Trabajo Práctico Probabilidad y Estadística

Bobe Julio y Grau Marianela

June 11, 2021

1 Presentación

En este informe se efectúa un estudio descriptivo que incluya: tablas de distribución de frecuencias, gráficos y medidas descriptivas con respecto a las dos bases de datos dadas: usuarios2.csv y recorridos2.csv. Se incluye un análisis univariado de cada una de las variables presentadas. A su vez, estará incluído un gráfico donde un análisis comparativo de una variable según los niveles de otra será llevado a cabo, con su adecuada interpretación bivariada.

2 Acerca de los datos

El par de datos que se presenta para su posterior estudio consta de dos unidades de análisis diferentes. Por un lado, existe información relativa a los usuarios que utilizaron el servicio durante el año 2020. Por el otro, se presenta información sobre los recorridos que realizó cada uno de los usuarios. Ambos conjuntos de datos se dan a conocer en dos bases de datos: usuarios2 y recorridos2.

- usuarios2: dicha base de datos provee los ids. de los usuarios que andan en bicicleta, su sexo y sus respectivas edades. La cantidad de personas que provee dicha base de datos es de 100.
- recorridos2: dicha base de datos también provee los ids. de los usuarios que andan en bicicleta, por lo tanto, acá se muestra la conexión o unión entre ambas bases de datos. Además, recorridos2 proporciona la dirección de la estación de origen y llegada, la distancia recorrida (en metros), el día y la duración del viaje (en segundos). La cantidad de recorridos que provee dicha base de datos es de 561.

2.1 Variables

Unidad de análisis: persona que vive en Buenos Aires que usa el servicio EcoBici.

Muestra: 100 personas que viven en Buenos Aires que usan el servicio de EcoBici.

Base de datos usuario2:

- género_usuario: variable cualitativa categórica.
- edad_usuario: variable cuantitativa continua.

Luego, para la base de datos recorrido2:

Unidad de análisis: recorrido que implementa una persona que usa el servicio EcoBici.

Muestra: 561 recorridos en total realizados por las 100 personas de la muestra de la base de datos de usuarios2.

Base de datos recorridos2:

- duración: variable cuantitativa continua.
- distancia: variable cuantitativa continua.
- día: variable cualitativa categórica.
- dirección_estación_origen: variable cualitativa categórica.
- dirección_estación_llegada: variable cualitativa categórica.

3 Análisis descriptivo

Dicho análisis se realiza utilizando tablas de frecuencia, gráficos y medidas resumen oportunas para cada variable considerada teniendo en cuenta si es cualitativa (nominal u ordinal) o cuantitativa (discreta o continua).
Primero, se estudia cada variable por separado. En una segunda parte, se analiza la relación que se crea mediante una variable según los niveles de otra.

3.1 Análisis univariado

Se comenzará a analizar cada variable por separado.

3.1.1 Género

Tabla de distribución de frecuencias:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual
Femenino	44	0.44	44
Masculino	31	0.31	31
Otro	25	0.25	25
Total	100	1.00	100

Figure 1: Tabla de frecuencias de la variable género.

Análisis de la segunda fila de la tabla:

- Los hombres representan el 31% de las personas que usan el servicio de EcoBici.
- El género masculino es el segundo en cuanto a cantidad de personas que usa EcoBici.

Gráfico de sectores:

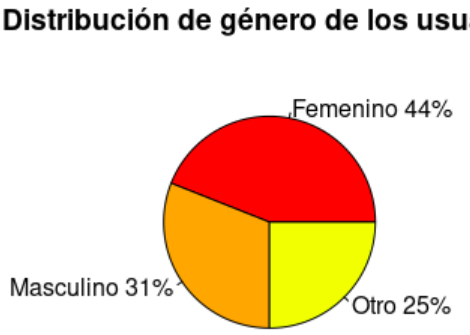

Figure 2: Gráfico de sectores de la variable género

En dicho gráfico se observa que el servicio de EcoBici es mayormente utilizado por el género femenino, seguido por el masculino y finalmente por otros.
Medida de interés:
Moda = Femenino.

3.1.2 Edad

Las edades fueron partidas en intervalos regulares de 7 años.

Tabla de distribución de edad (en años):

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual	Frecuencia Relativa Porcentual Acumulada
[18,25)	22	0.22	22.22	22.22
[25,32)	32	0.32	32.32	54.55
[32,39)	11	0.11	11.11	65.66
[39,46)	17	0.17	17.17	82.83
[46,53)	12	0.12	12.12	94.95
[53,60)	1	0.01	1.01	95.96
[60,67)	2	0.02	2.02	97.98
[67,74)	1	0.01	1.01	98.99
[74,81)	0	0.00	0.00	98.99
[81,88)	1	0.01	1.01	100.00
Total	99	1.00	100.00	100.00

Figure 3: Tabla de frecuencias de la variable edad (en años).

Análisis de la segunda fila de la tabla de edad:

- Las personas de entre los 25 años de edad hasta los 32 representan el 32.32% de las personas que usan el servicio de EcoBici.
- El 54.55% de los usuarios del servicio de EcoBici tienen hasta 32 años de edad.

Histograma de la variable edad:

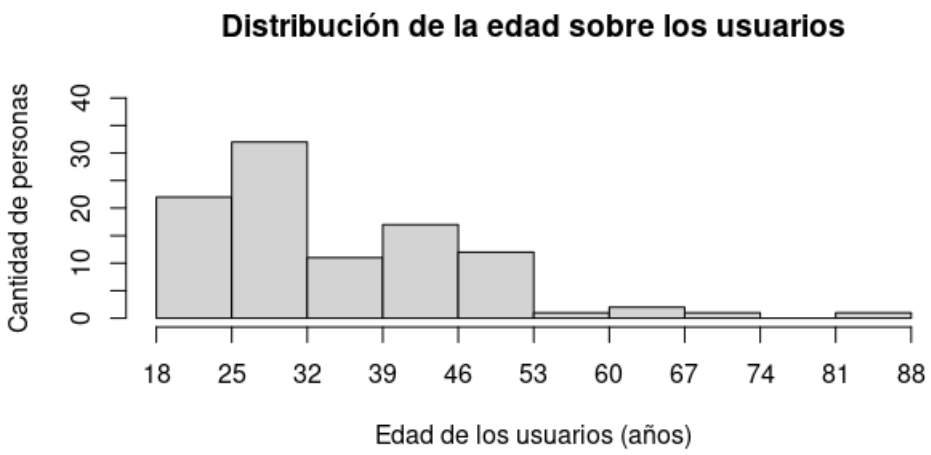

Figure 4: Histograma variable edad.

En el histograma se observa que la brecha de edad que más usa EcoBici es de los 25 hasta los 32 años.

Boxplot de la variable edad:

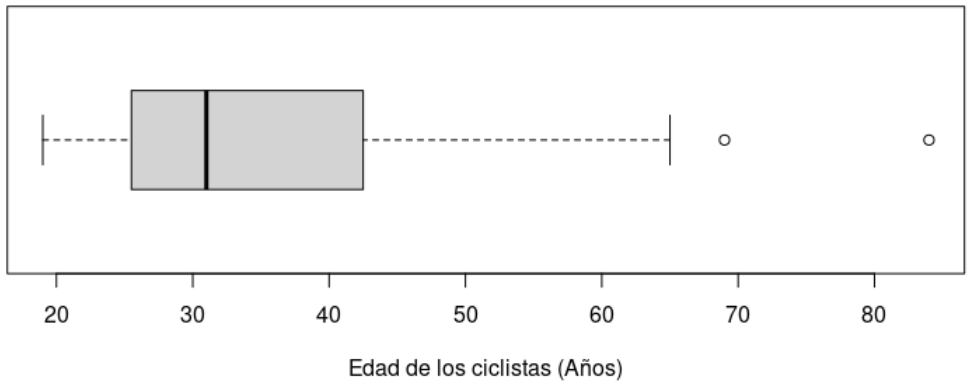

Figure 5: Boxplot de la variable edad.

En el boxplot se puede ver que la muestra posee dos outlaiers.

Medidas de interés:

- Mediana: 31.0
- Primer cuartil: 25.5

- Tercer cuartil: 42.5
- Rango intercuartil: 17.0

3.1.3 Día

Tabla de distribución de los días:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual
Domingo	65	0.12	11.59
Jueves	71	0.13	12.66
Lunes	109	0.19	19.43
Martes	80	0.14	14.26
Miércoles	73	0.13	13.01
Sábado	74	0.13	13.19
Viernes	89	0.16	15.86
Total	561	1.00	100.00

Figure 6: Tabla de frecuencias de la variable día.

Análisis de la segunda fila de la tabla de días:

- Los recorridos realizados los dias jueves por las personas que usan el servicio de EcoBici representan un 12.66%.

Gráfico de barras de la variable días:

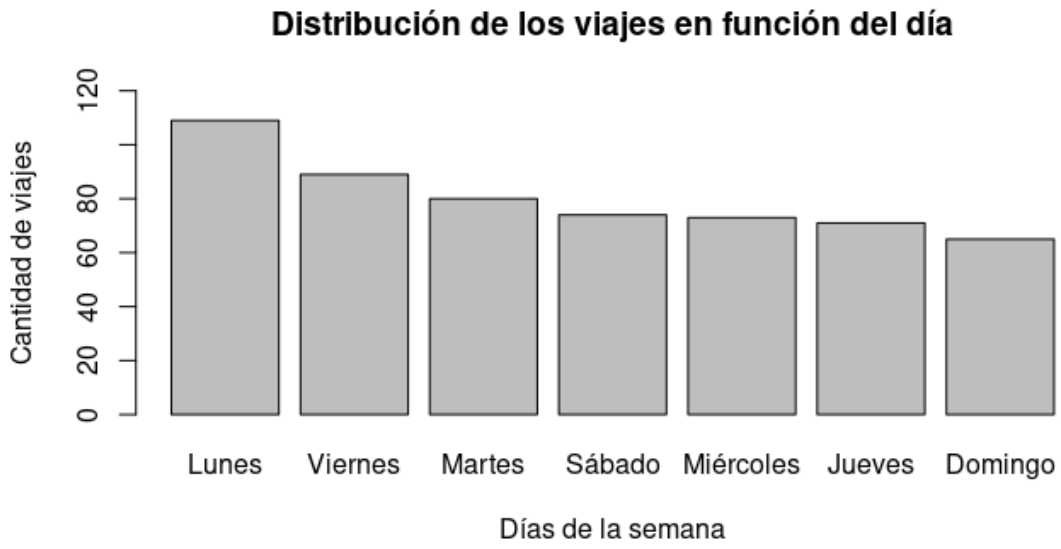

Figure 7: Gráfico de barras de la variable días.

En el gráfico de barras se puede ver que los días lunes y viernes son en los que el servicio de EcoBici tuvo mas uso.

Medida de interés:
Moda = Lunes.

3.1.4 Distancia

Se separó las distancias de los recorridos en intervalos de mil metros.

Tabla de frecuencias de las distancias recorridas:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual	Frecuencia Relativa Porcentual Acumulada
[0,1000)	181	0.32	32.26	32.26
[1000,2000)	229	0.41	40.82	73.08
[2000,3000)	68	0.12	12.12	85.20
[3000,4000)	50	0.09	8.91	94.12
[4000,5000)	17	0.03	3.03	97.15
[5000,6000)	5	0.01	0.89	98.04
[6000,7000)	8	0.01	1.43	99.47
[7000,8000)	0	0.00	0.00	99.47
[8000,9000)	2	0.00	0.36	99.82
[9000,10000)	1	0.00	0.18	100.00
Total	561	1.00	100.00	100.00

Figure 8: Tabla de frecuencias de las distancias recorridas.

Análisis de la segunda fila de la tabla de distancias:

- Se puede observar que 229 viajes usando el servicio de EcoBici recorrieron entre uno y dos kilómetros.
- El 73.08% de los viajes utilizando el servicio de EcoBici recorrieron hasta dos kilómetros.

Histograma de la variable distancia:

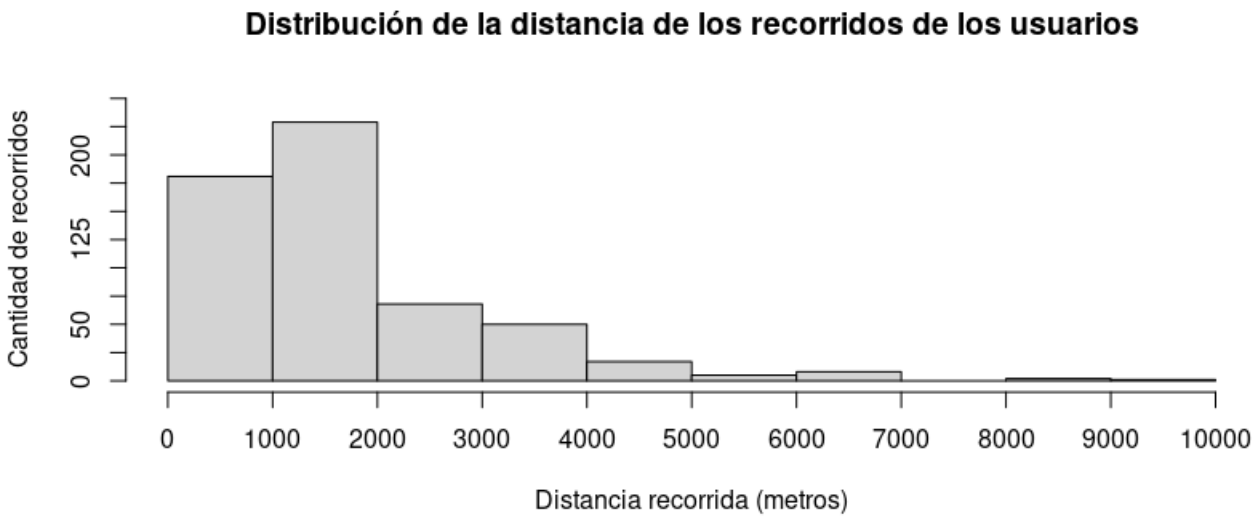

Figure 9: Histograma variable distancia.

En el histograma de la variable distancia se puede ver hay una poca cantidad de viajes que recorren más de cuatro kilómetros. Por otro lado, una gran parte de los viajes recorren hasta dos kilómetros.

Boxplot de la variable distancia:

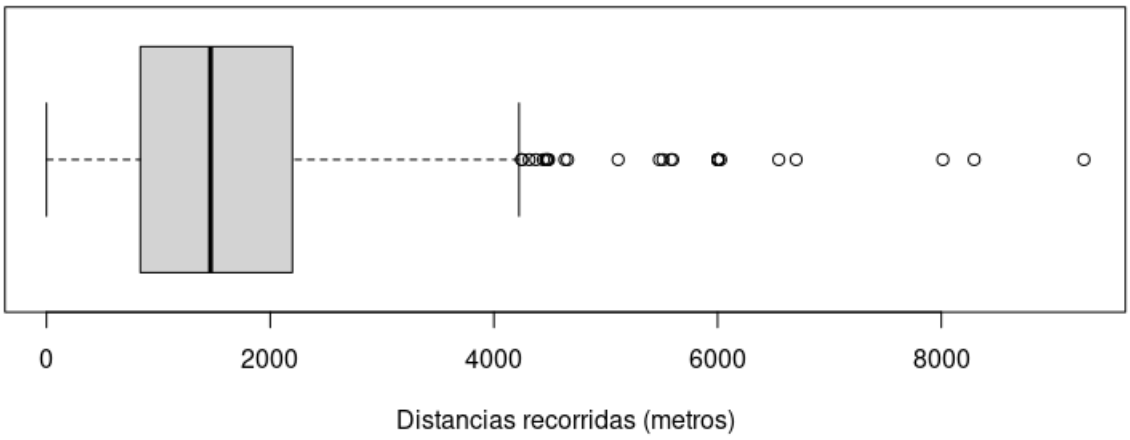

Figure 10: Boxplot variable distancia.

En el boxplot se pueden observar varios outliers lo que implica que hay ciertos viajes que constan de una cantidad superior de metros a comparación de los demás.

Medidas de interés:

- Mediana: 1466.68
- Primer cuartil: 838.78
- Tercer cuartil: 2198.54
- Rango intercuartil: 1359.76

3.1.5 Duración

Se pasó la información de la base de datos a minutos y luego se agrupó en intervalos de 15 minutos, hasta los 120 minutos. Los restantes se agruparon en un último intervalo.

Tabla de frecuencias de la duración de los viajes:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual	Frecuencia Relativa Porcentual Acumulada
[0,15)	235	0.42	41.89	41.89
[15,30)	179	0.32	31.91	73.80
[30,45)	76	0.14	13.55	87.34
[45,60)	34	0.06	6.06	93.40
[60,75)	16	0.03	2.85	96.26
[75,90)	8	0.01	1.43	97.68
[90,105)	7	0.01	1.25	98.93
[105,120)	1	0.00	0.18	99.11
[120, 540)	5	0.01	0.89	100.00
Total	561	1.00	100.00	100.00

Figure 11: Tabla de frecuencias de la variable duración.

Análisis de la segunda fila de la tabla de duración:

- 179 viajes duraron entre 15 y 30 minutos.
- El 73.80% de los viajes duran hasta 30 minutos.

Histograma de la variable duración:

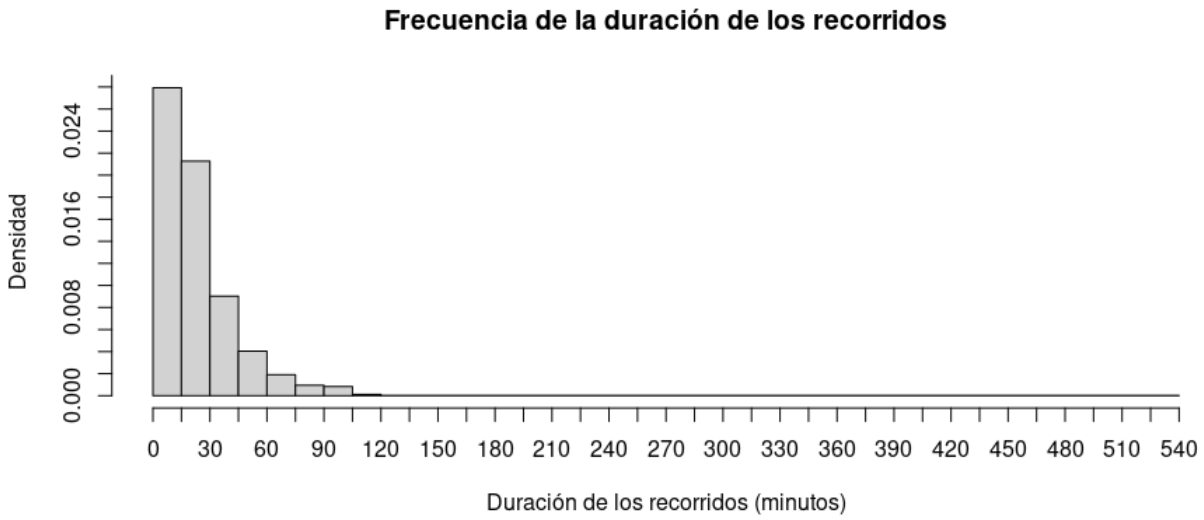

Figure 12: Histograma de la variable duración.

En el histograma se puede observar que hay una gran densidad de viajes que transcurren en a lo sumo 30 minutos. Luego la densidad decae abruptamente a partir de los 30 minutos en intervalos de 15 minutos.

Boxplot de la variable duración:

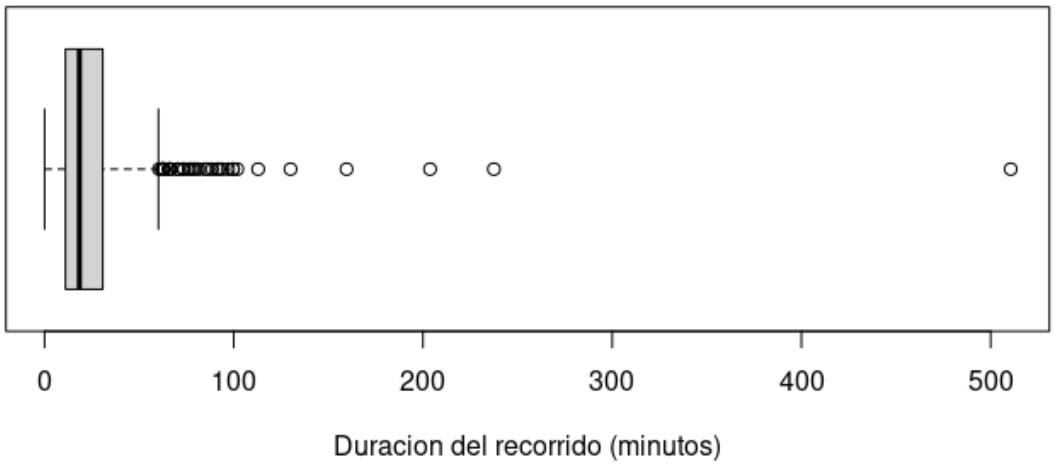

Figure 13: Boxplot de la variable duración.

En el boxplot se pueden apreciar una gran cantidad de outliers.

Medidas de interés:

- Mediana: 18.38
- Primer cuartil: 11.02
- Tercer cuartil: 30.77
- Rango intercuartil: 19.75

3.1.6 Estación de origen

Para esta variable se contaron cuántos viajes se originaron desde cada una de las estaciones y luego se agruparon por ese número.

Tabla de frecuencias de la variable estación de origen:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual	Frecuencia Relativa Porcentual Acumulada
1	44	0.32	31.88	31.88
2	31	0.22	22.46	54.35
3	12	0.09	8.70	63.04
4	19	0.14	13.77	76.81
5	6	0.04	4.35	81.16
6	3	0.02	2.17	83.33
7	4	0.03	2.90	86.23
8	4	0.03	2.90	89.13
9	5	0.04	3.62	92.75
10	2	0.01	1.45	94.20
>10	8	0.06	5.80	100.00
Total	138	1.00	100.00	100.00

Figure 14: Tabla de frecuencias de la variable estación de origen.

Análisis de la segunda fila de la tabla:

- Hay 31 estaciones de origen que fueron visitadas dos veces.
- El 54.35% de las estaciones origen fueron visitadas hasta a lo sumo dos veces.

Gráfico de Bastones:

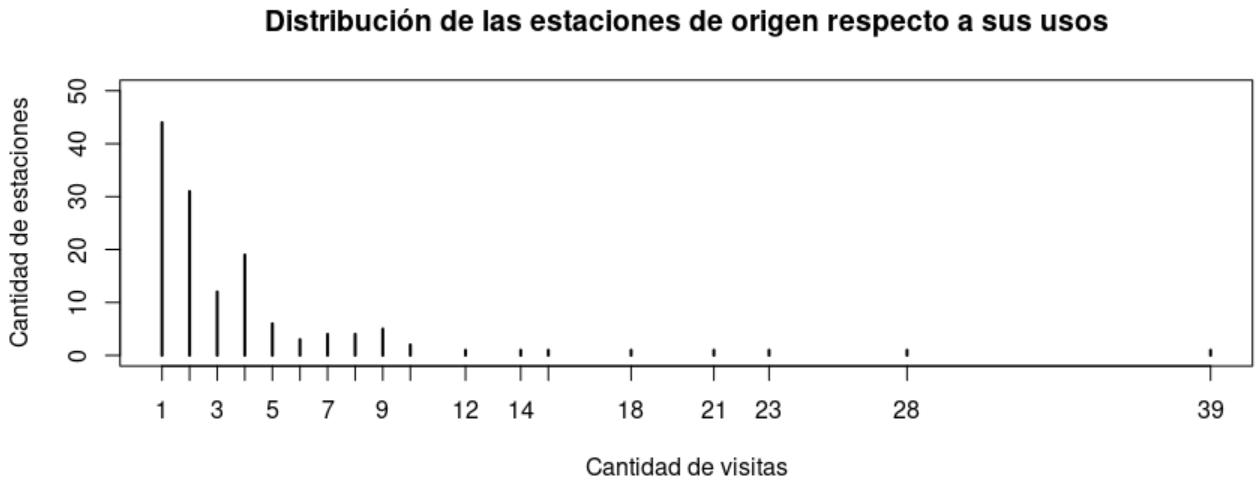

Figure 15: Gráfico de bastones de la variable estación origen

En dicho gráfico se observa que una gran parte de las estaciones solo son usadas para originar una viaje a lo sumo 2 veces. También se puede notar que son muy pocas las estaciones que se usan para originar un viaje más de 5 veces.

Medida de interés:
Media = 4.07.
Desvío = 5.25.
Estación más usada = Avenida Juan de Garay 1050. Fue usada 39 veces.

3.1.7 Estación de destino

Para esta variable se contaron cuántos viajes tuvieron como destino determinada estación y luego se agruparon por ese número.

Tabla de frecuencias de la variable estación de destino:

	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Porcentual	Frecuencia Relativa Porcentual Acumulada
1	46	0.33	33.33	33.33
2	28	0.20	20.29	53.62
3	20	0.14	14.49	68.12
4	10	0.07	7.25	75.36
5	10	0.07	7.25	82.61
6	3	0.02	2.17	84.78
7	4	0.03	2.90	87.68
8	3	0.02	2.17	89.86
9	2	0.01	1.45	91.30
10	1	0.01	0.72	92.03
>10	11	0.08	7.97	100.00
Total	138	1.00	100.00	100.00

Figure 16: Tabla de frecuencias de la variable estación de destino.

Análisis de la segunda fila de la tabla:

- Hay 28 estaciones de destino que fueron visitadas dos veces.
- El 53.62% de las estaciones destino fueron visitadas hasta a lo sumo dos veces.

Gráfico de Bastones:

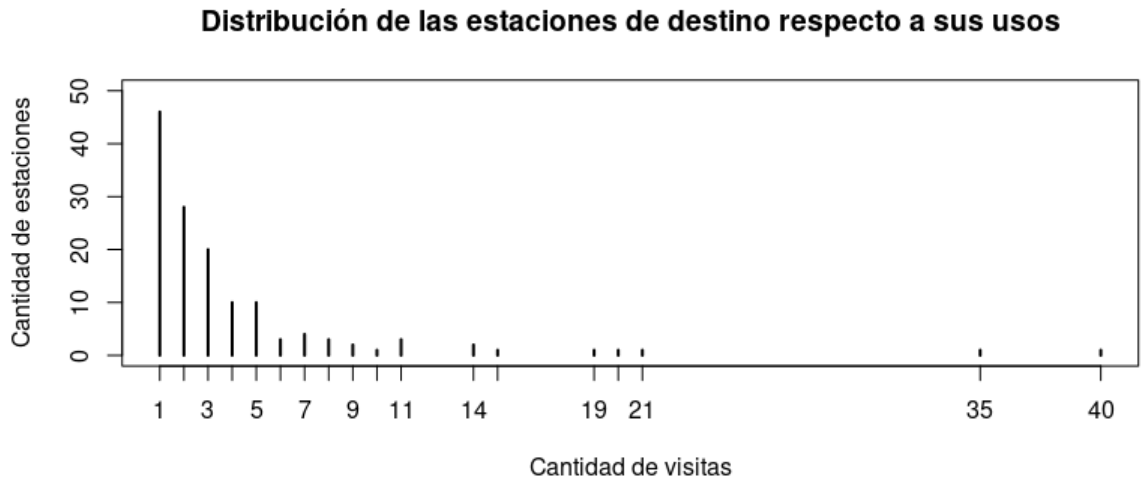

Figure 17: Gráfico de bastones de la variable estación de destino.

En dicho gráfico se observa que una gran parte de las estaciones solo son usadas como destino de un viaje a lo sumo 2 veces. También se puede notar que son muy pocas las estaciones que se como destino de un viaje más de 6 veces.

Medida de interés:
Media = 4.07.
Desvío = 5.56.
Estación más usada = AZOPARDO 700. Fue usada 40 veces.

3.2 Análisis bivariado

Se comentará el análisis de la cantidad de ciclistas en función del género y los días de la semana.

La tabla que se obtuvo fue la siguiente:

	Domingo	Jueves	Lunes	Martes	Miércoles	Sábado	Viernes	Total
Femenino	38	36	53	44	37	27	47	282
Masculino	17	23	32	20	16	24	29	161
Otro	10	12	24	16	20	23	13	118
Total	65	71	109	80	73	74	89	561

Figure 18: Tabla de la cantidad de ciclistas en función del género y los días de la semana.

Análisis de la segunda fila de la tabla:

- Los días Lunes y los días Viernes es cuando hay mayor cantidad de hombres usando el servicio EcoBici.
- Los días Miércoles y los días Domingos es cuando hay un menor uso por parte del género masculino en el servicio de EcoBici.

Gráfico de Barras:

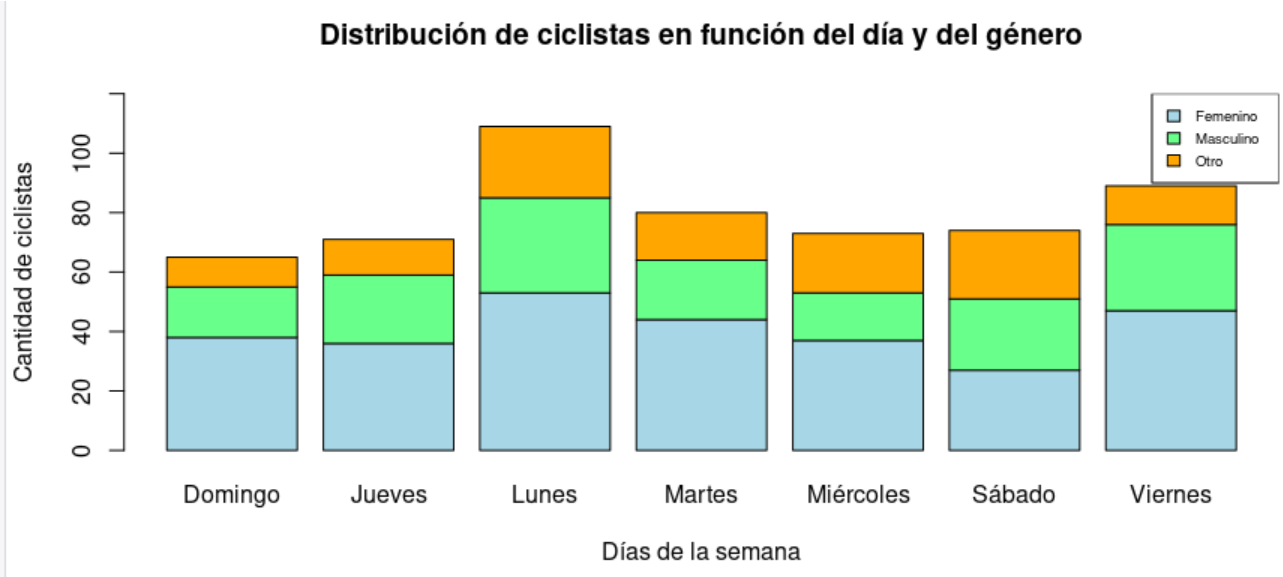

Figure 19: Gráfico de barras de las variables género y día.

En dicho gráfico se observa que los días sábados es el día en donde se presenta una cantidad parecida o cercana de personas de distinto género. Es decir, en los días sábados no hay mucha variación con respecto al género de los ciclistas.
Además, los días lunes es cuando hay más cantidad de hombres y mujeres.