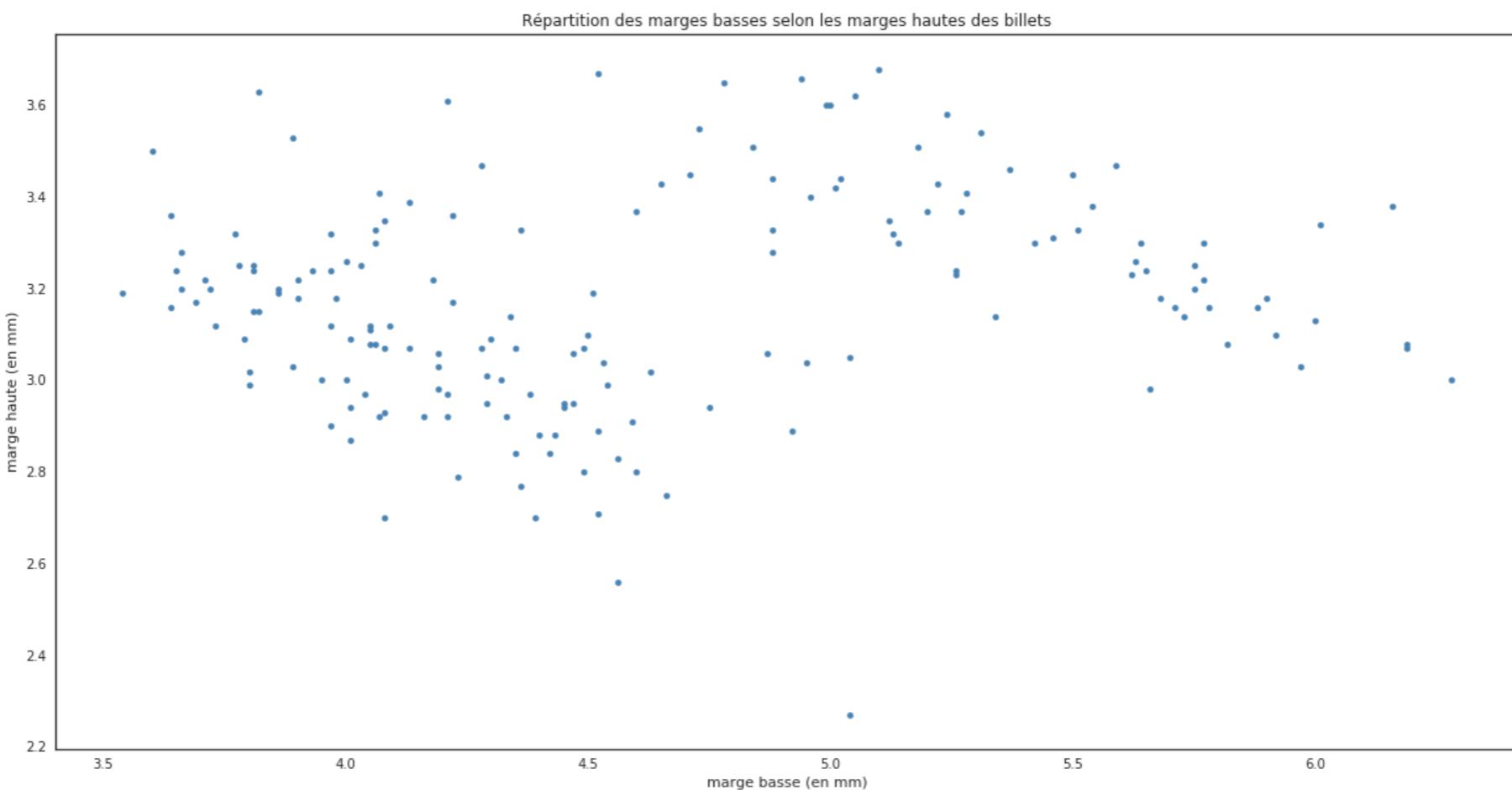
Sauvegarder l'image ? (y/n) :y

Indiquer le chemin du dossier.../presentation/images

Le coefficient de corrélation (Pearson) est égal à 0.08

Les variables sont pas corrélées

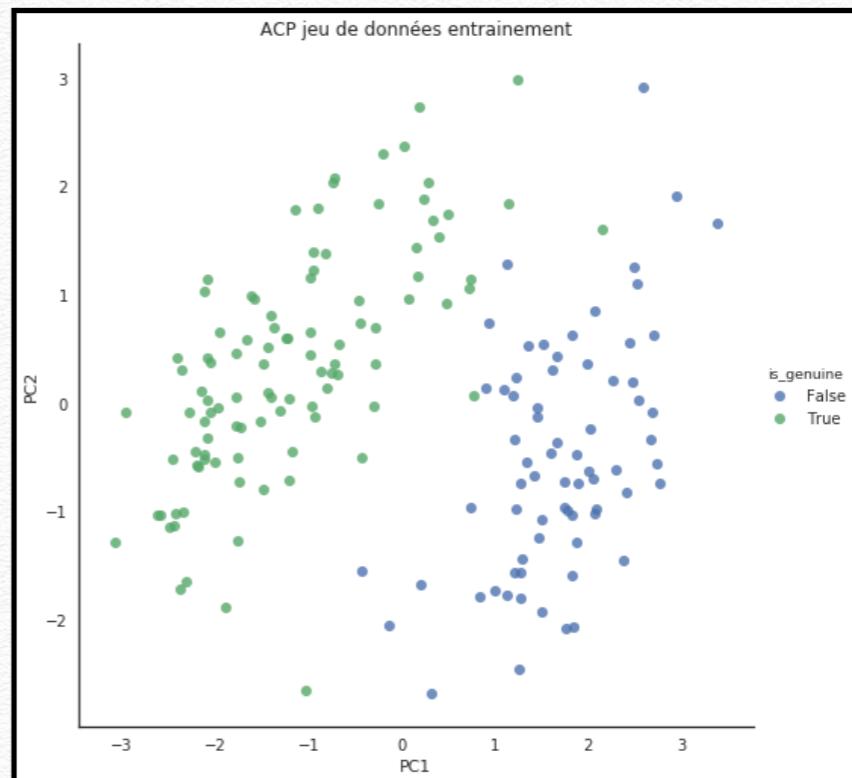
Analyses bivariées



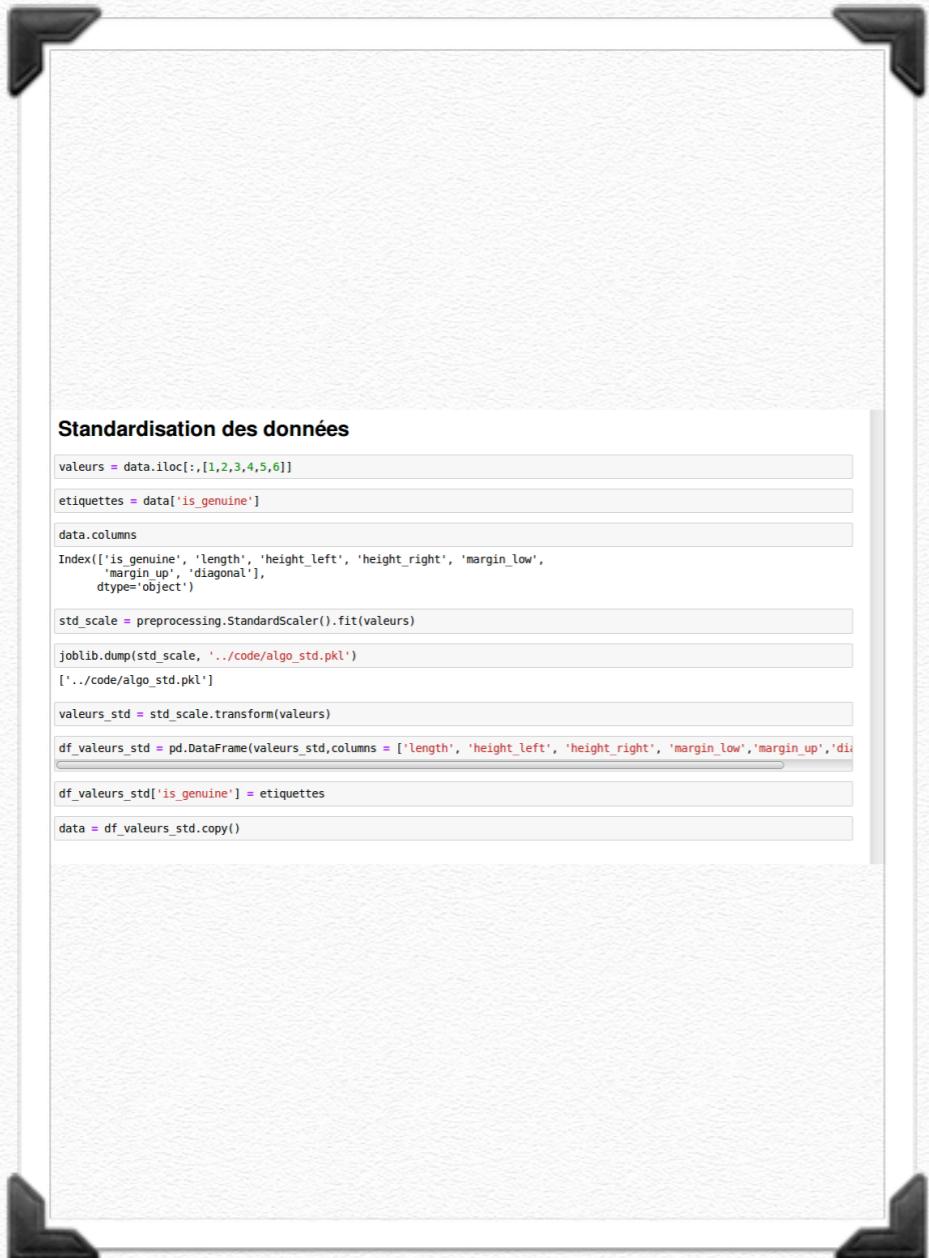
Sauvegarder l'image ? (y/n) : y
Indiquer le chemin du dossier.../présentation/images
Le coefficient de corrélation (Pearson) est égal à 0.17
Les variables sont pas corrélées

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

- ❖ Objectif : Résumé l'information contenu dans les 7 caractères contenus dans notre jeu de données à l'aide de deux axes principaux.
- ❖ Que représente « PC1 » et « PC2 » ?



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'])

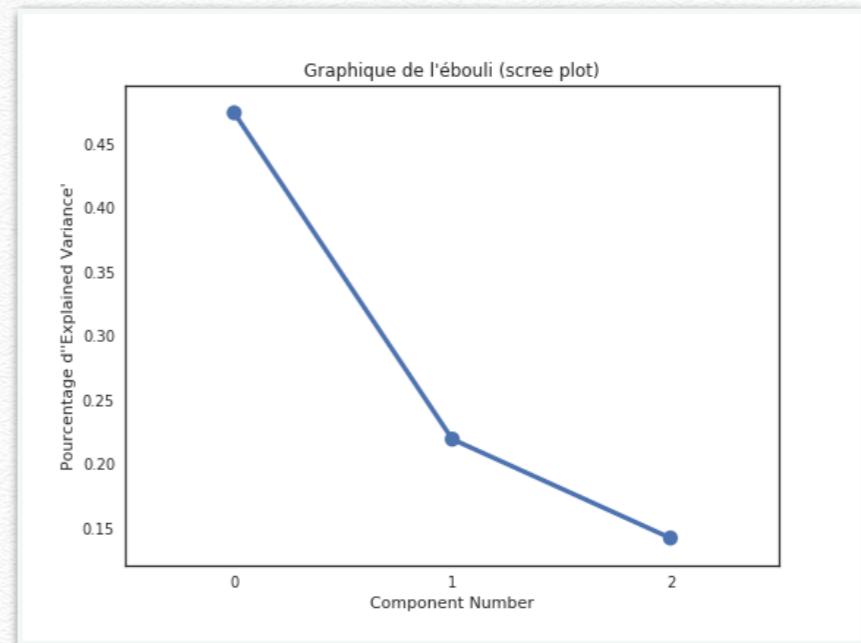
df_valeurs_std['is_genuine'] = etiquettes

data = df_valeurs_std.copy()
```

- ❖ Standardisation des données (centrer-réduire) : Evite que les algorithmes de Machine Learning mettent plus de temps à trouver un modèle prédictif quand les données sont dans des ordres de grandeurs différents.
- ❖ Pour cela : On transforme nos variables de sorte qu'elles répondent à un loi normale centrée réduite :
 - Moyenne = 0
 - Ecart-type = 1

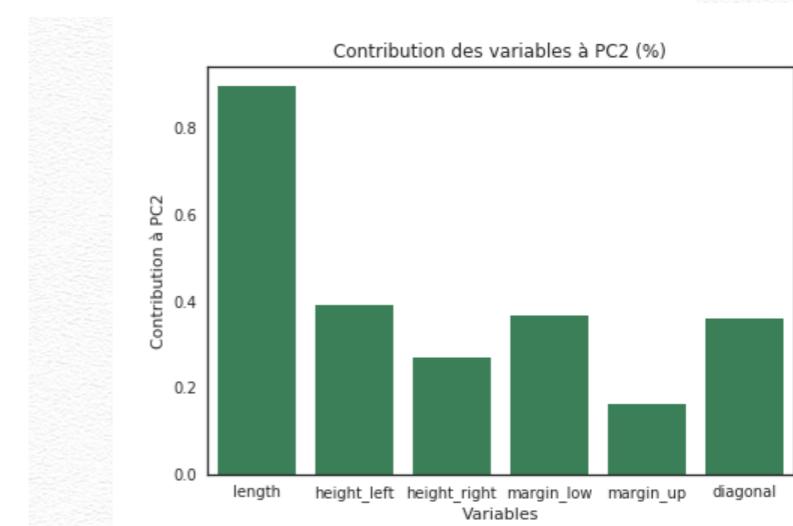
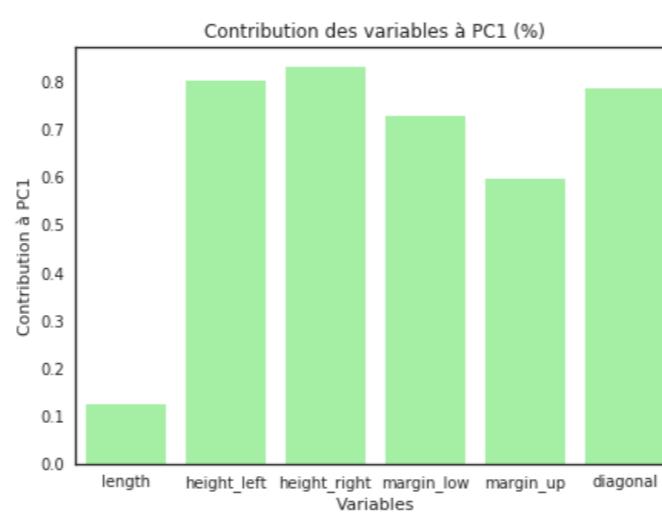
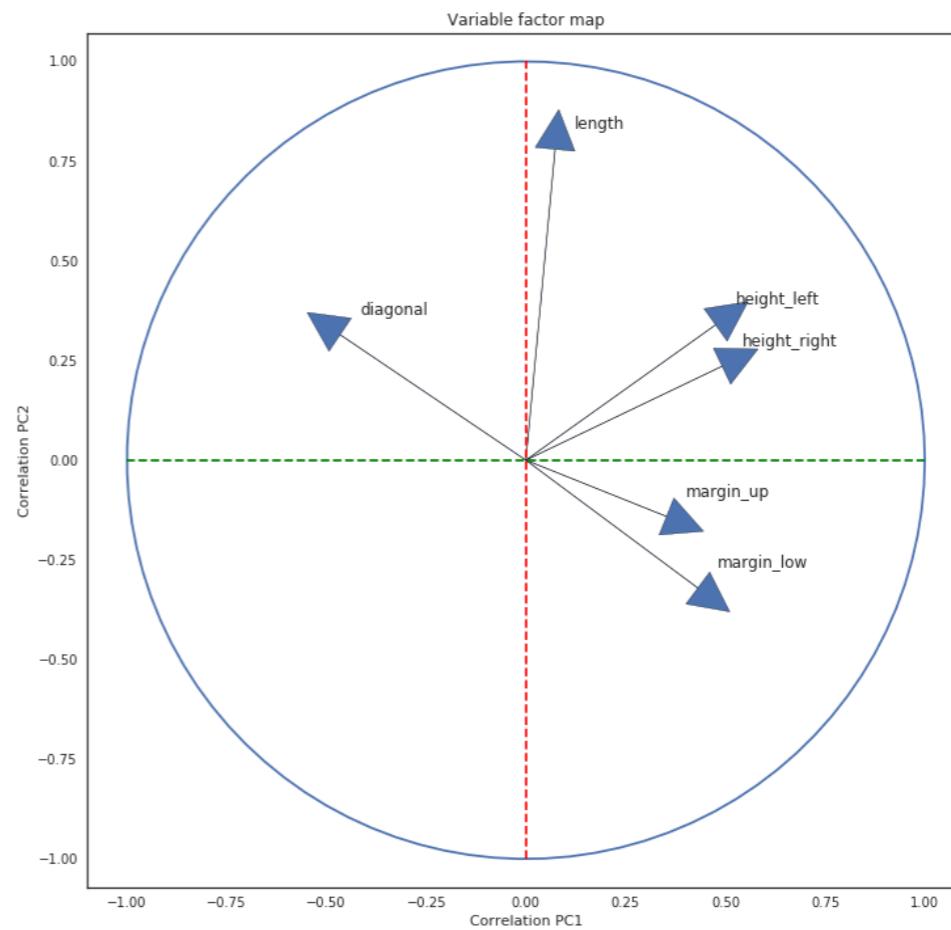
ACP : Choix du nombre de composantes principales (axes)

- ❖ Objectif : Sélectionner les axes qui maximise 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.



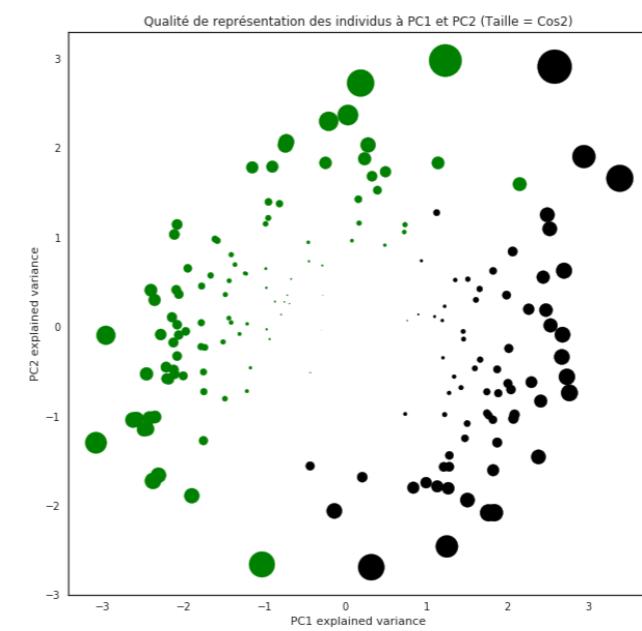
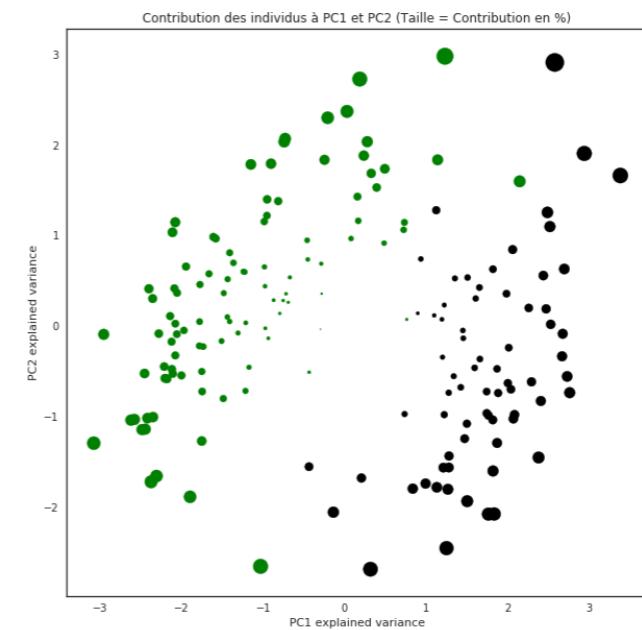
ACP : Cercle des corrélations

- ❖ Objectif : Montre 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 négative 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.

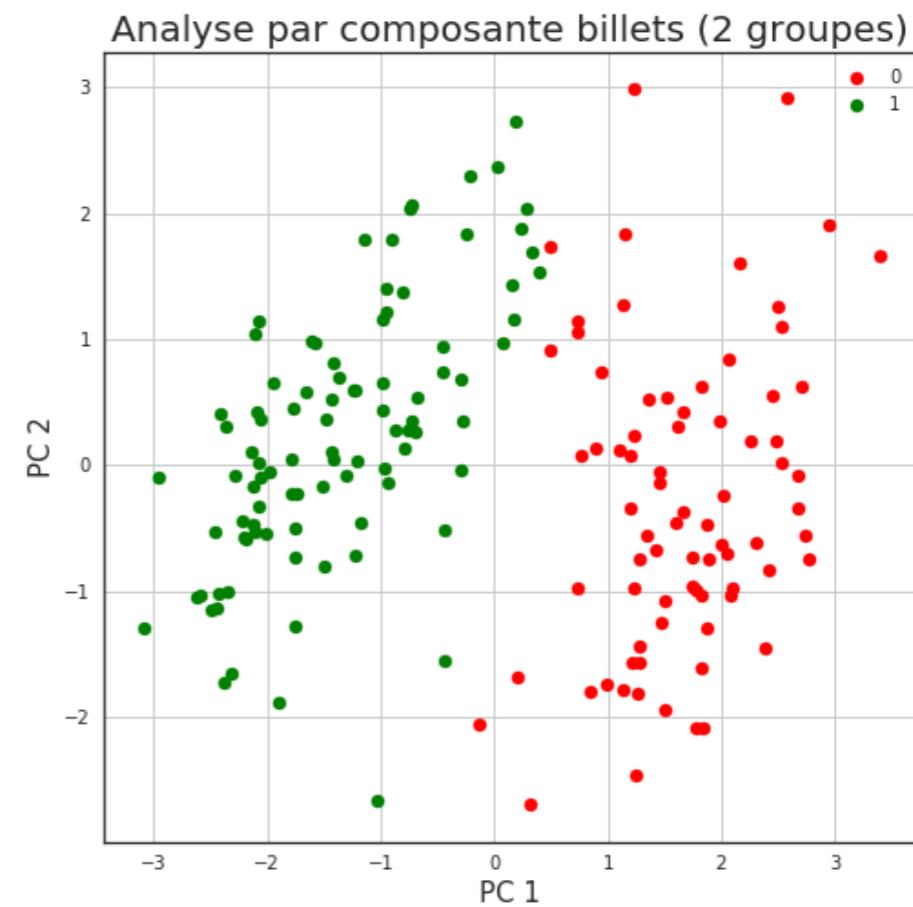
- ❖ En vert, les vrais billets et en noir les faux billets.
- ❖ Plus les points sont gros plus les individus contribuent aux deux composantes principales et sont mieux représentés.



Classification des billets

Classement des billets selon un algorithme non supervisé : Kmeans

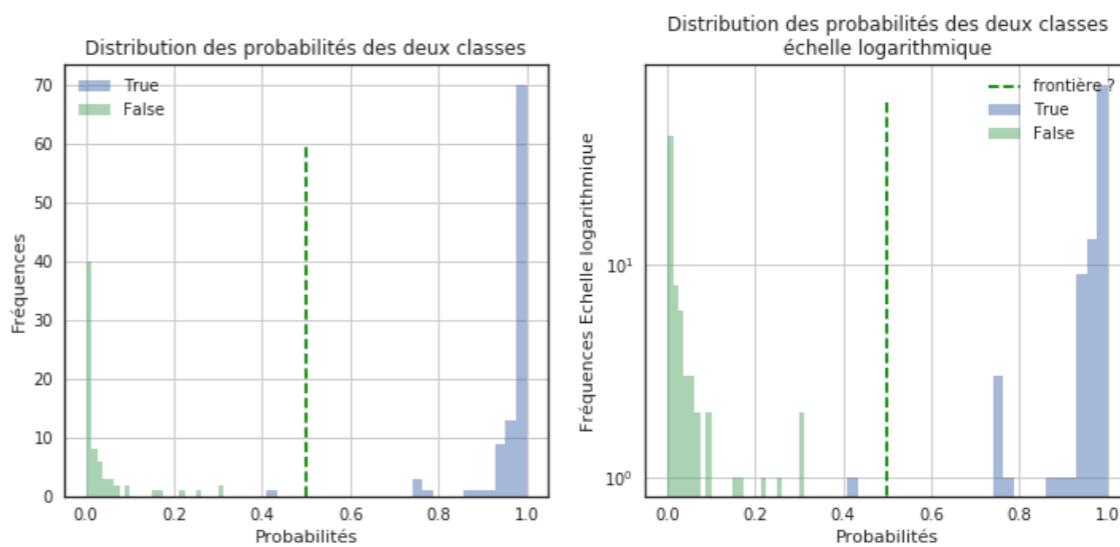
- ❖ Placement de points au hasard dans un plan, puis regroupement des individus autour de ces points.
- ❖ Calcule des centroïdes des groupes formé et regroupement des individus autour de ces centroïdes.
- ❖ Répétition de l'opération jusqu'à que les centroïdes ne bougent plus.



etiquette_cah	False	True	Total
is_genuine			
False	1.0	69.0	70.0
True	92.0	8.0	100.0
total	93.0	77.0	170.0

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

- ❖ La régression logistique est une régression linéaire qui a pour borne 0 et 1. Elle permet de relier la survenance ou la non survenance d'un événement au niveau de variables explicatives.
- ❖ Objectif : prédire si un billet est faux ou non.



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

Le modèle est confiant !

Test de l'algorithme de
classification ...

Conclusion

- ❖ L'analyse univarié des variables révèle la présences d'outliers.
- ❖ La standardisation des données limite l'impact des outliers
- ❖ L'ACP résume l'information correctement l'information via un premier axe représentant la longueur des billets et un second qui compile le reste des variables.
- ❖ La comparaison de résultat d'un algorithme non supervisé avec les données d'entraînement montre certaines confusions.
- ❖ Le modèle de régression logistique se montre plus confiant

Remise en cause du seuil de décision (0,5)

Seuil de décision optimale = 0.999659