Cahier des Charges

Projet d'Analyse de la Pollution de l'Air

Sommaire

I- Introduction

Contexte

La pollution de l'air est un problème de santé publique et environnemental majeur. Les particules fines (PM2.5, PM10), les oxydes d'azote (NO2), l'ozone (O3), le dioxyde de soufre (SO2) et le monoxyde de carbone (CO) sont des polluants courants qui affectent la qualité de l'air. Ce projet vise à analyser les niveaux de pollution de l'air dans plusieurs régions en examinant les corrélations avec des facteurs géographiques et démographiques tels que l'altitude, la densité de population et le revenu moyen. Une analyse approfondie permettra de mieux comprendre les facteurs contribuant à la pollution de l'air et d'identifier des opportunités pour améliorer la qualité de l'air.

Objectif

Le but du projet est de créer une solution complète permettant de :

1. Récupérer et intégrer des données de pollution de l'air via une API.
2. Combiner ces données avec des informations démographiques et géographiques.
3. Analyser les corrélations et tendances entre la pollution et les facteurs géographiques et démographiques.
4. Présenter les résultats à l'aide d'un tableau de bord interactif pour faciliter la prise de décision.

II- Développement du Projet

1. Récupération des Données de Pollution de l'Air

1.1 Sélection de l'API

- **API Choisie :** OpenWeatherMap
  - **URL :** OpenWeatherMap API
  - **Description :** OpenWeatherMap fournit des données sur la qualité de l'air, y compris les indices de qualité de l'air (AQI) et les concentrations de divers polluants tels que PM2.5, PM10, O3, NO2, SO2, et CO. Ces informations sont essentielles pour évaluer les niveaux de pollution dans différentes régions.

1.2 Développement du Script de Collecte

- **Langage de Programmation :** Python
  - **Bibliothèque :** `requests`
  - **Description :** Le script Python utilise la bibliothèque `requests` pour envoyer des requêtes HTTP à l'API OpenWeatherMap. Les réponses, obtenues au format JSON, sont converties en un format structuré tel que des DataFrames pour une analyse ultérieure. Le script inclut des

fonctionnalités pour gérer les erreurs et les réponses inattendues, et il est conçu pour s'exécuter à intervalles réguliers de 2 Heures pour une collecte automatique des données.

1.3 Stockage Temporaire

- **Format Actuel :** Fichiers CSV
    - **Emplacement :** Local
    - **Description :** Les données récupérées sont actuellement stockées dans des fichiers CSV locaux, ce qui facilite la manipulation et l'analyse lors des premières étapes. Toutefois, le développement d'une base de données relationnelle, en particulier PostgreSQL, est en cours. Cette future base de données assurera une gestion plus robuste et évolutive des données, permettant une récupération efficace et une manipulation optimisée à long terme.

2. Données Fournies

Données Démographiques et Géographiques : Les données fournies incluent des informations démographiques et géographiques essentielles pour l'analyse. Ces données sont reçues sous forme de fichiers CSV ou de bases de données et comprennent :

- **Population** : Le nombre total d'habitants dans chaque région.
- **Densité de population** : Le nombre de personnes par kilomètre carré.
- **Taux d'urbanisation** : Le pourcentage de la population vivant dans des zones urbaines.
- **Revenu Moyen** : Le revenu moyen par habitant.
- **Niveau d'Éducation** : Le niveau moyen d'éducation de la population.
- **Altitude** : L'altitude en mètres au-dessus du niveau de la mer pour chaque région.
- **Proximité des sources de pollution** : La distance en kilomètres par rapport aux sources majeures de pollution.

3. Développement du Programme

- **Automatisation de la collecte :** Le programme automatise la collecte des données de pollution en utilisant le script développé. Il inclut des tâches programmées avec Apache Airflow pour assurer la récupération régulière des données. Les DAGS d'Airflow permettent de gérer l'exécution des tâches de manière efficace et orchestrée, garantissant ainsi l'extraction automatique et ponctuelle des données de l'API.
- **Stockage des Données :** Actuellement, les données récupérées sont stockées dans des fichiers CSV locaux. Cependant, le développement d'une base de données relationnelle (PostgreSQL) est en cours. La future base de données permettra une récupération facile et une gestion efficace des données, offrant une solution plus robuste et scalable pour le stockage et la manipulation des informations.

- **Documentation :** Une documentation est fournie, expliquant la configuration et l'utilisation du programme [https://github.com/Rjonathan03t/WEATHER-DAG/blob/main/README.md].

4. Nettoyage et Transformation des Données

4.1.Nettoyage des Données : Les données sont nettoyées pour éliminer les valeurs manquantes, corriger les erreurs et harmoniser les formats. Les étapes incluent :

- **Identification et traitement des valeurs manquantes :** On identifie les valeurs manquantes et on applique des techniques de gestion telles que l'imputation ou la suppression.
- **Correction des incohérences :** On corrige les erreurs et les incohérences présentes dans les données, telles que les valeurs aberrantes ou les doublons.
- **Harmonisation des formats de données :** On harmonise les formats des données pour garantir leur cohérence, par exemple, en convertissant toutes les dates dans un format standard ou en harmonisant les unités de mesure.

4.2.Transformation des Données : Les données sont transformées pour préparer l'analyse. Cela inclut :

- **Agrégation des données de différentes sources :** On agrège les données provenant de diverses sources pour créer un ensemble de données complet et cohérent.
- **Normalisation des valeurs pour faciliter la comparaison :** On normalise les valeurs pour rendre les différentes variables comparables, comme la mise à l'échelle des valeurs dans une plage commune.
- **Fusion des données de pollution avec les informations démographiques et géographiques :** On fusionne les données de pollution avec les informations démographiques et géographiques en utilisant des clés communes, telles que les localisations, pour créer un DataFrame intégré prêt pour l'analyse.

5.Analyse des Données

5.1.Exploration des Corrélations :

Les corrélations entre les niveaux de pollution et les facteurs géographiques et démographiques sont analysées à l'aide de techniques statistiques avancées. Ce processus commence par l'utilisation de Jupyter Notebook, un environnement interactif qui facilite l'exploration et la visualisation des données.

- Calcul des corrélations:

    - Les coefficients de corrélation entre les variables numériques sont calculés à l'aide de méthodes statistiques telles que le coefficient de corrélation de Pearson.
    - Le coefficient de corrélation de Pearson mesure la force et la direction de la relation linéaire entre deux variables continues, avec une valeur allant de -1 à 1. Un coefficient proche de 1 ou -1 indique une forte corrélation positive ou négative, respectivement, tandis qu'un coefficient proche de 0 indique peu ou pas de corrélation.

5.2.Identification des Tendances :

Les tendances significatives dans les données sont identifiées en analysant les données à travers le temps et l'espace. Cette analyse se déroule comme suit :

- Analyse Temporelle :
    - Les séries chronologiques des niveaux de pollution sont examinées pour détecter les variations saisonnières et les tendances à long terme. Les graphiques linéaires montrent les fluctuations des indices de qualité de l'air (AQI) et des concentrations de polluants au fil du temps.
- Analyse Spatiale :
    - Les variations de la pollution de l'air sont étudiées en fonction des différents facteurs géographiques et démographiques. Les graphiques de dispersion sont utilisés pour visualiser comment la pollution de l'air varie selon la densité de population, l'altitude, et la proximité des sources de pollution.

Utilisation de Jupyter Notebook :

- **Jupyter Notebook** est utilisé tout au long de ce processus pour sa capacité à exécuter du code Python de manière interactive, permettant une exploration rapide des données et une visualisation instantanée des résultats. Les cellules de code permettent de tester différentes hypothèses et méthodes d'analyse, tandis que les cellules Markdown facilitent la documentation des résultats et des interprétations.
- Les graphiques et visualisations sont intégrés directement dans le notebook, ce qui permet d'examiner les résultats en temps réel et d'ajuster les analyses en fonction des observations.

6. Automatisation avec Apache Airflow

Développement des Pipelines ETL : Apache Airflow est utilisé pour automatiser l'extraction, la transformation et le chargement des données. Les pipelines ETL sont conçus pour gérer les flux de données et assurer leur intégration fluide dans le système d'analyse.

Planification des Tâches : Les tâches sont planifiées pour s'exécuter à des intervalles réguliers, garantissant ainsi la mise à jour continue des données et des analyses.

Gestion des Dépendances : Les dépendances entre les différentes étapes du pipeline ETL sont gérées pour assurer que les processus s'exécutent dans le bon ordre et sans erreurs.

7. Création du Tableau de Bord

7.1. Développement des Visualisations

- **Types de Visualisations :** Les visualisations comprennent des graphiques en secteurs, des graphiques en anneau, des histogrammes, des tableaux, et d'autres types pertinents pour représenter les données. Ces visualisations permettent de présenter les tendances, les distributions, et les corrélations de manière claire et informative.
- **Outils :** Pour le développement des visualisations, nous utilisons Looker Studio [https://lookerstudio.google.com/u/1/reporting/3d76ec8b-618a-4fd4-bc56-c2b07a60f3ff/page/p_ilxm05qyjd/edit]. Cet outil offre une gamme de fonctionnalités pour créer des rapports et des tableaux de bord interactifs adaptés à l'analyse des données de pollution de l'air.

7.2. Choix de l'Outil de Visualisation

- **Critères de sélection :** L'outil de visualisation a été sélectionné en fonction de plusieurs critères clés. La facilité d'utilisation permet une adoption rapide et une création efficace des visualisations. L'intégration fluide avec les données assure une synchronisation précise et en temps réel des informations. De plus, l'outil offre des options de personnalisation variées, permettant d'adapter les visualisations aux besoins spécifiques du projet et de représenter les données de manière claire et pertinente.

7.3. Conception de l'Interface

- **Exigences :** L'interface utilisateur du tableau de bord est conçue pour être claire, interactive, et responsive. L'objectif est de créer une expérience utilisateur fluide qui facilite la navigation et l'interaction avec les données.
- **Fonctionnalités :** Les fonctionnalités incluent des filtres pour affiner les données affichées, ainsi que des vues personnalisables permettant aux utilisateurs de créer des affichages adaptés à leurs besoins d'analyse. Ces fonctionnalités améliorent l'interaction et la compréhension des données.

III- Conclusion

Résultats Attendus:

Le projet devrait permettre d'obtenir une compréhension approfondie des relations entre la pollution de l'air et divers facteurs géographiques et démographiques. En

combinant les données issues de l'API de pollution de l'air avec les informations démographiques et géographiques fournies, nous visons à identifier des tendances et des corrélations significatives. Ces résultats devraient fournir des bases solides pour développer des recommandations visant à améliorer la qualité de l'air dans les régions analysées.

Critères d'Évaluation:

Le succès du projet est évalué selon plusieurs critères :

- **Qualité des Données** : La précision et la fiabilité des données collectées et traitées.
- **Analyse et Pertinence** : La pertinence des analyses effectuées et la précision des corrélations identifiées entre la pollution et les facteurs étudiés.
- **Efficacité du Pipeline ETL** : L'efficacité et la fiabilité du pipeline ETL automatisé avec Apache Airflow, assurant une intégration fluide et régulière des données.
- **Tableau de bord interactif** : La fonctionnalité, l'esthétique, et l'interactivité du tableau de bord, qui permet aux utilisateurs d'explorer et de visualiser les résultats de manière claire et intuitive.

En fournissant un cadre détaillé et structuré, ce cahier des charges constitue un guide essentiel pour le développement du projet. Il assure que toutes les phases du projet, de la collecte des données à l'analyse et à la visualisation, sont réalisées avec précision et efficacité. Ce niveau de détail permet de garantir que le projet sera mené à bien selon les standards élevés requis, avec une attention particulière portée à chaque étape du processus.

Contributeurs

- **Jonathan - STD 22105**
- **Toki - STD 22106**
- **Rojo Tiana - STD 22107**