Projet

Julien

2025-04-23

Chargement packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Chargement dataset :

```
data <- read_csv('dataset/daily_rent_detail.csv')
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 16086672 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Questions :

Question 2a : Y a-t-il une variation du nombre de trajets selon les jours de la semaine ?

Prétraitement des données pour les question 2a et 2b :

Pour cette analyse, nous avons choisi de ne conserver que la variable started_at, qui indique le moment précis où chaque vélo a été emprunté. À partir de cette information temporelle, nous avons dérivé d'autres variables selon les questions que nous souhaitions étudier :

- Pour la question 2a, nous avons utilisé la fonction wday() du package lubridate afin d'extraire le jour de la semaine correspondant à chaque date.
- Pour la question 2b, nous avons extrait le mois grâce à la fonction month().

Ces transformations nous ont permis de comprendre l'évolution du nombre d'emprunts selon les moments de la semaine ou de l'année.

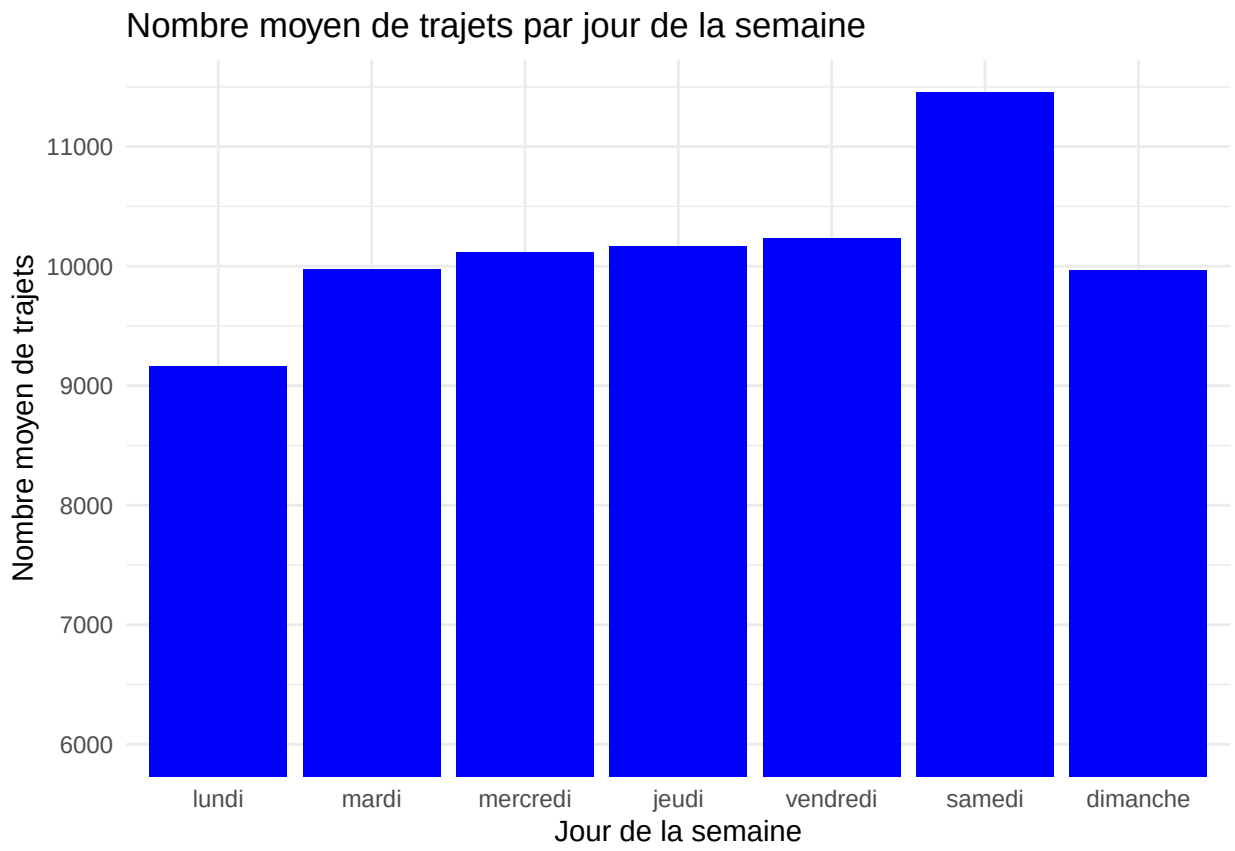
Hypothèse :

Nous imaginons que les utilisateurs empruntent davantage les vélos en semaine (pour se rendre au travail notamment) que le week-end.

Graphique :

```r
# Préparation des données
avg_rides_per_day <- data %>%
  transmute(day_of_week = wday(started_at, label = TRUE, abbr = FALSE, week_start = 1), date = as_date(s
  group_by(date, day_of_week) %>% # un groupe = une date unique avec son jour
  summarise(daily_count = n(), .groups = "drop") %>%
  group_by(day_of_week) %>%
  summarise(avg_per_day = mean(daily_count))

# Visualisation
avg_rides_per_day %>%
  ggplot(aes(x = day_of_week, y = avg_per_day)) +  geom_col(fill = "blue") +
  labs(title = "Nombre moyen de trajets par jour de la semaine",
       x = "Jour de la semaine",
       y = "Nombre moyen de trajets") +
  scale_y_continuous(breaks = seq(6000, 12000, by = 1000)) +
  coord_cartesian(ylim = c(6000, NA)) + # Limite inférieure à 6000
  theme_minimal()
```



Interprétation :

Contrairement à notre hypothèse initiale, les week-ends enregistrent en réalité un nombre très élevé de trajets, en particulier le samedi qui dépasse tous les autres jours de la semaine. Cela suggère que l'usage du vélo ne se limite pas à une fonction utilitaire (travail), mais qu'il est aussi très utilisé pour les loisirs. Le dimanche reste élevé, presque équivalent aux jours de travail.

En semaine, le nombre de trajets reste relativement stable autour de 10 000 trajets par jour. Cependant, on observe une baisse notable le lundi, où la moyenne descend à environ 9 000 trajets. Cela peut s'expliquer par le fait que de nombreux commerces, établissements culturels ou services (comme les banques) sont souvent fermés le lundi, ce qui réduit potentiellement le besoin de déplacement.

Cette observation pourrait être affinée par une analyse horaire (Les pics du matin et du soir existent-ils en semaine ?) et une analyse des catégories d'usagers (membres ou occasionnels)

Question 2b : Y a-t-il une variation du nombre de trajets selon les mois de l'année ?

Hypothèse :

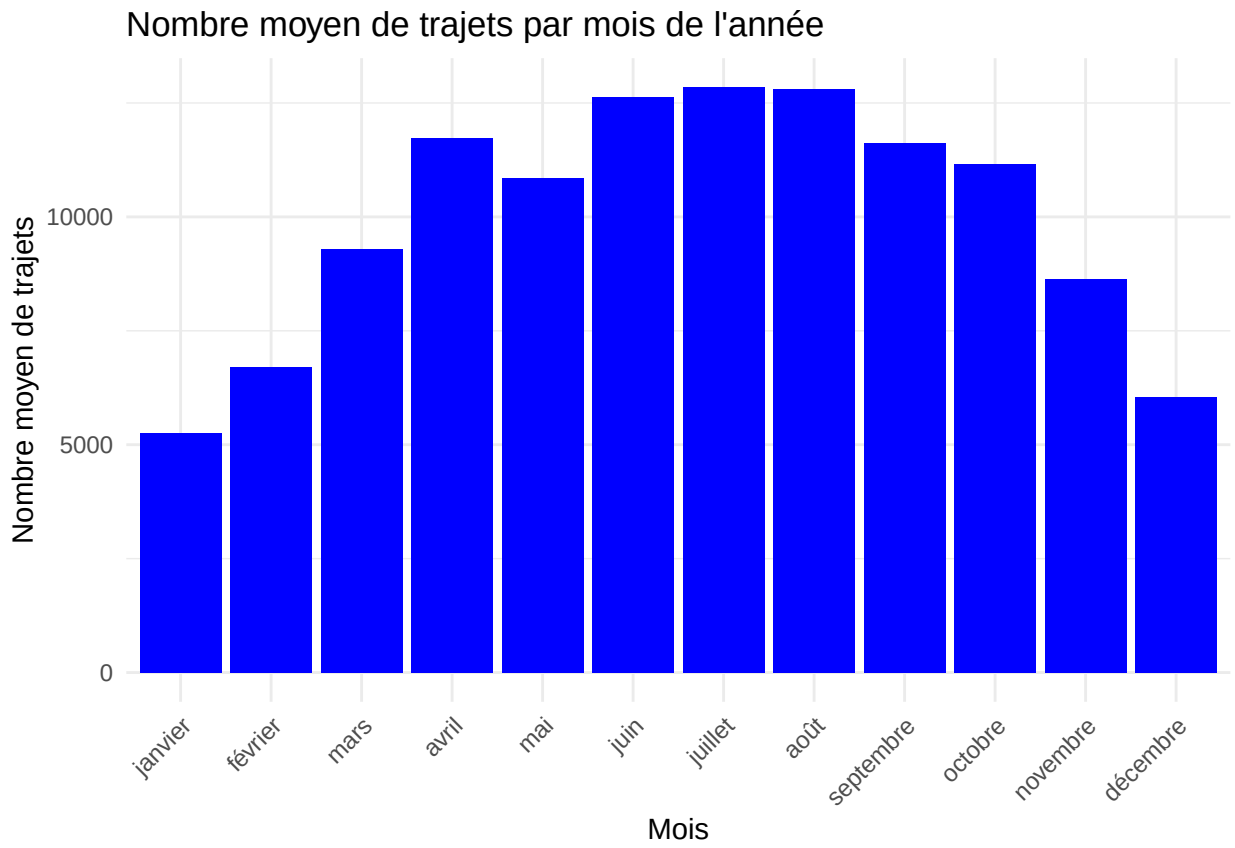
Nous supposons que la fréquentation des vélos est plus importante en été, en raison du beau temps, et qu'elle diminue pendant les mois froids et pluvieux.

Graphique :

```
# Préparation des données
avg_rides_per_month <- data %>%
  transmute(months = month(started_at, label = TRUE, abbr = FALSE), date = as_date(started_at)) %>% # o
  group_by(date, months) %>% # un groupe = une date unique avec son mois
```

```r
  summarise(monthly_count = n(), .groups = "drop") %>%
  group_by(months) %>%
  summarise(avg_per_month = mean(monthly_count))

# Visualisation
avg_rides_per_month %>%
  ggplot(aes(x = months, y = avg_per_month)) +  geom_col(fill = "blue") +
  labs(title = "Nombre moyen de trajets par mois de l'année",
       x = "Mois",
       y = "Nombre moyen de trajets") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interprétation : Le nombre moyen de trajets évolue clairement au rythme des saisons.

L'activité augmente fortement entre avril et septembre, avec un pic en juillet-août où l'on atteint près de 13 000 trajets en moyenne. Cette hausse s'explique sans doute par les vacances estivales et des conditions météorologiques plus favorables.

À l'inverse, les trajets diminuent nettement en hiver, notamment entre décembre et février, où le nombre moyen redescend autour de 5 000 trajets en décembre. Cette baisse reflète un usage moindre du vélo durant les mois les plus froids de l'année.