Trabajo Práctico 1

Laboratorio de Datos 2C 2023 (comision: Guillermo Solovey)

Integrantes

- Ariel Bakal, LU 1014/22
- Tomás Agustín Rivera Solari, LU 865/22
- Leandro Figueroa Isarrualde LU 213/17

Objetivo

Este trabajo consistira en el analisis descriptivo y exploratorio de dos datasets, y luego un analisis entre ambos, buscando que conclusiones podemos sacar al relacionarlos.

Trabajaremos con los siguientes datasets

Clima
Dataset de datos meteorologicos y climaticos de MeteoStat. Especificamente de Aeroparque.

Viajes EcoBici
Dataset de uso de sistema EcoBici de la Ciudad de Buenos Aires que consta de viajes realizados en bicicleta a lo largo de la ciudad.

Ambos datasets tienen datos unicamente del año 2022.

Dataset Clima

Tenemos los datos en un archivo csv, los importamos y los guardamos como tipo data frame:

```
require(readr)
data_clima <- data.frame(read_csv("data/clima_aeroparque_2022.csv"))
```

Tenemos 365 registros que corresponden a los 365 dias del año.

```
head(data_clima)
```

	date	tavg	tmin	tmax	prcp	snow	wdir	wspd	wpgt	pres	tsun
1	2022-01-01	24.8	20.7	29.1	2.3	NA	344	14.4	NA	1004.1	NA
2	2022-01-02	24.7	19.2	29.7	1.5	NA	16	9.6	NA	1005.1	NA
3	2022-01-03	27.9	24.1	31.7	6.7	NA	81	12.5	NA	1007.1	NA
4	2022-01-04	28.2	24.8	32.6	3.9	NA	236	18.3	NA	1005.9	NA
5	2022-01-05	21.7	16.7	31.7	0.0	NA	118	24.2	NA	1015.5	NA
6	2022-01-06	22.1	18.8	25.7	0.0	NA	78	22.7	NA	1016.7	NA

Este dataset contiene las siguientes variables:

- `date`, fecha del registro (tipo character).
- `tavg`, temperatura promedio en °C (tipo numeric).
- `tmin`, temperatura mínima en °C (tipo numeric).
- `tmax`, temperatura máxima en °C (tipo numeric).
- `prcp`, precipitacion total en mm (tipo numeric).
- `snow`, profundidad de la nieve en mm (tipo numeric).
- `wdir`, direccion del viento en Grados (tipo integer).
- `wspd`, velocidad del viento promedio en km/h (tipo numeric).
- `wpgt`, rafaga de viento maxima en km/h (tipo numeric).
- `pres`, presion del aire al nivel del mar en hPa (tipo numeric).
- `tsun`, duracion total de la luz del sol en min (tipo numeric).

Notamos que, en nuestro caso no tenemos datos de las variables `snow`, `wpgt` y `tsun`. Entonces procedemos a limpiarlos, ya que, no nos brindan información:

```
require(dplyr)
```

```
data_clima <- data_clima %>% select(!c(snow, wpgt, tsun))
head(data_clima)
```

	date	tavg	tmin	tmax	prcp	wdir	wspd	pres
1	2022-01-01	24.8	20.7	29.1	2.3	344	14.4	1004.1
2	2022-01-02	24.7	19.2	29.7	1.5	16	9.6	1005.1
3	2022-01-03	27.9	24.1	31.7	6.7	81	12.5	1007.1
4	2022-01-04	28.2	24.8	32.6	3.9	236	18.3	1005.9
5	2022-01-05	21.7	16.7	31.7	0.0	118	24.2	1015.5
6	2022-01-06	22.1	18.8	25.7	0.0	78	22.7	1016.7

Para empezar, consideramos la temperatura una de las variables más importantes para entender como se comportó el clima a lo largo de 2022.

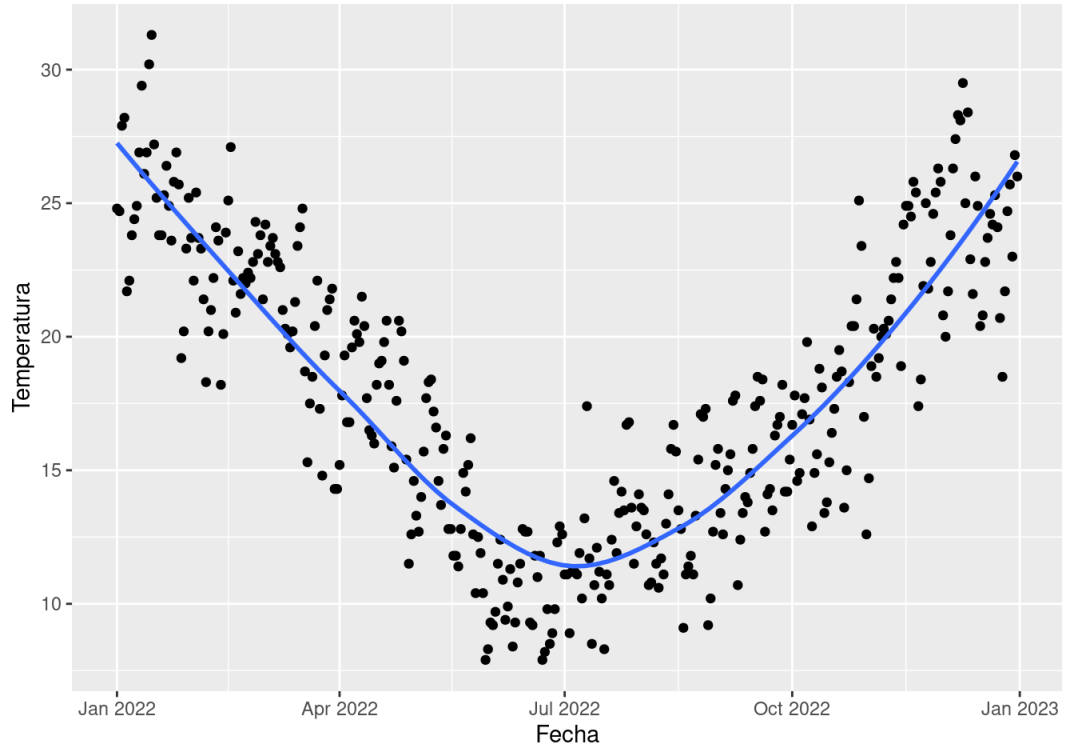
Por esto, veamos como cambió la temperatura a lo largo del año:

```
require(ggplot2)
```

```
Loading required package: ggplot2
```

```
ggplot(data=data_clima,aes(x=date,y=tavg),na.rm=T)+geom_point()+geom_smooth(se=F)+
  labs(x="Fecha",y="Temperatura")
```

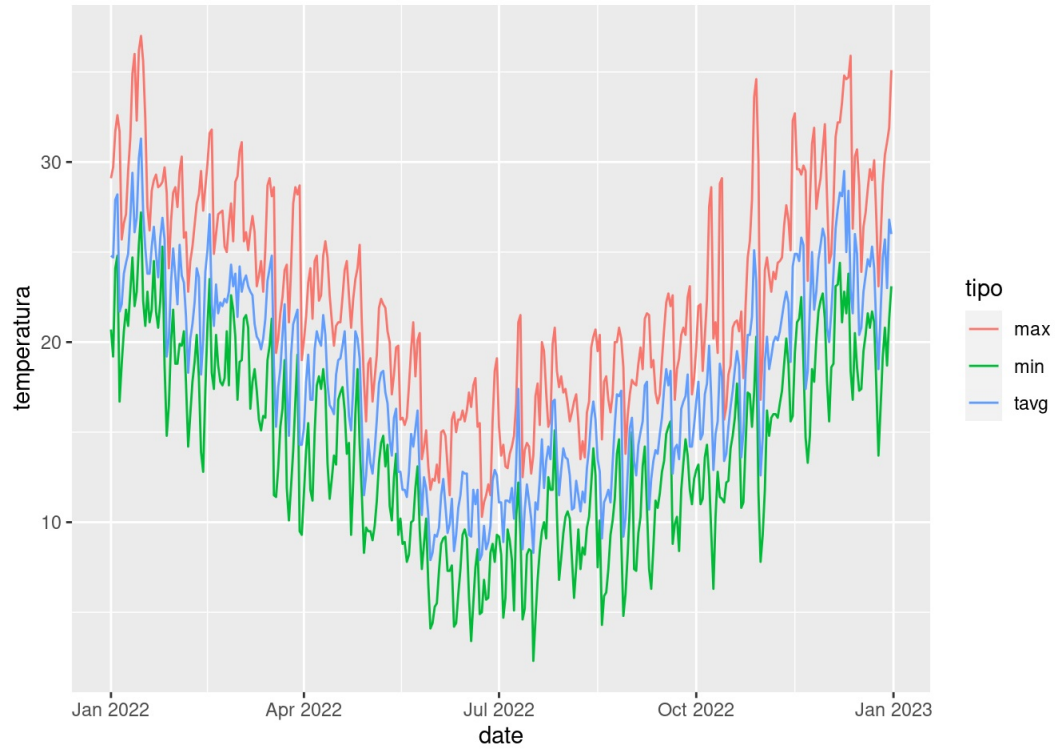
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Como se observa en el gráfico, la temperatura, como era esperable, se comporta de acuerdo a la estación del año. Durante los primeros meses del año, al estar en verano, la temperatura es alta. A medida que pasan los meses, la temperatura decrece, y llega a su punto más bajo en Julio. Durante pleno invierno. A partir de ahí, al temperatura vuelve a crecer, hasta valores similares a los del principio. Debido a que vuelve el verano.

Para tener una noción algo más precisa de como se comporta la temperatura, veamos los valores máximo y minimos, y comparemoslos con el promedio:

```
data_max<-data_clima%>%select(date,tmax)%>%rename(temperatura=tmax)%>%cbind(tipo="max")
data_tavg<-data_clima%>%select(date,tavg)%>%rename(temperatura=tavg)%>%cbind(tipo="tavg")
data_min<-data_clima%>%select(date,tmin)%>%rename(temperatura=tmin)%>%cbind(tipo="min")

data_temperaturas<-rbind(data_max,data_tavg,data_min)


ggplot(data_temperaturas,aes(x=date,y=temperatura,color=tipo))+
  geom_line()
```



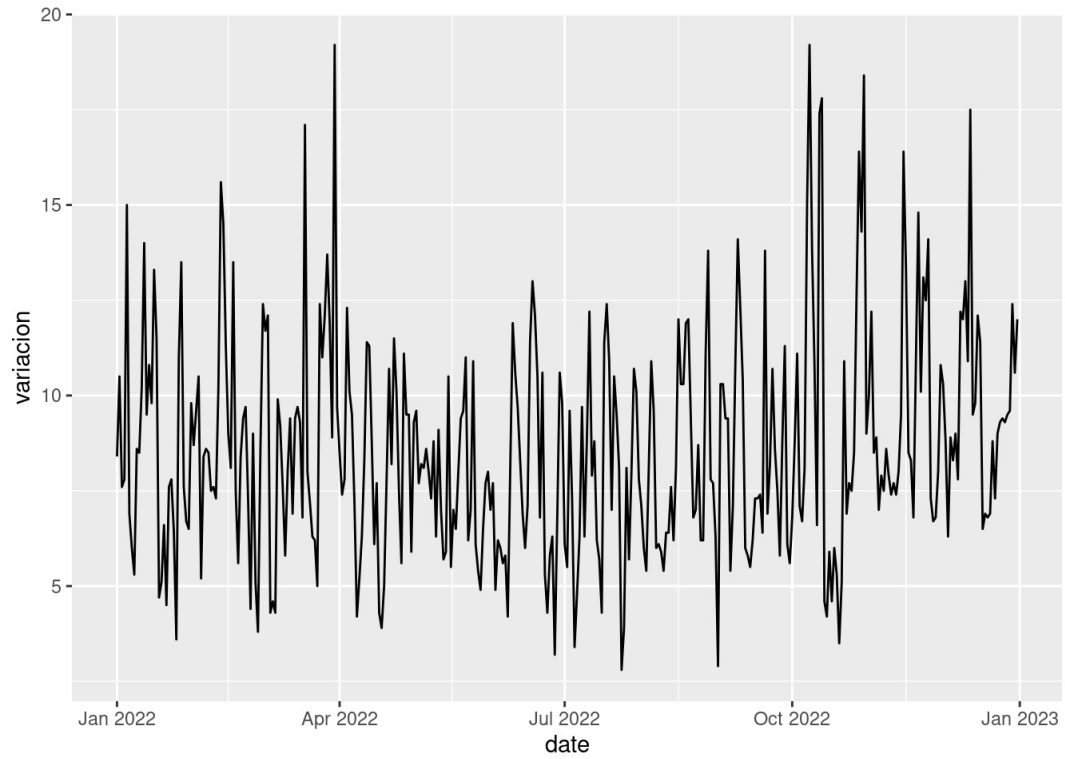
Este gráfico nos permite tener algunas precisiones sobre la temperatura y su comportamientos.

Por un lado, vemos como el máximo valor de la temperatura durante el 2022 fue superior a 35°C, y se dio durante los primeros meses del año. Además, durante el día más frío, la temperatura se encontró entre los 0°C y los 5°C.

También se puede ver que, durante el invierno, la temperatura alcanzaba máximas de hasta 20°C, número relativamente alto, considerando la estación del año. Es interesante que, este valor de 20°C, también se el valor de las temperatura mínimas durante el verano y los meses calurosos.

Ahora, veamos como varía la temperatura por día:

```
cambio_temp<-data_clima%>%mutate(variacion=tmax-tmin)%>%select(date,variacion)

ggplot(cambio_temp,aes(date,variacion))+geom_line()
```

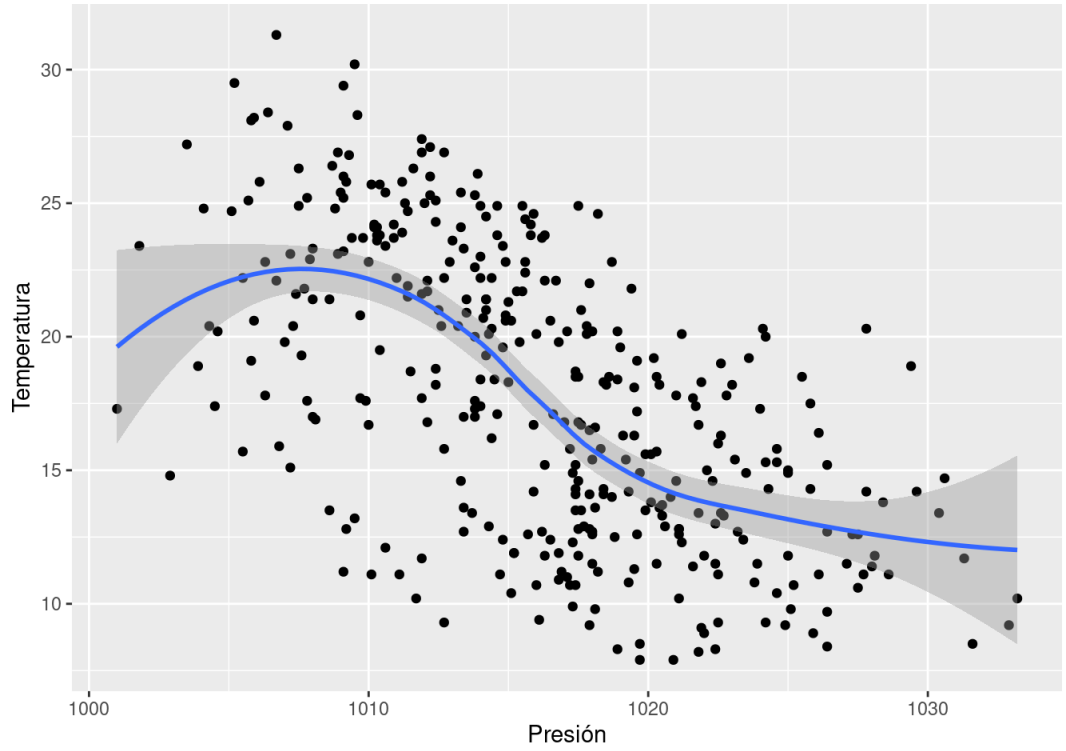


Observando el gráfico, notamos como las mayores diferencias de temperatura se encuetran concentradas en los meses que corresponden al otoño y a la primavera. Mientras que si nos centramos en los meses correspondientes al verano y al invierno, estos saltos de temperatura no son tan pronunciados.

Para terminar con el analisis de la temperatura, veamos como se comporta con respecto a la presión:

```
ggplot(data=data_clima,aes(x=pres,y=tavg),na.rm=T)+geom_point()+geom_smooth()+
  labs(x="Presión",y="Temperatura")
```

`geom_smooth()` using method = 'loess' and formula = 'y ~ x'

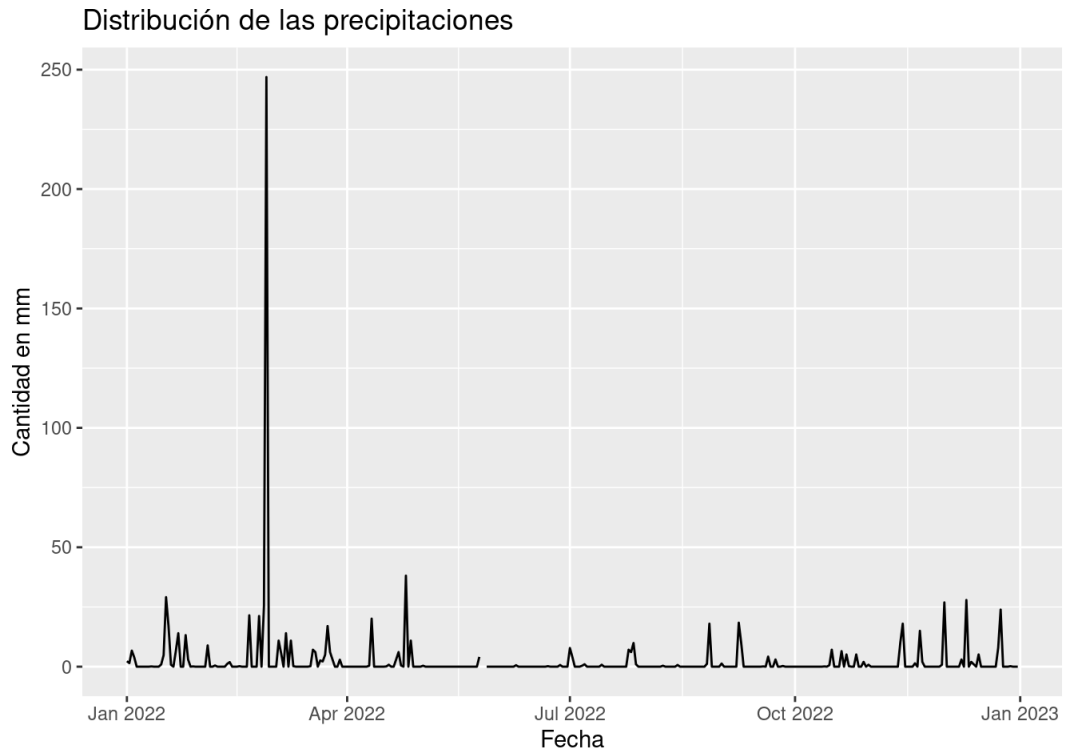


La relación entre la presión y la temperatura es muy clara al observar el comportamiento del gráfico. A medida que aumenta la presión, la temperatura disminuye.

Este es un comportamiento que llama la atención, ya que uno tiende a pensar que, a mayor presión, mayor temperatura. Sin embargo, para la temperatura atmosférica, esto no es así, y se comporta de manera inversa.

Conociendo como se comporta la temperatura, tomemos otra variable para analizar. Veamos como fueron las precipitaciones a lo largo del 2022:

```
ggplot(data_clima, aes(date, prcp)) + geom_line() +
labs(title = "Distribución de las precipitaciones",x="Fecha",y="Cantidad en mm"
)
```



Para empezar, vemos como las precipitaciones se concentran en los primeros meses del año. Durante el invierno son escasas, y ya a final de año, comienzan a incrementar.

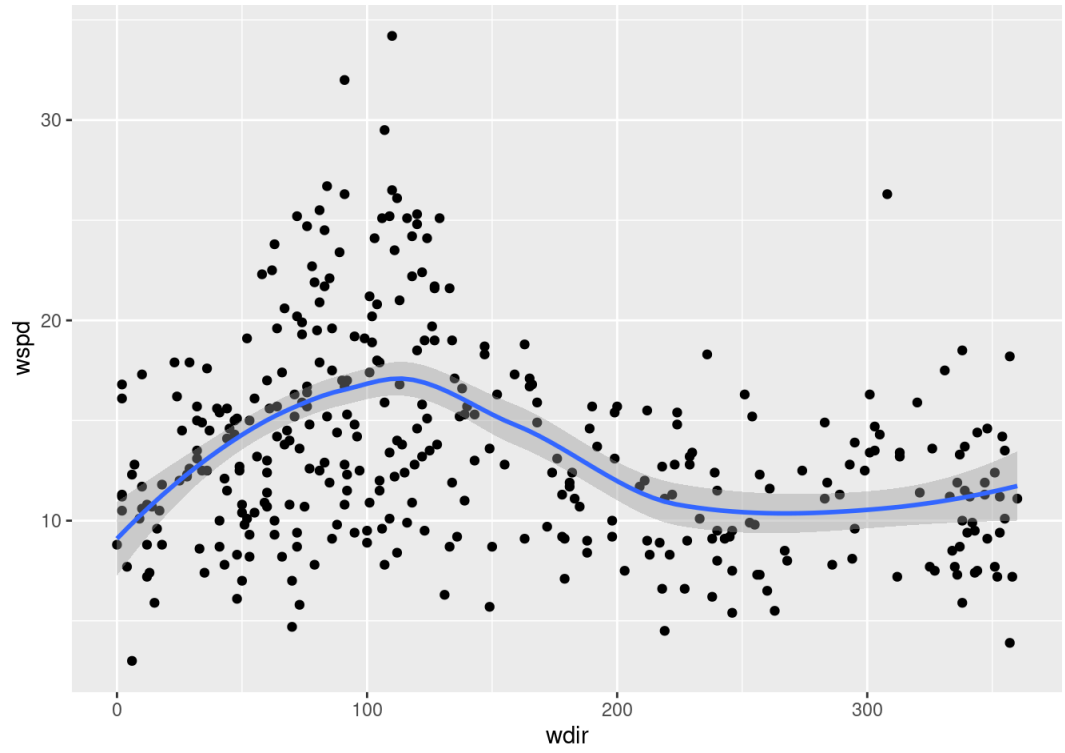
Ahora bien, rápidamente, notamos como existe un valor extraordinario, en el cual las precipitaciones alcanzan los 250mm. Veamos este valor más en detalle.

El valor corresponde al 27 de Febrero, y durante este día se registraron precipitaciones de 246.9 mm. Fue el día más lluvioso del año, con una amplia diferencia.

Ahora, veamos como se comportó el viento durante el año:

```
ggplot(data_clima,aes(wdir,wspd,na.rm=T))+geom_point()+geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Vemos como el viento se mantuvo en velocidades entre los 5km/h y los 25 km/h, con algunos valores excepcionales que no entran en este rango. Además, cuantos el viento viaja con dirección 100°, su velocidad aumenta. En otras direcciones, su velocidad se mantiene de forma constante.

Con estos datos, podemos entender como se comportó el clima en la Ciudad de Buenos Aires, durante el 2022.

Dataset EcoBici

Partimos de un csv con 10000 registros:

```
require(knitr)
require(kableExtra)
mostrar_df <- function(df)
{
  kable(head(df,n=3), format = "html", escape = FALSE) %>%

    kable_styling(full_width = FALSE)
}
```

```
data_bici_2022 <- data.frame(read_csv("data/trips_2022_reducido.csv"))
mostrar_df(data_bici_2022)
```

X.1	X	Id_recorrido	duracion_recorrido	fecha_origen_recorrido	id_estacion_origen	nombre_estacion_origen	direccion_estacion
1733135	1733135	15242490BAEcobici	2262	2022-08-14 18:16:59	95BAEcobici	095 - ESMERALDA	ESMERALDA 516
2722787	2722787	16737944BAEcobici	728	2022-12-25 15:57:31	260BAEcobici	371 - Paseo de las Americas	Av. Pres Figueroa A 6400
1538494	1538494	15315476BAEcobici	414	2022-08-22 14:12:37	132BAEcobici	132 - CORRIENTES	Reconquista & Corr Av.

Tenemos las siguientes variables:

- `id_recorrido`, Id que identifica el viaje (tipo character), .
- `duracion_recorrido`, duración, en segundos, del recorrido (tipo numeric).
- `fecha_origen_recorrido`, fecha y hora del inicio del recorrido (tipo POSIXct).
- `id_estacion_origen`, Id que identifica a la estación de origen (tipo character).
- `nombre_estacion_origen`, nombre de la estación de origen (tipo character).
- `direccion_estacion_origen`, dirección de la estación de origen (tipo character).
- `long_estacion_origen`, la longitud de la estación de origen (tipo numeric).
- `lat_estacion_origen`, la latitud de la estación de origen (tipo numeric).

- `fecha_destino_recorrido`, fecha y hora del final del recorrido (tipo POSIXct).
- `id_estacion_destino`, Id que identifica a la estación de destino (tipo character).
- `nombre_estacion_destino`, nombre de la estación de destino (tipo character).
- `direccion_estacion_destino`, dirección de la estación de destino (tipo character).
- `long_estacion_destino`, la longitud de la estación de destino (tipo numeric).
- `lat_estacion_destino`, la latitud de la estación de destino (tipo numeric).
- `id_usuario`, Id que identifica al usuario (tipo character).
- `modelo_bicicleta`, modelo de la EcoBici utilizada (tipo character).
- `genero`, genero del usuario (tipo character).
- `fecha`, fecha del viaje (tipo Date).

Vamos a trabajar unicamente con viajes de duracion entre 300 y 3600 segundos:

```
data_bici <- data_bici_2022 %>%
  filter(duracion_recorrido >= 300 & duracion_recorrido <= 3600)
```

De esta forma, obtenemos un data frame de 9253 registros, de todos los viajes entre 5 minutos y 1 hora.

Dividimos el analisis de este dataset en distintas categorís, que nos resultaron relevantes.

En primer lugar, observamos como varía el uso de la EcoBici, según al grupo que pertenece el usuario. En este caso, y por los datos otorgados, tomamos como grupo al género del usuario.

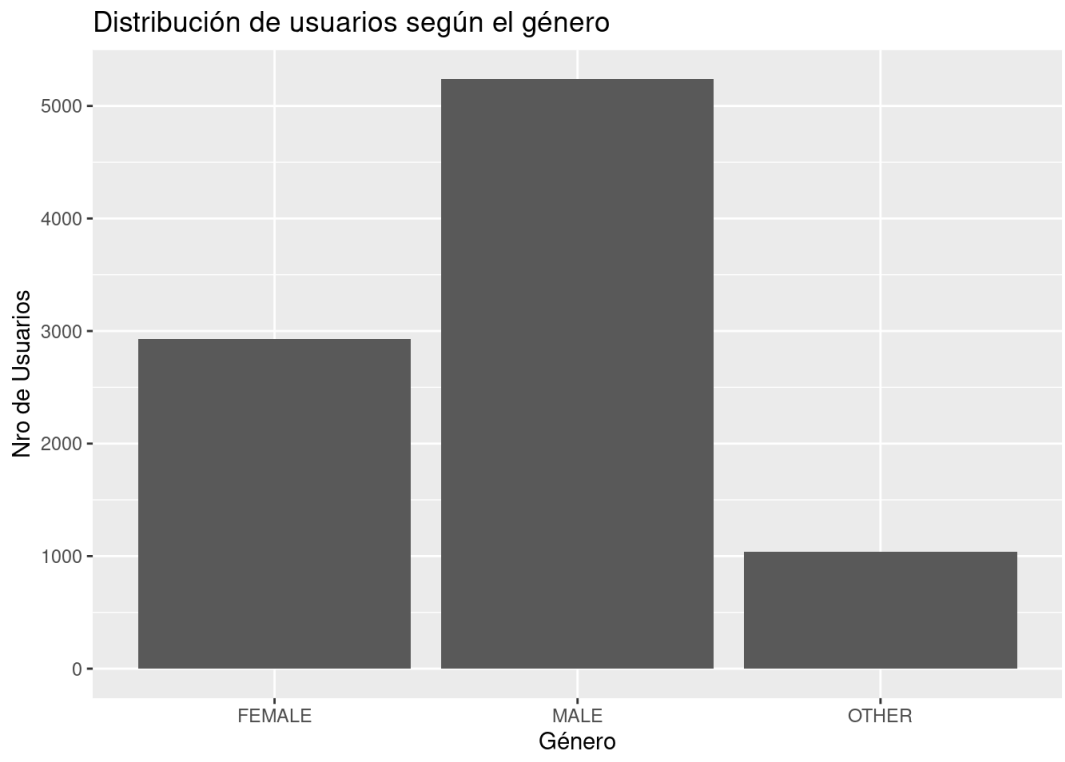
```
require(tidyr)
```

```
Loading required package: tidyr
```

```
genero <- data_bici %>%
  drop_na() %>%
  group_by(Género) %>%
  summarise(cantidad = n())
mostrar_df(genero)
```

Género	cantidad
FEMALE	2928
MALE	5239
OTHER	1038

```
ggplot(genero, aes(Género, cantidad)) +
  geom_bar(stat = "identity") +
  labs(title = "Distribución de usuarios según el género", x = "Género", y = "Nro de Usuarios"
)
```



Como se puede apreciar, existe una diferencia significativa en el uso de la EcoBici, según el género. Un 32% de los usuarios son mujeres, un 57% son hombres y el 11% restante no han especificado su género.

Por otro lado, observamos la frecuencia de los viajes. Es decir, cuales eran los orígenes más comunes, los destinos más comunes, y los recorridos más comunes.

Tomando los origenes más comunes, tenemos que:

```
origenFrecuente <- data_bici %>%
  group_by(nombre_estacion_origen) %>%
  count() %>%
  arrange(desc(n))

head(origenFrecuente)
```

```
# A tibble: 6 × 2
# Groups:   nombre_estacion_origen [6]
  nombre_estacion_origen           n
  <chr>                        <int>
1 147 - Constitución             119
2 014 - Pacifico                 111
3 130 - RETIRO II                110
4 054 - Acuña de Figueroa        101
5 008 - Congreso                  97
6 160 - Godoy Cruz y Libertador   86
```

Obsevamos que las estaciones Constitución, Pacífico y Retiro son las más comunes. Lo cual tiene sentido, teniendo en cuenta que son tres puntos de mucho tránsito en la Ciudad de Buenos Aires.

Tomando los destinos más frecuentes, tenemos:

```
destinoFrecuente <- data_bici %>%
  group_by(nombre_estacion_destino) %>%
  count() %>%
  arrange(desc(n))

head(destinoFrecuente)
```

```
# A tibble: 6 × 2
# Groups:   nombre_estacion_destino [6]
  nombre_estacion_destino          n
  <chr>                        <int>
1 147 - Constitución             125
2 014 - Pacifico                 106
3 008 - Congreso                 100
4 160 - Godoy Cruz y Libertador   95
5 005 - Plaza Italia              85
6 131- HOSPITAL DE CLÍNICAS       85
```

En este caso, Constitución es la estación de destino más concurrida, con una amplia diferencia.

Tomemos los recorridos más frecuentes:

```
frecuentes <- data_bici %>%
  group_by(nombre_estacion_origen, nombre_estacion_destino) %>%
  summarize(count = n(),
            duracion_promedio = mean(duracion_recorrido, na.rm = TRUE)) %>%

  arrange(desc(count))
```

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

```
head(frecuentes)
```

```
# A tibble: 6 × 4
# Groups:   nombre_estacion_origen [6]
  nombre_estacion_origen          nombre_estacion_dest…¹ count duracion_promedio
  <chr>                           <chr>                  <int>            <dbl>
1 391 - Plaza República de Ecuad… 014 - Pacifico            12             760.
2 014 - Pacifico                  391 - Plaza República…    11             625.
3 130 - RETIRO II                 393 - Barrio 31           11             495.
4 152 - JULIETA LANTERI           152 - JULIETA LANTERI     11            1846.
5 350 - Plaza Irlanda             350 - Plaza Irlanda       10            1331.
6 103 - MALBA                     001 - FACULTAD DE DER…     9             689.
# ℹ abbreviated name: ¹nombre_estacion_destino
```

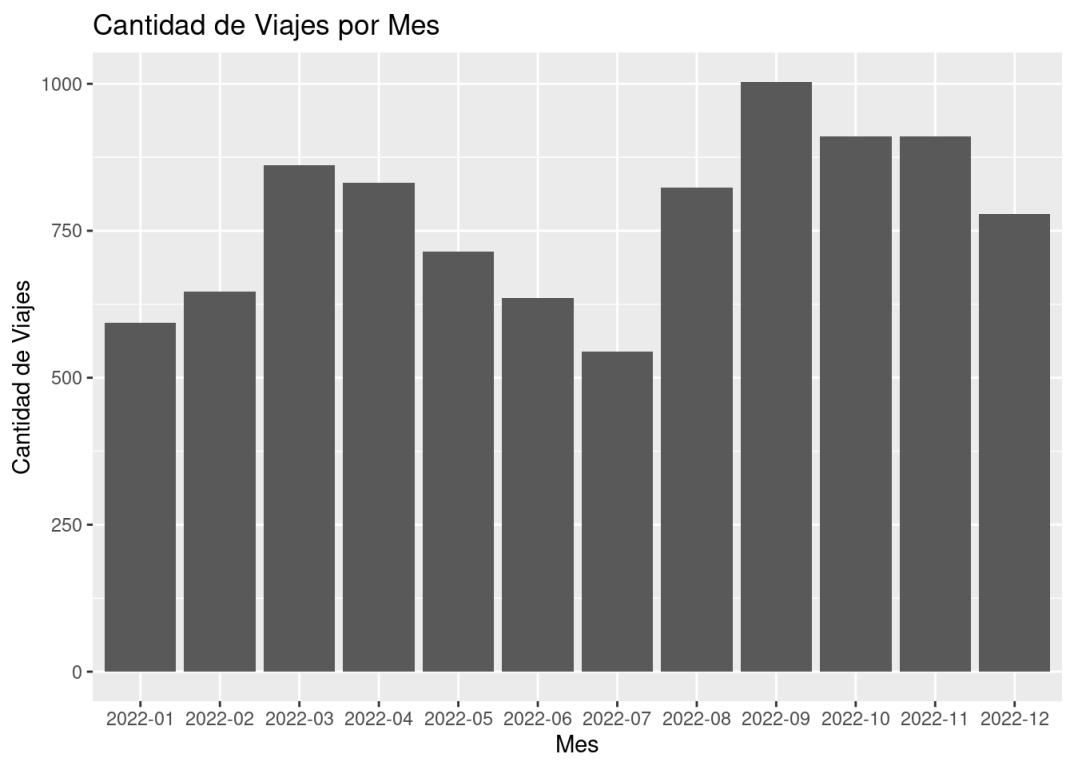
Al ver los recorridos más frecuentes, notamos que algunos de los destinos y orígenes más comunes, en particular, Pacífico y Retiro, también forman parte de los recorridos más comunes.

Por otra parte, un resultado curioso obtenido, es que hay dos recorridos entre los más frecuentes, con mismo origen y destino. Los viajes “Julieta Lanteri-Julieta Lanteri” y “Plaza Irlanda-Plaza Irlanda”, teniendo en cuenta la ubicación de las estaciones son viajes recreativos. Es decir, que no tienen el fin de trasladarse por la Ciudad, sino de dar un recorrido en bicicleta.

Siguiendo con el análisis, obsevamos la frecuencia de los viajes, repartidos en diferentes períodos de tiempo.

En primer lugar, observamos la distribución de los viajes a lo largo del año. Es decir, como se repartieron por mes:

```
viajes_por_mes <- data_bici %>%
  mutate(Mes = format(fecha, "%Y-%m")) %>%
  group_by(Mes) %>%
  count()

ggplot(viajes_por_mes, aes(Mes, n)) +

  geom_bar(stat = "identity") +
  labs(title = "Cantidad de Viajes por Mes", x="Mes", y="Cantidad de Viajes"
)
```

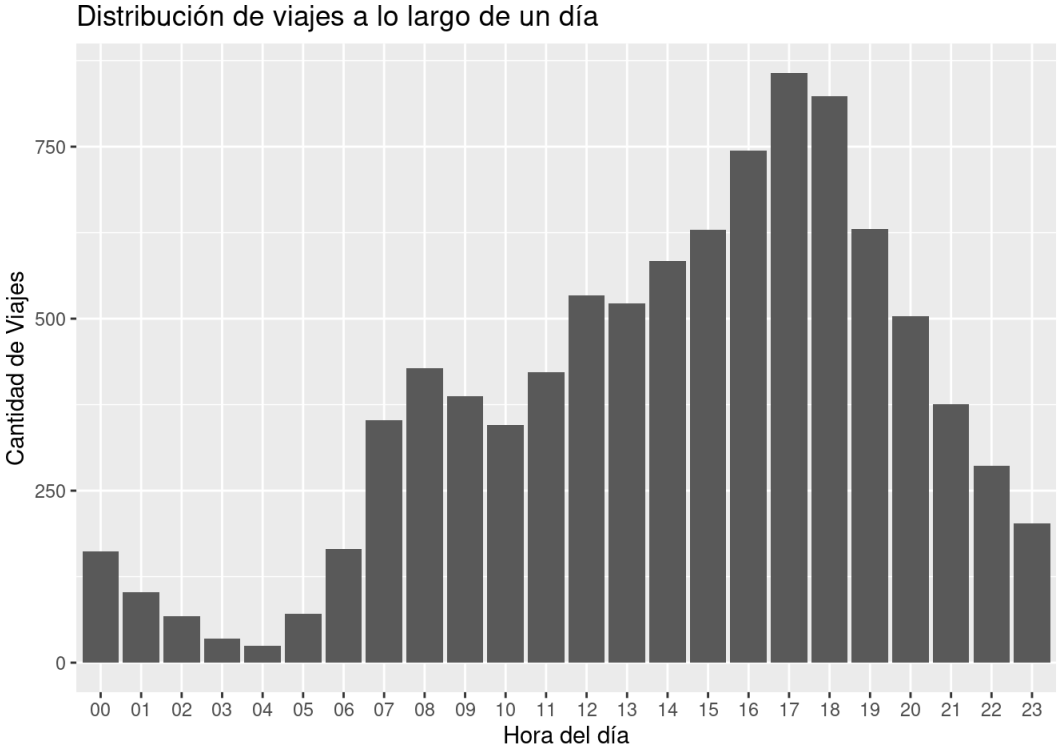


Observando el gráfico obtenido, notamos como los puntos más bajos de uso de la EcoBici fueron Julio y Enero. Mientras que el punto más alto fue Septiembre. Además observamos tendencias en cuanto a como crece el gráfico. De Enero a Marzo, y de Julio a Septiembre, el gráfico crece. Mientras que de Marzo a Julio y de Septiembre a Diciembre, el gráfico decrece.

Observando como varía con los meses, podemos intuir que parte de este comportamiento, tiene que ver con el clima. Podremos confirmar esto al analizar el dataset correspondiente a este apartado.

Además, siguiendo con la forma en la que se distribuyen los viajes, nos centramos en como lo hacen a lo largo de un día:

```
viajes_por_hora <- data_bici %>%
  mutate(Hora = format(fecha_origen_recorrido, "%H")) %>%
  group_by(Hora)

ggplot(viajes_por_hora,aes(Hora)) +
  geom_bar() +
  labs(title = "Distribución de viajes a lo largo de un día", x="Hora del día", y="Cantidad de Viajes"
)
```



Podemos ver como, la hora en la que comienza a utilizarse de forma significativa la EcoBici es las 7 de la mañana. Un resultado coherente, al ser un horario que se aproxima a los habituales horarios de entrada al trabajo o a los distintos centros educativos.

Desde las 7 hasta las 11, se mantiene de forma relativamente constante. No es hasta las 12 que su uso comienza a aumentar de forma significativa, llegando a su pico a las 17. Este comportamiento se explica debido a que esta hora es un horario habitual de fin de la jornada laboral. Esto implica una gran congestión en el resto de medios de transporte, por lo que es posible que mucha gente opte por utilizar la EcoBici. A esto se le pueden sumar los viajes recreativos, que deben ser más frecuentes a estas horas, al ser, por lo general, horarios no laborables.

Ahora bien, supusimos que el uso de la EcoBici puede cambiar, dependiendo de si el día tomado es un día de semana, o un fin de semana. Por lo tanto, analizamos ambos casos por separado.

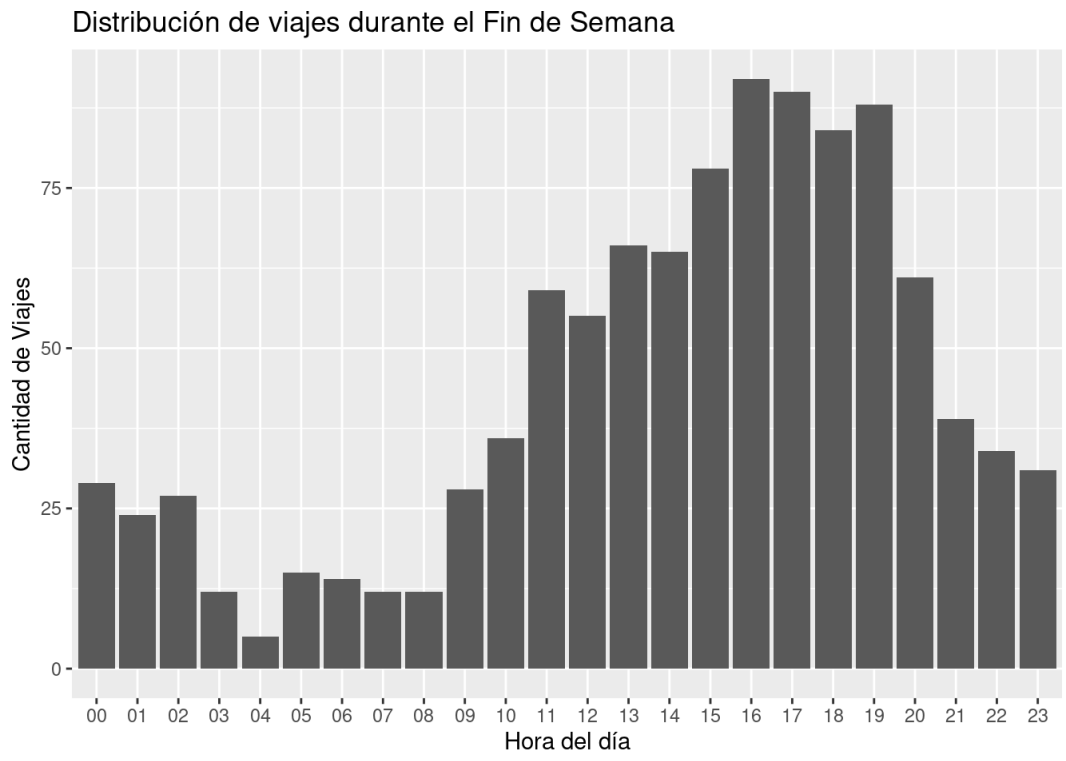
Primero, creamos un dataset, que separara los días, dependiendo de si eran Fines de Semana, o Días de Semana:

```
finesDeSemana <- data_bici %>%
  mutate(Dia = ifelse(weekdays(fecha) %in% c("Saturday","Sunday"), "Fin de Semana", "Día de Semana")
)

mostrar_df(finesDeSemana)
```

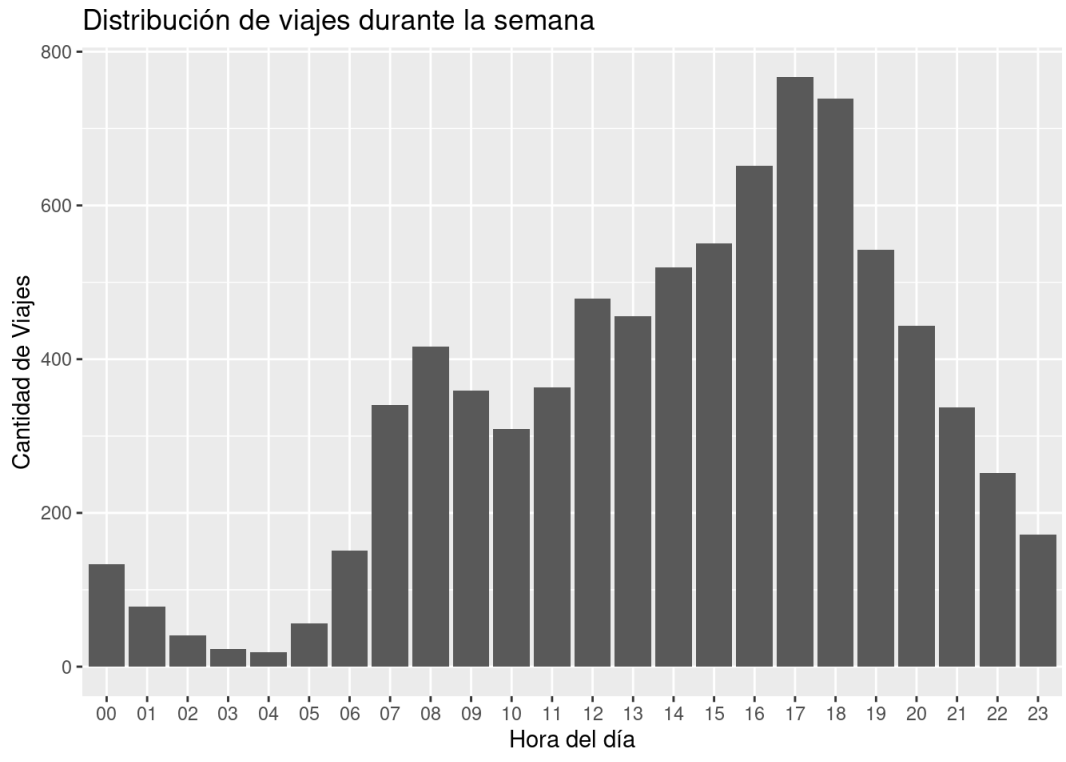
X.1	X	Id_recorrido	duracion_recorrido	fecha_origen_recorrido	id_estacion_origen	nombre_estacion_origen	direccion_estacion
1733135	1733135	15242490BAEcobici	2262	2022-08-14 18:16:59	95BAEcobici	095 - ESMERALDA	ESMERALDA 516
2722787	2722787	16737944BAEcobici	728	2022-12-25 15:57:31	260BAEcobici	371 - Paseo de las Americas	Av. Pres Figueroa A 6400
1538494	1538494	15315476BAEcobici	414	2022-08-22 14:12:37	132BAEcobici	132 - CORRIENTES	Reconquista & Corr Av.

A partir de estos, observamos el comportamiento del uso de la EcoBici:

```
FDS_por_hora <- finesDeSemana %>%
  mutate(Hora = format(fecha_origen_recorrido, "%H")) %>%
  group_by(Hora) %>%
  filter(Dia != "Día de Semana")

ggplot(FDS_por_hora,aes(Hora)) +
  geom_bar() +
  labs(title = "Distribución de viajes durante el Fin de Semana", x="Hora del día", y="Cantidad de Viajes"
)
```




```
WD_por_hora <- finesDeSemana %>%
  mutate(Hora = format(fecha_origen_recorrido, "%H")) %>%
  group_by(Hora) %>%
  filter(Dia != "Fin de Semana")

ggplot(WD_por_hora,aes(Hora)) +
  geom_bar() +
  labs(title = "Distribución de viajes durante la semana", x="Hora del día", y="Cantidad de Viajes"
)
```

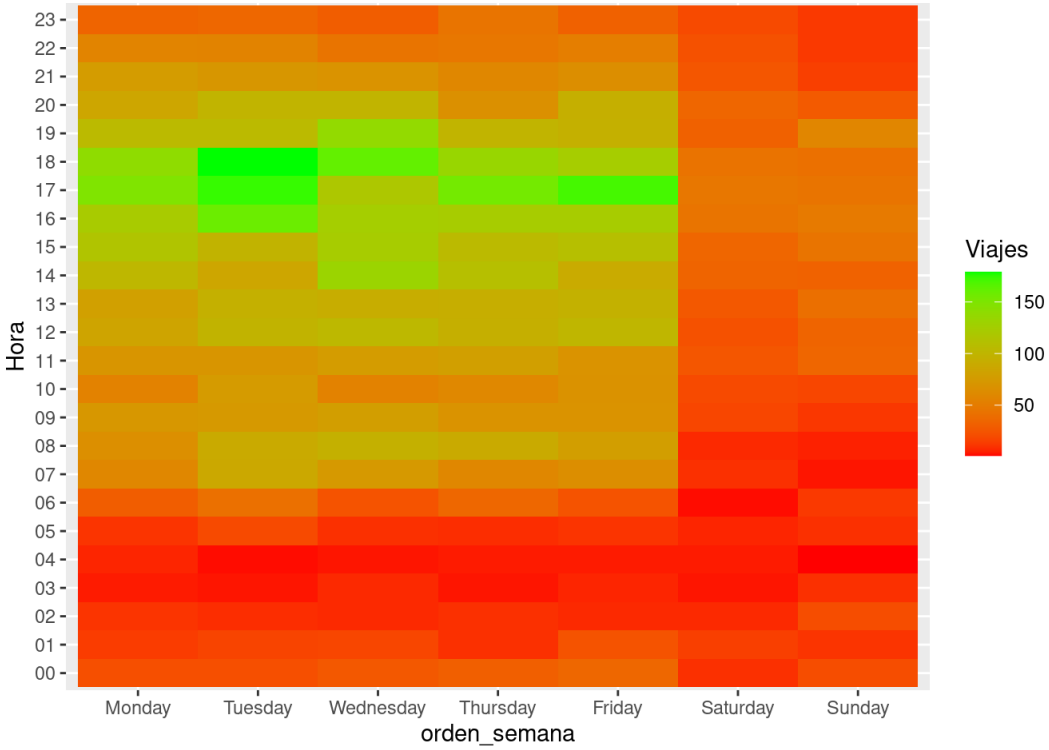


Como esperabamos, el ambos gráfico presenta notables diferencias. Por un lado, el gráfico que muestra la distribución durante los días de semana, tiene un comportamiento similar al gráfico realizado para un día, sin importar que sea o no de semana. Los picos se encuentran en los horarios de entrada y sálida a los trabajos y centros educativos. Además en uso de EcoBici decrece de forma constante desde las 18 hasta las 04.

No obstante, al ver la distribución de viajes durante los Fines de Semana, notamos un comportamiento diferente. Para empezar, el primer horario en el que hay un gran uso de la EcoBici son las 11. A partir de este momento, crece hasta las 16, y se mantiene en un número similar, hasta las 20, donde empieza a bajar significativamente. Algo curioso que podemos observar, es como desde las 21 hasta las 02, el uso de la Ecobici se mantiene constante. Esto no sucede durante los días de semana. Además, el uso durante esta franja horaria es igual al uso durante las 09 y las 10.

Todo este comportamiento, se encuentra resumido en el siguiente mapa de calor:

```
diasSemana <- data_bici %>%
  mutate(Dias_Semana = weekdays(fecha),
         Hora = format(fecha_origen_recorrido, "%H")) %>%
  group_by(Dias_Semana, Hora) %>%
  mutate(Viajes = n()) %>%
  drop_na()

orden_semana = factor(diasSemana$Dias_Semana, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")
)

ggplot(diasSemana, aes(orden_semana, Hora, fill = Viajes)) +

  geom_tile() +
  scale_fill_gradient(low = "red", high = "green"
)
```

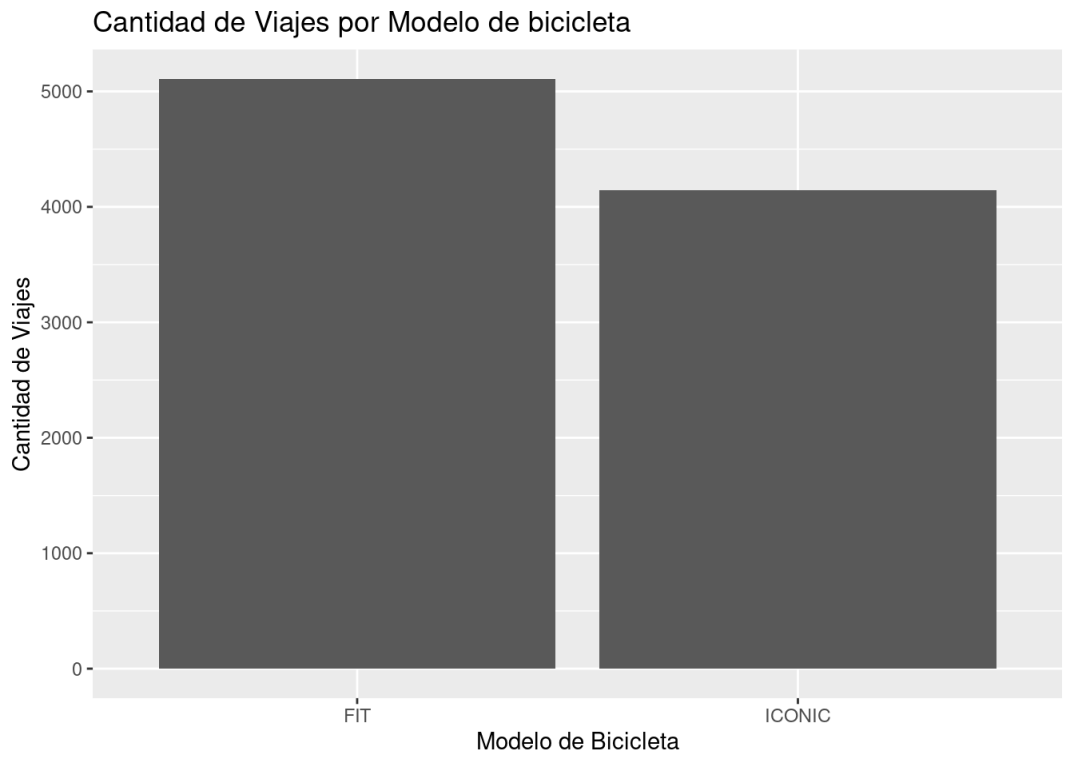


Finalmente, observamos cuanto se usa cada modelo de Ecobici. Durante el 2022, se usaron dos modelos: el FIT y el ICONIC:

```
modelo_màs_usado <- data_bici %>%
  group_by(modelo_bicicleta) %>%
  count()

head(modelo_màs_usado)
```

```
# A tibble: 2 × 2
# Groups:   modelo_bicicleta [2]
  modelo_bicicleta     n
  <chr>            <int>
1 FIT               5107
2 ICONIC            4146
```

```
ggplot(modelo_màs_usado, aes(modelo_bicicleta, n)) +
  geom_bar(stat = "identity") +
  labs(title = "Cantidad de Viajes por Modelo de bicicleta", x="Modelo de Bicicleta", y="Cantidad de Viajes"
)
```



Se observa como el FIT es el modelo más usado por los usuarios de EcoBici. Sin embargo, al no ser una diferencia significativa, este resultado puede deberse a la cantidad de modelos de bicicleta disponibles, y no a una preferencia particular de los usuarios.

Análisis Exploratorio

Ya habiéndonos familiarizado con ambos dataframes por separado, estamos en condiciones de relacionar el uso de la EcoBici con las condiciones climáticas partículares.

Lo primero que hicimos, fue juntar ambos dataframes, con la columna de Fechas como ancla:

```
data_clima$date <- as.Date(data_clima$date)

union_bici_clima<-data_bici%>%rename(date=fecha)%>%inner_join(data_clima,by="date")

mostrar_df(union_bici_clima)
```

X.1	X	Id_recorrido	duracion_recorrido	fecha_origen_recorrido	id_estacion_origen	nombre_estacion_origen	direccion_estacion
1733135	1733135	15242490BAEcobici	2262	2022-08-14 18:16:59	95BAEcobici	095 - ESMERALDA	ESMERALDA 516
2722787	2722787	16737944BAEcobici	728	2022-12-25 15:57:31	260BAEcobici	371 - Paseo de las Americas	Av. Pres Figueroa A 6400
1538494	1538494	15315476BAEcobici	414	2022-08-22 14:12:37	132BAEcobici	132 - CORRIENTES	Reconquista & Corr Av.

Este nuevo dataframe tiene las condiciones climáticas de todos los viajes que se registraron en el 2022, además de las características y variables de los mismos viajes.

Habiéndo analizado ambos dataframes por separado, suponemos que las variables climáticas que más podrían afectar al uso de la Ecobici son:

- La temperatura
- Las precipitaciones
- El viento

Por lo tanto, analizamos el uso de la Ecobici, teniendo en cuenta estas tres condiciones climáticas.

Temperatura

Una de las principales variantes del clima es la temperatura. Analizaremos la duracion y la frecuencia de los viajes en relacion al cambio de temperatura.

Partimos de tomar la cantidad de viajes en dos casos especiales, la temperatura mas baja y mas alta registradas.

```
print(data_clima[which.min(data_clima$tmin), ]$date
)
```

```
[1] "2022-07-17"
```

```
min(data_clima$tmin)
```

```
[1] 2.3
```

```
print(data_clima[which.max(data_clima$tmax), ]$date)
```

```
[1] "2022-01-15"
```

```
max(data_clima$tmax)
```

```
[1] 37
```

Dado a que son dos temperaturas extremas para nosotros, podriamos predecir que la cantidad de viajes es similar y baja.

```
nrow(subset(data_bici_2022, fecha == '2022-07-17'))
```

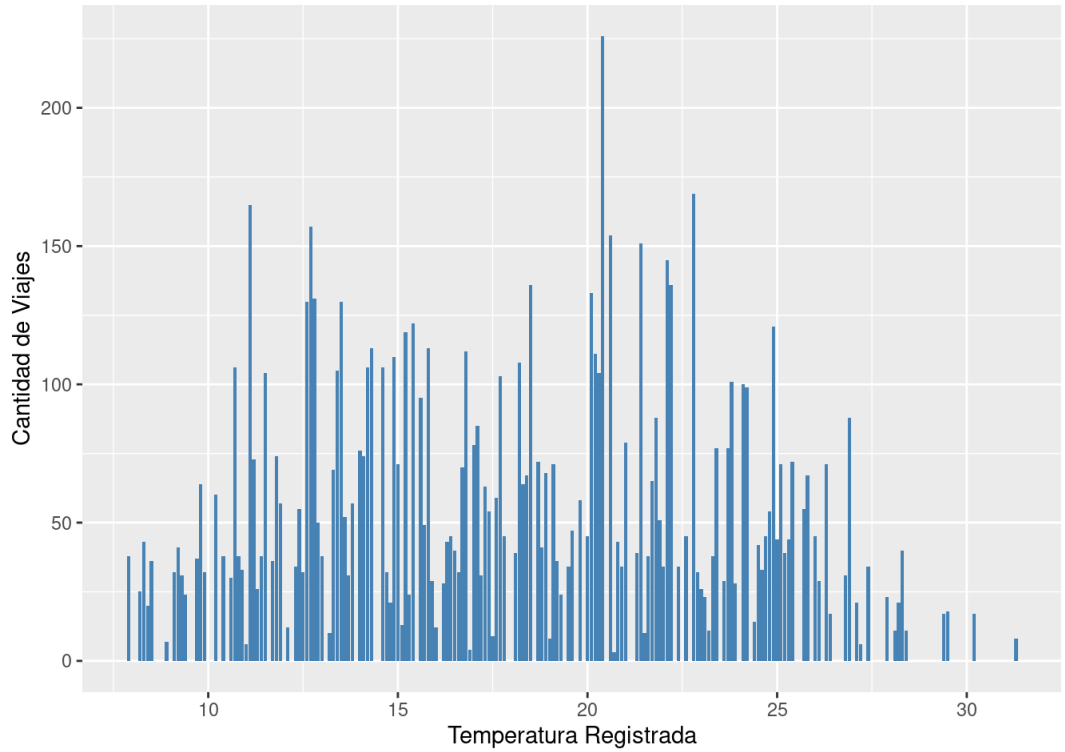
```
[1] 7
```

```
nrow(subset(data_bici_2022, fecha == '2022-01-15'))
```

```
[1] 8
```

Veamos la cantidad de viajes realizados en relacion de la temperatura registrada

```
require(ggplot2)
require(dplyr)

data_clima_2022 <- data_clima %>% rename(fecha = date)
datos_combinados <- merge(data_bici, data_clima_2022, by = "fecha")

resultados <- datos_combinados %>%
  group_by(tavg) %>%
  summarise(cantidad = n())

ggplot(resultados, aes(x = tavg, y = cantidad))
+
  geom_bar(stat = "identity", fill = "steelblue") +

  labs(x = "Temperatura Registrada", y = "Cantidad de Viajes")
```

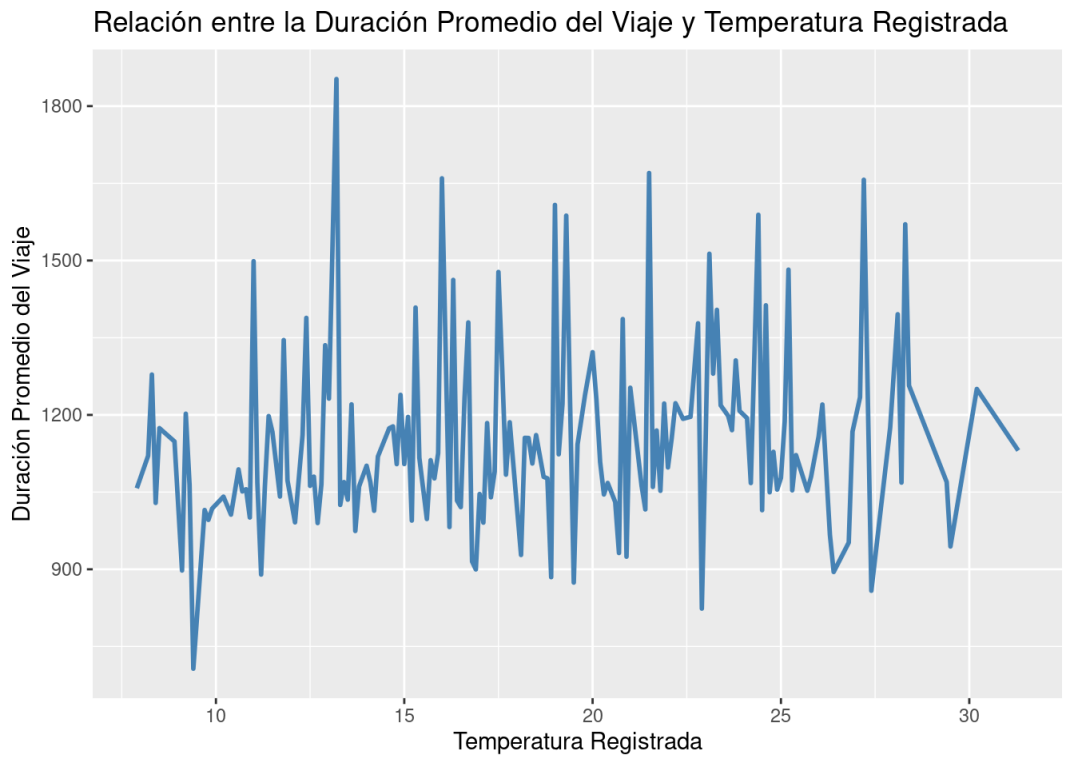


Efectivamente podemos ver como estas temperaturas extremas reducen el uso de bicicleta. Además vemos cual es la temperatura con la cual tenemos mas viajes, la cual es de 20°C.

Analogamente podriamos decir lo mismo de la duracion de los viajes.

```
duracion_promedio <- datos_combinados %>%
  group_by(tavg) %>%
  summarise(duracion_promedio = mean(duracion_recorrido, na.rm = TRUE)
)

ggplot(duracion_promedio, aes(x = tavg, y = duracion_promedio))
+
  geom_line(color = "steelblue", size = 1) +  # Usar geom_line() para un gráfico de líneas
  labs(x = "Temperatura Registrada", y = "Duración Promedio del Viaje")
+
  ggtitle("Relación entre la Duración Promedio del Viaje y Temperatura Registrada")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



Observamos una leve pero significativa diferencia con duraciones que no superan los 1500 segundos de duracion en los extremos de temperatura. Lo cual afirma nuestra analogia. Notar que tenemos un pico de 1852 segundos a los 13.2 °C, lo cual no es un clima ideal como lo es 20 °C, donde tenemos mas viajes. Esto es asi porque se trata del dia 9 de Julio, Dia de la Independencia. Este feriado es la justificacion mas razonable de este pico.

Podemos concluir que la temperatura es un gran factor en el uso de las bicicletas y en su duracion. Este analisis podría llevar a la toma de decision de ofrecer menor cantidad de bicicletas en estas temperaturas extremas, por ende menor mantenimiento y menores gastos.

Precipitaciones

Para analizar como afectan las precipitaciones al uso de la Ecobici, tomamos como referencia el género de las personas. Es decir que buscamos si las precipitaciones afectaban al uso de la Ecobici de forma diferente a los distintos géneros. De esta forma, no solo obtendríamos en comportamiento de esta relación en particular (para cada género), sino también en general (sin importar el género).

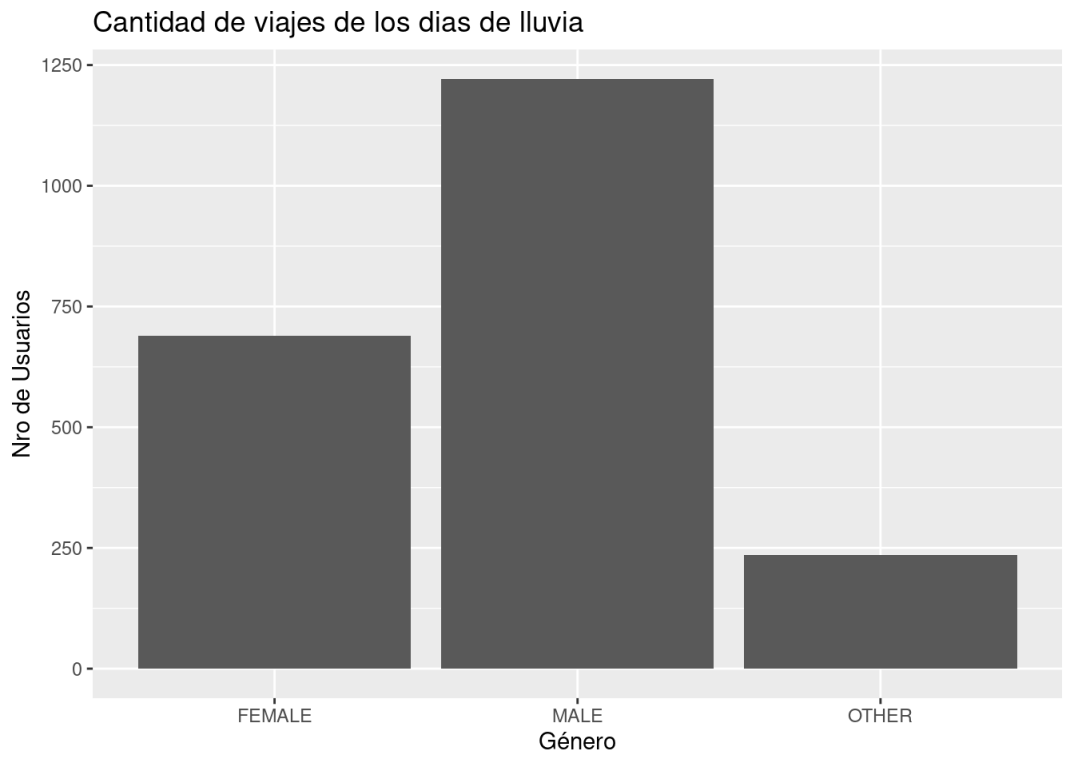
Para empezar, creamos un dataset, que indicara si ese día, había habido precipitaciones:

```
data_inner<-union_bici_clima%>%select(date,duracion_recorrido,Género,prcp)
data_inner<-data_inner%>%mutate(llueve=ifelse(prcp>0,"sí","no"))

mostrar_df(data_inner)
```

date	duracion_recorrido	Género	prcp	llueve
2022-08-14	2262	MALE	0.7	sí
2022-12-25	728	MALE	0.0	no
2022-08-22	414	MALE	0.0	no

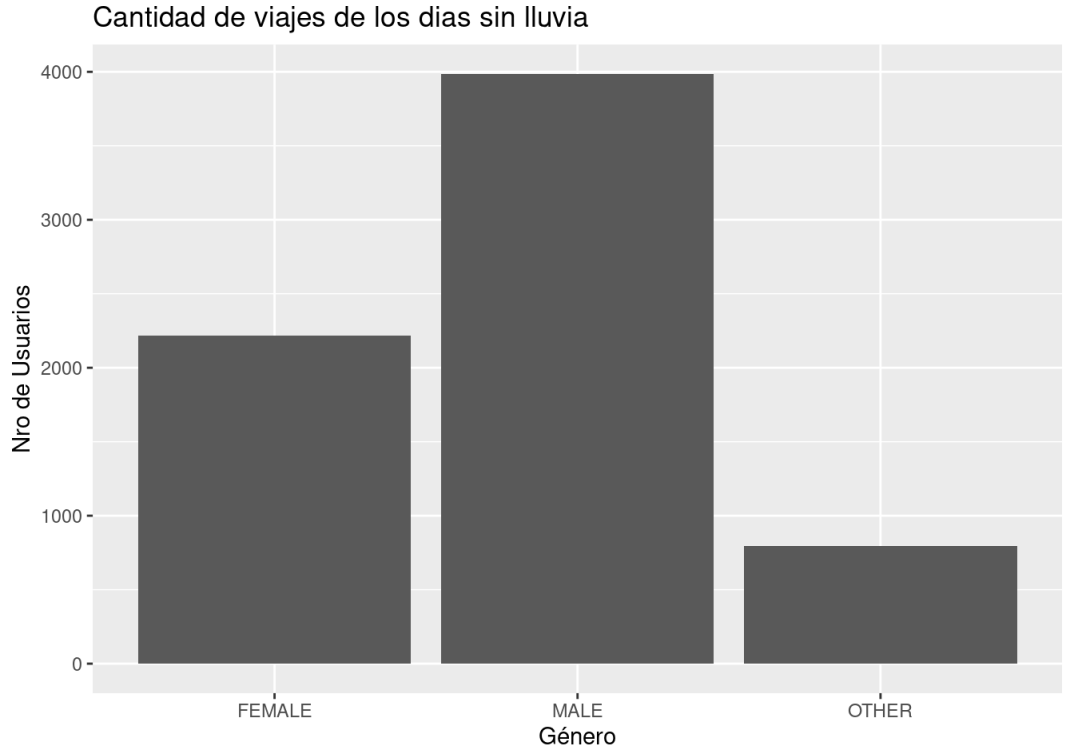
Luego, observamos la cantidad de usuarios, por género, que utilizaron la Ecobici durante los días lluviosos:

```
inner_dias.con.lluvia<-data_inner%>%filter(llueve=="sí")%>%drop_na()

ggplot(inner_dias.con.lluvia,aes(x=Género))+geom_bar()+
  labs(title="Cantidad de viajes de los dias de lluvia",y="Nro de Usuarios",x="Género"
)
```



Y la cantidad de usuarios, por género, que utilizaron la Ecobici durante lo días en lo que no llovió:

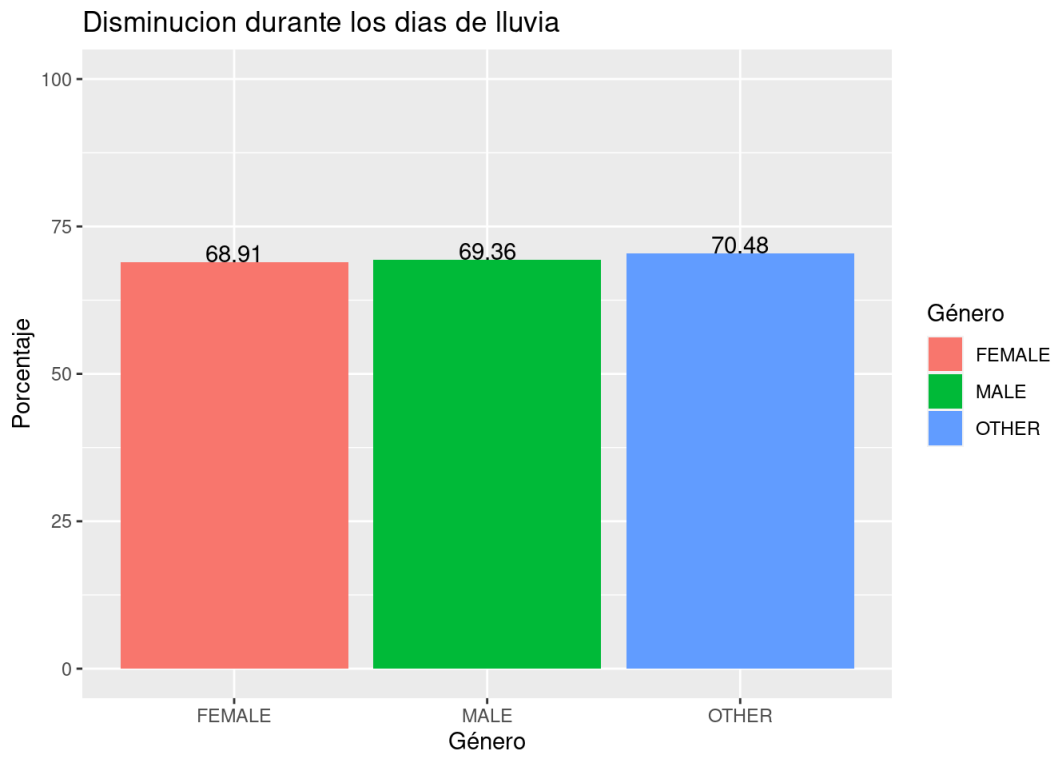
```
inner_dias.sin.lluvia<-data_inner%>%filter(llueve=="no")%>%drop_na(Género)

ggplot(inner_dias.sin.lluvia,aes(x=Género))+geom_bar()+
  labs(title="Cantidad de viajes de los dias sin lluvia",y="Nro de Usuarios",x="Género"
)
```



Comparando ambos gráficos, rapidamente notamos como existe una gran diferencia en la cantidad de viajes que se realizaron durante los días con lluvia y los días sin lluvia. Sin embargo, no parece haber una gran diferencia en la distribución según el género.

Veamos más detalladamente cuanto disminuye el uso de la Ecobici durante los días lluviosos:

```
cuenta1<-inner_dias.sin.lluvia%>%group_by(Género)%>%count(Género)%>%rename(cantidad1=n)
cuenta2<-inner_dias.con.lluvia%>%group_by(Género)%>%count(Género)%>%rename(cantidad2=n)


porcentaje<-cuenta1%>%inner_join(cuenta2,by="Género")
porcentaje<-porcentaje%>%mutate(porcentaje=100-(cantidad2/cantidad1)*100)


ggplot(porcentaje,aes(x=Género,y=porcentaje,fill=Género))+geom_bar(stat = "identity")+
  labs(y="Porcentaje",x="Género") + ylim(0,100)+
  geom_text(aes(label=round(porcentaje,2)), position=position_dodge(.9), vjust=0)+
  labs(title="Disminucion durante los dias de lluvia")
```



Como se puede ver, durante los días lluviosos, el uso de la Ecobici se reduce un 70%, aproximadamente, indistintamente del género del usuario.

Este número, al ser un porcentaje considerablemente grande, nos indica que las precipitaciones afectan enormemente, y de manera negativa, el uso de la Ecobici.

Ahora, veamos si la lluvia afecta la duración de los viajes:

```
duracion_con_lluvia<-inner_dias.con.lluvia%>%group_by(Género)%>%
  summarise(promedio_c.lluvia=mean(duracion_recorrido))

duracion_sin_lluvia<-inner_dias.sin.lluvia%>%group_by(Género)%>%
  summarise(promedio_s.lluvia=mean(duracion_recorrido))


porcentaje2<-duracion_con_lluvia%>%inner_join(duracion_sin_lluvia,by="Género")
porcentaje2<-porcentaje2%>%mutate(diferencia_en_porcentaje=100-(promedio_c.lluvia/promedio_s.lluvia)*100)

head(porcentaje2)
```

```
# A tibble: 3 × 4
  Género promedio_c.lluvia promedio_s.lluvia diferencia_en_porcentaje
  <chr>             <dbl>             <dbl>                    <dbl>
1 FEMALE            1176.             1184.                    0.710
2 MALE              1076.             1091.                    1.30
3 OTHER             1132.             1143.                    0.955
```

Como se observa en la tabla, la diferencia de duración de los viajes, en los días con lluvia y los días sin lluvia, no parece ser significativa, al rondar el 1%.

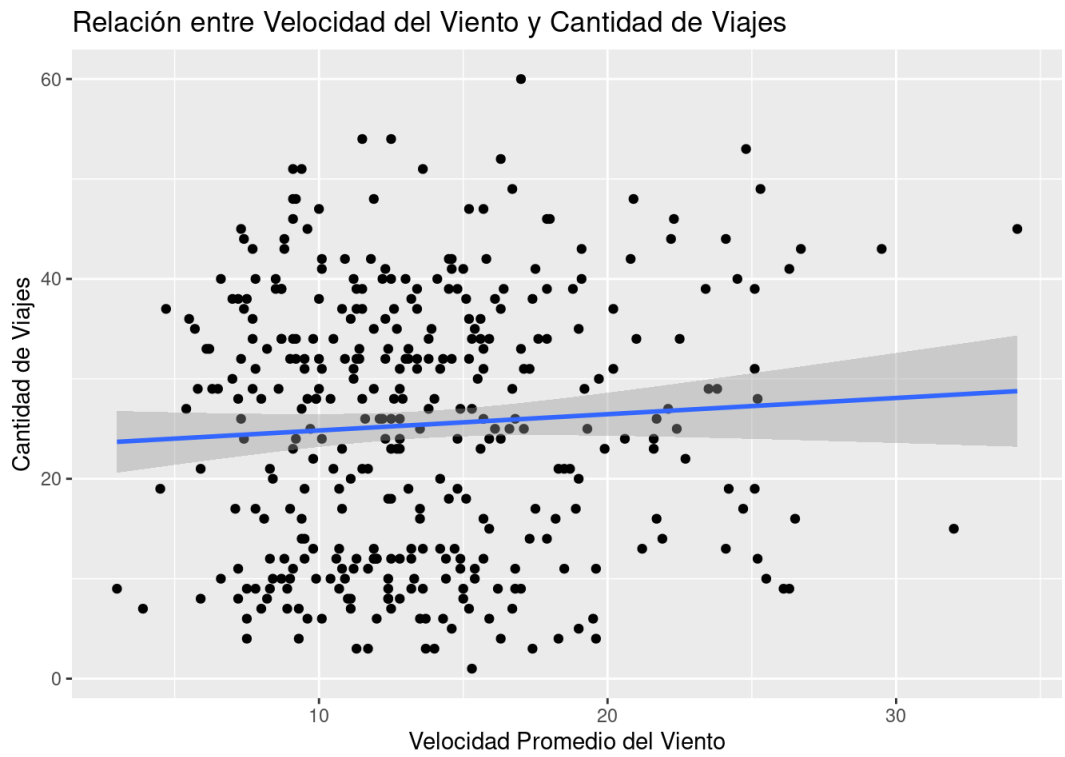
Por lo tanto, podemos decir, que las precipitaciones afectan enormemente al uso de la Ecobici, pero no modifican demasiado la duración de los viajes.

Viento

Lo primero que se observó, fue si la velocidad del viento afecta la cantidad de viajes que se realizan. Es decir, si existe un patrón del tipo: a mayor velocidad del viento, menos viajes. Para esto, observamos la velocidad promedio del viento por día, y la relacionamos con la cantidad de viajes que se hicieron durante ese mismo día.

```
wspd_promedio <- union_bici_clima %>%
  group_by(date) %>%
  summarise(viajes_totales = n(), velocidad_promedio_viento = mean(wspd), na.rm = TRUE
)

ggplot(wspd_promedio, aes(x = velocidad_promedio_viento, y = viajes_totales))
+
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Velocidad Promedio del Viento", y = "Cantidad de Viajes")
+
  ggtitle("Relación entre Velocidad del Viento y Cantidad de Viajes"
)
```

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



Por lo que se ve en el gráfico, no parece haber una relación directa entre la velocidad del viento y la cantidad de viajes realizados.

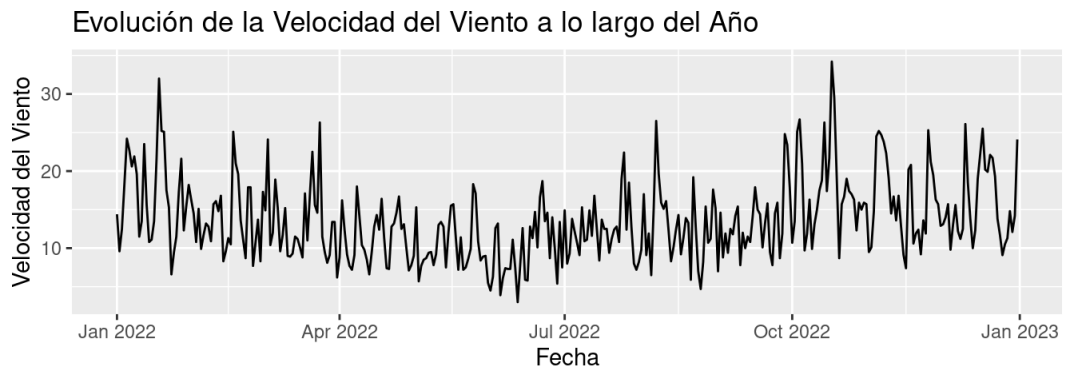
Luego, observamos si la velocidad del viento afecta la duración de los viajes. Para esto, comparamos la variación de la velocidad del viento a lo largo del año, con el promedio de duración de los viajes por mes:

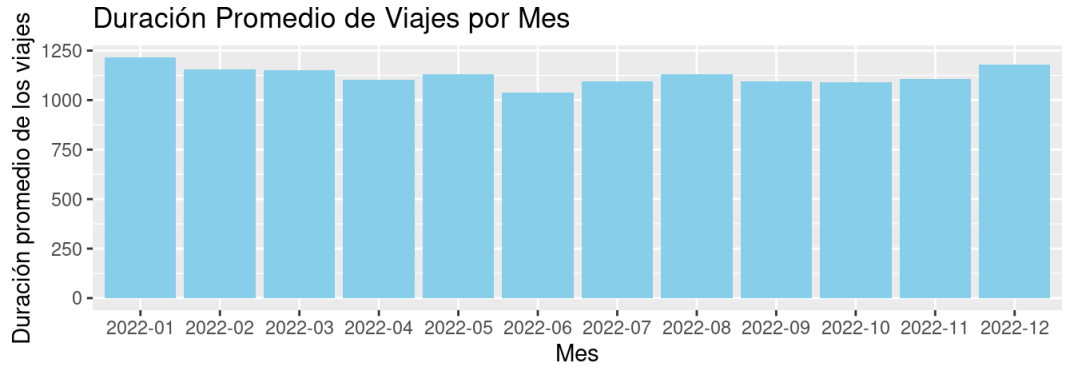
```
duracion_por_mes <- data_bici %>%
  mutate(Mes = format(fecha, "%Y-%m")) %>%
  group_by(Mes) %>%
  summarise(duracion_promedio = mean(duracion_recorrido))

evo_viento_año <- ggplot(union_bici_clima, aes(x = date, y = wspd))
+
  geom_line() +
  labs(x = "Fecha", y = "Velocidad del Viento")
+
  ggtitle("Evolución de la Velocidad del Viento a lo largo del Año")

mesPromedio <- ggplot(duracion_por_mes, aes(x = Mes, y = duracion_promedio))
+
  geom_bar(stat = "summary", fun = "mean", fill = "skyblue")
+
  labs(title = "Duración Promedio de Viajes por Mes", x = "Mes", y = "Duración promedio de los viajes"
)
```

```
require(gridExtra)
```

```
Loading required package: gridExtra


Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```
grid.arrange(evo_viento_año, mesPromedio, ncol = 1
)
```





Observando el comportamiento de ambos gráficos, notamos como la duración promedio de los viajes por mes no varía demasiado, y no tiene un patrón similar a la velocidad del viento por mes.

Es cierto que se aprecia como Julio es tanto el mes en el que se registró la menor velocidad del viento y la menor duración promedio del viajes. Sin embargo, esto parece ser más una casualidad estadística que un patrón, ya que con la temperatura se explica la menor duración de los viajes durante el mes de Julio.

Todo parece indicar que la velocidad del viento no afecta en demasía al uso de la Ecobici. Para terminar de confirmar esto, veamos que sucedió el día de 2022 en el que el viento viajo, en promedio, a mayor velocidad:

```
indice_fila_max <- which.max(union_bici_clima$wspd)
print(union_bici_clima[indice_fila_max, 20])
```

```
[1] "2022-10-17"
```

Sabemos que el 17 de Octubre fue el día en el que, en promedio, el viento viajó a una mayor velocidad.

Veamos cuantos viajes se realizaron durante este día:

```
fecha_deseada <- as.Date("2022-10-17")

viajes_en_fecha_deseada <- union_bici_clima[union_bici_clima$date == fecha_deseada,]
count(viajes_en_fecha_deseada)
```

```
  n
1 45
```

Durante el 17 de Octubre, se realizaron 45 viajes. Un número que puede parecer alto. Comparemoslo con el promedio de viajes que se realizan por día:

```
promedio_por_día <- union_bici_clima %>%
  group_by(date) %>%
  count()

print(mean(promedio_por_día$n))
```

```
[1] 25.42033
```

El promedio de viajes por día es de 25.42 viajes por día, y se observa como el 17 de Octubre hubo 20 viajes más que el promedio. Ahora bien, es importate considerar el contexto. El 17 de Octubre, al ser el Día de la Lealtad, se realizan muchas movilizaciones por el centro de la Ciudad de Buenos Aires. Esto puede explicar la elevada cantidad de viajes que se dan durante este día.

Por lo tanto, al ser el 17 de Octubre un caso especial, veamos que sucedió el segundo día en el que el viento viajo a una mayor velocidad.

Realizando un analisis similar al anterior, encontramos que el 18 de Enero es el día que buscamos. Veamos cuantos viajes se realizaron:

```
fecha_deseada <- as.Date("2022-1-18")

viajes_en_fecha_deseada <- union_bici_clima[union_bici_clima$date == fecha_deseada,]
count(viajes_en_fecha_deseada)
```

```
  n
1 15
```

Durante este día, se realizaron 15 viajes, 10 viajes menos que el promedio. Ahora bien, ¿es el viento el factor que hace que se utilice menos la Ecobici?

Por todo lo analizado anteriormente, no parece serlo. Ya vimos que durante Enero, el uso de la Ecobici disminuye de forma general, y este puede explicarse gracias a la temperatura.

Con todo esto, concluimos que la velocidad a la que viaja el viento no es un factor determinante que condicione el uso de la Ecobici. Esto se debe, muy probablemente, a que el viento no viaja a velocidades tan altas como para afectar de forma significativa los viajes en bicicleta.

Conclusión

Luego de haber analizado como varía la cantidad de viajes y su duración con respecto a la Temperatura, a las Precipitaciones y al Viento, pudimos observar que el Viento no juega un papel significativo y que solo podriamos ver casos muy especiales, como por ejemplo, algun dia de viento extremo, lo cual en 2022 no sucedió.

La Temperatura y las Precipitaciones son dos factores significantes en el uso de bicicleta y su duracion. Sin embargo solo la Temperatura debe ser considerada a la hora de tomar decisiones con respecto el servicio de alquiler de bicicletas. La Temperatura tiene un patron muy similar a lo largo de los años, entonces este factor se puede utilizar para, por ejemplo, reducir la cantidad de biciletas que se ofrecen en temperaturas extremas o temporadas donde las hay (Invierno/Verano), por ende menor mantenimiento y menores gastos. Los dias de lluvia, en cambio, no tienen cierto patron a diferencia de la Temperatura, sabemos que las temporadas mas lluviosas son en otoño y primavera, pero deberiamos hacer un seguimiento de la tormentas y un analisis meteorologico, el cual podria ser innecesariamente costoso.