Projet BIUM

voyager avec 200 € et 2 masques
Y a-t-il une vivante dans le globe ?

Auteur
Youva Addad

Client
Laure Soulier

Table des matières

1 Résumé

Dans ce travail de business intelligence nous avons réalisé ce projet centré sur la décision du choix de la destination pour les vacances en fonction de divers critére.Le coût de vie de la ville où bien la situation sanitaire du pays entre autre nos critéres de sélection.
Dans ce rapport nous allons abordé les differents outils utilisé ainsi que les différentes analyses de données, la façon dont nous avons extrait les données mais en premier lieu nous allons parlé de l'analyse des besoins.
Nous allons parlé des differentes etapes que nous avons suivi afin d'aboutir a un résultat final qui soit le plus adéquat possible. Schéma suivi :

1. Définition des objectifs et des exigences
2. Le choix de la méthodologie et des outils à utiliser
3. Etablissement d'un programme de travail
4. Mise en place du programme de travail
5. Organiser le reporting et transmettre l'information

2 Introduction

Depuis plus d'une année maitenant le monde connait un ralentissement du au covid, les déplacements de chacun est fortement trés limité, il y a eu de plus un ralentissement economique dans le monde entier. Avec l'approche des vacances on peu ce demandé quel ville/pays choisir qui minimisera le risque lieé au covid et qui est le moins chére possible, toute en profitant biensur des loisirs et points d'interets diverse et variée.
Nous allons analysé ce point sous plusieur dimension pour pouvoir prendre une décision la meilleur qui soit.

3 Difficultés rencontrées

S'agissant d'un sujet à l'échelle mondiale nous avions passé une partie non négligeable a la collecte de données, nous avons collecté une quantité de données suffisente pour pouvoir traité du sujet, mais ceci était une partie trés sérer nous avions des diffultés a trouvé des données englobant le tout, des données traitant suffisamment de ce sujet, ce qui nous a permit tout de même d'exploité de nouvelle technique que nous détaillerons plus tard.

4 Données

Nous avons donc séléctionner les données qu'il nous fallait afin d'axer notre fait dessus, s'agissant d'un sujet qui traite du covid et de la distination la moins chere nous avons choisis de regardé :

- La situation sanitaire, le nombre cas ainsi que le nombre de vaccination.
- Le coût de la vie, typiquement le coût d'un restaurant, taxis, location ....
- Le prix des hotels, le logement est une étape décisif dans le choix d'une ville
- Prix des vols
- POI, les points d'intérêt pour toutes les villes du monde, Point d'intérêt pour les restaurants, tourisme, choses a voir ou a faire

(nous avions aussi des données sur la criminalité des villes mais on les a pas exploités).

1. owid-covid-data.csv ce fichier recense des informations sur le nombre cas, nombre de vaccination, nombre de mort pour chaque pays et depuis le début de la pandémie 24/02/2020 à 30/04/2021, ce fichier a été extrait depuis 'Johns Hopkins Coronavirus Resource Center' une université américaine.
2. Cost_of_life.csv ce fichier recensse pour une ville divers coût par exemple le cout d'un restaurant, le coût d'un café, coût fruit & légume, coût taxi ... etc, ce fichier a été scrapper depuis numbreo.com. Numbeo est une base de données mondiale accessible à tous sur les prix à la consommation déclarés, les taux de criminalité perçus, la qualité des soins de santé, entre autres statistiques.
3. hostel.csv ce fichier nous fournit pour un hotel son nom le prix la nuit, le nombre de lit ...etc. Nous avons scrapper ces données directement depuis Booking.
4. vols.csv ce fichier nous fournit le nom de l'aeroport la date du vol, le pays destination, latitude & longitude et le prix du vol ...etc de meme ce fichier a été srapper depuis booking.
5. POI est un dossier comportant les points d'intérêt, ou chaque fichier est un csv pour un pays des points d'intérêt, ils sont répartie par categorie & sous-categorie dans le fichier le nom international. Nous avons récolté ce fichier dans plusieurs source dans la majorité est les sources gouvernementale

la figure suivante montre quels sont les pays et ville que nous avons étudier :

FIGURE 1 – Les villes en point noir

5 Extract Transform Load

Nous avons précédemment cité les sources et les extractions des données mais nous avons pas parler de la transformation et le chargement des données. Pour cela nous avons utilisé :

1. Talend pour gérer nos données
2. Dataiku pour pouvoir traité, joindre, filtrer, remplacer, amelioré nos données.
3. Notebook python
4. Tableau Desktop pour l'intégration de nos données
5. Power BI

Nous avons donc nettoyé nos données avec ces outils et nous les avons converties aux formats de rapport qui conviennent, Nous avons donc appliqué les régles suivante :

- Nous avons sélectionner uniquement certaines colonnes typiquement les colonnes que nous avons jugé pertinente pour notre analyse. Les colonnes ayant des valeurs nulles par exemple nous les avons pas séléctionnes pour une meilleur cohérence. Par exemple la base covid ayany 41 caractéristique nous avons du enlevé celle qui ne server a rien.
- Nous avons aussi traduit les valeur codée par exemple les longitude et latitude ont été traduite en Point qui est une forme généralle pour pouvoir faire une map a la suite.
- Nous avons Dériver de nouvelle colonnes calculée par combinaison d'autre caractéristique. par exemple pour la base coût de la vie nous avons créer une caratéristique regroupant la somme des prix de transport, de restauration, Sim prépayé.
- Nous avons trier les données en fonction de certaine colonnes pour améliorer les performances de recherche. Par exemple dans le cost of living nous les avons trier par pays.
- Nous avons joint des données provenant de plusieurs base de données, cette partie nous avons permit de fusionnée et dédupliquer les données. Par exemple dans la base cost of living nous avons les villes mais pas les coordonnées nous avons donc opérer une jointure pour pouvoir trouvé les coordonnées.
- Nous avons aggréger les données par cumul, moyenne, somme. cette étape nous a permit de ne pas prendre toute les données mais seulement l'aggrégation.
- Nous avons aussi utilisé la transposition les données étant dans le mauvais sens.
- Nous avons fractionner des colonnes en plusieurs colonnes. par exemples dans les POI nous avons le nom du monument en internationnale séparer du nom avec langue nationnalle.
- Nous avons utilisé le select distinct pour pouvoir ne prendre qu'une seul fois un attribut par exemple.
- Nous avons créer de nouvelle base, par exemple pour la date, nous avons donc fait un split de dates pour pouvoir avoir une plage bien précise.

Avec tout ces traitement nous avons définie nous table et data werehouse. Nous détaillerons plus tard comment nous avons créer le datawerehouse et la façon dont nous l'avons alimenter.

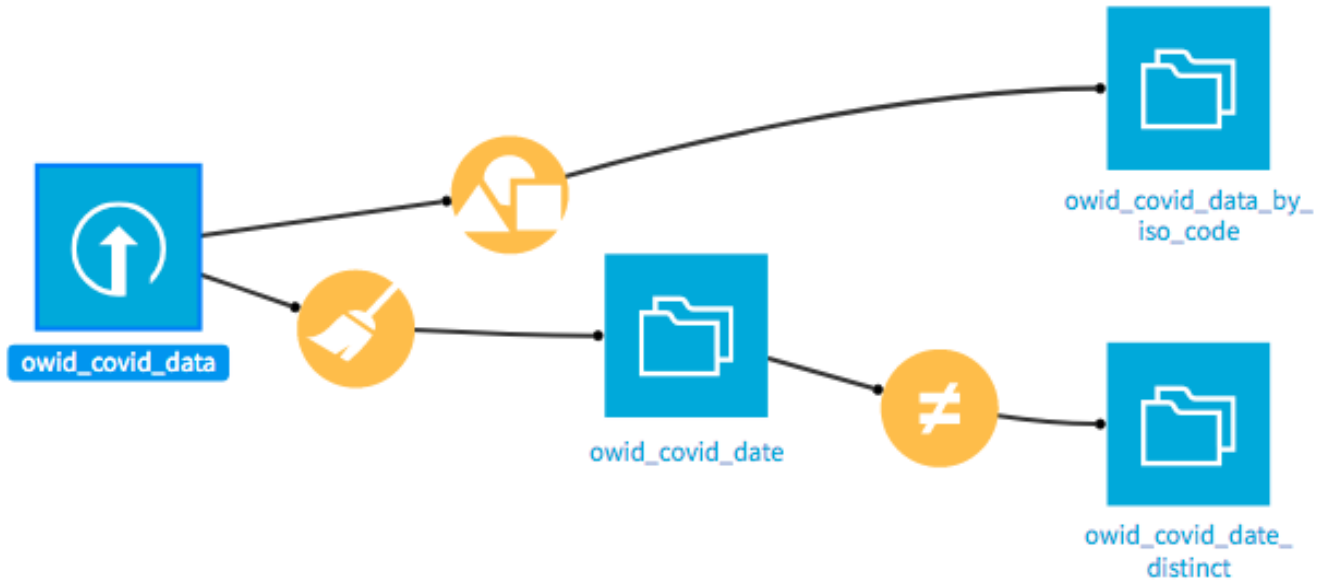

FIGURE 2 – Traitement des données covid

Ici donc nous avons deux paths, un pour la préparation des données (enlever les colonnes pas utiles ...Etc) et le select les valeurs distinct. De l'autre côté du chemin nous avons séléctionné les iso_code de pays.

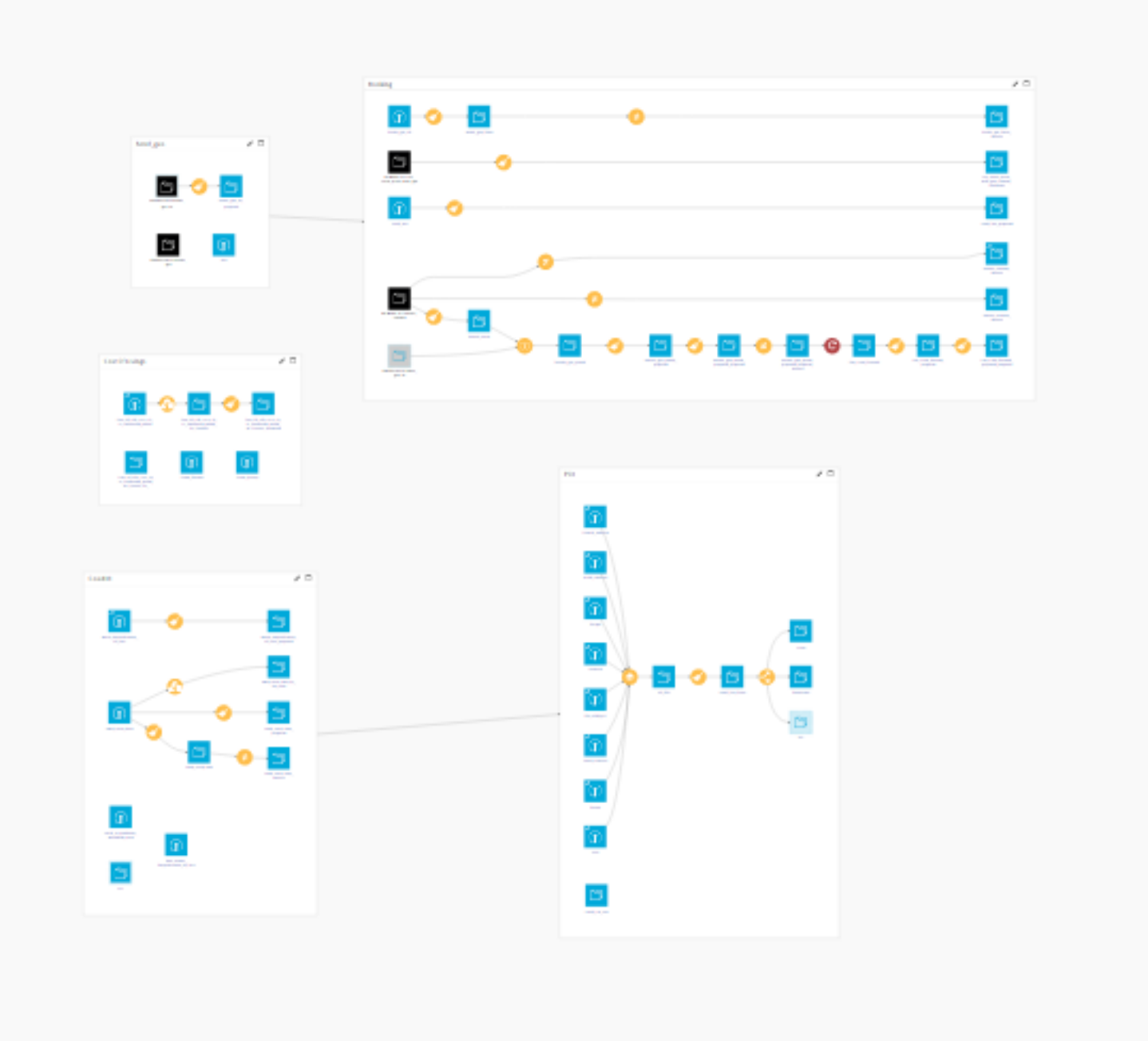

FIGURE 3 – Général des traitements effectué

6 Modélisation

6.1 Problématique

Nous souhaitant pouvoir prendre une décision de voyage pour cette été. Pour pouvoir répondre efficacement a cette problématique notre schéma doit donc répondre :

- le coût d'une ville, nourriture, transport et déplacement...etc
- la situation sanitaire en fonction du temps du nombre de vaccin, du nombre de test, et le nombre de cas.
- le prix d'un vol a une destination en fonction du temps.
- le prix d'un hotel d'une destination en fonction du temps.

6.2 Schéma Conceptuelle

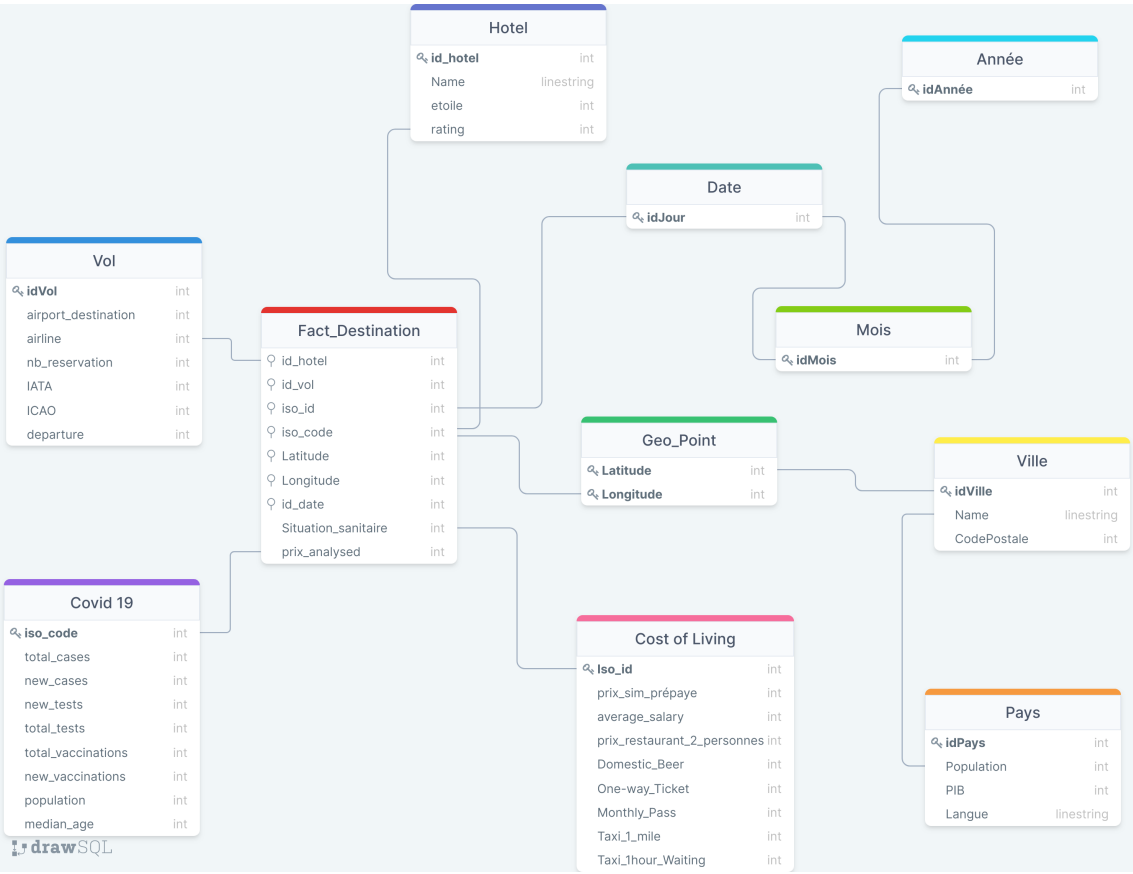

FIGURE 4 – Schéma en flocon

6.3 Schéma Normalisé

- Cost_of_living(iso_id, prix_sim_prepay, average_salary, prix_restaurants_2_personnes, Domestic_beer, ticket, monthly_pass, taxi_1_mile, taxi_1_hour).
- Covid 19(iso_code, total_cases, new_cases, new_tests, total_tests, total_vaccinations, new_vaccinations, population,median_age).
- Hotel(id_hote, Name,etoile, rating).
- Vol(id_vol, airport_destination, airline, nb_reservation, IOTA, ICAO, departure).
- Date(id_jour,#id_mois).
- Mois(id_mois,#Année).
- Année(id_année).
- Geo_Point(Latitude,Longitude,#id_ville).
- Ville(id_ville,#idpays,CodePostal).
- Pays(idPays,Population,PIB,Langue).
- Fact_Destination(#iso_id, #iso_code, #id_hote, #id_vol, #id_jour, #Latitude, #Longitude, Situation_sanitaire, prix_analysed)

7 Outil

1. **Talend** : Nous a permit de nettoyé et d'intégrer nous données afin de les exploités.
2. **Dataiku** : Nous a permit d'applique efficacement plein d'aggrégation, de jointure entre les données, unifié les données en splittant le tout, appliqué des filtres, créer de nouvelle colonne pré-calculée.
3. **Notebook python** : Nous a permit de collecté les données via le scrapping, afin de les visualisés pour déteminé la qualité des données collecter.
4. **Tableau Desktop** : Nous a permit de faire des visualisations pour pouvoir les intégrés dans le dashbord.

8 Analyse

Nous avons effectué nous analyse avec power bi et tableau desktop qui nous offre beaucoup de fonctionnalité, Nous avons effectué les analyses suivante :

1. **Analyse Cout de la vie :** pour pouvoir voir les pays ayant un cout de vie moins chére, nous avons donc commencé a intégré nos données dans tableau, nous avons choisis donc de ne voir que le coût d'un restaurant, d'un taxi, SIM prépayé, salaire moyen du pays ou de la ville.
2. **Analyse situation sanitaire :** nous avons effectue un comparatif du nombre de cas, test, vaccin réalisé pour un pays en suivant la granularité Année/Mois.
3. **Analyse vols :** Ici nous analysons les prix des vols pour voir la destination la plus propice pour voyagé.
4. **Analyse hotel :** Il s'agit ici d'avoir une idée généralle sur les tarifs appliqués dans ces villes en fonction du temps .

9 Prédiction

Dans cette section nous avons décidé de prédir le nombre moyen ou total de mort liée au covid dans un futur proche.

1. Lasso
2. Ridge
3. Linear Regression
4. SVM Regression

Nous avons décidé de travaillé avec les SVM Regression tout simplement parce qu'ils sont trés puissant, il sépare en suivant en maximisant une marge. De plus ils permettent d'apprend des solutions non lineair avec utilisation d'un noyau.
Donc les SVM's est un choix naturellement bon pour pouvoir faire des prédilection sur un future proche.