Trabajo Práctico
Estadística Descriptiva

Alumnas:

Cipullo, Inés

Sullivan, Katherine

Universidad Nacional de Rosario

2021

El presente informe tiene como objetivo la exposición de un análisis estadístico descriptivo sobre los datos recopilados del sistema de bicicletas compartidas de la Ciudad de Buenos Aires, EcoBici.

1 Sobre los datos

Los datos utilizados se encontraban divididos en dos unidades de análisis diferentes: una correspondiente a la información sobre los usuarios del sistema en el año 2020, y la otra, a la información sobre los recorridos realizados por los mismos, en el año 2020.

Se cuenta para el siguienete análisis con una muestra aleatoria de 100 usuarios tomados de las observaciones totales registradas por el Ministerio de Desarrollo Urbano y Transporte de la Ciudad de Buenos Aires, disponibles en `https://data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas`.

Todos los datos y gráficos presentados a continuación provienen de esta misma y única fuente.

2 Sobre las variables

Como fue mencionado en la sección anterior, la información se encontraba dividida en dos unidades de análisis. Cada una de ellas presenta diferentes variables que serán el objeto de interés de este informe.

En la primer unidad (referida a información de usuario) se cuenta con tres variables:

- ID de usuario (número de 6 dígitos que identifica un usuario),

- Género de usuario (pudiendo tomar las categorías Femenino, Masculino y Otro), y
- Edad de usuario (representada en años).

En la segunda unidad (referida a información de recorridos) se cuenta con 5 variables:

- Duración del recorrido(representada en segundos),
- Distancia (distancia entre la estación de origen y la de destino, representada en metros),
- Día (día de la semana en el que se realizo el recorrido),
- Dirección de origen (dirección de la estación de EcoBici desde donde se inició el recorrido), y
- Dirección de destino (dirección de la estación de EcoBici desde donde finalizó el recorrido).

3 Análisis univariado

3.1 Género de usuario

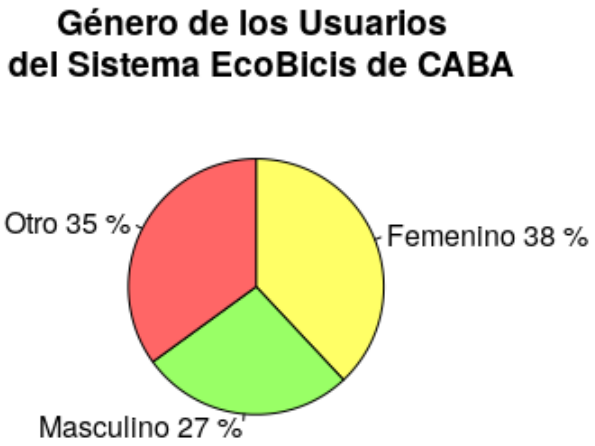
Cabe mencionar antes de proceder al análisis de la variable que la categoría "Otro" es el valor por defecto al ingresar los datos de usuario, por lo tanto resulta posible que usuarios que se identifiquen con cualquiera de las otras dos categorías hayan quedado bajo la categoría "Otro" por simplemente no modificar el valor por defecto.

Para comenzar el análisis, se puede observar la siguiente tabla de frecuencias sobre la variable género de usuario.

Género de los Usuarios del Sistema EcoBici de CABA

Género de usuario	**Frecuencia absoluta**	**Frecuencia relativa**
Femenino	38	0.38
Masculino	27	0.27
Otro	35	0.35
Total	100	1.00

Esta información, dada la condición cualitativa de la variable, se puede exponer en forma de gráfico de sectores. Así se puede visualizar claramente la porción del total que representa cada valor de la variable.



De lo descripto se puede observar que las categorías se encuentran bastante uniformemente divididas, y que la moda es Femenino, es decir, se cuenta con ms usuarios del gnero Feminino que de cualquiera de los otros.

3.2 Edad de usuario

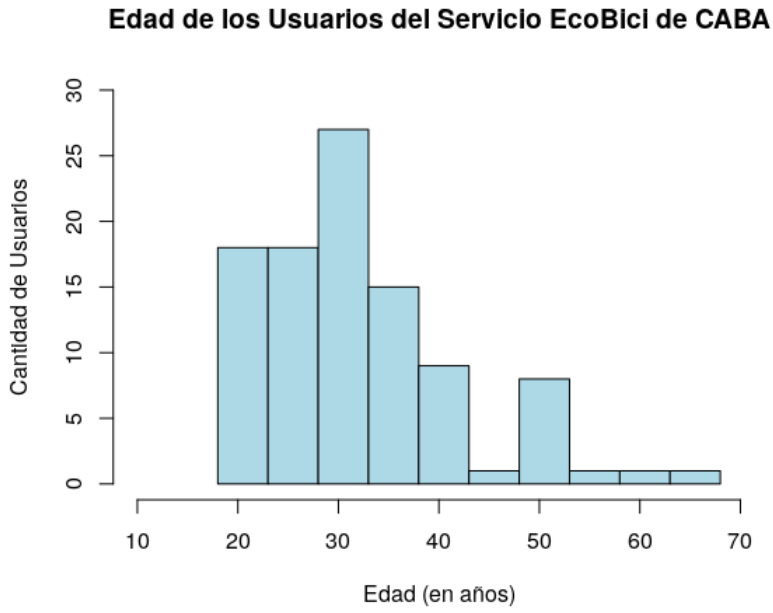
Es importante tener en cuenta que para este análisis univariado se cuenta con un total de 99 usuarios, puesto que se debió excluir de los datos recopilados un usuario cuyo valor de Edad se presentaba como faltante.

Se procedió a la división de la variable en intervalos de 5 años de edad quedando su tabla de frecuencias de la siguiente manera:

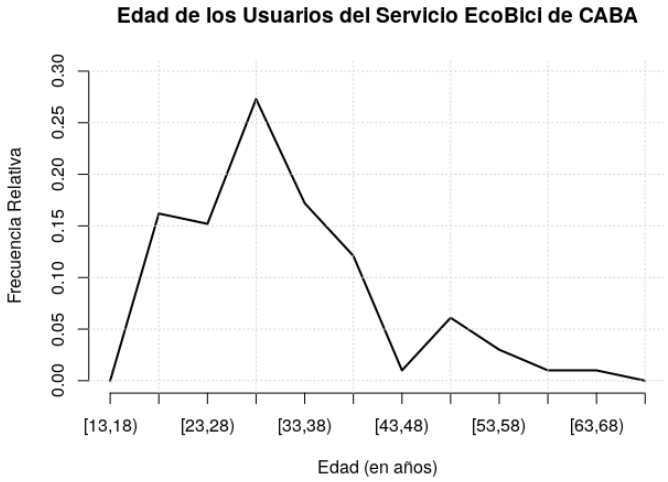
Edad de los Usuarios del Sistema EcoBici de CABA

Edad de usuario	**Frecuencia absoluta**	**Frecuencia relativa**	**Frecuencia absoluta acumulada**	**Frecuencia relativa acumulada**
[18,23)	16	0.16	16	0.16
[23,28)	15	0.15	31	0.31
[28,33)	27	0.27	58	0.58
[33,38)	17	0.17	75	0.75
[38,43)	12	0.12	87	0.88
[43,48)	1	0.01	88	0.89
[48,53)	6	0.06	94	0.95
[53,58)	3	0.03	97	0.98
[58,63)	1	0.01	98	0.99
[63,68)	1	0.01	99	1.00
Total	99	1.00	-	-

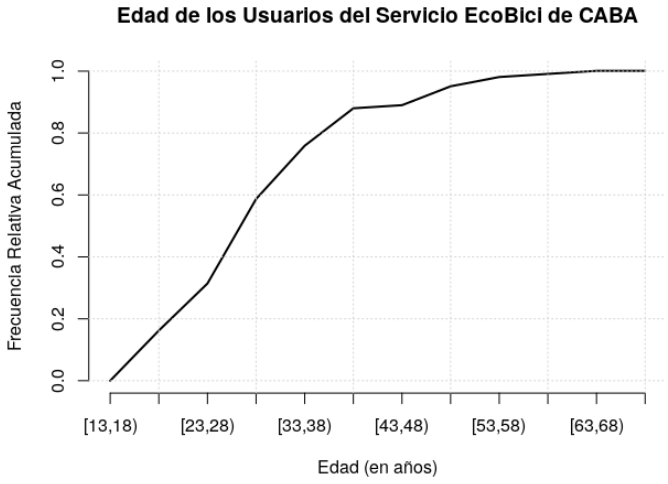
Manteniendo esta separación en intervalos se puede visualizar más cómodamente esta información en un histograma que toma como unidad el intervalo de 5 años, así pudiendo presentar la densidad de las áreas con la cantidad de usuarios.



Acompaando al histograma, también resulta útil la presentación del polígono de frecuencias (arriba) y el polígono acumulativo (abajo).



(a) Polígono de frecuencias



(b) Polígono acumulativo

Por último, dada la condición cuantitativa de la variable edad resulta interesante hablar sobre sus medidas resumen y respectivas medidas de dispersión.

La media es de 32.51 años con un desvío estándar de 9.95 años.

La mediana la marca la edad de 31 años. El primer cuartil, los 25 años y el tercer cuartil, los 37 años. Por lo tanto se cuenta con un rango intercuartil de 12 años.

La edad mínima presentada fue de 19 años y la máxima, de 66 años.

3.3 Día de recorrido

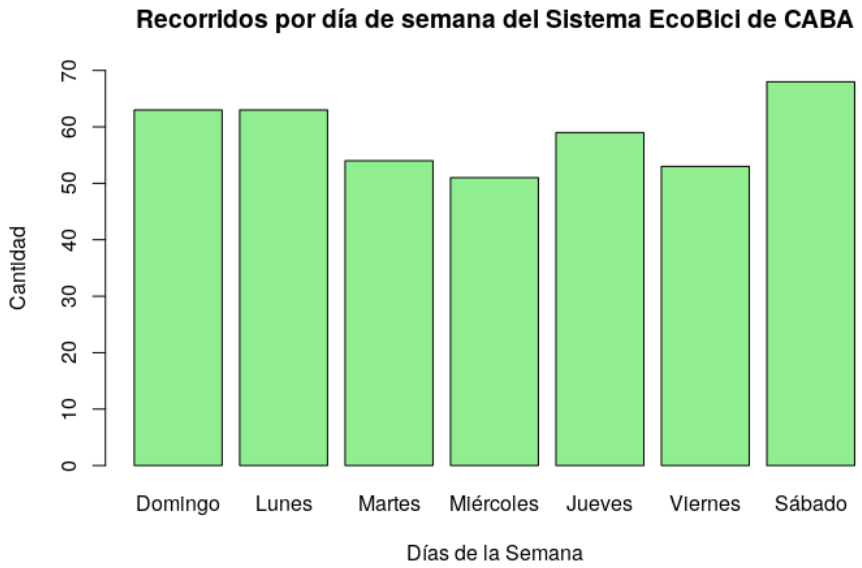
Para dar comienzo al análisis univariado de las variables referidas a los recorridos realizados por los 100 usuarios que comprenden la muestra, resulta pertinente mencionar que se cuenta con un total de 411 recorridos. Este será referido como total para estos análisis.

La variable Día de recorrido cuenta con 7 categorías: Domingo, Lunes, Martes, Miércoles, Jueves, Viernes y Sábado, y su tabla de frecuencia es la que se presenta a continuación.

Recorridos por día de semana del Sistema EcoBici de CABA

Día	**Frecuencia absoluta**	**Frecuencia relativa**
Domingo	63	0.15
Lunes	63	0.15
Martes	54	0.13
Miércoles	51	0.12
Jueves	59	0.14
Viernes	53	0.13
Sábado	68	0.17
Total	411	1.00

Se puede visualizar mejor esta información en el gráfico de barras que aparece a continuación.



De lo anterior resulta simple notar que la moda de la variable es Sábado y que la categoría con menor cantidad de recorridos es Miércoles, aunque, de cualquier manera, las categorías no presentan una gran diferencia entre sus valores.

3.4 Estación de origen de recorrido

Tomando en consideración que se cuenta con 142 estaciones se decidió presentar dentro del informe un cuadro con las 10 estaciones que fueron utilizadas más veces como origen por la muestra de usuarios. Si se desea obtener el cuadro de frecuencias completo puede acceder a él mediante el siguiente enlace `https://data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas`

Entonces, por un lado, recordando que no se llegan al total esperado de 411 recorridos porque solo presentamos las 10 con mayor frecuencia, se presenta la tabla a continuación.

Visualizando esta tabla resulta claro que la moda es la estación ubicada en Ramos Mejia, Av Dr Jose Maria Vargas & Av. Del Libertador.

Sin embargo, por otro lado, teniendo en cuenta la cantidad de estaciones se vio como pertinente el recategorizar la variable con respecto a la cantidad de estaciones utilizadas como origen una x cantidad de veces. Es decir, las categorías nuevas tendrán la forma de un número x que representa la cantidad de recorridos iniciados y su valor asociado será la cantidad de estaciones que hayan sido origen de esa x cantidad de recorridos.

Una vez hecha esta recategorización se hace fácil de reconocer el principio de Paretto que aparece: las categorías con números más bajos son las que agrupan la mayor cantidad de estaciones, lo cual se puede observar claramente en el siguiente gráfico de Paretto.

3.5 Estación de destino de recorrido

Tomando en consideración que se cuenta con 135 estaciones se decidió presentar dentro del informe un cuadro con las 10 estaciones que fueron utilizadas más veces como destino por la muestra de usuarios. Si se desea obtener el cuadro de frecuencias completo puede acceder a él mediante el siguiente enlace `https://data.buenosaires.gob.ar/dataset/estaciones-bicicletas-publicas`

Entonces, por un lado, recordando que no se llegan al total esperado de 411 recorridos porque solo presentamos las 10 con mayor frecuencia, se presenta la tabla a continuación.

Visualizando esta tabla resulta claro que la moda es la estación ubicada en Lavalle & Bouchard.

Sin embargo, por otro lado, teniendo en cuenta la cantidad de estaciones y lo realizado con la variable anterior se vio como pertinente el recategorizar

la variable con respecto a la cantidad de estaciones utilizadas como origen una x cantidad de veces. Es decir, las categorías nuevas tendrán la forma de un número x que representa la cantidad de recorridos iniciados y su valor asociado será la cantidad de estaciones que hayan sido origen de esa x cantidad de recorridos.

Otra vez, ya hecha esta recategorización se hace fácil de reconocer el principio de Paretto que aparece: las categorías con números más bajos son las que agrupan la mayor cantidad de estaciones, lo cual se puede observar en el siguiente gráfico de Paretto.