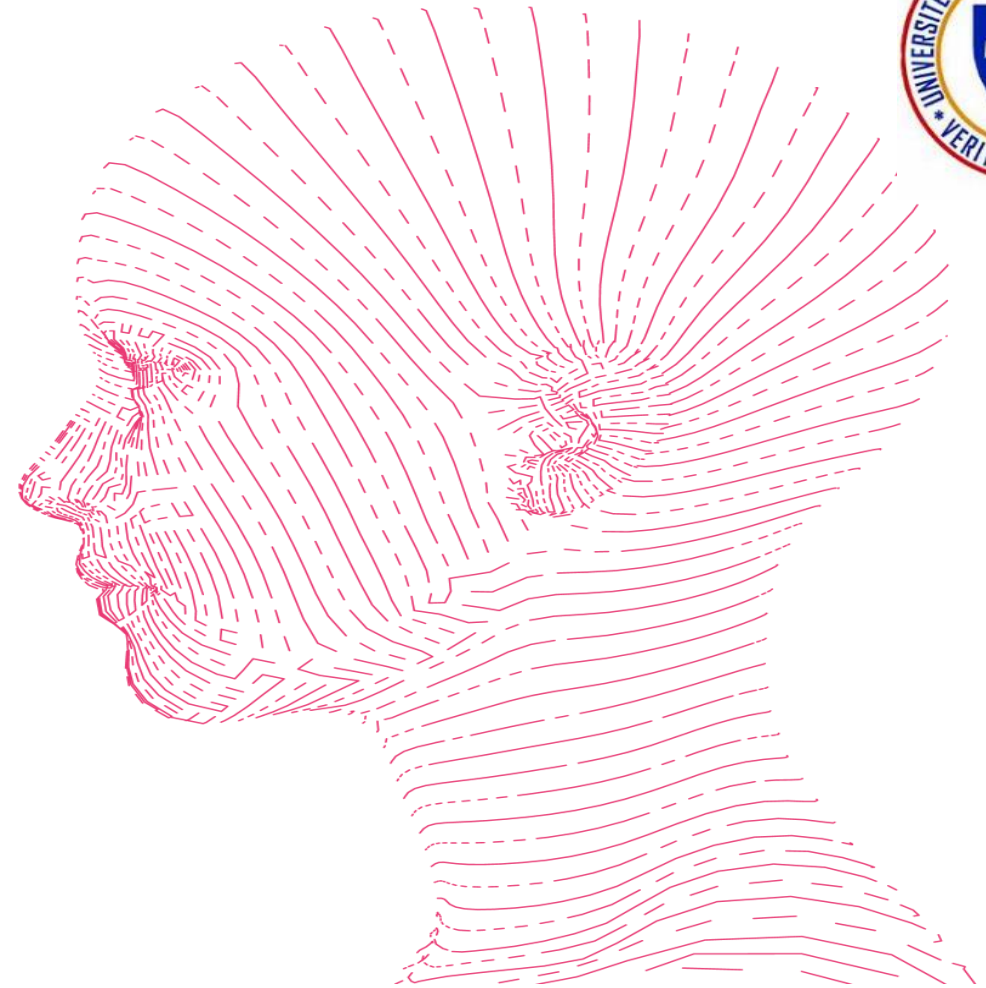


# Machine Learning



- Dispensé par **MWAMBA KASONGO Dahouda**
- Docteur en génie logiciel et systèmes d'information
- Machine and Deep Learning Engineer

➤ Assisté par Ass. **Daniel MBAYA**

- E-mail : [dahouda37@gmail.com](mailto:dahouda37@gmail.com)
- Tel.: **+243 99 66 55 265**

Heure : 8H00 – 12H00

## PLAN DU COURS

### CHAPITRE 0 Les données dans la sciences des données

0.1 Données traditionnelles et Big Data

0.2 Sources de données

0.3 Gestion de données

0.4 Besoin de science de données

0.5 Carriere en science de données

0.6 Cycle de vie en science de données

0.7 Outils et Bibliothèques

0.8 Business Intelligence



## PLAN DU COURS



## CHAPITRE 1 Introduction au Machine Learning

1.1. Qu'est-ce que Machine Learning?

1.2. Définition de Machine Learning

1.3. Différences entre l'IA, Machine Learning et Deep Learning

1.4. Exemples concrets du Machine Learning

1.5. Types de Machine Learning

1.5.1 Apprentissage supervisé

1.5.2 Apprentissage non-supervisé

1.5.3 Apprentissage semi-supervisé

1.5.4 Apprentissage par renforcement (juste une introduction)

1.6. Fonctionnement des algorithmes d'apprentissage automatique

1.7. Cycle de vie de l'apprentissage automatique



## PLAN DU COURS



## CHAPITRE 2 Concepts de base sur le Machine Learning

### 2.1. Terminologies d'apprentissage automatique

- **Features (Caractéristiques), Label (étiquettes) et Dataset (ensembles de données)**

### 2.2 Types de données

- **Catégorielles, Numériques, Textuelles, Images**
- Training set (Données d'entraînement), Validation set (Données de validation), Test set (Données de test)

### 2.3 Introduction aux algorithms

- **Qu'est-ce qu'un algorithme ?**
- **Les algorithms de Machine Learning**

### 2.4 Exemple Pratique de la Régression linéaire simple



## PLAN DU COURS



### CHAPITRE 3 Apprentissage supervisé : Régression et Classification

#### 3.1 Régression linéaire :

- Introduction à la régression linéaire multiple
- Prédire des valeurs continues (par exemple, prédire les prix des maisons)
- Const Function (Fonction de coût) et Mean Squared Error (Erreur quadratique moyenne)

#### 3.2 Classification :

- Régression logistique pour la classification binaire
- Métriques de classification dans le Machine Learning  
Accuracy (Exactitude), Precision (précision), Recall (rappel), F1-Score
- Prédiction de l'approbation des prêts à l'aide de Machine Learning
- Comparaison de performance des algorithmes



## PLAN DU COURS



### CHAPITRE 4 Apprentissage non supervisé : Clustering

#### 4.1. K-Means Clustering Algorithm

#### 4.2. Définition de Clustering k-Means

#### 4.3. Fonctionnement de K-Means Clustering Algorithm

- Comment choisir la valeur de « nombre K de clusters » ?

- Elbow Method

- Étapes de la méthode Elbow

#### 4.4. Implémentation Python de l'algorithme de clustering K-means

- Segmentation de la clientèle avec l'algorithme de clustering K-means



## PLAN DU COURS



### CHAPITRE 5 Projet pratiques

- **Projet 1 : Sur la Régression**
  - **Utiliser un ensemble de données, appliquer une régression linéaire, évaluer le modèle**
- **Projet 2 : Sur la Classification**
  - **Utiliser un ensemble de données, appliquer une régression logistique ou k-NN**



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### QU'EST-CE QUE LES DONNÉES ?

Les données sont une collection de valeurs discrètes qui transmettent des informations, décrivant la quantité, la qualité, les faits, les statistiques, d'autres unités de base de sens ou simplement des séquences de symboles qui peuvent être interprétées plus avant.

### QU'EST-CE QUE LA SCIENCE DES DONNÉES ?

La science des données est un terme qui échappe à toute définition complète unique, ce qui le rend difficile à utiliser, surtout si l'objectif est de l'utiliser correctement. La plupart des articles et publications utilisent l'expression librement, en partant du principe qu'elle est universelle ment comprise. Cependant, la science des données - ses méthodes, ses objectifs et ses applications - évolue avec le temps et la technologie.

Il y a 25 ans, la science des données faisait référence à la collecte et au nettoyage d'ensembles de données, puis à l'application de méthodes statistiques à ces données.

En 2018, la science des données est devenue un domaine qui englobe l'analyse de données, l'analyse prédictive, l'exploration de données, l'intelligence d'affaires, l'apprentissage automatique, l'apprentissage en profondeur et bien plus encore.





## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



La science des données est un domaine qui comprend de nombreux sous-domaines tels que l'intelligence artificielle, l'apprentissage automatique, les statistiques, la visualisation des données et l'analyse, et fournit des exemples et des exercices pratiques pour vous aider à appliquer ces concepts dans le monde réel. Au cours des dernières années, la demande de data scientists a été énorme.

✓ **Pour améliorer l'efficacité de l'entreprise, il devient important d'analyser les données.**

Dans ce chapitre sur la science des données, nous fournirons un aperçu complet des concepts, outils et techniques de base utilisés dans le domaine de la science des données.

La science des données est un domaine qui consiste à extraire des informations et des connaissances à partir de données à l'aide de diverses techniques et outils.

En résumé, l'apprentissage de la science des données implique la programmation, les statistiques, la visualisation de données, l'apprentissage automatique, la pratique, l'apprentissage continu.

Avec du dévouement et des efforts constants, vous pouvez maîtriser la science des données et commencer à élaborer des solutions à des problèmes du monde réel.

Les données sont le fondement de la science des données ; c'est le matériau sur lequel reposent toutes les analyses.

Dans le contexte de la science des données, il existe deux types de données : **les données traditionnelles et les méga données (Big Data).**



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.1 Données Traditionnelles et Big Data

#### 0.1.1 Données traditionnelles

Les données traditionnelles sont des données structurées et stockées dans des bases de données pouvant être gérées à partir d'un seul ordinateur ; il se présente sous la forme d'un tableau contenant des valeurs numériques ou textuelles.

Le terme « données traditionnelles » ne fait pas partie du vernaculaire officiel. C'est quelque chose que nous introduisons pour plus de clarté. Nous pensons que cela aide à souligner la distinction entre les méga données et les non méga données.

#### 0.1.2 Big Data

Le Big Data est plus volumineux que les données traditionnelles, mais pas au sens trivial. Il s'agit de données extrêmement volumineuses, réparties sur un réseau d'ordinateurs, mais elles ne se caractérisent pas seulement par leur volume.

Ces données peuvent être sous différents formats ; il peut être structuré, semi-structuré ou non structuré; et vous verrez souvent des méga données caractérisées par la lettre « V ». Cela découle des « 3V du big data » :

- **Variété** - chiffres, texte, mais aussi images, audio, données mobiles, etc. Les méga données peuvent être dans différents formats.
- **Vélocité** - elle est récupérée et calculée en temps réel
- **Volume** - les données volumineuses sont mesurées en téra-, péta-, exaoctets (c'est-à-dire 1 million de téraoctets).



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.1 Données Traditionnelles et Big Data

Indépendamment du fait qu'un data scientist ne reçoive que des données collectées ou qu'il doive suivre lui-même le processus de collecte, ces données seront au format brut. Ces données peuvent provenir d'enquêtes ou de paradigmes de collecte automatique de données, comme les cookies sur un site Web.

**Les données brutes sont des données intactes qui doivent être converties sous une forme plus compréhensible et utile pour un traitement ultérieur.** Le groupe d'opérations qui effectuent cela s'appelle le **prétraitement (Data Preprocessing)**.

**Le prétraitement** des données peut faire référence à la manipulation ou à la suppression de données avant qu'elles ne soient utilisées afin d'assurer ou d'améliorer les performances, et constitue une étape importante dans le processus d'exploration de données.

Pour les données traditionnelles, le prétraitement consiste en un étiquetage de classe (catégoriel et numérique), un nettoyage des données et un traitement des valeurs manquantes, tandis que pour les méga données, le prétraitement consiste en un étiquetage de classe (numéro, texte, images numériques, données vidéo, données audio numériques), le nettoyage des données et le traitement des valeurs manquantes.

- **Étiqueter les observations par classe** : cela consiste à organiser les données par catégorie ou à étiqueter les points de données selon le type de données (pour les données traditionnelles, cela peut être numérique/catégoriel ; pour les données volumineuses : texte, image, audio).

## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.1 Données Traditionnelles et Big Data

- **Nettoyage des données** : traitement des données incohérentes, telles que les catégories mal orthographiées et les valeurs manquantes.
- **Équilibrage des données** : si les catégories des données contiennent un nombre inégal d'observations, elles peuvent ne pas être représentatives de la population. Les méthodes d'équilibrage, comme l'extraction d'un nombre égal d'observations pour chaque catégorie et leur préparation pour le traitement, résolvent le problème.

### 0.2 Sources des données

Les données traditionnelles peuvent provenir d'enregistrements clients de base ou d'informations historiques sur le cours des actions. Les dossiers clients de base peuvent contenir des informations telles que l'identifiant du client, le nombre de fois qu'il a passé une commande, le montant d'argent dépensé, son adresse, ses coordonnées, etc.

Les big data peuvent provenir de partout; un nombre toujours croissant d'entreprises et d'industries utilisent et génèrent des méga données.

- ✓ Considérez les communautés en ligne, par exemple, Facebook, Google et LinkedIn ; ou des données de trading financier.
- ✓ Les données machine des capteurs des équipements industriels constituent également du Big Data.
- ✓ Et, bien sûr, la technologie portable.

Actuellement, le volume de données numériques s'élève à 3,2 zettaoctets, et 90 % de ces données ont été recueillies au cours des 2 dernières années seulement.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.3 Gestion des données

Les spécialistes des données qui traitent des données brutes et du prétraitement, créent des bases de données et les maintiennent peuvent porter des noms différents. Mais bien que leurs titres soient similaires, il existe des différences palpables dans les rôles qu'ils occupent.

Les architectes de données et les ingénieurs de données (les architectes Big Data et les ingénieurs Big Data, respectivement) sont cruciaux sur le marché de la science des données.

Le premier crée la base de données à partir de zéro ; ils conçoivent la manière dont les données seront récupérées, traitées et consommées. Par conséquent, l'ingénieur de données utilise le travail des architectes de données comme un tremplin, puis traite (prétraite) les données disponibles. Ce sont eux qui s'assurent que les données sont propres, organisées et prêtes à être prises en charge par les analystes.

L'administrateur de la base de données, quant à lui, est la personne qui contrôle le flux de données entrant et sortant de la base de données.

Bien sûr, avec le Big Data, la quasi-totalité de ce processus est automatisée, il n'y a donc pas vraiment besoin d'un administrateur humain.

L'administrateur de la base de données s'occupe principalement des données traditionnelles.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.3 Gestion des données

Si vous êtes débutant en science des données, voici quelques étapes que vous pouvez suivre pour commencer:

- ✓ **Apprendre la programmation**: La programmation est une compétence fondamentale pour la science des données. Python est le langage de programmation le plus couramment utilisé en science des données et il possède plusieurs bibliothèques utiles pour la science des données, telles que **NumPy**, **Pandas** et **Scikit-learn**.
- ✓ **Apprendre les statistiques** : Les statistiques sont le fondement de la science des données. Comprendre les concepts statistiques tels que la moyenne, la médiane, la variance et l'écart type est crucial pour travailler avec des données.
- ✓ **Apprenez la visualisation des données**: La visualisation des données est une compétence essentielle pour la science des données. Cela aide à comprendre les modèles et les tendances des données. Il existe plusieurs bibliothèques en Python utiles pour la visualisation de données, telles que **Matplotlib** et **Seaborn**.
- ✓ **Apprendre l'apprentissage automatique** : L'apprentissage automatique est au cœur de la science des données. Cela implique de créer des modèles capables d'apprendre des données et de faire des prédictions. Il existe plusieurs types d'algorithmes d'apprentissage automatique, tels que l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement (Reference Chapitre 1 Machine Learning).

## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.3 Gestion des données

- ✓ **Pratique avec des projets:** La pratique est essentielle pour apprendre la science des données. Vous pouvez commencer par travailler sur de petits projets tels que le nettoyage des données, l'analyse des données et les machine Learning.
- ✓ **Apprendre de la communauté:** La communauté de la science des données est très active et plusieurs ressources sont disponibles pour apprendre. Vous pouvez rejoindre des communautés en ligne telles que Reddit, LinkedIn ou Twitter. Vous pouvez également assister à des rencontres et événements locaux sur la science des données.
- ✓ **Apprendre en continu:** La science des données est un domaine en évolution rapide et de nouvelles techniques et outils émergent constamment. Par conséquent, il est essentiel de continuer à apprendre et de rester informé des dernières tendances et développements en matière de science des données.

En résumé, l'apprentissage de la science des données implique la programmation, les statistiques, la visualisation de données, l'apprentissage automatique, l'apprentissage continu. Avec du dévouement et des efforts constants, vous pouvez maîtriser la science des données et commencer à élaborer des solutions à des problèmes du monde réel.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.4 Besoin de science des données

Le besoin de science des données a augmenté rapidement dans divers secteurs en raison de la production massive de données et de la demande croissante de prise de décision basée sur les données. Voici quelques raisons clés pour lesquelles la science des données est nécessaire :

- ✓ **Explosion des données** : Avec la prolifération d'Internet, de l'IoT et des appareils numériques, les données sont générées à un rythme sans précédent. La science des données est essentielle pour gérer, analyser et tirer des informations de cette vaste quantité d'informations.
- ✓ **Amélioration de la prise de décision** : La science des données aide les organisations à prendre des décisions meilleures et plus rapides en fournissant des informations et des prévisions approfondies basées sur l'analyse des données. Cela peut être essentiel pour améliorer l'efficacité, améliorer l'expérience client et obtenir un avantage concurrentiel.
- ✓ **Automatisation** : Avec l'apprentissage automatique et l'intelligence artificielle, la science des données permet l'automatisation des processus, l'analyse prédictive et même des systèmes intelligents qui apprennent et s'adaptent au fil du temps.
- ✓ **Personnalisation améliorée** : Dans des secteurs comme le marketing, le commerce électronique, la science des données permet une expérience personnalisée en analysant le comportement, les préférences et les tendances de chacun pour personnaliser les produits et services.
- ✓ **Gestion des risques** : La science des données est utilisée pour identifier, évaluer et atténuer les risques dans des secteurs tels que la finance, la santé, l'assurance et la cybersécurité en analysant les tendances, en prédisant les défaillances potentielles et en détectant les anomalies.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.5 Carrières de science des données

La science des données est aujourd'hui considérée comme l'un des emplois les plus recherchés dans le domaine informatique. Les opportunités de croissance dans les emplois en science des données sont comparativement élevées que dans tout autre emploi. Les entreprises se concentrent désormais davantage sur les emplois en science des données pour élever leurs objectifs commerciaux, ce qui a également créé un flot d'emplois en science des données sur le marché.

### 0.6 Cycle de vie de la science des données

Le cycle de vie de la science des données fait référence au processus structuré utilisé pour guider les projets de science des données du début à la fin.

Il comprend généralement plusieurs étapes clés, chacune étant essentielle à la bonne exécution et à la fourniture d'informations.

Vous trouverez ci-dessous un aperçu des phases courantes du cycle de vie de la science des données:

1. Définition du problème
2. Collecte de données
3. Nettoyage et préparation des données
4. Exploration des données (EDA)
5. Modélisation des données
6. Évaluation du modèle
7. Déploiement du modèle
8. Surveillance et maintenance du modèle
9. Rapports et communication



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.6 Cycle de vie de la science des données

#### 0.6.1. Définition du problème

**Objectif** : Définir clairement le problème ou la question commerciale à résoudre.

**Activités** : Comprendre les objectifs, identifier la portée et formuler les questions auxquelles il faut répondre à l'aide de données.

**Résultats clés** : énoncé du problème bien défini, objectifs clairs, indicateurs de réussite.

#### 0.6.2. Collecte de données

**Objectif** : Rassembler toutes les données pertinentes nécessaires à l'analyse.

**Activités** : Identifier les sources de données, collecter les données brutes et s'assurer qu'elles sont correctement enregistrées.

**Résultats clés** : Données brutes provenant de diverses sources (bases de données, API, scraping Web, etc.).

#### 0.6.3. Nettoyage et préparation des données

**Objectif** : Nettoyer et transformer les données brutes pour les rendre utilisables pour l'analyse.

**Activités** : Gérer les données manquantes, supprimer les doublons, corriger les incohérences, effectuer la détection des valeurs aberrantes et formater les données de manière appropriée.

**Résultats clés** : Données nettoyées et prétraitées, prêtes à être explorées et analysées.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.6 Cycle de vie de la science des données

#### 0.6.4. Exploration des données (Analyse Exploratoire des Données)

**Objectif** : Comprendre les données et découvrir des modèles ou des informations.

**Activités** : Utiliser des méthodes statistiques et des outils de visualisation pour analyser la structure et les caractéristiques des données, vérifier les distributions, les corrélations et les tendances.

**Résultats clés** : Résumés de données, informations initiales, visualisations (graphiques, tracés)

#### 0.6.5. Modélisation des données

**Objectif** : Appliquer les algorithmes de machine Learning (apprentissage automatique) pour faire des prédictions ou des classifications.

**Activités** : Sélectionner des algorithmes appropriés (Régression, Classification, Clustering, etc.), entraîner le modèle sur les données et Affiner ou ajuster (Tuning) les hyperparamètres.

**Résultats clés** : Modèles entraînés, prédictions et informations issues des modèles.

#### 0.6.6. Évaluation du modèle

**Objectif** : Evaluer les performances du modèle.

**Activités** : Utiliser des mesures d'évaluation (exactitude, précision, rappel, score F1, AUC, etc.) pour vérifier les performances du modèle sur les données d'entraînement et de test. Les techniques de validation croisée sont également couramment utilisées.

**Résultats clés** : Rapports sur les performances du modèle, domaines potentiels d'amélioration.



## CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



### 0.6 Cycle de vie de la science des données

#### 0.6.7. Déploiement du modèle

**Objectif** : Intégrer le modèle dans un environnement réel pour l'utiliser.

**Activités** : Déployer le modèle en production, que ce soit dans le cadre d'une application Web, ou intégré dans des systèmes d'entreprise.

**Résultats clés** : Modèles déployés pouvant être utilisés pour les prédictions ou la prise de décision.

#### 0.6.8. Surveillance et maintenance du modèle

**Objectif** : S'assurer que le modèle continue de bien fonctionner au fil du temps.

**Activités** : Surveiller les performances du modèle, détecter les dérives (lorsque la précision du modèle se dégrade en raison de changements dans les modèles de données), mettre à jour ou recycler le modèle si nécessaire.

**Résultats clés** : Rapports sur les performances du modèle au fil du temps, ajustements ou modèles recyclés.

#### 0.6.9. Rapports et communication des résultats

**Objectif** : Présenter les informations, les prédictions ou les recommandations aux parties prenantes ou chef d'entreprise.

**Activités** : Créez des tableaux de bord, des rapports ou des présentations qui expliquent les résultats de manière claire et exploitable.

**Résultats clés** : Visualisations de données, rapports d'activité et informations exploitables (Ex. Power BI)

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.7 Outils et Bibliothèques

### ❑ Principes de base de Python

#### - Python pour la science des données

### ❑ Bibliothèques pour Machine Learning

- **Numpy** (tableaux),
- **Pandas** (manipulation de données)
- **Scikit-Learn** pour la création de modèles simples
- **Matplotlib** et **Seaborn** (visualization de données)

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### ➤ Qu'est-ce que l'analyse des données

L'analyse est un processus consistant à répondre « **Comment?** » et **pourquoi?** ».

- ✓ L'analyse des données est le processus d'**examen**, de **nettoyage**, de **transformation** et d'**interprétation** des données dans le but de découvrir des informations utiles, d'éclairer les conclusions et de soutenir la prise de décision.
  - Les données sont partout, dans les fiches, sur les plateformes de réseaux sociaux, dans les commentaires sur les produits, partout.
  - À l'ère de l'information, celle-ci est créée à une très grande vitesse et, lorsque les données sont analysées correctement, elles peuvent constituer l'atout le plus précieux d'une entreprise.
  - Pour développer votre entreprise et même grandir dans votre vie, **il suffit parfois d'analyser!**  
Si votre entreprise ne se développe pas, vous devez alors regarder en arrière, reconnaître vos erreurs et refaire un plan sans répéter ces erreurs.  
Et même si votre entreprise se développe, vous devez alors espérer la faire croître davantage.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### 0.8.1 Qu'est-ce que l'analyse des données

- ❑ L'analyse des données peut être effectuée à l'aide de diverses méthodes, outils et technologies, notamment:
  - ✓ L'analyse statistique,
  - ✓ L'apprentissage automatique (Machine Learning),
  - ✓ L'exploration de données et
  - ✓ Les techniques de visualisation (Streamlit et Power BI)

Elle est utilisée dans divers domaines et secteurs pour **extraire des informations précieuses** à partir des données, **éclairer les décisions stratégiques**, améliorer les processus et générer des résultats commerciaux.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### 0.8.2 Types de méthodes d'analyse des données

Il existe différents types de méthodes d'analyse de données, chacune adaptée à différents objectifs, types de données et contextes. Certains des types courants de méthodes d'analyse de données comprennent:

#### 1. Analyse descriptive

Une analyse descriptive examine les données et analyse les **événements passés** pour déterminer comment aborder les **événements futurs**.

Elle examine les performances **passées** et comprend les performances en exploitant des **données historiques** pour comprendre la cause du succès ou de l'échec dans le passé.

Presque tous les rapports de gestion tels que les ventes, le marketing, les opérations et les finances utilisent ce type d'analyse.



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### 0.8.2 Types de méthodes d'analyse des données

#### 2. Analyse diagnostique

L'analyse diagnostique fonctionne de pair avec l'analyse descriptive.

Alors que l'analyse descriptive découvre **ce qui s'est passé dans le passé**, l'analyse diagnostique, quant à elle, découvre **pourquoi cela s'est produit** ou **quelles mesures ont été prises à ce moment-là**, ou **à quelle fréquence cela s'est produit**.

Elle donne essentiellement une explication détaillée d'un scénario particulier en comprenant les modèles de comportement.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### 0.8.2 Types de méthodes d'analyse des données

#### 3. Analyse prédictive

Les informations que nous avons reçues de l'analyse descriptive et diagnostique, nous pouvons utiliser ces informations pour **prédire les données futures**.

Elle découvre essentiellement ce qui est susceptible de se **produire dans le futur**.

Désormais, lorsque les données futures ne signifient pas que nous sommes devenus des devins, en examinant les tendances et les comportements passés, nous prévoyons que cela pourrait se produire dans le futur.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.8 Analyse de données

### 0.8.2 Types de méthodes d'analyse des données

#### 4. Analyse prescriptive

Il s'agit d'une méthode avancée d'analyse prédictive.

Maintenant, lorsque vous prédisiez quelque chose, vous aurez certainement de nombreuses options, et nous ne savons alors plus quelle option fonctionnera réellement.

L'analyse prescriptive aide à déterminer quelle est **la meilleure option** pour y parvenir ou **fonctionner**.

L'analyse prédictive **prévoit les données futures**, l'analyse prescriptive, en revanche, **contribue à réaliser tout ce que nous avons prévu**.

L'analyse prescriptive est le niveau d'analyse le plus élevé utilisé pour choisir la meilleure solution optimale en examinant des données descriptives, diagnostiques et prédictives.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.1 La science des données explique le passé

Il existe également deux manières d'examiner les données : **dans le but d'expliquer un comportement qui s'est déjà produit** et pour lequel vous avez collecté des données, ou **pour utiliser les données dont vous disposez déjà pour prédire un comportement futur** qui ne s'est pas encore produit.

❑ Avant que la science des données ne se lance dans l'analyse prédictive, elle doit examiner les modèles de comportement fournis par le passé, les analyser pour en tirer des enseignements et éclairer la voie à suivre pour les prévisions.

➤ BI signifie généralement **Business Intelligence**.

Il fait référence aux stratégies et technologies utilisées par les entreprises pour **l'analyse des données commerciales**.

Les technologies BI fournissent des **vues historiques, actuelles et prédictives des opérations commerciales**.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.1 La science des données explique le passé

❑ Les fonctions courantes des technologies de business intelligence comprennent le **reporting**, le **traitement analytique en ligne**, l'**analyse**, l'**exploration de données**, l'**exploration de processus**, le **traitement d'événements complexes**, la **gestion des performances commerciales**, l'**analyse comparative**, l'**exploration de texte**, l'**analyse prédictive** et l'**analyse prescriptive**.

La BI se concentre pour fournir des réponses basées sur les données à des questions telles que :

- ✓ Combien d'unités ont été vendues ?
- ✓ Dans quelle région le plus de biens ont-ils été vendus ?
- ✓ Quels types de biens sont vendus où ?
- ✓ Comment le marketing par e-mail s'est-il comporté au dernier trimestre en termes de taux de clics et de revenus générés ?
- ✓ Comment cela se compare-t-il à la performance du même trimestre de l'année dernière ?

Même si la Business Intelligence n'a pas « data science » dans son titre, elle en fait partie, et ce n'est pas dans un sens trivial.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.2 Que fait un analyste du BI

Bien entendu, la science des données peut être appliquée pour mesurer les performances des entreprises. Mais pour que l'analyste du BI puisse y parvenir, il doit employer des techniques spécifiques de traitement des données.

**Le point de départ de toute science des données, ce sont les données.**

Une fois les données pertinentes entre les mains de l'Analyste BI (revenu mensuel, client, volume des ventes, etc.), il doit :

- ✓ quantifier les observations
- ✓ calculer les KPI (**Key Performance Indicator**),
- ✓ Examiner les mesures pour extraire des informations de leurs données.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.2 Que fait un analyste du BI

- La science des données consiste à raconter une histoire.
- Outre le traitement d'informations strictement numériques, la science des données, et **plus particulièrement la BI**, consiste à **visualiser les résultats** et à **créer des images faciles à digérer**, appuyées uniquement par les chiffres les plus pertinents.

Tous les niveaux de direction doivent être capables de **comprendre les enseignements des données** et **d'éclairer leur prise de décision**.

- Les analystes BI créent des **tableaux de bord** et des **rapports**, accompagnés de **graphiques**, **diagrammes**, **cartes** et autres visualisations comparables, pour **présenter les résultats pertinents** par rapport aux objectifs commerciaux actuels.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.3 Ou est le BI est utilisé

#### ➤ Optimisation de prix

La science des données est notamment appliquée pour éclairer des éléments tels que les techniques d'optimisation des prix.

#### ✓ Comment ça marche?

**Avec BI :** Les informations pertinentes sont extraites en temps réel et comparées aux informations historiques, et des actions sont prises en conséquence.

**Exemple:** Considérez le comportement de la direction de l'hôtel : les prix des chambres sont augmentés lorsque de nombreuses personnes souhaitent visiter l'hôtel et réduits lorsque l'objectif est d'attirer des visiteurs dans des périodes de faible demande.



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.3 Ou est le BI est utilisé

#### ➤ Gestion de l'inventaire

La science des données et la BI sont inestimables pour gérer les excédents et les sous-approvisionnements.

#### Comment ça marche?

Des analyses approfondies des transactions de vente passées identifient les modèles de saisonnalité et les périodes de l'année où les ventes sont les plus élevées, ce qui aboutit à la mise en œuvre de techniques de gestion des stocks efficaces qui répondent aux demandes à un coût minimum.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.4 Roles dans la BI

#### ☐ Un Analyste en Intelligence d'Affaires [Business Intelligence Analyst]

✓ Un analyste BI se concentre principalement sur **les analyses** et le **reporting** des données historiques passées.

#### ☐ Un Développeur en Intelligence d'Affaires [Business Intelligence Developer]

✓ Le développeur BI est la personne qui **manipule des outils de programmation plus avancés**, tels que Python et **SQL**, pour créer des analyses spécifiquement conçues pour l'entreprise.

**Il s'agit du poste le plus fréquemment rencontré dans l'équipe BI.**

#### ☐ Un Consultant en Intelligence d'Affaires [Business Intelligence Consultant]

✓ Le consultant BI n'est souvent qu'un « analyste BI externe ».

Les consultants BI seraient des analystes BI s'ils avaient été employés. L'analyste BI a des compétences et connaissances spécialisées, le consultant BI contribue à l'étendue de l'équipe data science.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

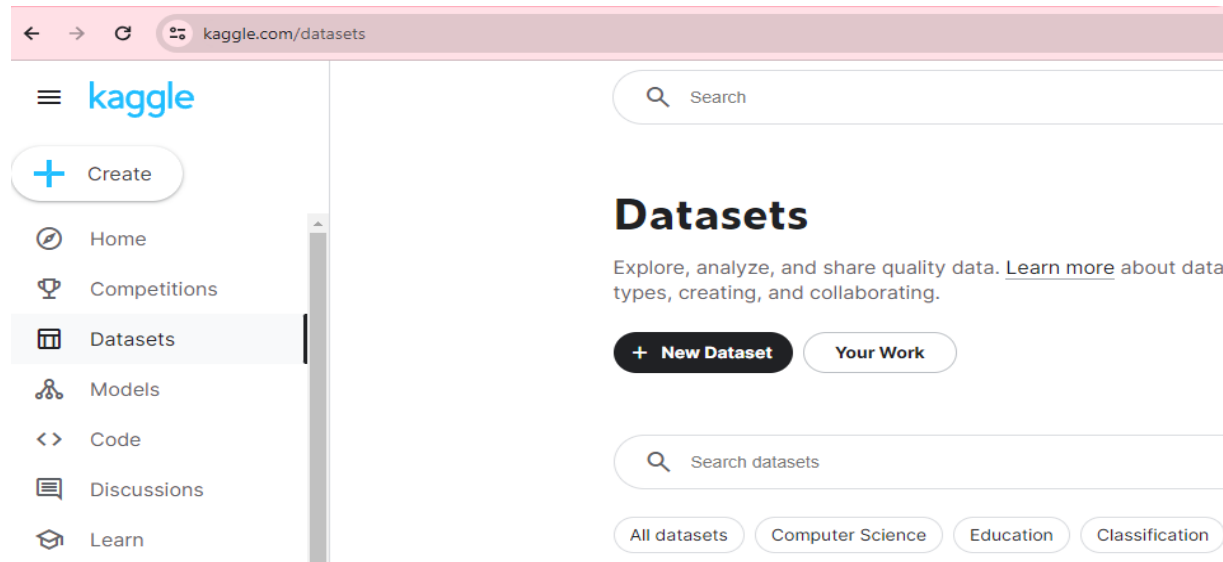


## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

Il existe de nombreuses sources où vous pouvez obtenir des données pour l'analyse de la science des données, en fonction de vos intérêts spécifiques et du type d'analyse que vous souhaitez effectuer. Voici quelques options populaires :

**1. Kaggle :** Kaggle est une plate-forme de concours de science des données, mais elle héberge également un grand nombre d'ensembles de données disponibles gratuitement pour l'exploration et l'analyse. Lien du Kaggle : <https://www.kaggle.com/datasets>



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

**2. Référentiel UCI Machine Learning :** le référentiel UCI Machine Learning est un ensemble de bases de données, de théories de domaine et de générateurs de données largement utilisés par la communauté de Machine Learning. Lien de UCI : <https://archive.ics.uci.edu/>


[Datasets](#) [Contribute Dataset](#) [About Us](#)

### Welcome to the UC Irvine Machine Learning Repository

We currently maintain 664 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!

[VIEW DATASETS](#) [CONTRIBUTE A DATASET](#)


#### Popular Datasets



**Iris**

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for ev...

[Classification](#) [150 Instances](#) [4 Features](#)




**Dry Bean**

Images of 13,611 grains of 7 different registered dry beans were taken with a high-resol...

[Classification](#) [13.61K Instances](#) [16 Features](#)


#### New Datasets



**PhiUSIIL Phishing URL (Website)**

PhiUSIIL Phishing URL Dataset is a substantial dataset comprising 134,850 legitimate an...

[Classification](#) [235.8K Instances](#) [54 Features](#)



**RT-IoT2022**

The RT-IoT2022, a proprietary dataset derived from a real-time IoT infrastructure, is intro...

[Classification, Regressi...](#) [123.12K Instances](#) [84 Features](#)

Cours-ML-Dahouda M., Ph.D.

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

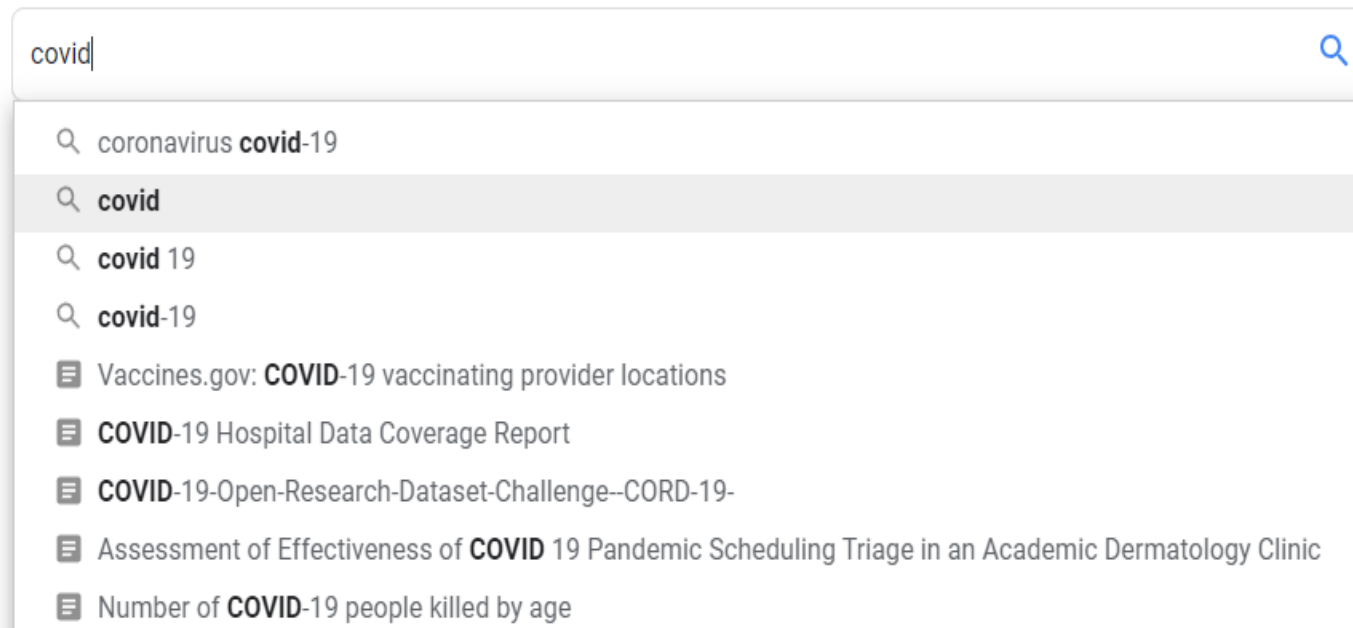
## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

3. Recherche d'ensembles de données Google : Google Dataset Search vous aide à trouver des ensembles de données stockés sur le Web.

C'est un outil utile pour découvrir des ensembles de données provenant de diverses sources : <https://datasetsearch.research.google.com/>

## Dataset Search



The screenshot shows the Google Dataset Search interface. The search bar at the top contains the text 'covid'. Below the search bar, a list of search results is displayed. The first result is 'coronavirus covid-19'. The second result is 'covid', which is highlighted. The third result is 'covid 19'. The fourth result is 'covid-19'. Below these are several results with document icons, including 'Vaccines.gov: COVID-19 vaccinating provider locations', 'COVID-19 Hospital Data Coverage Report', 'COVID-19-Open-Research-Dataset-Challenge--CORD-19-', 'Assessment of Effectiveness of COVID 19 Pandemic Scheduling Triage in an Academic Dermatology Clinic', and 'Number of COVID-19 people killed by age'.



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

4. **Portails de données ouvertes gouvernementaux** : de nombreux gouvernements proposent des portails de données ouvertes où vous pouvez trouver des ensembles de données liés à la démographie, à l'économie, à la santé, etc. Lien : <https://data.gov/>



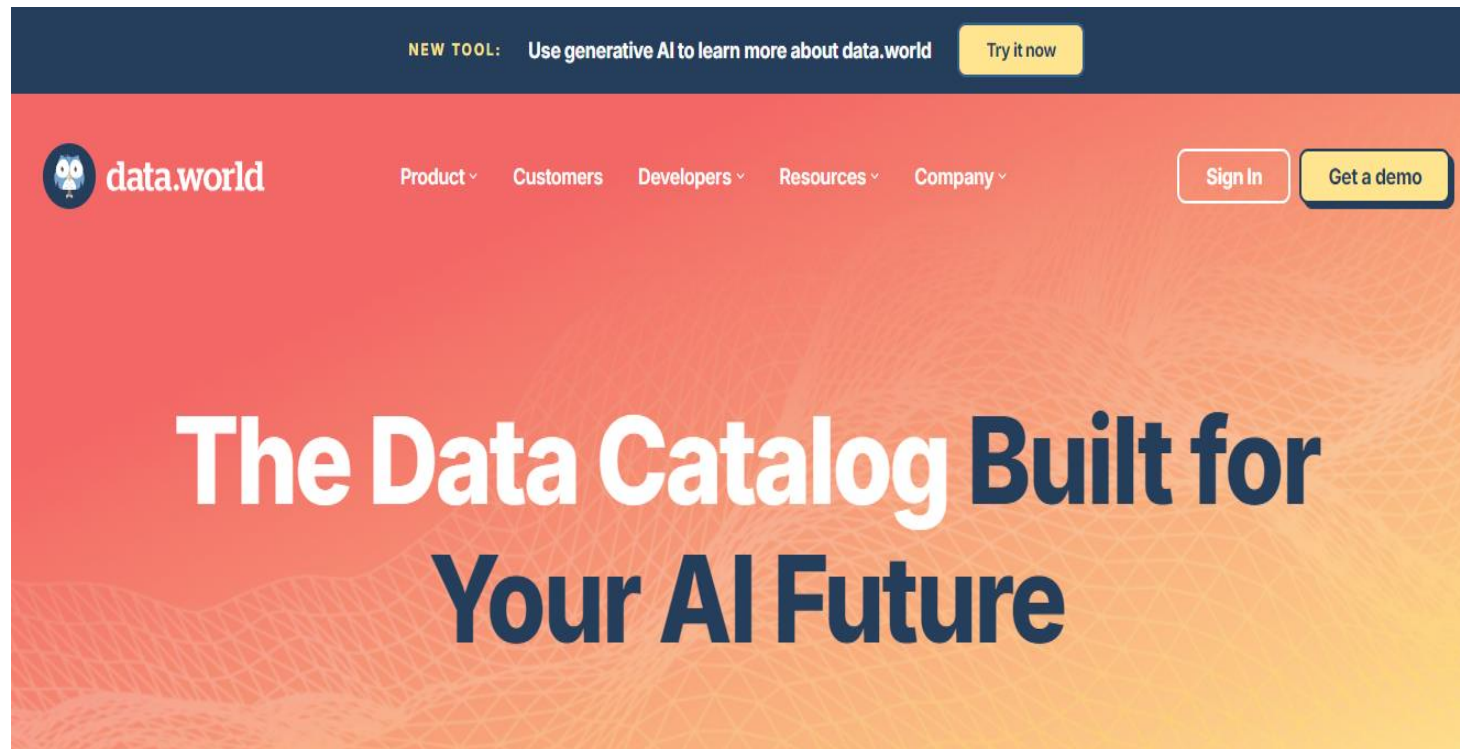
# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

5. **Data.world** : Data.world est une plateforme où vous pouvez trouver et partager des ensembles de données.

Il héberge une gamme diversifiée d'ensembles de données fournis par la communauté. Lien : <https://data.world/>



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES



## 0.9 Business Intelligence (BI)

### 0.9.4 Sources de donnees publiques

**6. Ensembles de données Reddit :** le sous-reddit r/datasets est une communauté où les gens partagent et demandent des ensembles de données. Vous pourriez trouver des ensembles de données intéressants ici <https://www.reddit.com/r/datasets/>

The screenshot displays the Reddit interface for the r/datasets subreddit. On the left, the navigation menu includes 'Home', 'Popular', 'RECENT' (with a dropdown arrow), and 'TOPICS' (with a dropdown arrow). Under 'TOPICS', there are links for 'Gaming', 'Sports', 'Business', 'Crypto', and 'Television'. The main content area shows the subreddit header for 'r/datasets' with a search bar and buttons for 'Create a post' and 'Join'. Below the header, a post by user 'u/Swat\_Sam2' is visible, titled 'Help with data analysis project (mysql online server help)'. The post text describes a problem with a MySQL server and asks for help. The right sidebar contains community statistics: '190K Members', '35 Online', and 'Top 1% Rank by size'. There are also links for 'Create a post', 'Join', and 'Community Bookmarks'.



# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

## 0.9 Business Intelligence (BI)

### 0.9.5. Labo : Power BI Desktop



**Lancer le Power Query Editor**

**1. Découverte de données/ Mise en forme des données.**

- ✓ Connexion aux données (Get Data)
- ✓ Nettoyage des données (Transform data)

**2. Modélisation des données**

- ✓ Création de relations
- ✓ Création de hiérarchies
- ✓ Calculs DAX( Expression de l'analyse des données)

**3. Visualisation des données**

- ✓ Création de rapports
- ✓ Visuels personnalisés

**Affichage du rapport**

**Affichage des données**

**Affichage du modèle**

**Add data to your report**

Once loaded, your data will appear in the Data pane.

Import data from Excel

Import data from SQL Server

Paste data into a blank table

Use sample data

Get data from another source →

# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

## 0.9 Business Intelligence (BI)

### 0.9.5. Labo : Power BI Destktop



New Table : Calendar = CALENDARAUTO()

New Column : Year = Year('Calendar'[Date])

New Column : Month Number = MONTH('Calendar'[Date])

New Column : Month = FORMAT([Date], "mmmm")

New Measure : Total Banks = COUNTROWS('banklist')

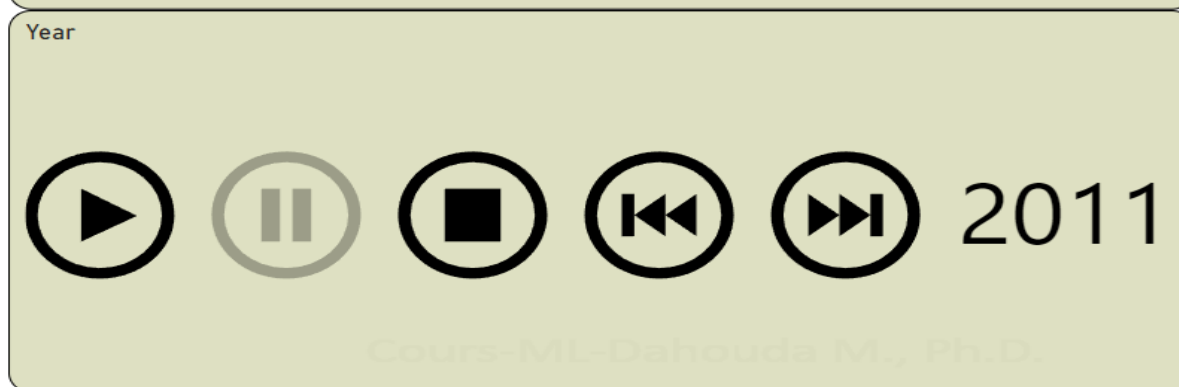
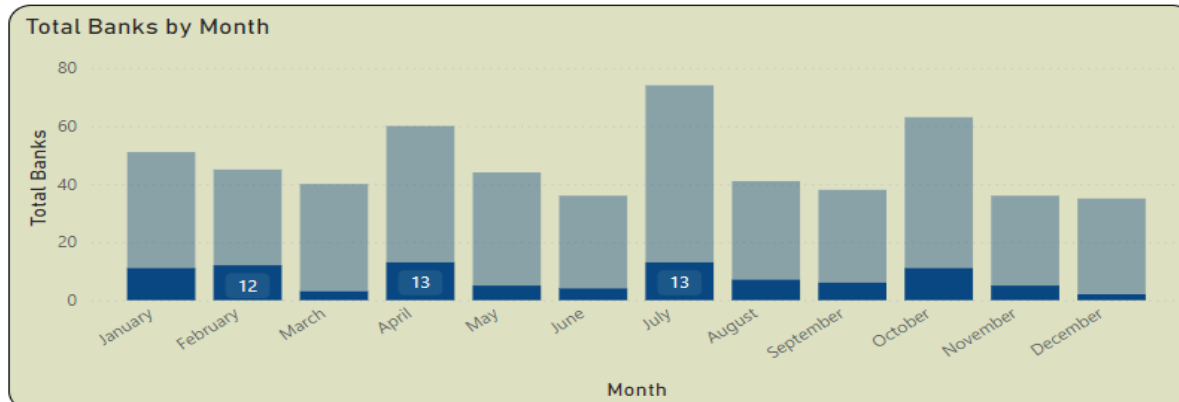
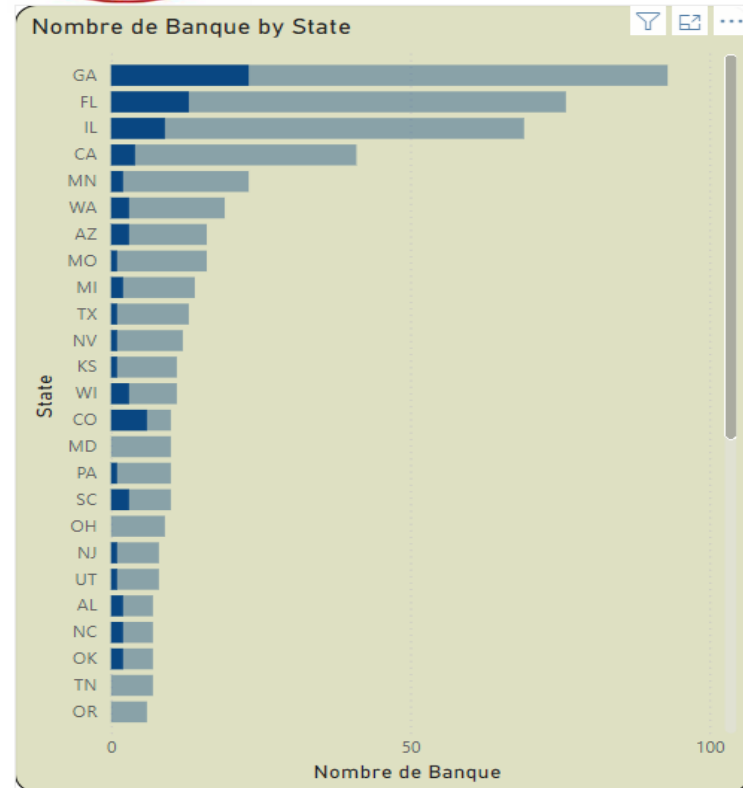
# CHAPITRE 0 LES DONNEES DANS LA SCIENC DE DONNEES

## 0.9 Business Intelligence (BI)

### 0.9.5. Labo : Power BI Desktop



## Intro to Power BI : Report Summary



# CHAPITRE 1 Introduction au Machine Learning

