# Portefolio Data Raphaël Belleil

Pendant ma formation de Data Analyst, j'ai réalisé 10 projets portant sur différents sujets.

Voici une présentation succincte de chacun de ces projets, vous trouverez davantage de détails ainsi que le code et les fichiers des projets sur mon compte GitHub : <https://github.com/raphaelbelleil/Projets_Data> .

Les intitulés des projets sont :

* [Projet 2 : Faites une analyse des ventes pour un e-commerce](#_Projet_2_:)
* [Projet 3 : Créez et utilisez une base de données immobilière avec SQL](#_Projet_3_:)
* [Projet 4 : Réalisez une étude de santé publique avec Python](#_Projet_4_:)
* [Projet 5 : Optimisez la gestion des données d'une boutique avec Python](#_Projet_5_:)
* [Projet 6 : Analysez les ventes d'une librairie avec Python](#_Projet_6_:)
* [Projet 7 : Analysez des indicateurs de l'égalité femme-homme avec Knime](#_Projet_7_:)
* [Projet 7 bis : Analysez des indicateurs de l'égalité femme-homme avec Talend](#_Projet_7_bis:)
* [Projet 8 : Faites une étude sur l'eau potable avec Tableau](#_Projet_8_:)
* [Projet 9 : Produisez une étude de marché avec Python](#_Projet_9_:)
* [Projet 10 : Détectez des faux billets avec Python](#_Projet_10_:)
* [Projet Bonus : Anticipez les besoins en consommation de bâtiments](#_Projet_Bonus_:)

# Projet 2 : Faites une analyse des ventes pour un e-commerce

## Contexte

Votre diplôme obtenu, cela fait maintenant 1 an que vous travaillez en tant que Data Analyst au service Marketing du Grand Marché, une entreprise de grande distribution dans plusieurs secteurs (nourriture, biens de consommation et high tech). Elle gère un entrepôt et livre à domicile les commandes effectuées par les clients sur son site Internet.

Aujourd’hui, comme tous les débuts de mois, vous avez prévu de travailler sur le rapport mensuel des actions marketing de votre équipe. Pour cela, il vous faudra :

* Préparer la présentation des chiffres clés généraux à partir des graphiques déjà générés ;
* Préparer le rapport des données spécifiques aux clients affiliés sur Excel à partir des données directement.

## Compétences mise en œuvre

💡 Générer des graphiques adaptés aux types de données (courbe, diagramme en barre, histogramme, boîte à moustache, nuage de points...)

💡 Synthétiser des résultats à destination d'un client

💡 Interpréter les informations provenant d'un Dashboard

💡 Utiliser des fonctionnalités avancées d'Excel

## Environnement technique

* Excel

## Résultat

|  |  |
| --- | --- |
|  |  |

# Projet 3 : Créez et utilisez une base de données immobilière avec SQL

## Scénario

Vous êtes Data Analyst chez Laplace Immo, un réseau national d’agences immobilières. Le directeur général est sensible depuis quelque temps à l’importance des données, et il pense que l’agence doit se démarquer de la concurrence en créant une nouvelle base de données puis en la requêtant pour avoir des informations sur le marché de l'immobilier en France.

## Etapes du projet

* Extraire les données nécessaires à l’analyse.
* Définir des règles de gestion de nettoyage des bases de données (formatage, suppression des doublons, typage valeurs aberrantes…).
* Définir des règles de gestion de structuration des différentes bases de données entre elles.
* Création d'un schéma relationnel sur PostgreSQL en respectant la 3NF.
* Création des fichiers à intégrer dans les tables.
* Création de la base de données (tables, champs, clés primaires et étrangères, contraintes, index)
* Importation des données dans les tables.
* Requêtage des données (jointure, requêtes imbriquées, Windows fonctions, création de vues, champs calculés...) pour obtenir des informations sur le marché de l'immobilier en France.

## Compétences évaluées

💡 Créer une base de données

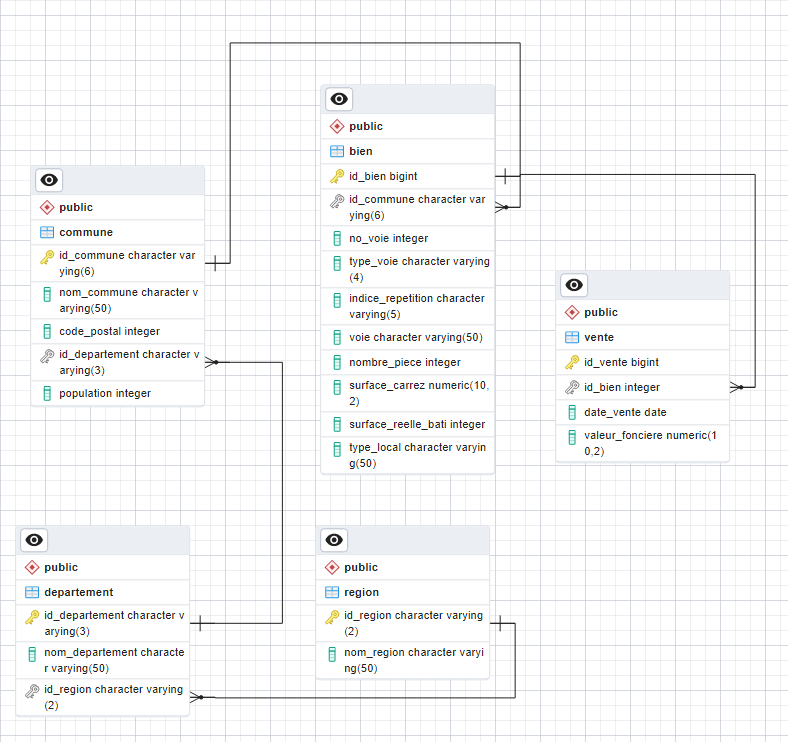
💡 Charger des données dans une base de données

💡 Effectuer des requêtes SQL pour répondre à une problématique métier

💡 Mettre à jour un catalogue de données

## Environnement technique :

* SQL, PostgreSQL, Excel, Algèbre relationnel



# Projet 4 : Réalisez une étude de santé publique avec Python

## Contexte

Le projet consiste à faire une analyse des ressources alimentaires au niveau mondial grâce à des données publiques de la FAO et à obtenir des informations sur les pays les plus dans le besoin au niveau de l'alimentation.

## Etapes

* Extraction des données du site de la FAO
* Nettoyage, préparation et exploration des données
* Analyse des données : pays, population, disponibilité alimentaire, aide alimentaire, sous-nutrition
* Utilisation des librairies spécialisées pour les traitements data (numpy, pandas)
* Utilisation des librairies spécialisées pour les visualisations (matplotlib, seaborn)
* Rédaction et présentation d'une méthodologie d'exploration et d'analyse des données
* Manipulation des DataFrames

## Compétences

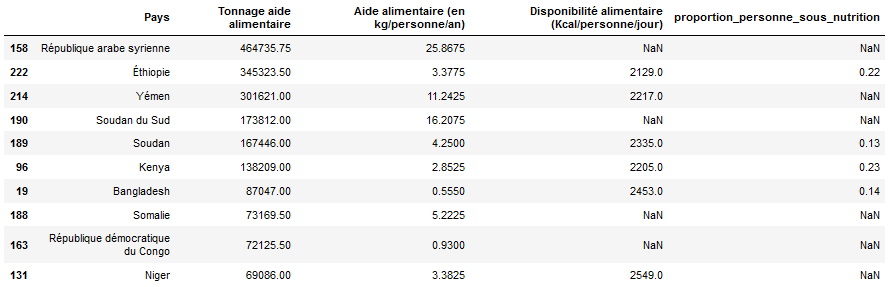
💡 Rédiger et présenter une méthodologie d'exploration et d'analyse des données

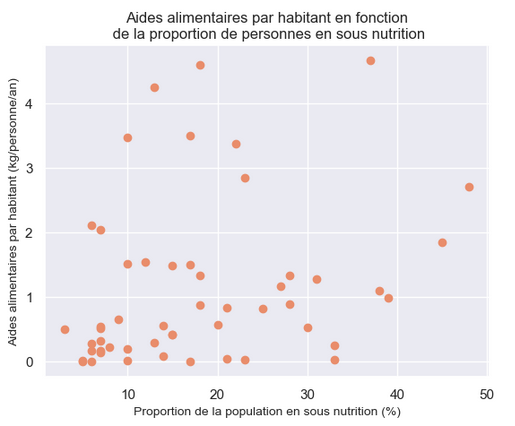
💡 Utiliser des librairies spécialisées pour les traitements data

💡 Manipuler des DataFrames

## Environnement technique

* Python (numpy, pandas, matplotlib, seaborn)





# Projet 5 : Optimisez la gestion des données d'une boutique avec Python

## Contexte

Le projet consiste à analyser les ventes en ligne d'un marchand de vin, de regrouper plusieurs sources de données (erp et web) et de réaliser l'analyse univariée des prix des produits.

## Etapes

* Nettoyage, préparation et exploration des données (gestion types, doublons, valeurs manquantes)
* Analyse univariée : décrire la répartition d'une variable
  + Mesures de tendance centrale (moyenne, mode, médiane, moyenne tronquée...),
  + Mesures de dispersion (variance, écart type, coefficient de variation),
  + Mesures de forme (asymétrie : skewness empirique, aplatissement : kurtosis)
  + Gestion des valeurs aberrantes (méthode de l'IQR et du zscore)
* Visualisation de l'analyse univariée (histogramme, boîte à moustaches)
* Tests statistiques de normalité (Kolmogorov-Smirnov, Shapiro-Wilk) d'une variable
* Transformation au logarithme et à la racine
* Tests paramétriques et non paramétriques
* Création de fonctions de synthèse automatisant toute l'analyse univariée

## Compétences

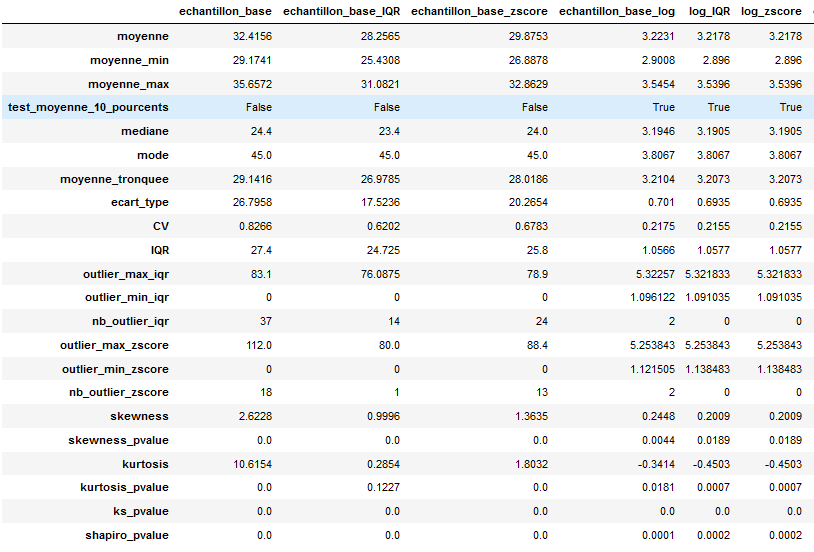
💡 Gérer les erreurs et les incohérences

💡 Classifier différents types de données

💡 Réaliser une analyse univariée pour interpréter des données

## Environnement technique :

* Python (numpy, pandas, matplotlib, seaborn, scipy)
* Jupyter notebook



# Projet 6 : Analysez les ventes d'une librairie avec Python

## Contexte

Le projet consiste à analyser les ventes en ligne d'une librairie.

Il se découpe en 2 parties :

* Analyse générale des données de ventes (Chiffre d'affaires, produits, clients...).
* Analyse du comportement des clients en ligne en étudiant les liens entre plusieurs variables (analyse bivariée).

## Etapes

### Analyse générale :

* Nettoyage, préparation et exploration des données (gestion types, doublons, valeurs manquantes, valeurs aberrantes)
* Analyse du chiffre d'affaires :
  + Année, mois, catégories de livres,
  + Moyennes mobiles, modèle ARIMA
  + Analyse de la saisonnalité
* Analyse du nombre de ventes
* Analyse des produits (classement, répartition, courbe de Lorenz et indice de Gini)
* Analyse des clients (classement, courbe de Lorenz et indice de Gini)

### Analyse du comportement des clients :

* Analyse bivariée entre 2 variables qualitatives (sexe et catégories de livres)
  + Tableau de contingence, Chi2, calcul des résidus standardisés
* Analyse bivariée entre 2 variables quantitatives (âge clients et montants des achats)
  + Visualisation (nuage de points, coefficient de Pearson, Régression Linéaire)
* Analyse bivariée entre une variable quantitative et qualitative (âge clients et catégories de livres)
  + ANOVA
* Réalisation de tests statistiques (Shapiro Wilk, Spearman, Levene, Kruskal-Wallis)

## Compétences

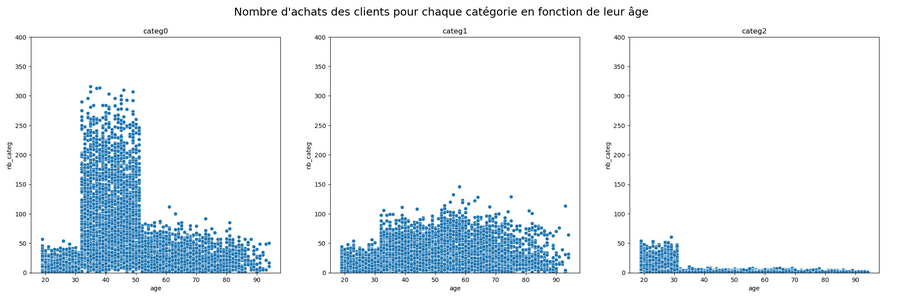
💡 Réaliser des tests statistiques

💡 Réaliser une analyse bivariée pour interpréter des données

💡 Analyser des séries temporelles

## Environnement technique

* Python (numpy, pandas, matplotlib, seaborn, scipy)
* Jupyter



# Projet 7 : Analysez des indicateurs de l'égalité femme-homme avec Knime

## Contexte

Le projet consiste à mettre en place un processus ETL (Extract, Transform, Load) qui automatise la création d'un rapport sur l'égalité Homme/Femme pour une entreprise. On utilisera le logiciel Knime, l'outil Diagnostic Egalité et données RH sur les employés.

## Etapes

* Mettre en place un processus ETL
* Collecter des données en respectant le RGPD
* Préparer des données pour l'analyse en respectant les normes internes à l’entreprise

Création d'un workflow sur Knime qui :

* Importe les données et les transfère vers une zone de préparation
* Nettoie et prépare les données
* Créer des indicateurs, tests statistiques et visualisations sur l'égalité Homme/Femme de l'entreprise
* Anonymise les données en respectant le RGPD
* Transfert le fichier pour des analyses sur Tableau

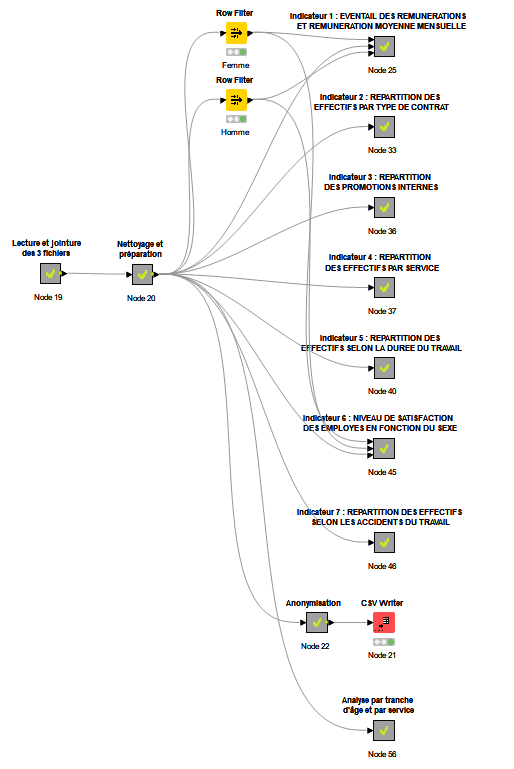
## Compétences

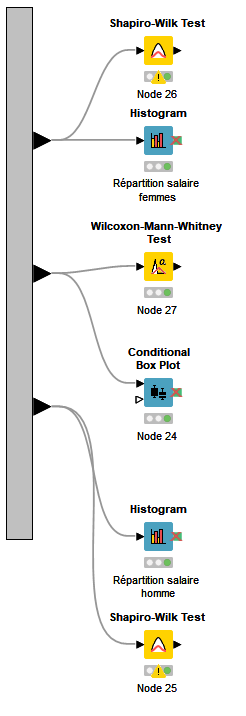
💡 Réaliser un processus ETL

💡 Collecter des données en respectant le RGPD

💡 Préparer des données en respectant les normes de l’entreprise

## Environnement technique

* ETL
* Knime
* Excel



# Projet 7 bis: Analysez des indicateurs de l'égalité femme-homme avec Talend

## Contexte

Le projet consiste à mettre en place un processus ETL (Extract, Transform, Load) qui automatise la création d'un rapport sur l'égalité Homme/Femme pour une entreprise. On utilisera le logiciel Talend, l'outil Diagnostic Egalité et données RH sur les employés.

## Etapes

* Mettre en place un processus ETL
* Collecter des données en respectant le RGPD
* Préparer des données pour l'analyse en respectant les normes internes à l’entreprise

Création d'un workflow sur Talend qui :

* Importe les données et les transfère vers une zone de préparation
* Nettoie et prépare les données
* Anonymise les données en respectant le RGPD
* Transfert le fichier pour des analyses sur Tableau

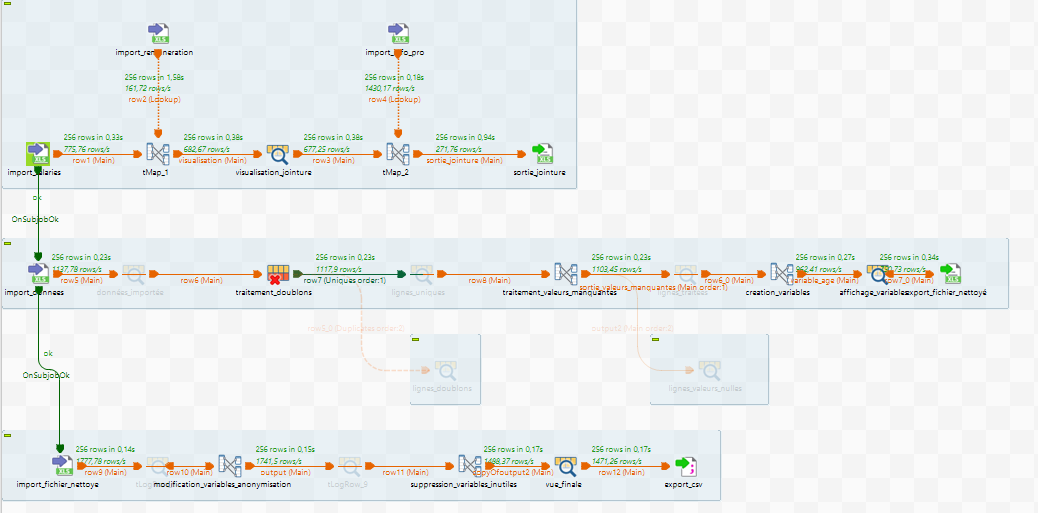
## Compétences

💡 Réaliser un processus ETL

💡 Collecter des données en respectant le RGPD

💡 Préparer des données en respectant les normes de l’entreprise

## Environnement technique

* ETL
* Talend
* Excel

# Projet 8 : Faites une étude sur l'eau potable avec Tableau

## Contexte

Le projet a pour objectif d'aider un investisseur à prendre une décision sur un projet d'infrastructure lié à la gestion de l'eau. Le tableau de bord devra permettre à l'investisseur :

* D'avoir une vision sur l’eau niveau mondial, continental et national
* De pouvoir prendre des décisions en termes d’investissement concernant la gestion de l’eau pour chaque pays
* De déterminer, à partir de certains critères, quels sont les pays qui doivent être aidés en priorité pour la création de nouveaux services, ceux qui doivent être modernisés et ce à qui on va demander des conseils

## Etapes

### Préparation des données 1ère méthode : Tableau Prep

* Créer un workflow de nettoyage et préparation des données sur Tableau Prep
* Export du fichier sur Tableau

### Préparation des données 2ème méthode : Python

* Nettoyage des données
* Restructuration des données et création d'une base de données relationnelle sur PostgreSQL avec pandas
* Liaison de la base de données à Tableau

### Réalisation du tableau de bord sur Tableau

* Création de champs calculés
* Choix des vues répondant aux problématiques
* Choix des types de visualisations, des indicateurs et des filtres
* Création d'une vue mondiale et continentale
* Création de 3 vues nationales permettant de filtrer les pays suivants différents critères

## Compétences

💡 Analyser un besoin client pour formuler des questions analytiques

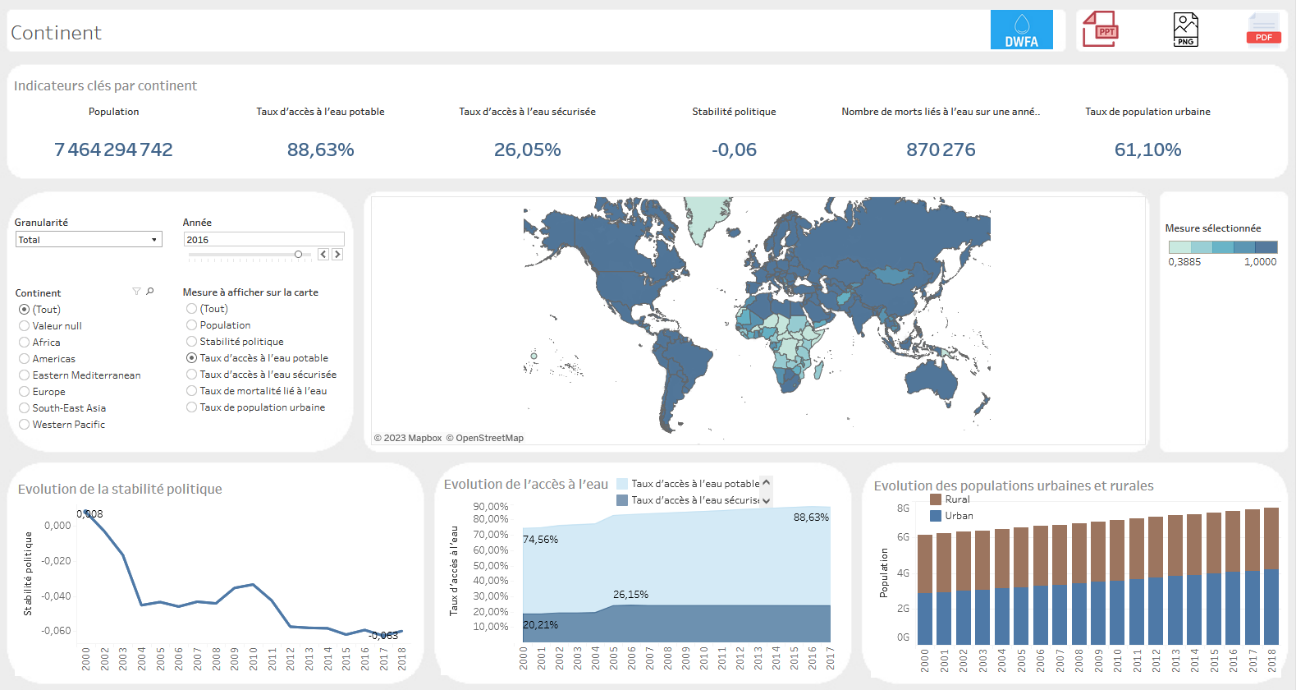
💡 Créer un tableau de bord répondant à des questions analytiques

💡 Générer des graphiques adaptés aux types de données

💡Synthétiser des résultats à destination d'un client

## Environnement technique

* Tableau Desktop, Tableau Prep, Tableau Public
* Excel, Python, PostgreSQL



# Projet 9 : Produisez une étude de marché avec Python

## Contexte

Le projet a pour objectif de sélectionner un pays pour implanter son entreprise de vente de poulet dans un nouveau pays. Il sera nécessaire de faire une étude de marché en sélectionnant des composantes économiques, sociales et politiques pour chaque pays et en analysant lequel convient le mieux pour s'y implanter. On réalisera cette étude de marché en faisant un clustering ascendant hiérarchique, une ACP et un KMeans.

## Etapes

### Création de son jeu de données :

* Analyse Pestel : choix des variables pertinentes à ajouter
* Import de fichier depuis plusieurs sources de données (données de base, FAO, banque mondiale)
* Exploration des données pour synthétiser des variables
* Nettoyage des données (gestion types, doublons, valeurs manquantes, valeurs aberrantes)
* Création d'un diagramme causale à partir des variables retenues

### Analyse :

* Réalisation de l'ACP (Analyse en composantes principales) : kmo, éboulis des valeurs propres, cercle des corrélations, représentation des individus sur le plan factoriel
* Réalisation du clustering ascendant hiérarchique
* Réalisation du KMeans
* Visualisation des clusters sur 2 et 3 dimensions
* Proposition de pays pour la nouvelle implantation de l'entreprise.

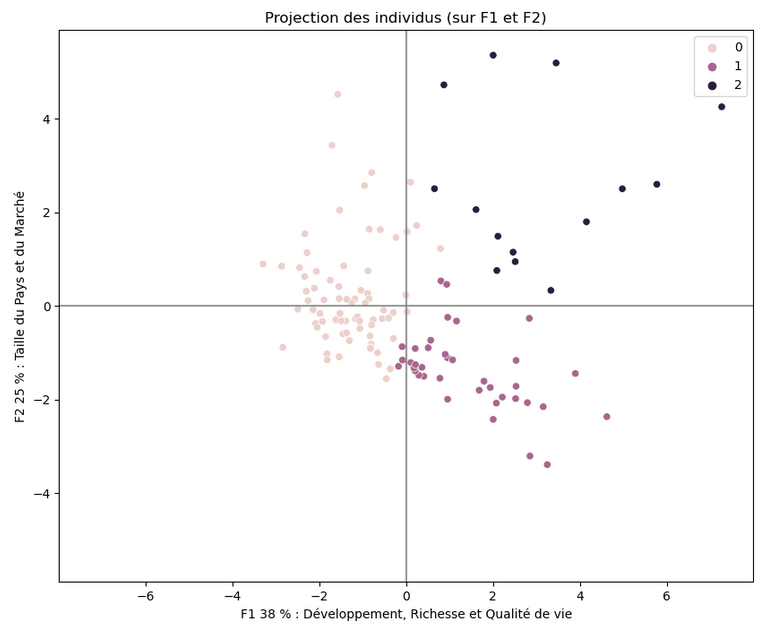
## Compétences

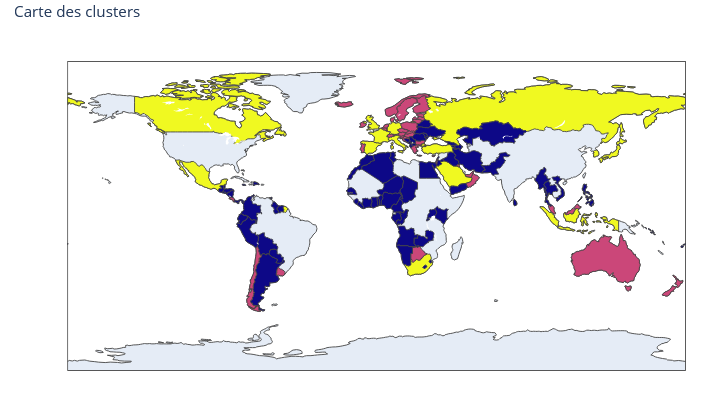
💡 Réaliser une étude de marché

💡 Effectuer un clustering et une ACP

💡 Explorer des données pour synthétiser des variables

## Environnement technique

* Python (numpy, pandas, matplotlib, seaborn, scipy, scikit-learn)
* Excel
* Algèbre linéaire

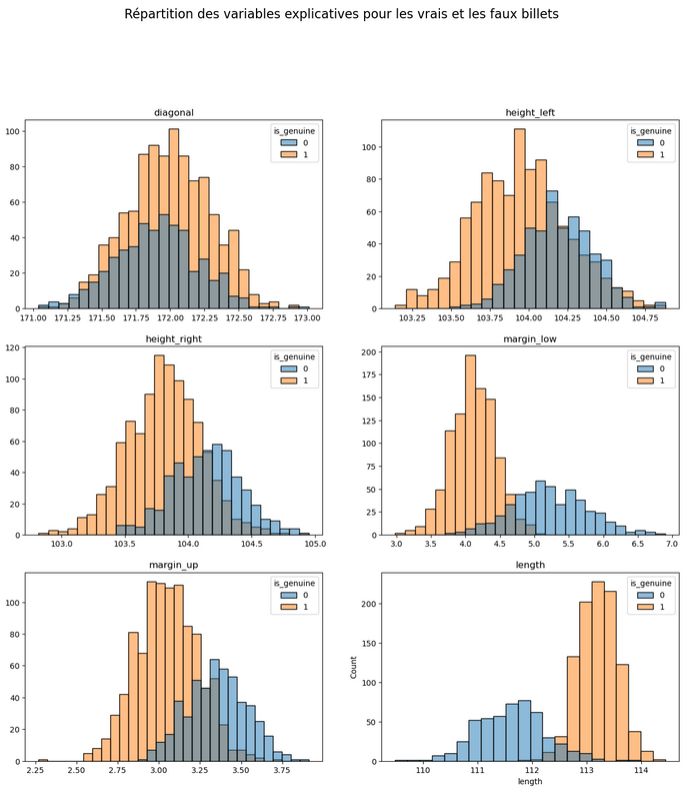


# Projet 10 : Détectez des faux billets avec Python

## Contexte

Le projet a pour objectif de mettre en place une modélisation pour identifier les faux billets grâce à leurs caractéristiques dimensionnelles. On imputera les valeurs manquantes grâce à une régression linéaire multiple puis on créera un modèle de régression logistique et un Kmeans pour la détection des faux billets.

## Etapes

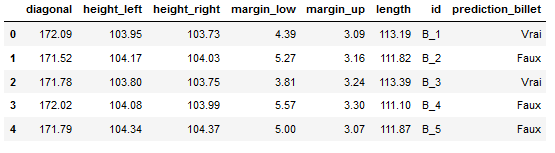
* Nettoyage des données
* Traitement des valeurs manquantes avec la régression linéaire
* Exploration (analyses univariées, matrice de corrélation, test d'hypothèse)
* Création du modèle de Régression Logistique, courbe ROC
* Kmeans
* Validation du modèle
* Création de la fonction de prédiction

## Compétences

* 💡 Réaliser une analyse prédictive
* 💡 Réaliser une régression logistique et linéaire
* 💡 Opérer des classifications automatiques pour partitionner les données

## Environnement technique

* Python (numpy, pandas, matplotlib, seaborn, scipy, scikit-learn)
* Algèbre linéaire
* Machine Learning, apprentissage automatique, Régression linéaire multiple, Régression Logistique, Kmeans



# Projet Bonus : Anticipez les besoins en consommation de bâtiments

## Contexte

Vous travaillez pour la ville de Seattle. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s’intéresse de près à la consommation et aux émissions des bâtiments non destinés à l’habitation.

Des relevés minutieux ont été effectués par les agents de la ville en 2016. Voici les données et leur source. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, vous voulez tenter de prédire les émissions de CO2 et la consommation totale d’énergie de bâtiments non destinés à l’habitation pour lesquels elles n’ont pas encore été mesurées. Vous sortez tout juste d’une réunion de brief avec votre équipe. Voici un récapitulatif de votre mission :

* Réaliser une courte analyse exploratoire.
* Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

## Compétences

💡 Mettre en place le modèle d'apprentissage supervisé adapté au problème métier

💡 Adapter les hyperparamètres d'un algorithme d'apprentissage supervisé afin de l'améliorer

💡 Transformer les variables pertinentes d'un modèle d'apprentissage supervisé

💡 Évaluer les performances d’un modèle d'apprentissage supervisé

💡 Mettre en place une démarche itérative d'optimisation du modèle (Transformation, Normalisation, Encodage, Imputation, Preprocessing, Feature Selection, Modélisation, Evaluation)

## Environnement technique

* Python (numpy, pandas, matplotlib, seaborn, scipy, statmodels, scikit-learn)
* Algèbre linéaire
* Machine Learning, apprentissage automatique, apprentissage supervisé
* Régression linéaire multiple, Ridge, Lasso, SVM, RandomForest, GridSearchCV