**Notes de présentation projet 9 – 21/04/2023**

**Introduction :**

Bonjour, je m’appelle Naziha, je suis data analyst au sein de l’entreprise « La Poule qui chante » qui est une entreprise agroalimentaire spécialisée dans la vente en gros de poulets prêts à la consommation.

2ème diapo : Notre entreprise souhaite étendre ses ventes à l’international. Dans ce contexte, nous avons mené une étude de marché pour répondre à la problématique : « Quels sont les pays qui rassemblent les meilleures conditions pour acheter nos volailles ? ».

3ème Ainsi, ma mission avait comme objectif de proposer une première analyse des groupements de pays que l’on peut cibler pour exporter nos poulets.

4ème diapo : Nous avons mené à bien cette étude de marché selon **4 grandes étapes** :

1. **D’abord,** lors du kick-off meeting, nous avons échangé et validé les besoins avec les différents métiers selon les critères de l’analyse PESTEL
2. **Ensuite**, nous avons importé nos jeux de données issus du site de la FAO pour pré-traitement, nettoyage et sélection de nos **« features »**
3. **Troisièmement,** nous avons mené une analyse multivariée à travers 2 méthodes de clustering et une analyse en composantes principales
4. **Et enfin, nous en avons extraits les premiers résultats**

5ème diapo : Pour ce faire, j’ai utilisé d’une part Visual Studio Code pour accueillir nos notebooks et d’autre part les bibliothèques de Python : Pandas et Numpy pour l’exploration, le nettoyage et les calculs, Matplotlib, seaborn et plotly pour la visualisation de nos résultats, et scikit learn qui est la bibliothèque moderne et la plus utilisée du Machine Learning pour l’analyse multidimensionnelle.

6ème diapo : Les critères PESTEL retenus pour cette première analyse, sont des critères politiques, économiques et sociaux. Nous avons alimenté notre jeu de données à partir du site de la FAO.

* + La disponibilité alimentaire et le PIB se rapportent aux critères économiques
  + La Sous-nutrition et la Stabilité politique aux critères politiques
  + Et la Population au critère social

7ème diapo :

* Nous avons d’abord importé nos données pour lecture :
  + Nous avons ciblé notre analyse sur l’année 2017 et le produit « viande fraîche ou surgelée de volaille prêt à la consommation »
  + Nous avons utilisé la méthode « Pivot\_table » pour conserver nos pays en index et faciliter la sélection de nos features.
* A partir du dataset sous-nutrition, nous avons supprimé toutes les valeurs extrêmes grâce au calcul de l’écart interquartile et du z-score.
* Nous avons fusionné nos jeux de données pour :
  + Traiter les valeurs manquantes :
    - Suppression des valeurs non renseignées des exportations,
    - Combler les valeurs manquantes renseignées sur le site de la FAO
  + Et faire les analyses univariée et bivariée
  + Nous obtenons un dataframe de 138 observations et 9 colonnes
* Nous avons créé des variables synthétiques telles que :
  + Le TAS qui correspond au taux d’autosuffisance d’un pays
  + Le TDI qui est le taux de dépendance des importations
  + Et la consommation moyenne de poulet par an et par habitant
* Enfin, nous avons sélectionné nos features qui sont au nombre de 7 :
  + TAS
  + TDI
  + Consommation moyenne de poulet
  + Population
  + PIB/hab.
  + Indice de stabilité politique
  + Exportations
* Enfin, avec IsolationForest, nous détectons et éliminons les anomalies à hauteur de 10% de contamination totale.
* Nous obtenons un dataframe final de **105 observations et 7 variables**

**8ème diapo :** Nous pouvons ensuite passer à l’analyse exploratoire multivariée. Pour ce faire, nous utilisons les packages de machine Learning scikit learn. **Dans notre cas, nous avons mené notre analyse exploratoire multidimensionnelle sur un apprentissage non supervisé.**

* *[Les algorithmes de machine Learning sont conçus pour déduire à partir des données les choix les plus optimaux à faire.* ***Les algorithmes s’appuient sur les données et un ensemble de méthodes statistiques pour apprendre quelles features/caractéristiques sont importantes dans les données***
* *Il existe* ***2 types de Machine Learning****: l’apprentissage supervisé et l’apprentissage non supervisé : l’apprentissage supervisé s’applique sur des données étiquetées et l’apprentissage non supervisé s’applique sur des données non étiquetées et permet de découvrir des patterns.*
* *L’analyse multivariée ou multidimensionnelle permet d’analyser rapidement un échantillon, en réduisant la dimension du nombre d'individus ou de variables.]*

**9ème diapo** : la première méthode exploratoire que nous avons appliquée est l’algorithme du K-Means. C’est une méthode de partitionnement qui a pour objectif de regrouper nos individus en groupes homogènes. Il travaille avec les centroïdes (ou centres de gravité des groupes). Une fois les centroïdes placés, on prend chaque point du nuage et on lui **associe le cluster du centroïde** dont il est le plus proche.

Une fois cette opération faite, on **calcule le centre de gravité** de chaque groupe de façon itérative jusqu'à la **convergence de l'algorithme**.

Pour déterminer le nombre de clusters à étudier, Nous avons 2 approches distinctes :

* + La méthode du "coude" qui calcul l’inertie intraclasse, c’est-à-dire la somme des distances au carré de chaque point par rapport à son centroïde. On cherche une "cassure" dans la courbe liant l’inertie intraclasse au nombre de clusters.
  + L’autre méthode, est le coefficient de silhouette moyen. Il est basé sur la distance moyenne entre le point et les autres points dans le même cluster (a) et la distance moyenne entre le point et les points dans le cluster le plus proche (b).
* Dans les 2 cas, nous obtenons un k optimal de 3 !

10ème diapo : Les clusters obtenus par le K-Means laissent apparaitre 3 types de pays.

* Les pays du Cluster 0 ont un TAS très élevé, TDI le plus bas, stabilité politique et économique très basse voir négative, une densité de population la plus élevée
* Les pays du Cluster 1 sont les plus grands exportateurs de viande de volaille, le PIB/habitant est le plus élevé ainsi que les indicateurs de stabilité politique
* Les pays du Cluster 2 ont un TAS bas, un TDI très élevé, stabilité politique et économique positive, et la consommation de volaille la plus élevée, la densité moyenne de la population est la moins élevée.

11ème diapo :

* + Les pays situés dans le cluster 0, sont composés principalement de pays avec un indice de stabilité négatif et une forte population
  + Les pays composant le cluster 1 sont essentiellement des pays industrialisés, en majorité en Europe. D’ailleurs, l’algorithme du K-Means a placé intelligemment la France dans ce cluster.
  + Et enfin, le cluster 2 est composé essentiellement d’îles et de pays de superficie de petite voir très petite taille, d’où la densité de population la moins élevée, un Taux d’autosuffisance bas, et un TDI très élevé.

12ème diapo : La 2ème méthode de clustering que nous avons appliquée est la Classification ascendante hiérarchique.

* L’approche **ascendante**, considère tout d'abord que chaque point est un cluster. Il y a donc autant de clusters que de points. Ensuite, on cherche les deux clusters les plus proches, et on les agglomère en un seul cluster. On répète cette étape de façon itérative jusqu'à ce que tous les points soient regroupés en un seul grand cluster
* Le dendrogramme est l’arbre qui regroupe tous ces clusters
* La méthode de Ward permet de calculer la distance entre les clusters en minimisant l’augmentation de l’inertie intraclasse.

Les 3 clusters produits par la CAH, nous donne des résultats très proches de ceux obtenus avec l’algorithme du K-Means.

* 13ème diapo : Enfin la 3ème méthode exploratoire multidimensionnelle appliquée sur nos données est l’ACP. Elle permet d’étudier les variabilités des individus, les ressemblances et différences et les liaisons entre les variables pour les regrouper en variables synthétiques (c’est la réduction) afin de supprimer des variables trop corrélées entre elles et pour réduire la dimensionnalité de nos données.
  + On réalise une ACP pour créer des axes principaux qui sont des éléments caractéristiques majeurs des 3 thèmes PESTEL (politique, économique, et social). Elle nous permet de connaître la dynamique des pays à l’intérieur de l’univers restreint des axes. **La Clé de l’ACP sera de rechercher la projection pour laquelle l’inertie des points est maximale**
  + On détermine le nombre d’axes à étudier avec le diagramme d’éboulis qui calcule la variance cumulée captée par chaque axe. Dans notre cas, nous obtiendrons 62% de la variance cumulée expliquée par les 2 premières composantes principales.

14ème diapo : nous avons projeté nos pays sur notre 1er plan factoriel normé.

15ème diapo :

16ème diapo :

17ème diapo :