BEDJOU Celina

RAPPORT DE PROJET : MINI WAREHOUSE POUR L'ANALYSE DES DONNÉES DES TAXIS À NEW YORK

M1 - ISD

SOMMAIRE

INTRODUCTION

Ce projet a pour objectif de créer un mini entrepôt de données (mini warehouse) à partir de deux sources de données : un fichier Parquet contenant les données des taxis de New York pour l'année 2023, et un fichier CSV avec des données météorologiques issues de Kaggle. Les données ont été préparées, insérées dans une base de données SQLite, et un tableau de bord interactif a été développé avec Dash pour exposer les résultats sous forme de graphiques. L'objectif final est d'effectuer des transformations sur les données, telles que des jointures et des agrégations, afin d'analyser les relations entre la demande de taxis et les conditions météorologiques.

OBJECTIFS DU PROJET

Les objectifs du projet sont les suivants :

1. Traitement des données à l'aide de Python (pandas,) pour effectuer des transformations comme des jointures et des calculs d'agrégations.
2. Visualisation interactive des résultats via un tableau de bord développé avec Dash, en utilisant des bibliothèques comme Plotly.
3. Exposition des résultats sous forme de graphiques permettant d'explorer la relation entre les taxis et les conditions météorologiques.

DESCRIPTION DES DONNÉES

3.1. DATASET DES TAXIS (FICHIER PARQUET)

Le dataset des taxis provient des fichiers Parquet disponibles sur le site NYC TLC (Taxi & Limousine Commission). Ces fichiers contiennent des informations détaillées sur les courses de taxis effectuées à New York. Les principales colonnes du fichier Parquet sont les suivantes :

- pickup_date: La date et l'heure de la prise en charge de la course.
- total_amount: Le montant total de la course.
- tip_amount: Le montant du pourboire donné par le passager.
- trip_distance: La distance de la course.

3.2. DATASET MÉTÉOROLOGIQUE (FICHIER CSV)

Le fichier CSV provient de Kaggle et contient des données météorologiques pour New York pendant l'année 2023. Les colonnes principales de ce fichier sont :

- date: La date de l'observation météorologique.
- temperature: La température en degrés Celsius.
- humidity: Le taux d'humidité relatif.

3.3. BASE DE DONNÉES SQLITE

Les données des deux fichiers ont été extraites et préparées pour être insérées dans une base de données SQLite, après avoir effectuer les transformations suivantes :

1. Jointure entre les données de taxis et les données météorologiques en fonction de la date de la course et de l'observation météorologique.
2. Agrégations pour obtenir des statistiques sur les courses, telles que le nombre de courses, les pourboires totaux, et la température moyenne pour chaque période donnée (par heure, par jour, ou par mois).

DÉMARCHE

4.1. PRÉPARATION ET INSERTION DES DONNÉES DANS SQLITE

Les étapes suivantes ont été suivies pour préparer et insérer les données dans la base SQLite :

1. **Lecture des fichiers CSV et Parquet :**
   - Le fichier Parquet des taxis a été chargé à l'aide de pandas. Les données ont été nettoyées et formatées, en particulier la conversion de la colonne pickup_date au format datetime pour faciliter les jointures et les analyses temporelles.
   - Le fichier CSV météorologique a été également chargé et nettoyé, en s'assurant que les dates étaient correctement formatées pour effectuer la jointure.
2. **Insertion des Données dans SQLite :**
   - Après avoir nettoyé les données, elles ont été insérées dans la base de données SQLite à l'aide de pandas et SQLAlchemy.
   - Trois tables ont été créées dans la base de données : taxi_data, weather_data, et taxi_weather.
3. **Jointure et Transformation des Données :**
   - Une jointure inner a été effectuée entre les tables taxi_data et weather_data sur la base de la colonne date pour relier les données de taxis avec les informations météorologiques correspondantes.
   - Des agrégations ont été appliquées pour calculer des valeurs comme le nombre total de courses par heure, les pourboires totaux, et la température moyenne par période.

4.2. VISUALISATION DES RÉSULTATS AVEC DASH

Le tableau de bord Dash a été conçu pour afficher les résultats des analyses sous forme de graphiques interactifs. Les étapes suivantes ont été réalisées pour la visualisation :

1. Création de Graphiques :
    - Utilisation de Plotly pour créer des graphiques interactifs, y compris des graphiques en barres, des graphiques en ligne, et des graphiques circulaires.
    - Les visualisations comprennent :
        - Nombre de courses par mois.
        - Total des pourboires par mois.
        - Température moyenne par mois.
        - Relation entre Nombre de Courses et Pourboires
        - Relation entre Température et Nombre de Courses
        - Répartition des courses par jour de la semaine.
        - Répartition des moyens de paiement.
        - Répartition des moyens de heure.
2. Interface Utilisateur :
    - Le tableau de bord présente les graphiques sur la page d'accueil.
    - Un filtre global a été intégré, permettant à l'utilisateur de sélectionner un mois et de filtrer les données dans tous les graphiques.

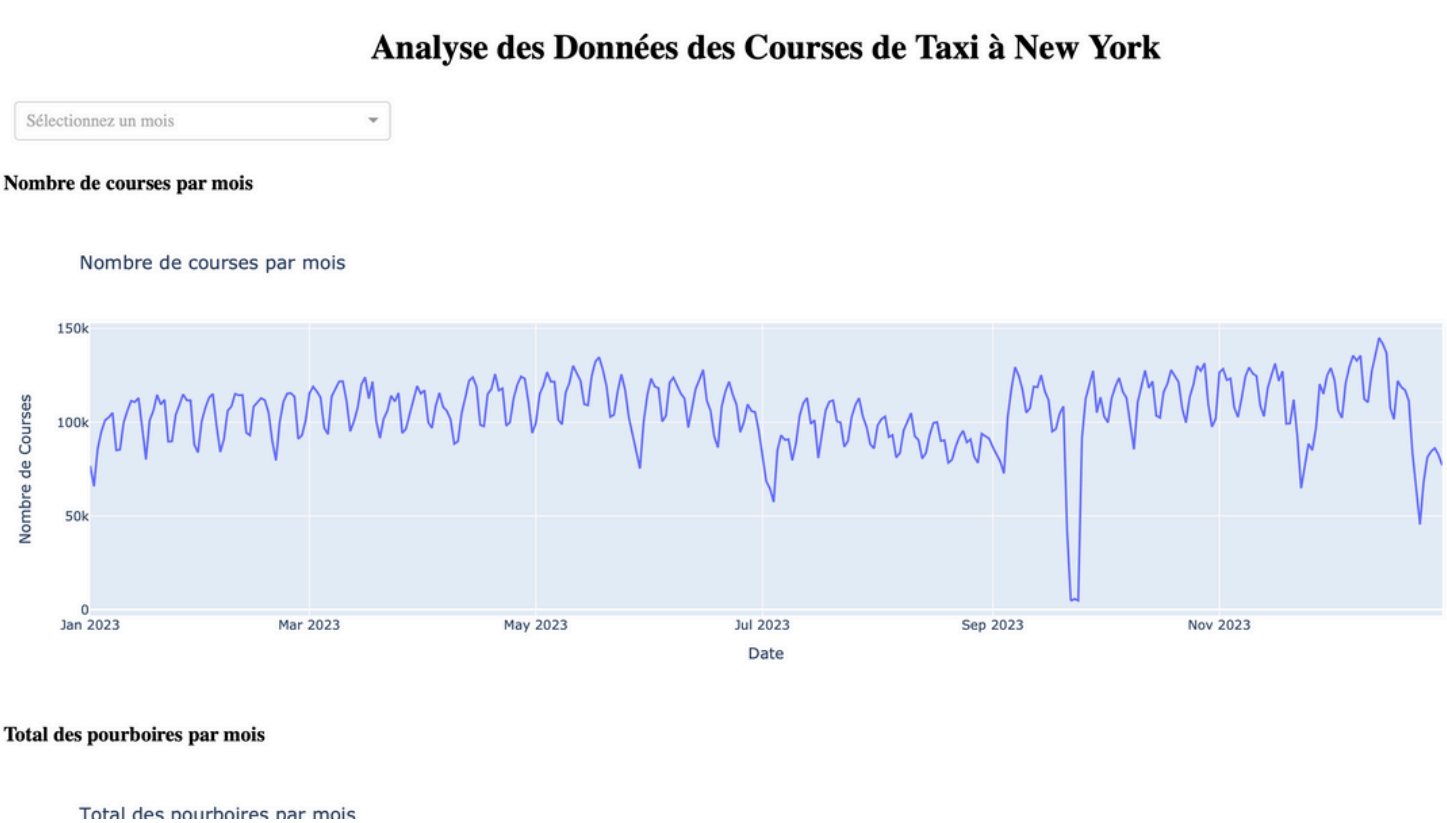

Figure 1 - L'interface de l'application

RÉSULTATS

Les résultats obtenus montrent plusieurs tendances intéressantes dans les données des taxis et des conditions météorologiques :

- Nombre de courses par heure : La demande de taxis est plus élevée aux heures de pointe le soir.
- Température et courses : L'analyse de la température montre une légère corrélation avec la demande de taxis. Les journées plus froides semblent avoir une demande plus élevée.
- Répartition des courses par jour de la semaine : Les courses sont réparties de manière inégale, avec des jours de semaine comme le vendredi et le jeudi affichant une demande plus élevée.

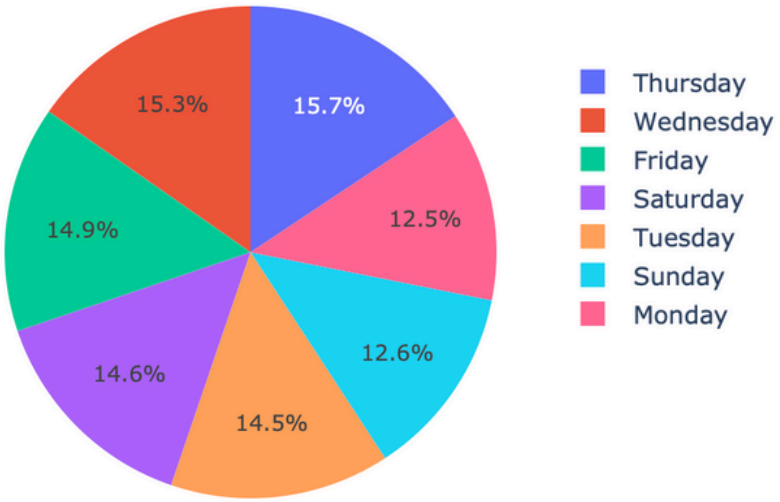

Figure 2 - Répartition des Courses par Jour de la Semaine

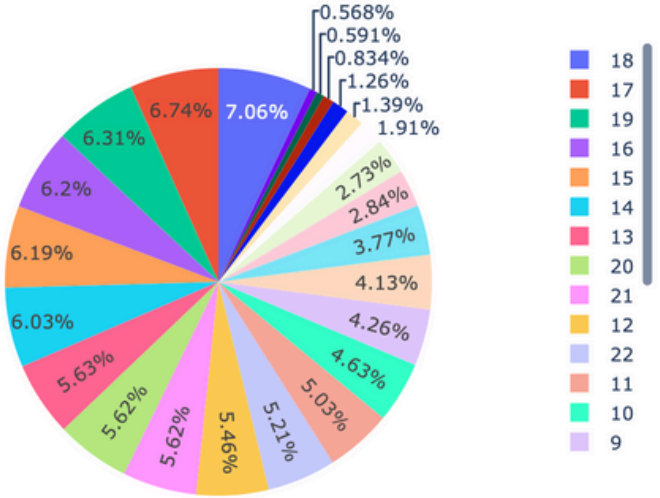

Figure 3 - Répartition des Courses par heure

CONCLUSION

Ce projet a permis de démontrer la puissance des outils de transformation et de visualisation de données, en intégrant des données de taxis et météorologiques dans un mini warehouse. L'utilisation de pandas et SQLite a facilité le traitement et l'agrégation des données, tandis que Dash et Plotly ont permis de créer une interface interactive pour l'analyse des résultats.
Les analyses ont permis de mettre en évidence des tendances intéressantes, telles que les variations de la demande de taxis en fonction de l'heure de la journée et de la température, ainsi que l'impact des conditions météorologiques sur les pourboires.