**Documentation du nettoyage ou de la manipulation des données**

- **Quels outils choisissez-vous et pourquoi ?** J'ai utilisé BigQuery pour la manipulation et le nettoyage de la base de données du 1er et 4e trimestre 2019 ainsi que le 1er trimestre 2020.
- **Avez-vous assuré l'intégrité de vos données ?** Cela sera fait en vérifiant les erreurs et en documentant le processus de nettoyage.

Je vous partage un fichier avec les codes SQL que j'ai utilisés pour nettoyer les fichiers.

- **Nettoyage :**

1- Renommage des colonnes du 1er trimestre 2020 pour faciliter l'union des tables
2- Convertir le format trip_id, bikeid,from_station_id,to_station_id (integer) en string
3- Unifier les fichiers trimestriels avec les colonnes communes
4- Ajouter les colonnes distinctes trimestielles à la table combinée
5- Changer les noms de deux observations dans la base
6- Vérifier s'il y a de doublon
7- Identifier les lignes vides : il y en une ligne avec 2 cellules vides
8- Compléter les espaces vides avec les données de la base
9- Calculer la durée de trajet en secondes en mettant une condition, si end_time est inférieur à start_time, le résultat doit être null
10- Faire un filtre sur les durées qui sont nulles, il y en a 130 observations nulles, qui sont des erreurs techniques
11- Supprimer ces 130 observations nulles
12- Faire un filtre croissant, il y a des durées de 0 secondes, ce sont des valeurs aberrantes
13- On va faire un filtre et supprimer les données inférieures à 120 secondes, car je suppose un utilisateur prend un vélo pour un trajet environ 120 secondes/ 2 minutes pour 420 secondes/7 minutes à pied. Il y a 22934 observations
14- Filtre décroissant sur la durée, la durée maximum est de 10632022 secondes de trajet, de manière logique c'est impossible, en plus comme c'est un service vélo en libre-service les utilisateurs ne peuvent passer plus d'une journée avec les vélos
15- Supprimer les durées de trajet supérieures à 10800 secondes /3heures. Il y a 4334 trajets qui ont une durée supérieure à 10800 secondes
16- Calculer la variable ride_length sur le format H :M :S
17- Calculer le jour de la semaine où chaque trajet à commencer
18- Calculer les mois de chaque trajet
19- Calculer la saison de chaque trajet
20- Calcul statistique descriptive sur la durée de trajet des utilisateurs
21- Statistiques descriptives par type d'utilisateurs
22- Mode du jour de la semaine
23- Statistiques descriptives par jour de la semaine et par type d'utilisateurs
24- Statistiques descriptives par saison et type d'utilisateur
25- Nombres de trajet et moyenne par le nom de la station, par usertype

--Avec une population de 1468612, les clients occasionnels représentent 11,60% et 88,40% les membres, il y a un fort écart entre les deux groupes. En catégorisant les durées de trajet,

Il semblerait que les clients occasionnels font en moyenne plus de temps avec les vélos que les membres soient 672 secondes contre 1803 secondes. La médiane nous montre que 50% des clients occasionnels a fait une durée de 1290 secondes contre 50% des membres avec 527 secondes.  Cette tendance montre que les clients occasionnels font plus de temps avec le vélo que les membres. Mais puisqu'il y a un écart important entre les deux groupes, le mieux c'est d'avancer les calculs statistiques pour tirer une conclusion réelle et logique.

--Pour affiner les résultats et analyser cette base, on va continuer sur le logiciel R.