# Exercises lecture 3 – data wrangling

Paolo Crosetto

octobre 2020

## `filter()` & `select()`

### Exercice 1

sauvegardez dans un nouvel objet tous les vols partis entre midi et deux heures, en gardant juste l'info sur l'aéroport de départ et d'arrivée

```
df %>%
  filter(dep_time >= 1200 & dep_time <= 1400) %>%
  select(dep_time, origin, dest) -> df_midday
```

### Exercice 2

isolez dans un nouvel objet tous les vols partis entre minuit et une heure du matin de JFK et de LGA. Quelle est, pour chacun de deux aéroport, la destionation la plus fréquente?

## `mutate()`

### Exercice 3

créez une variable qui montre la vitesse de chaque avion

### Exercice 4

créez une variable qui calcule l'impact (en %) du retard à l'arrivée sur le temps de vol

## `summarise()` and `group_by()`

### Exercice 5

calculez la moyenne, l'écart type, le min et le max du rétard à l'arrivée

### Exercice 6

même chose que l'exercice 6, mais par aéroport de départ

### Exercice 7

calculez la moyenne du retard par compagnie aérienne

### Exercice 8 – filter + select + mutate + summarise + group_by

quelle est la vitesse moyenne des vols qui partent entre 11h et 13h, par mois?

## meet the pipe: %>%

### meta-exercice 1

**re-faites** *tous* les exercices ci-dessus en utilisant l'opérateur 'et après' / pipe `%>%`

### Exercice 9

trouvez le maximum retard au depart par aéroport pour JFK et LGA pour chaque jour de l'an. Est-ce que les retards sont corrélés?

### Exercice 10

de quel aéroport partent les vols à plus longue distance?

## `join_...()` family of functions

first run this setup R code chunk. It will load in your workspace 3 data frames:

- **airports**: avec données sur les aéroports américains
- **flights**: qu'on connait déjà
- **planes**: avec les données pour chaque avion

```
planes <- planes
flights <- flights
airports <- airports
```

### Exercice 11

est-ce que les routes plus longues sont desservies apr les avions les plus modernes?

*notes*: utilisez `left_join()` et mergez les dataframes `flights` et `planes`

### Exercice 12

combien de vols qui partent des trois aéroport de NY atterrisent dans des destinations au dessus de 1000m s.n.m.?

## creating tidy data: reshape with `gather()` and `spread()`

### Exercice 13

tidy world_bank_pop dataset so that 'year' is a variable and for each country and each year you have urban population and urban population growh only. Plot as a geom_line the total population for each country over the years.