DETECTION DE FAUX BILLETS

SOUTENANCE DU PROJET 6

PARCOURS DATA ANALYSTE VERSION 1

TRAORE LASSINA JAMAL

Plan

1. Contexte
2. Echantillon-Statistique descriptive
3. ACP-Kmeans
4. Regression logistique
    a. approches
    b. resultats
5. Test du set
6. Concultion

Contexte

- *La France fait partie des **peleotons de tete***
- *Le trafic de faux billets cree une **économie parallèle***
- *Cela met à mal les **plans d'investissement***
- *Cela crée un manque à gagner au **fisc***
- *Et la non tracabilité de l'argent cash*
- ***Economie souterraine de plus d 60 milliards €***

Notre Mission

DATA ANALYSTE

L'echantillon

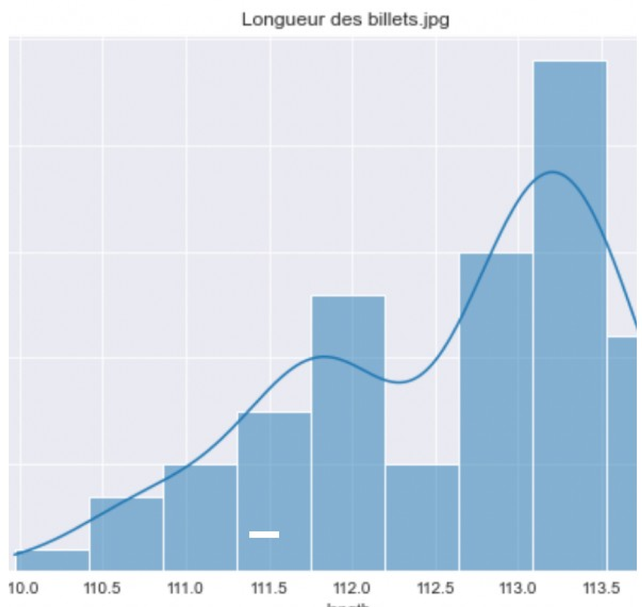
	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
0	True	171.81	104.86	104.95	4.52	2.89	112.83
1	True	171.67	103.74	103.70	4.01	2.87	113.29
2	True	171.83	103.76	103.76	4.40	2.88	113.84
3	True	171.80	103.78	103.65	3.73	3.12	113.63
4	True	172.05	103.70	103.75	5.04	2.27	113.55

- Echantillon constitué d'individu : des billets et leurs caracteristiques diverses : diagonale, hauteurs, marges et longueurs
- Ces attributs sont des variables continues
- La variable **booléenne indique l'authenticité du billet et constitue la variable illustrative**
- Le data set n'a aucune valeur aberrante ou d'anomalies

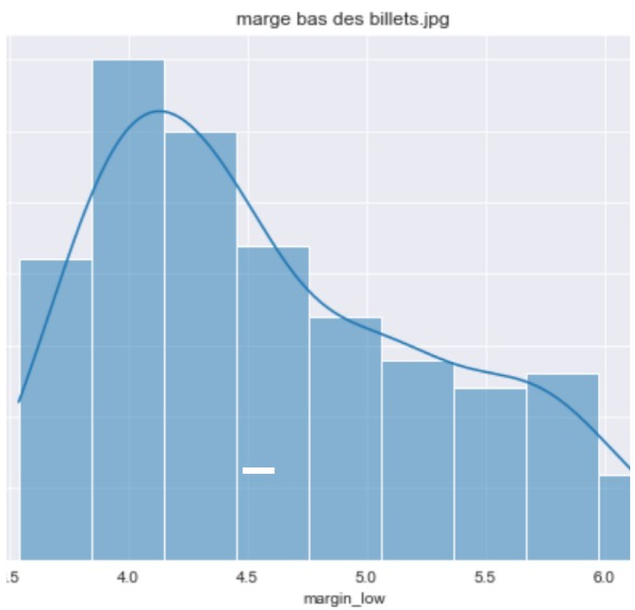
STATISTIQUE DESCRIPTIVE

Analyses univariées

- 2 classes indiquant l'authenticité du billet.
- La classe des **vrais** billets est plus grande que celle des **faux** .
- Mais il n'y a pas un grand desequilibre des deux classes

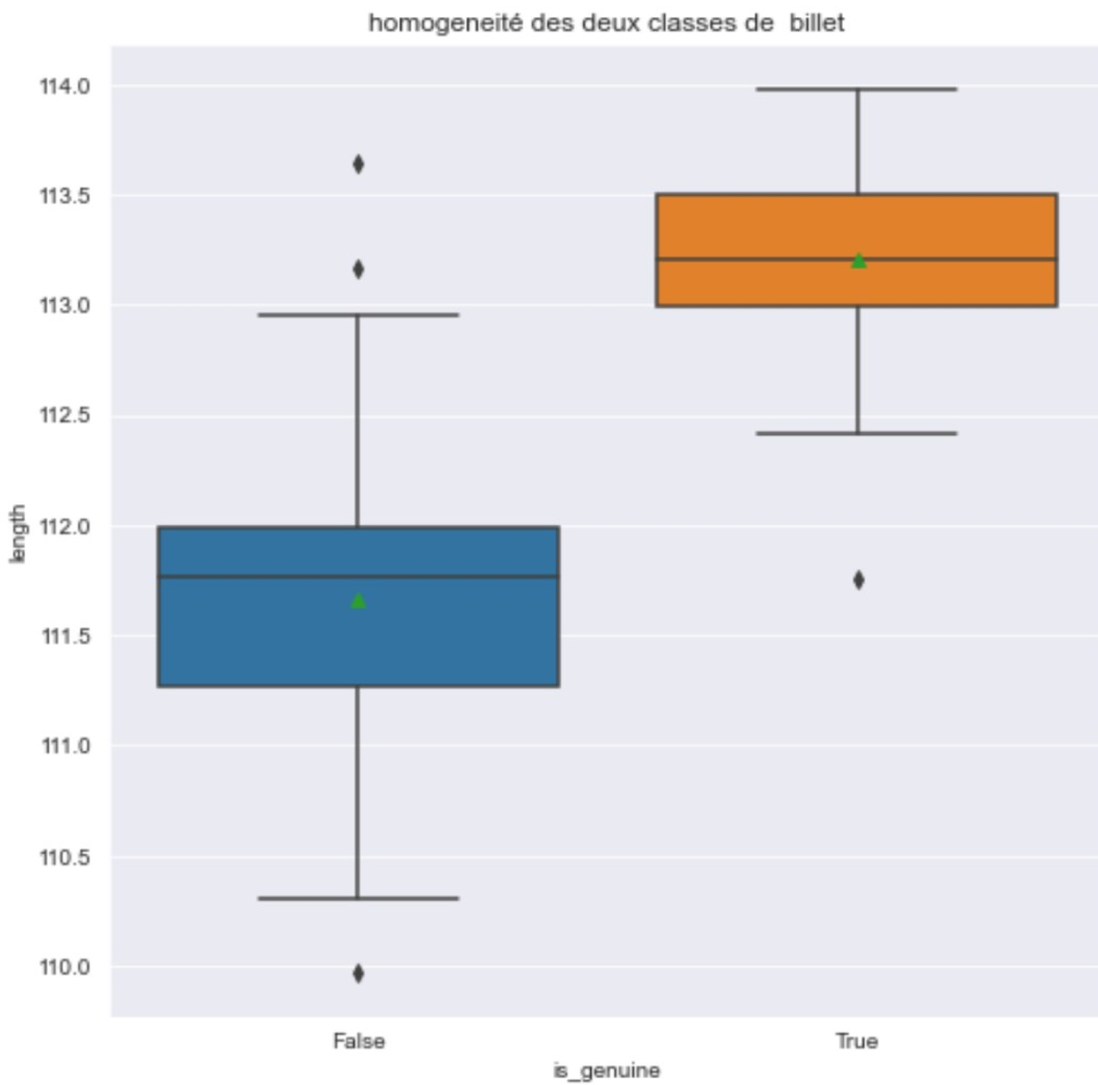


Les mesures de tendance centrale sur la variable length indiquent une moyenne et une mediane autour de la valeur 112 en dessous du mode de cette distribution
La distribution ne suit pas celle d'une loi gaussienne



Cette representation de la marge basse du billet indique un mode autour de 4 mm et au delà de cette mésure , on constate de moins en moins d'observation.

VARIANCE DES DEUX CLASSES

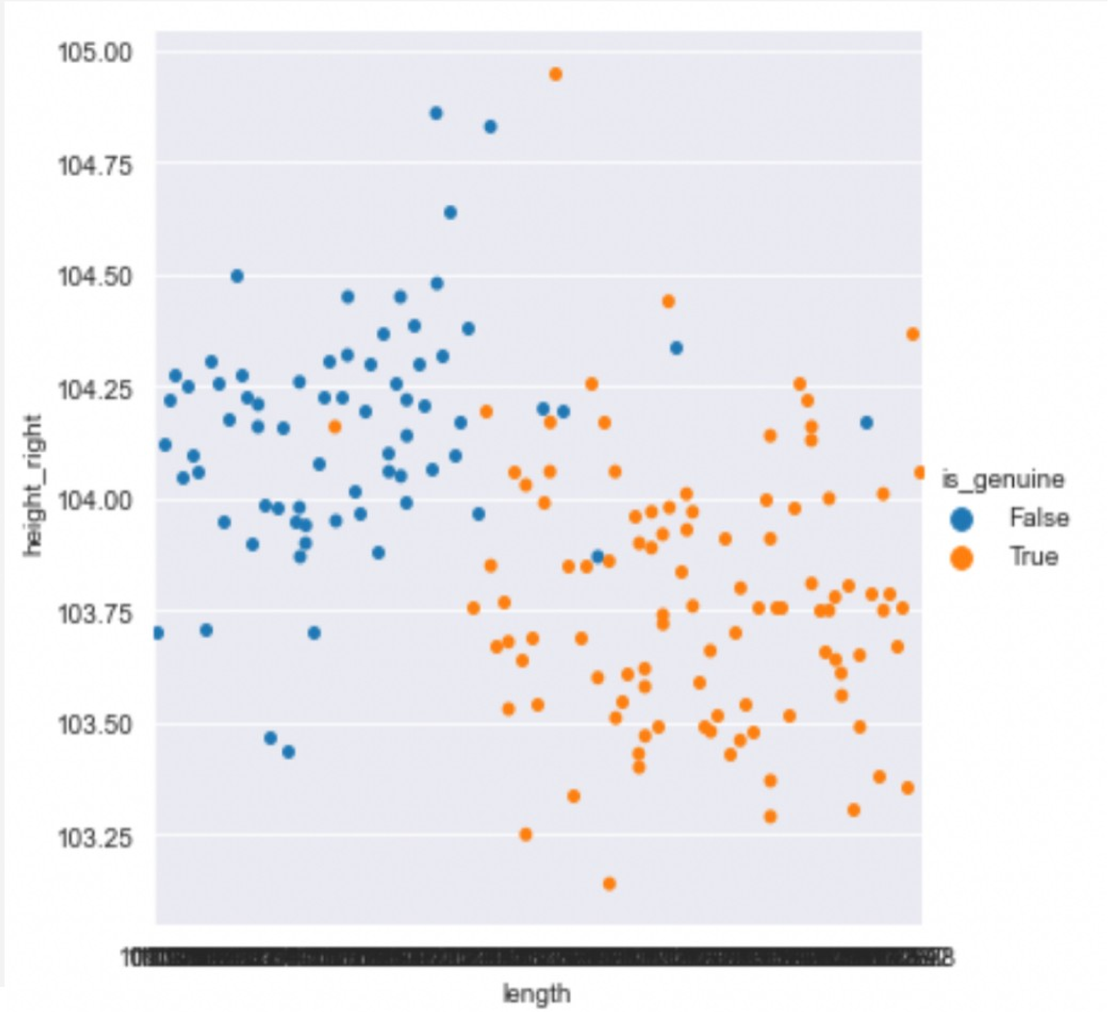


Homoscedasticité des deux classes

- Une plus grande variance quand le billet est faux .c'est a dire que la longueur du billet quand il est faux varie enormement. ce qui est comprehensible
- Globalement , on pourrait dire que les faux billets ont tendance à etre moins longs que les vrais.

CLASSES DISCTINCTES

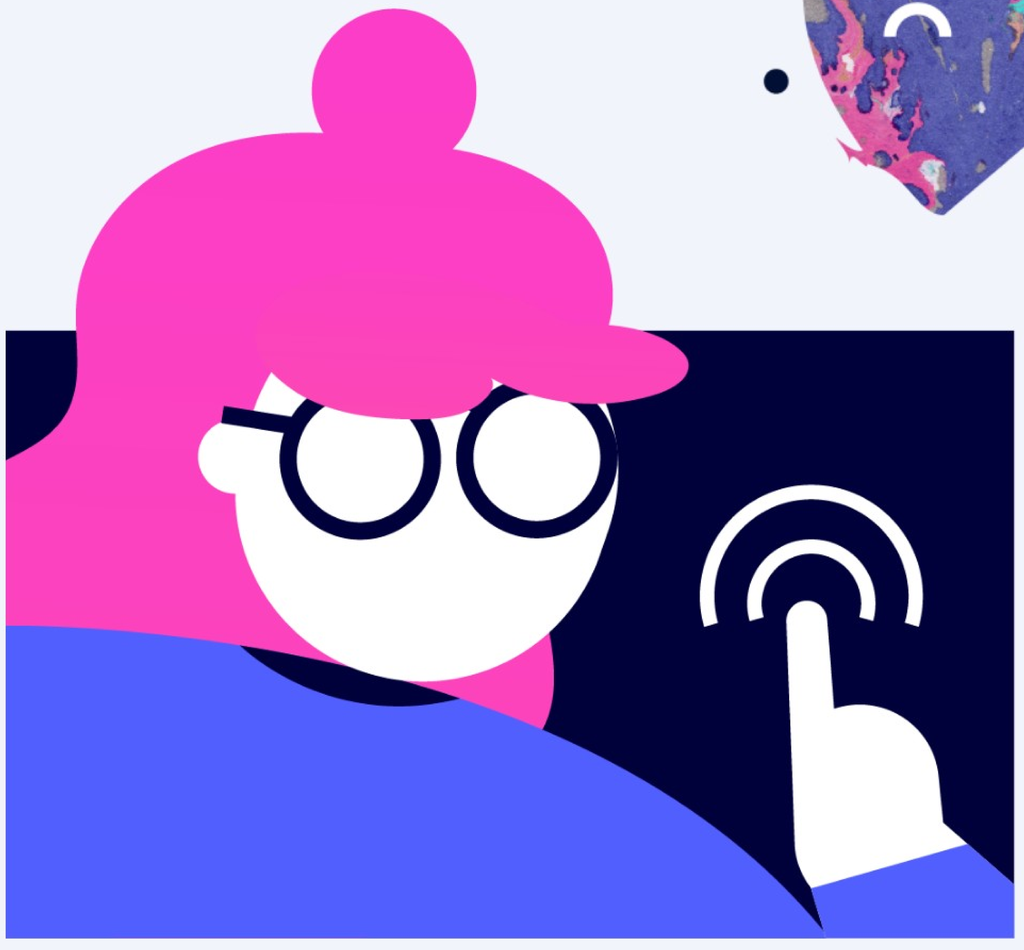
- longeur du billet en lien avec sa hauteur de droite selon qu'il soit vrai ou faux .
- les vrais billets sont generalement plus longs que les faux mais la hauteur de droite des faux billets depassent en moyenne ceux des vrais billets .
- On peut voir deux groupes distincts de faux et vrais billets avec les deux couleurs .



Bilan Mission 0

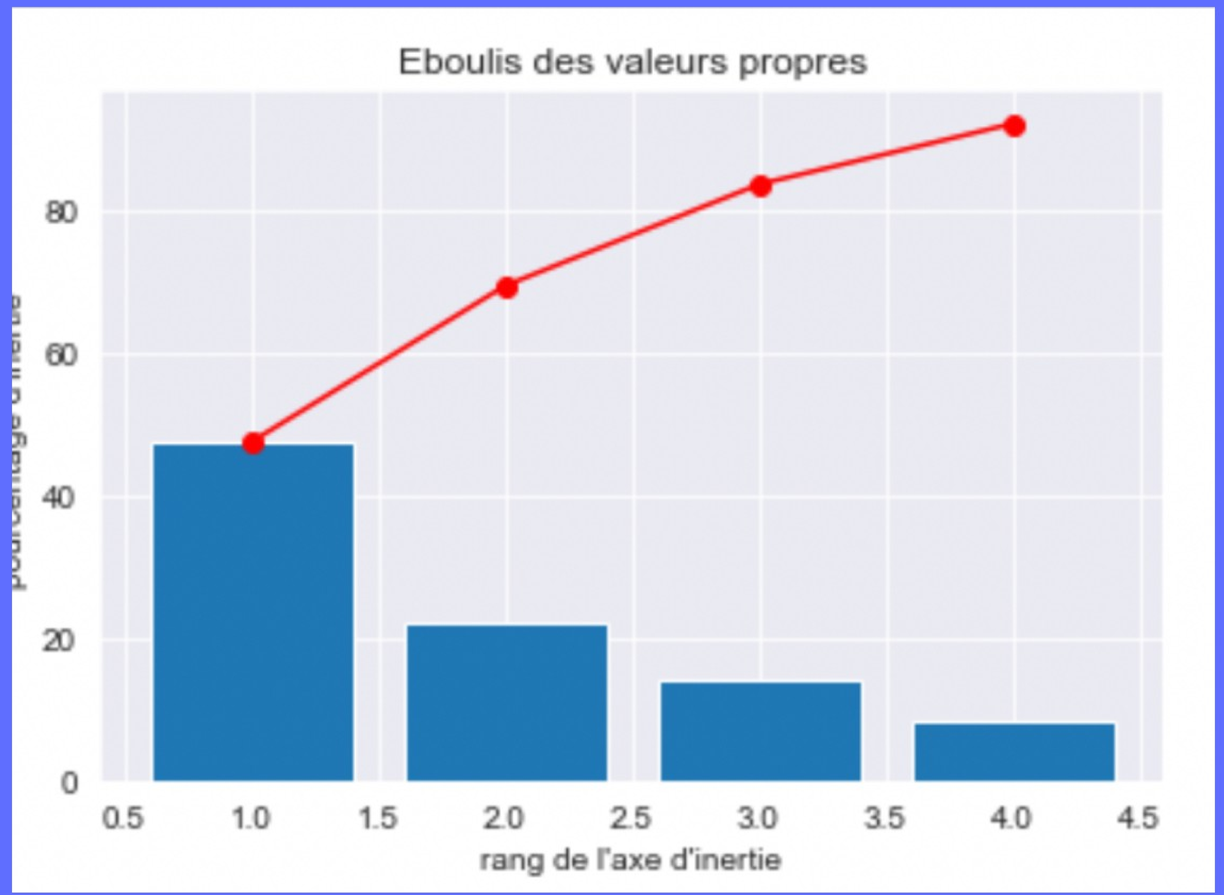
- La statistique descriptive de l'echantillon montre deux classes de billets indiquant l'authenticité de celui-ci

- Nous etudierons encore mieux notre echantillon avec l' ACP



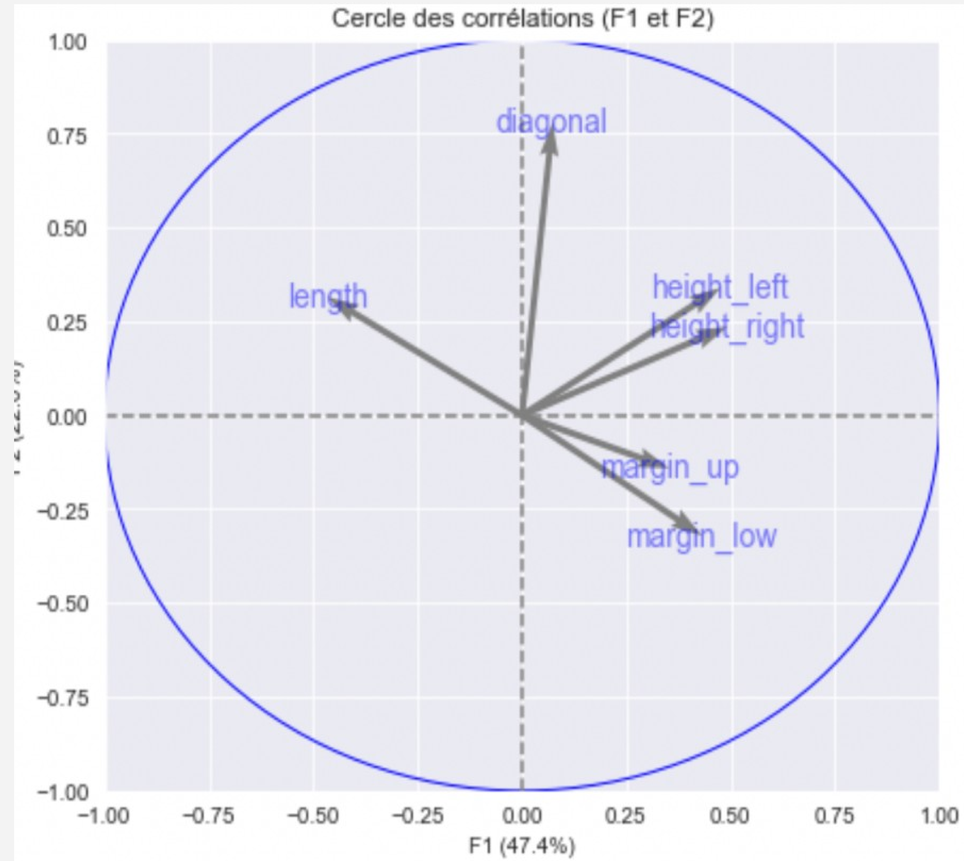
ACP/Eboullie des valeurs propres

*Le **premier plan factoriel** occupe 70% des informations . cela equivaut à 70% d'inertie du premier plan factoriel constitué de F1 et F2 .*
***70% des informations** sur notre echantillon sont fournies par les 2 premieèes dimensions ou axes . Les autres axes occupent les 30% d'informations restants.*



ACP/CERCLE DE CORRELATION

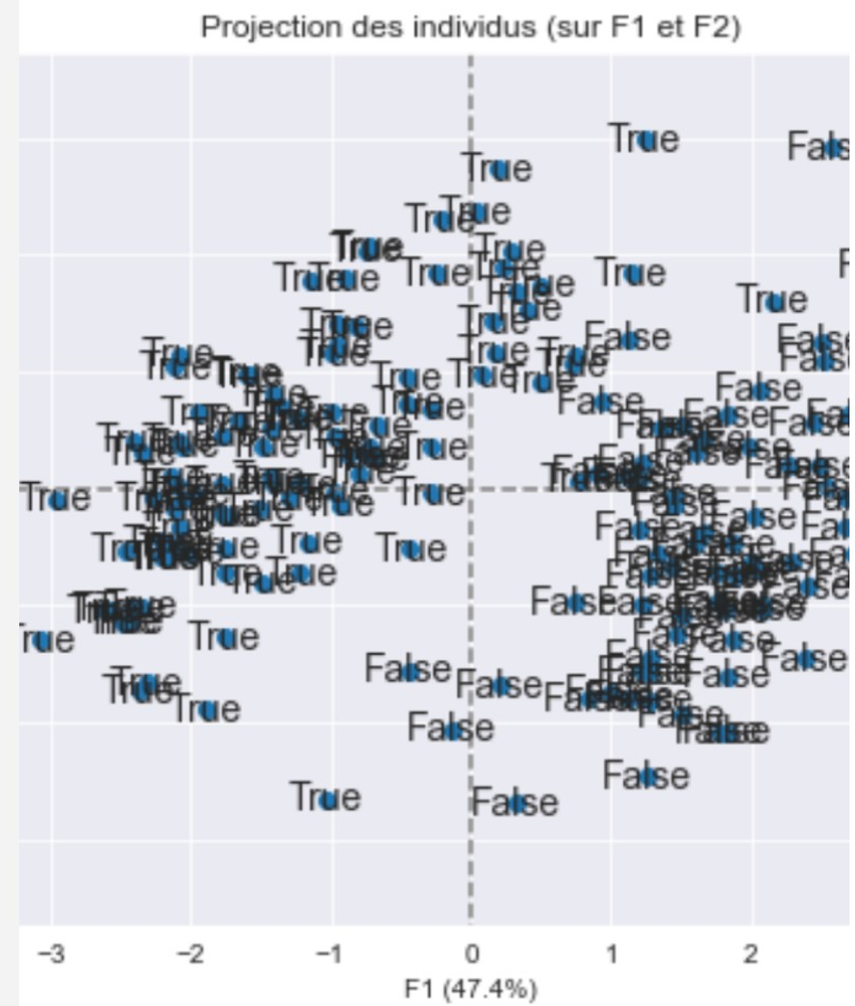
- *Plusieurs variables de contribution aux axes*
- *Les variables hauteurs et marges sont positivement correlées à F1 quand la var diagonal est fortement correlee à F2.*
- *La var length est negativement correlée à F1. Notons que les variables hauteurs et marges sont mal representées en regard à la longueur de leur rayons.*



ACP/PROJECTION DES INDIVIDUS

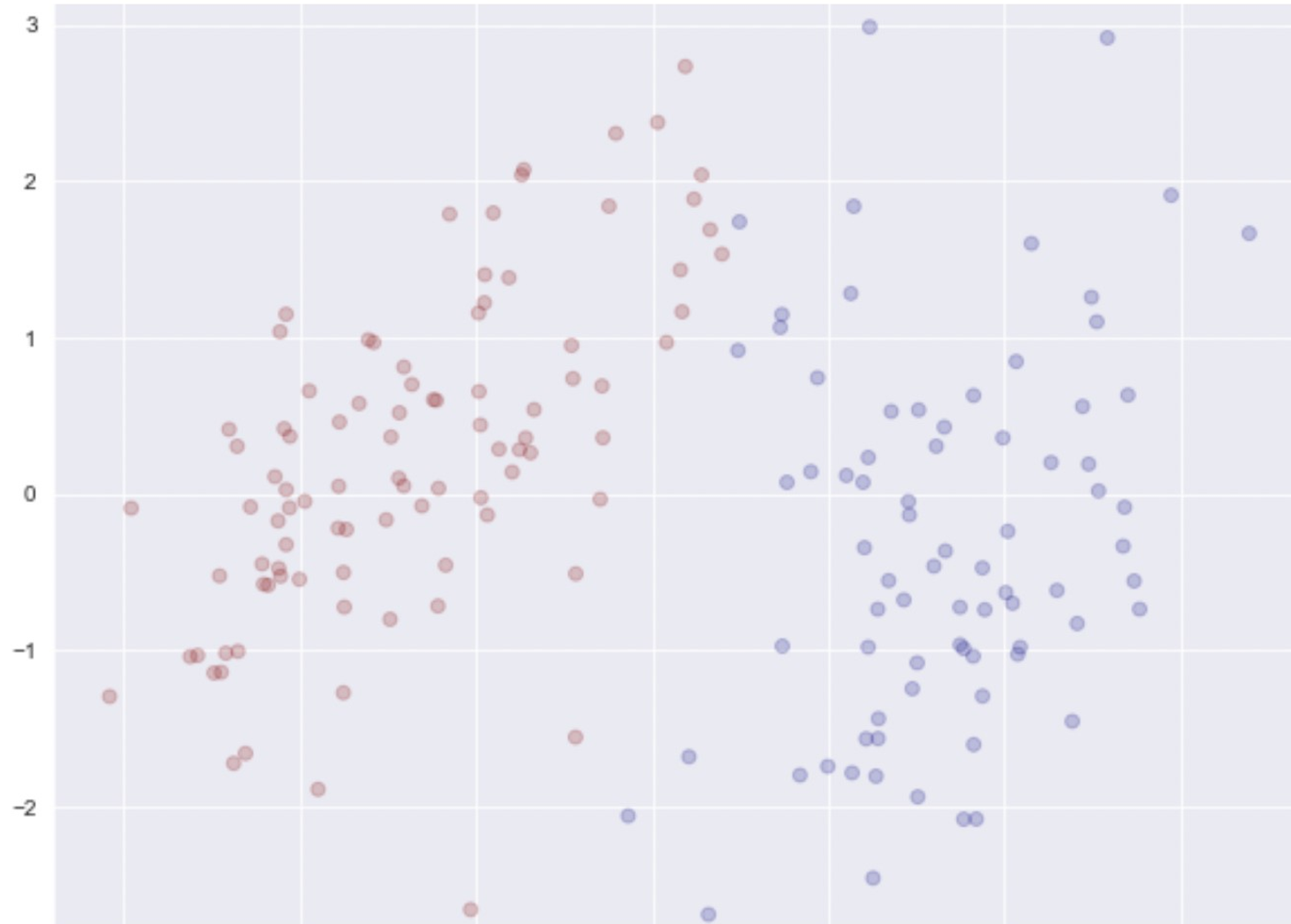
Deux classes distinctes

Nous voyons les deux modalités prises par la variable expliquée d'authenticité de billet .
Une forme de regroupement des modalités est visible de sorte que les modalités true et false qui ne se melangent que très peu.



ACP /contribution

Les variables length et diagonal sont fortement corrélées à F1 et à F2
La variable Length nous parait la plus representative
diagonal contribue fortement à F2



CLASSIFICATION AVEC KMEANS

Un nuage de points regroupés en deux clusters principaux .
Les points bleux representent les faux billets et l'autre couleur les bons billets .
On apercoit bien une grande variance inter-cclasse et une variance intra-classe petite.
Homoscedasticité semble correcte au sein d'une classe

REGRESSION LOGISTIQUE

Etapes de la modelisation

Nous classons avec le modèle de **regression logistique les billets selon leur authenticité.**
A partir d'un set et de la variable booleenne composée de deux modalites , **nous allons predire la qualité des billets**

1

Variables d'entrée
Creation des variables X et y

2

Scission du data set
Division du set en set de TTrain et TEST

03

Testing and learning
Application du modele de regression et de la variable predictive

04

Seuil de probabilite
Affichage des pourcentages selon le seuil fixé

MODELISATION :REGRESSION LOGISTIQUE

Avec statsmodel , variables significatives

On arrive avec la **P-valeur à discriminer les variables** .
Ici nous ne retenons que la variable *Length* qui est **significative au seuil de test de 5%**
Nous retenons que cette variable finaleemnt

```
alized Linear Model Regression Res
==================================
lse]', 'is_genuine[True]']   No. (
                       GLM   Df Re
                  Binomial   Df Mo
                     logit   Scale
                      IRLS   Log-L
         Fri, 19 Nov 2021   Devia
                 11:17:34   Pears
                        7
                nonrobust
==================================
          z      P>|z|      [0.025
----------------------------------
      6.292      0.000     378.028
     -6.298      0.000      -6.400
==================================
```

' et $\beta 2$ = -4.9

Modèle performant

Notre score du modele est au maximum et indique que les ppredictions sont justes et coherentes avec les observations reelles
Mais cele peut etre questionné

```
#Voyons le score de  notre modelisation
model.score(X_reg_test, y_reg_test)
```

```
1.0
```

```
#Evaluons la modelisation et la precision
print(classification_report(y_reg_test, y_reg_pred))
```

```
              precision    recall  f1-score   support

       False       1.00      1.00      1.00        14
        True       1.00      1.00      1.00        20

    accuracy                           1.00        34
   macro avg       1.00      1.00      1.00        34
weighted avg       1.00      1.00      1.00        34
```

Process/Testing

Mise en situation de test de notre algorithme

1

DIMENSION

Importation du set de test

Verification des diimensions des deux sets
Eventuel preprocessing

+

2

MODELISATION

Avec la fonction de probabilité sur le set de test

Creation de la var : y_test avec le code python
Affichage de la var y_test

+

3

JONCTION

Creation de colonne 'proba' au set initial

Transformation de cette variable en notre variable illustrative

+

4

TRANSFORMATION

Changer la var 'proba' en var de pourcentage

Utliser la fonction de python pour recuperer la deuxieme modalité

+

Conclusion

Economie resilliente

La data science peut aider à reduire la faux monnayge et ainsi optimiser la lutte contre ce crime.
Il faut que les decideurs publics s'initient aux usages du *machine learning* pour faire avancer l'économie

#TECHFORGOOD

Thanks!

www.companyweb.com