CONCOURS DE DATA-VISUALISATION
DES DONNEES
DE TRANSPORT AERIEN

Berkeley, 9 Mai 2018

Julie Nguyen (Etudiante française), Miyabi Ishihara (Etudiante japonaise),
Raphael Grillo Avila (Etudiant américain)

Nous répondons à 5 défis définis dans le règlement du concours.

Nous utilisons R – open langage pour répondre à ces questions, notamment les « tidyverse » et « sf » paquets, les plus récents paquets de R pour la visualisation graphiques des données temporelles et spatiales.

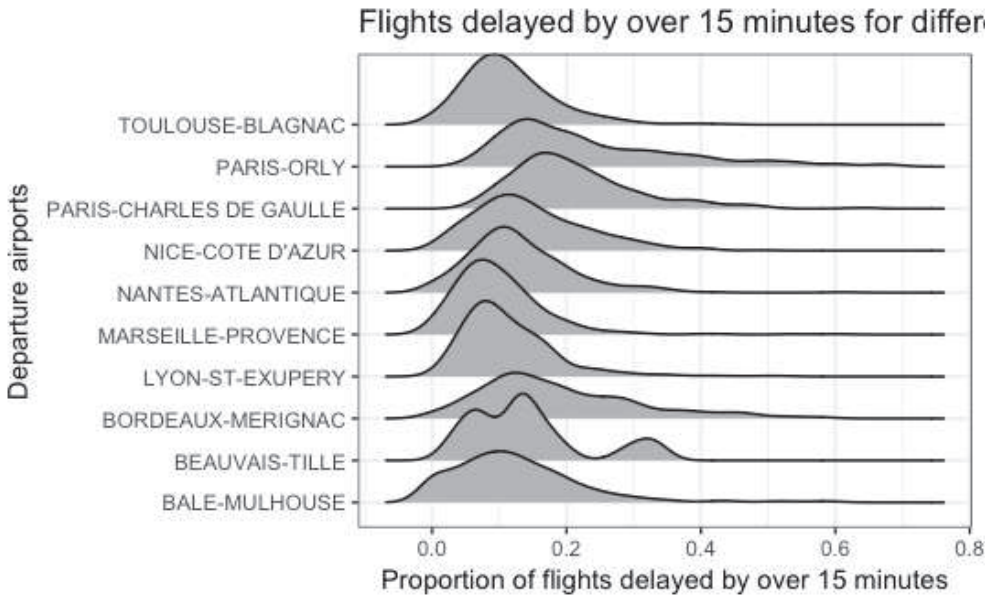
- Défi n°1 : accès à des données numériques avec sélections de période temporelle et/ou de zone

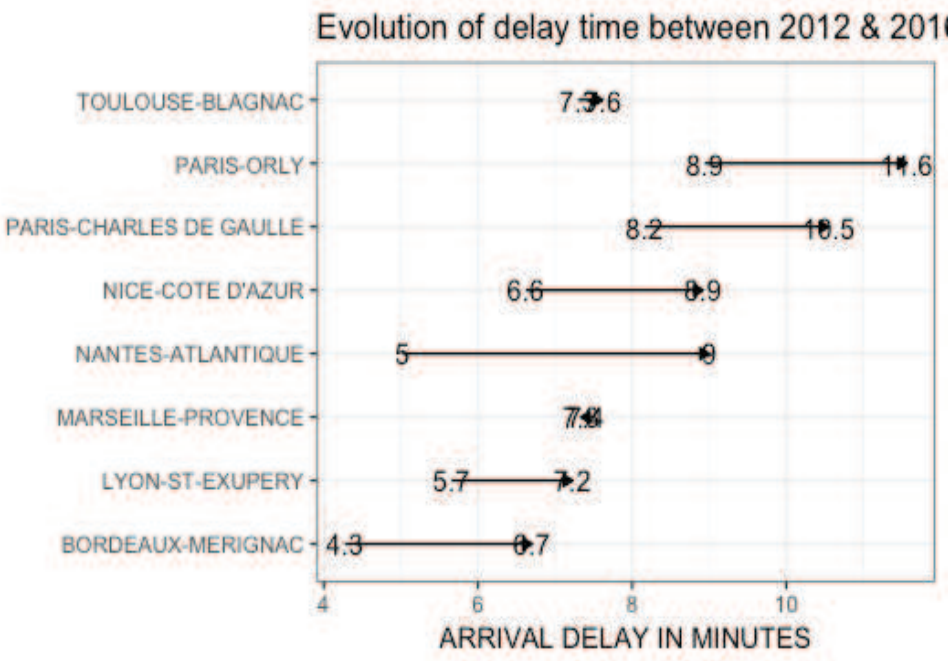
Le défi ici est de lire et nettoyer des données françaises, surtout utiliser les variables communs de différentes « cvs » dossiers EMI, TRA et RET.

Utilisant les plus innovants paquets nous permets de joindre, choisir, fixer les seuils de ces données très facilement sous forme de « data frame ».

- Défi n°2 : visualisation graphique des évolutions temporelles

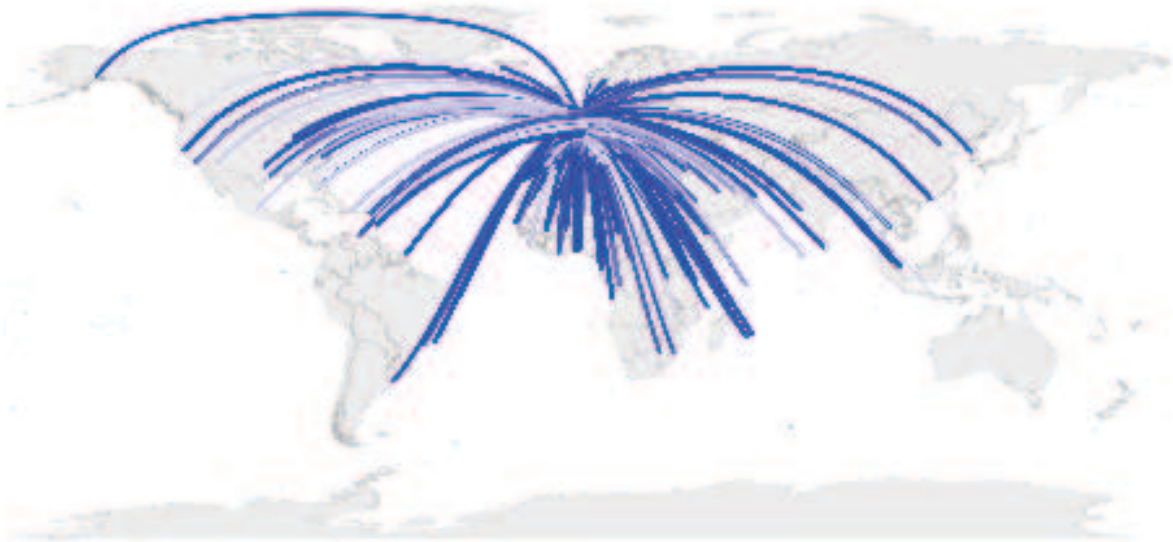
Plusieurs choix sont possibles. Ici, nous joignons deux graphes que nous travaillons sur les évolutions des retards.



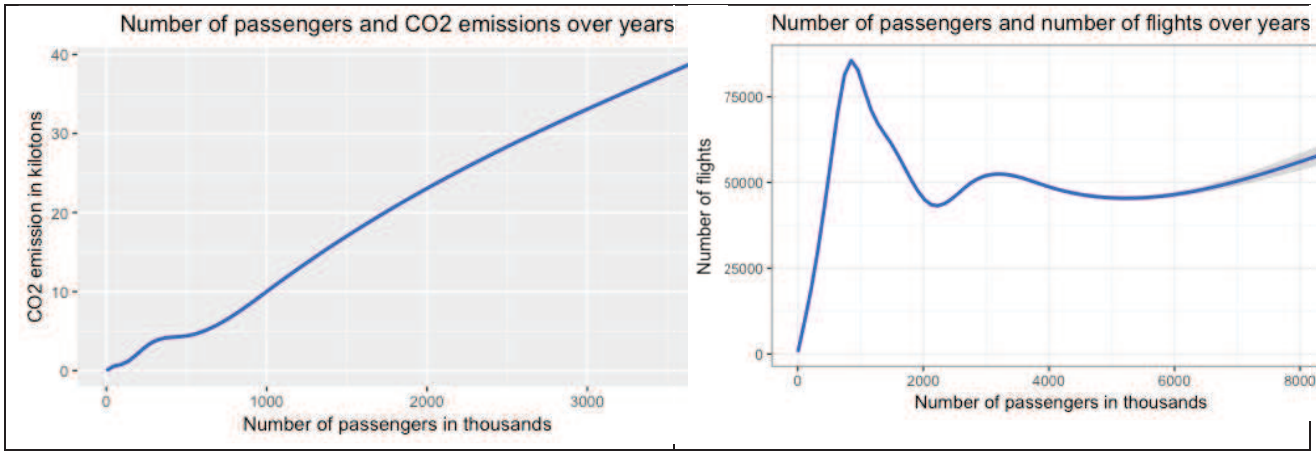


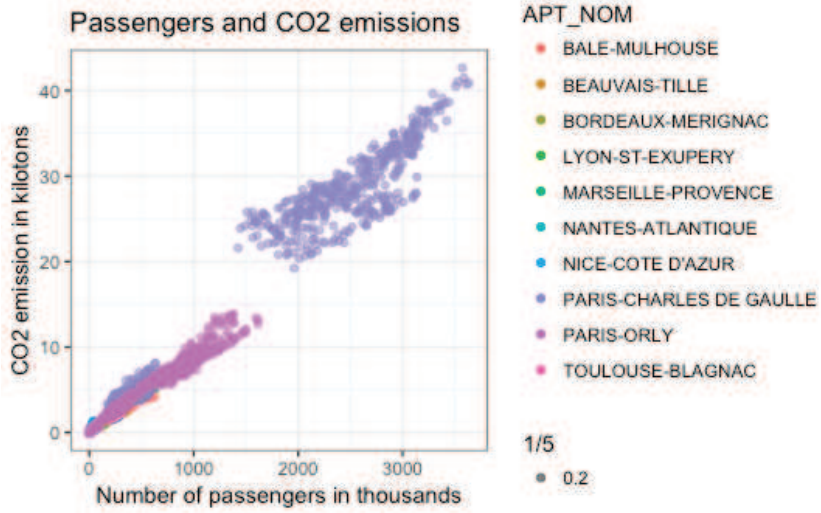
- Défi n° 3 : visualisation de la structure géographique

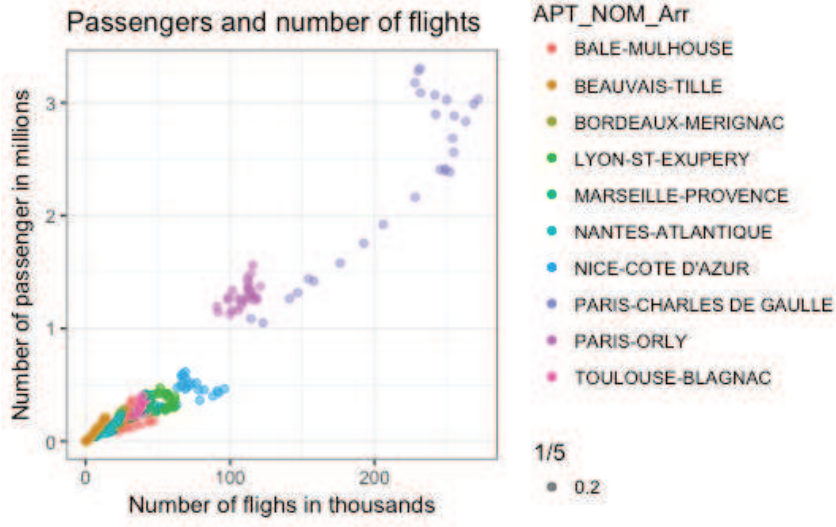
Cette partie pourrait être développée de manière plus interactive sur le territoire française ou européenne uniquement.



- Défi n° 4 : visualisation de corrélations éventuelles



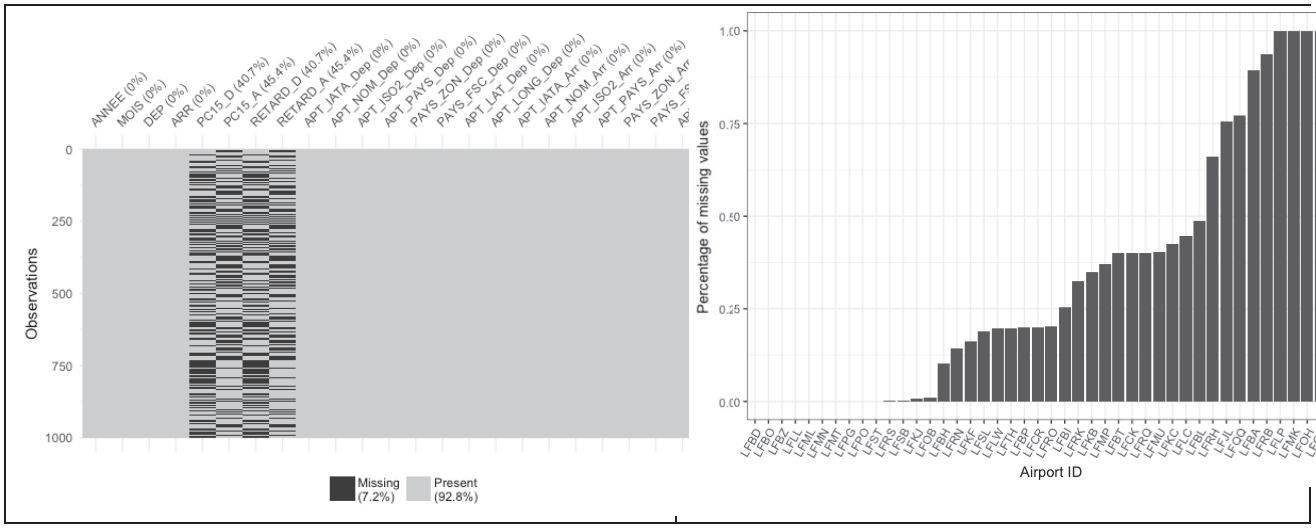




A partir des graphes, nous pouvons observer que le nombre de passagers et la quantité d'émissions de CO2 sont fortement corrélés. La relation est quasi linéaire.

Cependant, la relation entre le nombre de passagers et le nombre de vols n'est pas linéaire car la majorité des vols sont de courte ou moyenne distance, avec moins de passagers que de longs courriers.

- Défi n°5 : info/avertissement sur des données manquantes.
Encore une fois, les nouveaux paquets nous permettent d'identifier facilement des données manquantes :



On pourrait observer que les données sur les retards ont beaucoup de données manquantes. Nous essayons de trouver s'il y a des motifs dans les valeurs manquantes.

À partir de la courbe des valeurs manquantes ci-dessous, nous constatons qu'environ 40% des entrées sont manquantes sur quatre colonnes relatives au retard de vol. Nous voyons également que, naturellement, les deux variables sur le délai de départ (PC15_D et RETARD_D) ont les mêmes entrées manquantes et deux variables sur le délai d'arrivée (PC15_A et RETARD_A) ont les mêmes entrées manquantes. Cependant, lorsque les données sont limitées aux vols en France, il y a moins de données manquantes: environ 12% des entrées sont manquantes.