République Algérienne Démocratique et populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et Informatique

Département Informatique

Rapport Partie 1
Projet Data Mining

Techniques de Datamining

Réaliser par :

- ADMANE Hocine 171731054926
- HABCHI Lydia 161731064864
- DJAOUADI Yamina 161631080116
- TEBBANIMOHAMED WALID 171731088266
- SEDKAOUIAMINE 161731026140

Master spécialité : SII

Année universitaire : 2021/2022

Table des matières

1. Introduction :

Le data mining désigne le processus d'analyse de volumes massifs de données et du Big Data sous différents angles afin d'identifier des relations entre les data et de les transformer en informations exploitables.

En français, ce processus porte différents noms :

- Exploitation de données.
- Fouille de données.
- Forage de données.
- Ou encore extraction de connaissances à partir de données.

Avant d'arriver à l'application des tâches du Data Mining on doit passer par une étape primordiale, qu'est le prétraitement des données.

Le prétraitement des données consiste à nettoyer, intégrer, appliquer multiples transformations et réduire les données collectées de différentes sources. Avant d'entamer ces opérations, il est impérieux de bien étudier l'ensembles des données afin de les connaitre et les comprendre grâce à la phase exploratoire des données qui permet d'obtenir une vision globale et précise de l'ensemble de données.de plus elle facilite la détection des régularités telles que les corrélations et les dépendances entre les attributs, mais aussi les irrégularités telles que des données aberrantes ou du bruit.

Durant cette première partie du projet nous nous intéressons à l'étude exploratoire des données.et à l'implémentation d'un système qui permet de faire le prétraitement d'un Dataset.

2. Objectif du projet (partie 1) :

- Etudier le dataset seeds.txt, ainsi que rédiger une description globale du dataset et des attributs.
- Développer une application en java qui permet de réaliser les fonctionnalités suivantes :
  - Lecture du benchmark, manipulation de son contenu, extraire les attributs, afficher la description et les valeurs.
  - Affichage du dataset.
  - Ajout, Modification et suppression d'une instance.
  - Calculer les mesures de tendance centrale et en déduire les symétries.
  - Calculer les mesures de dispersion.

- Construire une boite à moustache pour chaque attribut et afficher les données aberrantes.
- Construire un histogramme et visualiser la distribution des données.
- Construire et afficher les diagrammes de dispersions et en déduire les corrélations entre les attributs.

3. Résumé sur la phase d'exploration des données :

La première manière consiste à définir ce que les variables représentent, identifier leurs types et la nature de leurs valeurs, la deuxième, c'est d'utiliser des méthodes statistiques qui aboutissent par la génération de graphes afin de clarifier les observations des différentes dimensions, techniquement parlant, il s'agit de calculer les paramètres de tendance centrale, qui sont la moyenne, la médiane et le mode qui donnent l'ordre de grandeur de l'ensemble des mesures, ainsi que les paramètres de dispersion, qui comportent le max, le min, la variance, les quartiles et d'autres. A la fin on cherche si elle existe une corrélation entres les attributs, et cela grâce aux diagrammes de dispersions et le coefficient de corrélation.

4. Analyse des données :

4.1 Description du Dataset « seeds »

Cette base de données représente différentes mesures des propriétés géométriques des grains appartenant à trois variétés différentes de blé.

Une visualisation de haute qualité de la structure interne du noyau a été détectée à l'aide d'une technique de rayons X doux.

Chaque instance est classée dans l'une des trois '3' classes suivantes :

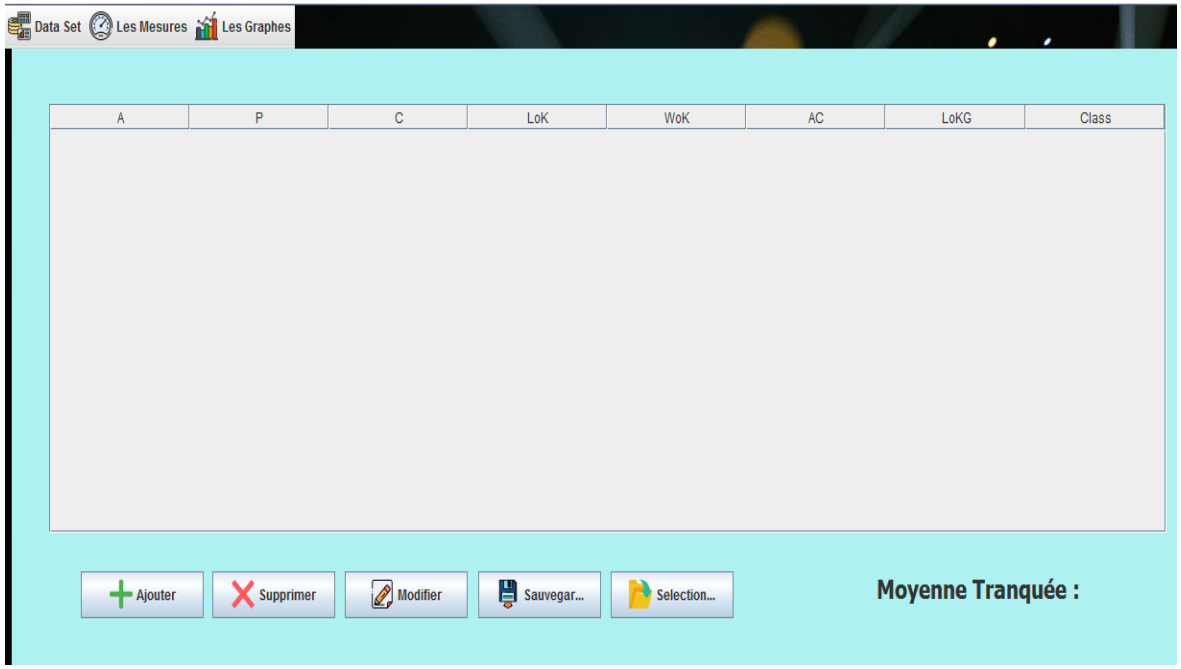
classe	**Nombre d'instances**
Rosa	69
Kama	65
Variétés canadiennes	76

- **Caractéristiques de l'ensemble de données :** Multivarité.
- **Nombre d'instance :** 210.
- **Nombre d'attribut :** 7.
- **Valeurs manquantes :** pas de valeurs manquantes.

4.2 Description des attributs :

N°	Nom	Description	Type	Valeurs possibles
1	A	la surface du grain	Numérique	[10.59 -21.18]
2	P	Le périmètre du grain ‘’	Numérique	[12.41-17.25]
3	C	Compacité du grain ‘C’est calculer à partir des attributs ‘A’et ‘P’avec la formule : C=4*pi*A/P^2.	Numérique	[0.808-0.9183]
4	LOK	La longueur du grain ‘’	Numérique	[4.899-6.675]
5	WOK	La largeur du grain ‘’	Numérique	[2.63-4.033]
6	AC	Coefficient d’asymétrie ‘’	Numérique	[0.7651-8.456]
7	LOKG	Longueur du sillon du noyau ‘’	Numérique	[4.519-6.55]

5. Présentation de L’interface de l’application :



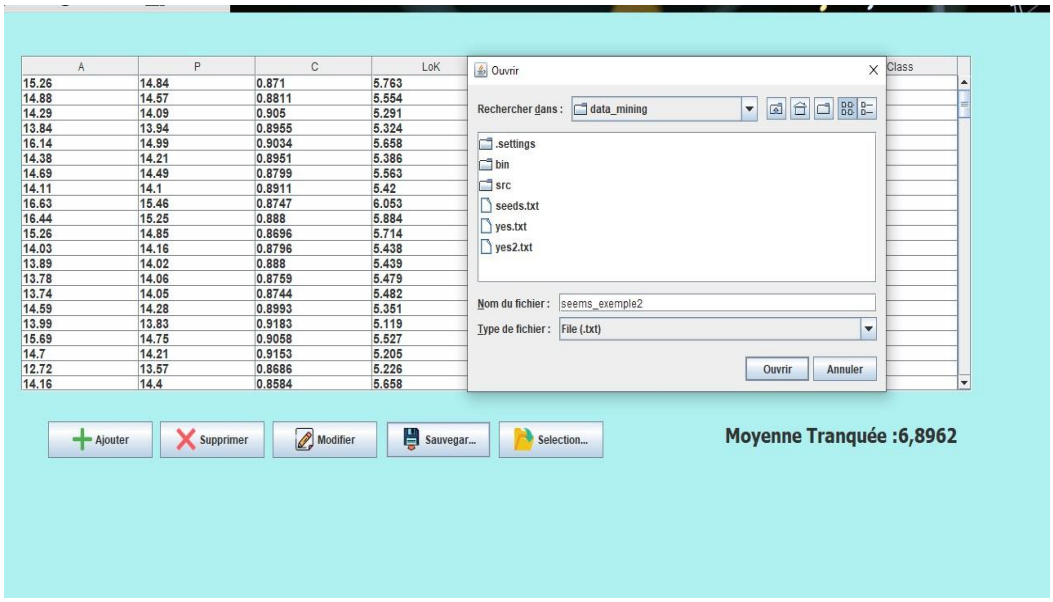
Sur la première fenêtre on a plusieurs boutons qui nous permettent d'afficher et de manipuler le dataset.

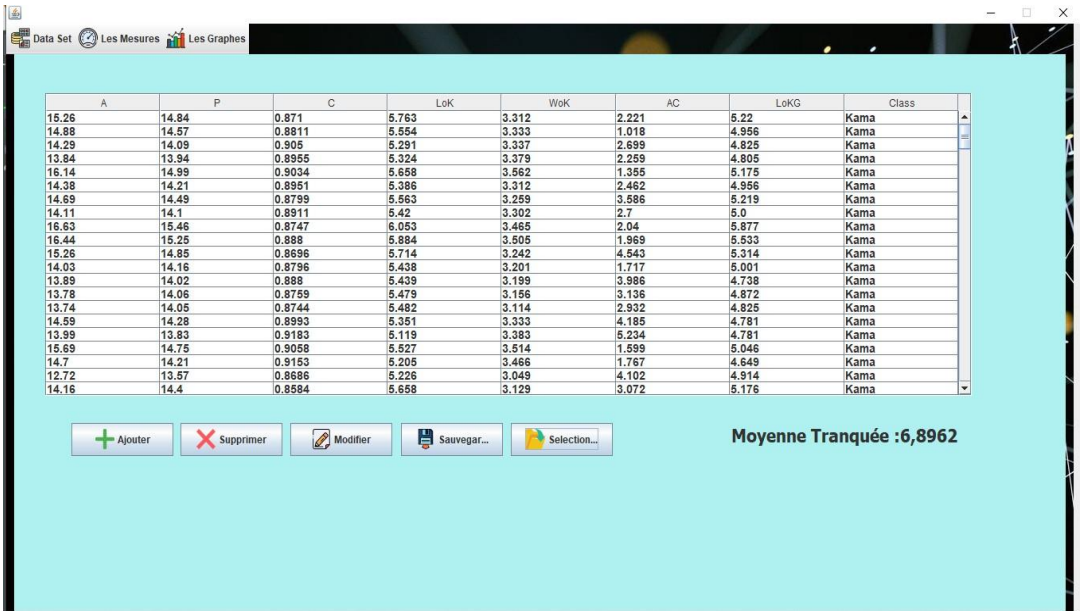
- Sur le bouton **« dataset »** on peut afficher soit :
  - Le data set (mais avant on doit sélectionner le fichier du data set).
  - La description du data set sélectionné.
  - Ou des informations sur les différents attributs.
- Le bouton **« Les Mesures »** nous permet d'afficher les résultats des calculs des paramètres de tendance centrale ainsi que les mesures de dispersion.
- Sur bouton **« Les Graphe »** on peut choisir d'afficher :
  - L'histogramme ou Box plot d'un attribut qu'on doit spécifier.
  - Scatter plot ou QQ-plot de deux attribut.
  - Ainsi que le diagramme de tous les box plots de nos attributs.
- Le bouton **« Ajouter »** nous permet d'ajouter une instance à la fin de noter fichier.
- Le bouton **« Supprimer »** pour supprimer une instance mais avant de cliquer sur se bouton il faut d'abord sélectionner l'instance à supprimer.
- Le bouton **« Modifier »** Sur se bouton on peut modifier les valeurs de n'importe quel attribut dans notre data set.
- Le bouton **« sauvegarder »** nous permet de sauvegarder le data set dans un fichier .txt.
- Et en dernier le bouton **« sélectionner »** pour sélectionner le fichier qui contient notre dataset.

6. Test de l'application sur Dataset :

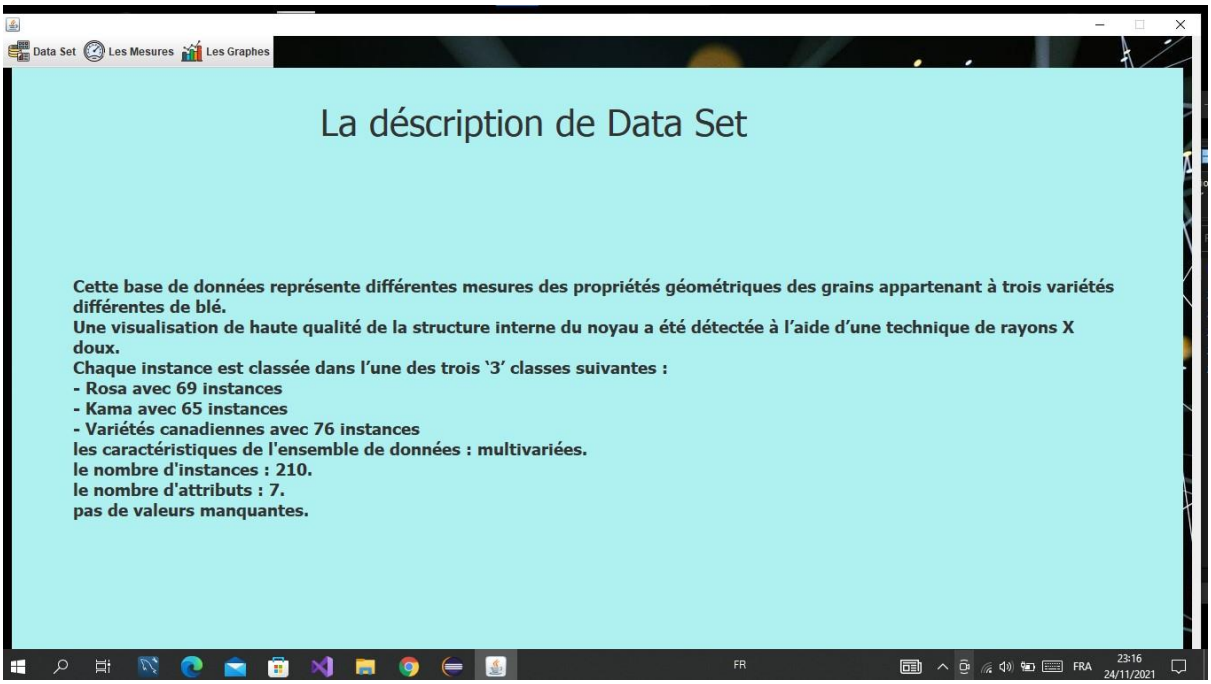
6.1 Lecture est affichage de dataset :

Le bouton dataset nous permet soit d'afficher le data set (mais avant on doit sélectionner le fichier du data set) et cette fenêtre va s'afficher :

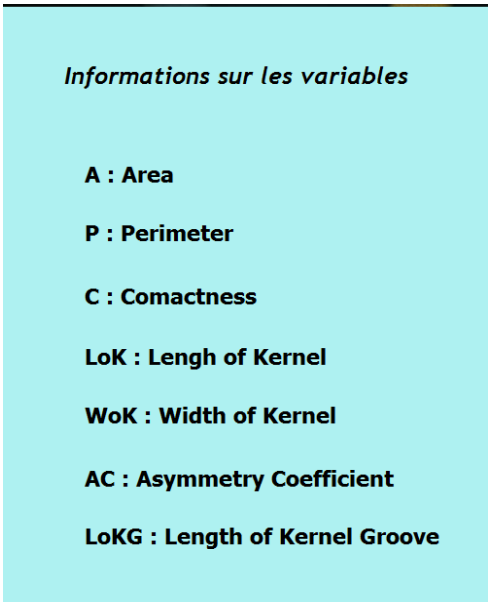




Soit on choisi d’afficher la description suivante de notre data set :



Soit on choisi d'afficher les Informations sur les différents attributs de notre data set :



6.2 Phase de prétraitement :

On a pas appliqué la normalisation sur nos données ni le traitement des valeurs manquantes car les valeurs de notre Dateset sont dans le même intervalle et ne contient pas de valeurs manquantes.

6. 3 Calcul des paramètres de tendance centrale :

a. Moyenne (Mean) :

La moyenne est obtenue en sommant l'ensemble des observations N d'un attribut X. Cette somme sera divisée par le nombre total des observations.

b. Moyenne tronquée :

Une moyenne tronquée, ou moyenne réduite, est une mesure statistique de centralité. Similaire à la moyenne arithmétique et à la médiane, qui consiste à calculer une moyenne arithmétique en éliminant les valeurs extremes.

c. Mode (Mode) :

Le mode représente la valeur la plus fréquente dans l'ensemble des observations. La série peut avoir un seul mode, appelée série unimodale, ou plusieurs, bimodale, trimodale ou multimodale.

d. Midrange :

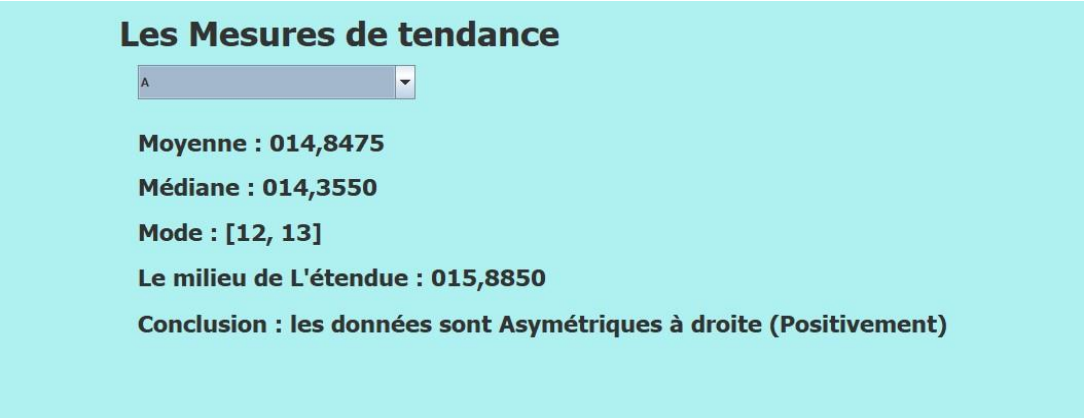
C'est la moyenne entre le max et le min de la source de données

Max – min /2.

Les calculs pour ce Data Set donneront :

	A	P	C	LOK	WOK	AC	LOKG
Moyenne	**14.8475**	**14.559**	**0.871**	**5.629**	**3.259**	**3.700**	**5.408**
Q2 médiane	**14.355**	**14.32**	**0.8735**	**5.5235**	**3.2370**	**3.599**	**5.223**
Mode	**[12-13]**	**[13-14]**	**[0-1]**	**[5-6]**	**[3-4]**	**[4-5]**	**[5-6]**
Midrange	**15.885**	**14.830**	**0.863**	**5.787**	**3.332**	**4.611**	**5.534**
Moyenne tronquée	**17.32**						

Exemple d'affichage des résultats de calculs des paramètres de tendance centrale de l'attribut 'A':



6.4 Calcul des mesures de dispersion :

a. Maximum (Max) :

C'est La plus grande valeur que peut prendre un attribut.

b. Minimum(Min) :

C'est La plus petite valeur que peut prendre un attribut**.**

c. L'Etendue :

Il représente la différence entre la valeur maximale et la valeur minimale de la source de données.

Max - min

d. Médiane (Median/Q2):

La médiane est la valeur du milieu dans la liste des observations ordonnées.

Si le nombre d'observation N est pair : N= 2p.

Alors on fait la moyenne entre la p ème et la $(p+1)^{ème}$ valeur.

Sinon Si le nombre d'observation N est impair : N=2p+1.

Alors Q2= $(p+1)^{ème}$ valeur.

e. Quartiles (Q1 et Q3) :

Les quartiles sont les valeurs qui divisent un ensemble de valeurs en 4 sous-ensembles de même taille.

Avant de calculer les quartiles, il faut d'abord ranger les valeurs par ordre croissant.

Q1= la (N/4) ème valeur de la liste.

Q3=la (3N/4) ème valeur de la série.

f. Ecart interquartile :

IQR = Q3 – Q1.

g. Variance :

Est calculé avec la formule suivante :

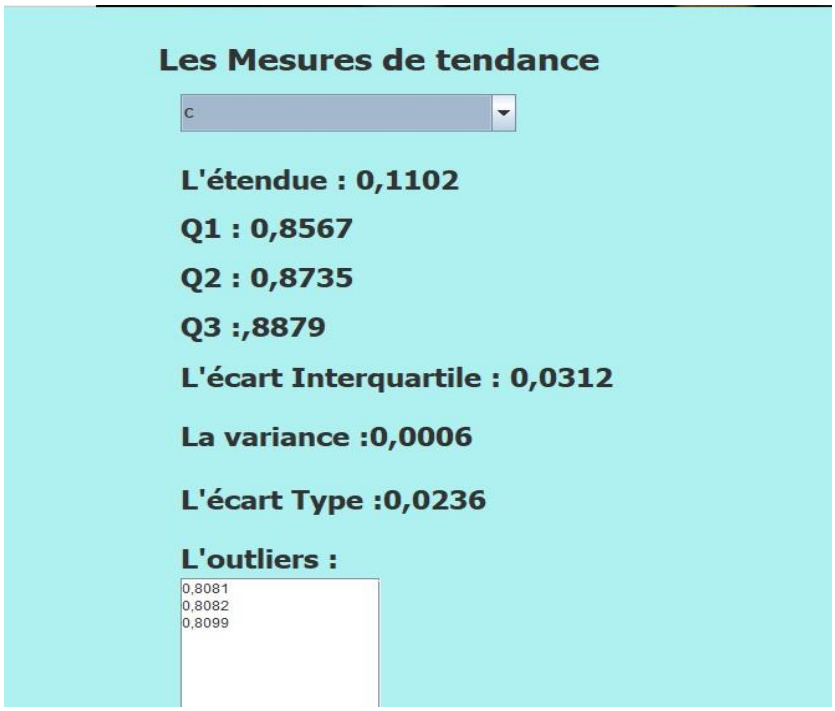
$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

h. Ecart-type.

c'est la racine carrée de la variance

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Exemple d'affichage des résultats des mesures de dispersion de l'attribut C sur l'interface de notre application java :



	A	P	C	LOK	WOK	AC	LOKG
Maximum	**21.18**	**17.25**	**0.9183**	**6.675**	**4.033**	**8.456**	**6.55**
Minimum	**10.59**	**12.41**	**0.8081**	**4.899**	**2.63**	**0.7651**	**4.519**
Etendue	**10,59**	**4,84**	**0,11**	**1,77**	**1,403**	**7,69**	**2,031**
Q1	**12.26**	**13.45**	**0.8567**	**5.2620**	**2.941**	**2.553**	**5.045**
Q2	**14.355**	**14.32**	**0.8735**	**5.5235**	**3.2370**	**3.599**	**5.223**
Q3	**17.32**	**15.73**	**0.8879**	**5.98**	**3.5620**	**4.773**	**5.877**
Ecart interquartile	**5.060**	**2.280**	**0.031**	**0.718**	**0.621**	**2.220**	**0.832**
Variance	**8.4664**	**1.7055**	**0.0006**	**0.1963**	**0.1427**	**2.2607**	**0.2416**
Ecart-type	**2.9097**	**1.3060**	**0.0236**	**0.4431**	**0.3777**	**1.5036**	**0.4915**

7. Mesures de dispersion, histogramme et diagrammes de dispersion :

a. Analyse et conclusion (Asymétrie):

Une distribution est dite symétrique si la moyenne, la médiane et le mode sont presque égaux ou égaux, elle est dite asymétrique sinon.

On dit qu'une distribution est asymétrique à gauche si la médiane est plus grande que le mode, dans le cas contraire on dit qu'elle est asymétrique à droite.

$Moyenne = Médiane = mode \Rightarrow Symétrique$
$Moyenne < Médiane < Mode \Rightarrow Asymétrique$ à $gauche$ (Négativement)
$Moyenne > Médiane > Mode \Rightarrow Asymetrique$ à $droite$ (Positivement)

Dans notre cas :

L'attribut A est symétrique.

L'attribut P est asymétrique à droite.

L'attribut C est symétrique.

L'attribut LOK est asymétrique à droite.

L'attribut WOK est asymétrique à droite .

L'attribut AC est asymétrique à droite.

L'attribut LOKG est asymétrique à droite.

b. Boite à moustache (box plot) :

Graphique résumant l'information fournie par l'étendue, ainsi que par les trois quartiles et les intervalles qui les séparent.

Un box-plot est un graphique simple composé d'un rectangle duquel deux droites sortent afin de représenter certains éléments des données.

Interprétation :

On remarque que l'attribut 'A' a le plus grand étendu par rapport aux autres, ses données sont légèrement asymétriques et ceux des attributs WoK et AC sont symétriques, on remarque qu'on a des outliers uniquement dans les attributs C et AC. On remarque également que l'attribut C comporte la plus petite plage de données sortent afin de représenter certains éléments des données.

Attribut	Valeurs aberrantes
A	Non constatés
P	Non constatés
C	Constatés 0,8081 - 0,8082 - 0,8099
LOK	Non constatés
WOK	Non constatés
AC	Constatés 08,3150 - 08,4560
LOKG	Non constatés

c. Histogrammes :

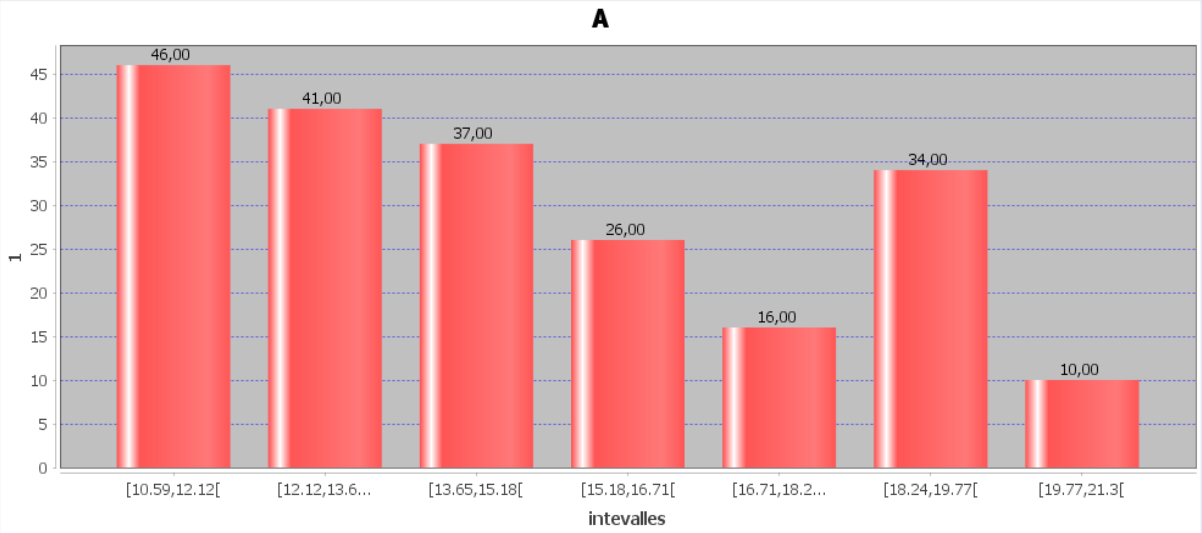
Représentation graphique des fréquences ou effectifs relatifs à un caractère quantitatif continu à l'aide d'une série de rectangles dont la base constitue un intervalle de variation des valeurs du caractère et la surface l'effectif correspondant.

Le nombre de classes sur lesquelles doivent être réparties les valeurs a été calculé par la formule suivante :

$$K = 1 + \log_2 N \approx 1 + \frac{10}{3} \log_{10} N$$

c. 1 histogrammes et interprétations :

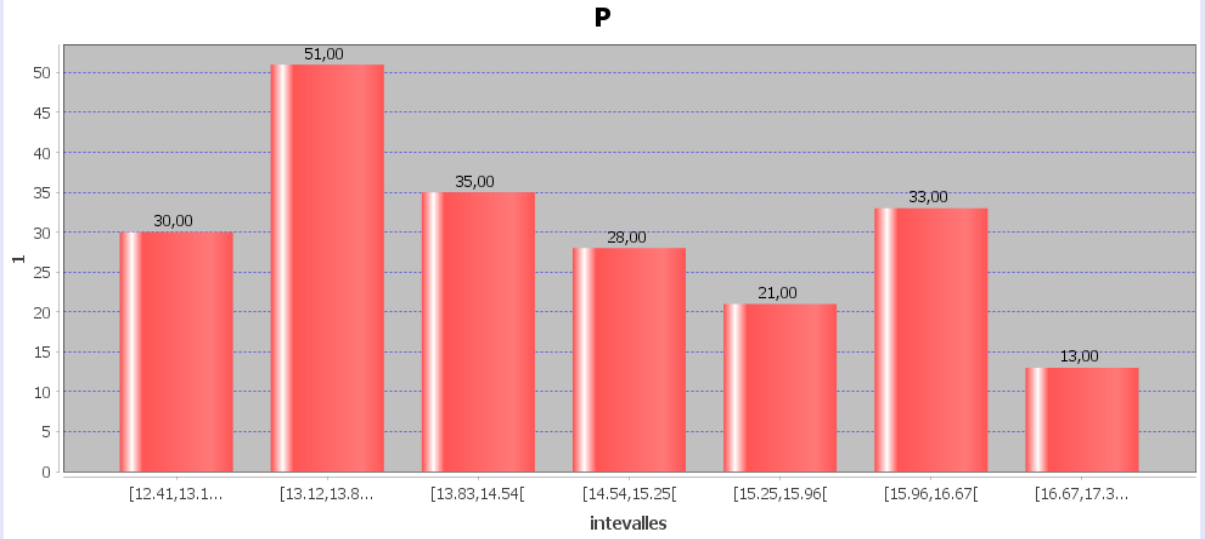
Histogramme du 1er attribut :



Interprétation :

D'après le diagramme à bâtons, on remarquera que nous avons 7 classes de valeurs pour cet attribut. Mais la majorité des valeurs se trouve dans la classe [10.59-12.12 [qui a une fréquence de 45 valeurs.

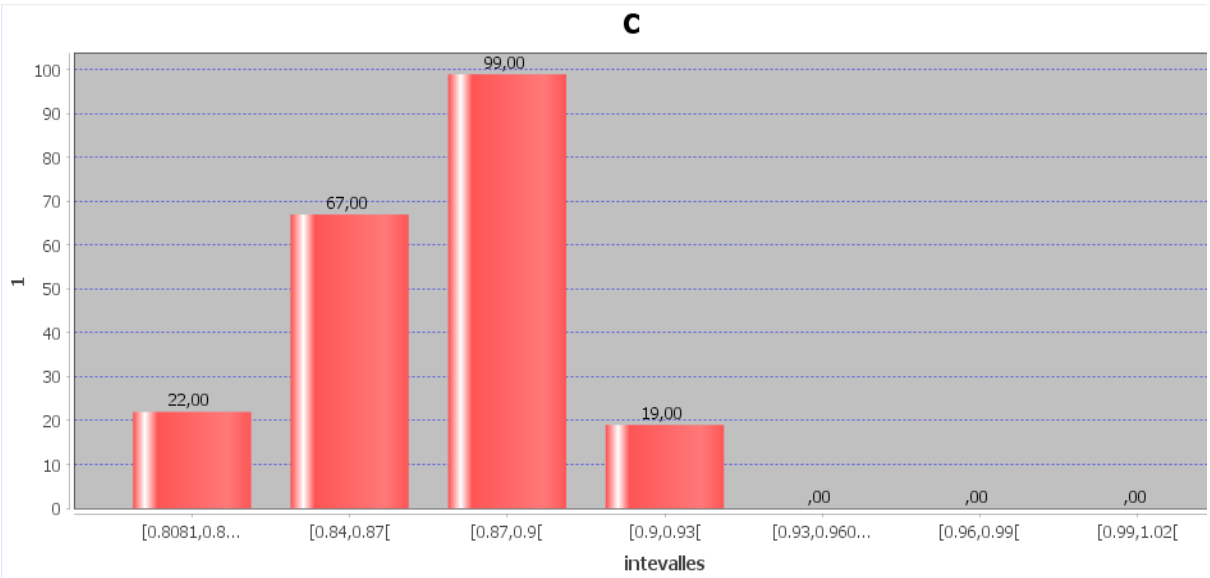
Histogramme du 2ème attribut :



Interprétation :

D'après le diagramme, on remarquera que nous avons 7 classes de valeurs pour cet attribut. Mais la majorité des valeurs se trouve dans la classe [13.12-13.8 [qui a une très grande fréquence de 51 valeurs.

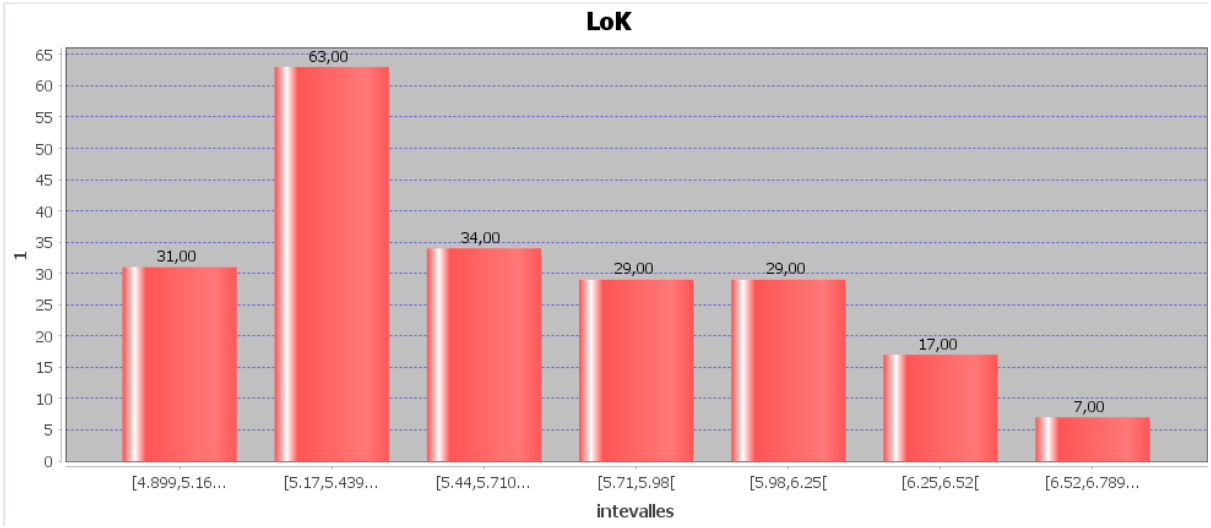
Histogramme du 3ème attribut :



Interprétation :

D'après le diagramme à bâtons, on remarquera que nous avons que 4 classes de valeurs pour cet attribut. Mais la majorité des valeurs se trouve dans la classe [0.87-0.9 [qui a une fréquence proche de 100 valeurs.

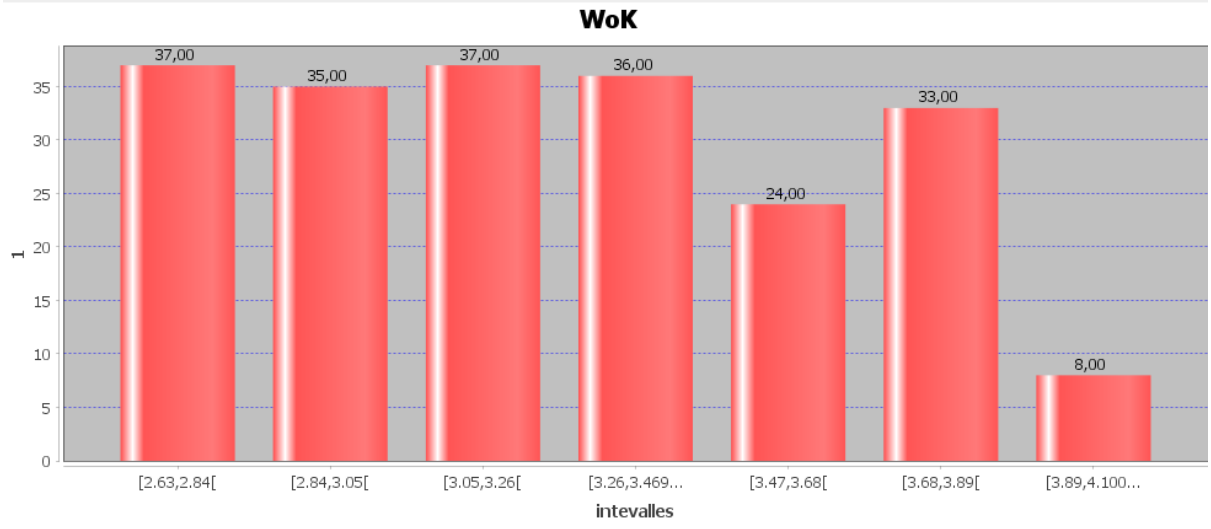
Histogramme du 4ème attribut :



Interprétation :

D'après le diagramme à bâtons, on remarquera que nous avons 7 classes de valeurs pour cet attribut. Mais la majorité des valeurs se trouve dans la classe [5.17-5.439 [qui a une fréquence importante par rapport au autre clase de 63 valeurs.

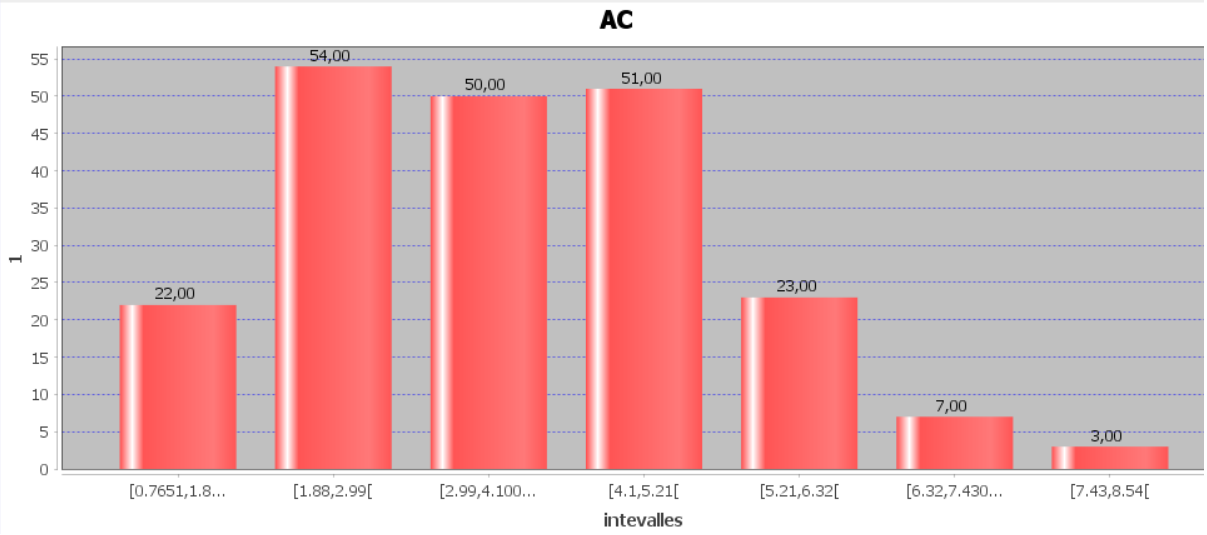
Histogramme du 5ème attribut :



Interprétation :

D'après le diagramme à bâtons, on remarquera que nous avons 7 classes de valeurs pour cet attribut. Mais la majorité des classes on une fréquence entre 35 36 et 37 valeurs.

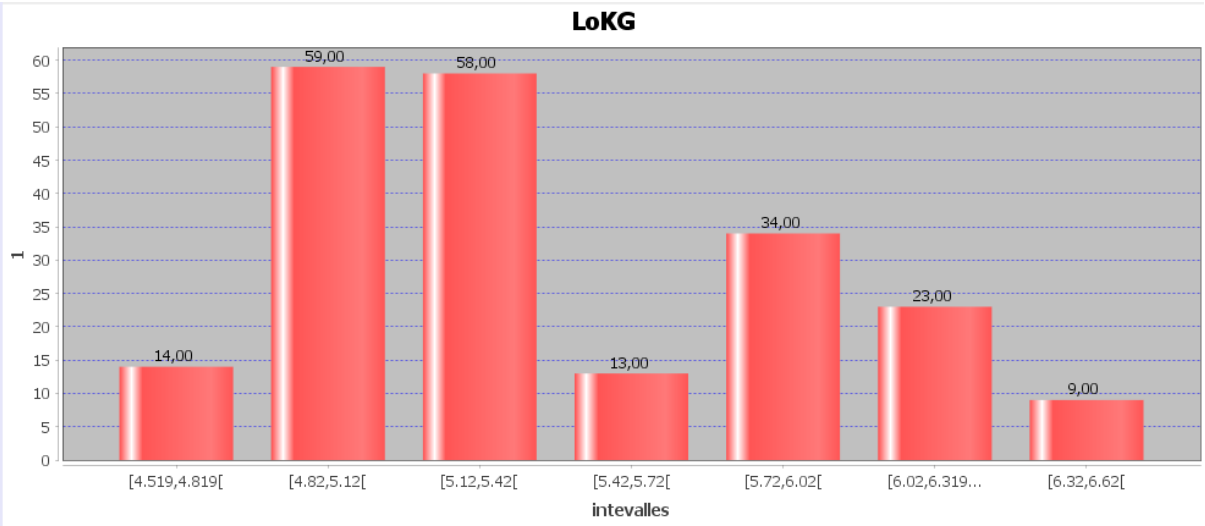
Histogramme du 6ème attribut :



Interprétation :

On remarquera que nous avons 7 classes de valeurs pour cet attribut. Mais la majorité des valeurs se trouve dans les 3 classes [1.88-2.99 [[2.99-4.1[[4.1-5.21[qui a une fréquence entre 50 et 54 valeurs.

Histogramme du 7ème attribut :



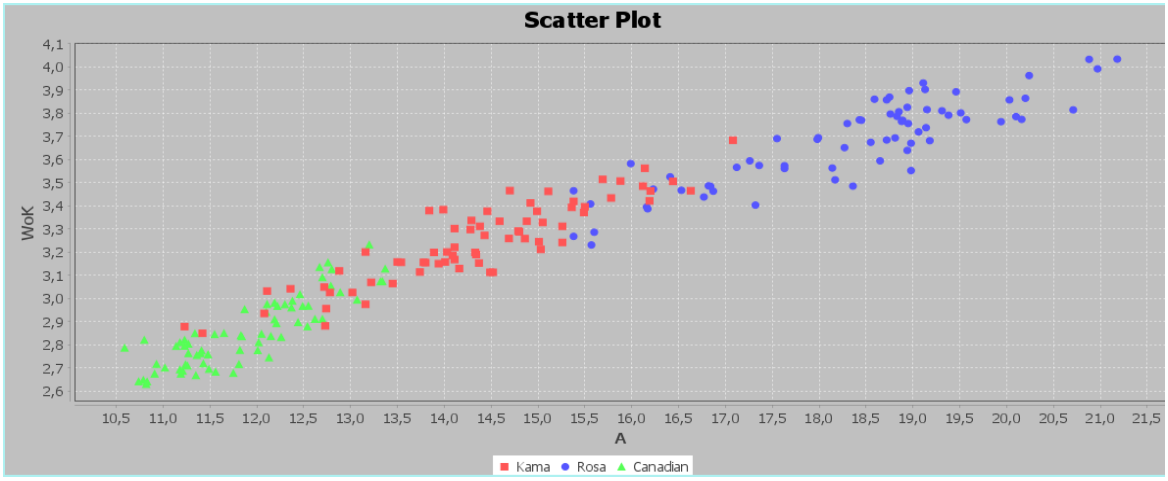
Interprétation :

D'après le diagramme à bâtons, on remarquera que nous avons 7 classes de valeurs pour cet attribut.et ca ce voix clairement que la majorité des valeurs se trouve dans les deux classe [4.82-5.12 [et [5.12-5.42 [avec une fréquence de 58 et 59 valeurs.

d. Diagramme de dispersion :

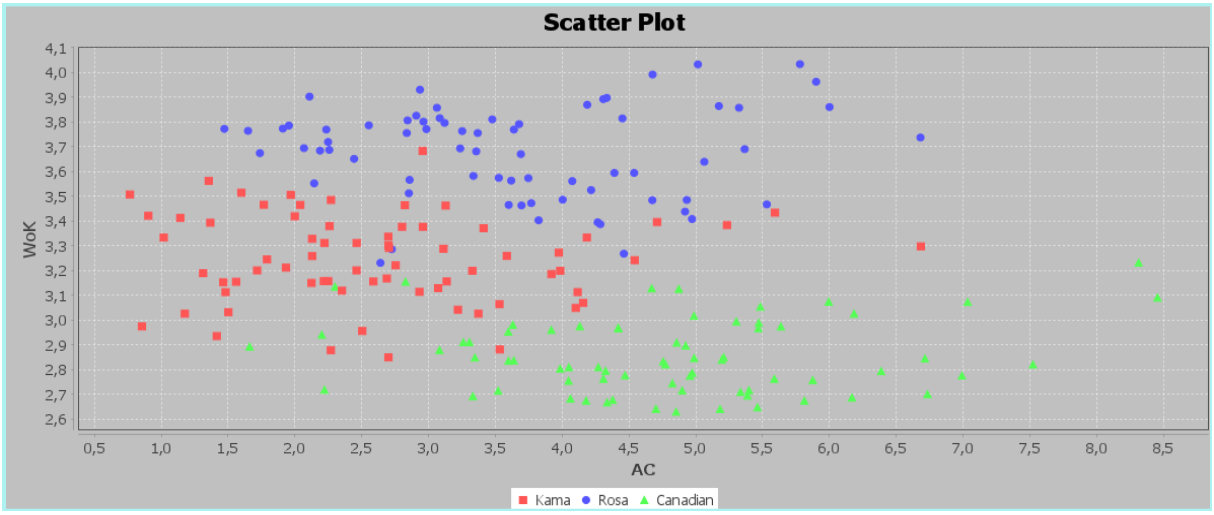
d.1 Scatter plot :

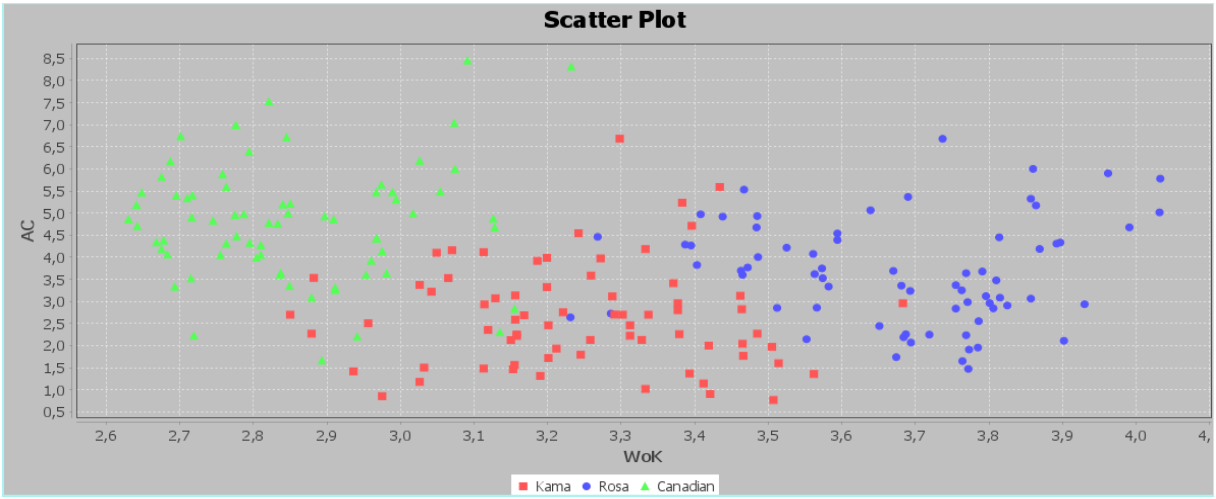
Scatter plot est appelé aussi nuage de points permet de visualiser les mesures de deux attributs qui peuvent être liés entre eux, les ensembles des valeurs des deux variables sont situées sur les axes x et y. scatter plot est très simples et efficace, à partir de ce graphe on peut déduire si il ya une corrélation entre les deux variables ou non.



Interprétation:

D'après le nuage de points ci-dessus on peut en conclure que les données des deux attributs sont corrélées et par conséquent on pourrait effectuer une réduction de données en éliminant les données (colonne) d'un des deux attributs.





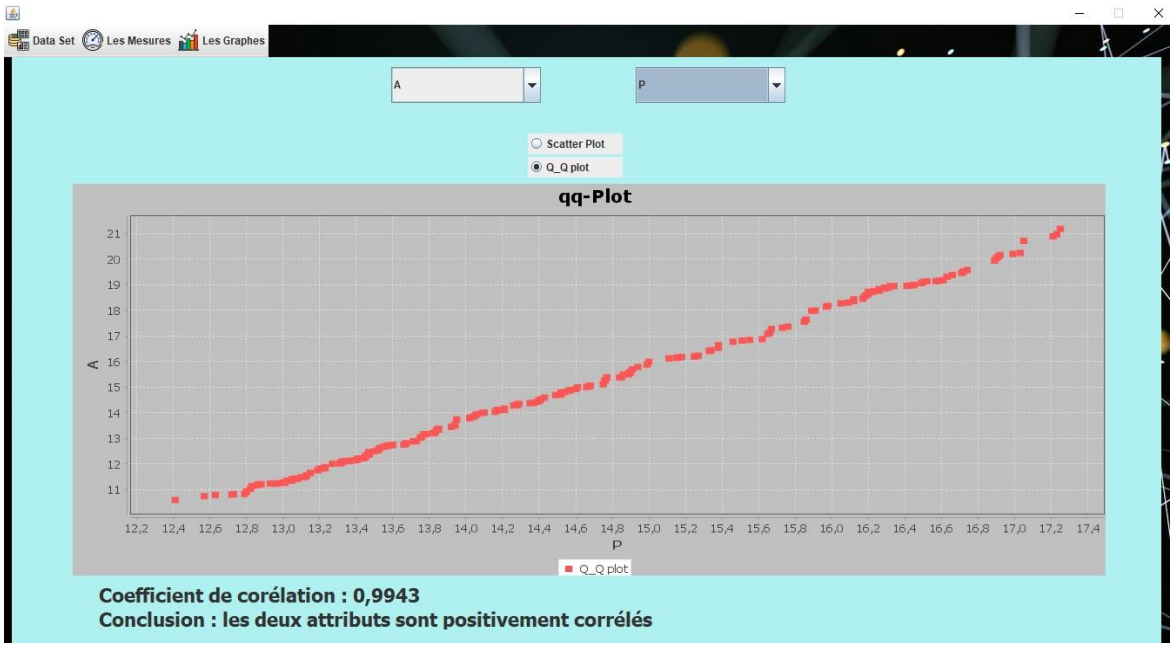
Interprétation:

On ne remarquera aucun cumul de points donc il n'existe aucune corrélation entre les deux attributs. On verra par la suite le coefficient de corrélation dans le tableau de calculs. On verra qu'il est négatif proche du 0 ce qui démontre la très faible corrélation ou bien l'absence de corrélation.

d.2 QQ-Plot :

Le diagramme Quantile-Quantile ou diagramme Q-Q ou Q-Q plot est un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique.

Le diagramme quantile-quantile permet également de comparer deux distributions que l'on estime semblables.



Interprétation :

On remarque qu'il existe une corrélation positive forte entre les deux attributs 'A' et 'P'. On verra par la suite le coefficient de corrélation dans le tableau de calculs. On verra qu'il est positif et supérieur à 0,5, il approche 1 ce qui démontre la forte corrélation détectée.

e. Analyse de corrélations :

Le coefficient de corrélation est calculé a en utilisant, la formule suivante :

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

Si $r_{A,B} > 0$ **alors** corrélation positive (forte, moyenne, faible)

Si $r_{A,B} < 0$ **alors** corrélation négative

Sinon pas de corrélation.

On calcule le coefficient de corrélation pour chaque paire d'attributs :

	A	P	C	LOK	WOK	AC	LOKG
A	1.0	0.9943	0.6083	0.95	0.9708	-0.2296	0.8637
P	0.9943	1.0	0.5292	0.9724	0.9448	-0.2173	0.8908
C	0.6083	0.5292	0.99	0.3679	0.7616	-0.3315	0.2268
LOK	0.95	0.9724	0.3679	0.99	0.8604	-0.1716	0.9328
WOK	0.9708	0.9448	0.7616	0.8604	1.0	-0.2580	0.7491
AC	-0.2296	-0.2173	-0.3315	-0.1716	-0.2580	1.0	-0.0111
LOKG	0.8637	0.8908	0.2268	0.9328	0.7491	-0.0111	1.0

8. Conclusion :

Lors de cette première partie du projet on a pu atteindre nos objectifs fixés au départ :

- L'extraction et l'affichage du contenu du dataset, extraction des attributs, affichage de la description.
- Calcul des mesures de tendance centrale et en déduire les symétries des attributs A, C et la non symétrie du reste des attributs P, LOK, WOK, AC, LOKG.
- Affichage des histogrammes pour chaque attribut et en déduire les valeurs les plus fréquentes.
- Affichage de la boîte à moustache pour chaque attribut et en déduire les données aberrantes. Dans notre cas les attributs C et AC contiennent des données aberrantes qu'on devra corriger dans le prétraitement.
- Affichage des diagrammes de dispersion, calcul du coefficient de corrélation et déduction des corrélations.

Nous avons détecté une forte corrélation positive entre quelque attributs Mais aussi une forte corrélation négative entre autre attributs.

D'après le travail qu'on a fait on déduit que la phase d'analyse et du prétraitement des données est très importante avant de faire du Data Mining.