Trabajo práctico 1: Estadística descriptiva

Introducción

Este trabajo consiste en el análisis descriptivo del sistema EcoBici, un sistema de bicicletas compartidas que funciona hace varios años en la Ciudad de Buenos Aires, y cuenta con 200 estaciones y 1200 rodados.

Fuente de los datos

Los datos analizados fueron extraídos de las bases de datos proporcionadas por la Ciudad de Buenos Aires (https://data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas). Se usarán dos bases de datos:

- **Usuarios**: personas que usan EcoBici
  - Tamaño de la muestra: 100 elementos
  - Variables:
    - id_usuario: identificador único que representa a cada usuario (cualitativa)
    - genero_usuario: género del usuario (cualitativa)
    - edad_usuario: edad del usuario (cuantitativa continua)
- **Recorridos**: datos sobre los viajes
  - Tamaño de la muestra: 380 elementos
  - Variables:
    - id_usuario: identificador correspondiente a la base Usuarios
    - direccion_estacion_origen: la estación de donde la bicicleta fue sacada expresada como cadena de caracteres (cualitativa)
    - direccion_estacion_destino: la estación donde la bicicleta fue devuelta expresada como cadena de caracteres (cualitativa)
    - duracion_recorrido: el tiempo, en segundos, que la bicicleta fue usada (cuantitativa continua)
    - distancia: la longitud, en metros, de la estación de origen hasta la estación de destino (cuantitativa continua)
    - dia: el día de la semana que la bicicleta fue usada

Análisis univariado de Usuarios

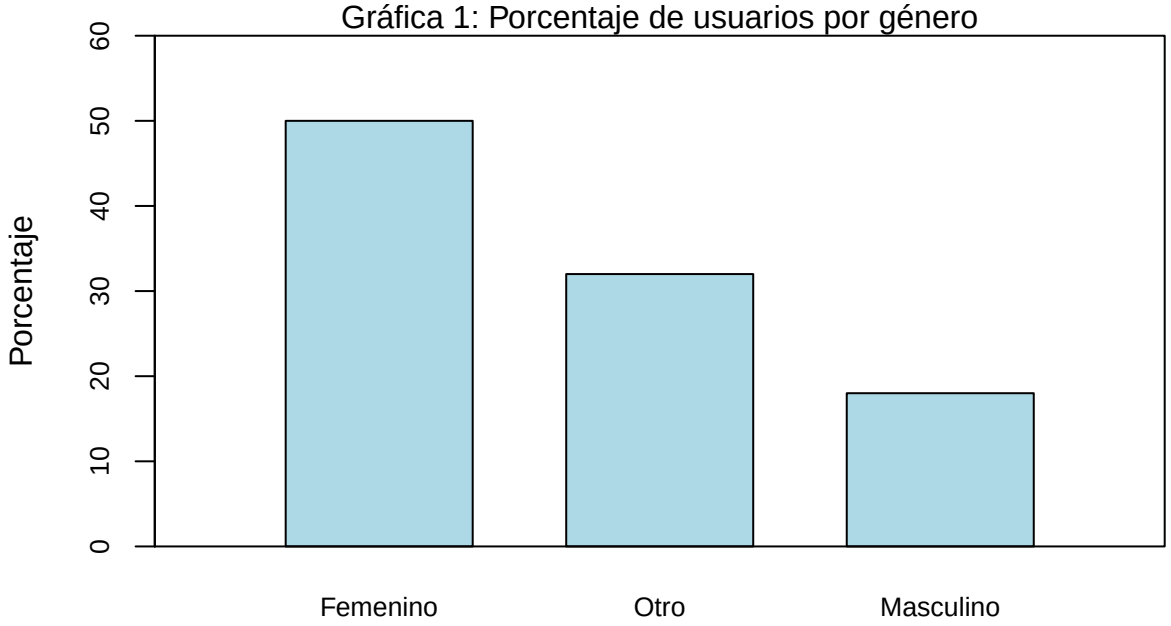
Género

```
##           Frec.absoluta Frec.relativa
## Femenino             50          0.50
## Masculino            18          0.18
## Otro                 32          0.32
## Total               100          1.00
```

Tabla 1: usuarios por género

```
##       Moda
## "Femenino"
```

Medidas descriptivas



Se puede observar que la mayoría de los usuarios de la muestra son mujeres, siendo la mitad de los mismos. También es llamativa la cantidad de personas que pertenecen a "Otro", pero quizás esto puede ser porque cuando los usuarios se registran y no especifican género, toma el valor "Otro" por defecto.

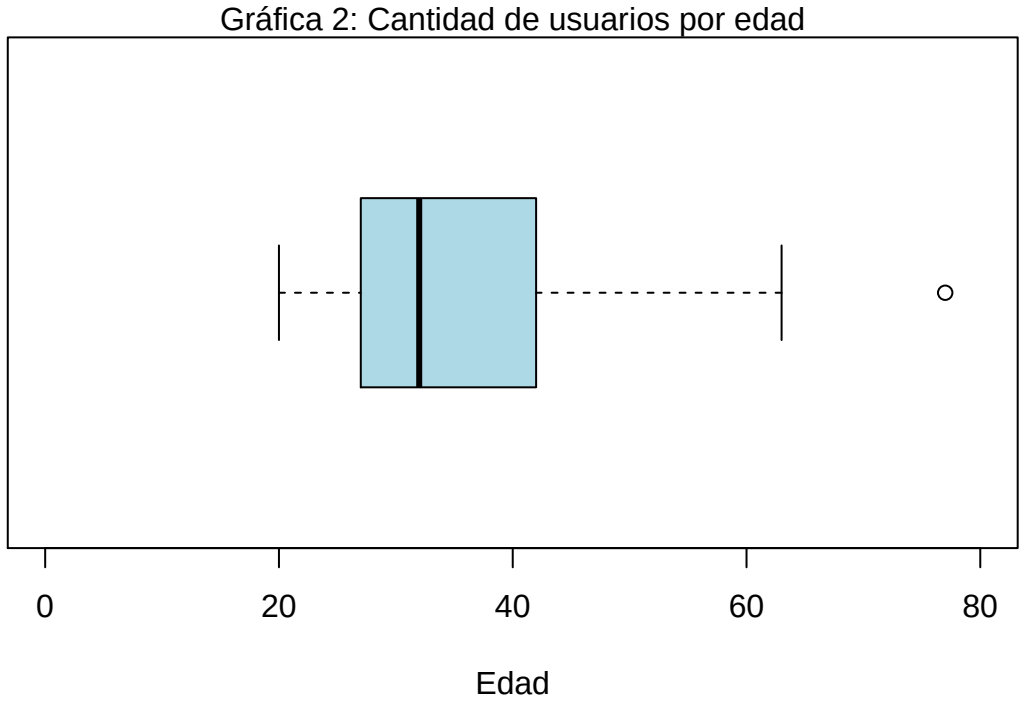
Edad

	Frec.absoluta	Frec.abs.acumulada	Frec.relativa	Frec.rel.acumulada
[20,25)	18	18	0.19	0.19
[25,30)	21	39	0.22	0.40
[30,35)	23	62	0.24	0.64
[35,40)	4	66	0.04	0.68
[40,45)	12	78	0.12	0.80
[45,50)	8	86	0.08	0.89
[50,55)	7	93	0.07	0.96
[55,60)	2	95	0.02	0.98
[60,65)	1	96	0.01	0.99
[65,70)	0	96	0.00	0.99
[70,75)	0	96	0.00	0.99
[75,80)	1	97	0.01	1.00

Tabla 2: usuarios por edad

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
20.00	27.00	32.00	34.52	42.00	77.00	3

Medidas descriptivas



Acá se ve una mayor concentración de usuarios entre los 25 y 35 años aproximadamente, siendo la mitad menor a 32 años. El 90% de los usuarios son menores a 50 años, como se ve en la tabla de frecuencias.

El rango va desde los 20 hasta los 77, aunque es posible que las edades mayores a 60 sean producto de errores en el registro de la información. En el gráfico se ve que hay un outlier en el valor máximo, 77 años.

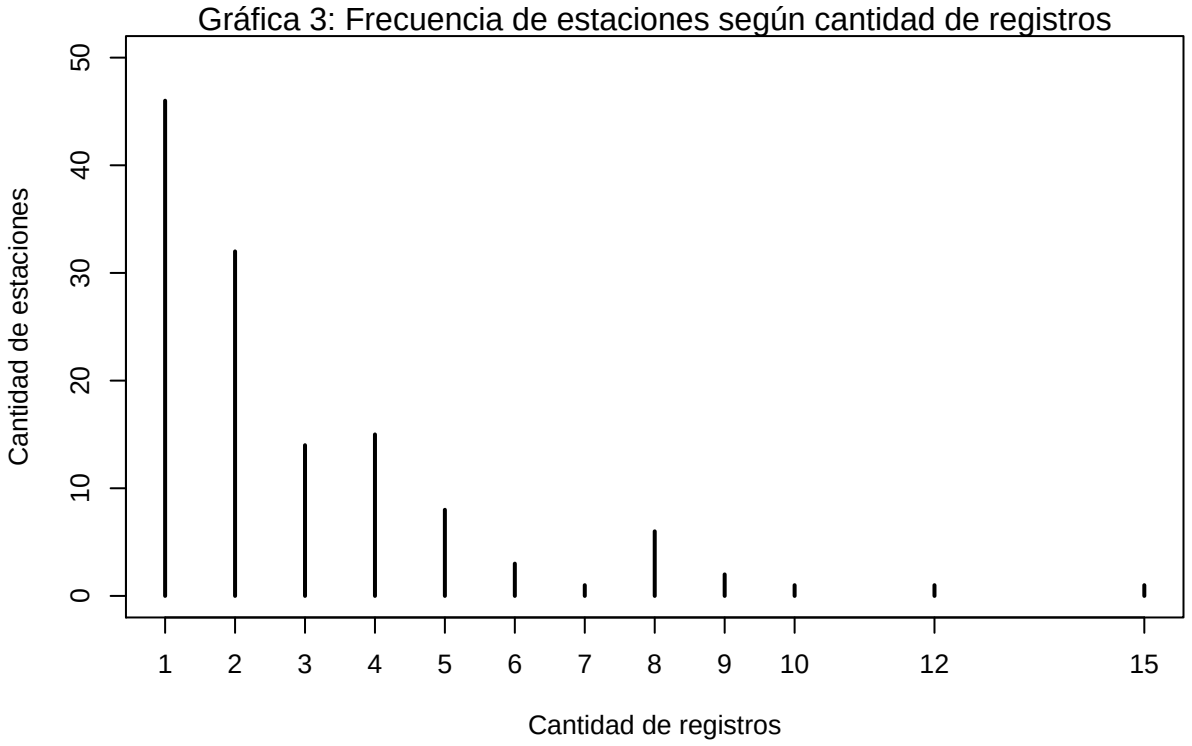
También hubo 3 registros que no tenían edad, y no se tuvieron en cuenta para el análisis.

Análisis univariado de Recorridos

Estación de origen

	Frecuencia
1	46
2	32
3	14
4	15
5	8
6	3
7	1
8	6
9	2
10	1
12	1
15	1
Total	130

Tabla 3: Estaciones de origen según cantidad de veces registradas



Dado que estación de origen toma demasiados valores distintos, se eligió agruparlas según la cantidad de veces que fueron registradas. La tabla 3 se entiende como: hubo 46 estaciones que fueron registradas 1 vez, 32 que fueron registradas 2 veces, etc. Se puede ver que la mayoría de las estaciones fueron usadas muy pocas veces según esta muestra.

A continuación se realizará el análisis reduciendo la muestra a las 10 estaciones más concurridas.

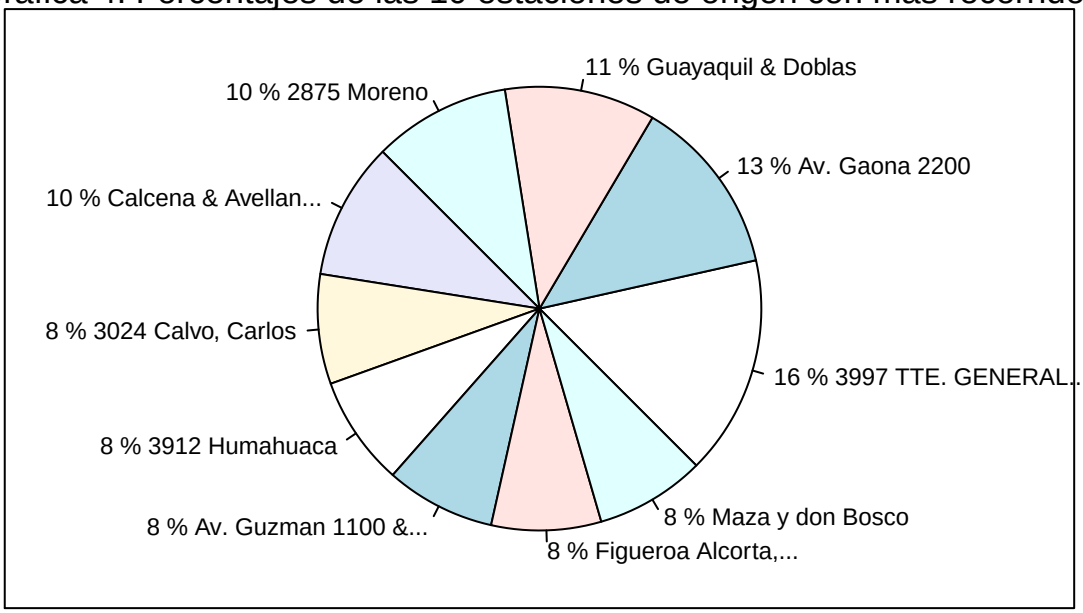
	Frec. absoluta	Frec. relativa
3997 TTE. GENERAL JUAN DOMINGO PERON	15	0.16
Av. Gaona 2200	12	0.13
Guayaquil & Doblas	10	0.11
2875 Moreno	9	0.10
Calcena & Avellaneda Av.	9	0.10
3024 Calvo, Carlos	8	0.08
3912 Humahuaca	8	0.08
Av. Guzman 1100 & Av. Corrientes	8	0.08
Figueroa Alcorta, Pres. Av. & Dorrego Av.	8	0.08
Maza y don Bosco	8	0.08
Total	95	1.00

Tabla 4: Las 10 estaciones de origen más comunes

Moda
"3997 TTE. GENERAL JUAN DOMINGO PERON"

Medidas descriptivas

Gráfica 4: Porcentajes de las 10 estaciones de origen con más recorridos

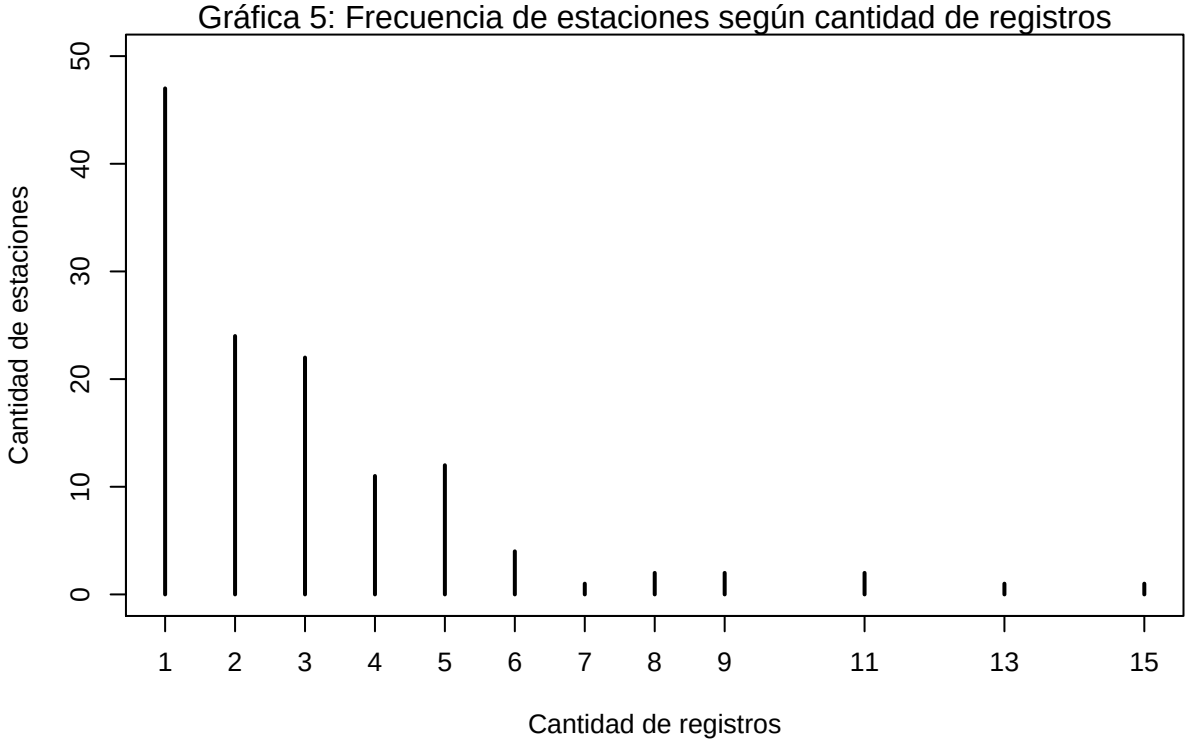


Las 10 estaciónes de origen más frecuentes representan aproximadamente el 25% de la muestra, es decir, el primer cuartil. De un total de 380 registros, se redujo a 98. Y la más común es “TTE. GENERAL JUAN DOMINGO PERON”.

Estación de destino

	Frecuencia
1	47
2	24
3	22
4	11
5	12
6	4
7	1
8	2
9	2
11	2
13	1
15	1
Total	129

Tabla 5: Estaciones de destino según cantidad de veces registradas



Se hizo el mismo agrupamiento de datos que en estaciones de origen. Acá también la mayoría de las estaciones registran poca frecuencia.

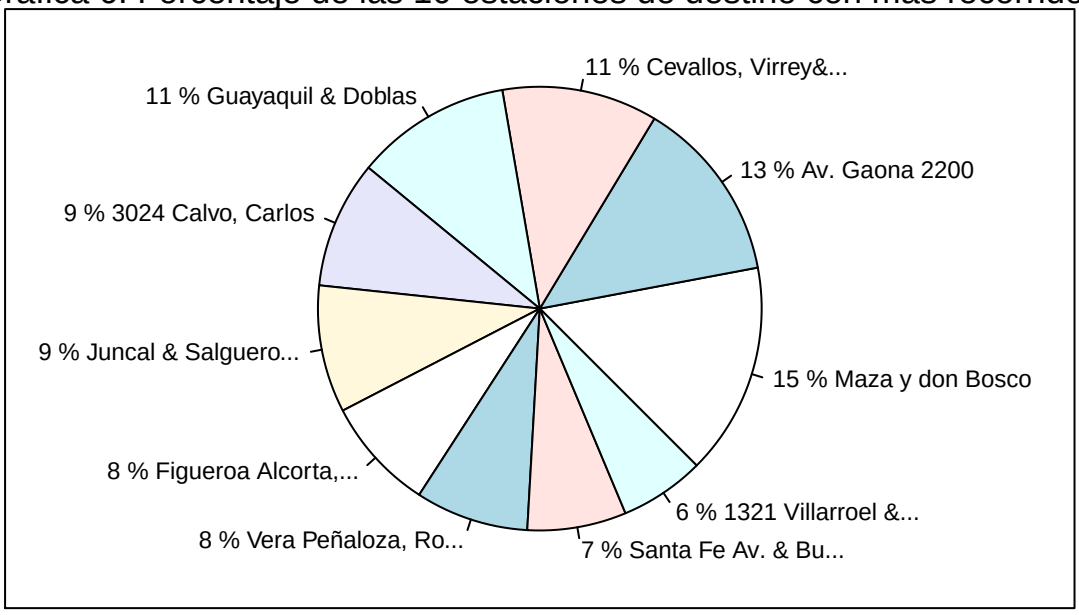
```
##                                          Frec. absoluta Frec. relativa
## Maza y don Bosco                                     15           0.15
## Av. Gaona 2200                                       13           0.13
## Cevallos, Virrey& Yrigoyen, Hipolito Av.             11           0.11
## Guayaquil & Doblas                                   11           0.11
## 3024 Calvo, Carlos                                    9           0.09
## Juncal & Salguero, Jeronimo                           9           0.09
## Figueroa Alcorta, Pres. Av. & Dorrego Av.             8           0.08
## Vera Peñaloza, Rosario 599 & Lanteri Julieta          8           0.08
## Santa Fe Av. & Bullrich, Int. Av.                     7           0.07
## 1321 Villarroel & Humboldt CABA                       6           0.06
## Total                                                97           1.00
```

Tabla 6: Las 10 estaciones de destino más comunes

```
##               Moda
## "Maza y don Bosco"
```

Medidas descriptivas

Gráfica 6: Porcentaje de las 10 estaciones de destino con más recorridos



Al igual que con las de origen, se muestran las 10 estaciones de destino más comunes, correspondientes también al primer cuartil. Siendo la más común “Maza y don Bosco”.

En este caso después de reducir la muestra quedaron 97 registros.

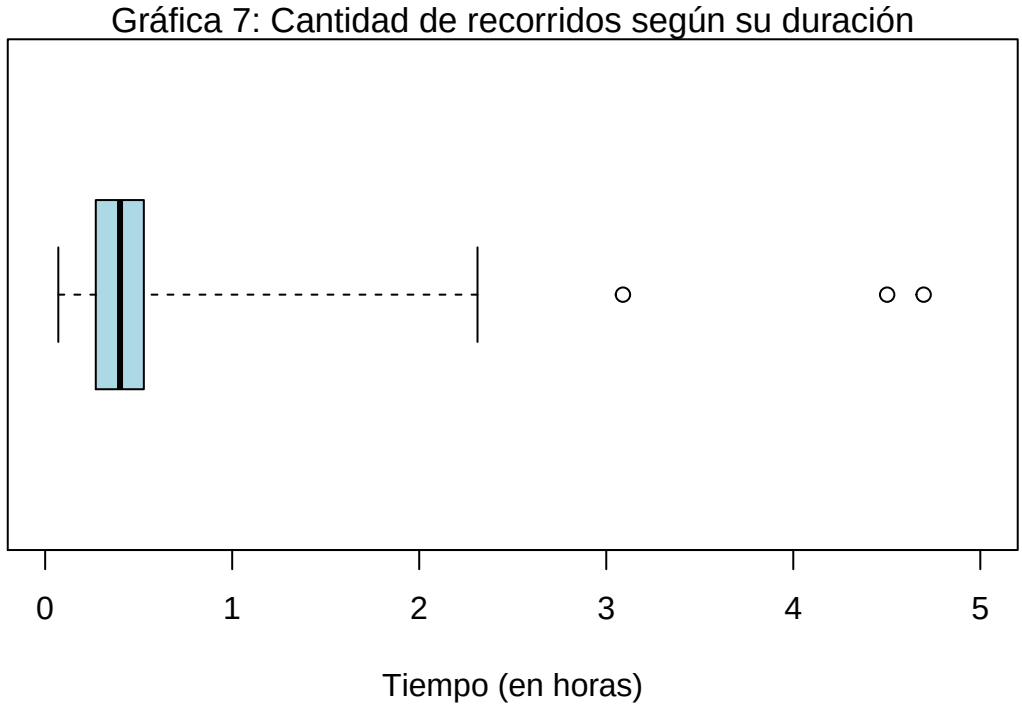
Duración

	Frec.absoluta	Frec.abs.acumulada	Frec.relativa	Frec.rel.acumulada
[0,0.5)	261	261	0.69	0.69
[0.5,1)	96	357	0.25	0.94
[1,1.5)	10	367	0.03	0.97
[1.5,2)	5	372	0.01	0.98
[2,2.5)	5	377	0.01	0.99
[2.5,3)	0	377	0.00	0.99
[3,3.5)	1	378	0.00	0.99
[3.5,4)	0	378	0.00	0.99
[4,4.5)	0	378	0.00	0.99
[4.5,5)	2	380	0.01	1.00

Tabla 7: Duración (en horas) de los recorridos

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.07	0.27	0.40	0.49	0.53	4.70

Medidas descriptivas



Los datos se agruparon en intervalos de media hora para mejor visualización.

Por lo que se ve en el gráfico y en la tabla, la gran mayoría de los viajes (69%) duraron media hora o menos, y casi todos (94%) duraron menos de 1 hora. También se observa que hubo dos de entre 4 y 5 horas, y uno de un poco más de 3 horas, que aparecen como outliers en el gráfico.

La mitad de las personas anduvieron 0.4 horas (24 minutos) o menos, y el total anduvo en promedio 0.49 horas (~30 minutos).

El 10% de los viajes de menor duración son de 10 minutos o menos. Podría ser el caso de estaciones que estan cerca una de la otra, o de usos en la zona que terminaron en lugar de partida.

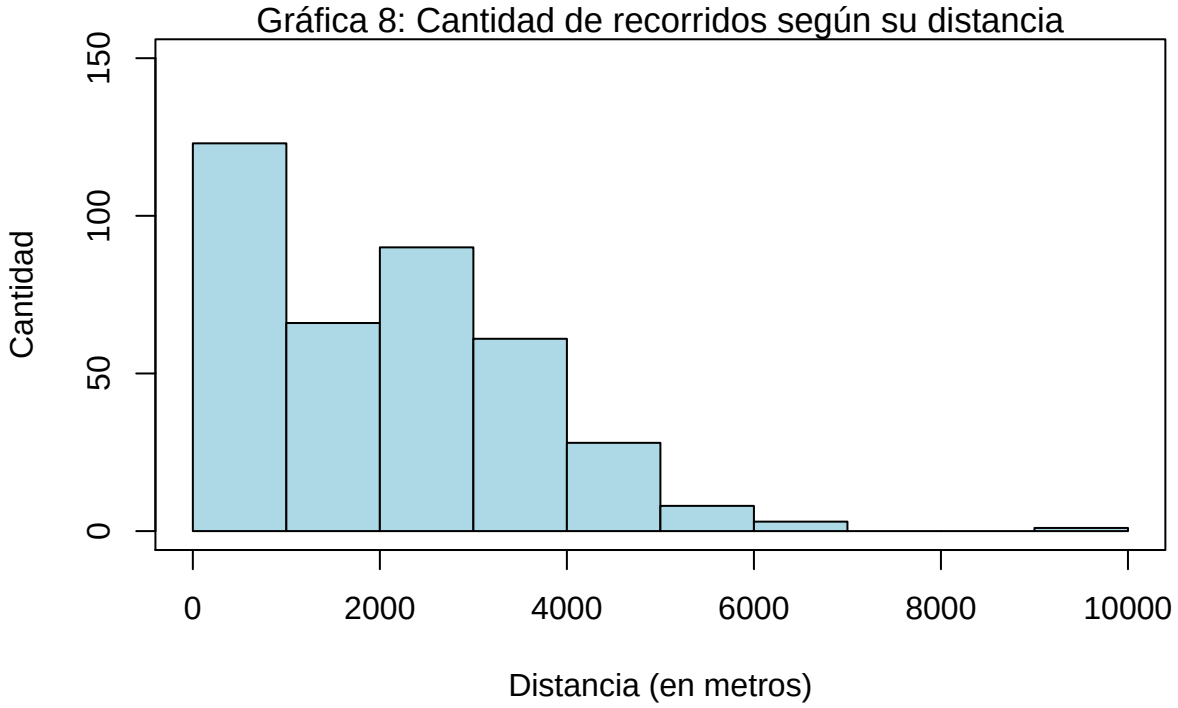
Distancia

	Frec.absoluta	Frec.abs.acumulada	Frec.relativa	Frec.rel.acumulada
[0,1000)	123	123	0.32	0.32
[1000,2000)	66	189	0.17	0.50
[2000,3000)	90	279	0.24	0.73
[3000,4000)	61	340	0.16	0.89
[4000,5000)	28	368	0.07	0.97
[5000,6000)	8	376	0.02	0.99
[6000,7000)	3	379	0.01	1.00
[7000,8000)	0	379	0.00	1.00
[8000,9000)	0	379	0.00	1.00
[9000,10000)	1	380	0.00	1.00

Tabla 8: Distancia (en metros) de los recorridos

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	758.3	2022.7	1998.2	3075.0	9756.6

Medidas descriptivas



Las distancias recorridas son proporcionales a las duraciones analizadas antes, que corresponden a viajes cortos. Lo que aparece acá, y que no se ve en el gráfico ni en la tabla, es que casi el 20% de los registros son de 0 metros. La distancia tomada es la de la estación de origen hasta la de destino, lo que da un mínimo de distancia recorrida, y no la distancia efectiva. Por lo que las muestras que registran distancia 0 son de recorridos que empiezan y terminan en la misma estación.

La mitad de los usuarios anduvo 2022 metros, y el promedio fue de 1998 metros.

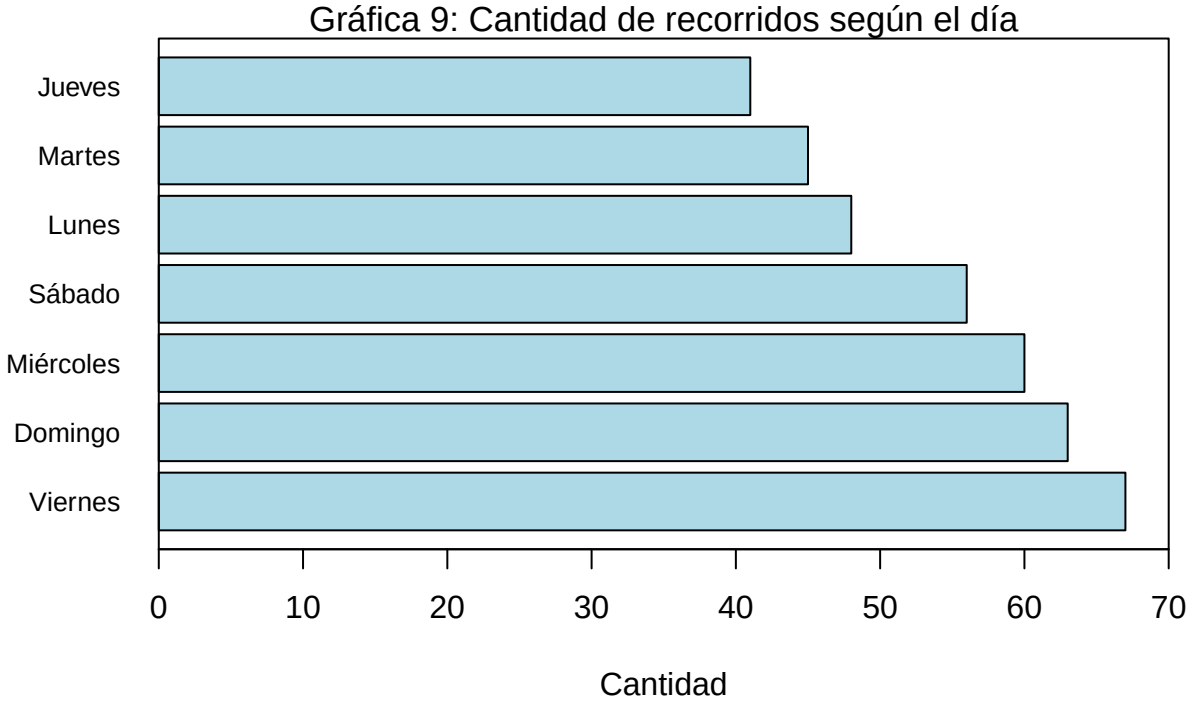
Días

```
##           Frec.absoluta Frec.relativa
## Viernes              67          0.18
## Domingo              63          0.17
## Miércoles            60          0.16
## Sábado               56          0.15
## Lunes                48          0.13
## Martes               45          0.12
## Jueves               41          0.11
## Total               380          1.00
```

Tabla 9: Días de los recorridos

```
##      Moda
## "Viernes"
```

Medidas descriptivas



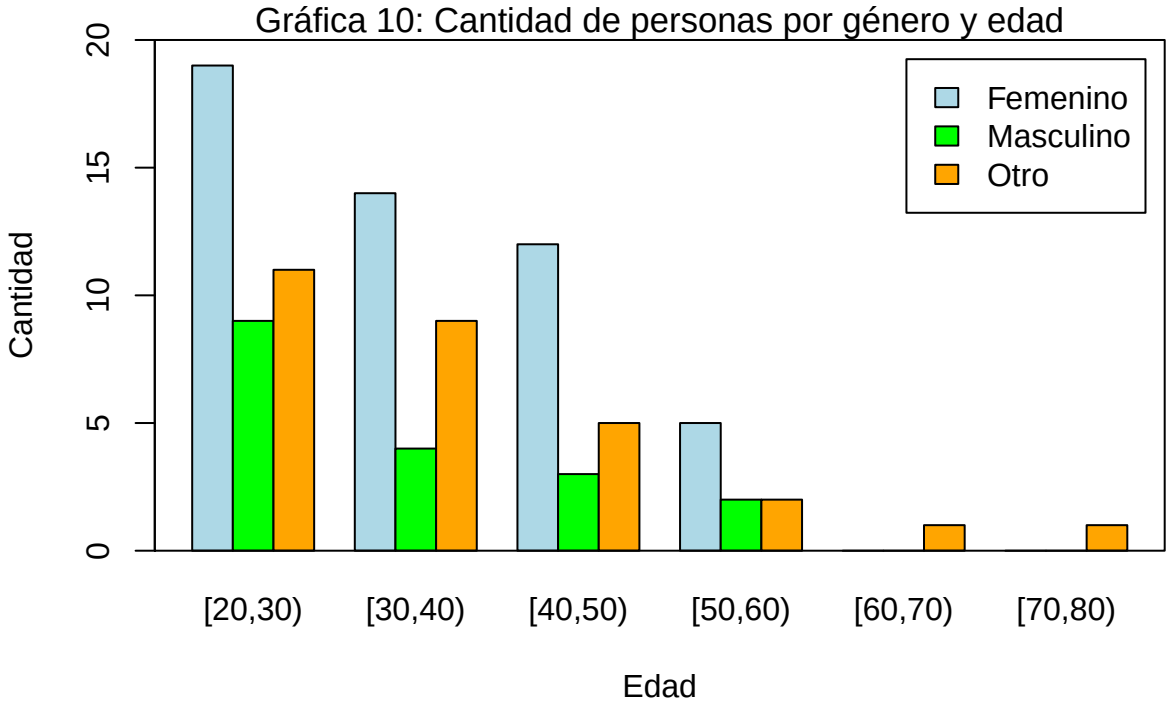
En los días viernes se ha registrado un mayor uso de bicicletas, pero no varía mucho la cantidad entre todos los días de la semana.

Análisis bivariado

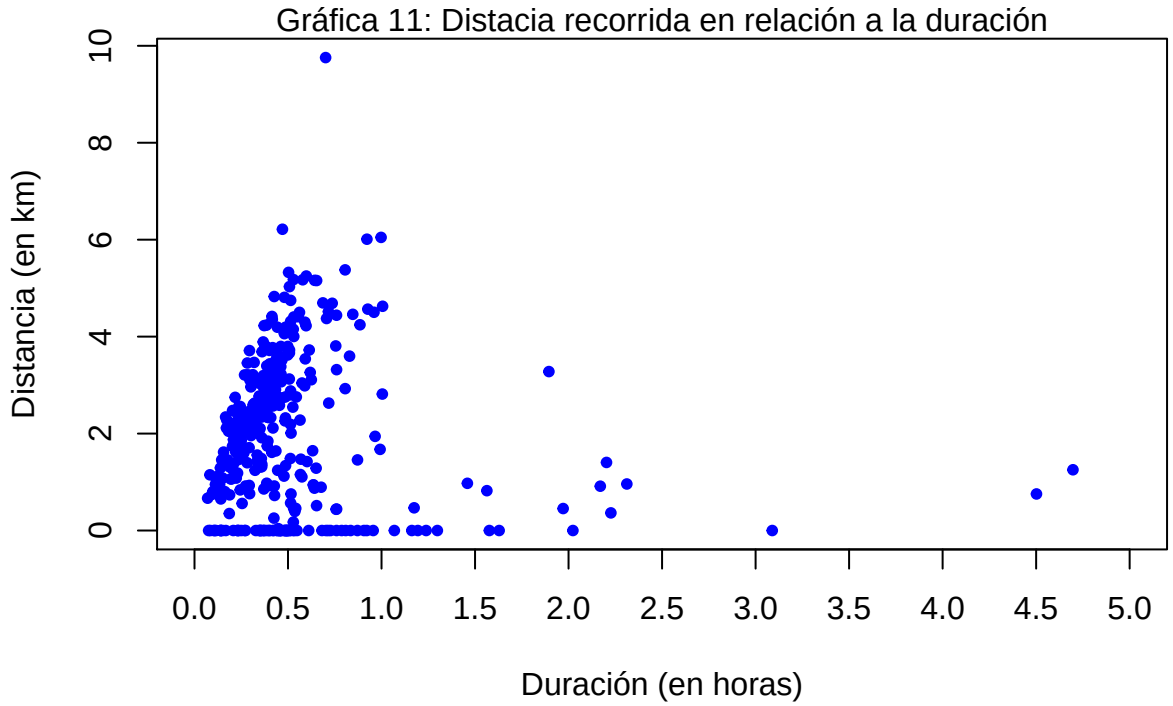
Género y edad

```
##           [20,30) [30,40) [40,50) [50,60) [60,70) [70,80)
## Femenino       19      14      12       5       0       0 50
## Masculino       9       4       3       2       0       0 18
## Otro           11       9       5       2       1       1 29
## Total          39      27      20       9       1       1 97
```

Tabla 10: Usuarios por género y edad



Duración y distancia



Se decidió hacer el análisis bivariado de estas dos mediciones para tratar de entender los valores extremos que surgían del análisis univariado.

En el caso de la distancia se podía ver un valor extremo de casi 10km, y acá se nota que la duración de ese recorrido se encuentra entre la media hora y la hora, más o menos a la mitad del intervalo, dando una velocidad aproximada de 13km/h. Se concluye que si bien es un valor alejado de los demás, no se debe a un error porque no es algo imposible que suceda.

Los valores extremos de la duración tampoco se podrían descartar como errores de medición, porque los usuarios pueden tomar las bicicletas, andar todo el tiempo que quieran, y después volver a la misma estación o a una cercana, lo que también explicaría por qué hay tantos recorridos de distancia 0.