

Estudio del sistema público de bicicletas de la CDMX

José Ángel Rodríguez

3/8/2022

Pregunta 1

Primero realizamos un breve análisis exploratorio de los datos que tenemos:

```
##          genero        edad        bici      resumen_ecobici
## 1      Length:1103273   Min.   :17.00   Min.   : 775   Min.   : 1.0
## 2      Class  :character  1st Qu.:28.00  1st Qu.: 7892  1st Qu.: 70.0
## 3      Mode   :character  Median :33.00  Median : 9456  Median :159.0
## 4      Mean    :35.96   Mean    : 9425   Mean    :187.3
## 5      3rd Qu.:41.00   3rd Qu.:11212  3rd Qu.:290.0
## 6      Max.    :86.00   Max.    :15339   Max.    :3002.0
## 7
## 8      estacion_arribo fecha_arribo   fecha_retiro hora_arribo
## 9      Min.   : 1.0   Min.   :2021-08-01   Min.   :2021-05-03   Min.   : 0.00
## 10     1st Qu.: 68.0  1st Qu.:2021-08-25  1st Qu.:2021-08-25  1st Qu.:11.00
## 11     Median :154.0  Median :2021-09-20  Median :2021-09-20  Median :14.00
## 12     Mean   :184.5  Mean   :2021-09-17  Mean   :2021-09-17  Mean   :14.21
## 13     3rd Qu.:285.0  3rd Qu.:2021-10-12 3rd Qu.:2021-10-12  3rd Qu.:18.00
## 14     Max.   :480.0  Max.   :2021-10-31  Max.   :2021-10-31  Max.   :23.00
## 15     hora_retiro dia_arribo   dia_retiro mes_arribo mes_retiro
## 16     Min.   : 0   Min.   : 1.00   Min.   : 1.00   Min.   : 8.000   Min.   : 5.000
## 17     1st Qu.:10  1st Qu.: 9.00  1st Qu.: 9.00  1st Qu.: 8.000  1st Qu.: 8.000
## 18     Median :14  Median :16.00  Median :16.00  Median : 9.000  Median : 9.000
## 19     Mean   :14  Mean   :16.25  Mean   :16.25  Mean   : 9.065  Mean   : 9.064
## 20     3rd Qu.:18  3rd Qu.:24.00 3rd Qu.:24.00  3rd Qu.:10.000 3rd Qu.:10.000
## 21     Max.   :23  Max.   :31.00  Max.   :31.00  Max.   :10.000  Max.   :10.000
## 22
## 23
## 24
## 25
## 26
## 27
## 28
##          duracion misma_estacion
## 1      Min.   : 2.00   Min.   :0.00000
## 2      1st Qu.: 7.00  1st Qu.:0.00000
## 3      Median :12.00  Median :0.00000
## 4      Mean   :72.18  Mean   :0.05605
## 5      3rd Qu.:21.00  3rd Qu.:0.00000
## 6      Max.   :169873.00  Max.   :1.00000
```

De aquí podemos ver dos cosas extrañas. Primero, el valor máximo de la variable ‘duracion’ que corresponde a aproximadamente 117 días y además tenemos un promedio de 72 minutos por viaje cuando el 75% de las observaciones está abajo de 21 minutos. Como segundo punto tenemos que el valor máximo de la variable ‘estacion_retiro’ es 3002, cuando se supone que solamente hay 480 estaciones. Sobre este segundo punto solamente se tienen dos observaciones:

```
##      genero edad  bici estacion_retiro estacion_arribo fecha_arribo fecha_retiro
## 1      M   26  9592       3002           443 2021-08-12 2021-08-12
## 2      M   26 11740       3002           113 2021-10-14 2021-10-14
##      hora_arribo hora_retiro dia_arribo dia_retiro mes_arribo mes_retiro duracion
```

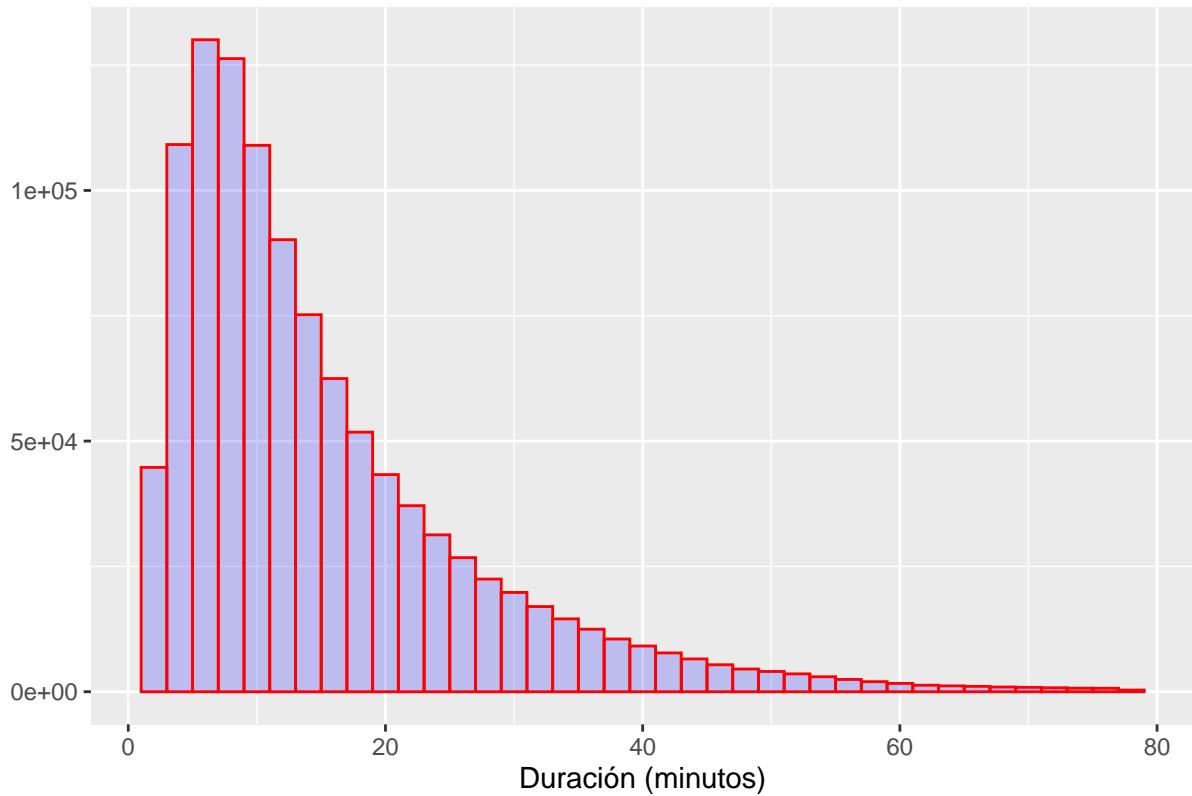
```

## 1      13      8      12      12      8      8      280
## 2      14      8      14      14     10      10      390
##   misma_estacion    z_scores
## 1          0 0.09093706
## 2          0 0.13907141

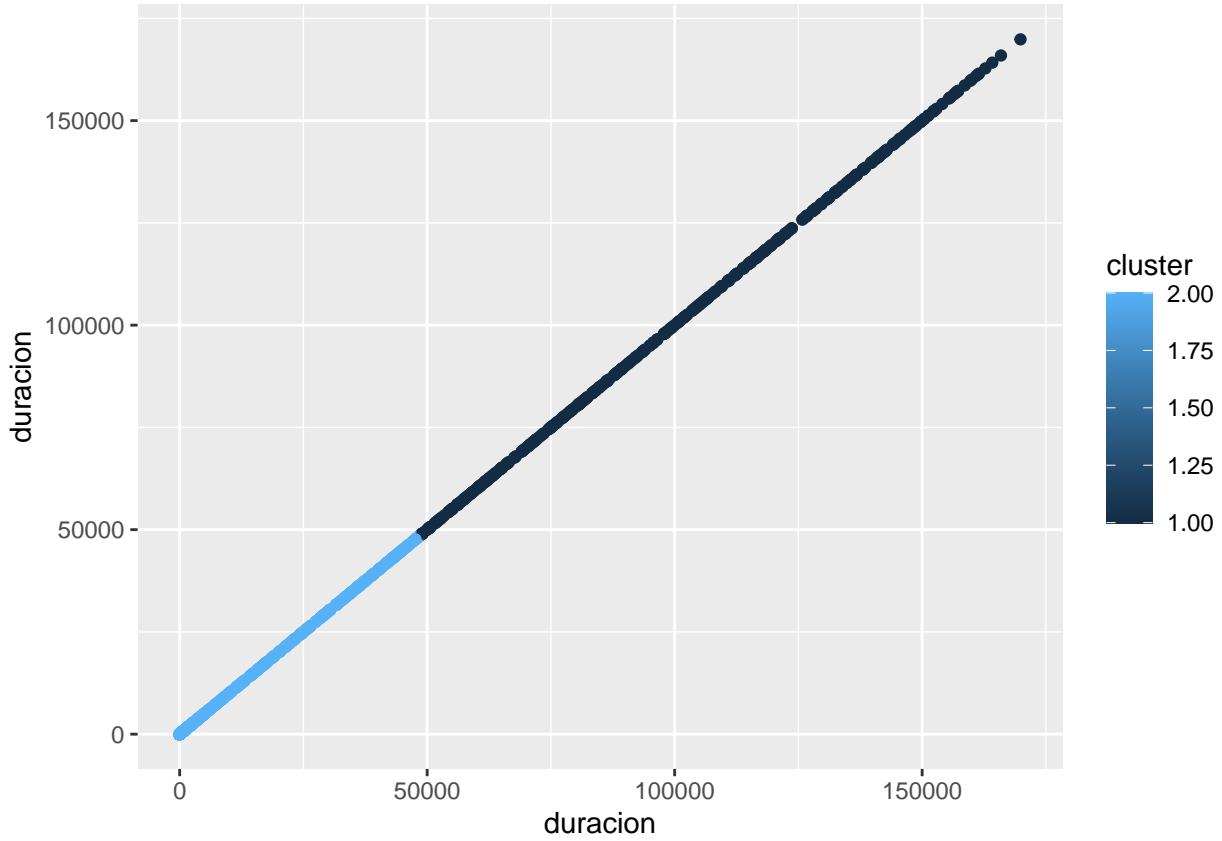
```

y como claramente corresponden a datos que están mal registrados podemos ignorarlos para el resto del estudio. Sin embargo, para el primer punto no es tan fácil decidir que hacer. Por el sistema de tarifas y multas que tiene ECOBICI podríamos suponer que cualquier observación que tiene una duración de uso de más de 3 horas corresponde a un accidente, extravío de la bicicleta o un error del sistema en el registro, que en cualquier caso podemos considerar como outliers, pero vamos a revisarlo con un poco de más cuidado. Primero vemos como se ve la distribución de la duración para observaciones en los que esta variable es menor a 79 (99% de las observaciones)

Histograma de Duración de Viaje



Como la duración está tan sesgada a la izquierda algunas de las técnicas tradicionales para la detección de outliers como usar el rango intercuartílico no son tan apropiadas, ya que tendríamos un rango intercuartílico muy pequeño y perderíamos muchas de las observaciones con duración mayor a 10 minutos. Algo que podría dar mejores resultados es usar técnicas de proximidad de los datos como agrupar en clusters usando el algoritmo k-means. Como sabemos que hay mucha mayor densidad para observaciones con duraciones pequeñas podemos tomar k=2 para ver hasta donde se consideran observaciones como parte de las de duración corta. Alternativamente se podría usar el ‘elbow-method’ para encontrar un valor óptimo para k, pero para el uso del algoritmo que tenemos tomar k=2 funciona. Como solamente nos interesa agrupar los datos de acuerdo a la duración solamente consideraremos esta variable.



Vemos que los clusters se separan en aproximadamente 50,000 minutos que corresponde a 34 días, que es un buen primer paso y reafirma lo que se creía pero nos gustaría tener un límite un poco más estricto. Para esto usamos el método de ‘z_scores’ que indica que tanto (en términos de la desviación estandar) se alejan las observaciones de la media y por lo tanto es una mejor alternativa para datos que están muy lejos de ser simétricos (como que los que tenemos) que usar el rango intercuartílico. Después de quitar las observaciones con valores nos quedamos con los siguientes datos (más del 99.9% de los datos originales)

```
##                                     resumen_ecobici_sin_outliers
## 1             genero          edad        bici estacion_retiro
## 2 Length:1083157   Min.   :17.00   Min.   : 775  Min.   : 1.0
## 3 Class  :character 1st Qu.:28.00  1st Qu.: 7891 1st Qu.: 69.0
## 4 Mode   :character Median :33.00  Median : 9452  Median :156.0
## 5                   Mean   :35.98  Mean   : 9420  Mean   :184.4
## 6                   3rd Qu.:41.00  3rd Qu.:11207 3rd Qu.:286.0
## 7                   Max.   :86.00  Max.   :15339  Max.   :474.0
## 8 estacion_arribo fecha_arribo    fecha_retiro      hora_arribo
## 9 Min.   : 1.0  Min.   :2021-08-01  Min.   :2021-07-31  Min.   : 0.00
## 10 1st Qu.: 67.0 1st Qu.:2021-08-25  1st Qu.:2021-08-25  1st Qu.:11.00
## 11 Median :152.0  Median :2021-09-20  Median :2021-09-20  Median :14.00
## 12 Mean   :181.6  Mean   :2021-09-17  Mean   :2021-09-17  Mean   :14.22
## 13 3rd Qu.:281.0 3rd Qu.:2021-10-12 3rd Qu.:2021-10-12 3rd Qu.:18.00
## 14 Max.   :474.0  Max.   :2021-10-31  Max.   :2021-10-31  Max.   :23.00
## 15 hora_retiro dia_arribo    dia_retiro      mes_arribo      mes_retiro
## 16 Min.   : 0   Min.   : 1.00   Min.   : 1.00   Min.   : 8.000  Min.   : 7.000
## 17 1st Qu.:10  1st Qu.: 9.00  1st Qu.: 9.00  1st Qu.: 8.000  1st Qu.: 8.000
## 18 Median :14  Median :16.00  Median :16.00  Median : 9.000  Median : 9.000
## 19 Mean   :14  Mean   :16.25  Mean   :16.25  Mean   : 9.064  Mean   : 9.064
```

```

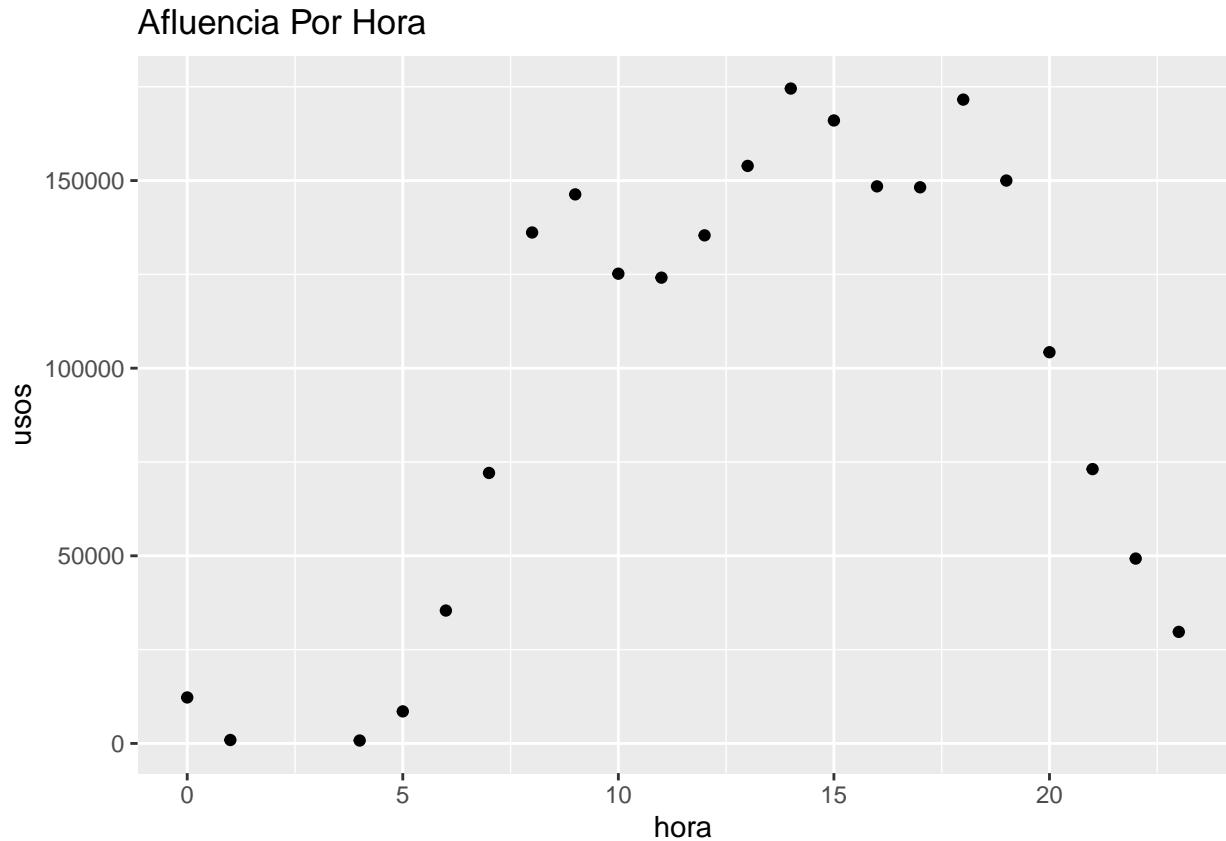
## 20 3rd Qu.:18   3rd Qu.:24.00   3rd Qu.:24.00   3rd Qu.:10.000   3rd Qu.:10.000
## 21 Max.     :23   Max.     :31.00   Max.     :31.00   Max.     :10.000   Max.     :10.000
## 22
## 23           duracion      misma_estacion      z_scores
## 24           Min.    : 2.00    Min.    :0.0000000  Min.    :0.00000806
## 25           1st Qu.: 7.00    1st Qu.:0.0000000  1st Qu.:0.0228350
## 26           Median  :12.00    Median  :0.0000000  Median  :0.0263357
## 27           Mean    :17.27    Mean    :0.05646   Mean    :0.0250041
## 28           3rd Qu.:21.00    3rd Qu.:0.0000000  3rd Qu.:0.0285236
## 29           Max.    :1213.00   Max.    :1.0000000  Max.    :0.4992038

```

Aun hay observaciones con valores de duración alta, que podemos ver por el máximo de esta variable. Esto se tiene que tomar en cuenta más adelante ya que la media de este valor puede no ser tan representativo como la mediana. Ahora vamos a la primera pregunta.

La mayor afluencia se tiene a las 18, 14, 15 horas. Esto se encontró agrupando la información por hora para hacer un conteo de el número de retiros y arribos que se tienen y se puede ver en la siguiente gráfica.

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Las 10 estaciones más ocupadas (considerando tanto número de retiros como número de arribos) son las siguientes estaciones:

```

## # A tibble: 10 x 4
##       estacion numero_de_arribos numero_de_retiros usos
##       <int>            <int>            <int> <int>
## 1          1              8722             8374 17096
## 2          2              8693             8320 17013
## 3          3              7541             8633 16174

```

##	4	27	7885	7798	15683
##	5	64	8011	7546	15557
##	6	182	7098	7101	14199
##	7	36	6889	6618	13507
##	8	38	6614	6209	12823
##	9	266	7642	4626	12268
##	10	136	6253	5982	12235

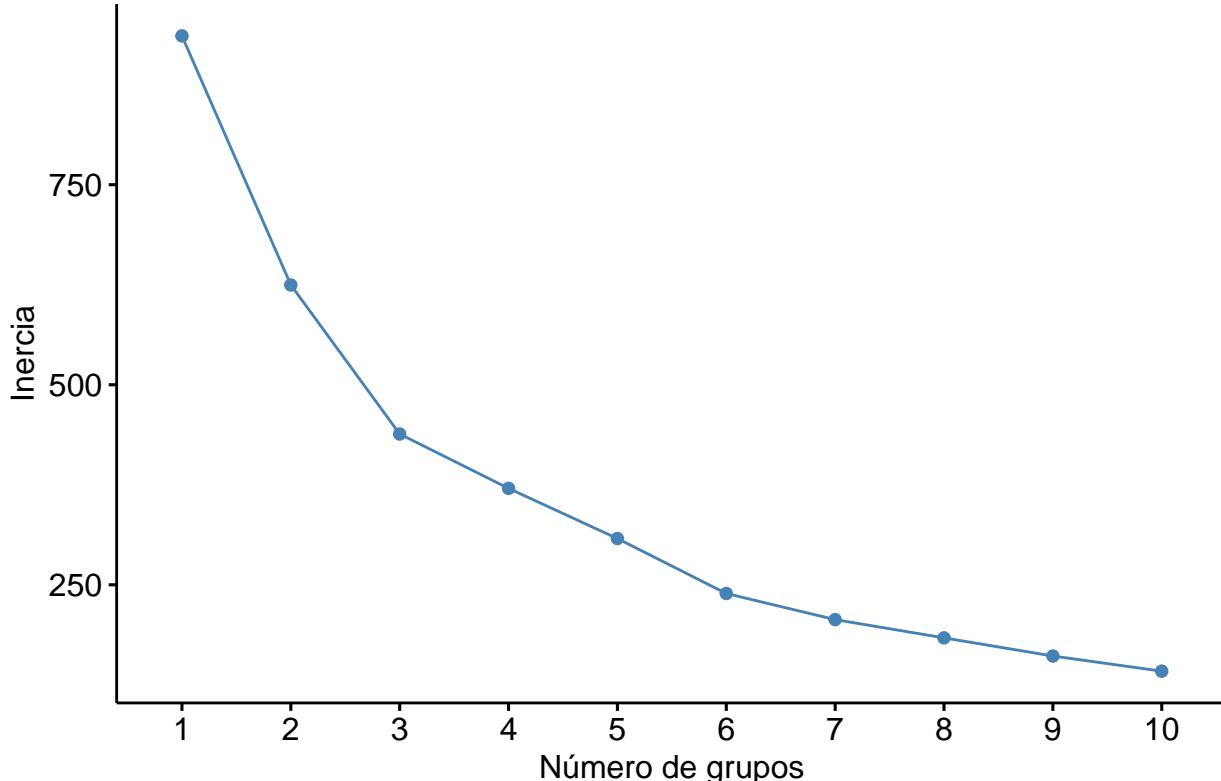
Como el mínimo de edad es 17 años podemos suponer que la mayoría de los usuarios usan las bicicletas para transportarse entre su casa y trabajo, su casa y universidad o preparatoria o para ir del trabajo a comer. A esto se puede deber que las horas de mayor uso sean las que se tienen ya que son horas que corresponden a horas de salida del trabajo, horas de ir a comer y tal vez horas de salida de clases.

Pregunta 2

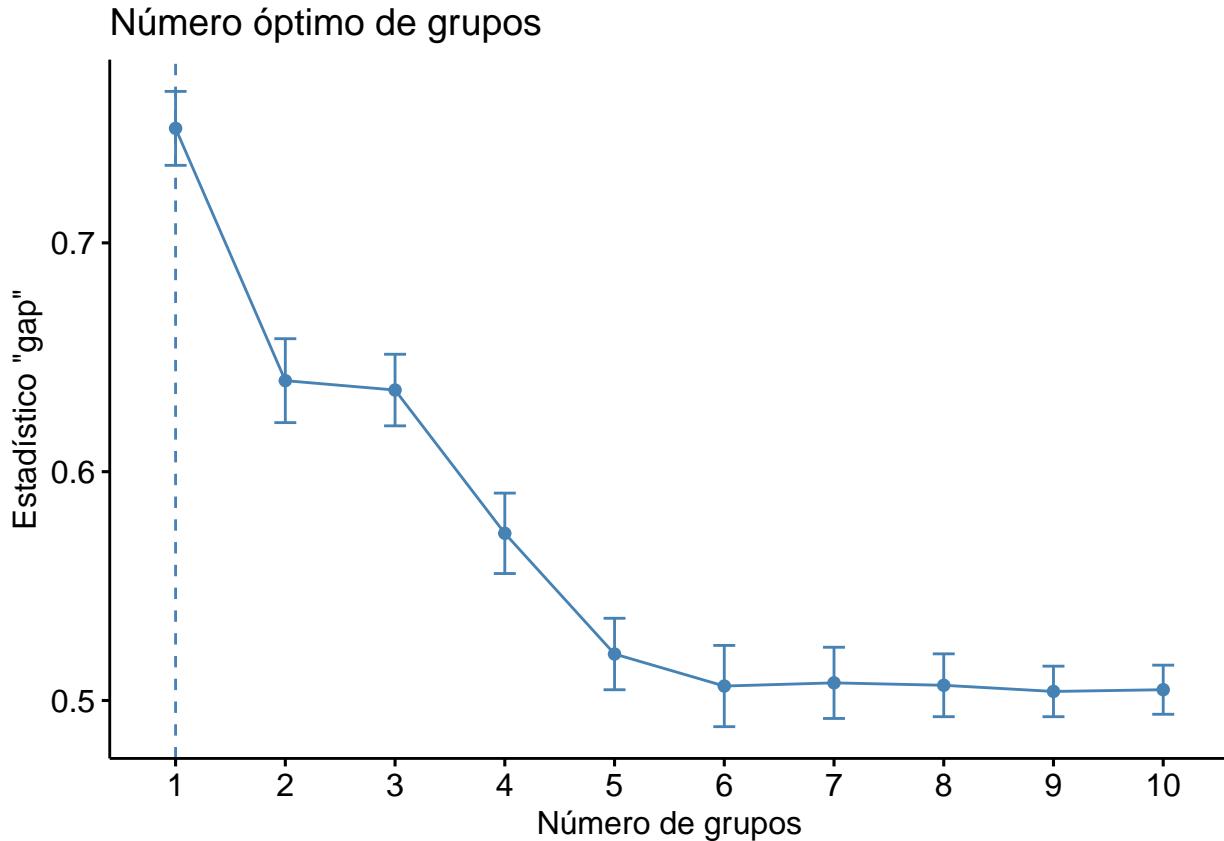
Ahora vamos a analizar el comportamiento de la edad de los usuarios contra la hora a la que terminan los viajes a nivel estación para poder encontrar algo acerca de las estaciones y sus usuarios. Como vamos a trabajar esto a nivel estación es necesario agrupar la información y considerar algún estadístico que represente la información de cada variable. Usamos el promedio en lugar de la moda o mediana ya que aunque queremos estudiar el comportamiento de una estación y para eso tendría sentido considerar moda para tener la información que representa a los usos de bicicletas típicos de esa estación, eso llevaría a una menor variación en los datos y sería más difícil para el método encontrar información relevante.

El método de aprendizaje no supervisado que vamos a usar es el algoritmo ‘k means clustering’ que funciona muy bien para encontrar grupos con respecto a 2 variables. Para escoger el número de grupos primero comparamos la inercia (suma de distancias de observaciones en cada grupo) con el número de grupos y buscamos el punto en que el decremento en la inercia empieza a ser lineal.

Inercia por número de grupos

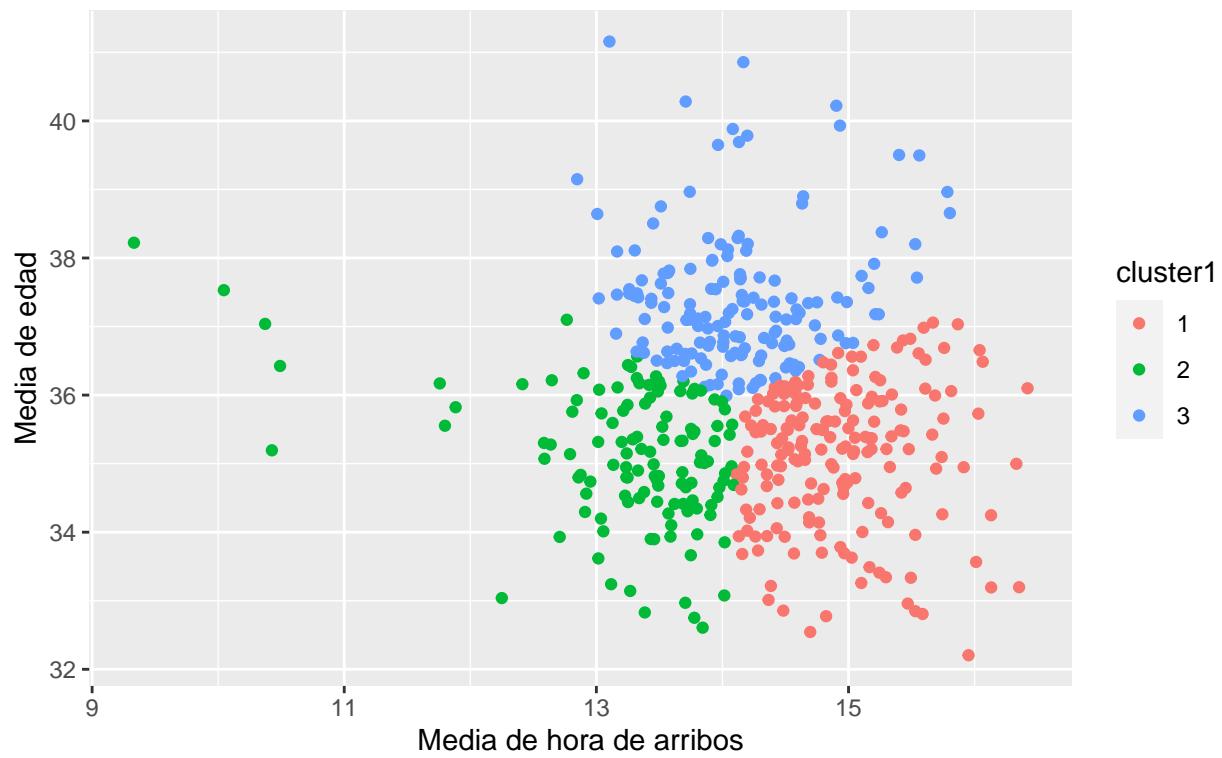


Vemos que ese comportamiento se manifiesta primero en 3 grupos. Por lo que buscamos nos gustaría tener más grupos para poder discriminar mejor unas estaciones de otras y por eso también será de interés considerar 6 grupos donde también vemos un corte aunque mucho menos pronunciado. Comparamos los resultados obtenidos anteriormente con otro método llamado ‘gap statistic’ que toma información similar a lo que se hace en el ‘elbow method’, pero también se considera la variación esperada de los datos sin agrupación alguna. Obtenemos lo siguiente:

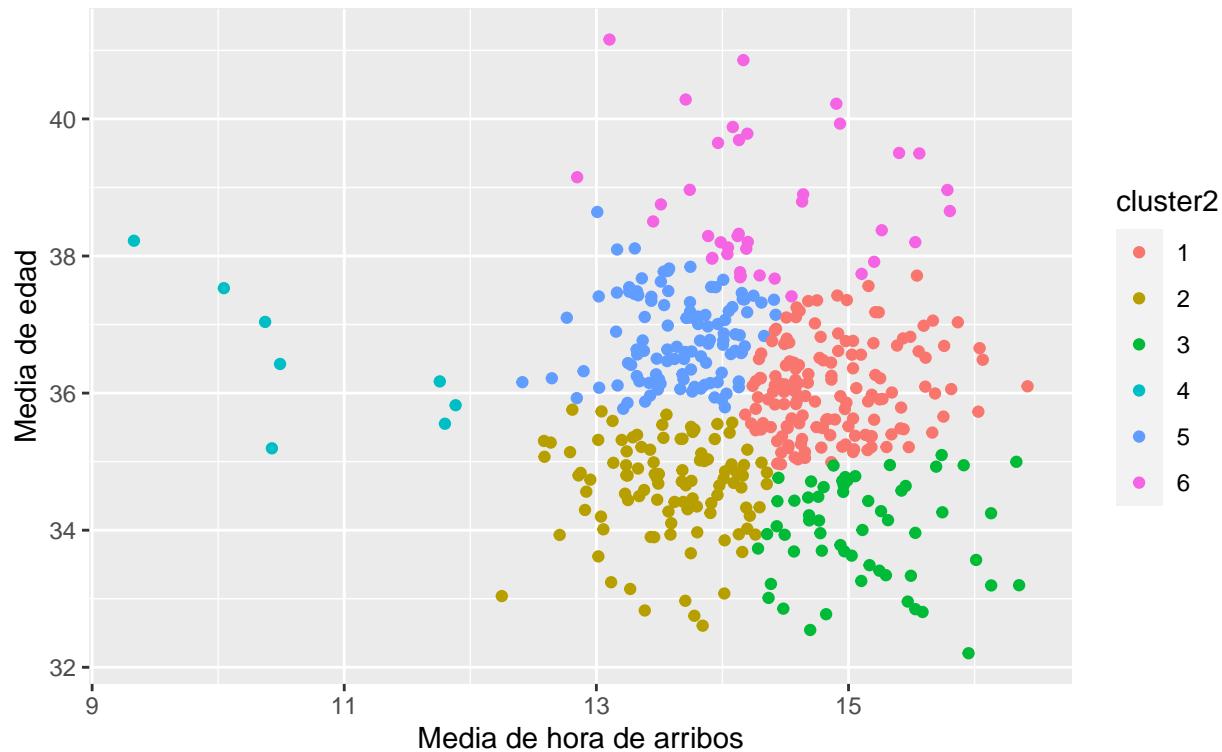


Vemos que aunque hay un máximo local en 3 grupos (cómo se vio con el otro método) después de los 6 grupos el decremento empieza a ser más pequeño cada vez. En lo que sigue mostraremos los resultados obtenidos con 3 grupos pero nos enfocaremos más en el caso de 6 grupos.

Agrupación de Estaciones tomando los retiros de las estaciones



Agrupación de Estaciones tomando los retiros de las estaciones



Las estaciones de los grupos 1,3 y 6 tienen horas de arribo altas (recordamos que aunque los valores están alrededor de las 3:00 p.m. estos son promedios) esto se puede deber a que se encuentran principalmente en zonas universitarias, oficinas o comerciales, ya que al salir de ahí las personas llegan a su destino tarde. Por el otro lado podemos considerar que las estaciones de los grupos 2 y 5 se encuentran en zonas residenciales, ya que las personas que tomaron bicicletas de esas estaciones tuvieron que haber llegado a las mismas bastante temprano, podemos decir lo mismo con mayor seguridad para el grupo 4.

Aunque en cuanto a hora de arribo los grupos 2 y 5 son similares, las edades de los usuarios son muy diferentes, tomando en cuenta que estas estaciones se encuentran en zonas residenciales estas se pueden discriminar por niveles de ingreso considerando que en promedio las personas que usan las estaciones del grupo 2 tienen menor ingreso en promedio que las del grupo 5. Con respecto a las estaciones 1,3 y 6 podemos usar la edad para discriminar el tipo de industria al que corresponden las personas que salieron de esas estaciones.

Pregunta 3

El modelo estadístico paramétrico que vamos a usar para encontrar tendencias en el número de usos por estación por día va a ser Regresión Lineal. Para cada estación construiremos un modelo de regresión lineal simple y consideraremos la pendiente (el coeficiente que estima el cambio en usos a lo largo del tiempo). Como este valor va a depender del nivel de usos que haya por estación es necesario relativizar esta cantidad y para hacer eso dividimos la pendiente entre el promedio de usos diarios por estación.

Aunque hay efectos de estacionalidad que puedan meter ruido a los modelos al considerar los mismos períodos para todas las estaciones podemos suponer que estos efectos impactan de la misma forma a todas las estaciones y trabajar con la información como se tiene.

Las 3 estaciones con mayor tendencia a la alta y menor tendencia a la baja son las siguientes:

```
## # A tibble: 6 x 3
```

```

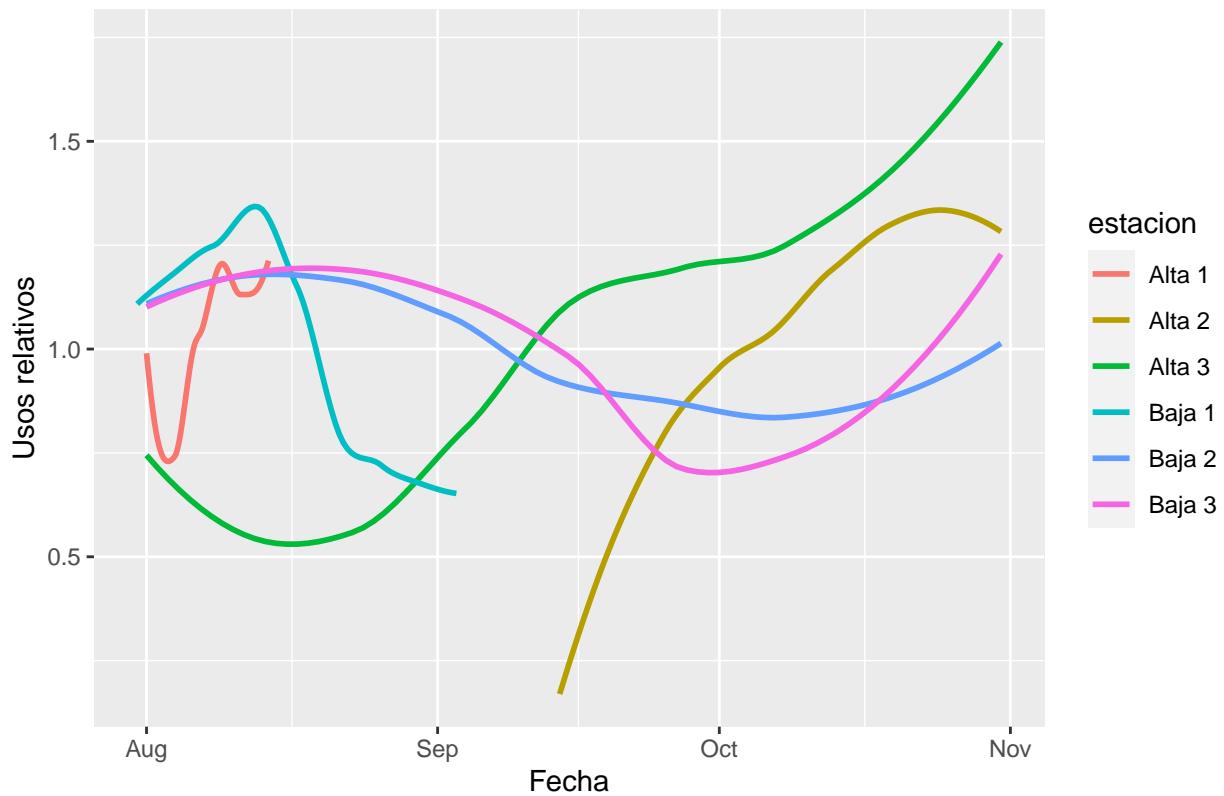
##   estacion `Indice de Tendencia` Clasificación
##   <int>           <dbl> <chr>
## 1     442          0.0350 Alta 1
## 2     208          0.0222 Alta 2
## 3     192          0.0130 Alta 3
## 4     193         -0.0198 Baja 1
## 5     260         -0.00420 Baja 2
## 6     148         -0.00404 Baja 3

```

donde el Indice de Tendencia representa la proporción del promedio de usos diarios que representa el cambio de usos de un día a otro. Ahora hacemos una comparación de las tendencias gráficamente.

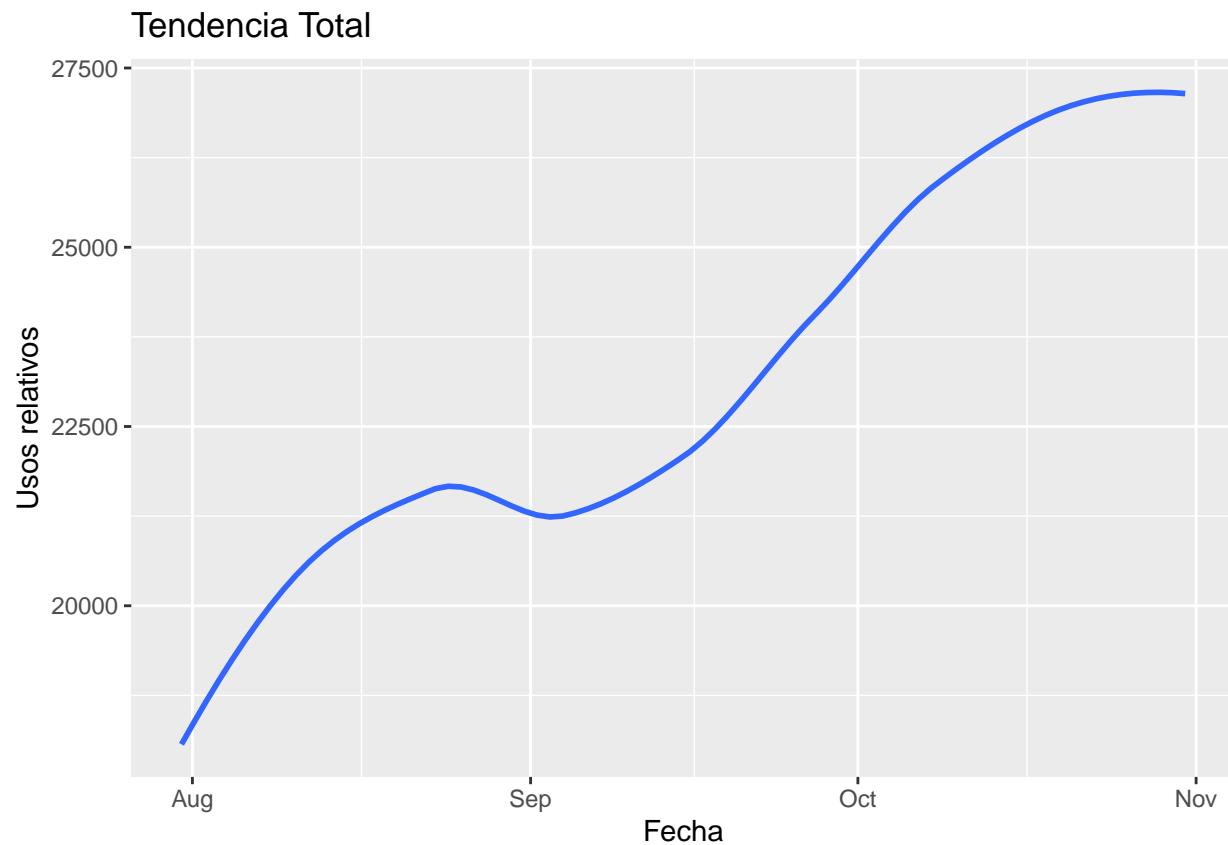
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Comparación de tendencia



Podemos ver que esta forma de comparar la tendencia da buenos resultados si nos interesa el comportamiento un poco más a largo plazo, pero que tal vez no le asigna la importancia que se merece la información más reciente, como se ve en los casos de las estaciones Baja2 y Baja 3. También es importante mencionar que en general se ha presentado un comportamiento de tendencia a la alza como muestra la última gráfica y que por eso no está mal que nuestras estaciones con tendencia a la baja no todas tengan un comportamiento como el de la estación Baja 1.

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



Esta información puede ser útil para saber en qué tipo de lugares sería conveniente abrir nuevas estaciones o dar mayor capacidad a las que existen.

Pregunta 4

Para mostrar el mapa se puede usar la API de Google Maps para JavaScript.