Détecter les faux billets

Septembre 2021

PLAN

Introduction

1. Présentation du jeu de données
2. Analyse univariée
3. Analyse bivariée
4. Analyse en composantes principales ACP
5. Classification non supervisée: K-means
6. Classification supervisée: la régression logistique
7. Test du programme avec le jeu de données

Conclusion

Introduction

Dans le cadre de la lutte contre la criminalité organisée, à l'Office central pour la répression du faux monnayage au Ministère de l’intérieur, ma société de consulting informatique me propose de créer un algorithme de détection de faux billets.

1. Présentation du jeu de données

dataframe data contient :

- 170 lignes
- 7 colonnes

	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

Détections des outliers

Grâce à l'écart interquartile

on a détecté 5 outliers

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
34	True	172.75	104.33	103.97	4.34	3.14	113.12
70	True	171.04	103.84	103.64	4.22	3.36	112.70
166	False	173.01	104.59	104.31	5.04	3.05	110.91
0	True	171.81	104.86	104.95	4.52	2.89	112.83
4	True	172.05	103.70	103.75	5.04	2.27	113.55

Nouveau dataframe

Le nouveau dataframe a :

- 165 lignes
- 7 colonnes

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
0	True	171.67	103.74	103.70	4.01	2.87	113.29
1	True	171.83	103.76	103.76	4.40	2.88	113.84
2	True	171.80	103.78	103.65	3.73	3.12	113.63
3	True	172.57	104.65	104.44	4.54	2.99	113.16
4	True	172.38	103.55	103.80	3.97	2.90	113.30
...	...	...	...	...	...	...	...
160	False	171.43	104.26	103.97	5.73	3.14	111.82
161	False	172.11	104.23	104.45	5.24	3.58	111.78
162	False	172.47	104.27	104.10	4.88	3.33	110.68
163	False	171.82	103.97	103.88	4.73	3.55	111.87
164	False	171.96	104.00	103.95	5.63	3.26	110.96

165 rows × 7 columns

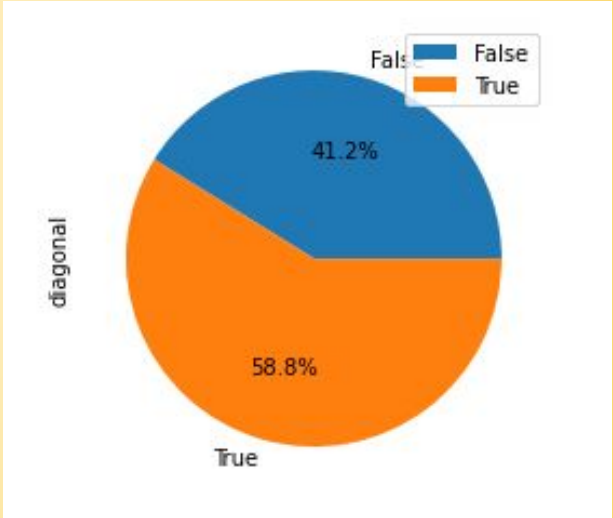
2. Analyse univariée

- Ona 165 individus
- la diagonale moyenne est 171.93
- hauteur moyenne à gauche est 104.06
- hauteur moyenne à droite est 103.92
- marge moyenne du bas est 3.17
- marge moyenne du haut est 4.61
- longueur moyenne est 112.56

	diagonal	height_left	height_right	margin_low	margin_up	length
count	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000
mean	171.934788	104.060364	103.922182	4.611515	3.177333	112.568848
std	0.283232	0.290726	0.323860	0.710143	0.227564	0.925008
min	171.130000	103.230000	103.140000	3.540000	2.560000	109.970000
25%	171.730000	103.850000	103.690000	4.050000	3.020000	111.850000
50%	171.940000	104.050000	103.950000	4.450000	3.180000	112.850000
75%	172.130000	104.280000	104.170000	5.140000	3.330000	113.290000
max	172.590000	104.720000	104.860000	6.280000	3.680000	113.980000

Proportion des individus

is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
False	69	69	69	69	69	69
True	96	96	96	96	96	96

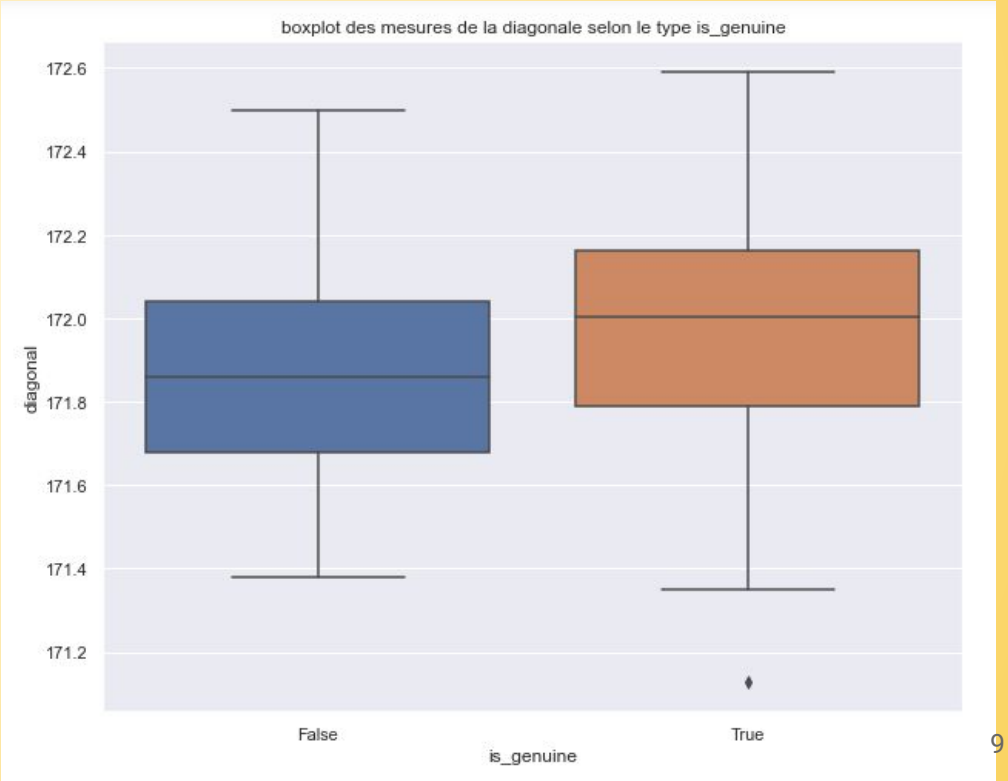


3. Analyse bivariée

Boxplot des diagonales des billets

- False
  min: 171.4
  médiane: 171.84
  max: 172.5

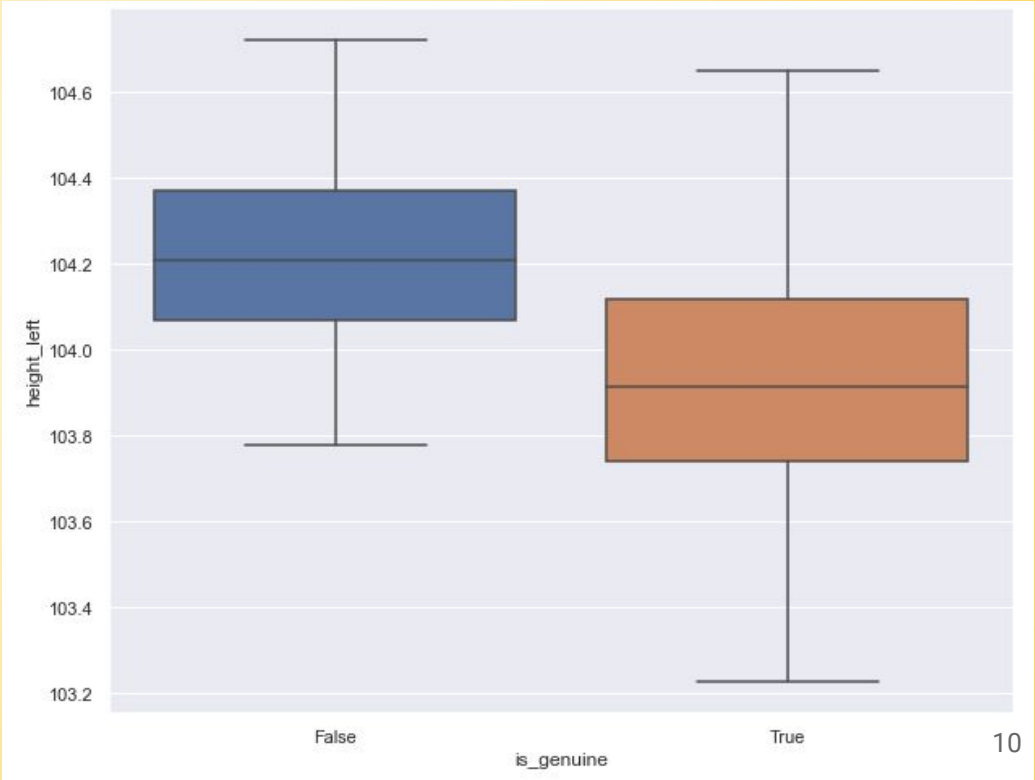
- True
  min: 171.38
  médiane: 172
  max: 172.6



Boxplot de height_left

- False
  min: 103.8
  mediane: 104.2
  max: 104.67

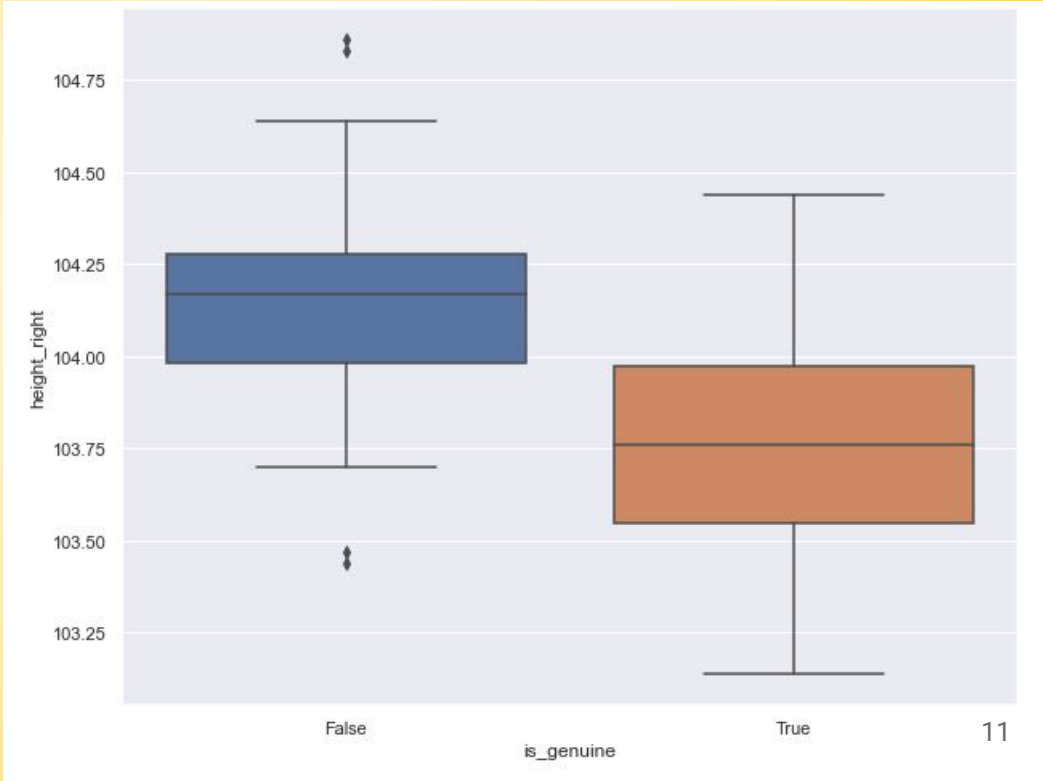
- True
  min: 103.22
  mediane:103.85
  max: 104.62



Boxplot de height_right

- False
  min: 103.73
  mediane: 104.20
  max: 104.70

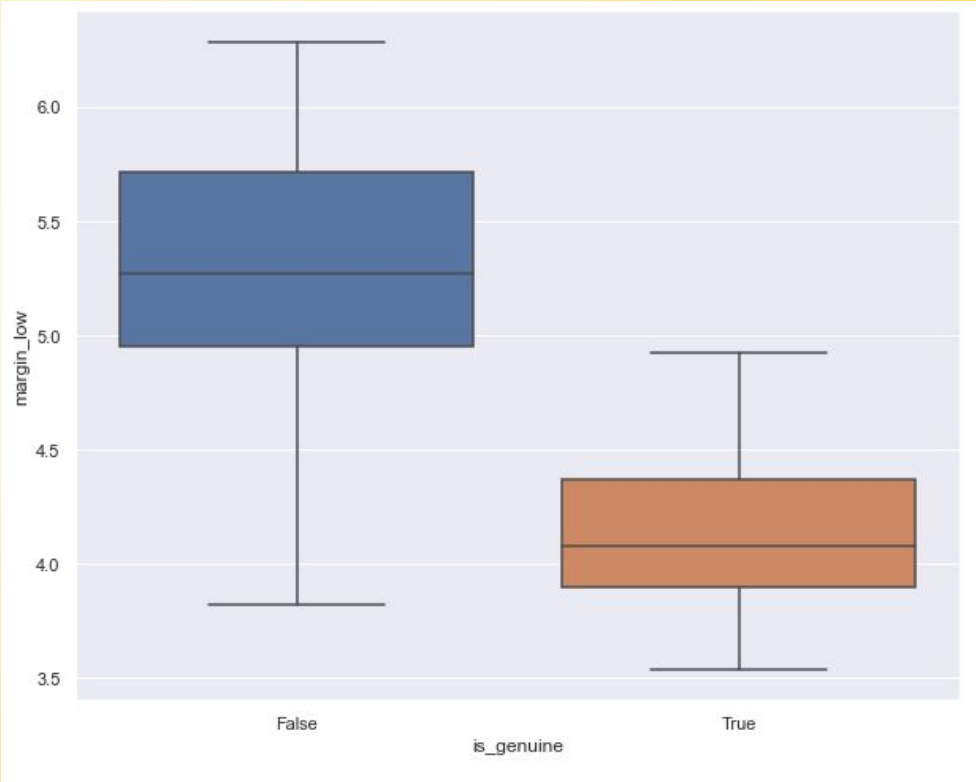
- True
  min: 103.15
  mediane: 103.75
  max: 104.48



Boxplot de margin_low

- False
  min: 3.7
  mediane: 5.25
  max: 6.4

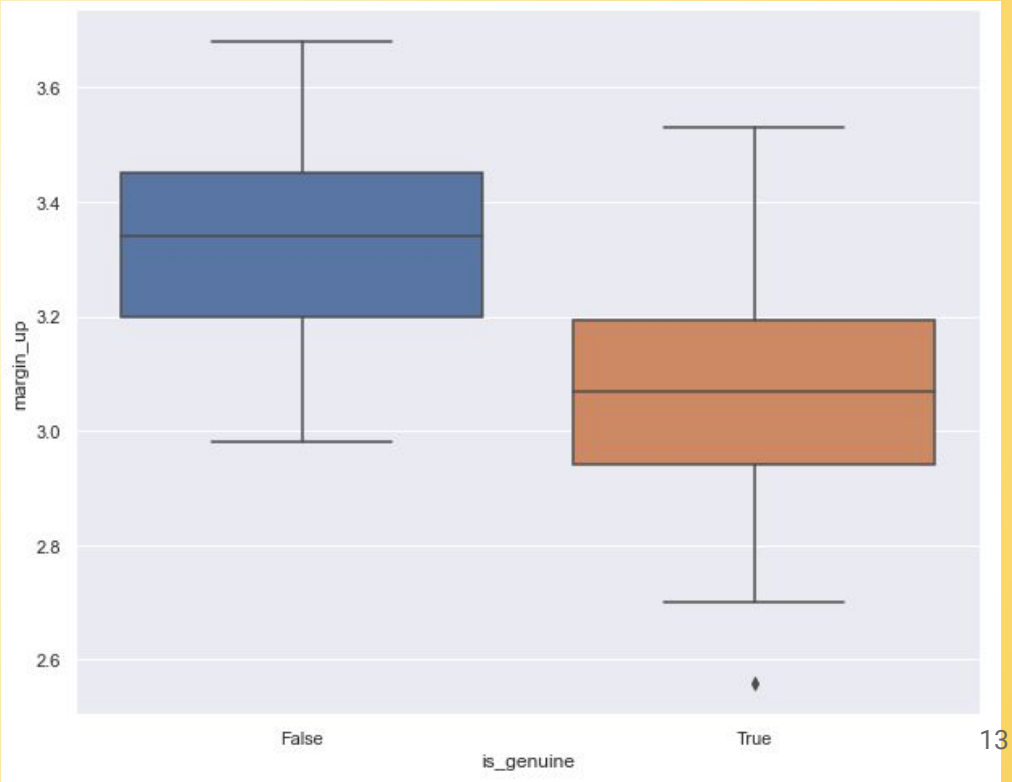
- True
  min: 3.5
  mediane: 4.1
  max: 4.9



Boxplot de margin_up

- False
  min: 3
  mediane: 3.35
  max: 3.7

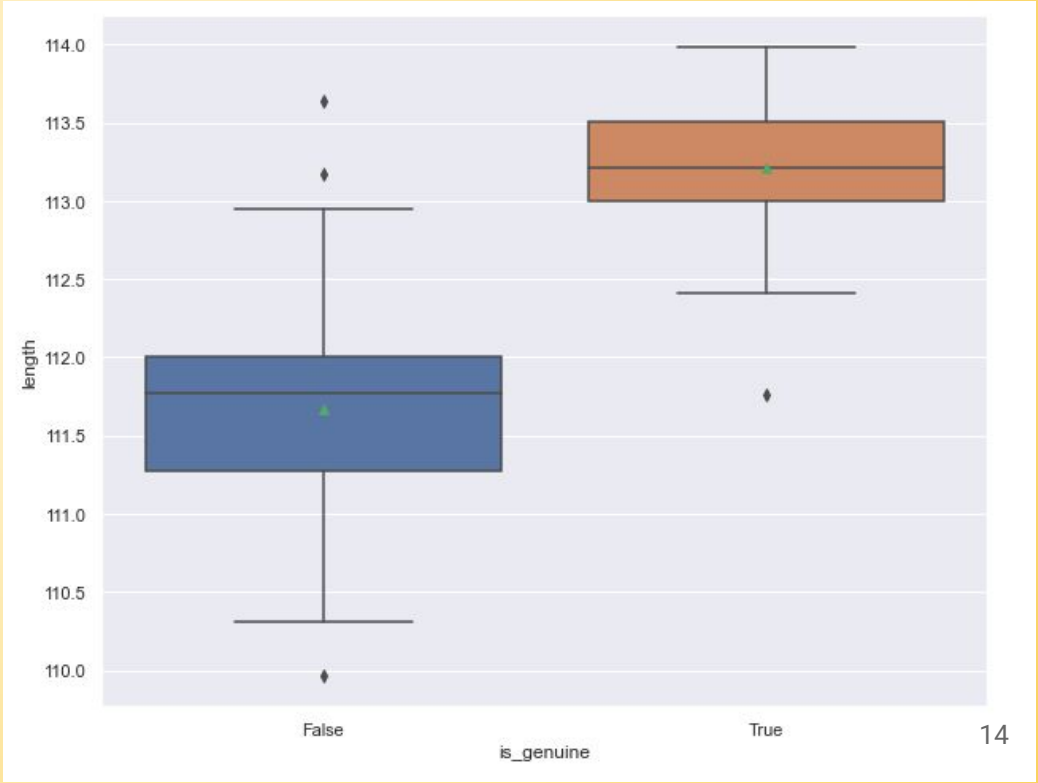
- True
  min: 2.7
  mediane: 3.1
  max: 3.5



Boxplot de length

- False
  min: 110.4
  mediane: 111.8
  max: 113

- True
  min: 112.45
  mediane: 113.25
  max: 114



Matrice de corrélation

On note une bonne corrélation entre **Height_left** et **hight_right**.

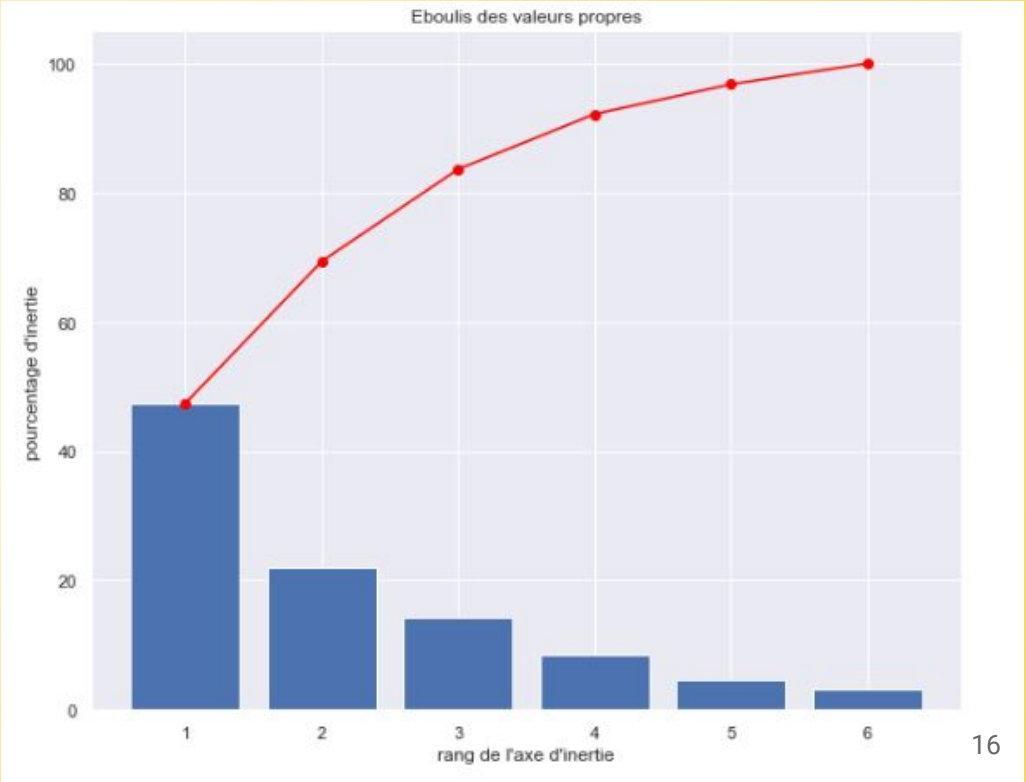
	diagonal	height_left	height_right	margin_low	margin_up	length
diagonal	1.000000	0.319584	0.220418	-0.181020	-0.027366	0.080295
height_left	0.319584	1.000000	0.734390	0.424530	0.324788	-0.421387
height_right	0.220418	0.734390	1.000000	0.509375	0.366918	-0.417021
margin_low	-0.181020	0.424530	0.509375	1.000000	0.171113	-0.637352
margin_up	-0.027366	0.324788	0.366918	0.171113	1.000000	-0.525284
length	0.080295	-0.421387	-0.417021	-0.637352	-0.525284	1.000000

4. ACP

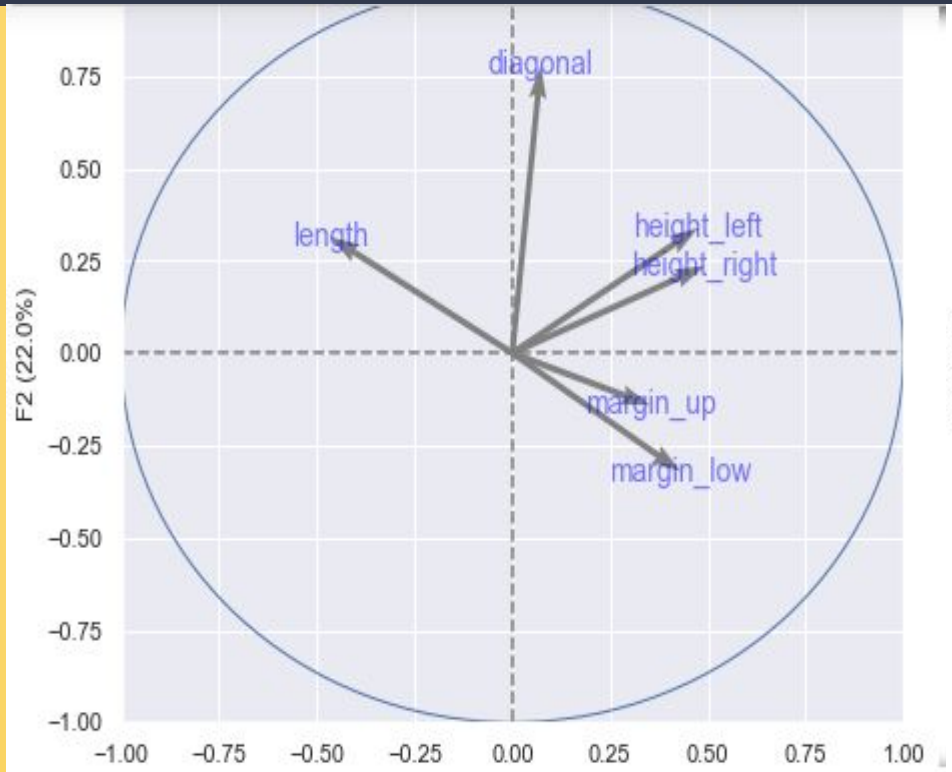
Eboulis des valeurs propres:

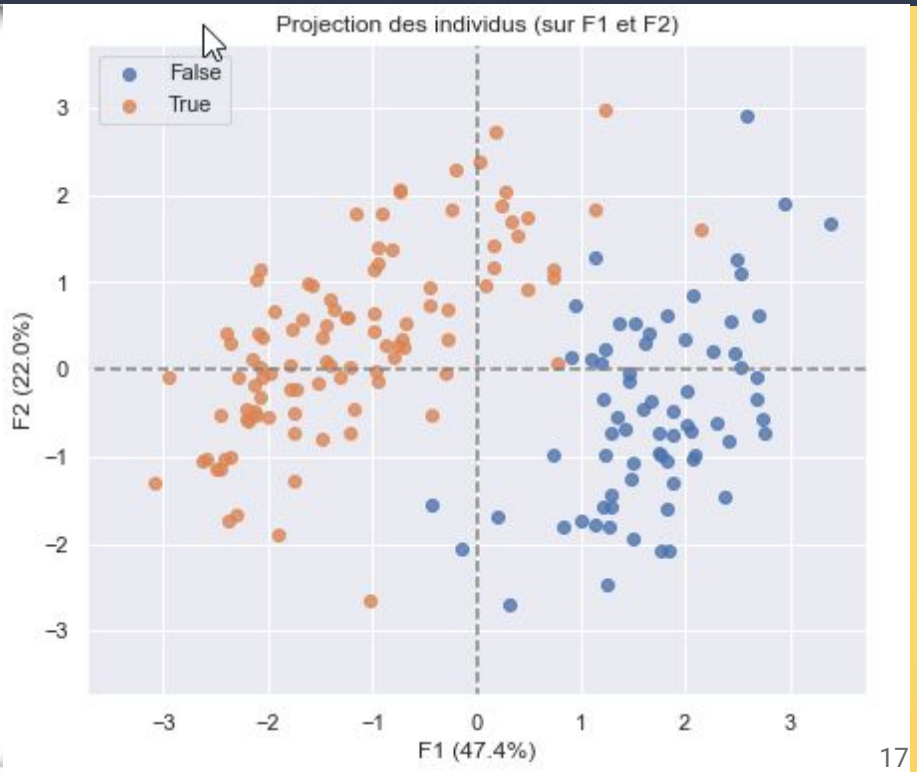
L'axe 1 et 2 représentent 70% de l'inertie totale.

On va baser notre analyse sur le 1er plan factoriel



Cercle de corrélations et projection des individus





Qualité de représentation des individus

	is_genuine	COS2_1	COS2_2	premier_plan
35	True	2.5	0.0	3.5
160	False	6.9	4.3	11.2
43	True	3.9	6.3	11.2
88	True	2.3	13.6	16.9
52	True	7.7	10.6	19.3

Les 5 billets avec la pire qualité de représentation

	is_genuine	COS2_1	COS2_2	premier_plan
90	True	89.8	6.1	96.9
45	True	63.1	33.1	97.2
46	True	96.2	0.3	97.5
143	False	97.3	0.8	98.1
148	False	57.8	41.1	98.9

Les 5 billets avec la meilleure qualité de représentation

Contribution des individus aux axes

	is_genuine	CTR_1	CTR_2	premier_plan
49	True	1.96	0.75	3.71
34	True	0.01	3.33	4.34
70	True	0.22	3.15	4.37
166	False	1.38	3.79	5.17
5	True	0.32	3.97	5.29

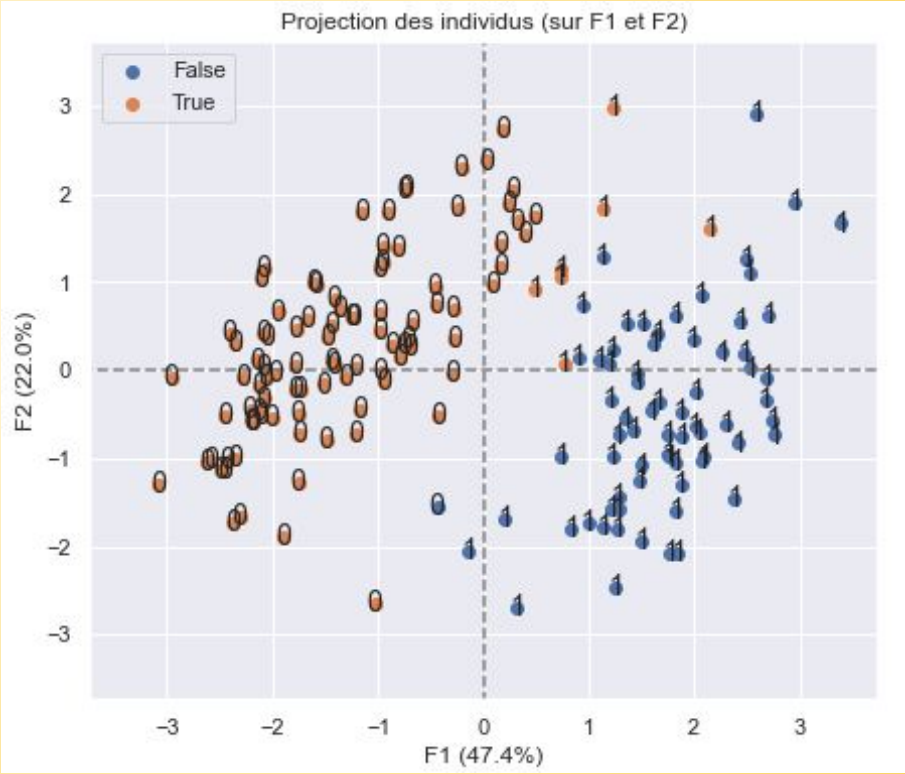
Les 5 billets avec la meilleure contribution dans le 1er plan factoriel

	is_genuine	CTR_1	CTR_2	premier_plan
102	False	0.17	0.01	0.18
118	False	0.25	0.01	0.26
114	False	0.30	0.00	0.30
128	False	0.31	0.02	0.33
150	False	0.30	0.05	0.35

Les 5 billets avec la pire contribution dans le 1er plan factoriel

5. Classification non supervisée K-Means

- les 2 ensembles sont nettement séparés
- on constate 7 faux positifs
- et 1 faux négatif



ANOVA

H0: diagonal est corrélée à la variable is_genuine

H1: il n'y a pas de corrélation entre diagonal et is_genuine.

La p_valeur de l'ANOVA entre diagonal et is_genuine est 0.07. Donc p_valeur > 0.05.

On rejette l'hypothèse H0

Variables explicatives	**variable à expliquer**	**p_valeur**
heigth_right	is_genuine	6.66e-15
height_left	is_genuine	2.33e-10
margin_low	is_genuine	3.94e-39
margin_up	is_genuine	7.56e-17
length	is_genuine	1.23e-43
diagonal	is_genuine	**0.07**

6. Régression logistique

L'objectif de la régression logistique est de **modéliser,** de **classifier,** une variable binaire prenant ses valeurs dans {0,1} en fonction de variables explicatives quantitatives (et potentiellement qualitatives).
Ce modèle permet de prédire la probabilité qu'un événement se produise (p=1) ou pas (p=0)

Fonction logit

La régression logistique utilise une fonction logistique comme fonction de lien.

La fonction qui remplit le mieux ces conditions est la fonction sigmoïde, définie sur R à valeurs dans [0,1]. Elle s'écrit de la manière suivante :

$$Y = \frac{1}{1 + e^{-X}}$$

Matrice de confusion

Vrai négatif: 17
Faux négatif: 1
Vrai positif: 23
Faux positif: 2

	0	1
0	17	2
1	1	23

Les metrics

Taux de succès : probabilité de bon classement du modèle
Sensibilité : capacité du modèle à retrouver les positifs
Précision : proportion de vrais positifs parmi les individus classés positifs

	Données Test	Données Train
Taux de succès	0.93	1
Précision	0.92	1
Rappel ou Sensibilité	0.95	1

Quelques étapes de la régression logistique

- Partitionnement des données (apprentissage et test)

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

- Création de la fonction Régression logistique

```
reg_log = LogisticRegression(solver='liblinear')
reg_log.fit(x_train, y_train)
```

- Sauvegarde de prédictions des données test

```
y_pred = reg_log.predict(x_test)
```

- Probabilités des événements de x_test via la méthode predict_proba()

```
y_pred_proba = reg_log.predict_proba(x_test)
```

7. Test du programme avec le jeu de test donné par l'évaluateur

Conclusion

L’étude de ce projet nous a permis de revoir l’analyse univariée et bivariée, l’ACP et la classification non supervisée avec le K-means.
En plus de cela, nous avons découvert un nouvel outil qui nous permet de prédire des modèles avec la régression logistique.

Sources

Cours OpenClassrooms

http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf

https://whatis.techtarget.com/fr/definition/Regression-logistique

https://fr.wikipedia.org/wiki/Fonction_logistique_(Verhulst)

https://pingouin-stats.org/generated/pingouin.anova.html?highlight=anova#pingouin.anova

https://www.youtube.com/watch?v=hNRDLVpzYgI&ab_channel=INTECHWETRUST

etc

Merci de votre attention

Présenté par Yaya CISSE