Analyse de l'activité

Oumaima Bouchiba

2025-03-19

Charger les données

```
activity_data <- read.csv("/Users/bekkaryounes/Downloads/activity.csv", stringsAsFactors = FALSE)

# Vérifier les premières lignes
head(activity_data)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```
# Vérifier la structure
str(activity_data)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : chr  "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

```
# Convertir la colonne date
activity_data$date <- as.Date(activity_data$date, format="%Y-%m-%d")
```

Analyser les valeurs manquantes

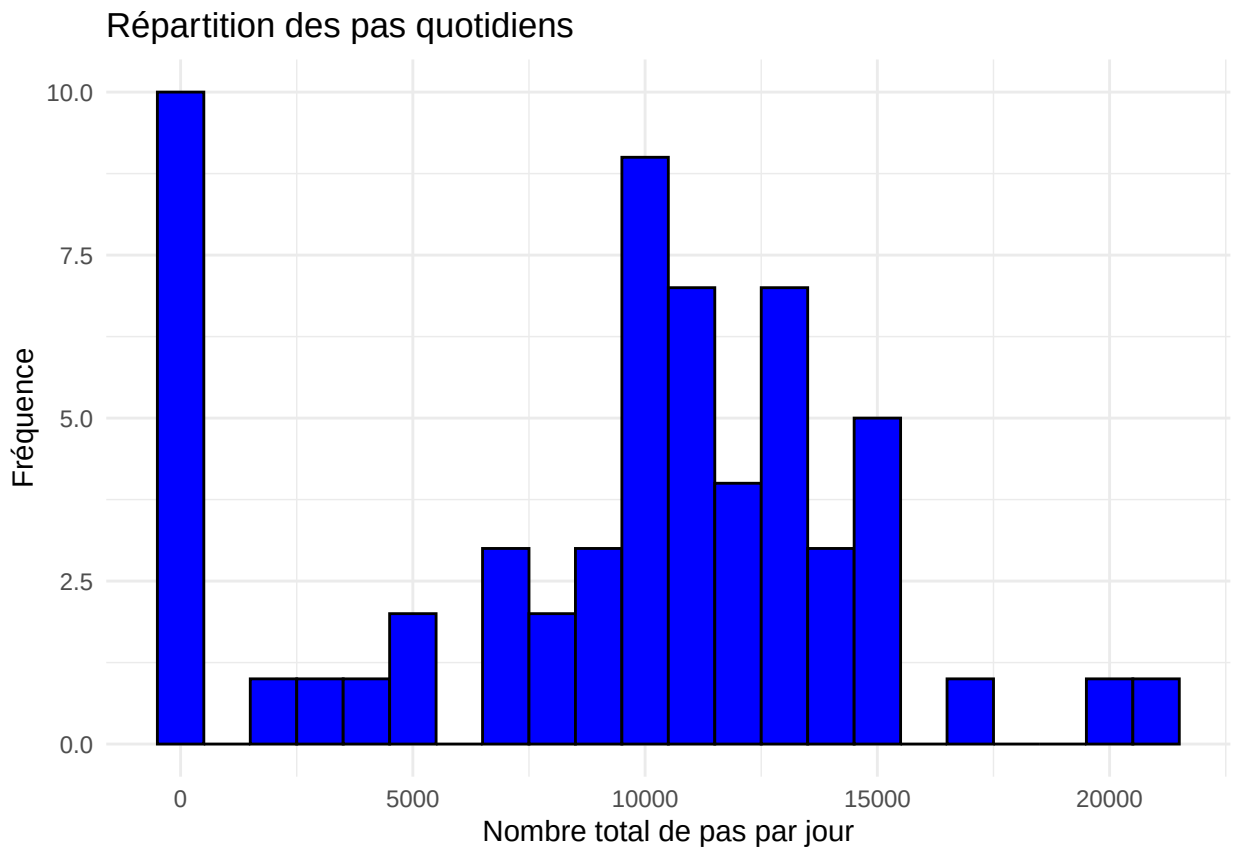
```
sum(is.na(activity_data$steps))
```

```
## [1] 2304
```

Nombre total de pas par jour

```r
daily_steps <- activity_data %>%
  group_by(date) %>%
  summarise(total_steps = sum(steps, na.rm = TRUE))

# Histogramme
ggplot(daily_steps, aes(x = total_steps)) +
  geom_histogram(binwidth = 1000, fill = "blue", color = "black") +
  labs(title = "Répartition des pas quotidiens",
       x = "Nombre total de pas par jour",
       y = "Fréquence") +
  theme_minimal()
```



Moyenne et médiane des pas quotidiens

```r
mean_steps <- mean(daily_steps$total_steps, na.rm = TRUE)
median_steps <- median(daily_steps$total_steps, na.rm = TRUE)
mean_steps
```

```
## [1] 9354.23
```

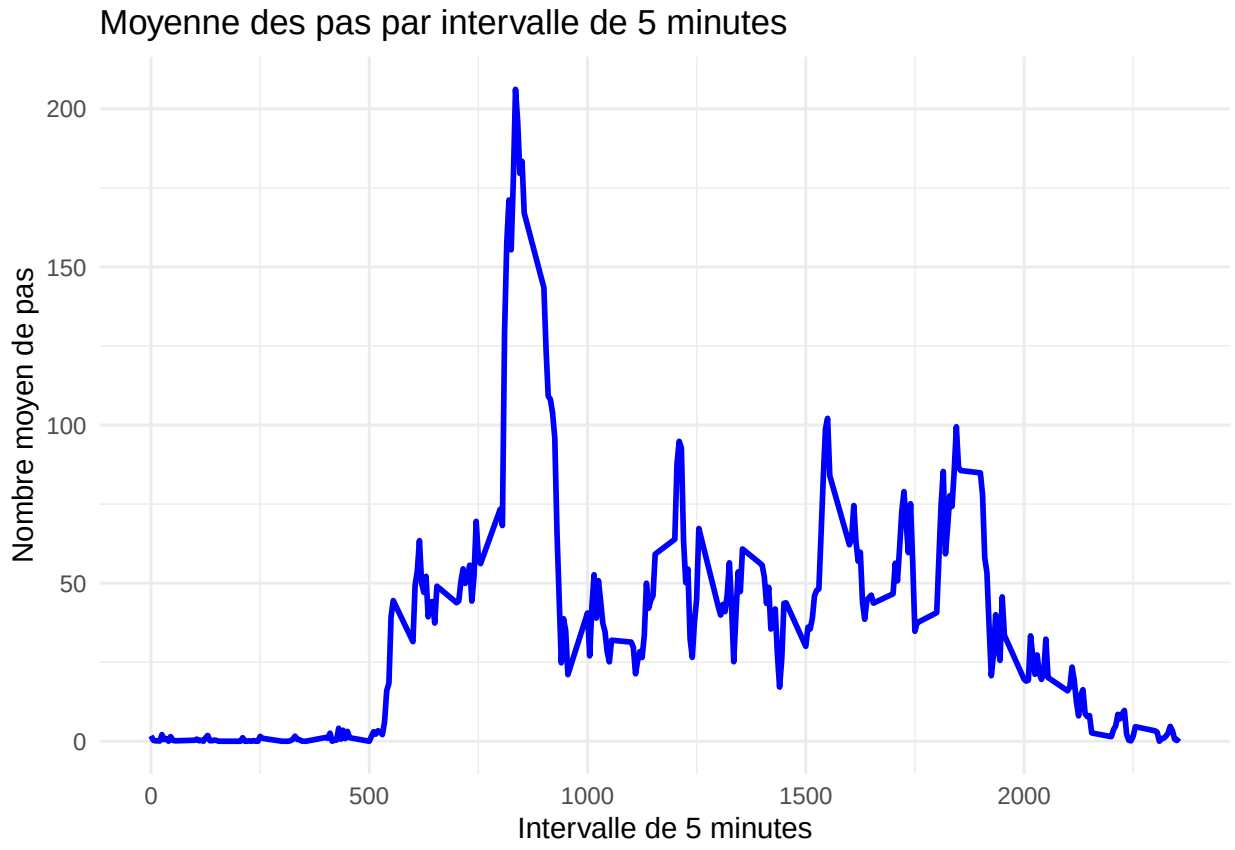
```
median_steps
```

```
## [1] 10395
```

Analyse des intervalles de 5 minutes

```
interval_avg <- activity_data %>%
  group_by(interval) %>%
  summarise(avg_steps = mean(steps, na.rm = TRUE))

# Série temporelle
ggplot(interval_avg, aes(x = interval, y = avg_steps)) +
  geom_line(color = "blue", linewidth = 1) +
  labs(title = "Moyenne des pas par intervalle de 5 minutes",
       x = "Intervalle de 5 minutes",
       y = "Nombre moyen de pas") +
  theme_minimal()
```



Imputation des valeurs manquantes

```r
activity_data_imputed <- activity_data %>%
  left_join(interval_avg, by = "interval") %>%
  mutate(steps = ifelse(is.na(steps), avg_steps, steps)) %>%
  select(-avg_steps)

# Vérification
sum(is.na(activity_data_imputed$steps))
```

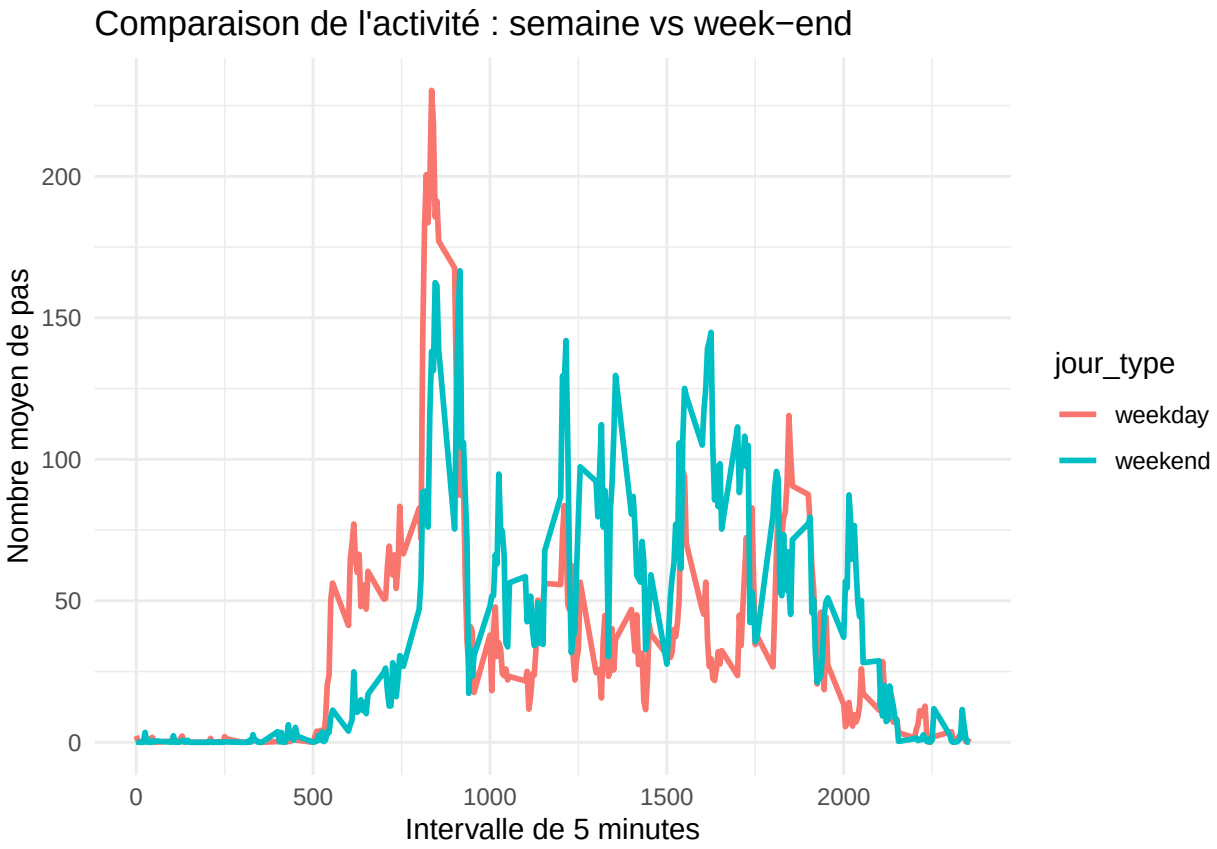
```
## [1] 0
```

Comparaison semaine vs week-end

```r
activity_data_imputed$jour_type <- ifelse(weekdays(activity_data_imputed$date) %in% c("Saturday", "Sunda

interval_weekly <- activity_data_imputed %>%
  group_by(interval, jour_type) %>%
  summarise(avg_steps = mean(steps), .groups = 'drop')

# Graphique comparatif
ggplot(interval_weekly, aes(x = interval, y = avg_steps, color = jour_type)) +
  geom_line(linewidth = 1) +
  labs(title = "Comparaison de l'activité : semaine vs week-end",
       x = "Intervalle de 5 minutes",
       y = "Nombre moyen de pas") +
  theme_minimal()
```


Comparaison de l'activité : semaine vs week−end
Nombre moyen de pas
0
50
100
150
200
0
500
1000
1500
2000
Intervalle de 5 minutes
jour_type
weekday
weekend