Projet BIUM

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

Laure Soulier

Table des mati`eres

1 R´esum´e **2**

2 Introduction **2**

3 Difficult´es rencontr´ees **2**

4 Donn´ees **3**

5 Extract Transform Load **5**

6 Mod´elisation **8**
6.1 Probl´ematique 8
6.2 Sch´ema Conceptuelle 8
6.3 Sch´ema Normalis´e 9

7 Outil **9**

8 Analyse **9**

9 Pr´ediction **10**

1 R´esum´e

Dans ce travail de business intelligence nous avons r´ealis´e ce projet centr´e sur la d´ecision du choix de la destination pour les vacances en fonction de divers crit´eres. Le choix de la ville o`u bien la situation sanitaire du pays entre autre nos crit´eres de s´election.
Dans ce rapport nous allons abord´e les differents outils utilis´e ainsi que les diff´erentes analyse de donn´ees, la fa¸con dont nous avons extrait les donn´ees mais en premier lieu nous allons parl´e de l'analyse des besoins.
Nous allons parl´e des differentes etapes que nous avons suivi afin d'aboutir a un r´esultat final qui soit le plus ad´equat possible. Sch´ema suivi :

1. D´efinition des objectifs et des exigences
2. Le choix de la m´ethodologie et des outils `a 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 ann´ee maitenant le monde connait un ralentissement du au covid, les d´eplacements de chacun est fortement tr´es limit´e, il y a eu de plus un ralentissement economique dans le monde entier. Avec l'approche des vacances on peu ce demand´e quel ville/pays choisir qui minimisera le risque lie´e au covid et qui est le moins ch´ere possible, toute en profitant biensur des loisirs et points d'interets diverse et vari´ee.
Nous allons analys´e ce point sous plusieur dimension pour pouvoir prendre une d´ecision la meilleur qui soit.

3 Difficult´es rencontr´ees

S'agissant d'un sujet `a l'´echelle mondiale nous avions pass´e une partie non n´egligeable a la collecte de donn´ees, nous avons collect´e une quantit´e de donn´ees suffisente pour pouvoir trait´e du sujet, mais ceci´etait une partie tr´es s´erer nous avions des diffult´es a trouv´e des donn´ees englobant le tout, des donn´ees traitant suffisamment de ce sujet, ce qui nous a permit tout de mˆeme d'exploit´e de nouvelle technique que nous d´etaillerons plus tard.

4 Donn´ees

Nous avons donc s´el´ectionner les donn´ees qu'il 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 regard´e :

- 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 ´etape d´ecisif dans le choix d'une ville
- Prix des vols
- POI, les points d'int´erˆet pour toutes les villes du monde, Point d'int´erˆet pour les restaurants, tourisme, choses a voir ou a faire

(nous avions aussi des donn´ees sur la criminalit´e des villes mais on les a pas exploit´es).

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 d´ebut de la pand´emie 24/02/2020à 30/04/2021,ce fichier a ´et´e extrait depuis 'Johns Hopkins Coronavirus Resource Center' une universit´e am´ericaine.
2. Cost_of_life.csv ce fichier recensse pour une ville diverses pâr exemple le cout d'un restaurant,le coût d'un caf´e,coût fruit & l´egume,coût taxi ... etc, ce fichier a ´et´e scrapper depuis numbreo.com. Numbeo est une base de donn´ees mondiale accessible `a tous sur les prix `a la consommation d´eclar´es, les taux de criminalit´e per¸cus, la qualit´e des soins de sant´e, 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 donn´ees 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 ´et´e srapper depuis booking.
5. POI est un dossier comportant les points d'int´erˆet, chaque fichier est un csv pour un pays des points d'int´erˆets sont r´epartie par categorie & sous-categorie dans le fichier le nom international. Nous avons r´ecolt´e ce fichier dans plusieurs source dans la majorit´e est les sources gouvernementale

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

Figure 1 – Les villes en point noir

5 Extract Transform Load

Nous avons pr´ec´edemment cit´e les sources et les extractions des donn´ees mais nous avons pas parler de la transformation et le chargement des donn´ees. Pour cela nous avons utilis´e :

1. Talend pour g´erer nos donn´ees
2. Dataiku pour pouvoir trait´e, joindre, filtrer, remplacer, amelior´e nos donn´ees.
3. Notebook python
4. Tableau Desktop pour l'int´egration de nos donn´ees
5. Power BI

Nous avons donc nettoy´e nos donn´ees avec ces outils et nous les avons converties aux formats de rapport qui conviennent, Nous avons donc appliqu´e les r´egles suivante :

- Nous avons s´electionner uniquement certaines colonnes typiquement les colonnes que nous avons jug´e pertinente pour notre analyse. Les colonnes ayant des valeurs nulles par exemple nous les avons pas s´el´ectionnes pour une meilleur coh´erence. Par exemple la base covid ayany 41 caract´eristique nous avons du enlev´e celle qui ne server a rien.
- Nous avons aussi traduit les valeur cod´ee par exemple les longitude et latitude ont ´et´e traduite en Point qui est une forme g´en´eralle pour pouvoir faire une map a la suite.
- Nous avons D´eriver de nouvelle colonnes calcul´ee par combinaison d'autre caract´eristique par exemple pour la base cout de la vie nous avons cr´eer une carat´eristique regroupant la somme des prix de transport, de restauration, Sim pr´epay´e.
- Nous avons trier les donn´ees en fonction de certaine colonnes pour am´eliorer les performances de recherche. Par exemple dans le cost ofliving nous les avons trier par pays.
- Nous avons joint des donn´ees provenant de plusieurs base de donn´ees, cette partie nous avons permit de fusionn´ee et d´edupliquer les donn´ees. Par exemple dans la base cost of living nous avons les villes mais pas les coordonn´ees nous avons donc op´erer une jointure pour pouvoir trouv´e les coordonn´ees.
- Nous avons aggr´eger les donn´ees par cumul, moyenne, somme. cette ´etape nous a permit de ne pas prendre toute les donn´ees mais seulement l'aggr´egation.
- Nous avons aussi utilis´e la transposition les donn´ees ´etant 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 s´eparer du nom avec langue nationnalle.
- Nous avons utilis´e le select distinct pour pouvoir ne prendre qu'une seul fois un attribut par exemple.
- Nous avons cr´eer de nouvelle base, par exemple pour la date, nous avons donc fait un split de dates pour pouvoir avoir une plage bien pr´ecise.

Avec tout ces traitement nous avons d´efinie nous table et data werehouse. Nous d´etaillerons plus tard comment nous avons cr´eer le datawerehouse et la fa¸con dont nous l'avons alimenter.

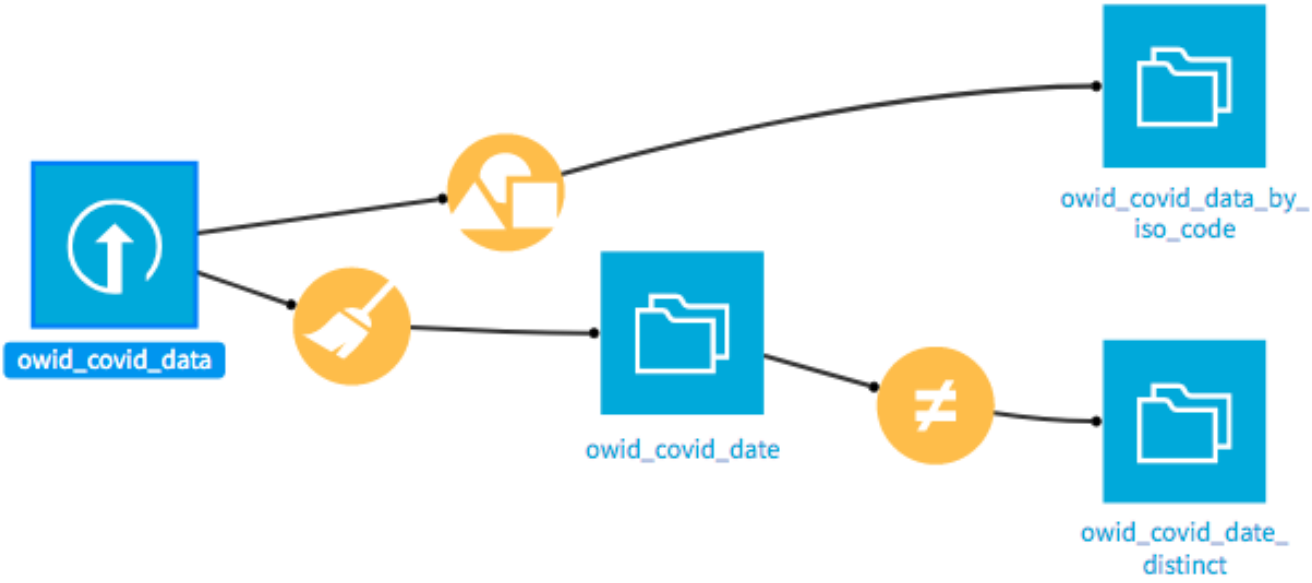

Figure 2 – Traitement des donn´ees covid

Ici donc nous avons deux paths, un pour la pr´eparation des donn´ees (enlever les colonnes pas utiles ...Etc) et le select les valeurs distinct. De l'autre cˆot´e du chemin nous avons s´el´ectionn´e les iso code de pays.

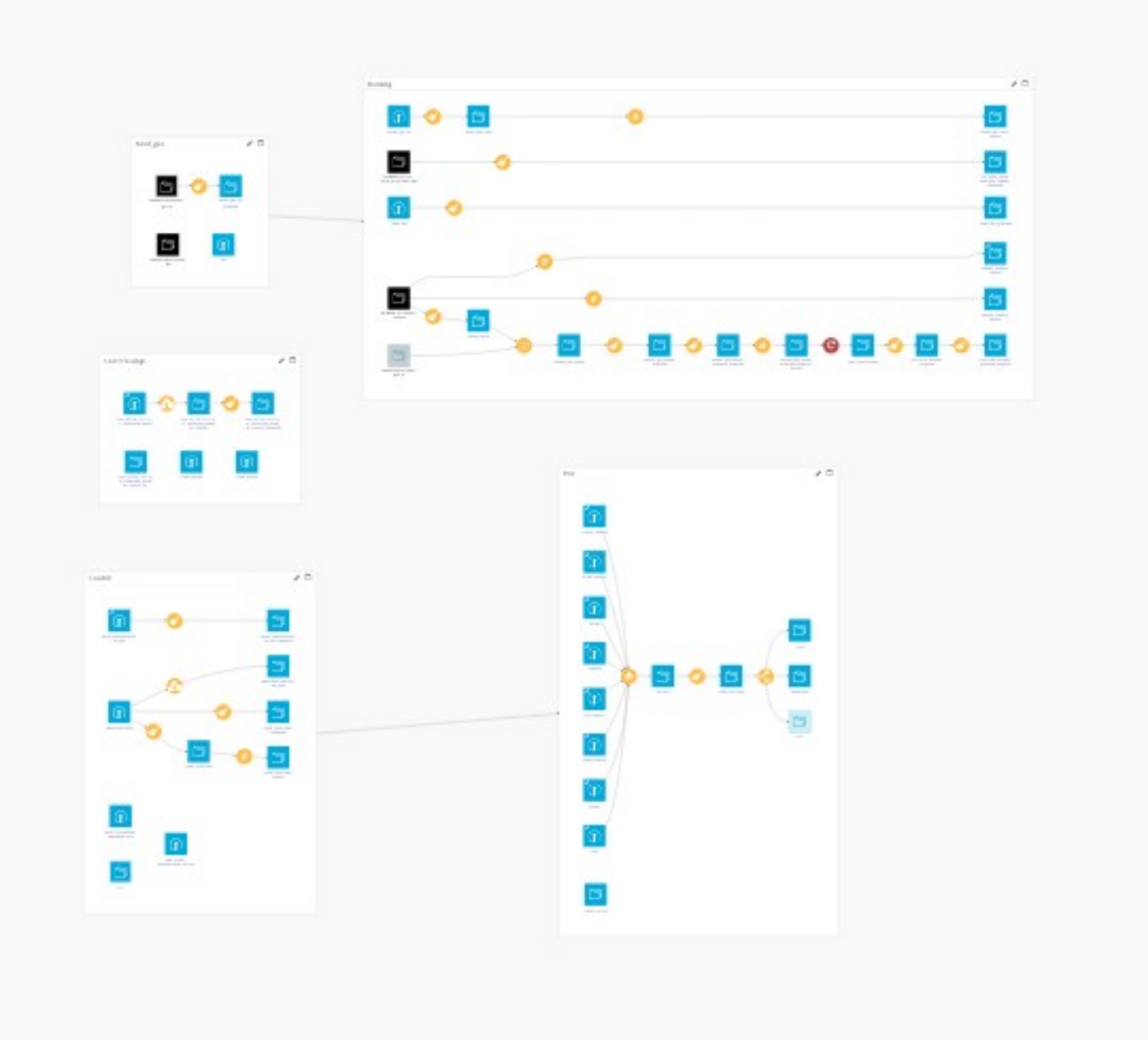

Figure 3 – G´en´eral des traitements effectu´e

6 Mod´elisation

6.1 Probl´ematique

Nous souhaitant pouvoir prendre une d´ecision de voyage pour cette ann´ee. Pour pouvoir r´epondre efficacement a cette probl´ematique notre sch´ema doit donc r´epondre :

- le coût d'une ville, nourriture, transport et d´eplacement...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 Sch´ema Conceptuelle

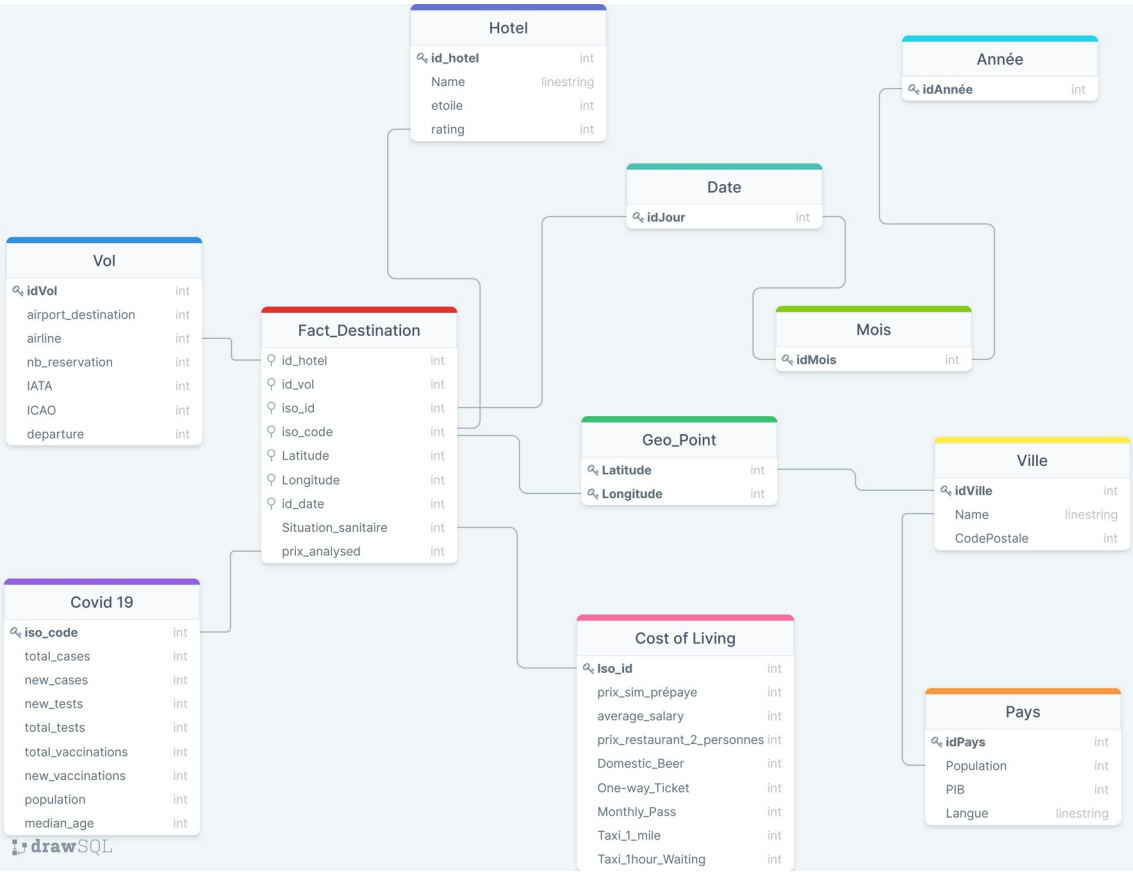

Figure 4 – Sch´ema en flocon

6.3 Sch´ema Normalis´e

— 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 vacci 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,#Ann´ee).
— Ann´ee(id ann´ee).
— 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 nettoy´e et d'int´egrer nous donn´ees afin de les exploit´es.
2. **Dataiku** : Nous a permit d'applique efficacement plein d'aggr´egation, de jointure entre les donn´ees, unifi´e les donn´ees en splittant le tout, appliqu´e des filtres, cr´eer de nouvelle colonne pr´e-calcul´ee.
3. **Notebook python** : Nous a permit de collect´e les donn´ees via le scrapping, afin de les visualis´es pour d´etemin´e la qualit´e des donn´ees collecter.
4. **Tableau Desktop** : Nous a permit de faire des visualisations pour pouvoir les int´egr´es dans le dashbord.

8 Analyse

Nous avons effectu´e nous analyse avec power bi et tableau desktop qui nous offre beaucoup de fonctionnalit´e, Nous avons effectu´e les analyses suivante :

1. **Analyse Cout de la vie :** pour pouvoir voir les pays ayant un cout de vie moins ch´ere, nous avons donc commenc´e a int´egr´e nos donn´ees dans tableau, nous avons choisis don de ne voir que le coût d'un restaurant, d'un taxi, SIM pr´epay´e, salaire moyen du pays ou de la ville.
2. **Analyse situation sanitaire :** nous avons effectue un comparatif du nombre de cas, test, vaccin r´ealis´e pour un pays en suivant la granularit´e Ann´ee/Mois.
3. **Analyse vols :** Ici nous analysons les prix des vols pour voir la destination la plus propice pour voyag´e.
4. **Analyse hotel :** Il s'agit ici d'avoir une id´ee g´en´eralle sur les tarifs appliqu´es dans ces villes en fonction du temps .

9 Pr´ediction

Dans cette section nous avons d´ecid´e de pr´edir le nombre moyen ou total de mort li´ee au covid dans un futur proche.

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

Nous avons d´ecid´e de travaill´e avec les SVM Regression tout simplement parce qu'ils sont tr´es puissant, il s´epare 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 pr´edilection sur un future proche.