Université de Caen Normandie
IUT Grand Ouest Normandie - Pôle de Caen
Département Science des Données
*Antenne de Lisieux*

Diplôme Bachelor Universitaire de Technologie
SCIENCE DES DONNÉES
*Parcours Science des Données : Exploration et modélisation statistique*

Troisième année

# PORTFOLIO

Oscar JOSEPH--GENESLAY

Année universitaire 2022-2025

# Sommaire

# SAÉ Création de reporting à partir de données stockées dans un SGBD relationnel

## A. Présentation du projet

Cette SAÉ était divisée en deux parties. Dans la première partie, le chef de police de Gotham City nous a confié la mission de **retrouver la véritable identité de Batman**. Dans la deuxième partie, le maire de Gotham City nous a demandé de réaliser des **représentations graphiques** sur les caractéristiques de sa population. Le travail devait être effectué à l'aide de Python, de fichiers CSV fournis par notre enseignant et du langage SQL. Nous devions rendre un rapport rédigé en LaTeX. Nous avions des heures dédiées en cours pour avancer sur le projet. Pour la réalisation de ce travail nous avions besoin des notions de programmation python ainsi que de connaissances dans le langage SQL.

## B. Organisation du travail

Nous étions un groupe de trois pour ce projet. J'avais pour mission de réaliser la première partie sur la recherche de l'identité de Batman, tandis que les autres membres devaient s'occuper de la deuxième partie sur les représentations graphiques de la population de Gotham City. Nous avons réalisé notre travail en cours ainsi que dans les salles informatiques mais aussi pendant nos heures personnelles chez nous. Nous avons profité des heures de cours mises à notre disposition pour effectuer des réunions pour voir l'avancement du projet.

## C. Démarche et résolution du problème

J'ai d'abord commencé par ouvrir tous nos fichiers CSV (contenant des informations sur les habitants de Gotham City) afin de les préparer pour la manipulation des données. Nous avons dû faire une étape d'agrégation de certaines données. Des informations sur l'identité de Batman nous ont été communiquées, et nous avons donc croisé ces informations avec nos fichiers CSV à l'aide de programmes **Python** et d'une base de données **SQLite**. Nous avons regardé pour chaque information qui nous étaient fournis, lesquelles ne correspondaient pas à notre individu. À la fin, nous avons obtenu une seule personne correspondant aux informations communiquées par le chef de police de Gotham City. La difficulté était de travailler sur des jeux de données avec un grand nombre de lignes et d'informations à traiter, ce qui rendait l'exécution longue. Ensuite, Nous nous sommes occupés de créer des représentations graphiques de la population de Gotham City (cf **Figure 1**). Encore une fois, la difficulté était de travailler avec de grands jeux de données. La réalisation des graphiques a été effectuée à l'aide de **R Studio** et **Excel**, ce qui a été complexe.

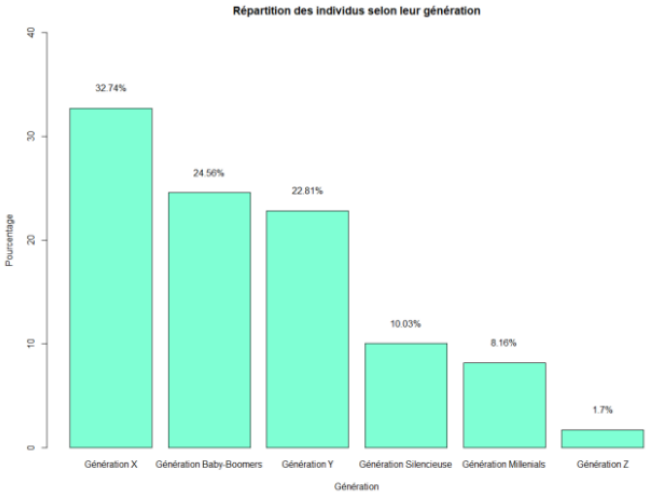



*Figure 1 : Représentation graphique sur des caractéristiques de la population*

## D. Mes acquis

On peut conclure que cette SAÉ nous a montré que le traitement des données peut s'avérer plus difficile que prévu lorsque les jeux de données sont volumineux. Il est donc important de réfléchir en amont à la technique la plus optimisée pour éviter de perdre du temps sur le projet. Cette expérience nous a permis d'apprendre à mieux travailler en équipe, à se répartir les tâches efficacement et à collaborer pour atteindre un objectif commun.

# SAÉ Écriture et lecture de fichiers de données

## A. Présentation du projet

Pour cette SAÉ, Madame Vermeersch, la directrice d'un collège, nous a demandé de créer un fichier contenant les moyennes des élèves dans plusieurs matières, afin que le conseil de classe de la commune de Tilloy-lez-Marchiennes puisse se tenir malgré un problème informatique. Nous avions pour mission **de créer un fichier CSV comportant les noms, prénoms et moyennes des élèves** dans chaque matière, ainsi qu'un carnet de notes pour chaque élève de la classe, avec leurs moyennes, leur mention et les notes de chaque matière. Nous avions à notre disposition des fichiers au format .txt et .csv. Les données nous ont été fournies sur la plateforme E-campus. Nous devions également rendre un rapport en **LaTeX**.

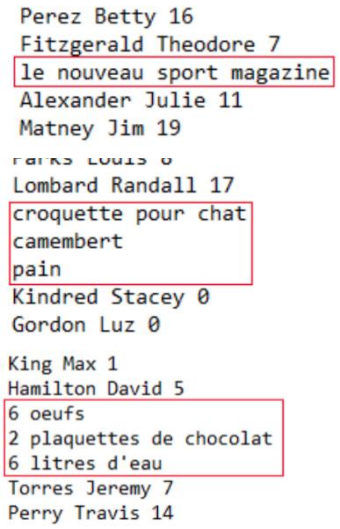
## B. Organisation du travail

Notre groupe était composé de trois étudiants, chacun ayant une fonction spécifique. Un membre s'occupait du traitement des fichiers CSV, pour ma part je m'occupais des fichiers au format TXT et l'autre membre réalisait le rapport en LaTeX au fur et à mesure de notre avancement dans le projet. Nous avons principalement travaillé dans les salles informatiques de l'IUT où nous avons mené des réunions d'avancement de projet, pendant les heures prévues dans notre emploi du temps, ainsi que sur notre temps libre. Les notions de **programmation Python** nous ont été utile pour mener à bien ce projet.

## C. Démarche et résolution du problème

Nous n'avons pas rencontré de difficultés particulières pour ouvrir et traiter les fichier **CSV**. Le travail était plus compliqué quand il s'agissait d'obtenir les moyennes des matières des élèves car les notes étaient au format ''caractère'', ce qui ne permettait pas de faire directement une moyenne. On a donc utilisé la fonction "map", qui permet de transformer les valeurs énumérées en format nombre. J'avais la tâche de retirer les données indésirables comme la liste de courses (Figure 2 : Données atypiques (liste de course)) qui a été le plus délicat et m'a pris beaucoup de temps pour comprendre comment les supprimer. J'avais remarqué que mes données inutiles commençaient par une minuscule. J'ai donc utilisé la fonction ".lower()" pour les repérer. Mon deuxième problème était de supprimer ces valeurs. Ma première méthode consistait à remplacer ces données par des chaînes de caractères vides (""). Cependant, cela créait des données vides qu'il fallait ensuite supprimer. J'ai donc opté pour une autre technique. J'ai créé une nouvelle variable et dès que je voyais une donnée importante, je l'ajoutais dans cette variable. À la fin de mon programme, j'obtenais toutes les bonnes données sans les données indésirables. L'objectif final a été atteint, nous avons finalement réussi à créer nos fichiers afin que le conseil de classe puisse se tenir dans les meilleures conditions.

*Figure 2 : Données atypiques (liste de course)*




## D. Mes acquis

Dans cette SAÉ, nous avons pu remarquer que le traitement et la préparation des données avant de les utiliser étaient primordiaux. En effet, si nous ne réalisons pas ces étapes, nous risquons d'être bloqués par la suite. Pour ma part, j'aurais pu être plus efficace en prenant plus de temps pour réfléchir à mes erreurs, afin de les résoudre plus rapidement plutôt que d'essayer, sans résultat, pendant des heures. La communication et l'organisation dans le groupe sont les clés pour réussir un tel projet.

# SAÉ Préparation et synthèse d'un tableau de données

## A. Présentation du projet

Dans cette SAÉ, notre objectif était de réaliser une **analyse exploratoire** sur la location de vélos dans la ville de New-York pour le mois de juin 2022, à l'aide du **langage R**. Nous avons récupéré les données mises à disposition sur le site de Citibike de New-York et nous avons travaillé sur le fichier "202206-citibike-tripdata.csv". Nous avons utilisé **RStudio** pour créer des graphiques et effectuer une analyse sur les données souhaitées. Un rapport sur Word nous a été demandé (20 pages).

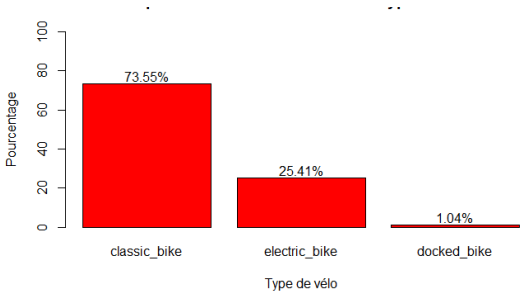
## B. Organisation du travail

Étant en groupe de 2, nous avons mis en place un système de programmation en binôme. Pour ma part, j'étais à côtés pour corriger d'éventuelles erreurs et réfléchir à la meilleure façon de procéder pour obtenir un résultat cohérent, ainsi que pour la création du code demandé. Nous avons travaillé ensemble dans les salles informatiques de l'IUT ainsi que chez nous.

## C. Démarche et résolution du problème

Le principal problème que nous avons pu constater était lié à l'utilisation d'un logiciel que nous avions vu très peu auparavant, ainsi qu'à un nouveau langage de programmation et une nouvelle structure. Nous avons dû régulièrement consulter la documentation de **R** afin de comprendre les différentes fonctions qui nous étaient demandées pour parvenir à la création d'un graphique. Le plus difficile n'était pas de réaliser les graphiques en eux-mêmes, mais plutôt d'organiser et de filtrer les données de manière à les stocker dans des objets pour nos représentations graphiques. Nous avons quand même réussi à réaliser la totalité des graphiques attendus.

*Figure 3 : représentation graphique sous R sur la répartition des types de vélos*




## D. Mes acquis

Dans le cadre de cette SAÉ, nous avons pu constater que le traitement et la préparation des données avant leur utilisation étaient primordiaux pour éviter toute analyse faussée. Le langage R est un outil essentiel à maîtriser pour des poursuites d'études ou pour entrer dans le domaine des statistiques. La programmation en binôme est particulièrement pertinente dans ce type de travail, où la moindre erreur peut biaiser nos résultats. Elle permet d'éviter les erreurs d'inattention et facilite la compréhension du programme et du sujet. Ce projet nous a permis de comprendre l'importance de la rigueur dans la préparation des données et de la maîtrise de l'outil informatique pour obtenir des analyses fiables. De plus, le travail en binôme s'est avéré très efficace pour la réussite de ce projet. J'ai pu développer mes compétences en langage R.

# SAÉ Régression sur données réelles

## A. Présentation du projet

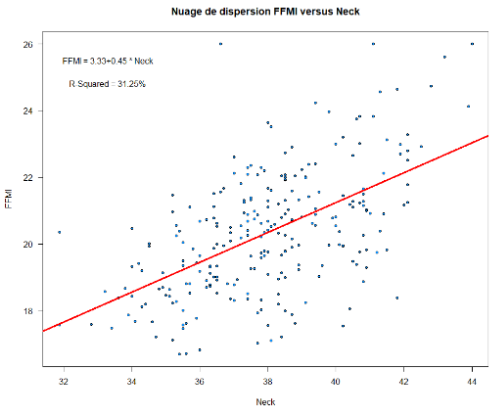
Le but de cette SAÉ était d'**étudier l'indice de masse maigre** en relation avec certaines variables morphologiques telles que la circonférence du cou, des hanches ou de l'abdomen, afin de déterminer si ces variables pouvaient **expliquer une variation de l'indice**. Pour cela, nous avons effectué une analyse exploratoire des données et des modélisations pour mieux comprendre cette relation. Nous avons importé des données issues d'un fichier **CSV**. L'objectif était de rendre un **rapport** sur ces analyses et de faire une **présentation orale** sous PowerPoint.

## B. Organisation du travail

Nous étions un groupe de trois étudiants, et j'étais responsable de la création du code **R** sous **RStudio**. J'ai donc préparé les données, créé des représentations graphiques, ainsi que des tableaux pour comprendre la relation entre les variables morphologiques et notre indice de masse maigre. Les autres membres du groupe ont réalisé les rapports écrits et PowerPoint.

## C. Démarche et résolution du problème

*Figure 4 : Nuage de dispersion entre FFMI et Neck*




Pour débuter notre projet d'analyse de l'indice de masse maigre, nous avons commencé par préparer les données en sélectionnant les valeurs qui nous intéressaient et en ajoutant des indicateurs, tels que l'indice de masse maigre que nous avons calculé. Nous avons également supprimé des données atypiques pour éviter qu'elles ne faussent nos représentations graphiques. Pour cela, nous avons utilisé l'écart interquartile pour obtenir des valeurs minimales et maximales. Nous avons ensuite réalisé des représentations graphiques pour étudier l'association entre notre indice et nos variables morphologiques à l'aide de nuages de points. Nous avons cherché les coefficients de détermination pour identifier la meilleure variable à étudier, ce qui s'est avéré être la circonférence du cou. Enfin, nous avons créé un nuage de dispersion croisant notre meilleur modèle avec l'indice de masse maigre et avons inséré la droite des moindres carrés (**Figure 4 : Nuage de dispersion entre FFMI et Neck**). Les étapes de préparation des données ainsi que les analyses ont été réalisé sous le langage **R** avec le logiciel **RStudio**.

## D. Mes acquis

Cette SAÉ nous a permis de comprendre l'importance de bien analyser les graphiques pour comprendre le sujet. Si nous ne préparons pas nos données en amont, nous risquons d'avoir des représentations faussées et donc des analyses erronées. Le travail en groupe était également crucial car cela nous a permis de réfléchir plus rapidement au code à intégrer dans notre script, nous faisant ainsi gagner un temps précieux. Personnellement, j'ai également pu améliorer ma compréhension sur le langage **R** qui était nouveau pour moi ainsi que les différentes fonctions mises à notre disposition.

# SAÉ 3.02 – Intégration de données dans un datawarehouse

## A. Présentation du projet

Cette SAÉ nous a fait travailler sur les données météorologiques et sur la qualité de l'air de la ville de Sheffield. Le projet était divisé en deux parties. Dans la première partie, nous devions **importer automatiquement des données** présentent dans des fichiers de type *CSV* et les traiter pour les insérer dans une base de données *MySQL*. La seconde partie consistait à **effectuer des représentations graphiques** à l'aide du langage de programmation *Python* pour analyses les données. Le travail devait ainsi être effectué sur *Python* et en *SQL*, sur une machine virtuelle utilisant le système d'exploitation *Ubuntu*. Les données ont été récolté sur le site *Airviro*. Nous devions rendre un rapport écrit et faire une soutenance orale pour présenter nos analyses. Pour la réalisation de ce travail nous avons dû utiliser les notions de programmation, d'analyse de données et de communication.

## B. Organisation du travail

Pour réaliser ce projet nous étions en groupe de trois étudiants. J'avais pour mission de coder les scripts notebook d'importation et d'analyse des données, ainsi que de réaliser les rapports (écrit et soutenance orale). Les autres membres du groupe ont également travaillé sur ces tâches. Nous avons réalisé ce travail sur les heures prévues à cet effet, mais également sur notre temps personnel.

## C. Démarche et résolution du problème

J'ai d'abord commencé par importer les données des fichiers *CSV* dans *Python* pour avoir créé deux objets de traitements : un représentant les données météorologiques et un autre pour la pollution. L'objectif de passer par ce langage de programmation était de **nettoyer et réorganiser les données** à l'aide des packages *Pandas* et *Numpy*. À la fin du traitement, j'ai obtenu une liste *Python* avec les différentes données concaténées, ce qui sera optimisé pour les insérer ensuite dans une base de données. J'ai finalement créé deux listes *Python* : une résumant les paramètres avec les unités des valeurs et une autre rassemblant les observations. Ensuite, j'ai inséré l'intégralité des données dans une base de données *MySQL* avec *Python*. Les deux tables *SQL* étaient liées par une clé d'identification unique pour chaque donnée. Avec ce même langage, j'ai enfin entrepris l'étape d'analyse de données en réalisant différents graphiques avec le package *Matplotlib (exemple :* **Figure** *)*. Avec ces données, nous avons analysé l'évolution des températures moyennes par année, puis **nous avons étudié l'influence du vent sur la pollution** pour voir une potentielle **corrélation** et enfin **nous avons examiné la qualité de l'air de Sheffield** selon les années en fonction des molécules de pollution.

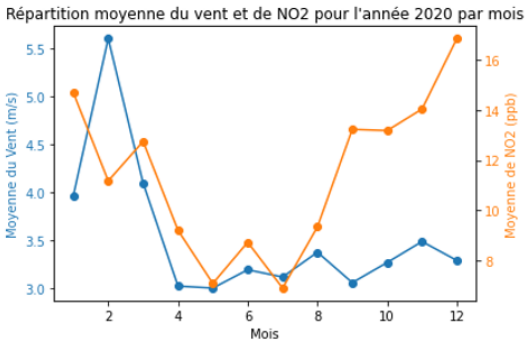



*Figure 5 : Représentation d'une analyse réalisée*

La complexité de ce projet était dans la durée nécessaire pour insérer les données dans la base de données. En cas d'erreur non détectée dans nos données, cela aurait exigé de recommencer le processus, entraînant ainsi une perte significative de temps.

## D. Mes acquis

Cette SAÉ nous a montré que le traitement des données avant les analyses est une étape importante dans la réalisation d'un projet. En effet, si on ne prend pas du temps sur cette partie, on risque d'obtenir des données erronées. La voluminosité des données est également à prendre en compte lors d'un projet car elle peut prendre beaucoup de temps qu'il faut éviter de perdre. Cette expérience m'a permis de développer mes compétences en programmation *Python* et en requêtage *SQL*.

# SAÉ 3.03 – Description et prévision de données temporelles

## A. Présentation du projet

Ce projet avait pour but de travailler sur les **séries chronologiques**. Nous avons analysé les données mises à disposition par *l'INSEE* sur les nuitées dans l'hôtellerie en Normandie sur la période de janvier 2011 à septembre 2023. L'objectif était d'**analyser les différentes composantes** ainsi que de **faire une prédiction des données.** Pour réaliser cette SAÉ, nous avons utilisé le logiciel *RStudio* pour les analyses. Les données étaient sous le format *CSV*. L'objectif final était de concevoir un rapport à l'aide d'un fichier *R-Markdown* sur ce sujet ainsi qu'une soutenance orale réalisée sur *PowerPoint*. Nous avons eu besoin des notions en **programmation *R*** pour ce projet, mais également des connaissances sur la thématique des **données temporelles**.

## B. Organisation du travail

Afin de mener à bien ce projet, j'étais avec deux autres étudiants. J'avais pour mission de réaliser le code d'analyse ainsi que le rapport. Les autres membres de mon groupe m'ont aidé à faire cette tâche et ont réalisé le diaporama pour la soutenance orale. Nous avons réalisé ce projet sur notre temps personnel. Nous nous sommes réunis à plusieurs reprises pour répartir les tâches à accomplir.

## C. Démarche et résolution du problème

J'ai d'abord commencé par ouvrir le fichier *CSV* présent dans un dossier zippé que je devais dézipper sous le langage *R*. Puis, j'ai entrepris l'étape de **l'analyse exploratoire** qui nous permet d'avoir une vision globale des données. Ensuite, j'ai réalisé **un modèle de régression linéaire** par morceaux et un autre modèle de régression permettant d'estimer la composante saisonnière, ce qui nous a permis de visualiser les mois les plus favorables et défavorables au nombre de nuitées dans l'hôtellerie. Puis, j'ai **analysé les résidus**. Pour finir, j'ai **réalisé une prévision** de données sur un an pour voir l'évolution dans le temps du nombre de nuitées.

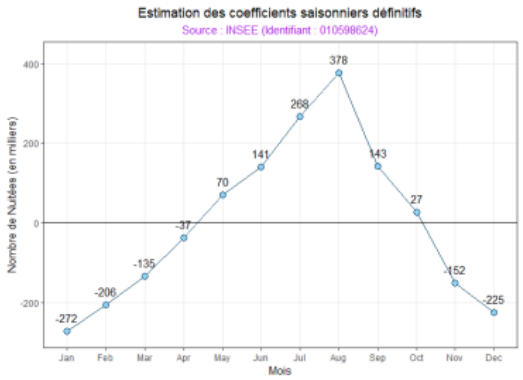



*Figure 6 : Représentation d'une analyse réalisée*

La difficulté dans ce projet était d'analyser des graphiques que nous n'avions pas l'habitude de voir et qui nécessitaient d'effectuer des recherches pour ne pas réaliser une analyse faussée.

## D. Mes acquis

La préparation des données et la conception d'un modèle de régression sont primordiales dans le cadre de ces analyses, car un modèle mal ajusté ou mal réalisé peut amener à des analyses erronées. Ce projet nous a permis de faire des analyses très variées. Cette SAÉ m'a permis de développer mes compétences en programmation *R* et en analyses de données dans un contexte d'une série chronologique.

# SAÉ 4.01 – Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs

## A. Présentation du projet

Dans le cadre de cette SAÉ, j'ai abordé la **prédiction d'une variable quantitative** à partir de divers facteurs, tant quantitatifs que qualitatifs. L'objectif principal était de **prédire le poids d'un manchot** des îles Palmer en Antarctique **en utilisant des modèles de régression linéaire**. Pour mener à bien ce projet, j'ai utilisé le logiciel *RStudio* et le langage de programmation *R*. Le rapport a été rédigé en *LaTeX* et la présentation orale a été réalisée grâce à la programmation avec *Marp*. Les données ont été récupérées en utilisant la librairie *palmerpenguins* de *R*. Plusieurs scripts *RMarkdown* ont été développés pour analyser, traiter et modéliser les données. La réalisation de ce projet a nécessité une solide compréhension de la programmation en R ainsi que des compétences en conception de modèles statistiques.

## B. Organisation du travail

Ce projet a été réalisé en collaboration avec deux autres étudiants. Ma responsabilité principale était de développer les scripts de programmation en *R* pour prédire et analyser le poids des manchots. De plus, j'ai conçu le support visuel pour la soutenance orale en utilisant *Marp*. La majorité du travail a été effectuée en dehors des heures de cours. Par ailleurs, des Travaux Pratiques et Dirigés nous ont permis d'acquérir les compétences nécessaires à la réalisation de ce projet.

## C. Démarche et résolution de problème

Pour aborder cette SAÉ, nous avons exploré quatre techniques de prédiction d'une variable quantitative, suivant une méthodologie structurée en trois étapes : **analyses exploratoires, conception des modèles** (mathématiques et en *R*), et **vérification des modèles** (linéarité, homoscédasticité, gaussianité et identification des points influents). Voici les différentes approches adoptées :

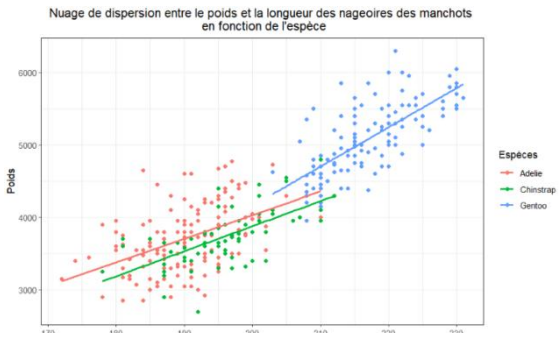



*Figure 7 : Visualisation d'un modèle*

- **Modèles de régression linéaire simple** : Cette étape visait à analyser les variations du poids en fonction de variables spécifiques telles que la longueur de la crête supérieure du bec, la profondeur de la crête, et la longueur des nageoires, afin d'identifier d'éventuelles associations.
- **Modèles de régression linéaire de type ANOVA** : Nous avons étudié l'influence des variables qualitatives, comme l'espèce et le sexe des manchots, sur le poids en réalisant une analyse de variance pour évaluer les différences significatives entre les groupes.
- **Modèles de régression linéaire de type ANCOVA :** Cette étape avait pour objectif d'expliquer les variations du poids en utilisant la longueur de la nageoire en combinaison avec l'espèce ou le sexe, afin de mettre en lumière d'éventuelles associations.
- **Modèles de régression linéaire multiple** : Enfin, nous avons cherché à prédire le poids en utilisant le modèle le plus robuste, sélectionné par une approche ascendante et descendante manuelle, en exploitant le maximum de variables significatives disponibles.

## D. Mes acquis

Au cours de ce projet, j'ai renforcé mes compétences en programmation avec *R*, en particulier dans la conception, l'analyse et la vérification de modèles de régression linéaire. De plus, j'ai développé une meilleure compréhension des méthodes statistiques pour l'analyse de données quantitatives et qualitatives, ainsi que la capacité à présenter mes résultats de manière claire et structurée. Aussi, ce projet m'a permis d'approfondir ma capacité à travailler en équipe et à gérer efficacement les ressources et le temps pour atteindre les objectifs fixés.

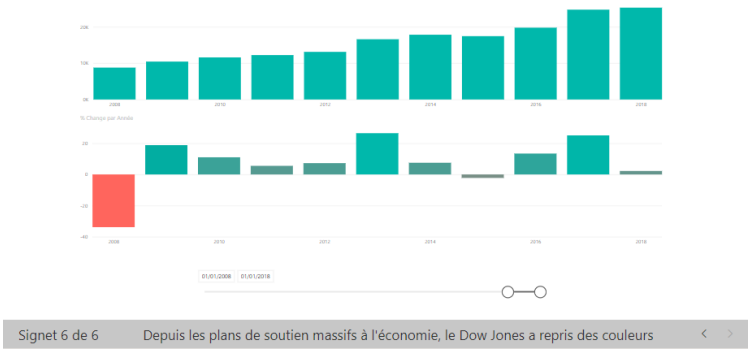
# Storytelling sous PowerBI

## A. Présentation du travail pratique

Ce travail pratique visait à nous familiariser avec la création d'un **storytelling** autour du Dow Jones Industrial Average, un indice boursier. L'objectif était d'acquérir des compétences dans l'art du storytelling, avec des graphiques et des données préalablement mis à notre disposition. De cette manière, nous avons exploré les diverses fonctionnalités de **PowerBI** pour concevoir une narration captivante.

## B. Démarches

*Figure 8 : storytelling*




L'objectif de cette séance pratique était de développer les compétences nécessaires pour élaborer une narration dynamique sur l'indice boursier étudié. J'ai acquis la capacité de saisir des moments clés en utilisant la fonction de "signet", qui permet de figer un instant précis de l'histoire, tel qu'un effondrement du marché financier. En outre, j'ai appris à mettre en évidence un graphique spécifique parmi plusieurs présents sur le tableau de bord, offrant ainsi une clarification visuelle. L'ajout d'un filtre d'année a également été exploré pour examiner des segments spécifiques des données.

## C. Mes acquis

L'objectif global de créer un storytelling est d'expliquer de manière accessible des éléments complexes présents dans les graphiques. Ce processus permet de simplifier l'interprétation des données et d'offrir une narration compréhensible autour d'événements significatifs tels que les crises boursières. Ce travail pratique m'a permis de m'entrainer et de découvrir le storytelling sous **PowerBI**. J'ai appris à manipuler les fonctions du storytelling qui permettent de créer un storytelling attrayant.

# SAÉ Mener une étude statistique dans un domaine d'application

## A. Présentation du projet

Ce projet tutoré m'a fait travailler sur de la **statistique épidémiologique** en étudiant les facteurs de risques et protecteurs liés à l'incidence du cancer du poumon. Ce travail était en lien avec le Centre François Baclesse qui a fourni les données de la **cohorte AGRICAN**, contenant un suivi de santé d'agriculteurs français, et qui nous a aidés tout au long de cette SAE. Pour étudier ces facteurs, j'ai utilisé les **tests du Chi-2 et de Student**, ainsi que le **modèle de Cox**, en utilisant les langages de **programmation *SAS* et *R***.

## B. Organisation du travail

- **Modalité :** travail de groupe (3 étudiants)
- **Suivi :** réunions entre nous et notre tuteur
- **Ressources :** cours de statistiques épidémiologiques, TP sur le modèle de Cox, programmation R et SAS, articles scientifiques
- **Outils :** RStudio, SAS, Excel
- **Livrables :** rapport (Word), présentation orale (25min, PowerPoint)

## C. Démarche et résolution de problème

Un traitement de données avait été effectué par les équipes de Baclesse avant de nous faire travailler dessus afin de nous faire gagner du temps. Toutefois, nous avons dû préparer un peu notre base pour qu'elle corresponde à notre étude. Nous avons donc calculé, à l'aide du langage *SAS*, la durée d'élevage bovin de chaque agriculteur et exclu les résidents du département de la Côte-d'Or, qui ne faisaient pas partie de l'étude. Nous avons créé d'autres variables, comme la variable d'incidence (malade ou non malade), ainsi qu'une variable définissant les classes d'âge. Puis, nous avons exporté cette base de données. Suite à cela, nous avons créé un **flowchart** représentant le nombre total d'individus traités et restants, ainsi que les différentes étapes de traitement. Ensuite, nous avons utilisé le langage R pour décrire les caractéristiques sociodémographiques, les habitudes de vie et les activités agricoles. Nous avons ainsi **réalisé des tableaux (en *R*)** comparant les deux groupes de la variable d'incidence avec d'autres variables (l'âge, le sexe, le nombre de paquets de cigarettes par an, etc.), en réalisant des **tests statistiques** du Chi-2 et de Student, afin de déterminer s'il y avait des différences significatives entre les cas et les non-cas. Enfin, nous sommes retournés sur *SAS* pour étudier les **associations entre les variables** sociodémographiques, les habitudes de vie et les variables d'exposition professionnelle agricole avec le cancer du poumon. Pour ce faire, nous avons utilisé le **modèle de Cox** avec l'âge comme échelle de temps. Cela nous a permis d'obtenir les Hazard Ratios, l'intervalle de confiance à 95% et les p-values pour chaque facteur étudié. Nous devions présenter ces résultats sous forme de tableaux Excel, mais nous avons aussi réalisé des graphiques pour certaines variables, afin d'apporter plus de clarté. Pour ce qui est du rendu du projet, nous avons écrit un rapport sur les résultats que nous avons obtenus, nous les avons comparés à des articles scientifiques et nous avons présenté des résultats d'études, notamment dans notre introduction.

| | | Modèle de Cox | |
|---|---|---|---|
| | N cas | HR (IC95%) | P-Value |
| **Sexe** | | | |
| Hommes | 967 | Référence | |
| Femmes | 255 | **0,28** [0,25-0,33] | **< 0,0001** |
| Valeurs manquantes (n = 4) | | | |
| **IMC** | | | |
| <18.5 | 22 | **1,87** [1,22-2,89] | **0,0043** |
| [18.5-25[ | 379 | Référence | |
| [25-30[ | 501 | **0,58** [0,38-0,88] | **0,0115** |
| >= 30 | 167 | **0,54** [0,34-0,84] | **0,0062** |
| Valeurs manquantes (n = 17 872) | | | |

*Figure 9 : Extrait d'un tableau d'analyse*

## D. Mes acquis

Ce projet m'a offert une expérience enrichissante en **statistiques épidémiologiques**, en manipulant le modèle de Cox. J'ai également pu manipuler les langages *SAS* et *R*, et **développer mes compétences** dans ces outils. Ce projet m'a aussi permis de **découvrir les articles scientifiques** dans ce domaine. Cela a été un bon entraînement en **communication**, lors de la présentation orale ou des réunions avec notre tutrice.

# SAÉ Migration de données vers ou depuis un environnement NoSQL

## E. Présentation du projet

Dans ce projet, j'ai travaillé sur de la **programmation *NoSQL*** en étudiant plusieurs aspects clés de l'activité commerciale du site e-commerce Olist, tels que les tendances d'achat, les délais de livraison ou encore les scores de satisfaction, dont les données étaient stockées dans différents fichiers au format *CSV*. Ce travail a été réalisé à l'aide du système de gestion de base de données, **MongoDB**, ainsi que du langage de programmation **Python** en réalisant un **notebook documenté**. J'ai ainsi dû utiliser plusieurs méthodes et fonctions que propose ce SGBD afin d'obtenir les résultats demandés.

## F. Organisation du travail

- **Modalité :** travail individuel
- **Ressources :** cours de gestion de base de données NoSQL, TP sur MongoDB et Python
- **Outils :** Python (Notebook)
- **Livrables :** rapport (Word), Notebook

## G. Démarche et résolution du problème

La première étape de ce projet était de **créer la base de données MongoDB** à l'aide du package **pymongo** de *Python*, notamment en construisant les différentes tables qui devront être utilisées. Les données étaient disponibles sous 7 fichiers CSV différents, tous contenant un type de données spécifique (les clients, les produits, les fournisseurs, etc.). J'ai donc ensuite inséré automatiquement les données dans les différentes tables créées, à l'aide d'une **fonction *Python* réalisée**. Comme l'objectif était de produire des analyses sur ces données, j'ai donc créé une nouvelle fonction qui **affiche les résultats d'une requête et qui stocke en même temps les données** dans une liste *Python*.

```
prix_moyen = products.aggregate([
    {
        "$lookup": { # Jointure avec la table order_items
            "from": "order_items",
            "localField": "product_id",
            "foreignField": "product_id",
            "as": "product_info"
        }
    },
    {"$unwind": "$product_info"},
    {"$match": {"product_info.price": {"$ne": None}}},  # Filtrer les documents sans prix
    {
        "$group": {
            "_id": "$product_category_name", # Grouper par les catégories de produits
            "moyenne_prix": {"$avg": "$product_info.price"} # Calcul de la moyenne des prix
        }
    },
    {"$sort": {"moyenne_prix": -1}} # Trier les prix dans l'ordre décroissant
])

liste_prix_moyen = [] # Assignation d'une nouvelle liste
listes_ajout(prix_moyen, liste_prix_moyen) # Affichage des résultats + insertion des données dans une liste
```

*Figure 10 : Exemple de requête utilisée*

Les analyses que je devais faire étaient imposées par mon enseignant. J'ai donc commencé à faire des **jointures entre les tables**, à l'aide d'identifiants uniques qui permettaient cette liaison, et utilisé différentes **méthodes de filtrage** (group, match, limit, etc.) des données afin d'obtenir des résultats. Je suis ainsi tombé sur un problème de temps d'exécution des requêtes qui me semblait très long (plusieurs dizaines de minutes). En me renseignant sur des forums d'aide, je me suis aperçu que plusieurs cas comme le mien avaient été découverts. Le problème se trouvait dans le typage des colonnes des identifiants pour les jointures qui étaient sous le format caractère. Ainsi, la requête comparait caractère par caractère chaque identifiant pour lier les données. J'ai donc trouvé qu'il fallait appliquer le typage « index », en utilisant la fonction create_index() de pymongo, pour définir des colonnes comme des ID. Par la suite, cette méthode a été concluante puisque mes requêtes ne prenaient que quelques secondes à s'exécuter.

J'ai donc terminé en réalisant des **analyses graphiques** en utilisant le package **matplotlib** de *Python* mettant à disposition la création de visualisations. Dans le même temps, j'utilisais le format **Notebook** de *Python* afin de créer un **code documenté et lisible** permettant d'identifier clairement les parties d'analyses.

## H. Mes acquis

Grâce à ce projet, j'ai **renforcé ma maîtrise de *Python*** et **approfondi mes connaissances** en bases de données *NoSQL*, particulièrement **MongoDB**. J'ai également exploré les techniques d'indexation pour **optimiser les performances** du code. L'utilisation de Jupyter Notebook m'a permis d'adopter une approche structurée et de produire un code bien documenté et facile à comprendre.

# SAÉ Mise en œuvre d'un processus de Datamining

## D. Présentation du projet

Dans ce projet, l'objectif était **d'analyser les avis des utilisateurs** sur certains produits du site e-commerce Olist. Cette analyse visait à évaluer la satisfaction client pour ensuite **développer un modèle de Deep Learning et de clustering.** Ce travail devait être effectué avec un **Notebook *Python***. Les données, sous le format *CSV*, étaient fournies par notre enseignant. Plusieurs packages ont été utilisés afin de donner accès à différentes fonctions pour mener à bien ce projet telles que **pandas** (pour la gestion de données), **nltk** (pour le traitement de texte), **matplotlib** (pour les visualisations graphiques) ou encore **sklearn** (pour la prédiction).

## E. Organisation du travail

- **Modalité :** travail individuel
- **Ressources :** cours et TP sur le traitement, la prédiction de données, le Deep Learning et le clustering
- **Outils :** Python (Notebook)
- **Livrables :** rapport (Word), Notebook

## F. Démarche et résolution du problème

Pour commencer le projet, il a fallu **créer une variable cible** (pour la prédiction de données) permettant de définir la satisfaction des clients. J'ai donc défini des modalités qualitatives (très satisfait, satisfait, mécontent, très insatisfait) à partir de la note de satisfaction (de 0 à 5) attribuée par les clients. Ensuite, la deuxième étape devait **traiter les données** qui étaient sous forme de texte libre. Pour ce faire, j'ai utilisé différentes méthodes de **traitement en langage naturel (NLP)** afin de nettoyer correctement la donnée :

- **Suppression des valeurs manquantes :** conservation uniquement des données exploitables
- **Élimination des valeurs atypiques :** filtrage des données pour garantir la fiabilité des analyses
- **Conversion en minuscules :** uniformisation du texte pour faciliter l'analyse
- **Suppression de la ponctuation :** élimination des caractères non informatifs
- **Élimination des stop words :** retrait des mots courants sans valeur sémantique significative
- **Lemmatisation :** réduction des mots à leur forme canonique pour regrouper les variantes
- **Suppression des chiffres :** élimination des caractères numériques non pertinents pour l'analyse

Puis, j'ai réalisé un **clustering et une prédiction de données** en comparant deux méthodes de traitement textuel : **Bag of Words** (comptage simple des occurrences) et **TF-IDF** (pondération selon l'importance des mots). Pour optimiser les performances d'exécution, j'ai travaillé sur un **échantillon** représentatif de 20 % des données initiales avec un vocabulaire limité à 5 000 mots. La décomposition SVD a permis de conserver l'information essentielle tout en réduisant la dimensionnalité. L'efficacité des méthodes a été évaluée via le **score ARI**. Pour la partie prédictive, j'ai développé un modèle de **Deep Learning** en séparant les données en ensembles d'apprentissage (80 %) et de test (20 %), permettant ainsi d'entraîner le modèle sur des données distinctes de celles utilisées pour l'évaluation. Le but était de **prédire les quatre modalités qualitatives**. J'ai obtenu des taux de performance d'environ 68 %. Afin de visualiser ces prédictions, j'ai créé **une matrice de confusion** et un graphique représentant les **courbes ROC** de chaque modalité prédite afin de voir les prédictions les mieux réalisées.

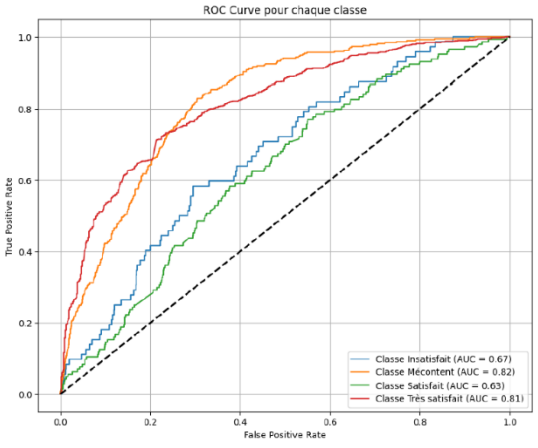



*Figure 11 : Visualisation des courbes ROC*

## G. Mes acquis

Cette SAÉ m'a permis de développer des compétences en **traitement et prédiction de données** et particulièrement dans le **Natural Language Processing (NLP)**. J'ai également développé un **esprit critique** au regard des résultats que j'obtenais afin de paramétrer correctement les modèles pour avoir des résultats cohérents. Et j'ai enfin développé des **qualités d'organisation** du code en utilisant le Notebook *Python*.

# SAÉ Modélisation statistique pour les données complexes et le Big Data

## A. Présentation du projet

L'objectif de cette SAÉ était de **créer différentes méthodes de prédiction** de données afin de prédire l'état de maladie (malade ou non malade) d'un agriculteur à partir de ratios de pratique d'activités agricoles (pourcentage de temps sur une activité). Pour ce faire, différentes étapes de **traitements et de modélisations** ont été réalisées. Ce travail a été réalisé sous le langage *R* à l'aide du logiciel *RStudio*. Les données provenaient de la **cohorte AGRICAN** de 1950, sous le format rds, spécifique à *R*.

## B. Organisation du travail

- **Modalité :** travail en groupe de 3 étudiants
- **Ressources :** cours et TP (R et SAS) sur les méthodes d'analyses, de traitement et de prédiction de données sur de la Big Data
- **Suivi :** heures dédiées, travail personnel
- **Outils :** Langage R ; logiciel RStudio
- **Livrables :** rapport (Word), code R

## C. Démarche et résolution de problème

La première étape du projet consistait à **traiter et décrire les données**. Nous avons recherché les **valeurs atypiques par des analyses univariées et multivariées**. Pour l'analyse univariée, nous avons examiné **les courbes de densités lissées** afin d'identifier d'éventuels pics anormaux et compilé dans un tableau Excel **les statistiques descriptives** (quantiles, médiane, écart-type, etc.) pour chaque variable. L'analyse multivariée s'est appuyée sur une **Analyse en Composantes Principales**, permettant de visualiser le cercle des corrélations et de projeter les individus dans un espace bidimensionnel pour identifier ceux qui se démarquaient significativement des autres. Suite à l'analyse des valeurs propres de l'ACP, nous avons conservé les 12 premières composantes principales. Cette réduction de dimensionnalité nous a permis **d'optimiser la base de données** tout en préservant l'essentiel de l'information. La deuxième étape du projet était consacrée à la **sélection des variables les plus influentes**. Pour ce faire, nous avons mis en œuvre une approche **bootstrap** en générant 50 échantillons aléatoires sur lesquels nous avons effectué **des régressions logistiques**. Cette méthode nous a permis d'évaluer les p-values de chaque variable dans les données initiales et d'identifier les 5 variables les plus significatives pour notre modèle. Nous avons aussi utilisé la méthode **Backward** pour connaître automatiquement les variables les plus significatives. La dernière étape de ce projet consistait à **créer des modèles prédictifs**. Nous avons utilisé séparément les 5 variables significatives des données initiales et les 5 premières composantes de la base de données réduite pour entraîner et tester nos modèles. Ces modèles ont été exécutés sur 40 échantillons de données. Voici les différents modèles que nous avons conçus :

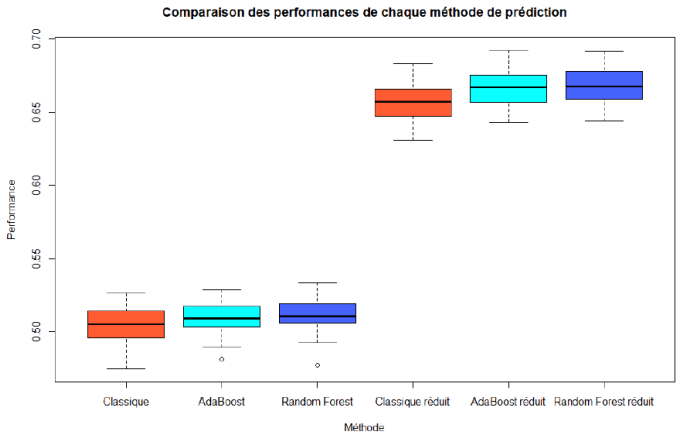



*Figure 12 : Visualisation des performances des arbres de décisions*

- **La régression logistique :** avec la méthode classique, bagging et boosting
- **Les arbres de décisions :** avec la méthode classique, Ada Boost et forêts aléatoires
- **Les réseaux de neurones :** avec une couche cachée et 2, 4 et 6 nœuds

Les performances ont été représentées sous forme de **boxplots** pour chaque méthode avec chaque type de données (initiales / réduites).

## D. Mes acquis

Avec ce projet, j'ai acquis des compétences en **traitement et analyse de données**, ainsi qu'en **modélisation prédictive** en utilisant des techniques avancées sous le langage *R*. J'ai également développé des aptitudes en **travail collaboratif** et en gestion de projets.