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 3 : Créez et utilisez une base de données immobilière avec SQL
- Projet 4 : Réalisez une étude de santé publique avec Python
- Projet 5 : Optimisez la gestion des données d'une boutique avec Python
- Projet 6 : Analysez les ventes d'une librairie avec Python
- Projet 7 : Analysez des indicateurs de l'égalité femme-homme avec Knime
- Projet 7 bis : Analysez des indicateurs de l'égalité femme-homme avec Talend
- Projet 8 : Faites une étude sur l'eau potable avec Tableau
- Projet 9 : Produisez une étude de marché avec Python
- Projet 10 : Détectez des faux billets avec Python
- Projet Bonus : Anticipez les besoins en consommation de bâtiments

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

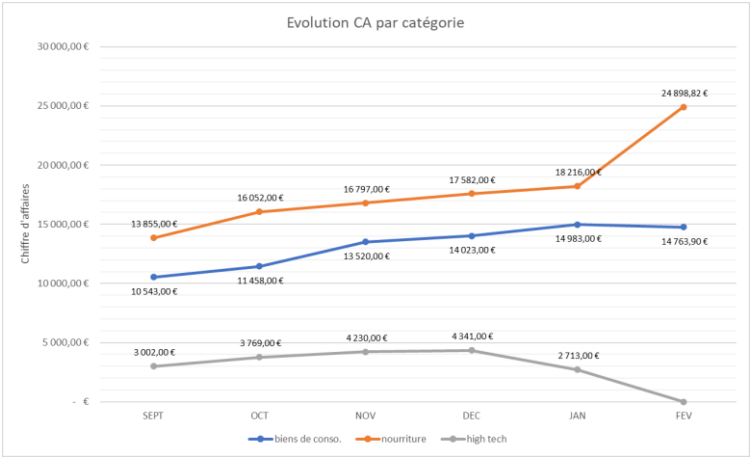
- Excel

Résultat

Analyse des ventes mensuelles des clients affiliés

CATEGORIE	SEPT	OCT	NOV	DEC	JAN	FEV	TOTAL
biens de conso.	10 543,00 €	11 458,00 €	13 520,00 €	14 023,00 €	14 983,00 €	14 763,90 €	79 290,90 €
nourriture	13 855,00 €	16 052,00 €	16 797,00 €	17 582,00 €	18 216,00 €	24 898,82 €	107 400,82 €
high tech	3 002,00 €	3 769,00 €	4 230,00 €	4 341,00 €	2 713,00 €	- €	18 055,00 €
TOTAL	27 400,00 €	31 279,00 €	34 547,00 €	35 946,00 €	35 912,00 €	39 662,72 €	204 746,72 €

TEMPS D'ACHAT	NB TRANSACTIONS	CA TOTAL
Inférieur à 4 min.	47	1 562,73 €
Supérieur à 9 min 30	91	7 577,32 €



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

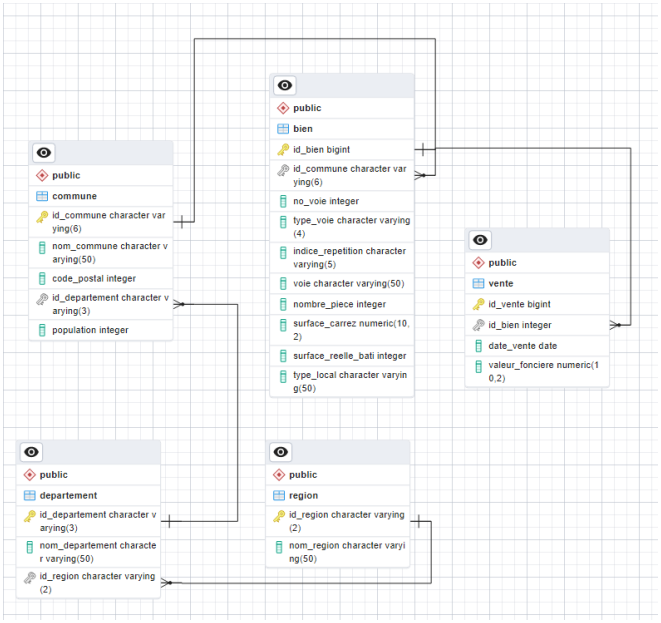
💡 Charger des données dans une base de données

💡 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

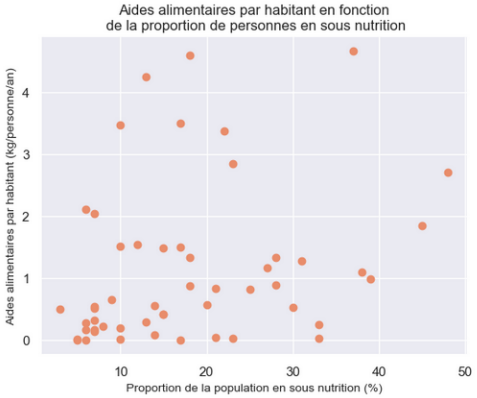
💡 Utiliser des librairies spécialisées pour les traitements data

💡 Manipuler des DataFrames

Environnement technique

- Python (numpy, pandas, matplotlib, seaborn)

	Pays	Tonnage aide alimentaire	Aide alimentaire (en kg/personne/an)	Disponibilité alimentaire (Kcal/personne/jour)	proportion_personne_sous_nutrition
158	République arabe syrienne	464735.75	25.8675	NaN	NaN
222	Éthiopie	345323.50	3.3775	2129.0	0.22
214	Yémen	301621.00	11.2425	2217.0	NaN
190	Soudan du Sud	173812.00	16.2075	NaN	NaN
189	Soudan	167446.00	4.2500	2335.0	0.13
96	Kenya	138209.00	2.8525	2205.0	0.23
19	Bangladesh	87047.00	0.5550	2453.0	0.14
188	Somalie	73169.50	5.2225	NaN	NaN
163	République démocratique du Congo	72125.50	0.9300	NaN	NaN
131	Niger	69086.00	3.3825	2549.0	NaN



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

	echantillon_base	echantillon_base_IQR	echantillon_base_zscore	echantillon_base_log	log_IQR	log_zscore
moyenne	32.4156	28.2565	29.8753	3.2231	3.2178	3.2178
moyenne_min	29.1741	25.4308	26.8878	2.9008	2.896	2.896
moyenne_max	35.6572	31.0821	32.8629	3.5454	3.5396	3.5396
test_moyenne_10_pourcents	False	False	False	True	True	True
mediane	24.4	23.4	24.0	3.1946	3.1905	3.1905
mode	45.0	45.0	45.0	3.8067	3.8067	3.8067
moyenne_tronquee	29.1416	26.9785	28.0186	3.2104	3.2073	3.2073
ecart_type	26.7958	17.5236	20.2654	0.701	0.6935	0.6935
CV	0.8266	0.6202	0.6783	0.2175	0.2155	0.2155
IQR	27.4	24.725	25.8	1.0566	1.0577	1.0577
outlier_max_iqr	83.1	76.0875	78.9	5.32257	5.321833	5.321833
outlier_min_iqr	0	0	0	1.096122	1.091035	1.091035
nb_outlier_iqr	37	14	24	2	0	0
outlier_max_zscore	112.0	80.0	88.4	5.253843	5.253843	5.253843
outlier_min_zscore	0	0	0	1.121505	1.138483	1.138483
nb_outlier_zscore	18	1	13	2	0	0
skewness	2.6228	0.9996	1.3635	0.2448	0.2009	0.2009
skewness_pvalue	0.0	0.0	0.0	0.0044	0.0189	0.0189
kurtosis	10.6154	0.2854	1.8032	-0.3414	-0.4503	-0.4503
kurtosis_pvalue	0.0	0.1227	0.0	0.0181	0.0007	0.0007
ks_pvalue	0.0	0.0	0.0	0.0	0.0	0.0
shapiro_pvalue	0.0	0.0	0.0	0.0001	0.0002	0.0002

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 en 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

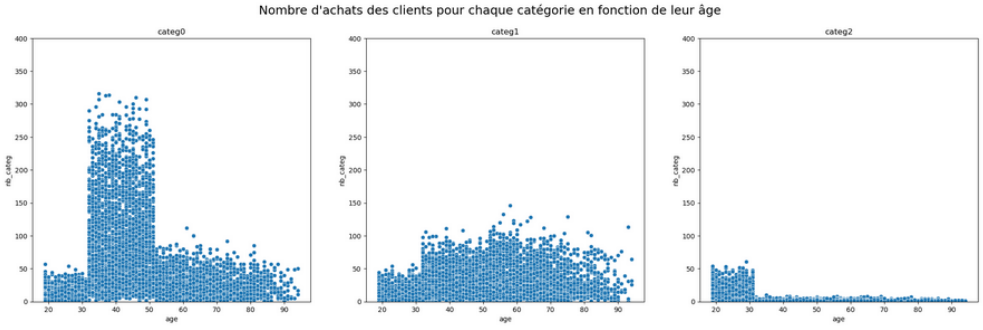
💡 Réaliser des tests statistiques

💡 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 :

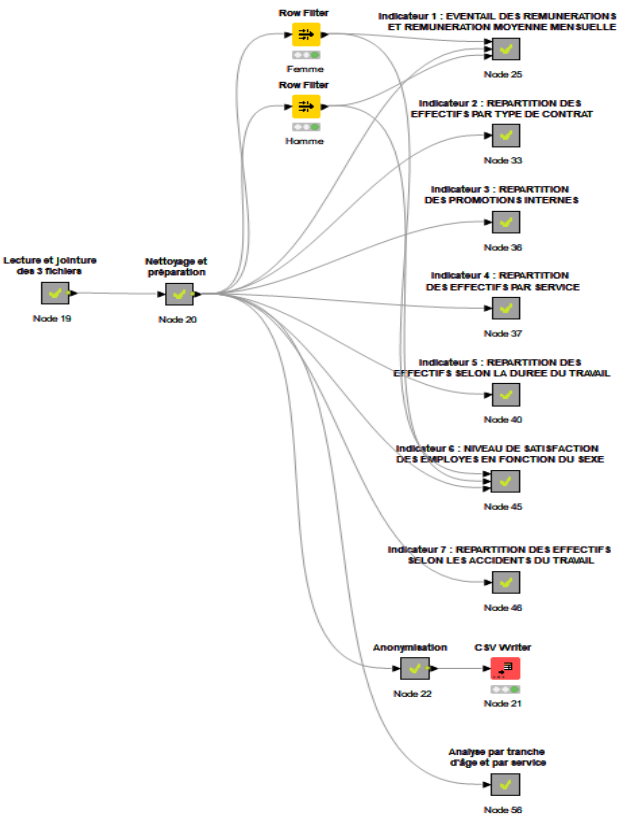
- Importe les données et les transfère vers une zone de préparation
- Nettoie et prépare les données
- Créer des indicateurs, tests statistiques et visualisations sur l'égalité Homme/Femme de l'entreprise
- Anonymise les données en respectant le RGPD
- Transfert le fichier pour des analyses sur Tableau

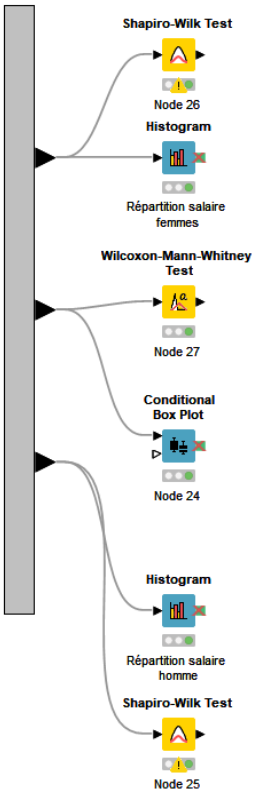
Compétences

- 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

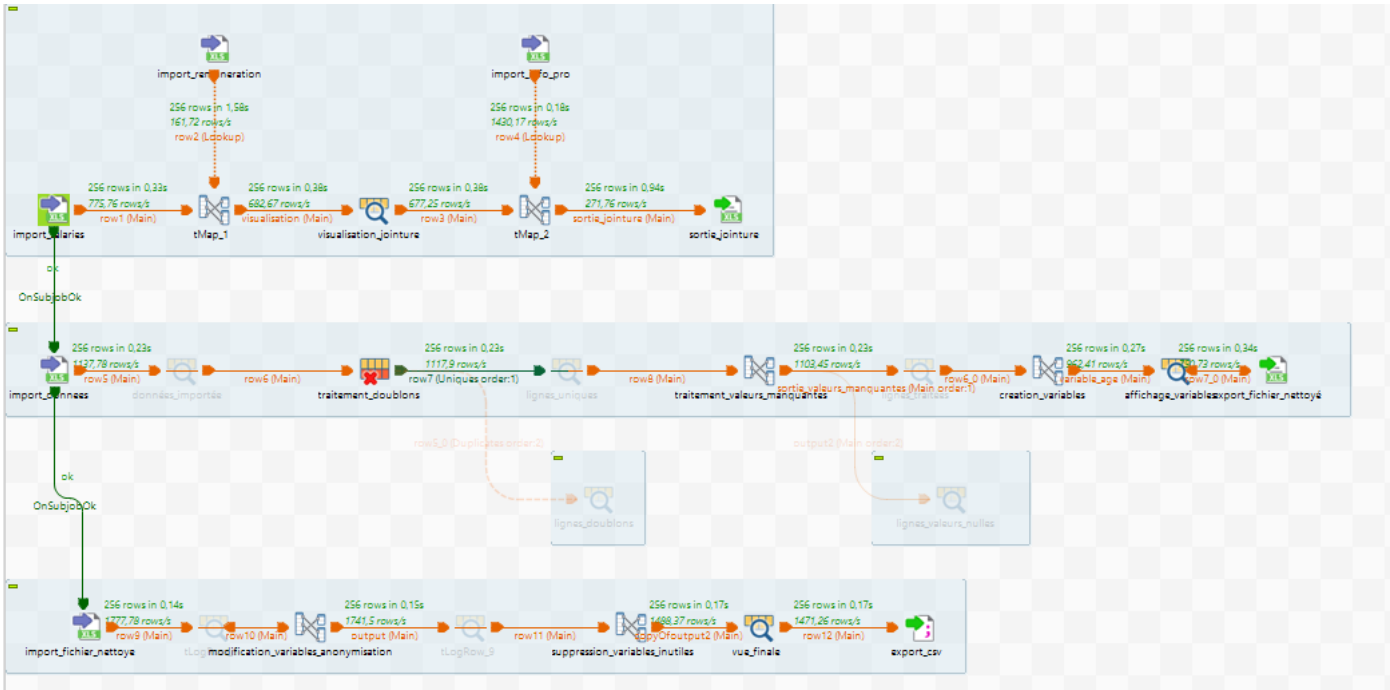
💡 Réaliser un processus ETL

💡 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

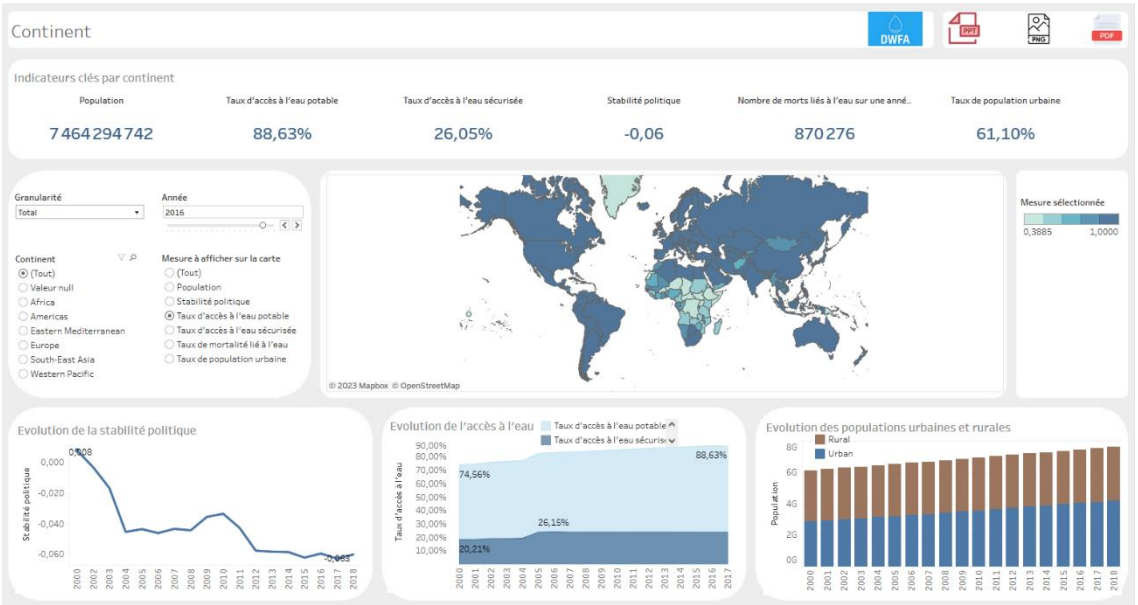
Créer un tableau de bord répondant à des questions analytiques

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

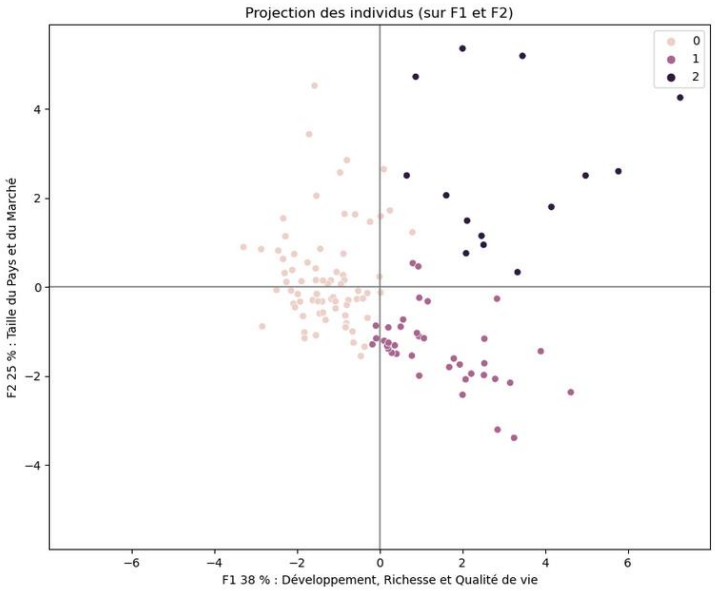
💡 Réaliser une étude de marché

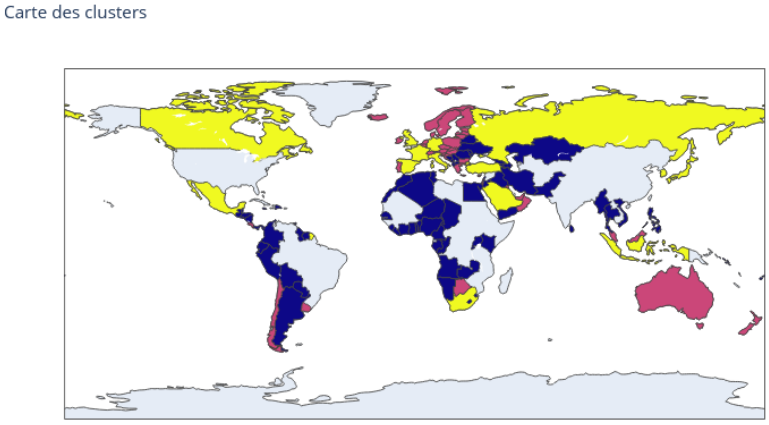
💡 Effectuer un clustering et une ACP

💡 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

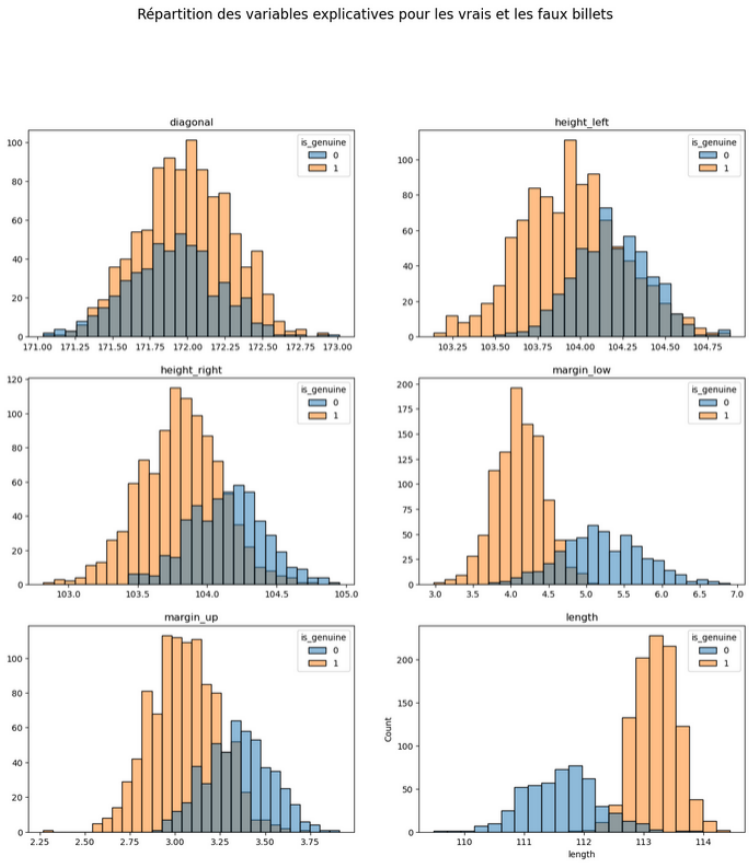
- Nettoyage des données
- Traitement des valeurs manquantes avec la régression linéaire
- Exploration (analyses univariées, matrice de corrélation, test d'hypothèse)
- Création du modèle de Régression Logistique, courbe ROC
- Kmeans
- Validation du modèle
- Création de la fonction de prédiction

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



	diagonal	height_left	height_right	margin_low	margin_up	length	id	prediction_billet
0	172.09	103.95	103.73	4.39	3.09	113.19	B_1	Vrai
1	171.52	104.17	104.03	5.27	3.16	111.82	B_2	Faux
2	171.78	103.80	103.75	3.81	3.24	113.39	B_3	Vrai
3	172.02	104.08	103.99	5.57	3.30	111.10	B_4	Faux
4	171.79	104.34	104.37	5.00	3.07	111.87	B_5	Faux

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