Analyse de l'utilisation des vélos libre-service à Washington

Julien Schieler, Icham Lecorvaisier, Mathis Reslinger, Lila Mortier

Introduction

Données

Nous utilisons 4 jeux de données pour analyser l'utilisation des vélos libre-service à Washington :

Information sur les locations de vélos (source :https://www.kaggle.com/datasets/taweilo/capital-bikeshare-dataset-202005202408?select=station_list.csv)

- **Nom du dataset :** daily_rent_detail.csv
- **Origine des données** : Ces données proviennent de l'entreprise Capital Bike Share, l'entreprise de location de vélos libre service de Washington.
- **Pourquoi ces données ?** Elles permettent d'observer l'utilisation de ces vélos, en analysant les dates et heures d'utilisation, les lieux d'emprunt et de dépôt notamment.
- **Nombre d'observations** : 16 086 673
- **Nombre de colonnes** : 13 colonnes incluant :
  - Identifiant du trajet ("ride_id")
  - Type du vélo ("rideable type") : vélo classique, électrique, autre (cargo)
  - Date et heure de l'emprunt ("started_at")
  - Date et heure du retour ("ended_at")
  - Nom de la station de départ ("start_station_name")
  - Identifiant de la station de départ ("start_station_id)
  - Nom de la station d'arrivée ("end_station_name")
  - Identifiant de la station d'arrivée ("end_station_id")
  - Latitude de départ ("start_lat")
  - Longitude de départ ("start_lng")
  - Latitude d'arrivée ("end_lat")
  - Longitude d'arrivée ("end_lng")
  - Type d'abonnement du client ("member_casual") : utilisateur en abonnement casual ou membre.
- **Créateur et éditeur** : Entreprise Capital Bike Share
- **Format** : CSV
- **Sous-groupes** :
  - Type de vélo : vélo classique, électrique et cargo
  - Type d'utilisateur : occasionnel et membre
  - Stations : possibilité d'analyser par station d'origine ou de destination
  - Périodes temporelles : heures, jours, saisons, etc...

Liste des stations de Washington (source :https://www.kaggle.com/datasets/taweilo/capital-bikeshare-dataset-202005202408?select=station_list.csv)

- **Nom du dataset :** station_list.csv
- **Origine des données** : Ces données proviennent de l'entreprise Capital Bike Share, l'entreprise de location de vélos libre-service de Washington.
- **Pourquoi ces données ?** Elles permettent d'avoir la liste des stations à Washington
- **Nombre d'observations :** 912 stations différentes
- **Nombre de colonnes :** 2 colonnes incluant **:**
  - Nom de la station

  - Identifiant de la station
- **Format :** CSV
- **Sous-groupes** :
  - Catégorisation géographique (ex: quartiers)

Emprunts et dépôts des vélos par station (source :https://www.kaggle.com/datasets/taweilo/capital-bikeshare-dataset-202005202408?select=station_list.csv)

- **Nom du dataset :** usage_frequency.csv
- **Origine des données** : Ces données proviennent de l'entreprise Capital Bike Share, l'entreprise de location de vélos libre-service de Washington.
- **Pourquoi ces données ?** Elles permettent d'avoir le nombre d'emprunt et de dépôt pour chaque station à Washington
- **Nombre d'observations :** 873 318
- **Nombre de colonnes :** 4 colonnes incluant **:**
  - Date
  - Nom de la station
  - Nombre d'emprunt sur cette station
  - Nombre de dépôt sur cette station
- **Format :** CSV
- **Sous-groupes** :
  - Date
  - Stations : possibilité d'analyser par stations
  - Activité : Emprunt et dépôt
  - Utilisation : stations très utilisées et peu utilisées

Météo (source : https://www.kaggle.com/datasets/taweilo/capital-bikeshare-dataset-202005202408?select=weather.csv)

- **Nom du dataset :** weather.csv
- **Origine des données** : Ces données proviennent de l'entreprise Visual Crossing qui est un service de météo de Virginia aux Etats Unis.
- **Pourquoi ces données ?** Elles permettent d'obtenir la météo de chaque jour afin de pouvoir croiser ces données avec l'utilisation des vélos à Washington
- **Nombre d'observations :** 1584, une pour chaque jour entre mai 2020 et août 2024
- **Nombre de colonnes :** 32 colonnes dont 7 que nous allons utiliser **:**
  - Date ("*datetime*")
  - Moyenne de température sur la journée ("*temp*")
  - Moyenne de température ressentie sur la journée ("*feelslike*")
  - Précipitations ("*precip*") : quantité de liquide tombé / prévu dans la période
  - Epaisseur de neige au sol ("*snowdepth*")
  - Couverture nuageuse ("*cloudcover*") : entre 0 et 100%
  - Vitesse du vent ("*windspeed*") : vitesse moyenne du vent soutenu, mesurée comme la vitesse moyenne du vent survenant au cours de la à deux minutes précédentes
- **Créateur et éditeur** : Visual Crossing (fournisseur de données météorologiques)
- **Format :** CSV
- **Sous-groupes** :
  - Conditions météorologiques : pluie, neige, vent, etc...
  - Températures : froid, tempéré, chaud
  - Saisons : printemps, été, automne, hiver
  - Jours : semaine ou week-end

Plan d'analyse

Avant de commencer l'analyse, nous nous posons plusieurs questions que l'on peut catégoriser :

1. Questions exploratoires
    1. Utilisation générale
        - Quelle est la tendance générale de l'utilisation des vélos au cours de l'année ? Y a-t-il des pics saisonniers ?
          Graphique : Lineplot du nombre de trajet en fonction du temps
          Variables : Nombre de trajets, date
        - Combien de trajets sont effectués en moyenne par jour/mois ?
          Graphique : barplot du nombre de trajets par jour de la semaine, par mois de l'année
          Variables : Nombre de trajets, date d'emprunt
        - Quelles stations sont les plus utilisées pour les départs/arrivées ?
          Graphique : Carte avec points de taille différentes pour le nombre de départs/arrivées
          Variables : Nom des stations, Nombre de départs/arrivées, Coordonnées des stations
        - Quel type de vélo (électrique, classique ou cargo) est le plus utilisé ?
          Graphique : Barchart avec pourcentage/nombre de vélos par catégorie
          Variables : Type de vélo, Nombre de trajets par type de vélos
        - Quel est le comportement moyen selon le type d'utilisateur (occasionnel ou membre) ?
          Graphique : Radar chart (potentiellement plusieurs)
          Variables : moyenne de durée de trajet, distance, heure d'emprunt et de dépôt, date (jour de la semaine), type d'abonnement, type de vélo
        - Quelles sont les durées moyenne des trajets ? Varient-elles selon le jour de la semaine ou la météo ?
          Graphique : Boxplot ou violin plot
          Variables : Durée du trajet, type d'utilisateur, date, condition météo
        - Les vélos électriques sont-ils plus prisés pendant les heures de pointe ?
          Graphique : Bar plot des différents types de vélo
          Variables : Type du vélo, Date
    2. Spatio-temporel
        - Quelles stations sont les plus actives à différentes périodes de la journée (matin ou soir) ?
          Graphique : Heatmap avec X= heure, Y = station et couleur = nombre de trajets
          Variables : Heure, Nom de la station, Nombre de trajets
        - Quelles sont les stations les plus utilisées le week-end, en semaine ?
          Graphique : Stacked barplot
          Variables : Jour de la semaine, nom station, nombre de trajets
        - Existe-t-il des stations avec un fort déséquilibre entre départs et arrivées ?
          Graphique : Barplot (différence départ, arrivée)
          Variables : Nom de la station, Nombre de départs/arrivées
        - Quels sont les trajets les plus fréquents (station de départ vers station d'arrivée) ?
          Graphique : Barplot (top 10 des paires de stations les plus fréquentées)
          Variables : Nom de la station de départ, Nom de la station d'arrivée, Nombre de trajets
2. Questions explicatives
    1. Lien entre météo et usage
        - Comment la météo influence-t-elle l'usage des vélos ? (pluie, température, vent, etc...)
          Graphique : Lineplot ou scatterplot par variable météorologique
          Variables : Conditions météo, Nombre de trajet par jour
        - En cas de météo extrême (fortes pluies ou froid extrême), est-ce que le volume de trajets diminue fortement ?
          Graphique : Boxplot (jours normaux vs météo extrême)
          Variables : Conditions météo, nombre de trajets

- Quel est l'impact de la neige ou du vent ou de la couverture nuageuse sur la durée moyenne des trajets ?
  Graphique : Lineplot ou scatterplot par variable météorologique
  Variables : Conditions météo, Durée moyenne des trajets

2. Différences selon les utilisateurs
   - Les membres abonnés utilisent-ils plus les vélos que les usagers occasionnels quand il fait mauvais ?
     Graphique : Lineplot avec comparaison entre les types d'utilisateur
     Variables : météo, nombre de trajet, type d'utilisateur
   - Selon les saisons, la fréquence de cyclistes occasionnels change-t-elle comparé aux membres ?
     Graphique : Lineplot par type d'utilisateur
     Variables : nombre de trajets, type d'utilisateurs, mois
3. Temporalité
   - En se focalisant sur les utilisateurs, y a-t-il des effets "heures de pointe" dans la journée ?
     Graphique : Lineplot (heure en abscisse et nombre de trajets en ordonnée)
     Variables : nombre de trajets, heure
4. Évènements
   - Est ce qu'il est possible de voir une différence d'utilisation des vélos lors de la prise du Capitol le 6 janvier 2021?
     Graphique : Lineplot (date en abscisse, nombre de trajets en ordonnée, avec mise en évidence du 6 janvier 2021)
     Variables : Date de début du trajet, Nombre de trajets par jour

Contraintes et limites :

- **Données agrégées** : L'absence d'identifiant unique pour chaque vélo limite les possibilités de reconstitution précise des itinéraires ou des distances parcourues.
- **Données contextuelles limitées** : Les événements exceptionnels (grèves, manifestations, travaux, etc.) ou les changements d'infrastructure (nouvelles pistes cyclables, réaménagements urbains) ne sont pas pris en compte dans les données disponibles.
- **Intention de l'usager** : bien qu'il soit riche en information, les jeux de données ne reflète pas nécessairement les raisons des choix des usagers (par ex: s'il n'y a plus de vélos électriques disponibles, l'usager va être contraint de prendre un vélo classique contre son gré)

L'objectif est d'aboutir à une compréhension approfondie de la pratique du cyclisme à Washington à partir de ces jeux de données, en identifiant les tendances d'usage, les comportements des usagers, et les facteurs (comme la météo ou la localisation des stations) influençant l'utilisation du service. Cette analyse vise à fournir des insights pertinents pour les décideurs publics, les urbanistes, les acteurs de la mobilité durable, ainsi que pour les citoyens investis dans l'amélioration des mobilités douces.

Partie Exploration

Importation du jeu de données et des librairies

Question 1

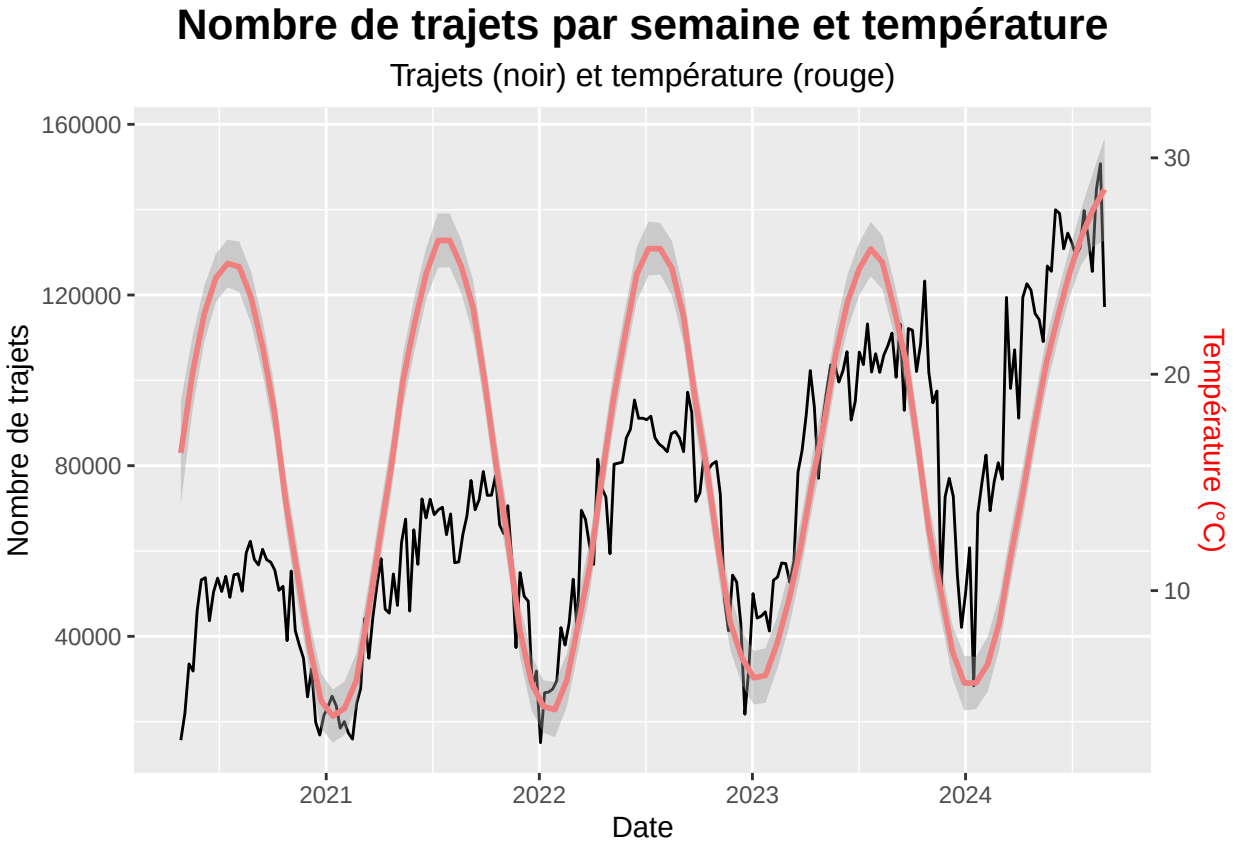
Hypothèse :

En premier lieu, une observation intéressante à faire est d'observer la tendance générale de l'utilisation des vélos au cours des années étudiées. Nous pouvons imaginer une corrélation avec la température pour représenter les saisons, et l'utilisation des vélos.

Graphique :

```r
#Préparation des données
# Nombre trajets par semaine
df$date <- as.Date(df$started_at)

df_semaine <- df %>%
  mutate(semaine = floor_date(date, "week", week_start = 1)) %>%
  group_by(semaine) %>%
  summarise(nombre_trajets_semaine = sum(n()))

#Température moyenne par semaine
weather_semaine <- weather %>%#
  mutate(semaine = floor_date(datetime, "week", week_start = 1)) %>%
  group_by(semaine) %>%
  summarise(temp = mean(temp, na.rm = TRUE))

#Facteur d'agrandissement Nombre trajet / Température
facteur_trajet_temperature <- max(df_semaine$nombre_trajets_semaine, na.rm = TRUE) / max(weather_semaine

# Visualisation
ggplot() +
  geom_line(data = df_semaine,
            aes(x = semaine, y = nombre_trajets_semaine),
            color = "black") +
  geom_smooth(data = weather_semaine,
              aes(x = semaine, y = temp * facteur_trajet_temperature),
              color = "lightcoral",
              method = "loess",
              span= 0.2,) +
  scale_y_continuous(
    name = "Nombre de trajets",
    sec.axis = sec_axis(~ . / facteur_trajet_temperature, name = "Température (°C)") ) +
  labs(title = "Nombre de trajets par semaine et température",
       x = "Date",
       subtitle = "Trajets (noir) et température (rouge)") +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    axis.title.y.right = element_text(color = "red")
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interprétation :

Nous constatons une augmentation au fil des années de l'utilisation des vélos et surtout une cyclicité. Nous observons une corrélation très forte entre la température et l'utilisation des vélos. Les utilisateurs privilégient l'utilisation des vélos lors de températures chaudes. Ainsi, nous pouvons voir que les utilisations majoritaires se font pendant le printemps, l'été et l'automne.

Certaines semaines ne suivent pas le paterne comme la semaine du 3 janvier 2022. Avec une température moyenne de 0°, l'utilisation du vélo est la plus faible du dataset avec 15 083 utilisations. Ceci peut s'expliquer par une chute de neige tombée le 3 janvier où 17,5 cm sont tombés en une journée. On observe des phénomènes similaires lors de la semaine du 19 décembre 2022, en raison de conditions météorologiques défavorables comprenant des grelons, de la neige et de la pluie. Pendant la semaine du 15 janvier 2024, on observe une chute de température importante, passant de 7°C la semaine précédente à -2,5°C. Cette semaine-là, on enregistre également des grelons, de la neige et de la pluie, ce qui contribue à une utilisation réduite des vélos.

Ces exemples nous amène à nous demander si une corrélation forte peut être associée entre la météo et l'utilisation des vélos.

Question 2

Question 2a : Y a-t-il une variation du nombre de trajets selon les jours de la semaine ? Traitement des données pour les questions 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().

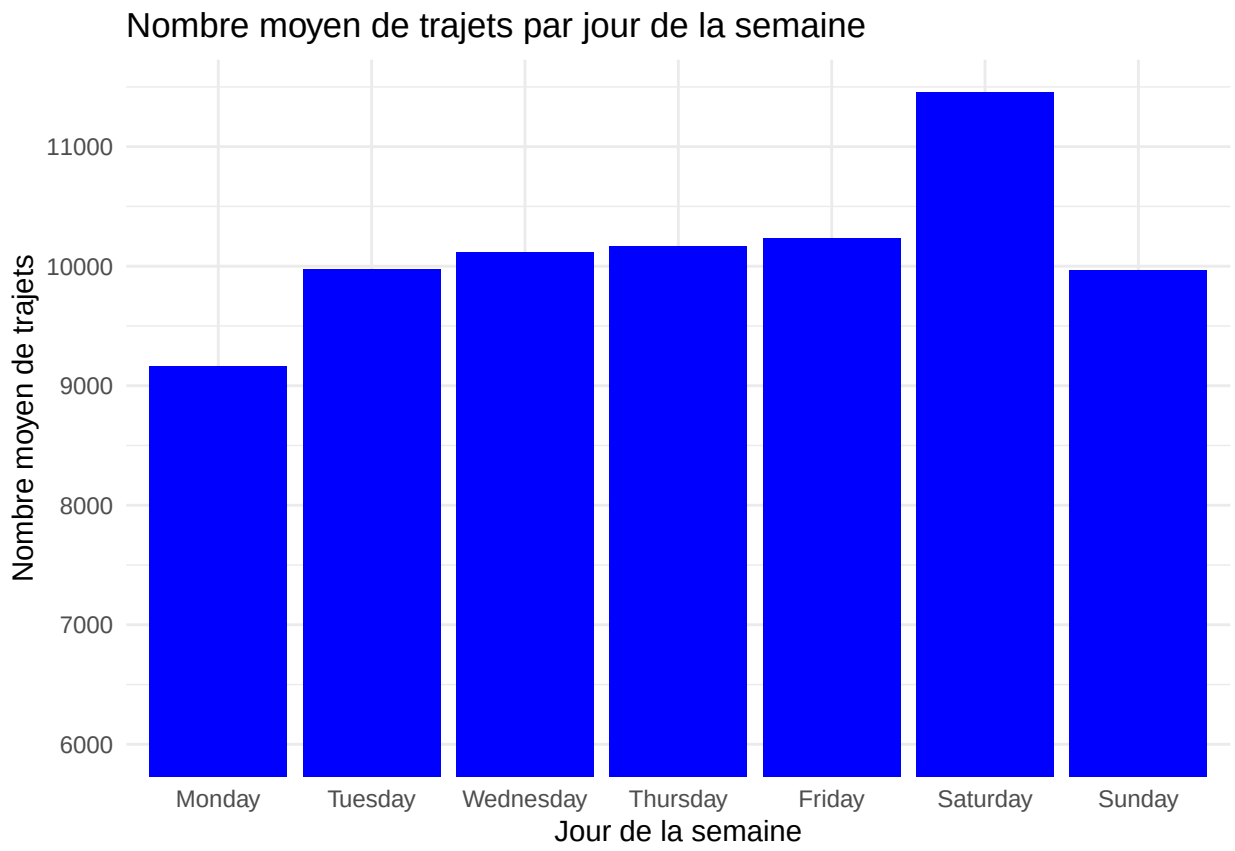
Ces transformations nous ont permis de comprendre l'évolution du nombre d'emprunts selon les moments de la semaine ou de l'année.

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 <- df %>%
  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)

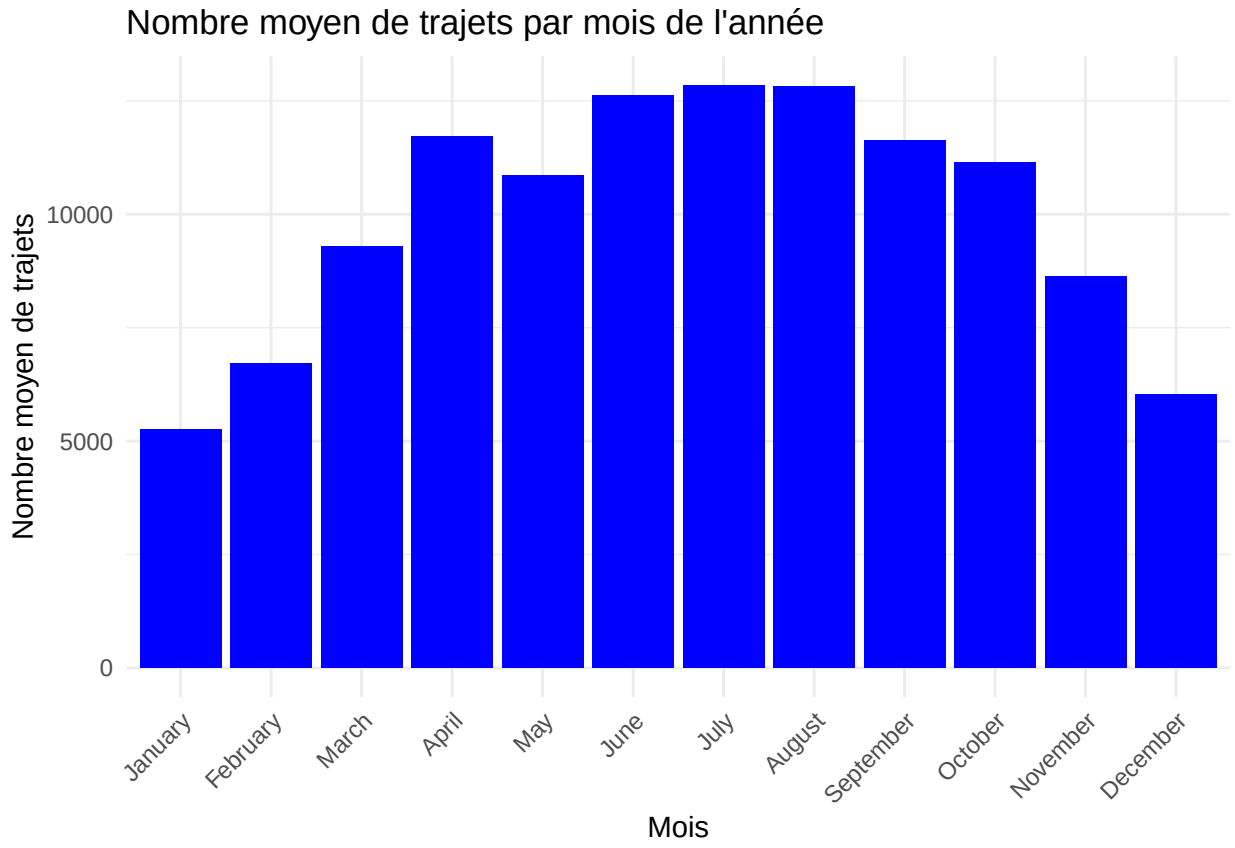
Question 2b : Y a-t-il une variation du nombre de trajets selon les mois de l'année ? Hypothèse :

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 <- df %>%
  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
  summarise(monthly_count = n(), .groups = "drop") %>%
```

```r
  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.

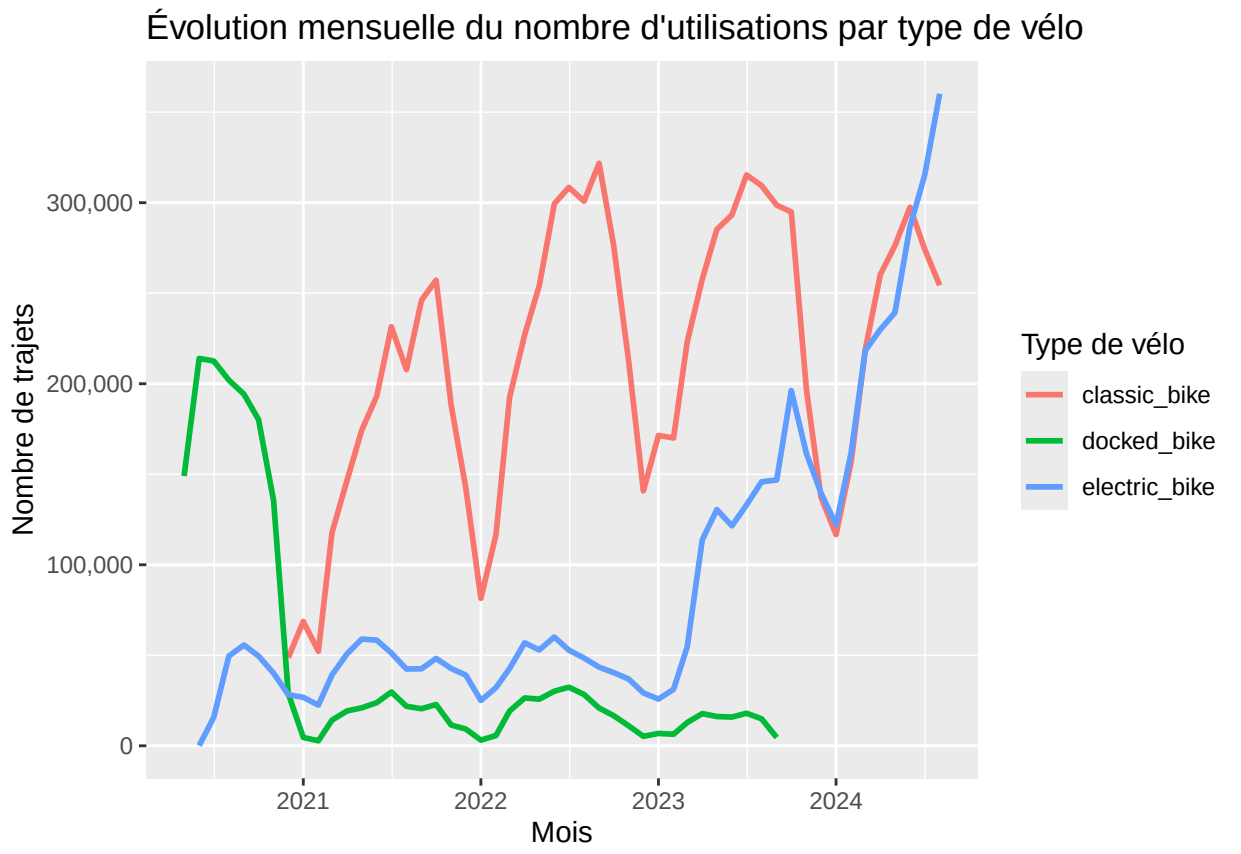
À l'inverse, les trajets diminuent nettement en hiver, notamment entre décembre et février, où le nombre moyen redescend autour de 5 000 trajets en décembre. Cette baisse reflète un usage moindre du vélo durant les mois les plus froids de l'année.

Question 4 : Evolution de l'usage des différents types de vélos

Quelle est l'évolution de l'usage des différents type de vélos ?

On peut supposer une hausse de l'utilisation de vélos électriques.

```r
monthly_usage <- df %>%
  mutate(month = floor_date(as.Date(started_at), unit = "month")) %>%
  group_by(month, rideable_type) %>%
  summarise(nb_trajets = n(), .groups = "drop")

# Lineplot
ggplot(monthly_usage, aes(x = month, y = nb_trajets, color = rideable_type)) +
  geom_line(size=1) +
  labs(
    title = "Évolution mensuelle du nombre d'utilisations par type de vélo",
    x = "Mois",
    y = "Nombre de trajets",
    color = "Type de vélo"
  ) +
  scale_y_continuous(labels = scales::comma)
```



Interprétation

L'évolution mensuelle du nombre d'utilisations par type de vélos permet de montrer plusieurs tendances intéressantes. On peut tout d'abord voir que les vélos « docked » c'est-à-dire les vélos qui se prennent et rendent à des stations ont été les plus populaires avant 2021 mais ont connu une grosse diminution au profit de vélos « classiques ». Ces vélos peuvent être posés et rendus n'importe où et ne nécessitent pas d'être rattachés à une station. Ensuite entre 2021 et 2024, le vélo classique a été largement le plus utilisé avant d'être dépassé par les vélos « électriques » qui ont connu une hausse significative à partir de 2023. Ce graphique montre également que l'utilisation des vélos suit un cycle saisonnier, aspect qui sera détaillé

dans une autre visualisation. Ces données suggèrent une modernisation de la flotte de Capital Bike Share, et une transition des utilisateurs vers des vélos plus confortables qui demandent moins d'efforts (électriques) ainsi que des vélos qui leur permettent plus de flexibilité en terme de localisation d'emprunt et de rendu (diminution des « docked »)

Question 8 :

Quelles stations sont les plus actives à différentes périodes de la journée ?

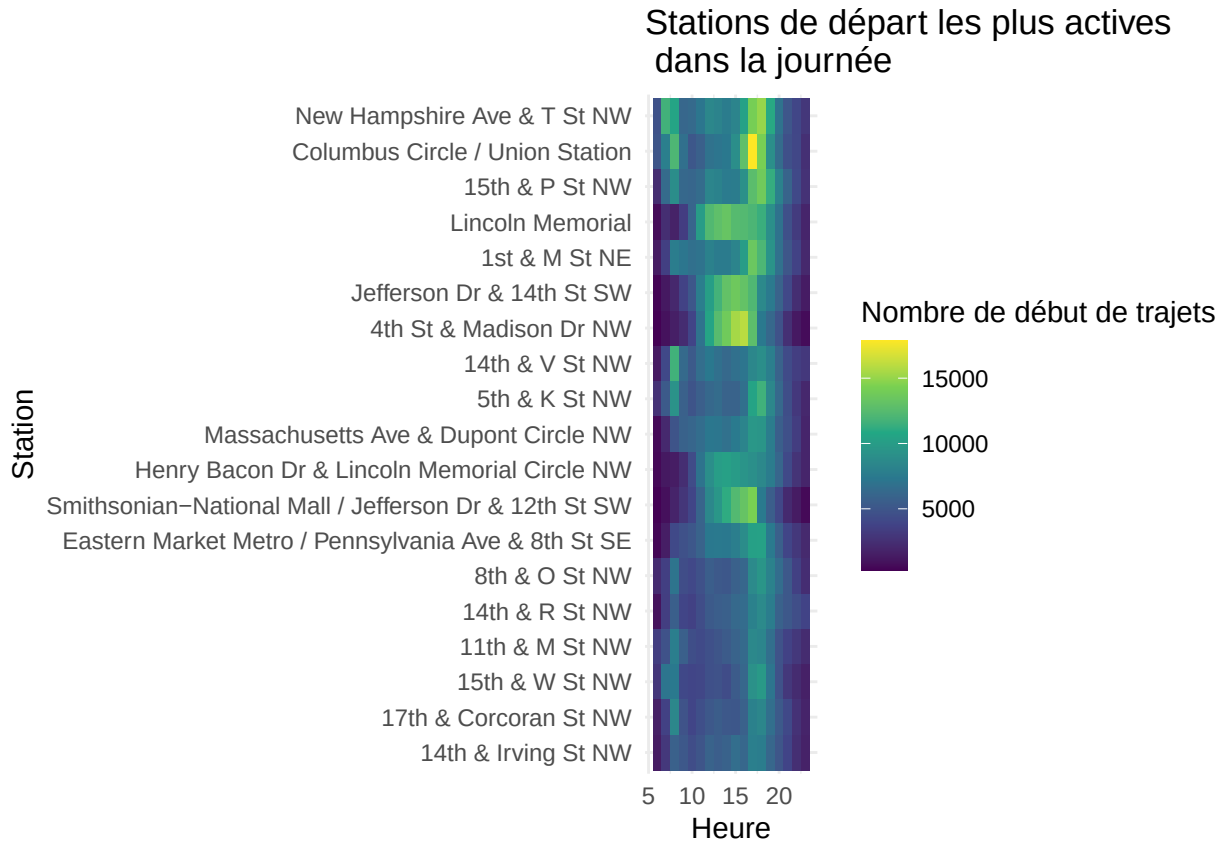
Pour cette analyse, nous avons examiné comment l'activité des stations varie selon les heures de la journée. Cette question est essentielle pour comprendre les cycles d'utilisation quotidiens et optimiser la redistribution des vélos aux moments stratégiques.

Notre démarche a consisté à analyser finement la distribution horaire des départs et des arrivées. Pour cela, nous avons :

- Extrait l'heure de départ et d'arrivée à partir des variables started_at et ended_at
- Agrégé les données par heure et par station pour quantifier le nombre de trajets
- Identifié les 20 stations les plus actives pour améliorer la lisibilité
- Créé des heatmaps visualisant l'intensité d'utilisation selon l'heure
- Filtré la période nocturne (0h-5h) pour mieux faire ressortir les tendances principales

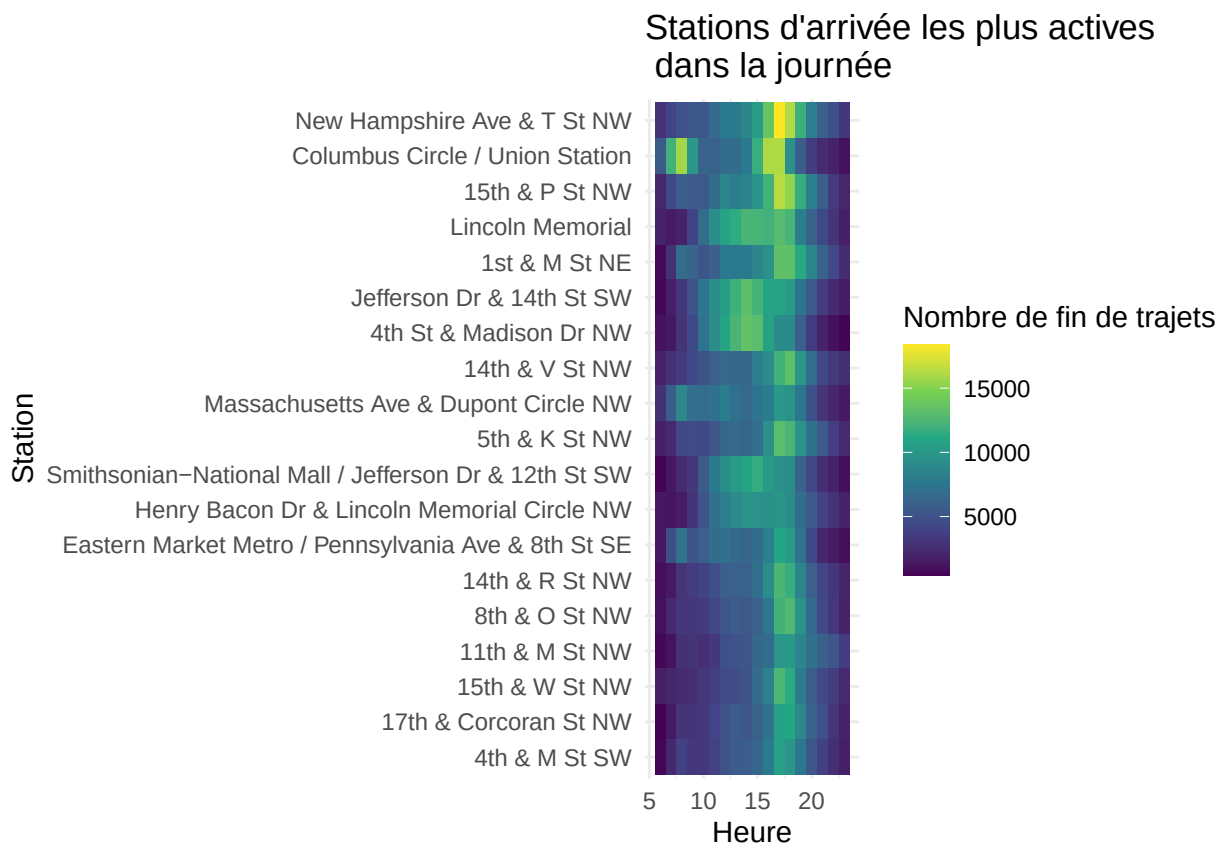
```r
# Extract heure
df$heure_start <- hour(ymd_hms(df$started_at))

# Agréger par heure et station de départ
df_heat <- df %>%
  group_by(start_station_name, heure_start) %>%
  summarise(nb_trajets = n(), .groups = "drop")

# TOP 20
top_stations <- df_heat %>%
  group_by(start_station_name) %>%
  summarise(total = sum(nb_trajets)) %>%
  top_n(20, total) %>% pull(start_station_name)

df_heat_top <- df_heat %>% filter(start_station_name %in% top_stations)

# Suppression NA (les autres stations apres filtre)
df_heat_top <- df_heat_top %>% filter(!is.na(start_station_name))

# Calculer l'ordre des stations par popularité
station_order <- df_heat_top %>%
  group_by(start_station_name) %>%
  summarise(total_trajets = sum(nb_trajets)) %>%
  arrange(desc(total_trajets)) %>%
  pull(start_station_name)

df_heat_top <- df_heat_top %>%
  mutate(start_station_name = factor(start_station_name, levels = station_order))

df_heat_top_filtre <- df_heat_top %>%
  filter(!(heure_start >= 0 & heure_start <= 5))
```

```r
# HEATMAP filtre sur 2 a 5h du matin
ggplot(df_heat_top_filtre, aes(x = heure_start, y = reorder(start_station_name, desc(start_station_name
  geom_tile() +
  scale_fill_viridis_c() +
  labs(title = "Stations de départ les plus actives \n dans la journée",
       x = "Heure", y = "Station", fill = "Nombre de début de trajets") +
  theme_minimal()
```




```r
# Extract heure
df$heure_end <- hour(ymd_hms(df$ended_at))

# Agréger par heure et station de arrivée
df_heat <- df %>%
  group_by(end_station_name, heure_end) %>%
  summarise(nb_trajets = n(), .groups = "drop")

# TOP 20
top_stations <- df_heat %>%
  group_by(end_station_name) %>%
  summarise(total = sum(nb_trajets)) %>%
  top_n(20, total) %>% pull(end_station_name)

df_heat_top <- df_heat %>% filter(end_station_name %in% top_stations)

# Suppression NA (les autres stations apres filtre)
```

```r
df_heat_top <- df_heat_top %>% filter(!is.na(end_station_name))

# Calculer l'ordre des stations par popularité
station_order <- df_heat_top %>%
  group_by(end_station_name) %>%
  summarise(total_trajets = sum(nb_trajets)) %>%
  arrange(desc(total_trajets)) %>%
  pull(end_station_name)

df_heat_top <- df_heat_top %>%
  mutate(end_station_name = factor(end_station_name, levels = station_order))

df_heat_top_filtre <- df_heat_top %>%
  filter(!(heure_end >= 0 & heure_end <= 5))

# HEATMAP filtre sur 2 a 5h du matin
ggplot(df_heat_top_filtre, aes(x = heure_end, y = reorder(end_station_name, desc(end_station_name)), fi
  geom_tile() +
  scale_fill_viridis_c() +
  labs(title = "Stations d'arrivée les plus actives \n dans la journée",
       x = "Heure", y = "Station", fill = "Nombre de fin de trajets") +
  theme_minimal()
```



Interprétation

L'analyse des heatmaps représentant l'activité des stations de départ et d'arrivée en fonction de l'heure de la journée révèle des modèles d'utilisation intéressants. Les stations les plus actives varient en fonction de

l'heure, avec des pics d'activité pendant les heures de pointe (7-9h et 16-18h).

L'analyse des stations révèle deux profils d'utilisation distincts. "New Hampshire Ave & T St NW", entourée de logements sans parking, présente un pic d'activité le matin pour les départs et le soir pour les arrivées, caractéristique d'un flux pendulaire marqué. "Columbus Circle/Union Station", connectée aux transports en commun, montre une utilisation dominante aux heures de pointe, liée aux horaires de travail.

À l'opposé, "Lincoln Memorial" et "Jefferson Dr & 14th St SW", situées dans un même parc touristique près d'agences gouvernementales, connaissent une activité plus étalée sur la journée, avec un pic en milieu de journée. Cette différence s'explique par leur fonction récréative/touristique versus professionnelle.

Les autres stations, comme "5th & K St NW", suivent également un schéma professionnel avec une activité concentrée autour des heures de pointe, reflétant une utilisation principalement liée aux déplacements domicile-travail.

Question 9 :

Quelles sont les stations les plus utilisées le week-end, en semaine ?

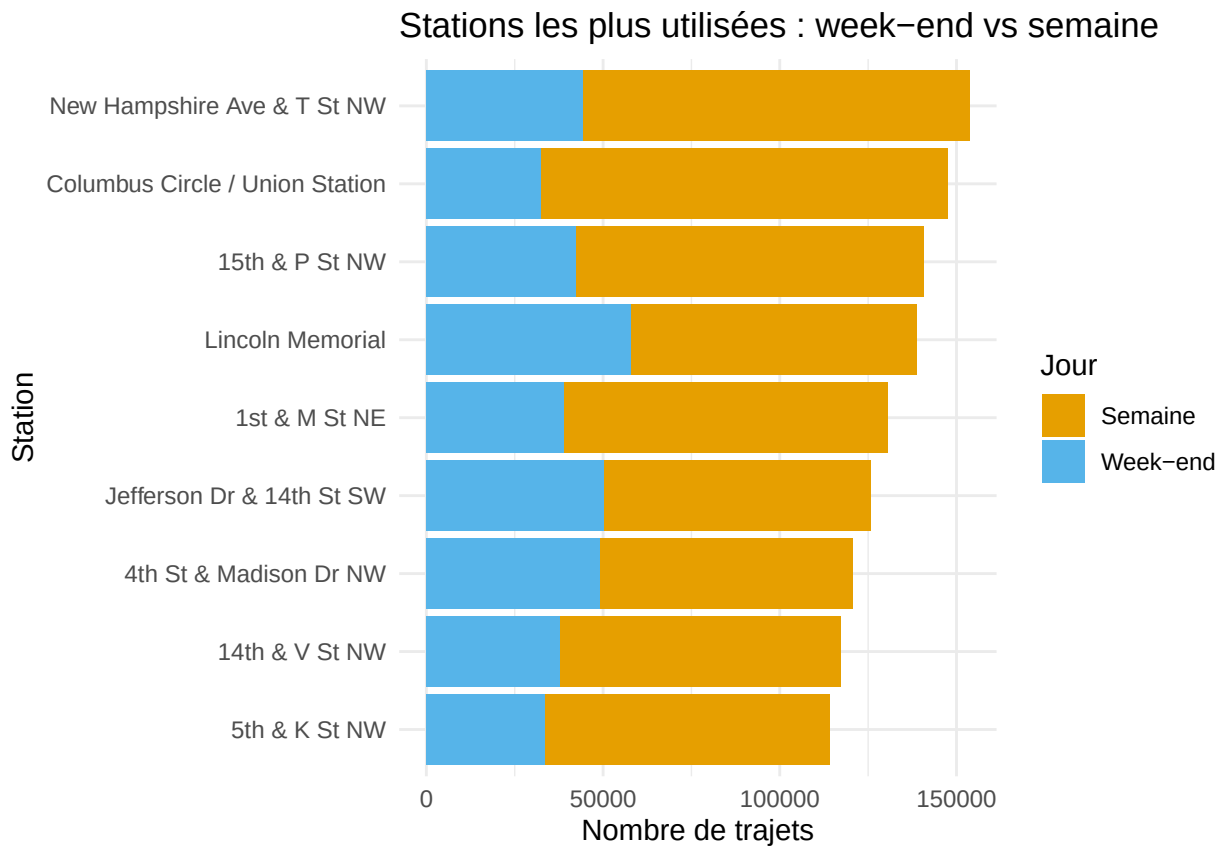
Pour cette analyse, nous avons voulu examiner comment l'utilisation des stations varie entre la semaine et le week-end. Cette question est pertinente car elle permet de distinguer les stations principalement utilisées pour les trajets domicile-travail (en semaine) de celles fréquentées pour des activités de loisirs (week-end).

Notre démarche a consisté à transformer les données temporelles pour catégoriser chaque trajet selon qu'il a lieu en semaine ou le week-end. Pour cela, nous avons :

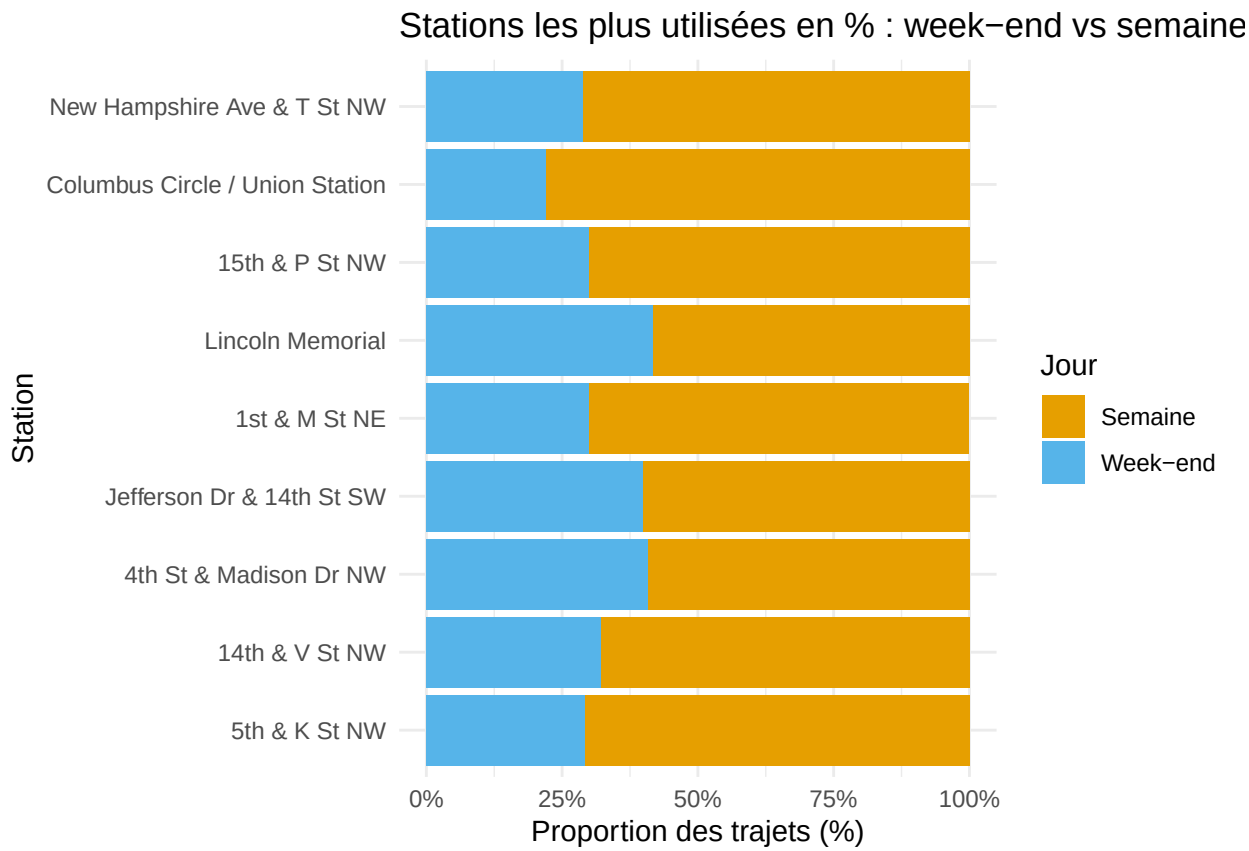
- Extrait le jour de la semaine à partir de la variable started_at
- Créé une variable binaire "type_jour" (Semaine/Week-end)
- Agrégé les données par station et par type de jour
- Filtré les 10 stations les plus actives pour améliorer la lisibilité

```
# Extract jour
df$date <- ymd_hms(df$started_at)
df$jour_semaine <- wday(df$date, label = TRUE, week_start = 1) # 1 = lundi, 7 = dimanche
df$type_jour <- ifelse(df$jour_semaine %in% c("Sat", "Sun"), "Week-end", "Semaine")

# Agréger par station et type jour
df_bar <- df %>%
  group_by(start_station_name, type_jour) %>%
  summarise(nb_trajets = n(), .groups = "drop")

# TOP 10 stations
top_stations2 <- df_bar %>%
  group_by(start_station_name) %>%
  summarise(total = sum(nb_trajets)) %>%
  top_n(10, total) %>%
  pull(start_station_name)

# Filtre du TOP 10
df_bar_top <- df_bar %>% filter(start_station_name %in% top_stations2)

# Suppression NA (les autres stations apres filtre)
df_bar_top <- df_bar_top %>% filter(!is.na(start_station_name))
```

```r
ggplot(df_bar_top, aes(x = reorder(start_station_name, nb_trajets), y = nb_trajets, fill = type_jour)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Stations les plus utilisées : week-end vs semaine",
       x = "Station", y = "Nombre de trajets", fill = "Jour") +
  scale_fill_manual(values = c("Week-end" = "#56B4E9", "Semaine" = "#E69F00")) +
  theme_minimal()
```




```r
# STACKED BARPLOT Pourcentage
ggplot(df_bar_top, aes(x = reorder(start_station_name, nb_trajets), y = nb_trajets, fill = type_jour)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  coord_flip() +
  labs(title = "Stations les plus utilisées en % : week-end vs semaine",
       x = "Station", y = "Proportion des trajets (%)", fill = "Jour") +
  scale_fill_manual(values = c("Week-end" = "#56B4E9", "Semaine" = "#E69F00")) +
  theme_minimal()
```



Interprétation

L'analyse des stations révèle deux profils d'utilisation distincts. "New Hampshire Ave & T St NW", entourée de logements sans parking, génère le plus grand volume total avec des flux pendulaires marqués. "Columbus Circle/Union Station", connectée aux transports, montre une utilisation dominante en semaine (75%) liée aux horaires de travail.

À l'opposé, "Lincoln Memorial" et "Jefferson Dr & 14th St SW", situées dans un même parc touristique près d'agences gouvernementales, présentent une forte proportion d'utilisation le week-end (40-45%). Cette différence s'explique par leur fonction récréative/touristique versus professionnelle.

Les autres stations comme "5th & K St NW" suivent également un schéma professionnel avec une faible utilisation le week-end.