

# Cyclistic viajes 2022: Explorando la disparidades entre socios anuales y casuales

Francisco Valam Cortes

27 de Noviembre de 2023

## **Cyclistic informe de viajes 2022**

Me complace presentar el informe anual de viajes correspondientes al año 2022, en el cual se detallan las tendencias, los patrones y los hallazgos significativos que nos muestran las diferencias entre los socios anuales y los ciclistas ocasionales con respecto al uso de las bicicletas de Cyclistic.

### **Marco**

El año 2022 fue un periodo interesante para los objetivos de nuestra compañía, con el aumento de usuarios con respecto a los pasados, con nuestro enfoque en el futuro, la evolución de la movilidad urbana y el constante crecimiento de nuestra base de clientes, nuestro equipo de análisis de datos ha realizado una exaustiva evaluación de los datos de los viajes recopilados a lo largo del año para ofrecer información valiosa que pueda direccionar nuestras estrategias futuras.

### **Objetivos del informe**

El objetivo de este informe es proporcionar una visión general e integral de los patrones y hallazgos que diferencian a nuestros socios anuales de los ciclistas ocasionales durante el transcurso del año 2022 para identificar área de oportunidad y presentar recomendaciones para optimizar nuestra oferta de servicio de bicicletas.

### **Alcance**

El informe abarca diferentes aspectos de los viajes realizados en 2022, incluyendo la distribución de viajes por tipo de bicicleta elegida, análisis mensual y diario, distribución de los miembros anuales frente a los usuarios casuales, y la utilización de estaciones específicas.

## Preparando ambiente para las visualizaciones

Para trabajar con los datos se utilizaron los siguientes paquetes, que son utilizados para conexiones a bases de datos SQL, visualizaciones, manejos de datos y estadísticas.

```
library(RODBC)      # Conexiones SQL
library(conflicted) # Tratar conflictos en el código
library(DescTools)   # Estadísticas descriptivas

# Solución a ciertas sobreescrituras de funciones con el paquete tidyverse

conflict_prefer("filter", "dplyr")

## [conflicted] Will prefer dplyr::filter over any other package.

conflict_prefer("lag", "dplyr")

## [conflicted] Will prefer dplyr::lag over any other package.

library(tidyverse) # Manipulación de datos

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4    v readr     2.1.4
## vforcats   1.0.0    v stringr   1.5.1
## v ggplot2   3.4.4    v tibble    3.2.1
## v lubridate 1.9.3    v tidyr    1.3.0
## v purrr    1.0.2

## Carga de datos para trabajar desde una base SQL

con_str <- "DRIVER={SQL Server};SERVER=NIGGALAP;DATABASE=Cyclistic;Trusted_Connection=yes"
con <- odbcDriverConnect(con_str)
query <- "SELECT * FROM Cyclistic.dbo.Dim_Trips_2022"
df_base <- sqlQuery(con, query) # Hacemos una petición a la base de datos
odbcClose(con) # Una vez realizada la petición cerramos la conexión

## Mostramos algunos de los datos obtenidos en nuestra variable.
glimpse(head(df_base, n = 10))

## Rows: 10
## Columns: 13
## $ ride_id          <chr> "12E2DF2F1134D6BE", "12E2E6A5E77F1D3E", "12E436E837~"
## $ rideable_type    <chr> "electric_bike", "classic_bike", "electric_bike", "~"
## $ started_at        <chr> "2022-01-18 08:33:17.0000000", "2022-01-09 07:53:43~"
## $ ended_at          <chr> "2022-01-18 08:50:58.0000000", "2022-01-09 08:03:05~"
## $ start_station_name <chr> "Prairie Ave & Garfield Blvd", "Mies van der Rohe W~"
## $ start_station_id  <chr> "TA1307000160", "13338", "13193", "TA1305000017", "~"
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "Kingsbury St & Erie S~"
## $ end_station_id    <chr> "TA1306000003", "13265", "Hubbard Bike-checking (LB~"
## $ start_lat          <dbl> 41.93000, 41.89695, 41.92182, 41.86732, 41.97000, 4~
## $ start_lng          <dbl> -87.70000, -87.62176, -87.64402, -87.64863, -87.700~
## $ end_lat            <dbl> 41.88872, 41.89381, 41.95000, 41.87219, 41.96000, 4~
## $ end_lng            <dbl> -87.64445, -87.64170, -87.68000, -87.66150, -87.680~
## $ member_casual      <chr> "member", "member", "member", "casual", "~"
```

## Exploración de los datos

En esta sección nos encargaremos de explorar los datos para familiarizarnos con la información y su estructura.

```
dim(df_base) #Número de registros y variables.
```

```
## [1] 5660383      13
```

```
colnames(df_base) #Nombre de todas las variables
```

```
## [1] "ride_id"          "rideable_type"    "started_at"
## [4] "ended_at"          "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"   "start_lat"
## [10] "start_lng"         "end_lat"        "end_lng"
## [13] "member_casual"
```

Concluimos que hay más de 5 millones y medio en registros de viajes para el año 2022, los cuales tiene variables como su ID de viaje, tipo de bicicleta utilizada para el viaje, fecha y hora de inicio y fin del viaje, así como nombres, IDs, coordenadas en latitud y longitud de las estaciones dónde inició y finalizó el mismo, y por último el tipo de cliente que realizó el viaje.

## Estadísticas descriptivas

Para poder conocer la distribución de los datos es útil un breve resumen de las estadísticas de los datos que nos permitirán conocer, por ejemplo, los rangos en que se encuentran los tiempos de viaje, o en qué coordenadas se concentran la mayor parte de los viajes registrados.

```
#Convertiremos las columnas de fecha, ya que son leídas como tipo <Chr> al ser importadas
df_base$started_at <- as.POSIXlt(df_base$started_at,format = "%Y-%m-%d %H:%M:%S")
df_base$ended_at <- as.POSIXlt(df_base$ended_at,format = "%Y-%m-%d %H:%M:%S")

#Breve resumen estadístico de cada variable numérica
summary(df_base[,c("started_at","ended_at","start_lat", "start_lng", "end_lat","end_lng")])
```

```
##      started_at                  ended_at
## Min.   :2022-01-01 00:00:05.00   Min.   :2022-01-01 00:01:48.00
## 1st Qu.:2022-05-28 19:28:51.50  1st Qu.:2022-05-28 19:49:08.00
## Median :2022-07-22 15:18:48.00  Median :2022-07-22 15:37:35.00
## Mean   :2022-07-20 08:01:46.68  Mean   :2022-07-20 08:18:06.33
## 3rd Qu.:2022-09-16 07:43:34.00  3rd Qu.:2022-09-16 07:56:29.00
## Max.   :2022-12-31 23:59:26.00  Max.   :2023-01-01 18:09:37.00
## NA's    :88                      NA's    :88
##      start_lat                 start_lng
## Min.   :41.65      Min.   :-87.83   Min.   :41.65      Min.   :-87.83
## 1st Qu.:41.88     1st Qu.:-87.66   1st Qu.:41.88     1st Qu.:-87.66
## Median :41.90     Median :-87.64   Median :41.90     Median :-87.64
## Mean   :41.90     Mean   :-87.65   Mean   :41.90     Mean   :-87.65
## 3rd Qu.:41.93     3rd Qu.:-87.63   3rd Qu.:41.93     3rd Qu.:-87.63
## Max.   :45.64     Max.   :-73.80   Max.   :42.06     Max.   :-87.53
##
```

Este breve resumen nos muestra que nuestro registros de viajes comienza a subir durante el mes de MAYO, y caen hasta noviembre y siendo que la mayor concentración de servicios se registran en el mes de Julio, teniendo en cuenta de que son recuentos totales, incluyendo miembros anuales y casuales.

## Transformación, limpieza y visualización

Ahora que se conoce la estructura de los datos y existe familiaridad, se buscará un enfoque nos permita responder la siguiente pregunta:

**¿En qué se diferencian los socios anuales y los ciclistas ocasionales con respecto al uso de las bicicletas de Cyclistic?**

Para responder realizaremos una serie de transformaciones al set de datos, así como la exploración de diversos comportamientos entre la relación de ciertas variables.

Algunas de las variables que crearemos serán las siguientes:

- Duración del viaje en min.
- Día en que se realizó el viaje.
- Mes en que se realizó el viaje

Para la creación de estas variables realizamos lo siguiente:

```
#Relizamos la diferencia entre las 2 fechas para obtener la duración y
# es ajustada a minutos

df_base <- df_base %>%
  mutate(ride_length_min = round((ended_at - started_at)/60,digits = 2))

#Ajustamos el tipo de variable y su formato, ya que el valor final lleva texto.

df_base$ride_length_min <- as.numeric(sub(" secs", "", df_base$ride_length_min))

## Extraemos el día en que se realizó cada viaje y lo asignamos a una columna

df_base <- df_base %>%
  mutate(day_of_week=weekdays(started_at))

#Extraemos el mes en que se realizó el viaje y lo asignamos a una columna

df_base <- df_base %>%
  mutate(month=months(started_at))
```

## Gráficas

Generadas las columnas necesarias, realizamos la representación grafica de los viajes a través del año para observar las tendencias en el uso un tipo de vehículo sobre otro.

Una forma de comenzar a ver lo anterior es con una gráfica de caja:

```
##Quantiles y Límites
df_quantiles <- quantile(df_base$ride_length_min, na.rm = TRUE)
```

```

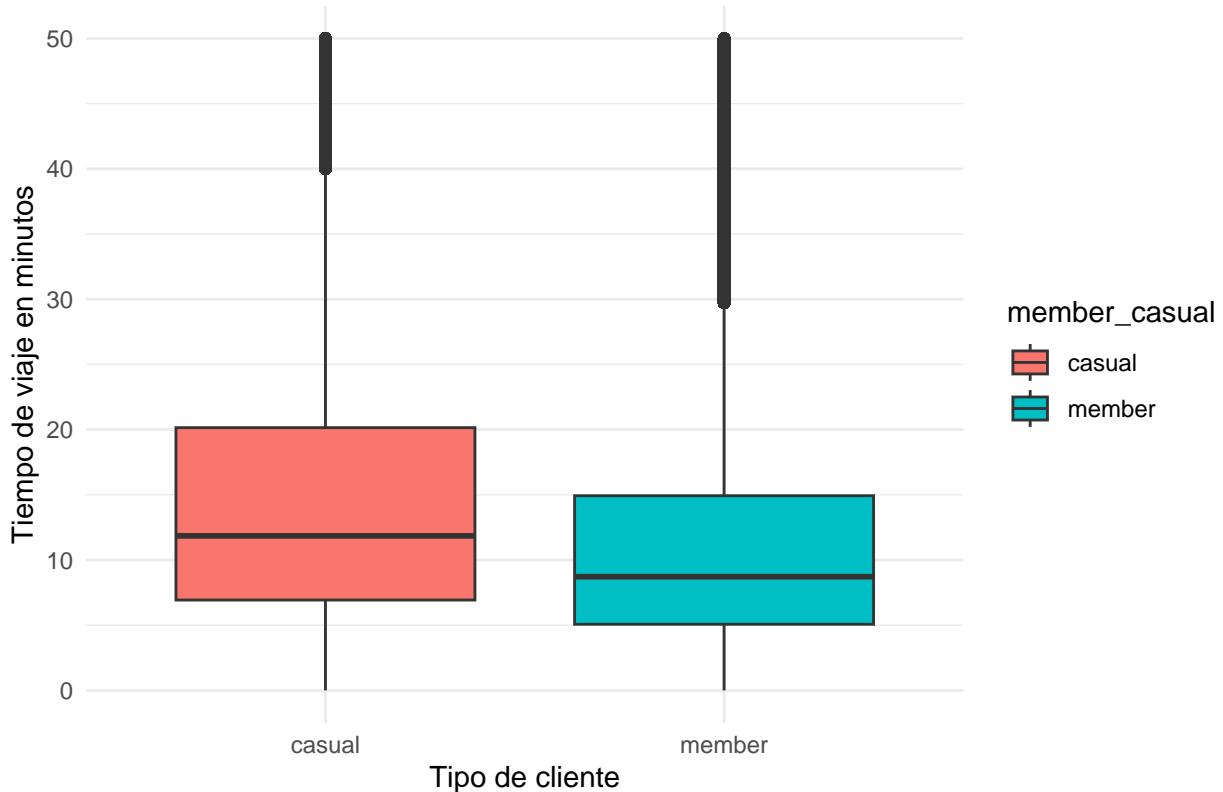
RIC <- df_quantiles["75%"] - df_quantiles["25%"]
LS <- df_quantiles["75%"] + (RIC * 1.5)
LI <- df_quantiles["25%"] - (RIC * 1.5)

# Obtiene los datos que cumplen con los límites
df_clean <- df_base %>% filter(ride_length_min >= LI &
                                ride_length_max <= LS)

#Boxplot
ggplot(data = df_base, aes(x=member_casual, y=ride_length_min, fill=member_casual))+ 
  geom_boxplot() +
  theme_minimal() +
  ylim(0, 50) +
  labs(title = "Gráfico de caja por minutos de viaje", x= "Tipo de cliente",
       y="Tiempo de viaje en minutos")

```

Gráfico de caja por minutos de viaje



```
rm(df_clean)
```

Ahora podemos observar la distribución de la duración de los viajes, tanto para los miembros anuales como para los casuales, mostrando que usualmente los miembros anuales suelen tener recorridos mas cortos con un rango entre los 5 y  $\pm 15$  minutos, mientras que los clientes casuales suelen tener recorridos entre los  $\pm 12$  y  $\pm 18$  minutos.

Ahora continuamos con mostrar el número de viajes realizados a lo largo del año.

```

# Filtración de información necesaria para el gráfico

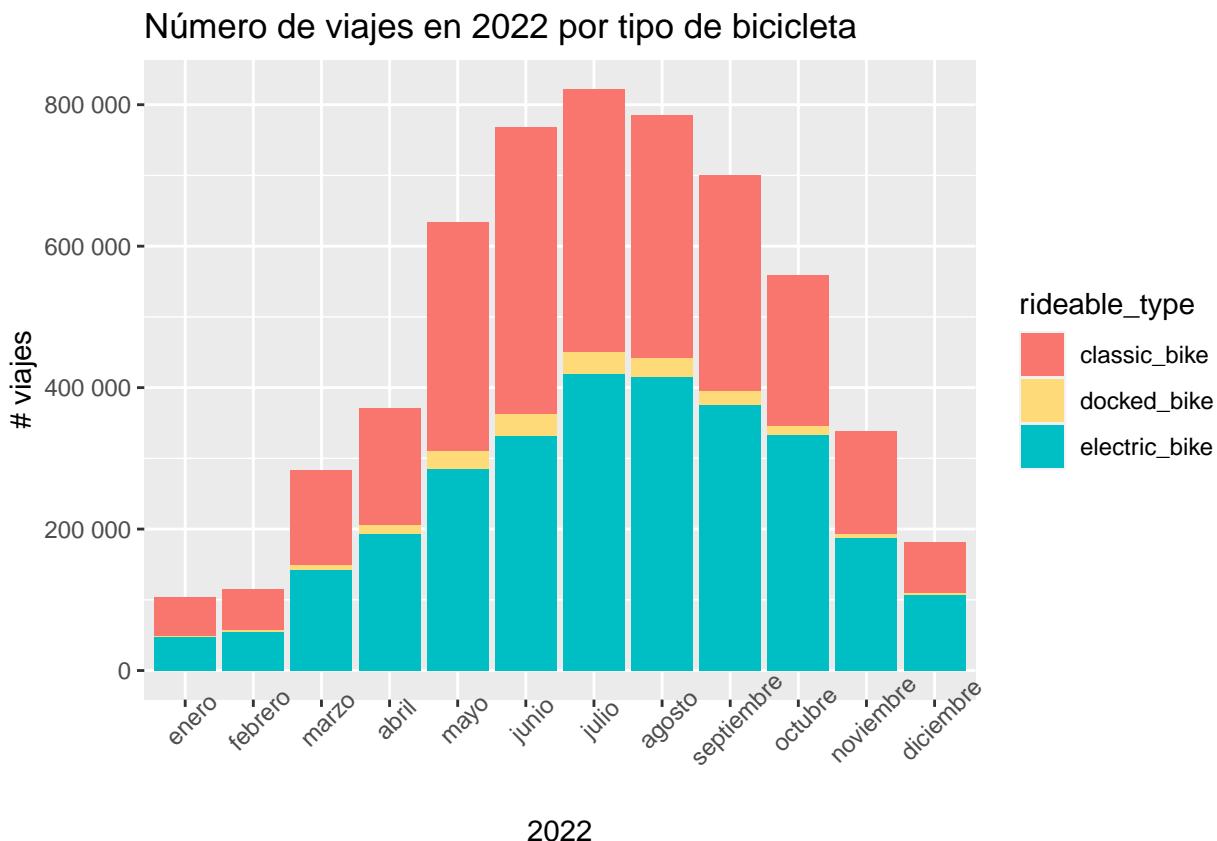
df_by_dates <- df_base[,c("rideable_type", "month", "day_of_week", "member_casual")] %>%
  group_by(rideable_type, month, day_of_week, member_casual) %>%
  summarise(observations = n())

## `summarise()` has grouped output by 'rideable_type', 'month', 'day_of_week'.
## You can override using the '.groups' argument.

#Grafica de barras a lo largo del año y rellena por los tipos de bicicletas.

ggplot(data = df_by_dates, aes(x = factor(month, levels = c("enero", "febrero",
                                                               "marzo", "abril", "mayo",
                                                               "junio", "julio", "agosto",
                                                               "septiembre", "octubre",
                                                               "noviembre", "diciembre")),
                                 y = observations, fill=rideable_type)) +
  geom_bar(stat = "identity", position = "stack") +
  theme(axis.text.x = element_text(angle = 45)) +
  scale_fill_manual(values = c("classic_bike"="#f8766d", "docked_bike"="#ffda78",
                               "electric_bike"="#00bfc4")) +
  scale_y_continuous(labels = scales::number_format(scale=1, accuracy = 1)) +
  labs(title = "Número de viajes en 2022 por tipo de bicicleta", x="2022", y= "# viajes")

```



Cómo podemos observar nuestros datos siguen una distribución normal, dónde el número de viajes alcanza

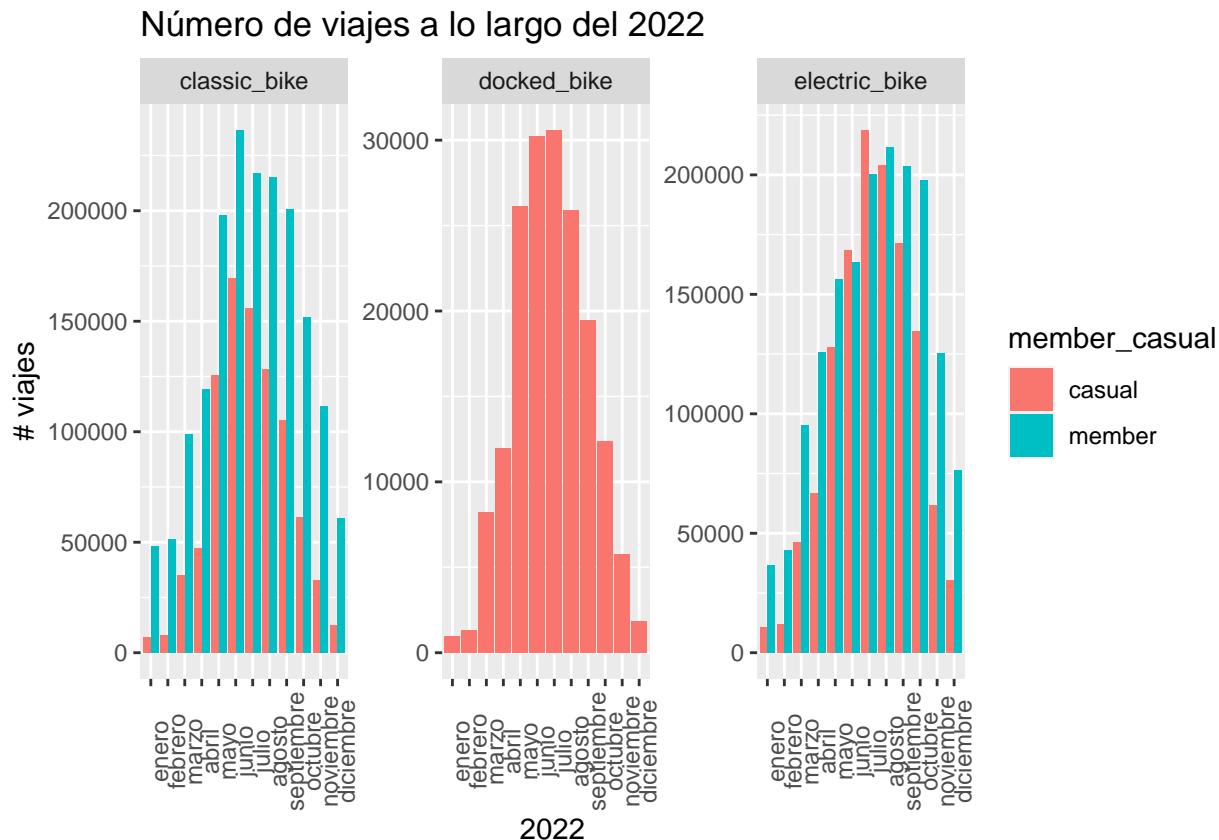
su punto máximo entre los meses de Junio y Agosto. Específicamente, las bicicletas, tanto eléctricas como acopladas, alcanzan su punto máximo en JULIO, mientras que las bicicletas clásicas lo hacen en JUNIO.

Ahora que conocemos la tendencia de los viajes, mostraremos como se ve la misma para ambos tipos de clientes a lo largo del año 2022.

```
# Seleccionamos sólo la información que necesitamos para nuestra gráfica
df_by_clients <- df_base[,c("rideable_type", "ride_length_min", "day_of_week", "month",
                           "member_casual")]

# Ordenamos la información
df_by_clients$month <- factor(df_by_clients$month, levels = c("enero", "febrero",
                                                               "marzo", "abril", "mayo",
                                                               "junio", "julio", "agosto",
                                                               "septiembre", "octubre",
                                                               "noviembre", "diciembre"))

# Graficamos
ggplot(data = df_by_clients, aes(x = month, fill=member_casual)) +
  geom_bar(position = "dodge", stat = "count") +
  facet_wrap(~rideable_type, scales = "free_y") +
  labs(title = "Número de viajes a lo largo del 2022", x="2022", y="# viajes") +
  theme(axis.text.x = element_text(angle = 90))
```

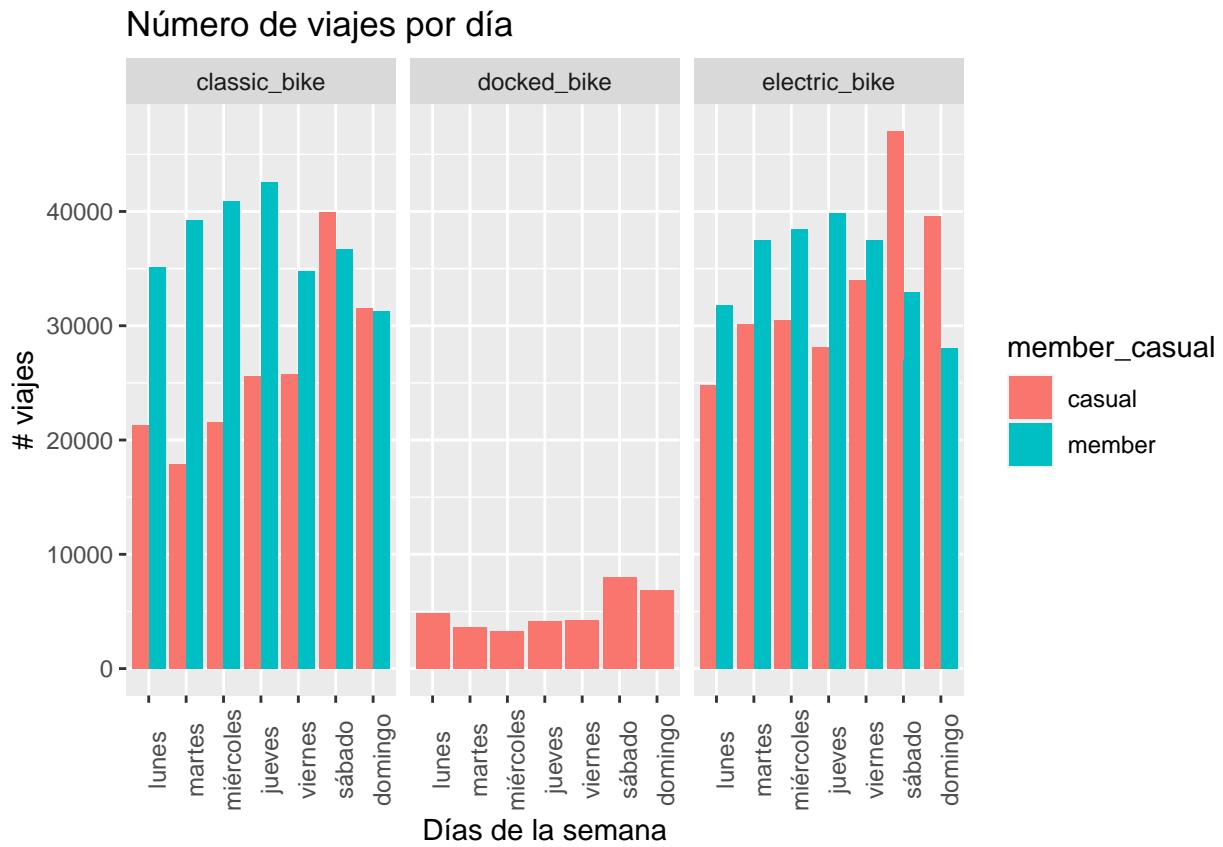


```
# limpieza de datos innecesarios
rm(df_by_clients)
```

Observamos que, por lo general los socios anuales usualmente suelen utilizar mucho mas las bicicletas clásicas que los usuarios casuales. Para el caso de las bicicletas eléctricas notamos un uso muy similar entre ambos tipos de clientes. Sin embargo, la verdadera diferencia radica en las bicicletas acopladas dónde los clientes casuales son indudablemente quienes las utilizan con mayor frecuencia. De igual manera, es sencillo deducir las tendencias para cada tipo de bicicleta basado en el tipo de cliente. Por lo general, los socios anuales tienden a usar mucho más el servicio a partir del mes de mayo, con un declive en el mes de octubre. Mientras tanto , los usuarios casuales tienden a crecer a partir del mes de mayo y caer durante el mes septiembre.

Para indagar aún mas en los datos y las diferencias entre los clientes daremos un paso más para observar la distribución de los viajes basándos únicamente en los días de la semana:

```
# graficamos
ggplot(data = df_by_dates, aes(x = factor(day_of_week, levels= c("lunes", "martes", "miércoles",
                                                               "jueves", "viernes", "sábado",
                                                               "domingo"))),
       y = observations, fill = member_casual)) +
  geom_bar(position = "dodge", stat = "identity") +
  facet_wrap(~rideable_type) +
  theme(axis.text.x = element_text(angle=90))+
  labs(title = "Número de viajes por día", x="Días de la semana", y= "# viajes")
```



Dentro de está gráfica observamos un nuevo patrón. Mientras que nuestros socios anuales utilizan comunmente las bicicletas en días entre semana y reducen su uso los fines de semana, los usuarios casuales representan el comportamiento contrario, ya que sus viajes se concentran en el fin de semana y bajan en días entre semana. Además, reafirmamos que son los clientes casuales quienes son los principales usuarios de las bicicletas acopladas.

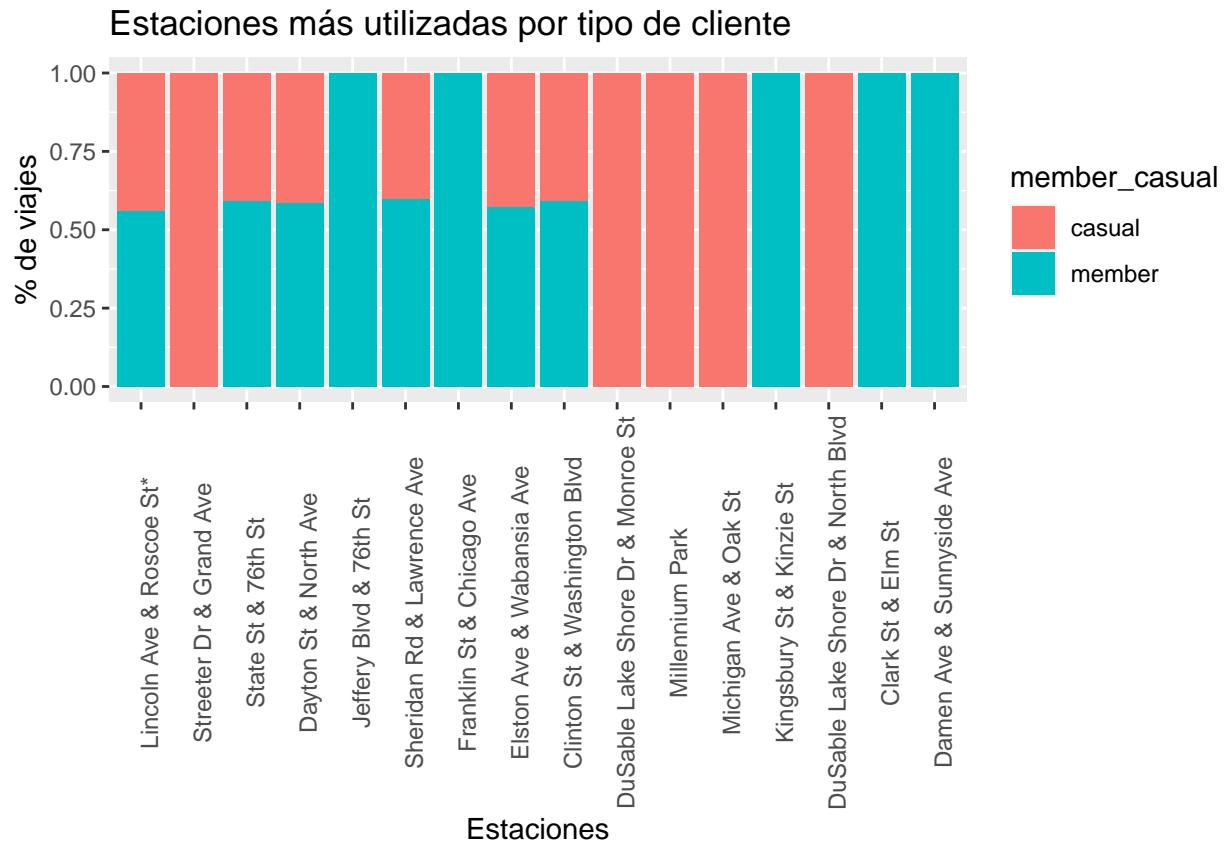
Hasta este punto, podemos observar claramente las diferencias entre los socios anuales y los usuarios casuales

con respecto al servicio de Cyclstc, así que para poder ubicar en qué lugares se concentran los viajes mostraremos las estaciones que concentran la mayor afluencia de clientes y como se distribuyen estos:

```
# Filtramos la información que necesitamos
df_estaciones_group <- df_base[c("start_station_name", "member_casual")] %>%
  group_by(start_station_name, member_casual) %>%
  summarize(observations = n()) %>%
  arrange(desc(observations), member_casual) %>%
  head(22)

## `summarise()` has grouped output by 'start_station_name'. You can override
## using the '.groups' argument.

# Graficamos
ggplot(data = df_estaciones_group, aes(x = factor(start_station_name,
  levels = unique(df_estaciones_group$start_station_name)),
  y=observations, fill = member_casual)) +
  geom_bar(position = "fill", stat = "identity") +
  labs(title = "Estaciones más utilizadas por tipo de cliente", x="Estaciones", y= "% de viajes") +
  theme(axis.text.x = element_text(angle = 90))
```



En el gráfico anterior, podemos observar que la mayoría de las estaciones con mayor afluencia son estaciones compartidas, tanto para miembro anuales como casuales. Se debe tener en cuenta que el conjunto superior de estaciones pertenecen a lugares muy visitados, como zonas turísticas, parques o lugares de entretenimiento y a medida que las estaciones se alejen de estas zonas, el numero de usuarios casuales tiende a reducirse, mientras que el de los miembros anuales aumentará.

## Conclusiones

Después de completar el análisis, se pueden extraer las siguientes conclusiones clave sobre el comportamiento estudiado entre los miembros anuales y casuales.

A continuación se presentan las más destacadas:

- Frecuencia de uso:
  - Los miembros anuales tienden a realizar un mayor uso de nuestros servicios en días entre semana con una duración media de 8 min. Mientras que los miembros casuales realizan la gran parte de sus viajes en fines de semana con una duración media de 12 min.
- Patrones anuales:
  - El uso de bicicletas por cada tipo de cliente a lo largo del año es muy claro, nuestros miembros anuales optan por utilizar la bicicleta clásica con mayor frecuencia siendo los meses entre Junio y Octubre donde incrementan sus viajes. Por su parte los usuarios casuales prefieren las eléctricas y son los principales usuarios de las bicicletas acopladas, mayoritariamente entre los meses de Mayo y Septiembre.
- Distribución de estaciones:
  - Los datos muestran una significativa diferencia entre el uso de estaciones, mientras que en las estaciones más utilizadas comparten ambos tipos de miembros, esto sigue un patrón. Mientras más turística o concurrida sea la zona, mayor será la proporción de clientes casuales que utilicen el servicio. En contraste, a medida que una estación se aleje de estas zonas, aumentará el número de miembros anuales.

## Recomendaciones

- Campaña de encuestas:

Basándonos en los hallazgos del análisis, se recomienda diseñar una campaña de encuestas para obtener las razones por las que nuestros clientes optan por un servicio casual, en las principales zonas de afluencia de este grupo.
- Ofertas o Membresías:

Con el objetivo de impulsar a los clientes casuales a transicionar a nuestros planes anuales, se pueden introducir: descuentos o membresías temporales para atraer a este grupo e incluso algunos especiales por temporadas.
- Implementación de un programa de lealtad:

Como parte de un objetivo de retener a nuestros clientes actuales e incentivar a los clientes casuales por obtener nuestro servicio, un plan de recompensas y beneficios por contar con nuestro servicio puede ser llamativo, como acceder a descuentos graduales o programas de puntos.

## Referencias:

Divvy. (2022). Divvy Trip Data. Recuperado de <https://divvy-tripdata.s3.amazonaws.com/index.html> (consultado el 20 de noviembre de 2023)