**Interprétation :**

**Interprétation ETL :**

**ETL (Extract, Transform, Load)** sont des processus qu’on les applique à des données pour avoir des données qui ont disponibles, prête pour l’utiliser dans les analyses comme dans notre sujet de l’agriculture pour l’avoir appliqué à la visualisation.

**1ér Etape: Cleaning d’une data Set**

L’objectif de cleaning c’est pour nettoyer les données alors éliminer les valeurs manquantes et de remplacer valeurs aberrantes par la moyenne et aussi de corriger les erreurs de saisie et supprimer les doublons, car ces valeurs peuvent fausser les analyses et les résultats ; et de le mettre dans un nouveau fichier qui contient juste les valeurs avec un sens.

**2éme Etape : Extract**

C’est permet de récupérer les données qui ont déjà nettoyé dans la partie Cleaning à partir de notre nouvelle data Set en les préparer pour la phase de transformation pour les utiliser dans les analyses.

**3éme Etape : Transformation**

C’est permet de préparer les données extraites dans la phase Extract après leur Cleaning pour qu’elles soient prêtes à charger dans la partie Load.

**4éme Etape : Load**

C’est permet à intégrer les données nettoyés dans la partie Cleaning et qui ont transformé dans la partie Transformation après leur extraction, et les mettre dans une base de donnée SQL Workbench. Après la création d’une base de donnée (agriculture) et d’avoir connecté à la base de donnée, et de créer les tables nécessaire (T\_crop).On l’importe les données dans notre base de donnée.

**Analyse Exploratoire :**

L’objectif de l’analyse exploratoire des données (AED) d’avoir comprendre les données et les explorer, et de découvrir des tendances, des schémas pour générer des hypothèses, de valider des idées et de préparer les données pour des analyses plus approfondies.

D’abord, nous avons fait **shape**: pour afficher le nombre d’observations c’est 2200 et le nombre de variables qui est 8. **Info :** permet de résumer le type des données (N, P, K => **int**) ; (température, humidity, ph, rainfall => **Float**) et label de type **object. Descripe :** permet d’afficher les statistiques telles que la moyenne de chaque variable, medium, etc...Et aussi **mean** pemet d’afficher la moyenne de chaque variable (N=> 50.55 ; K=> 48,14 ; P=>53,36 ; température=>25,61 ; humidity=> 71,48 ; ph=>6,46 ; rainfall=>103,46).

**Pour afficher chaque variable dans un graphe :**

**Pour rainfall :**

Nous avons fait un histograme pour **rainfall** en horizontal présente la variable rainfall et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (rainfall de 50 à 100 est augmenté).

**Pour N :**

Nous avons fait un histograme pour **N** en horizontal présente la variable **N** et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (N de 0 à 40 est augmenté).

**Pour P :**

Nous avons fait un histograme pour **P** en horizontal présente la variable **P** et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (P de 10 à 80 est augmenté).

**Pour K :**

Nous avons fait un histograme pour **K** en horizontal présente la variable **K** et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (K de 25 à 50 est augmenté).

**Pour température :**

Nous avons fait un histograme pour **température** en horizontal présente la variable **température** et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (Température de 23 à 30 est augmenté).

**Pour humidité :**

Nous avons fait un histograme pour **humidité** en horizontal présente la variable **humidité** et on vertical présente le compte des observations **=>** Nous avons dans le graphe que la distribution augmente indique une dispersion plus élevée des valeurs autour de la moyenne. (humidité de 80 à 95 est augmenté).

**Pour la matrice de corrélation**

Tout d’abord nous avons encodé la variable label pour être numérique.

**(N, P) ; (K, N) ; (labe, N) ; (Température, P) ; (humidity, P) ;(Ph, P) ;** **(rainfall, P) ; (label, P) ; (température, K) ; (rainfall, K) ;(label, K) ; (ph, température) ; (rainfall, température) ;**

**(Ph, humidity) ; (rainfall, ph) ; (label, ph) =>** ils ont des valeurs négative (proche de -1), cela signifie que la tendance à varier de manière opposé. Une fois la valeur d’une variable augmente alors la valeur de l’autre variable diminue.

**(Température, N) ; (humidity, N) ; (ph, N) ; (rainfall, N) ; (K, P) ; (humidity, K) ; (humidity, température) ; (label, température) ; (rainfall, humidity) ; (label, humidity) ; (label, rainfall) =>** ils ont des valeurs positives proche de 1, cela signifie que la tendance à augmenter. La valeur d’une variable augmente alors l’autre variable augmente.

**Des histogrammes de chaque variable avec label :**

**Rainfall avec label :**

Un histogramme qui contient la moyenne de rainfall avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(pigeonpas, mothbeans, lentil, jute)** => ont de moyenne très élevé par rapport au rainfall ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides. **(Orange, papaya, coconut, cotton, rice, maiz, pomegranate)** => ont de moyenne au moyen**. (chikpea, kidneybeans, mungbean, blackgram, banana, watermelon, muskmelon, apple, coffe)** => ont de moyenne très faible alors nous pouvons indiquer que cette zone géographique n’est pas sujette à des conditions.

**N avec label :**

Un histogramme qui contient la moyenne de N avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(Maiz, mothbeans, mugbean, apple, jute, coffe)** => ont de moyenne très élevé par rapport au N ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides. **(Lentil, mango, papaya)** => ont de moyenne au moyen**. (rice, chikpea, kidneybeans, blackgram, pigeonpas, pomegranate, banana, grapes, watermelon, muskmelon, orange, cocunut, cotton,)** => ont de moyenne très faible alors nous pouvons indiquer que cette zone géographique n’est pas sujette à des conditions.

**P avec label :**

Un histogramme qui contient la moyenne de P avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(Rice, blakcgrame)** => ont de moyenne très élevé par rapport au P ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides. **(maiz, chikpea, kidneybeans, pomegranate, banana, papaya, coconut)** => ont de moyenne au moyen**. (pigeonpas,mothbeans,mugbean, lentil, grapes, watermelon, muskmelon, apple, orange, cotton,jute, coffe)** => ont de moyenne très faible alors nous pouvons indiquer que cette zone géographique n’est pas sujette à des conditions.

**Ph avec label :**

Un histogramme qui contient la moyenne de ph avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(Rice, blakcgrame, pigeonpas,mothbeans,mugbean, lentil, grapes, watermelon, muskmelon, apple, orange, cotton,jute, coffe, maiz, chikpea, kidneybeans, pomegranate, banana, papaya, coconut)** => ont de moyenne très élevé par rapport au P ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides**.**

**Température avec label :**

Un histogramme qui contient la moyenne de température avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(rice, Maiz, chikpea, pigeonpas, mthbeans, mugbean, blackgram, lentil, mango, papaya, chikpea, kidneybeans, blackgram, pigeonpas, pomegranate, banana, grapes, watermelon, muskmelon, orange, cocunut, cotton,)** => ont de moyenne très élevé par rapport à la température ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides. **(pigenpos)**=> a une moyenne au moyen**.**

**Humidité avec label :**

Un histogramme qui contient la moyenne de température avec chaque groupe de label c’est pour observer les variations des moyenne par rapport au chaque groupe c’est pour comprendre la relation entre les variables et les données. **(rice, Maiz, chikpea, pigeonpas, mothbeans, mugbean, blackgram, lentil, banana, mango, grapes, watermelon, muskmelon,apple, orange,papaya, ,cocunut, cotton,jute,coffe)** => ont de moyenne très élevé par rapport à l’humidité ; nous pouvons indiquer que cette zone géographique est sujette à des conditions météorologique humides. **(kidneybeans, ponegrate)**=> a une moyenne au moyen**.**

**Un diagramme de dispertion :**

Un diagramme de dispersion qui montre la relation entre les précipitations (sur l'axe horizontal on a rainfall) et la température (sur l'axe vertical). Les points sont colorés en fonction des valeurs de la variable "label" pour distinguer différentes catégories ou groupes de données. C’est pour afficher toutes les caractéristiques de **label** en fonction de la **température** et **rainfall.** Nous avons trouvé que Papaya besoin de plus de la température et de rainfall par rapport au d’autre caractéristiques. Le maiz besoin de plus de rainfall. Alors que grappe et orange ont besoin moins que les autres soit pour rainfall ou pour température. Mango besoin de plus de température que de rainfall. Pour les autres caractéristiques c’est à peu près de la même chose pour la température et rainfall.

**Des diagrammes de label avec les d’autres variables :**

Avec Pairplot nous avons affiché chaque variable avec label pour voir les caractéristiques de label à propos de la variable (N, P, K, humidité, température, rainfall, PH). Par exemple : apple, mungbean sont besoin de plus de N que les autres caractéristiques.

**Créer des graphiques catégoriels :**

C’est permet de visualiser la distribution d’une variable numérique en fonction de catégories spécifique.

Un diagramme en boîte qui montre la distribution de la température pour chaque catégorie spécifiée par la variable "label". Les boîtes représentent les quartiles de la distribution, la ligne médiane à l'intérieur de chaque boîte représente la médiane des valeurs, et les moustaches montrent l'étendue des données. C’est dans notre exemple : Nous avons **la température** avec les caractéristiques de label c’est grapes, papaya, orange, pigeonpeas ont besoin de plus de la température.

**Pour humidité :** rice, mungbean, pomegaraante, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, coffee besoin de plus d-humidité que les autres caractéristiques.

**Pour température :** pigeonpeas,grapes ,orange, papaya besoin de plus d’humidité que les autres caractéristiques.

**Pour N :** banana, watermelon, watermelon, muskmelon, cotton, coffee besoin de plus de N que les autres caractéristiques.

**Pour PH :** mothbeans, lentil, orange, cotton, jute, coffee besoin de plus de ph que les autres caractéristiques.

**Pour P :** grapes, apple besoin de plus de p que les autres caractéristiques.

**Pour K :** grapes, apple besoin de plus de K que les autres caractéristiques.

**Partie ACP :**

**Etape 1 : Choix de nombre d’axe :**

D’abord une variable qualitative c’est label et l’ACP doit avoir juste les variables quantitatives donc on fait quali.Sub

On choisit le nombre des axes à partir des valeurs propres qui ont >1 alors on a 5 axes

**Relancer l’ACP avec 5 axes :**

**Etape 2 : Interprétation des axes :**

Pour les individus on choisit qui ont > 1/n = 1/2200 = 0,0004 =0,04%

Pour les variables on choisit qui ont > 1/p = 1/8 =0,125=12,5%

**1er Axe :**

**Les individus :**

1212,1215, 1217, 1296, 1226, 1265, 1589, 1235, 1264, 1294, 1300, 1283, 1291, 1271, 1530, 1255, 1289, 1218, 1243, 1232, 1209, 1208, 1292, 1544, 1214, 1234, 1523, 1556, 1237, 1273, 1529, 1259, 1228, 1588, 1549, 1295, 1583, 1574, 1520, 1285, 1207, 1535, 1225, 1251, 1525, 1240, 1570, 1581, 1241, 1533, 1256, 1515, 1514, 1567, 1552, 1526, 1507, 1244, 1502, 1275, 1573, 1586, 1534, 1277, 1569, 1575, 1564, 1527, 1298, 1579, 1516, 1600, 1528, 1548, 1508, 1524, 1594, 1577, 1287, 1545, 1592, 1555, 1253, 1284, 1205, 1536, 1506, 1572, 1220, 1202, 1576, 1262, 1236, 1565, 1247, 1509, 1557, 1597, 1518, 1543, 1542, 1587, 1505, 1510, 1546, 1252, 1559, 1286, 1582, 1566, 1560, 1257, 1250, 1222, 1595, 1227, 1547, 1551, 1501.

**Les variables :**

On a P et K.

**2éme Axe :**

**Les individus :**

234, 243, 223, 210, 241, 254, 296, 214, 221, 260, 235, 219, 297, 272, 233, 315, 290, 293, 288, 376, 256, 289, 246, 270, 212, 222, 244, 280, 294, 253, 271, 247, 201, 242, 327, 273, 238, 218, 286, 215, 226, 229, 259, 335, 275, 352, 281, 237, 378, 299, 298, 307, 519, 225, 283, 265, 395, 251, 202, 255, 388, 282, 278, 292, 300, 312, 328, 1759, 355, 276, 279, 295, 205, 339, 323, 274, 340, 211, 261, 262, 325, 341, 375, 217, 372, 377, 208, 209, 393, 266, 321, 380, 305, 230, 551, 391, 204, 263, 316, 379, 1751, 231, 250, 318, 287, 329, 338, 301, 330, 344, 284, 236, 264, 216, 356, 267, 1749, 359, 1779, 248, 203, 268, 1248, 302, 239, 324, 232, 228, 353, 400, 1798, 1711, 227, 220, 306, 360, 348, 350, 351, 396, 285, 334, 252, 1288, 1710, 392, 333, 322, 1733, 345, 269, 245, 439, 277, 368, 346, 240, 1745, 1792, 309, 1776, 365, 347, 366, 258, 213, 249, 354, 317, 373, 1211, 1280, 331, 390, 319, 308, 1276, 304, 357, 362, 313, 398, 1701, 206, 257, 374, 383, 311, 524, 361, 332, 1723, 1778, 394, 1744, 364, 537, 596, 224, 508, 585, 1748, 578, 326, 367, 381, 1743, 1705, 382, 55, 1263, 1799, 349, 1762, 384, 389, 320, 370, 387, 314, 1270, 358, 598, 1704, 303, 569, 1221, 207, 13, 1266, 363, 343, 399, 1716, 1230, 547, 573, 483, 336, 1753, 1281, 371, 385, 1775, 337, 1729, 310, 1785, 397, 1788, 418, 568, 11, 1260, 342, 1279, 1796, 802, 1706, 877, 1229, 560, 521, 1204, 37, 291, 27, 38, 1730, 540, 1707, 514, 1702, 1261, 548, 1738, 1787, 1727, 591, 847, 541, 1767, 1715, 1210, 594, 1725, 1206, 1231, 1233, 20, 369, 1772, 512, 526, 1770, 822, 1562, 595, 386, 505, 1519, 809, 841, 1239, 58, 44, 1596, 34, 22, 62, 599, 552, 863, 124

**Les variables :**

Température, humidité.

**3éme Axe :**

**Etape 3 : La qualité**

**Sur le plan (1 :2)**

**Les individus :**

Nous avons les individus suivant :1592, 1552, 1549, 1577, 1535, 1505, 1501, 1583, 1543, 1516, 1531, 1574, 1542, 1553, 1518, 1597, 1507, 1565, 1530, 1546, 1523, 1536, 1525, 1522, 1506, 1556, 1589, 1547, 1529, 1576, 1557, 1590, 1567, 1524, 1520, 1533, 1514, 1588, 1564, 1539, 1550, 1515, 1551, 1519, 1528, 1548, 1526, 1596, 1584, 1562, 1591, 1581, 1594, 1511, 1237, 1575, 1582, 1561, 1586, 1510, 1578, 1205, 1585, 1579, 1560, 1234, 1502, 1262, 1545, 1563, 1512, 1537, 1569, 1570, 1555, 1587, 1558, 1554, 1544, 1508, 1580, 1259, 1566, 1593, 1532, 1600, 1521, 1541, 1242, 1527, 1286, 1517, 1595, 1225, 1572, 1568, 1275, 1503, 1272, 1268, 1573, 1244, 1219, 1222, 1277, 1534, 1249, 1559, 1258, 1287, 1283, 1232, 1598, 1243, 1513, 1253, 1227, 1214, 1271, 1571, 1223, 1282, 1218, 1509, 1293, 1599, 1247, 1238, 1298, 1540 de bon qualité et bien représenté car leur seuil est > 0,9.

**Les variables :**

On a P, K, humidité de bon qualité et bien représenté car leur seuil est >0,71.

**Sur le plan (1 :3)**

**Les individus :**

Nous avons les individus suivant :1592, 1552, 1549, 1577, 1535, 1505, 1501, 1583, 1543, 1516, 1531, 1574, 1542, 1553, 1518, 1597, 1507, 1565, 1530, 1546, 1523, 1536, 1525, 1522, 1506, 1556, 1589, 1547, 1529, 1576, 1557, 1590, 1567, 1524, 1520, 1533, 1514, 1588, 1564, 1539, 1550, 1515, 1551, 1519, 1528, 1548, 1526, 1596, 1584, 1562, 1591, 1581, 1594, 1511, 1237, 1575, 1582, 1561, 1586, 1510, 1578, 1205, 1585, 1579, 1560, 1234, 1502, 1262, 1545, 1563, 1512, 1537, 1569, 1570, 1555, 1587, 1558, 1554, 1544, 1508, 1580, 1259, 1566, 1593, 1532, 1600, 1521, 1541, 1242, 1527, 1286, 1517, 1595, 1225, 1572, 1568, 1275, 1503, 1272, 1268, 1573, 1244, 1219, 1222, 1277, 1534, 1249, 1559, 1258, 1287, 1283, 1232, 1598, 1243, 1513, 1253, 1227, 1214, 1271, 1571, 1223, 1282, 1218, 1509, 1293, 1599, 1247, 1238, 1298, 1540, 1274, 1284, 1278 de bon qualité et bien représenté car leur seuil >0,9.

**Les variables :**

On a P, K de bon qualité et bien représenté car leur seuil est >0,77.

**Sur le plan (1 :4)**

**Les individus :**

Nous avons les individus suivant : 1256, 1237, 1207, 1273, 1289, 1214, 1259, 1277, 1234, 1243, 1291, 1257, 1271, 1217, 1218, 1235, 1258, 1226, 1236, 1225, 1241, 1282, 1232, 1285, 1283, 1290, 1222, 1530, 1265, 1255, 1296, 1253, 1219, 1212, 1294, 1244, 1205, 1589, 1299, 1529, 1264, 1295, 1262, 1523, 1284, 1300, 1250, 1274, 1286, 1240, 1556, 1544, 1515, 1272, 1251, 1209 de bien qualité et bien représenté car leur seuil >0,9

**Les variables :**

On a P, K de bon qualité et bien représenté car leur seuil >0,77.

**Sur le plan (1 :5)**

**Les individus :**

1237, 164, 1234, 1225, 1244, 1205, 1259, 1530, 1589, 1570, 1277, 1523, 1253, 1529, 1283, 1544,

1262, 1600, 1243, 1232, 1572, 1258, 1218, 1271, 1219, 1555, 1534, 1222, 1520, 1515, 1214, 1556,

1594, 1287, 1573, 1272, 1588, 1227, 1502, 1579, 1533, 1551

**Les variables :**

On a P, K de bon qualité et bien représenté car leur seuil >0,74.

**Etape 4 : Représentation graphique**

**Représentation graphique pour les individus :**

**1er groupe :** 1277, 1259, 1258, 1253, 1214, 1243, 1271, 1283, 1232, 1237, 1225, 1283, 1259.

**2éme groupe :** 1592, 1552, 1549, 1577, 1535, 1505, 1501, 1583, 1543, 1516, 1531, 1574, 1542, 1553, 1518, 1597, 1507, 1565, 1530, 1546, 1523, 1536, 1525, 1522, 1506, 1556, 1589, 1547, 1529, 1576, 1557, 1590, 1567, 1524, 1520, 1533, 1514, 1588, 1564, 1539, 1550, 1515, 1551, 1519, 1528, 1548, 1526, 1596, 1584, 1562, 1591, 1581, 1594, 1511, 1237, 1575, 1582, 1561, 1586, 1510, 1578, 1205, 1585, 1579, 1560, 1234, 1502, 1262, 1545, 1563, 1512, 1537, 1569, 1570, 1555, 1587, 1558, 1554, 1544, 1508, 1580, 1259, 1566, 1593, 1532, 1600, 1521, 1541, 1242, 1527, 1286, 1517, 1595, 1225, 1572, 1568, 1275, 1503, 1272, 1268, 1573, 1244, 1219, 1222, 1277, 1534, 1249, 1559, 1258, 1287, 1283, 1232, 1598, 1243, 1513, 1253, 1227, 1214, 1271, 1571, 1223, 1282, 1218, 1509, 1293, 1599, 1247, 1238, 1298, 1540.

**Représentation graphique pour les individus :**

K et P sont très corrélées positivement car l’ongle est 0 (R proche de 1).

K et P sont proche de cercle alors sont bien représenté.

Puisque les deux variables sont proche de cercle alors sont intéressante.

Il y’a les 2dimensions Dim1 a 18,5% et Dim2 a 27,6% alors la somme est 36,7% pour les deux dimensions.

**Etape 5 : Synthèse**

1er groupe : humidité, N, température avec 1759, 1711, 413, 170, 2087, 1049, 204, 417, 170, 727, 2004, 1438, 1626, 188, 1626, 216,90, 98, 392, 175, 11, 392.

2éme groupe : K et P avec 1288, 1248, 1276, 1221, 1288, 1245, 1589, 121, 1215, 1259, 121, 1267, 1268, 230, 1267.