cyclistic case study

Comment une société de vélos en libre-service peut-elle connaître un succès rapide ?

intro

executive summary

sources

résumé

En quoi diffèrent les usages des abonnés et des cyclistes occasionnels ?

propositions

saisonnalité
type d'usage
heure d'usage

annexes

intro

Pour cette étude de cas qui clôt le programme de formation certifiante sur l'analyse des données proposée par Google sur le site Coursera « Google Data Analytics », **j'assume le rôle d'un Data Analyst junior chez Cyclistic**, une entreprise de vélos en libre-service basée à Chicago.

Cyclistic propose plus de 5 800 vélos dans plus de 600 stations d'accueil réparties sur Chicago. La société propose deux types de vélos classiques et électriques et les commercialise à **deux segments de clientèle : les occasionnels et les membres abonnés**. Les cyclistes occasionnels sont les clients qui achètent des laissez-passer pour un trajet simple ou une journée complète. **Les membres sont les clients qui achètent des abonnements annuels de la société Cyclistic, ces derniers sont plus rentables que les cyclistes occasionnels.**

La stratégie future de la société vise à **convertir les usagers occasionnels en membres annuels (abonnés)**. Pour ce faire, l'équipe d'analystes marketing est chargée de mieux comprendre en quoi les cyclistes membres et les cyclistes occasionnels diffèrent en analysant des données historiques sur les trajets à vélo de la société de l'année 2023 afin d'identifier les tendances et de proposer une compagne publicitaire appropriée à la stratégie de l'entreprise.

Mes interlocuteurs :
Lily Moreno : La directrice du marketing et ma N+1. Lily Moreno est responsable du développement de campagnes et d'initiatives visant à promouvoir le programme de vélos en libre-service. Il peut s'agir d'e-mails, de médias sociaux et d'autres canaux.

Équipe d'analytique marketing Cyclistic : Une équipe d'analystes de données qui sont responsables de la collecte, de l'analyse et du reporting des données qui aident à guider la stratégie marketing de Cyclistic.

Équipe de direction Cyclistic : L'équipe de direction, notoirement orientée vers les détails, décidera d'approuver ou non le programme de marketing recommandé.

executive summary

J'ai analysé le data disponible pour l'année 2023 afin d'**identifier les variations de tendances** entre l'**usage des membres annuels et des utilisateurs occasionnels**.

Dans ces deux groupes, les usages diffèrent sur plusieurs points :

- **La proportion d'utilisateurs occasionnels augmente au printemps et pendant l'été**, de juin à août.
- Toute l'année, nos membres payants sont plus nombreux.
- **L'usage varie selon le jour de la semaine :**
    - Les abonnés vont favoriser un usage tout au long de la semaine avec des durées homogènes.
    - Les occasionnels iront favoriser le week-end et une durée d'utilisation extrêmement variée.
- En regardant les horaires d'utilisation :
    - **les deux groupes favorisent la soirée avec un pic à 17h.**
    - **Les trajets matinaux sont l'apanage des abonnés.**

Les recommandations que je ferai en fin de ce document se basent sur ces trois éléments de différentiation :

- **la saisonnalité**
- **le type d'usage**
- **l'heure d'usage**

sources

Pour cette analyse, j'utilise le **data mis à disposition par Cyclistic** que l'on peut retrouver ici.
(Remarque : Les jeux de données ont un nom différent car Cyclistic est une entreprise fictive).

Les données ont été mises à disposition par **Motivate International Inc.** sous cette licence.

J'utilise les données de **Janvier 2023 à Décembre 2023** pour baser mon analyse sur une année pleine.

Le Data a été téléchargé et des copies ont été stockées en local ainsi que sur google Drive.

Le data se présente sous la forme de **fichiers CSV (comma-separated values), de 13 colonnes.**

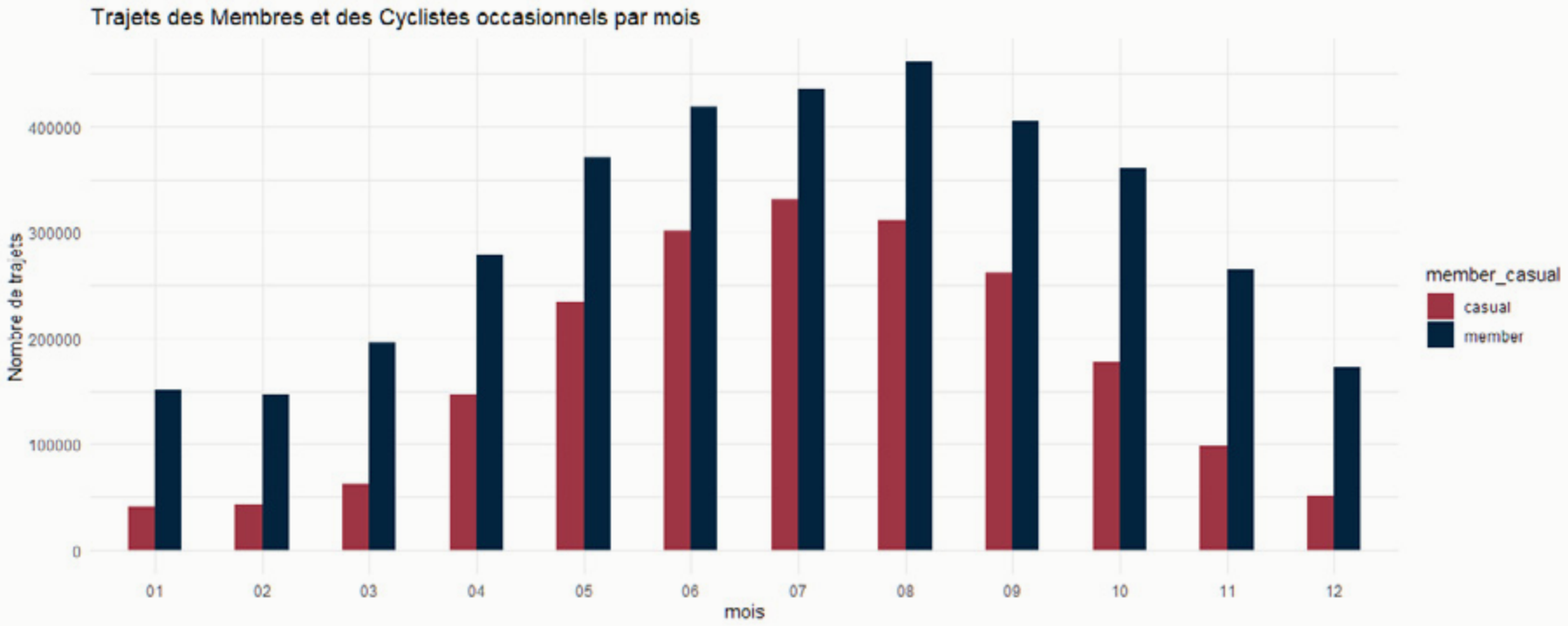
Dans l'optique du Case Study, ce dataset me permettra de répondre à la demande.
Mais des questions de confidentialité m'empêcheront d'utiliser des éléments identifiables
afin de **savoir si les utilisateurs occasionnels sont des occurences uniques**
ou si ce sont des clients récurrents.

note : vous pouvez retrouver les liens vers le code utilisé pour l'analyse ici.

En quoi diffèrent
les usages
des abonnés
et des cyclistes
occasionnels ?

résumé de l'analyse

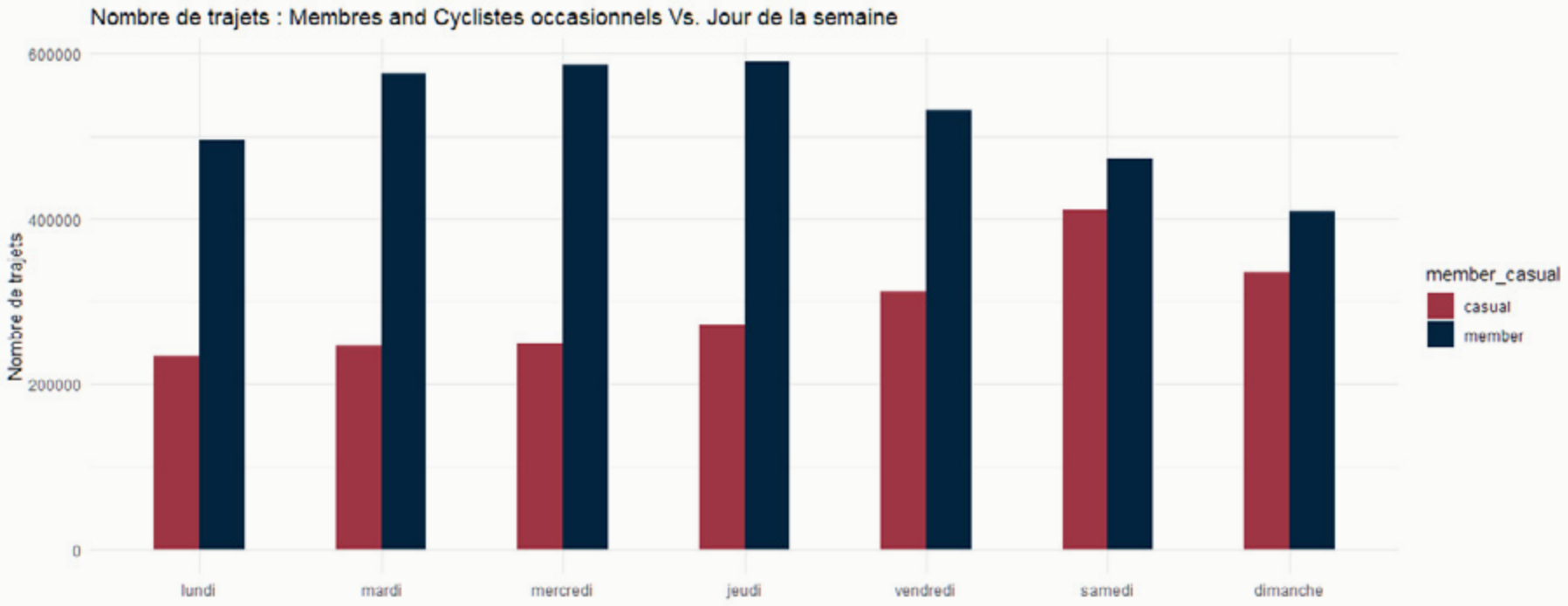
Augmentation de l'usage des cyclistes occasionnels pendant l'été.



source: cyclistic bike rides jan2023 to dec2023

résumé de l'analyse

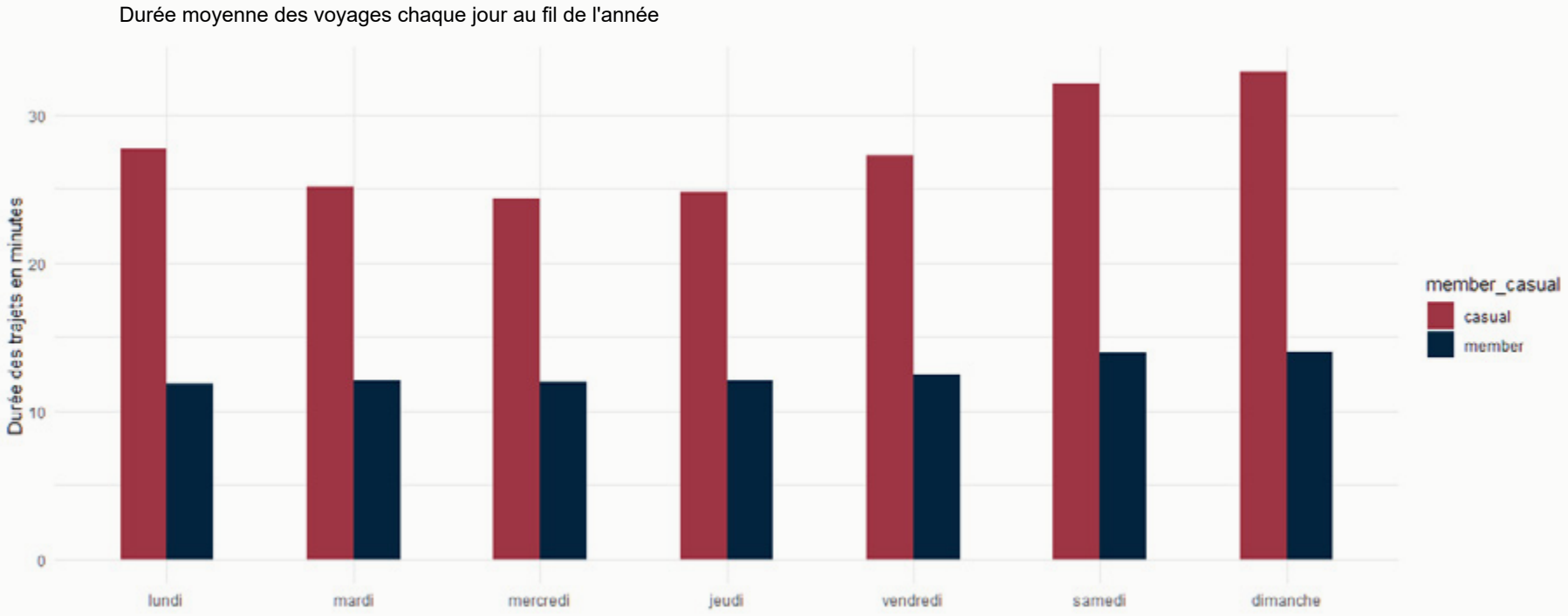
Jours semaine / Weekends
Les abonnés préfèrent **le milieu de la semaine**
alors que les cyclistes occasionnels voyagent majoritairement **le weekend**.



source: cyclistic bike rides jan2023 to dec2023

résumé de l'analyse

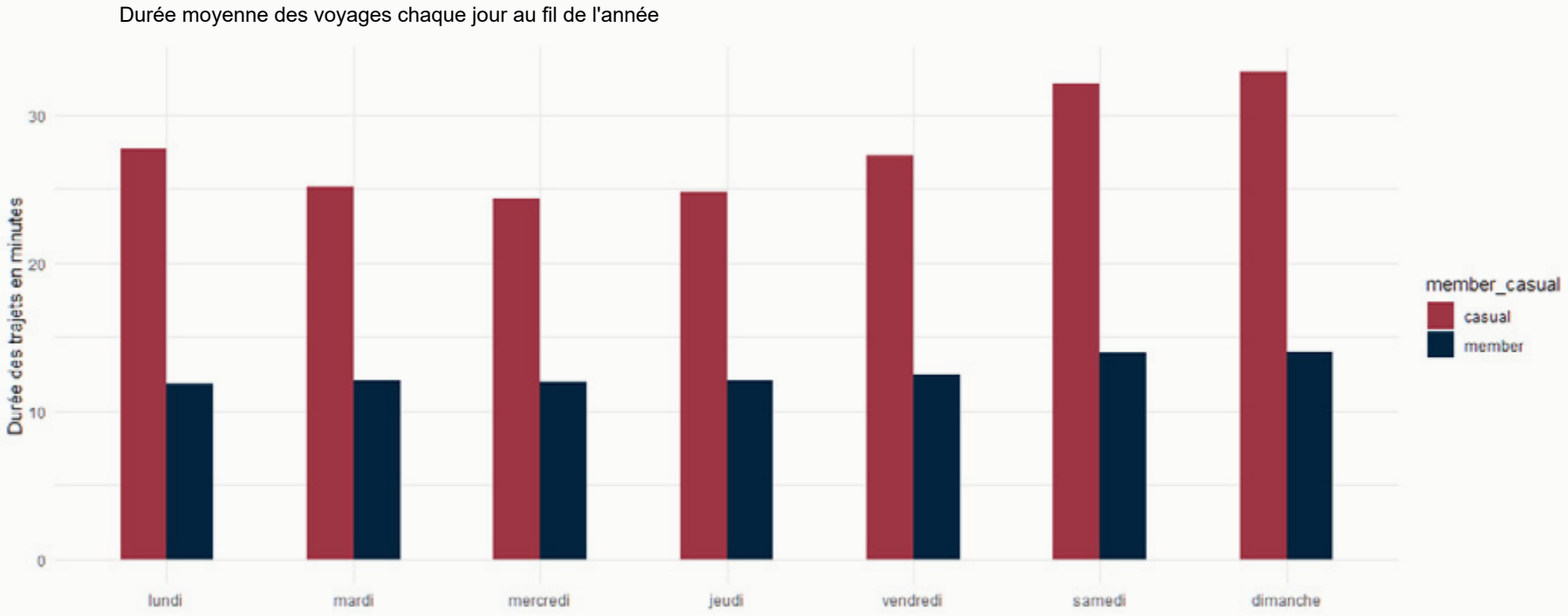
Des trajets stables pour les abonnés.
Les abonnés font des **trajets courts** et avec **globalement la même durée** tout au long de la semaine.



source: cyclistic bike rides jan2023 to dec2023

résumé de l'analyse

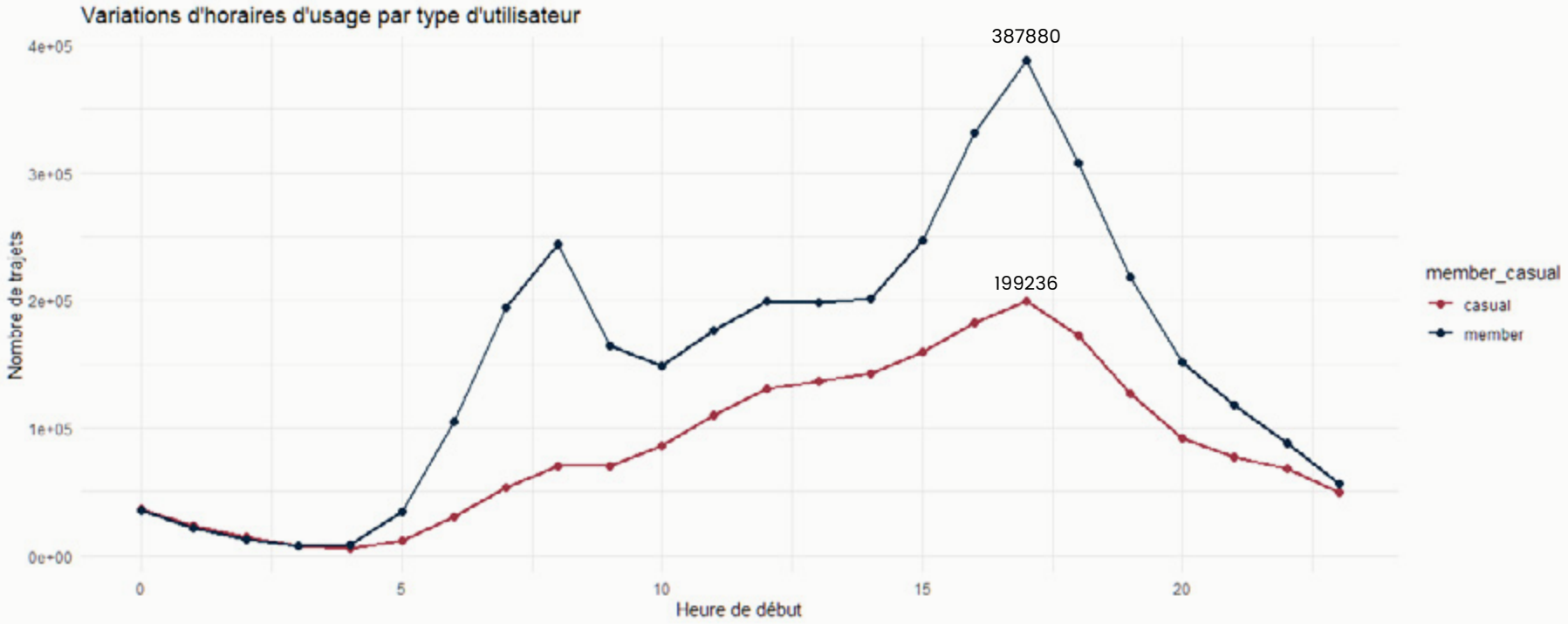
Une utilisation "week-end" pour les cyclistes occasionnels.
De manière générale, les occasionnels font des **trajets à la durée plus longue** et avec **une augmentation de la durée** pendant le week-end.



source: cyclistic bike rides jan2023 to dec2023

résumé de l'analyse

Des courbes semblables sauf le matin.
La fréquence d'utilisation est semblable pour les abonnés comme pour les cyclistes occasionnels **avec la période du matin (8h) comme différence majeure**.



source: cyclistic bike rides jan2023 to dec2023

Comment ces insights vont transformer ces cyclistes occasionnels en abonnés annuels ?

recommandations

saisonnalité : l'augmentation du printemps et de l'été

L'augmentation de la demande de nos vélos pendant le printemps et l'été nous offre une fenêtre de tir toute trouvée pour des campagnes de promotions et de publicité.

proposition :

Créer une campagne de mailing ciblant spécifiquement les cyclistes occasionnels pendant la fin du printemps et tout au long de l'été.

Prévoir des promotions et du matériel promotionnel en lien : Tarifs réduits sur l'abonnement annuel sur les 3 mois de l'été, des propositions de balades été et leurs contreparties hiver pour promouvoir l'usage all year long, faire des ventes flash de l'abonnement à prix cassés sur un an, etc.

recommandations

type d'usage : le même vélo, des habitudes différentes

Les abonnés utilisent nos vélos pour aller bosser : pendant la semaine, des pics le matin et le soir, avec des durées stables.
Les cyclistes occasionnels ont un usage récréatif : le week-end, pour aller se balader en soirée l'été.

proposition :

Créer des variations "sur mesure" de l'abonnement annuel.

Offre spéciale «Riders du week-end» pour pousser l'usage en week-ends.
Offre spéciale «Long rider» qui va chercher les aficionados des grandes balades.
Offre spéciale «Half Week» avec un tarif réduit le week-end selon l'usage en pleine semaine.
Bref, des offres qui pivotent autour de l'usage du week-end ou des variations de distance.
Organiser des événements spécifiquement pour les abonnés pendant les weekends.
Jouer sur le FOMO des occasionnels pour les amener à s'abonner.

recommandations

l'heure d'usage : l'opportunité de la tranche matinale

La seule différence notable dans les horaires d'utilisation des vélos
se situe dans la tranche matinale : entre 05:00 et 10:00.

proposition :

Communiquer sur les bienfaits sanitaires de Cyclistic pour les trajets quotidiens.

Appuyer sur les bienfaits du vélo pour la santé avec des campagnes marketing ciblées.
Les cyclistes abonnés et occasionnels ne diffèrent pas tant que ça en horaire d'usage, l'occasion présentée par le vide de la tranche matinale en terme de croissance est non négligeable.

annexes

packages Rstudio

```
tidyverse
lubridate
ggplot2
dplyr
readr
tibble
```

code pour traiter (langage : R)

```
jan_23_data <- read_csv(«202301-divvy-tripdata.csv»)
fev_23_data <- read_csv(«202302-divvy-tripdata.csv»)
mar_23_data <- read_csv(«202303-divvy-tripdata.csv»)
avr_23_data <- read_csv(«202304-divvy-tripdata.csv»)
mai_23_data <- read_csv(«202305-divvy-tripdata.csv»)
jun_23_data <- read_csv(«202306-divvy-tripdata.csv»)
jul_23_data <- read_csv(«202307-divvy-tripdata.csv»)
aug_23_data <- read_csv(«202308-divvy-tripdata.csv»)
sept_23_data <- read_csv(«202309-divvy-tripdata.csv»)
oct_23_data <- read_csv(«202310-divvy-tripdata.csv»)
nov_23_data <- read_csv(«202311-divvy-tripdata.csv»)
dec_23_data <- read_csv(«202312-divvy-tripdata.csv»)

# maintenant on va vérifier les noms de colonnes
# avant de pouvoir tout réunir pour avoir du data sur l'année

colnames(jan_23_data)
colnames(fev_23_data)
colnames(mar_23_data)
colnames(avr_23_data)
colnames(mai_23_data)
colnames(jun_23_data)
colnames(jul_23_data)
colnames(aug_23_data)
colnames(sept_23_data)
colnames(oct_23_data)
```

annexes

```
colnames(nov _ 23 _ data)
colnames(dec _ 23 _ data)

# c'est bon, il y a bien 13 colonnes partout et toutes ont les mêmes intitulés
# on réunit tout ça en un seul jeu maintenant

trip _ data <- bind _ rows(jan _ 23 _ data, fev _ 23 _ data, mar _ 23 _ data, avr _ 23 _ data, mai _ 23 _ data, jun _ 23 _
data, jul _ 23 _ data, aug _ 23 _ data, sept _ 23 _ data, oct _ 23 _ data, nov _ 23 _ data, dec _ 23 _ data)

# on découpe les dates en jours / mois / années

trip _ data$date <- as.Date(trip _ data$started _ at)
trip _ data$month <- format(as.Date(trip _ data$date), «%m»)
trip _ data$day <- format(as.Date(trip _ data$date), «%d»)
trip _ data$year <- format(as.Date(trip _ data$date), «%Y»)
trip _ data$day _ of _ week <- format(as.Date(trip _ data$date), «%A»)
colnames(trip _ data) #to get the names of all the columns

# on rajoute une colonne de 'durée de voyage'

trip _ data$ride _ length <- difftime(trip _ data$ended _ at, trip _ data$started _ at)
str(trip _ data) #on inspecte l'ajout des colonnes

trip _ data$ride _ length <- as.numeric(as.character(trip _ data$ride _ length))#on passe en numérique pour faciliter
les calculs
is.numeric(trip _ data$ride _ length) #le [1] TRUE nous le confirme
str(trip _ data)#on le revérifie parce que

#maintenant on rajoute une colonne distance parcourue en km
trip _ data$ride _ distance <- distGeo(matrix(c(trip _ data$start _ lng, trip _ data$start _ lat), ncol=2), matrix
(c(trip _ data$end _ lng, trip _ data$end _ lat), ncol=2))
trip _ data$ride _ distance <- trip _ data$ride _ distance/1000 #distance en km

# petit nettoyage des entrées où le ride _ length est négatif ou égal à 0
trip _ data _ clean <- trip _ data[!(trip _ data$ride _ length <= 0),]
glimpse(trip _ data _ clean) #création nouvelle table avec les données clean

View(trip _ data _ clean) #on retourne vérifier la version nettoyée avant de passer à la phase d'analyse
```

code pour analyser (langage : R)

```
str(trip _ data _ clean) #on reprend (depuis la veille) avec la structure du data
summary(trip _ data _ clean) #on revérifie les détails

#on va regrouper les cyclistes occasionnels et les membres et comparer leurs min, max, average et median ride _
length (durée)
trip _ data _ clean %>%
  group _ by(member _ casual) %>%
  summarise(average _ ride _ length = mean(ride _ length), median _ length = median(ride _ length),
            max _ ride _ length = max(ride _ length), min _ ride _ length = min(ride _ length))

#on va regrouper les cyclistes occasionnels et les membres et comparer le nombre de rides totaux
trip _ data _ clean %>%
  group _ by(member _ casual) %>%
  summarise(ride _ count = length(ride _ id))

#on aimerait bien avoir ce tibble en ggplot, donc je le mets dans une nouvelle table
ride _ count <- trip _ data _ clean %>%
  group _ by(member _ casual) %>%
  summarise(ride _ count = length(ride _ id))

# je visualise ma nouvelle table qui devrait être comme le tibble
View(ride _ count)

ggplot(ride _ count, aes(x = member _ casual, y = ride _ count, fill = member _ casual)) +
  geom _ bar(stat = «identity») +
  geom _ text(aes(label = ride _ count), vjust = -0.5) +
  labs(title = «Nombre de trajets : membres Vs cyclistes occasionnels»,
       x = «Type d'utilisateur»,
       y = «Nombre de trajets») +
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))
# on remarque la large différence entre le nombre de voyages des utilisateurs abonnés et des occasionnels

# maintenant on va découper ces résultats par jour

# organisation des jours
trip _ data _ clean$day _ of _ week <- ordered(trip _ data _ clean$day _ of _ week, levels=c(«lundi», «mardi», «mercredi»,
«jeudi», «vendredi», «samedi», «dimanche»))
```

annexes

```
trip _ data _ clean %>%
  group _ by(member _ casual, day _ of _ week) %>%  #groupage par member _ casual
  summarise(number _ of _ rides = n() #calcul du total de voyages et de leur durée moyenne
  ,average _ ride _ length = mean(ride _ length),.groups=»drop») %>% # calcul de la durée moyenne
  arrange(member _ casual, day _ of _ week) #sort

#on met ça dans une nouvelle table qu'on visionne
ride _ count _ daily <- trip _ data _ clean %>%
  group _ by(member _ casual, day _ of _ week) %>%  #groupage par member _ casual
  summarise(number _ of _ rides = n() #calcul du total de voyages et de leur durée moyenne
            ,average _ ride _ length = mean(ride _ length/60),.groups=»drop») %>% # calcul de la durée moyenne en mi-
nutes
  arrange(member _ casual, day _ of _ week) #sort
View(ride _ count _ daily)

# on ggplot le tout
ggplot(ride _ count _ daily, aes(x = day _ of _ week, y = number _ of _ rides, fill = member _ casual)) +
  labs(title =»Nombre de trajets : Membres and Cyclistes occasionnels Vs. Jour de la semaine»,
       x = «»,
       y = «Nombre de trajets») +
  geom _ col(width=0.5, position = position _ dodge(width=0.5)) +
  scale _ y _ continuous(labels = function(x) format(x, scientific = FALSE))+
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))

# tant qu'à faire, avec cette table on profite pour regarder la durée moyenne

ggplot(ride _ count _ daily, aes(x = day _ of _ week, y = average _ ride _ length, fill = member _ casual)) +
  labs(title =»Durée moyenne des trajets : Membres and Cyclistes occasionnels Vs. Jour de la semaine»,
       x = «»,
       y = «Durée des trajets en minutes») +
  geom _ col(width=0.5, position = position _ dodge(width=0.5)) +
  scale _ y _ continuous(labels = function(x) format(x, scientific = FALSE))+
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))

# maintenant on va vérifier l'utilisation mensuelle

trip _ data _ clean %>%
  group _ by(member _ casual, month) %>%
  summarise(number _ of _ rides = n(),.groups=»drop») %>%
  arrange(member _ casual, month)
```

annexes

```
number _ of _ rides <- trip _ data _ clean %>%
  group _ by(member _ casual, month) %>%
  summarise(number _ of _ rides = n(),.groups=»drop») %>%
  arrange(member _ casual, month)

ggplot(number _ of _ rides, aes(x = month, y = number _ of _ rides, fill = member _ casual)) +
  labs(title =»Trajets des Membres et des Cyclistes occasionnels par mois»,
       x = «mois»,
       y = «Nombre de trajets») +
  theme(axis.text.x = element _ text(angle = 45)) +
  geom _ col(width=0.5, position = position _ dodge(width=0.5)) +
  scale _ y _ continuous(labels = function(x) format(x, scientific = FALSE))+
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))


# comparons la distance moyenne parcourue entre membres et cyclistes occasionnels

trip _ data _ clean %>%
  group _ by(member _ casual) %>% drop _ na() %>%
  summarise(average _ ride _ distance = mean(ride _ distance))

average _ ride _ distance <- trip _ data _ clean %>%
  group _ by(member _ casual) %>% drop _ na() %>%
  summarise(average _ ride _ distance = mean(ride _ distance))
View(average _ ride _ distance)

ggplot(average _ ride _ distance, aes(x = member _ casual, y = average _ ride _ distance, fill = member _ casual)) +
  geom _ bar(stat = «identity») +
  geom _ text(aes(label = round(average _ ride _ distance, 2)), vjust = -0.5) +
  labs(title = «Distance moyenne des trajets par type d'utilisateur»,
       x = «»,
       y = «Distance moyenne (km)») +
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))

# testons voir les heures de trajet

# Ajouter une colonne pour l'heure de début des trajets
trip _ data _ clean <- trip _ data _ clean %>%
  mutate(start _ hour = lubridate::hour(started _ at))

# Calculer le nombre de trajets par heure pour les membres et les utilisateurs occasionnels
```

annexes

```
hourly _ usage <- trip _ data _ clean %>%
  group _ by(member _ casual, start _ hour) %>%
  summarise(ride _ count = n()) %>%
  ungroup()

# Visualiser les variations d'horaires d'usage
#Calculer les valeurs maximales pour chaque groupe
max _ values <- hourly _ usage %>%
  group _ by(member _ casual) %>%
  summarize(max _ ride _ count = max(ride _ count),
            max _ hour = start _ hour[which.max(ride _ count)])

ggplot(hourly _ usage, aes(x = start _ hour, y = ride _ count, color = member _ casual, group = member _ casual)) +
  geom _ line(size = 1) +
  geom _ point(size = 2) +
  labs(title = «Variations d'horaires d'usage par type d'utilisateur»,
       x = «Heure de début»,
       y = «Nombre de trajets») +
  theme _ minimal() +
  scale _ color _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))

daily _ avg _ duration <- ride _ count _ daily %>%
  group _ by(day _ of _ week, member _ casual) %>%
  summarize(avg _ duration = mean(duration))

view(daily _ avg _ duration)

ggplot(number _ of _ rides, aes(x = day _ of _ week, y = avg _ duration, fill = member _ casual)) +
  labs(title =»Durée moyenne des voyages chaque jour au fil de l'année»,
       x = «mois»,
       y = «Nombre de trajets») +
  theme(axis.text.x = element _ text(angle = 45)) +
  geom _ col(width=0.5, position = position _ dodge(width=0.5)) +
  scale _ y _ continuous(labels = function(x) format(x, scientific = FALSE))+
  theme _ minimal() +
  scale _ fill _ manual(values = c(«member» = «#042540», «casual» = «#A13646»))
```

annexes

sites webs et documentation

datanovia.com
palletton.com

documentation RStudio

Github

Reddit/R

kaggle

retrouvez moi sur www.charlesgrillet.com

merci