UNIVERSITÉ CAEN NORMANDIE

Université de Caen Normandie

GRAND OUEST NORMANDIE

IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie

SCIENCE DES DONNÉES

Campus de Lisieux

Science des données 1ère année :

PORTFOLIO

Maxime GAMONDELE

Année universitaire 2023-2024

Table des matières

Introduction

À travers ce portfolio, je présente différents projets que j'ai réalisés au cours de ma scolarité. Ces projets incluent des SAÉ (Situations d'Apprentissage et d'Évaluation) ainsi que des Projets Personnels (PP). Les deux types de projets ont pour objectif de mettre en pratique les notions apprises afin de résoudre des problématiques concrètes. Chaque projet est accompagné d'une description détaillée, expliquant les objectifs, les méthodes utilisées, et les résultats obtenus.

SAÉ 1-01 Création de reporting à partir de données stockées dans un SGBD relationnel

A. Présentation du projet

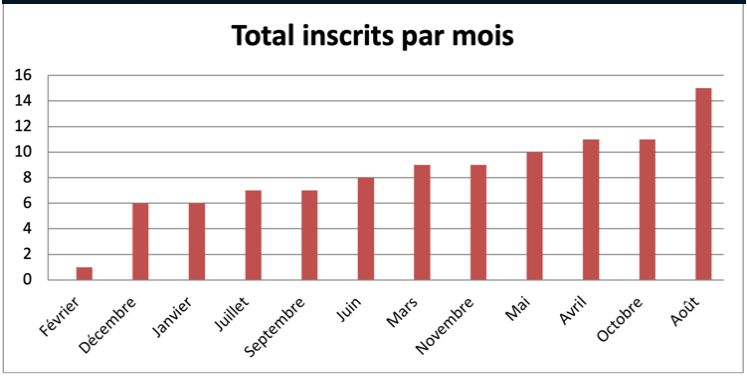
Le projet « création de reporting » vise à générer des graphiques et des tableaux dans Excel à l'aide de Python. Ces graphiques nous permettent d'étudier les données d'un réseau social que nous avons manipulées dans une base de donnée relationnelle et nous en avons tirée des informations en vue d'améliorer cette plateforme.

B. Organisation du travail

Le travail a été réalisé par groupes de 3 étudiants. La première partie, la structuration de la base de données a été effectuée individuellement par chaque membre du groupe et a ensuite été mise en commun afin de tirer le meilleur de chaque base de données. Par la suite, les tâches ont été réparties de manière à ce que chaque étudiant puisse avancer sur le projet à son rythme et à son niveau, tout en veillant à ce que cela soit bénéfique pour l'ensemble du groupe. À la fin de leur avancement respectif, les membres de l'équipe m'ont remis leur script Python. Cela m'a permis de créer un fichier final comprenant un script Python, une base de données, et un rapport détaillant les résultats obtenus. Pour ce projet, je me suis appuyé sur les cours de base de programmation et les cours de bases de données relationnelles.

C. Démarche et résolution du problème

```
    ####DIAGRAMME pour les mois
diag = BarChart()
diag.title = "Total inscrits par mois"
#DATA on sélectionne les zones de la feuille excel dont on a besoin pour faire les graphiques
data = Reference(f_total_inscrits_par_M, min_col=1, max_col=2, min_row=2, max_row=f_total_inscrits_par_M.max_row)
diag.add_data(data,titles_from_data=False) #avoir uniquement les mois en x et pas les valeurs (meilleur visuel)
cats = Reference(f_total_inscrits_par_M, min_col=1, max_col=1, min_row=2, max_row=f_total_inscrits_par_M.max_row)
diag.set_categories(cats)
diag.legend = None #retire la legende qui ici n'est pas pertinante (meilleur visuel)
#ajout diag à la feuille
f_total_inscrits_par_M.add_chart(diag,'D2')
```



La première partie du projet consistait à créer une base de données à l'aide de DB browser. Pour ce faire nous avons utiliser des requêtes SQL. Une fois que la base de données a été construite, nous avons utilisé Python afin d'y faire des requêtes SQL, nous permettant de récupérer des informations pertinentes sur la base de données telles que la liste des utilisateurs inactifs, les inscriptions en fonctions des jours, les photos les plus likées, les plus commentées, etc. Une fois tous ces éléments collectés, nous avons proposé des représentations graphiques pour certains. Pour ce faire, nous avons utilisée la bibliothèques openpyxl.chart. Cette dernière nous permet de créer des graphiques dans Excel directement.

D. Acquis

Ce projet m'a permis de consolider mes connaissances en SQL et sur les bases de données relationnelles. L'implémentation du langage SQL sur Python a été intéressante, notamment avec la découverte des requêtes paramétrées et des sous-requêtes. Grâce à ces dernières, j'ai pu définir des critères spécifiques lors de l'appel de fonctions Python qui modifie directement la requête SQL pour répondre aux critères définis.

En enrichissant ma compréhension de ces deux langages, j'ai également exploré une nouvelle bibliothèque, «openpyxl», en particulier son module dédié aux graphiques. Grâce à cette compétence supplémentaire j'ai compris comment créer des graphiques sur Excel à l'aide de Python.

SAÉ 1-02 Écriture et lecture de fichiers de données

A. Présentation du projet

Le projet « écriture et lecture de fichier de données » vise à créer des fonctions pour manipuler un fichier CSV sans bibliothèques, afin d'améliorer notre compréhension des listes, des dictionnaires et de la gestion des fonctions. La première étape consiste à nettoyer le fichier pour pouvoir analyser et traiter efficacement les données contenues dans celui-ci. Les fonctions nous on servi pour analyser les données du fichier et nous retourner des ratios, indicateurs et informations pertinentes.

B. Organisation du travail

Travail de groupe réalisé par groupe de 4 étudiants. En tant que leader de groupe, j'ai eu à répartir les tâches et à suivre l'avancement de mon équipe. Pour ce faire, je me suis appuyé sur Excel pour avoir une vue globale sur le travail de chacun et l'avancée de la SAÉ. Le projet a été réalisé à l'aide de Python, chaque semaine, notre groupe se retrouvait pour avancer sur les différentes parties et pour s'entraider si besoin. Pour optimiser l'avancement du projet, j'ai utilisé mes TD, cela m'a permis de revoir les exercices faits en cours et d'appliquer la logique de ces derniers dans le projet. Le rendu du devoir est divisé en deux parties. La première est un fichier Python regroupant les fonctions créées par chaque membre du groupe, et la deuxième partie est un rapport mettant en avant le travail que l'on a effectué, mais aussi les difficultés rencontrées et la manière dont nous les avons surmontées, offrant ainsi une perspective détaillée sur notre démarche.

C. Démarche et résolution du problème

Afin de démarrer le projet, il a fallu créer une fonction capable de lire le fichier CSV. Je fus responsable de la création de cette fonction, essentiel au commencement du projet. Durant la programmation de cette fonction, j'ai rencontré une erreur qui était la gestion du curseur. De ce fait à la place d'avoir des descripteurs qui indiquent le nom de chaque colonne, j'avais directement des valeurs. Par exemple à la place d'avoir « âge », j'allais avoir « 42 ». J'ai donc cherché à résoudre le problème au plus vite et je me suis servi d'un de mes scripts que j'avais fait en TP et j'ai compris mon erreur qui était due au fait que mon curseur était mal placé et lisait directement la ligne 1 à la place de la ligne 0. Une fois ce problème résolu, j'ai pu transmettre la fonction à mes camarades pour qu'ils puissent travailler leur partie. Les fonctions suivantes que nous nous sommes réparties sont des fonctions de création, modification de données et de calculs statistiques pour une meilleure compréhension des données.

D. Acquis

```python
def ajouter_col_mineur_adulte(donnees):
    """
    Ajoute une colonne 'mineur_adulte' aux données en fonction de l'age (age)
    des passagers.

    Args:
        donnees (list): Une liste de dictionnaires représentant le jeu de données.

    Returns:
        None: La fonction modifie les données directement en ajoutant la colonne
        'mineur_adulte'.
    """
    for passager in donnees:
        if passager["age"] is None:
            passager["mineur_adulte"] = "Inconnue"
        elif passager["age"] < 18:
            passager["mineur_adulte"] = "mineur"
        else:
            passager["mineur_adulte"] = "adulte"


def taux_de_survie_global(donnees):
    """
    Calcule le taux de survie global à partir d'un jeu de données de passagers.

    Args:
        donnees (list): Une liste de dictionnaires représentant le jeu de données des passagers.
            Chaque dictionnaire devrait contenir au moins la clé "survived", indiquant si le passager a survécu.
    Returns:
        float: Le taux de survie global en pourcentage.

    """

    total_survivant = 0
    for passager in donnees:
        if passager["survived"] == 1:
            total_survivant += 1

    tx_survie= (total_survivant/len(lst_survivants))*100
    return round((tx_survie),2)  #arrondissement du résultat au centième
```

Au cours de ce projet, j'ai enrichi ma maîtrise des bases du langage Python, notamment à travers la gestion des boucles, des listes, des fonctions et des chaînes de caractères. Le projet m'a permis d'approfondir mes connaissances en manipulation de fichiers CSV, un domaine qui me semblait complexe à son abord. Notamment avec les différentes étapes pour accéder au fichier à l'aide du module « os » (os.path.dirname, os.path.join). Ci-joint, un extrait de deux fonctions que j'ai réalisées. La première me permet d'ajouter une colonne qui identifie la catégorie de chaque passager en fonction de son âge. La deuxième permet de calculer un taux de survie. Ces deux fonctions m'ont fait travailler la gestion et explorations des listes. Bien que ce travail aurait pu être effectué à l'aide de la bibliothèque Pandas ou Numpy, il m'a permis de comprendre les bases de la programmation Python et de me créer des fondations solides pour la suite de mon apprentissage de ce langage.

SAÉ 1-03 Préparation et synthèses d'un tableau de données en vue d'une analyse exploratoire simple

A. Présentation du projet

Le projet « préparation et synthèse de données » a pour but de mener une étude statistique sur les locations de vélos Citibike au cours du mois de septembre 2023 dans la ville de New York. L'objectif de cette étude est d'étudier les comportements de location de vélos en appliquant les concepts de statistique descriptive nous permettant de comprendre les tendances dans ces données.

B. Organisation du travail

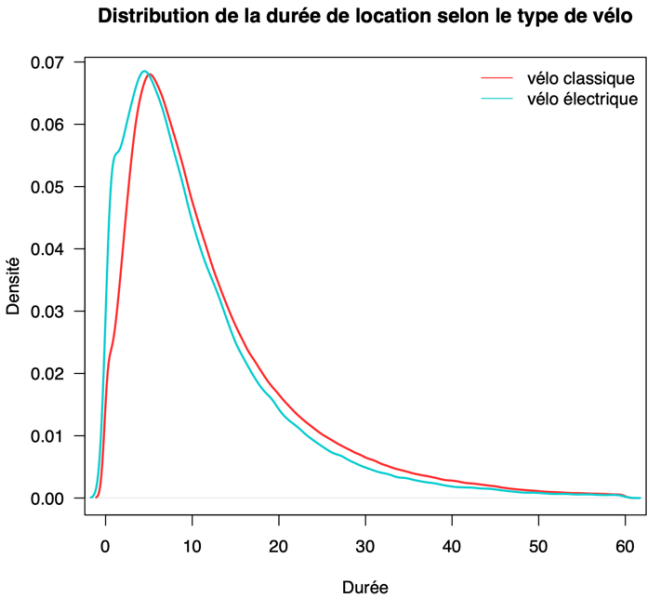
Ce projet a été réalisé par groupe de 2 étudiants, à l'aide du logiciel RStudio, préalablement étudié en classe. Pour réussir notre étude, nous nous sommes appuyés sur les cours de statistique descriptives et les cours de programmation en R. Le rendu de ce projet est divisé en deux parties. Une première contenant le script R avec les fonctions permettant de faire les calculs et les représentations graphiques. La seconde partie, quant à elle, était un rapport. Pour améliorer nos connaissances des différents langages de programmation, nous avons décidé d'utiliser LaTeX, un langage et un système de composition de documents qui est notamment utilisé pour les publications scientifiques.

C. Démarche et résolution du problème

La première étape du projet consistait à préparer les données. Pour ce faire, nous avons importé la librairie «lubridate», celle-ci permet de formater les dates et les heures. Nous avons aussi vérifié les types de données que nous avions dans chaques colonne et modifié si nécessaire. Une fois toutes ces étapes de préparation et chargement des données effectuées, nous avons pu commencer l'analyse.

Chaque question à laquelle nous devions répondre correspondait à une modélisation de graphique ou une analyse statistique. Une fois les tâches réparties, à chaque section terminée, je la partageais avec mon camarade pour avoir un avis extérieur sur mon étude.

Parmi différentes représentations, celle ci-contre est intéressante, car elle a été faite en plusieurs étapes. Premièrement, il a fallu séparer les vélos électriques et classiques dans deux objets différents et ensuite faire leur représentation sur un graphique commun pour pouvoir les comparer. Cela a nécessité d'avancer étape par étape pour garantir la clarté et la compréhension des résultats.



D. Acquis

Cette SAÉ m'a permis de mieux comprendre comment effectuer des représentations graphiques en R. Mon ressenti sur ce langage de programmation a vraiment évolué au fur et à mesure de la SAÉ. Initialement, j'avais des difficultés avec les différentes étapes pour créer un graphique, et vers la fin, j'avais non seulement compris le processus, mais j'étais aussi capable de proposer des modifications sur le graphique et de les appliquer.

SAÉ 1-04 Apprendre en situation la production de données en entreprise

A. Présentation du projet

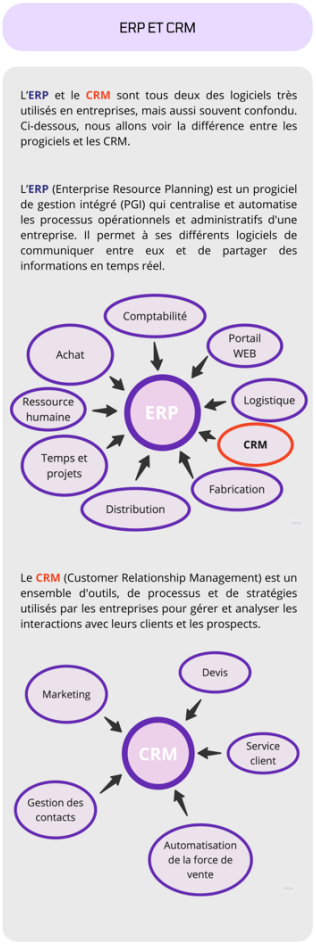
Le projet suivant, consistait à réaliser un poster synthétisant les points essentiels à savoir sur la production de données en entreprise. Le contenu du poster a abordé des aspects spécifiques tels que les différents types de données et les systèmes d'informations, fournissant ainsi une vision détaillée des données en entreprise.

B. Organisation du travail

Ce travail a été effectué par groupe de deux. Chaque membre de l'équipe a choisi les parties qu'ils souhaitaient présenter et synthétiser. Par la suite, toutes ces parties ont été assemblées entre elles pour faire le poster final. Pour faire ce poster nous nous sommes servies des outils PowerPoint, Canva et Miro pour les schémas.

C. Démarche et résolution du problème

L'essentiel de ce projet est la synthétisation d'information pertinentes. Chaque élément est un élément clé qui en amène un autre. Ils a fallu savoir créer un fil conducteur, gérer l'espace imposé, tout en ayant un maximum d'informations claires, compréhensibles, aérées et agréables à lire pour chaque lecteur. Par exemple, pour expliquer la différence entre l'ERP et le CRM, j'ai opté pour donner une définition de chacun, mais aussi les présenter par 2 schémas synthétisant leur rôle. Nous avons cherché à créer un poster non seulement informatif, mais également facile à mémoriser. Les images ci-joint représentent des extraits du poster avec les schémas réalisés sur Miro.



D. Acquis

Ce projet m'a permis de m'améliorer sur plusieurs éléments, mais deux d'entre eux se démarquent en particulier.

Tout d'abord, j'ai été poussé à comprendre les points de vue extérieures, pour savoir ce qui serait attrayant sur le poster. La transmission de l'information est devenue le point central. Pour transmettre, il faut assimiler et s'imprégner du contenu que l'on veut partager. J'ai donc cherché à comprendre chaque élément en rapport avec la donnée et avec l'entreprise. Cette démarche m'a permis d'approfondir mes connaissances sur des sujets tels que le fonctionnement de l'IOT, de l'ERP, des veilles, des systèmes d'informations, etc. En résumant ces points, j'ai amélioré ma capacité de synthèse et perfectionné ma capacité à transmettre des informations de manière simple. Deuxièmement, le travail sur le design était essentiel. L'objectif était de susciter l'intérêt du lecteur et de lui donner envie de parcourir le poster. En conclusion, ce projet m'a permis de transmettre des informations de façons simple et synthétisé, mais aussi de concevoir des présentations attractives visuellement.

SAÉ 1-05 Présentation en anglais d'un territoire économique et culturel

A. Présentation du projet

Le projet « Présentation en anglais d'un territoire économique et culturel » avait pour objectif de présenter des données sur les différences de salaires en fonction du genre en Normandie.

B. Organisation du travail

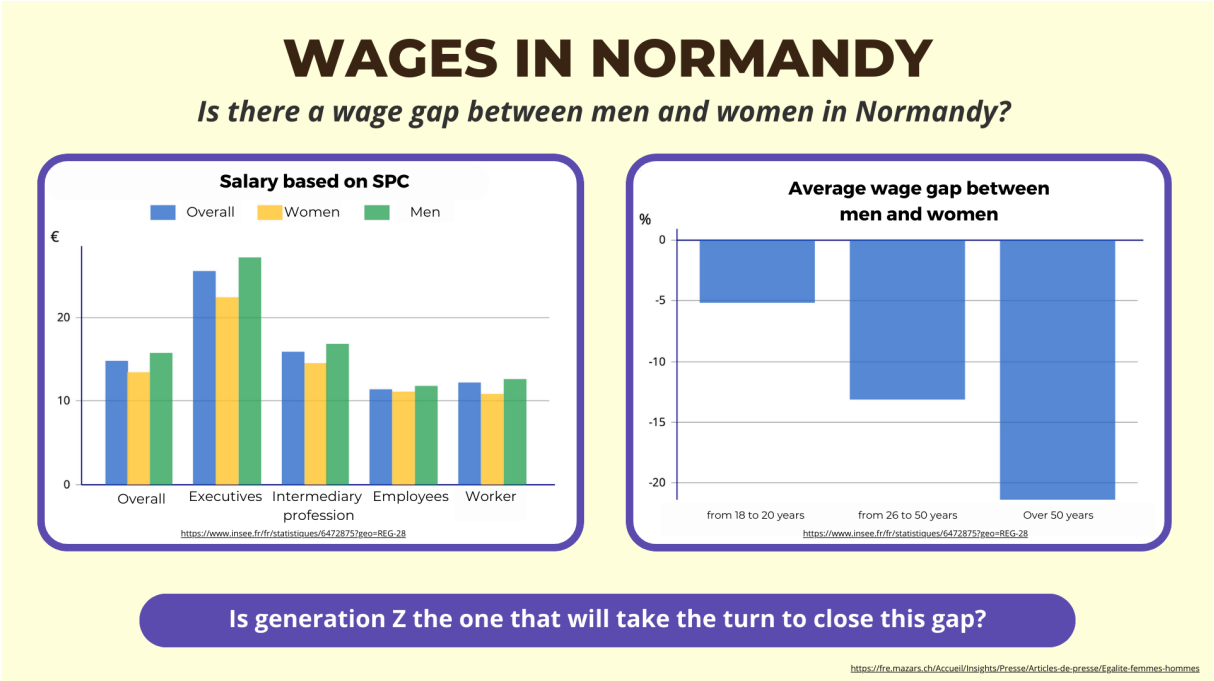
Ce projet a été réalisé individuellement. Les données utilisées pour répondre à la problématique proviennent des sources de l'INSEE. Pour présenter mes résultats, j'ai utilisé les outils de présentation PowerPoint et Canva. L'évaluation de ce projet portait sur le poster et sa présentation orale en anglais.

C. Démarche et résolution du problème

Pour répondre à la problématique, une phase de recherche de données a d'abord été nécessaire. À travers le site de l'INSEE, j'ai exploré différentes représentations graphiques mettant en avant les écarts de salaire en fonction du genre. J'ai opté pour deux visuels : le premier sur les salaires en fonction des catégories socioprofessionnelles (CSP) et le second en fonction de l'âge. Ces deux angles d'approche sont intéressants car ils permettent de voir les écarts de salaire de manière globale.

D. Acquis

Ce projet m'a permis de développer mon sens de la recherche d'informations et de comprendre ce qui pourrait avoir un impact sur les auditeurs. À travers ce travail, je n'ai pas seulement cherché à exposer de simples données, mais à les utiliser pour transmettre un message. De plus, j'ai exercé mon esprit critique pour convaincre mon public, le tout en anglais.



SAÉ 1-06 Mise en œuvre d'une enquête

A. Présentation du projet

Le projet "Mise en œuvre d'une enquête" consistait à concevoir une enquête sur un domaine spécifique et à en faire une étude. Nous avons choisi « Les lycéens et la pratique du sport ». Cette étude nous a permis de mettre en avant des faits concrets sur un échantillon d'une population réelle, ici les lycéens de la ville de Lisieux.

B. Organisation du travail

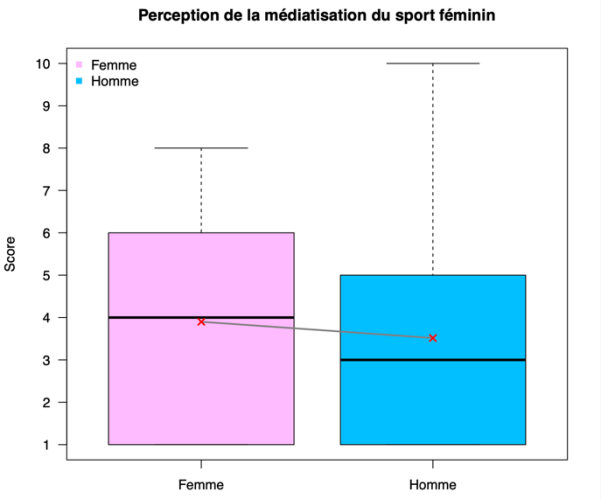
Ce travail a été réalisé par groupe de 4 étudiants. Pour ce projet, nous avons utilisé plusieurs logiciels, notamment Word pour faire le questionnaire, Excel pour mettre en forme les données collectés, RStudio pour l'analyse des données, Overleaf pour faire un rapport en LatTeX et enfin Powerpoint pour présenter nos résultats. Nous devions livre notre étude sous deux formats, une présentation et un rapport. Pour effectuer ce projet, je me suis servi des TP effectué sur RStudio et des TP Word pour la présentation du questionnaire.

C. Démarche et résolution du problème

Ce projet était divisé en trois grandes parties : le questionnaire, le traitement des données et la présentation des résultats.

Pendant la conception du questionnaire, l'accent a été mis sur la formulation des questions de manière à ce que ces dernières soient simples et compréhensibles. Cela pour garantir la précision des réponses lors du sondage des lycéens.

La seconde partie fut le traitement des données. Une fois, les questionnaires remplis, il a fallu reporter les réponses dans un fichier CSV afin de pouvoir traiter les données. Durant l'analyse avec RStudio, je ne trouvais pas le bon moyen pour analyser une variable quantitative contre une variable qualitative. En revenant sur mon cours de statistique descriptive, j'ai cherché quelle graphique était le plus adapté pour cette situation. J'ai d'abord essayé le diagramme de Cleeveland. Cependant, je n'ai pas aimé le visuel et trouvais qu'il manquait des indicateurs, donc au final j'ai opté pour des boxplot parallèles. Lors de mes premières tentatives de réalisations du graphique, je ne comprenais pas comment nous utilisions la fonction boxplot() jusqu'à lire la documentation R de la fonction et comprendre que les variables devaient être séparées par « ~ ». Le rendu final était plus intéressant, car on avait un aperçu de la médiane, de la moyenne, des quartiles un et trois, mais aussi des valeurs extrtêmes.



Pour la présentation des résultat nous avons cherché à synthétiser notre analyse pour respecter de temps de passage imposé pour la présentation. Il a donc fallu analyser quelles étaient les points clés de chaque représentation graphique et en présenter l'essentiel. Pour ce faire, on a donc travaillé avec chaque membre du groupe notre présentation en se chronométrant et en se conseillant les uns les autres pour échanger nos idées sur les points à améliorer.

D. Acquis

Durant ce projet, j'ai appris à formuler des questions de la bonne manière pour qu'elles soient compréhensibles de tous. De plus, j'ai approfondi mes connaissances sur R, notamment sur les études de deux variables en faisant des tableaux profilés et des graphiques leur correspondant. Et enfin, je continue de progresser sur mon niveau de présentation en cherchant à transmettre des informations complexes de manière simple.

SAÉ 2-01 Conception et implémentation d'une base de données

En cours.

SAÉ 2-02 Estimation par sondage simple

A. Présentation du projet

Le projet " Estimation par sondage simple " consistait à déterminerla loi de probabilité qui suivait les échantillons à notre disposition.

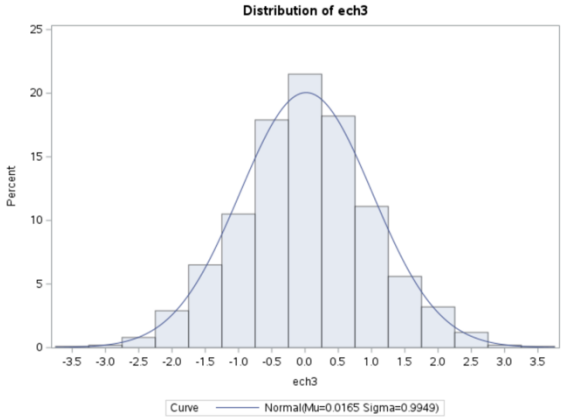
B. Organisation du travail

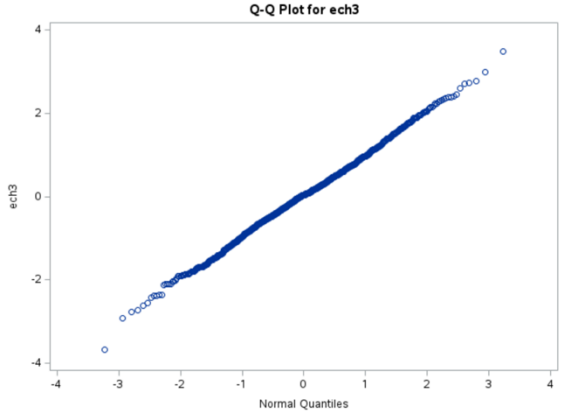
Ce travail a été réalisé par groupe de 2 étudiants. Pour ce projet, nous avons utilisé SAS (Statistical Analysis System) qui est un logiciel utilisé pour l'analyse statistique, la gestion de données, la visualisation et la prise de décision basée sur des données. Nous devions livrer notre étude sous forme d'un rapport, pour se faire, nous avons utilisé le langage LaTeX.

C. Démarche et résolution du problème

Pour déterminer la loi de probabilité, nous avons utilisé 2 méthodes principalement. La première, une méthode empirique, c'est-à-dire que nous utilisions la forme de distribution de l'échantillon à l'aide d'un histogramme comme nous pouvons le voir sur l'image de droite ci-dessous. À travers cette première méthode, nous comparons la distribution d'un échantillon à des distributions de probabilité classique comme la loi normale, la loi uniforme, etc. Pour identifier rapidement le type de loi, nous regardons en premier la symétrie de la distribution. Si elle l'est nous nous orientons vers les lois telles que Laplace-Gauss, Cauchy ou Student en revanche dans le cas contraire, nous nous penchions sur les lois Log-normales, Gamma ou Weibull.

La seconde méthode est celle du Q-Q Plot. Cette dernière consiste à étudier les quantiles théoriques d'une loi et le quantile observé de notre échantillon afin de voir si nous pouvons observer une droite à travers le nuage de point. Si c'est le cas alors notre échantillon se rapproche de la loi à laquelle nous avons comparé les quantiles. Ci-dessous, nous pouvons voir un nuage de dispersion des quantiles d'un échantillon et les quantiles théoriques de la loi normale. On voit que ces derniers s'alignent. On peut alors dire que notre échantillon semble suivre une loi normale ce qui confirme notre conjoncture précédente avec l'étude de l'histogramme.





D. Acquis

À travers ce projet, j'ai pu observé les différentes formes de répartions des lois de probabilité classique. Et à mettre en place une analyse à l'aide d'un Q-Q plot.

SAÉ 2-03 Régression sur données réelles

A. Présentation du projet

Le projet « régression sur données réelles » est un projet d'études statistiques cherchant à expliquer le poids des manchots provenant des archipels Palmer en Antarctique en fonction d'autres variables quantitatives. L'objectif est de savoir quelle variable explique le mieux la variabilité du poids, notamment en fonction du sexe et de l'espèce des manchots. Pour ce faire, nous utilisons des modèles de régression linéaires que nous allons comparer entre eux.

B. Organisation du travail

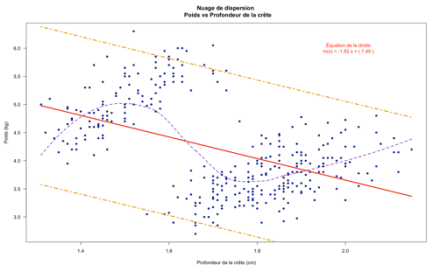
Ce projet a été réalisé par groupe de 3 étudiants. Au sein de notre groupe nous avons effectué la préparation des données ensemble afin de travailler sur le même jeu de données et avoir des résultats cohérents. Par la suite, j'ai divisé l'analyse et attribué une fonction à réaliser pour chaque membre du groupe. Cela nous a permis de travailler efficacement et surtout de s'entraider sur certains points si l'un d'entre nous rencontrait des difficultés. Le projet a été réalisé à l'aide du logiciel RStudio. Pour optimiser les résultats notre analyse, nous nous sommes appuyés sur les cours de statistique descriptive et les TP de cette SAÉ. Le rendu du projet est divisé en trois parties. Une première contenant le script R, une deuxième, le rapport rédigé en LaTeX et la dérnière, un script Marp pour la présentation des résultats.

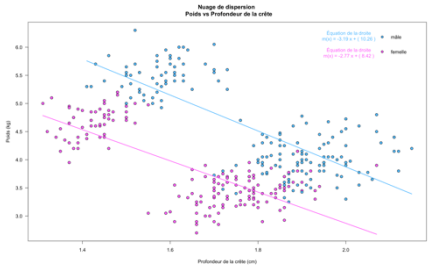
C. Démarche et résolution du problème

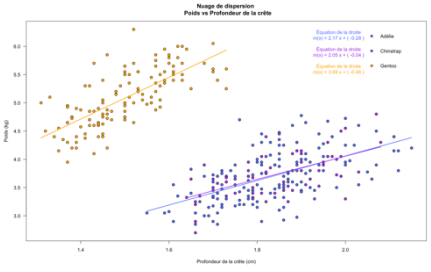
Pour répondre à l'objectif de ce projet nous avons décidé d'utiliser 3 fonctions. La première permettant de faire une simple analyse du poids des manchots en fonctions d'autre variables quantitatives. La deuxième fonction permet de faire une analyse similaire mais selon deux niveaux, ici, selon le sexe. Et enfin la dernière fonction une analyse selon 3 niveaux, ici, les espèces des manchots. Ce découpage en trois parties nous a aidé à mieux aborder le sujet et nous a permit de répondre correctement à la problématique.

Un des problèmes majeurs de notre projet était le fait de comprendre comment faire une étude selon une certaine variable par exemple le sexe ou bien l'espèce des manchots. Pour y arriver, nous avons dû créer des sous-ensembles du dataset principal (ex : pour pour la variable sexe nous avons crée les datasets mâle et femelle). Cela nous a permis d'étudier la variable poids en fonction de différents niveaux.

D. Acquis







Selon nos premières études, l'augmentation de la profondeur de la crête semblait entraîner une diminution du poids des manchots. Cependant, cela s'est avéré être faux. Dès que notre étude a été portée sur les espèces, l'analyse s'est inversée. Nous étions confrontés à un cas de paradoxe de Simpson qui stipule qu'un phénomène observé dans plusieurs groupes s'inverse lorsque les groupes sont combinés. L'augmentation de la profondeur de la crête induisait dorénavant une augmentation du poids chez les manchots. Pour conclure, cette SAÉ m'a apporté une compréhension de ce paradoxe, une meilleur approche et gestion des fonctions en R et aussi la capacité à étudier les associations entre des variables quantitatives. De plus, en approfondissant mes recherches, j'ai découvert que le coefficient de Spearman peut être utilisé pour mesurer l'association dans le cas de relations non linéaires, en se basant sur les rangs des observations. Cependant, dans le cadre de notre étude, nous n'avons pas eu besoin d'utiliser cette méthode de mesure.

SAÉ 2-04 Datavisualisation (challenge)

En cours.

SAÉ 2-05 Construction et présentation d'indicateurs de performance

A. Présentation du projet

Le projet intitulé « Construction et présentation d'indicateurs de performance » est un projet de business intelligence visant à présenter une analyse démographique et géographique des écureuils de Central Park.

B. Organisation du travail

Ce travail a été réalisé en groupe de quatre étudiants. Nous avons utilisé le logiciel Tableau Desktop pour créer un tableau de bord (dashboard) de notre étude.

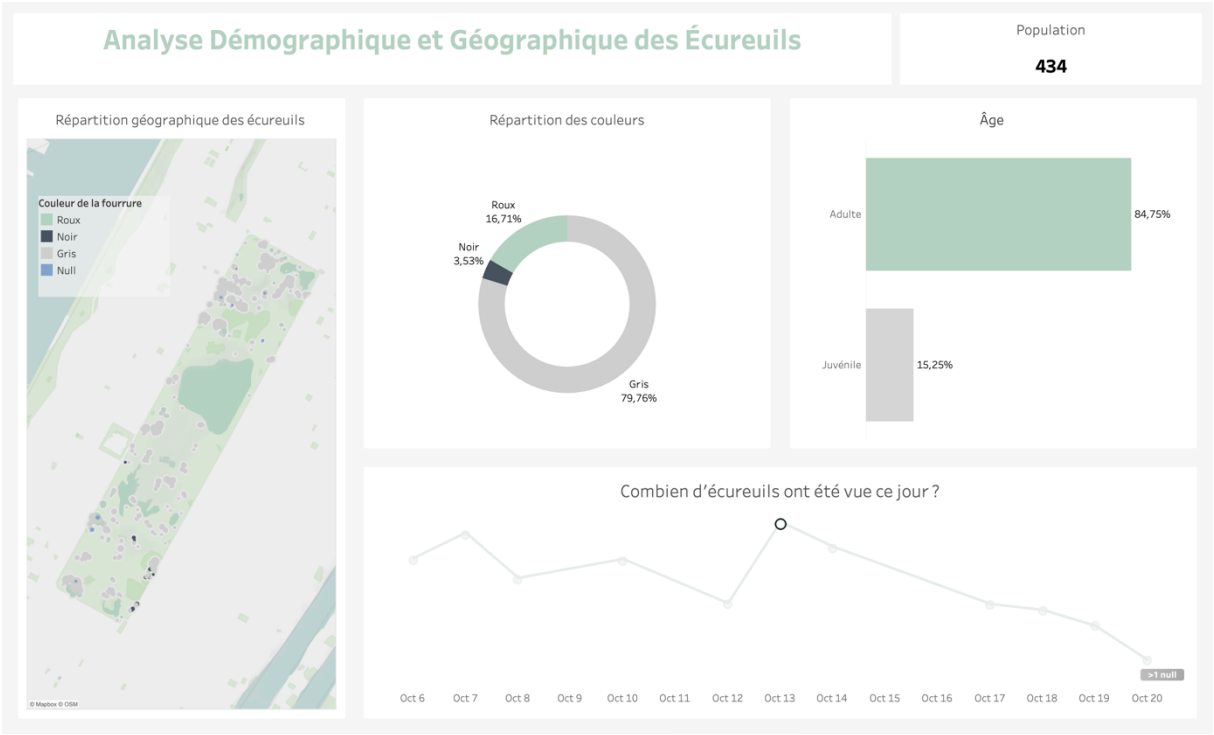
C. Démarche et résolution du problème

Pour pouvoir présenter notre tableau de bord, plusieurs étapes ont été nécessaires. Tout d'abord, il a fallu comprendre le jeu de données et déterminer ce que nous voulions mettre en avant. Ensuite, nous avons procédé à une étape de préparation des données afin d'assurer que chaque colonne avait le bon type de données. Ce n'est qu'après ces étapes que nous avons pu commencer à créer les visuels.

À cette fin, chaque membre du groupe était responsable de produire une représentation visuelle qui serait ensuite ajoutée au tableau de bord.

D. Acquis

Ce projet m'a permis de prendre en main le logiciel Tableau et appliquer les concepts théoriques vues en cours sur un jeu de données réel. De ce fait j'ai acquis une meilleure compréhension du logiciel mais surtout la dextérité de créer un dashboard.



SAÉ 2-06 Analyse de données, reporting et datavisualisation

A. Présentation du projet

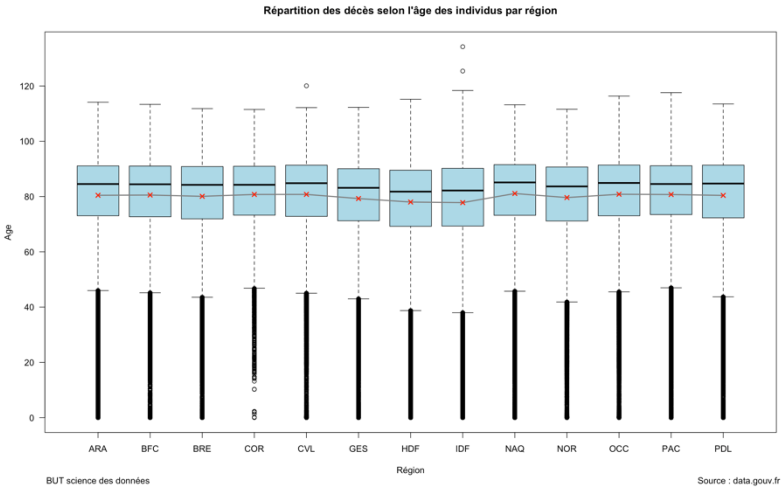
Le projet « Analyse de données, reporting et datavisualisation » est une étude portant sur les décès survenus en France au cours des années 2019 à 2022. L'objectif principal de cette étude était de mettre en lumière des informations à partir d'un large ensemble de données. Pour y répondre, nous avons utilisé différents types de représentations graphiques, telles que des histogrammes, des diagrammes à barres empilées, des boxplots, etc.

B. Organisation du travail

Ce projet a été réalisé par groupe de 3 étudiants. Chaque membre du groupe avait des représentations graphiques à réaliser en R. Pour mener à bien ce projet nous nous sommes servis de l'ensemble des ressources R étudié depuis le début de l'année et des cours de statistique descriptive. Le projet à rendre était divisé en trois parties : le script R, le rapport PDF fait en LaTeX et un code Marp pour la présentation orale de nos analyses.

C. Démarche et résolution du problème

Le point clé de notre analyse consistait à répondre à la question suivante pour chaque graphique : « Quelle représentation graphique serait la plus pertinente pour visualiser nos données ? ». Pour ce faire, nous avons étudié le type de nos variables (quantitatives ou qualitatives). Une fois cette étape terminée, nous avons sélectionné les graphiques appropriés que nous avons ensuite testés afin d'évaluer leur facilité de compréhension et leur capacité à mettre en évidence toutes les données que nous souhaitions présenter. De plus, pour améliorer notre travail, une fois un graphique terminé, nous l'analysions en groupe afin de mettre en lumière les points manquants et de définir les axes d'amélioration de ce dernier. Cette méthode de travail s'est révélée intéressante, car elle nous a permis de mieux comprendre les informations contenues dans les données et de les exploiter de manière plus efficace pour d'autres graphiques. Par exemple, sur le graphique ci-contre les segments reliant les moyennes sur les boxplots n'était pas présents avant ce débat au sein du groupe. Maintenant, qu'ils sont représentés sur notre graphique, il est plus explicite et nous permet d'émettre facilement une hypothèse de faible association entre les variables présentées.



D. Acquis

À travers ce projet, j'ai pu acquérir et développer un ensemble diversifié de compétences techniques, analytiques et de communication. Tout d'abord, en utilisant R, je commence à avoir de plus en plus d'automatisme me permettant d'éviter les erreurs, améliorer les rendus des visuels, être précis dans les titres et les indicateurs clés transmis par les graphiques. Ces représentations me permettent d'appliquer les concepts théoriques étudiés en cours et de m'en imprégner par cette mise en pratique. En rédigeant des rapports détaillés et en présentant mes résultats à mes camarades et mon jury, j'ai amélioré ma capacité à communiquer efficacement des informations complexes de manière claire et concise.

PP 1-01 Machine Learning à l'aide de la bibliothèque scikit-learn

En cours.

Forces et faiblesses : analyse réflexive

Compétences	Forces	Faiblesses
C1. Traiter des données à des fins décisionnelles	Python : - Gestion des listes, dictionnaires, chaînes de caractères. - Gestion des boucles - Gestion des fonctions - Matplotlib, Seaborn, Numpy SQL : - Requêtes simples - Création de modèle physique de données à partir d'un modèle logique de données	Python : - Pandas (exploration des colonnes et des fonctions associés à la bibliothèque) SQL : - Sous-requêtes - Types de jointures - Création de modèle logique et conceptuel de données
C2. Analyser statistiquement les données	R : - Créer les graphiques adaptés aux types d'études - Gestion d'erreur à l'aide du «Show Traceback »	SAS : - Compréhension des différentes procédures
C3. Valoriser une production dans un contexte professionnel	Tableau : - Création de Dashboard Soft skills : - Bonne communication orale (français et anglais) - Esprit critique	Tableau : - Gestion des erreurs - Maitrise de paramètres avancées (champs calculés) - Efficacité de réalisation du projet (Rapport équilibré entre la qualité du travail et le temps nécessaire à sa réalisation)

Axes d’amélioration

À travers l’ensemble de ces SAÉ et des ressources, j’ai pu enrichir mes compétences en traitement des données, en analyse statistique et en valorisation des données. Cette expérience m’a beaucoup appris, mais elle a également mis en évidence plusieurs points clé sur lesquels je dois encore m’améliorer.

Tout d'abord, en programmation Python, je dois améliorer ma compréhension de la bibliothèque Pandas, qui est très importante et pratique dans le domaine de la Data Science. Plus précisément, je dois me familiariser davantage avec les fonctions associées à cette bibliothèque. En ce qui concerne la programmation sur SAS, j’ai des difficultés à associer le type de procédure approprié en fonction de l’étude. En gestion de bases de données, je dois perfectionner ma réflexion sur la conception des MLD et MCD, qui sont les premières étapes de la construction d'une BDD SQL.
Deuxièmement, dans le bloc scientifique, en statistique inférentielle, je dois améliorer ma compréhension des estimations avec intervalle de confiance. En statistique descriptive, je dois retravailler la programmation des tableaux de contingence en R. Un point clé sur lequel j’avais de grosses difficultés, mais que j’ai déjà commencé à travailler, concerne les probabilités, notamment la compréhension des différentes lois de probabilité et leurs fonctions associées. Enfin, en mathématiques, même si j’ai compris les bases de l’algèbre linéaire, j’aimerais approfondir ce sujet, car je sais qu’il est essentiel pour mon futur professionnel.
Dernièrement, en ce qui concerne la valorisation des données, j’ai rencontré des difficultés avec la conception de visuels sur Tableau, ce qui me semble important à corriger. Enfin, ma prise de parole est trop rapide lorsque je présente des données, et cela nécessite également une amélioration.

Pour répondre à tous ces points que j’aimerais améliorer, je vais m’entraîner sur LeetCode.com pour la programmation, notamment en SQL et Python. De plus, j’ai repéré des formations sur Udemy.com en Python et R qui pourraient être intéressantes à suivre pendant les vacances scolaires. Pour le bloc scientifique, notamment pour les probabilités et l’algèbre linéaire, je compte m’entraîner sur les TD que nous avons déjà effectués et sur Bibmath.net. En ce qui concerne la valorisation des données, je compte m’exercer sur Tableau, en profitant des nombreuses vidéos intéressantes disponibles sur YouTube.

Pour conclure, les efforts investis dans les SAÉ et les ressources constituent des éléments clé pour notre préparation au monde professionnel. Cependant, je suis conscient qu'il est nécessaire d'intensifier mon travail personnel afin de mieux répondre aux défis et aux problématiques que je rencontrerai.