

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 siguiente 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 realizó 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.

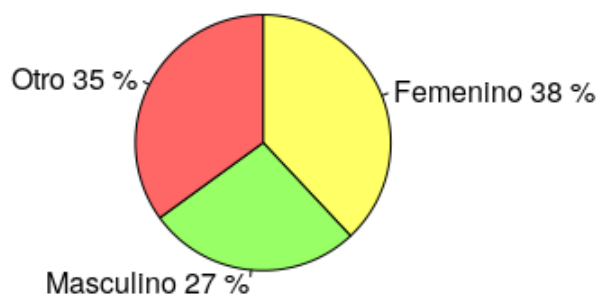
Para comenzar el análisis, se puede observar la siguiente tabla de frecuencias sobre la variable género de usuario.

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.

Género de los Usuarios del Sistema EcoBicis de CABA



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 Femenino que de cualquiera de los otros.

3.2 Edad de usuario

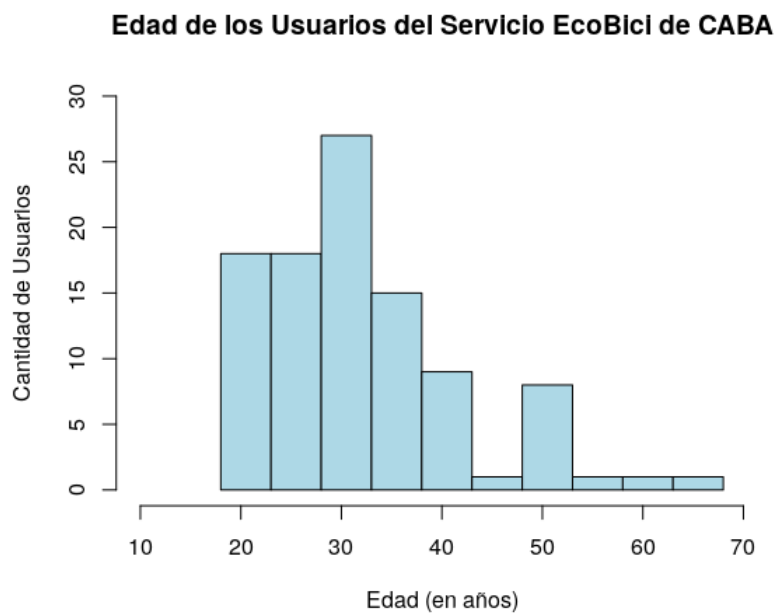
Es importante tener en cuenta que para este análisis univariado se cuenta con un total de 99 usuarios, puesto que se debió excluir de los datos recopilados un usuario cuyo valor de Edad se presentaba como faltante.

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

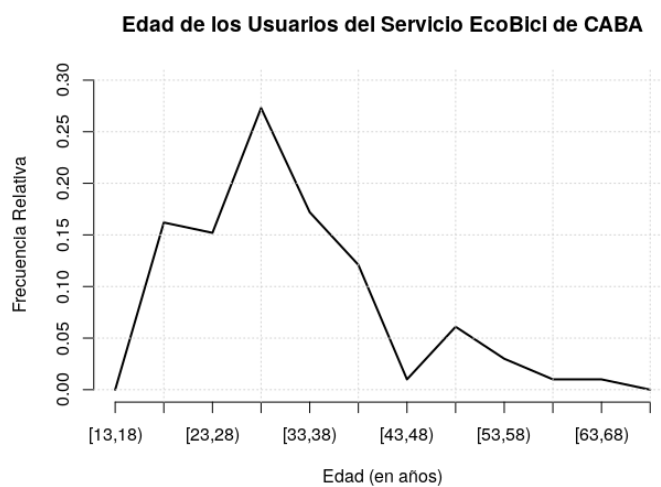
Edad de los Usuarios del Sistema EcoBici de CABA

| Edad de usuario | Frecuencia absoluta | Frecuencia relativa | Frecuencia absoluta acumulada | Frecuencia relativa acumulada |
|-----------------|---------------------|---------------------|-------------------------------|-------------------------------|
| [18,23) | 16 | 0.1616 | 16 | 0.1616 |
| [23,28) | 15 | 0.1515 | 31 | 0.3131 |
| [28,33) | 27 | 0.2727 | 58 | 0.5858 |
| [33,38) | 17 | 0.1717 | 75 | 0.7575 |
| [38,43) | 12 | 0.1212 | 87 | 0.8888 |
| [43,48) | 1 | 0.0101 | 88 | 0.8989 |
| [48,53) | 6 | 0.0606 | 94 | 0.9494 |
| [53,58) | 3 | 0.0303 | 97 | 0.9899 |
| [58,63) | 1 | 0.0101 | 98 | 0.9999 |
| [63,68) | 1 | 0.0101 | 99 | 1.0000 |
| Total | 99 | 1.00 | - | - |

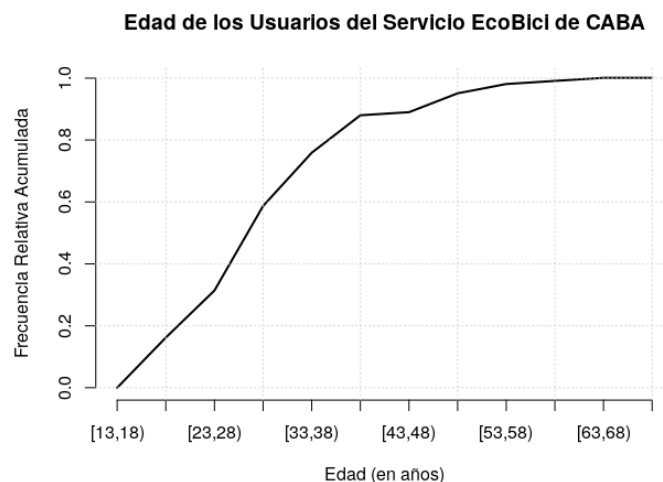
Manteniendo esta separación en intervalos se puede visualizar más cómodamente esta información en un histograma que toma como unidad el intervalo de 5 años, así pudiendo presentar la densidad de las áreas con la cantidad de usuarios.



Acompañando 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

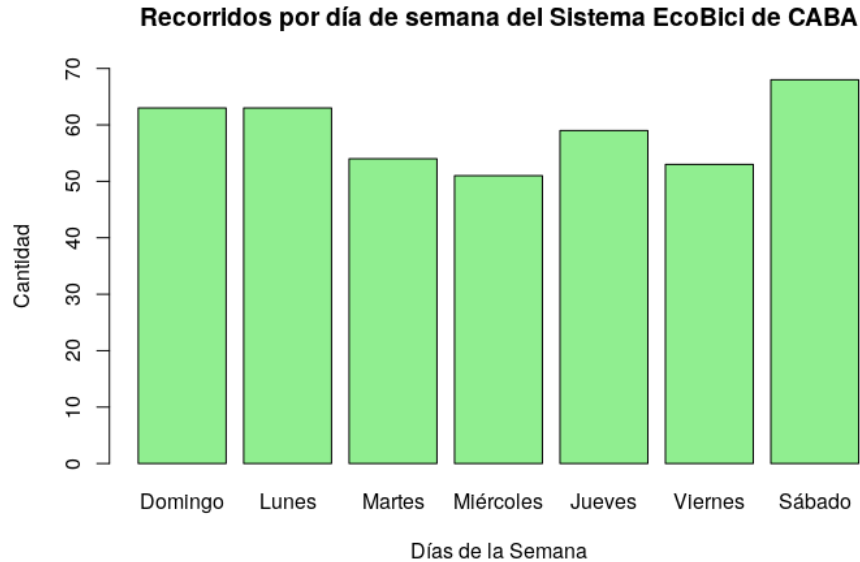
Para dar comienzo al análisis univariado de las variables referidas a los recorridos realizados por los 100 usuarios que comprenden la muestra, resulta pertinente mencionar que se cuenta con un total de 411 recorridos. Este será referido como total para estos análisis.

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 | 0.1533 |
| Lunes | 63 | 0.1533 |
| Martes | 54 | 0.1314 |
| Miércoles | 51 | 0.1241 |
| Jueves | 59 | 0.1436 |
| Viernes | 53 | 0.1290 |
| Sábado | 68 | 0.1653 |
| 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.

**10 estaciones de EcoBicis de CABA
más frecuentadas como origen**

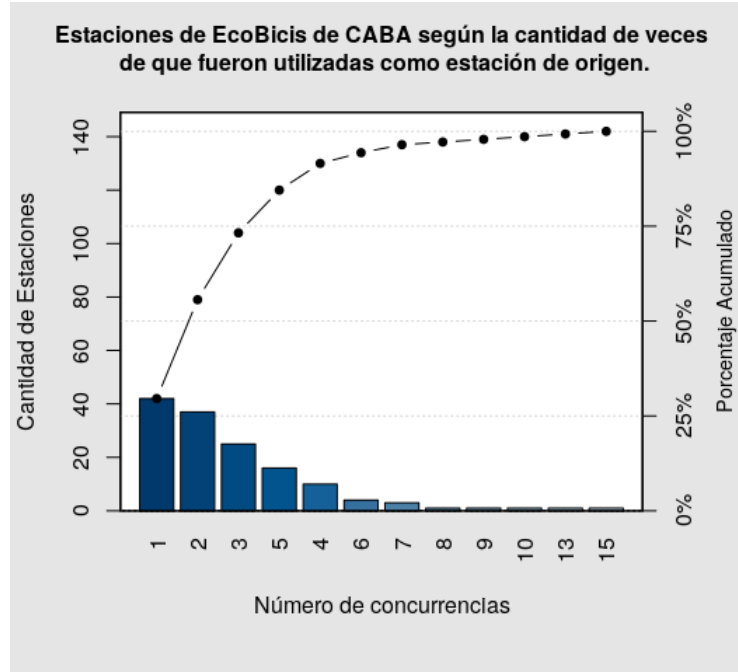
| Estación | Frecuencia absoluta | Frecuencia relativa |
|---|--------------------------------|--------------------------------|
| Ramos Mejia, Av Dr Jose Maria Vargas & Av. Del Libertador | 15 | 0.0365 |
| 2292 Montaeses | 13 | 0.0316 |
| 3912 Humahuaca | 10 | 0.0243 |
| 300 Almafuerce Av. & Los Patos | 9 | 0.0219 |
| 441 Bulnes & Peron, Juan Domingo, Tte. General | 8 | 0.0195 |
| 1785 Espinosa | 7 | 0.0170 |
| 3084 Agrelo | 7 | 0.0170 |
| Cordoba 6599 | 7 | 0.0170 |
| Av. Del Libertador, 3260 | 6 | 0.0146 |
| Lavalle & Acua De Figueroa, Francisco | 6 | 0.0146 |

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 Pareto 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 Pareto.



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.

**10 estaciones de EcoBicis de CABA
más frecuentadas como destino**

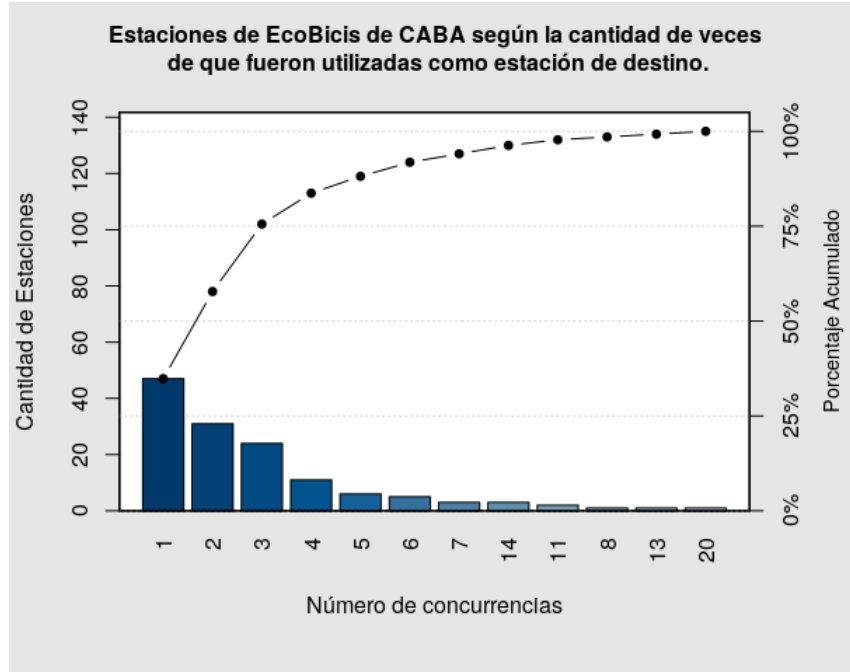
| Estación | Frecuencia absoluta | Frecuencia relativa |
|---|--------------------------------|--------------------------------|
| Lavalle & Bouchard | 20 | 0.0487 |
| 441 Bulnes & Peron, Juan Domingo, Tte. General | 14 | 0.0341 |
| Amenabar y Mendoza | 14 | 0.0341 |
| Quintino Bocayuva y Don Bosco | 14 | 0.0341 |
| Cevallos, Virrey & Yrigoyen, Hipolito Av. | 13 | 0.0316 |
| Culpina 121 | 11 | 0.0268 |
| 1355 San Martin Av. | 11 | 0.0268 |
| Av. Patricias Argentinas & Estivao | 8 | 0.0195 |
| 3084 Agrelo | 7 | 0.0170 |
| 3817 Trafal | 7 | 0.0170 |

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 Pareto 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 Pareto.



3.6 Distancia de recorrido

Antes de proceder con el análisis resulta importante notar el cambio en la unidad de medida de la distancia respecto a la fuente de los datos. En la fuente las distancias se representan en metros, mientras que en el presente informe se representan en kilómetros.

Además, dado que por la continuidad de la variable existen una gran cantidad de valores posibles para que tome la variable, se procede a agrupar las categorías de la variable en intervalos de un kilómetro.

Su tabla de frecuencias queda como sigue:

**Distancia de Recorridos en km
del Sistema EcoBici de CABA**

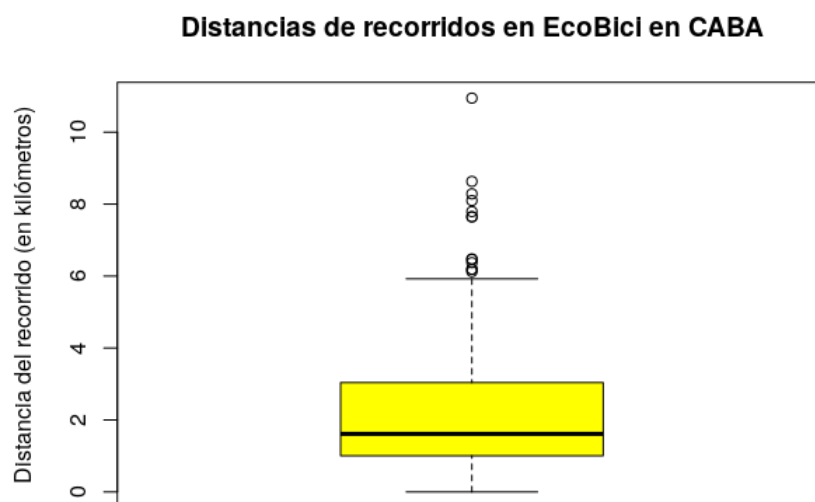
| Distancia de recorrido | Frecuencia absoluta | Frecuencia relativa | Frecuencia absoluta acumulada | Frecuencia relativa acumulada |
|-------------------------------|----------------------------|----------------------------|--------------------------------------|--------------------------------------|
| [0,1) | 103 | 0.2506 | 103 | 0.2506 |
| [1,2) | 136 | 0.3309 | 239 | 0.5815 |
| [2,3) | 66 | 0.1606 | 305 | 0.7421 |
| [3,4) | 55 | 0.1338 | 360 | 0.8759 |
| [4,5) | 23 | 0.0560 | 383 | 0.9319 |
| [5,6) | 15 | 0.0365 | 398 | 0.9684 |
| [6,7) | 6 | 0.0146 | 404 | 0.9830 |
| [7,8) | 3 | 0.0073 | 407 | 0.9903 |
| [8,9) | 3 | 0.0073 | 410 | 0.9976 |
| [9,10) | 0 | 0.0000 | 410 | 0.9976 |
| [10,11) | 1 | 0.0024 | 411 | 1.0000 |
| Total | 411 | 1.00 | - | - |

Resulta interesante presentar las medidas resúmenes de la variable y sus respectivas medidas de dispersión.

- La media es de 2.0880 km con un desvío estándar de 1.7148 km.
- El primer cuartil toma el valor de 1.0060 km, mientras que el segundo cuartil (o mediana) toma el valor de 1.6130 km y el tercer cuartil, de 3.0390 km.
- Se cuenta con un rango intercuartil de 2.0330 km.
- El valor máximo que toma la variable es de 10.9460 km y el mínimo es de 0 km (un distancia de recorrido es de 0 km si se devuelve la bicicleta a la misma estación de donde se la sacó).

- El intervalo de distancia que cuenta con más recorridos es el $[1,2)$.

Esta información de la variable se puede ver clara y resumida en el siguiente boxplot:



3.7 Duración de recorrido

Al igual que con la distancia, previo al análisis de esta variable se debe aclarar que se modificó la unidad de medición de la variable, pasando de segundos a minutos.

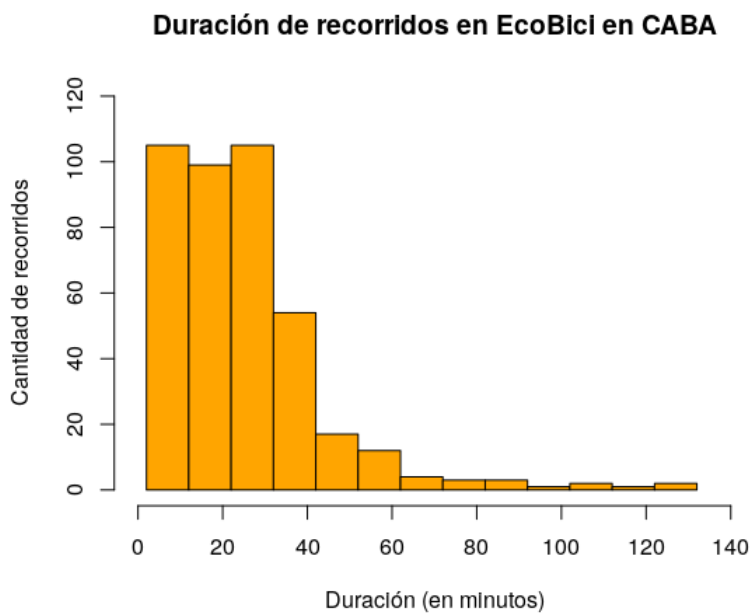
A su vez, también por la continuidad de la variable, se decidió dividirla en intervalos de 10 minutos hasta llegar a los 132 minutos, donde se agrupó a 3 valores muy extremos dentro de un solo intervalo.

Duración de Recorridos en minutos del Sistema EcoBici de CABA

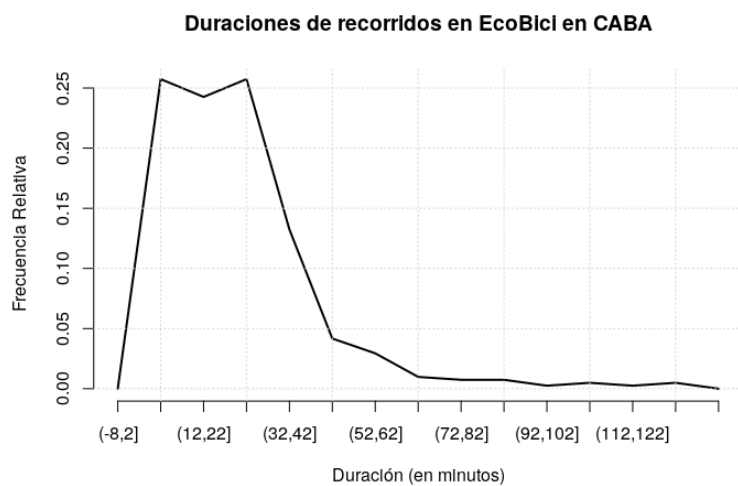
| Duración de recorrido | Frecuencia absoluta | Frecuencia relativa | Frecuencia absoluta acumulada | Frecuencia relativa acumulada |
|-----------------------|---------------------|---------------------|-------------------------------|-------------------------------|
| (2,12] | 105 | 0.2555 | 105 | 0.2555 |
| (12,22] | 99 | 0.2409 | 204 | 0.4964 |
| (22,32] | 105 | 0.2555 | 309 | 0.7518 |
| (32,42] | 54 | 0.1314 | 363 | 0.8832 |
| (42,52] | 17 | 0.0414 | 380 | 0.9246 |
| (52,62] | 12 | 0.0292 | 392 | 0.9538 |
| (62,72] | 4 | 0.0097 | 396 | 0.9635 |
| (72,82] | 3 | 0.0073 | 399 | 0.9708 |
| (82,92] | 3 | 0.0073 | 402 | 0.9781 |
| (92,102] | 1 | 0.0024 | 403 | 0.9805 |
| (102,112] | 2 | 0.0049 | 405 | 0.9854 |
| (112,122] | 1 | 0.0024 | 406 | 0.9878 |
| (122,132] | 2 | 0.0049 | 408 | 0.9927 |
| (132,485] | 3 | 0.0073 | 411 | 1.0000 |
| Total | 411 | 1.00 | - | - |

Se puede ver esta información de una forma más clara y ordenada si es presentada en un histograma. Para que este cumpla su propósito de verse así fue necesario excluir del mismo al último intervalo que presenta solo 3 recorridos con duraciones muy extremas.

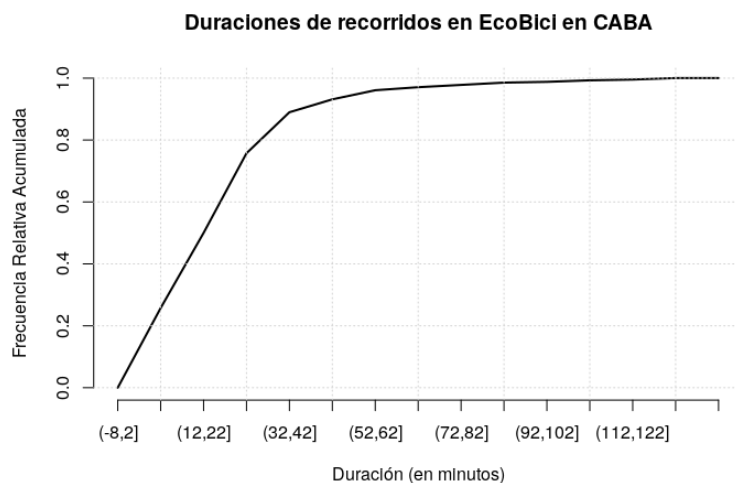
Entonces, el nuevo total de recorridos con el que se trabaja es de 408 y este sería el histograma que representa la duración de los recorridos (menores a 132 minutos) tomando como unidad los 10 minutos para que así la densidad representa la cantidad de recorridos.



Acompañando al histograma y manteniendo este total de 408 recorridos se presentan el polígono de frecuencias y el polígono acumulativo.



(a) Polígono de frecuencias



(b) Polígono acumulativo

Resulta útil, también, el hacer un análisis de las medidas resumen de la variable y sus respectivas medidas de dispersión. Para su análisis volvemos a considerar el total de 411 recorridos.

La media es de 27.3000 minutos con un desvío estándar de 32.5626 minutos.

En relación a los cuartiles se puede observar lo siguiente:

- Primer cuartil: 11.67 minutos.
- Segundo cuartil o mediana: 22.43 minutos.
- Tercer cuartil: 31.91 minutos.
- Rango intercuartil: 20.24 minutos.

Y, por último, el recorrido con mayor duración fue de 484.55 minutos y el de menor duración fue de 2.20 minutos.

4 Análisis bivariado

4.1 Duración de recorrido respecto a Día de recorrido