Rapport 3

Etude sur la fiabilité de Activité_évènement :

Pour avoir des idées sur les anomalies activité_évènement et combien de fois apparaissent dans les données, on calcule le nombre de toutes les associations possibles:

Table 1: Extrait de correspondances avitivité_évènement avant l'identification des anomalies

activity	event	participant_virtual_id	Nbr_Occurrences
Bureau	Allumage De CheminÃ©e	9999915	641
Bureau	Allumage De CheminÃ©e	9999965	150
Bureau	ArrÃªter De Cuisiner	9999944	75
Bureau	ArrÃªter De Cuisiner	9999946	2
Bureau	ArrÃªter De Cuisiner	9999964	468
Bureau	Cuisiner	9999944	417
Bureau	Fermeture De FenÃªtre	9999916	1607
Bureau	Fermeture De FenÃªtre	9999946	2403
Bureau	Fermeture De FenÃªtre	9999964	1027
Bureau	Fermeture De FenÃªtre	9999965	913

Après qu'on a créé un fichier csv et on l'a rempli manuellement avec toutes les associations possibles et logiques entre l'activité et évènement, puis on a importé dans un dataframe pour exécuter une requête NOT EXISTS qui va permettre de retourner toutes les anomalies sur les données.

Nombre total d'anomalies s'élève à 58046 tenant compte des cas qui peuvent être un peu logiques, pourtant on les a compté parmi les anomalies.

activity	event
NULL	{Valeur}
NULL	NULL
Bureau	Marcher
Inconnu	Ouverture De Fenêtre
Inconnu	Fermeture De Fenêtre

{valeur} : {Marcher,Repos,Fermeture De Fenêtre ,Ouverture De Fenêtre }

Vous trouverez le résultat avec précision de date et identifant de participant dans la table AnomaliesActivityEvent.

Le tableau suivant montre le nombre d'anomalies pour chaque activity/event

Table 3: Les anomalies entre Activité et évènement

activity	event	Nombre_occurrences
Bureau	Allumage De CheminÃ©e	791
Bureau	ArrÃªter De Cuisiner	545
Bureau	Cuisiner	417
Bureau	Marcher	3155
Bus	ArrÃªter De Cuisiner	585
Bus	Cuisiner	154
Bus	Fumer	16
Inconnu	Fermeture De FenÃªtre	60
Inconnu	Ouverture De FenÃªtre	1
Magasin	Fermeture De FenÃªtre	62
Magasin	Ouverture De FenÃªtre	20
NULL	Fermeture De FenÃªtre	143
NULL	Marcher	1
NULL	NULL	46981
NULL	Ouverture De FenÃªtre	1584
NULL	Repos	975
Parc	ArrÃªter De Cuisiner	2
Parc	Fermeture De FenÃªtre	21
Rue	ArrÃªter De Cuisiner	47
Rue	Cuisiner	760
Rue	Fermeture De FenÃªtre	506
Rue	Ouverture De FenÃªtre	129
Train	ArrÃªter De Cuisiner	26
Voiture	Allumage De CheminÃ©e	222
Voiture	ArrÃªter De Cuisiner	39
Voiture	Marcher	703
Voiture	Sport	41
VÃ©lo	Allumage De CheminÃ©e	24
VÃ©lo	Fermeture De FenÃªtre	36

Analyse des moyennes des polluants par activity

Du fait qu'on utilise les requêtes sous RStudio, la fonction Pivot n'est pas valable là-dessus, du coup on a du chercher une alternative, et on a utiliser une clause Filter sur la colonne activity pour selectionner l'activité pour chaque Avg.Voici la requête :

```
req2 <- "SELECT participant_virtual_id
     , AVG(\"PM10\") FILTER (WHERE activity = 'Domicile') PM10_Domicile
     , AVG(\"PM10\") FILTER (WHERE activity = 'Bureau') PM10_Bureau
     , AVG(\"PM10\") FILTER (WHERE activity = 'Rue') PM10_Rue
     , AVG(\"PM10\") FILTER (WHERE activity = 'Voiture') PM10_Voiture
     , AVG(\"PM10\") FILTER (WHERE activity = 'VÃ©lo') PM10_Vélo
     , AVG(\"PM10\") FILTER (WHERE activity = 'Bus') PM10_Bus
     , AVG(\"PM10\") FILTER (WHERE activity = 'Train') PM10_Train
     , AVG(\"PM10\") FILTER (WHERE activity = 'Restaurant') PM10_Restaurant
     , AVG(\"PM10\") FILTER (WHERE activity = 'MÃ©tro') PM10_Métro
     , AVG(\"PM10\") FILTER (WHERE activity = 'Magasin') PM10_Magasin
     , AVG(\"PM10\") FILTER (WHERE activity = 'Parc') PM10_Parc
     , AVG(\"PM10\") FILTER (WHERE activity = 'CinÃ©ma') PM10_Cinéma
```

```
    , AVG(\"PM10\") FILTER (WHERE activity = 'Tramway') PM10_Tramway
       FROM df
       GROUP BY participant_virtual_id;"
 ActivitePM10<-sqldf(req2)
 # Arrondir l'avg à deux chiffres après la virgule
ActivitePM10<-ActivitePM10 %>% mutate_if(is.numeric, round, digits=2)
```

On a fait le même traitement pour PM10, PM1.0, NO2, BC.

Pour savoir le lieu ou l'activité qui produit le plus PM2.5, on calcule la moyenne pour chaque activité_polluant, on prenant les grandes valeurs des moyennes et on obtient ce tableau :

Polluant	Activité	Moyenne
PM2.5	Restaurant	44.07
PM2.5	Rue	17.72
PM10	Restaurant	50.23
PM10	Rue	22.60
PM10	Bus	14.74
PM1.0	Restaurant	28.13
PM1.0	Train	7.60
PM1.0	Bus	7.50
PM1.0	Tramway	7
NO2	Bus	27.99
NO2	Magasin	23.66
NO2	Parc	24.16
NO2	Train	22.94
BC	Bus	27.99
BC	Parc	24.16
BC	Magasin	23.66

On remarque que le restaurant est l'endroit ou il y a une grande émission de polluants PM10 et PM2.5 et moins gravement pour le PM1.0, et vient en deuxième degré la rue,bus pour PM2.5/PM10 et train pour PM1.0.

Le bus, parc et magasin sont les lieu les plus importants qui produisent NO2 et BC.

Etude sur la corrélation :

On essaye de voir si il y a une corrélation entre les polluants, l'humidité ,la température et les donneés de Gps, pour se faire,il faut joindre selon le time et id_participant les tables df et dfGPS.

Ensuite on arrondit la partie de secondes de la colonne time à la plus proche minute pour les deux tables df et dfGPS qui contiennent les données de mesures et données GPS respectivement .

Puis on élimine toutes les redondances de cette colonne dans la table dfGPS .Ces valeurs dupliquées viennent du fait que les valeurs de longitude et latitude ont été prises par rapport d'un intervalle de secondes qui n'est pas fixe.

Table 5: Echantillon de la table des données GPS

	participant_virtual_id	timestamp	lat	lon
1	9999915	2019-10-19 09:24:00	48.80136	2.130643
3	9999915	2019-10-19 09:25:00	48.80139	2.130562

	participant_virtual_id	timestamp	lat	lon
5	9999915	2019-10-19 09:26:00	48.80181	2.129938
7	9999915	2019-10-19 09:27:00	48.80054	2.128678
9	9999915	2019-10-19 09:28:00	48.79890	2.127610
11	9999915	2019-10-19 09:29:00	48.80001	2.123287
13	9999915	2019-10-19 09:30:00	48.80096	2.119687
15	9999915	2019-10-19 09:31:00	48.80248	2.114200

maintenant on a la même structure des colonnes à joindre, on applique la jointure avec ce code :

```
GPS_Mesures<-merge(dfGPS, df, by.x = c("participant_virtual_id","timestamp"),
             by.y = c("participant_virtual_id","time"), all.y = TRUE)
```

L'option all.y est en TRUE pour garder aussi les mesures qui n'ont pas de valeurs GPS car on en aura besoin pour calculer les corrélations entre les autres variables.

Table 6: Echantillon de la table de Mesures_GPS

	participant_virtual_id	timestamp	lat	lon	PM2.5	PM10	PM1.0	Temperature
417	999993	2019-10-22 14:07:00	48.94401	2.250030	16	17	9	27
418	999993	2019-10-22 14:08:00	48.94391	2.249798	15	16	10	27
419	999993	2019-10-22 14:09:00	48.94388	2.249927	12	12	9	27
420	999993	2019-10-22 14:10:00	48.94371	2.249795	13	14	9	27
421	999993	2019-10-22 14:11:00	48.94342	2.249470	16	16	9	27
422	999993	2019-10-22 14:12:00	48.94375	2.250000	12	14	8	27
423	999993	2019-10-22 14:13:00	48.94385	2.250018	12	13	8	27
424	999993	2019-10-22 14:14:00	48.94374	2.250075	12	13	8	27
425	999993	2019-10-22 14:15:00	48.94373	2.249952	13	14	9	27

Sachant que la corrélation s'applique uniquement entre des variables quantitatives, on va enlever les variables qualitatives (event,activity) et on convertit NO2, BC, Humidite, Temperature en valeur numerique puisque ils sont stockés en R sous forme de character ou factor

```
matriceCorr<-round(cor(Correlation, use = "complete.obs"),2)
```

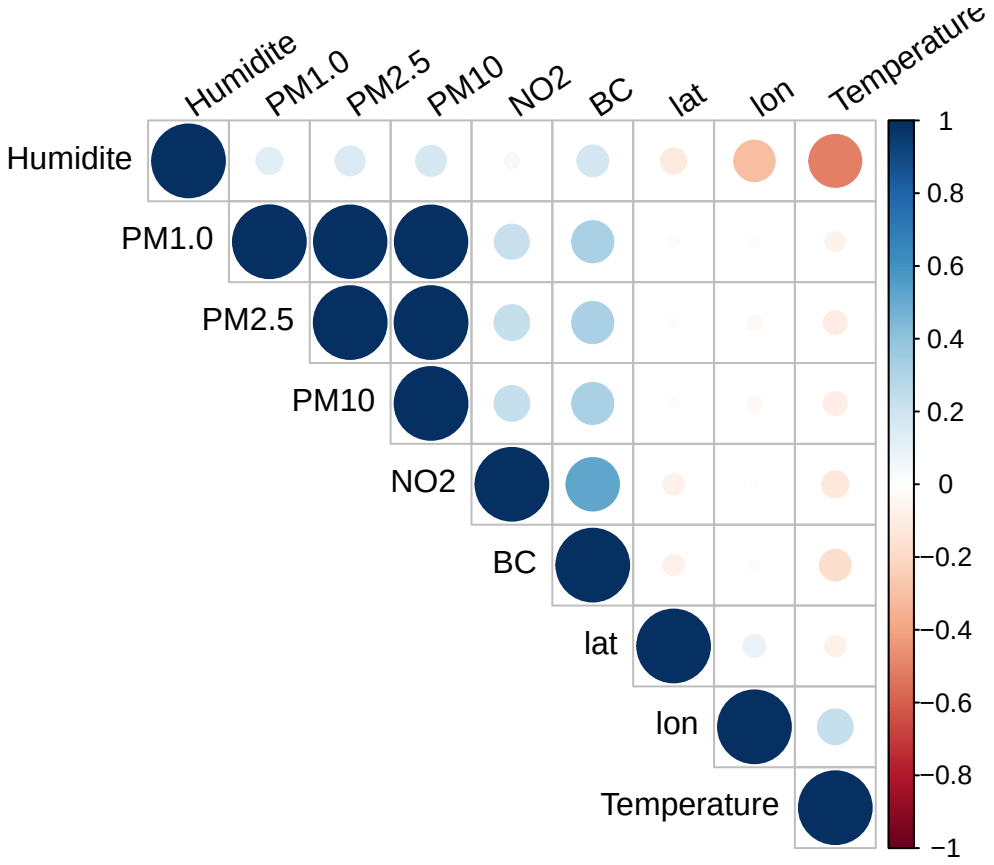
use="complete.obs" afin de ne pas prendre en compte les valeurs manquantes(NA)

Table 7: Matrice de corrélation entre chaque pair de variables

	lat	lon	PM2.5	PM10	PM1.0	Temperature	Humidite	NO2	BC
lat	1.00	0.09	-0.02	-0.02	-0.02	-0.08	-0.12	-0.08	-0.08
lon	0.09	1.00	-0.04	-0.04	-0.02	0.23	-0.31	0.01	0.02
PM2.5	-0.02	-0.04	1.00	1.00	0.99	-0.10	0.16	0.23	0.32
PM10	-0.02	-0.04	1.00	1.00	0.99	-0.10	0.17	0.23	0.32
PM1.0	-0.02	-0.02	0.99	0.99	1.00	-0.07	0.13	0.22	0.32
Temperature	-0.08	0.23	-0.10	-0.10	-0.07	1.00	-0.51	-0.13	-0.18
Humidite	-0.12	-0.31	0.16	0.17	0.13	-0.51	1.00	0.04	0.18
NO2	-0.08	0.01	0.23	0.23	0.22	-0.13	0.04	1.00	0.52
BC	-0.08	0.02	0.32	0.32	0.32	-0.18	0.18	0.52	1.00

Visualisation :

Pour bien analyser la corrélation on va l'afficher sous forme d'un plot



On constate qu'il y a plusieurs niveaux de corrélation :

- une parfaite corrélation entre les polluants PM10 et PM2.5, PM10 et PM1.0, PM2.5 et PM1.0, ce qui évident puisque ils sont différents que par rapport leurs tailles, donc ils peuvent exister au même endroit si la source de pollution en provoque.
- Une corrélation forte entre : d'une part NO2 et BC mais il retse toujours pas intéréssent d'étudier leur dépendance vu que cela ne va pas ajouter des informations sur l'exposition des participant à ces polluants, et d'autre part entre température et humidité ce qui va permettre de donner des regressions sur des variables atmosphériques.
- Une corrélation moyenne entre BC et les particules PM et entre NO2 et les PM, et également entre humidité et longitude et de moins degré entre température et longitude
- Une corrélation faible ou très faible entre latitude et humidité, latitude et NO2/BC, humidité et les autres polluants sauf NO2 qui a presque une null corrélation avec l'humidité