

Sección 1: Bebés (1 hora) Obligatorio.

1. Para cada AGEB de la delegación Álvaro Obregón estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. Enlista tus fuentes y presenta los resultados:

Estimaremos primero el número medio de hijos por edad y AGEB en un intervalo de seis meses, utilizando la tasa de natalidad en cada AGEB y el número medios de hijos por edad en un intervalo de seis meses. Con este resultado, y una estimación para cada AGEB de la distribución de mujeres por edades, obtendremos el número de bebés menores a seis meses que habitan el cada AGEB. Todos los datos fueron proporcionados por el INEGI, a través del censo de 2010 y de los registros administrativos (en <http://www.inegi.org.mx/est/contenidos/proyectos/registros/vitales/consulta.asp?c=11781&s=est#>)

Sea  $N_{t,i}$  la variable aleatoria que modela el número de hijos que tiene una mujer del AGEB  $i$  durante el periodo  $t$  de su vida ( $t$  es discreta y denota semestres). La suma  $\sum_t N_{t,i}$  modela entonces entonces el número total de hijos. Del censo 2010 utilizaremos el promedio de hijos nacidos vivos por AGEB,  $T_i = E[\sum_t N_{t,i}]$ .

Nos interesa obtener los nacimientos medios por edad y AGEB,  $E[N_{t,i}]$ . Con estos valores podremos entonces estimar el número de nacimientos en los ultimos seis meses en cada AGEB como

$$\sum_{t \in \{\text{edades de todas las mujeres en el AGEB con índice } i\}} E[N_{t,i}].$$

Para obtener los nacimientos medios por edad  $E[N_{t,i}]$  a partir del promedio de hijos  $T_i$ , utilizaremos una distribución de la edad de las madres en Álvaro Obregón reescalada, es decir

$$E[N_{t,i}] \sim \frac{\# \text{nacimientos de madres de edad } t}{\# \text{nacimientos totales}} \times \text{promedio de hijos en AGEB } i.$$

Al correr el modelo con los datos del INEGI obtenemos los nacimientos estimados en los últimos seis meses para cada una de las 199 AGEB. Estas estimaciones tienen una media de 33, un mínimo de 0 (parece haber AGEB despoblados) y un máximo de 108 nacimientos por AGEB.

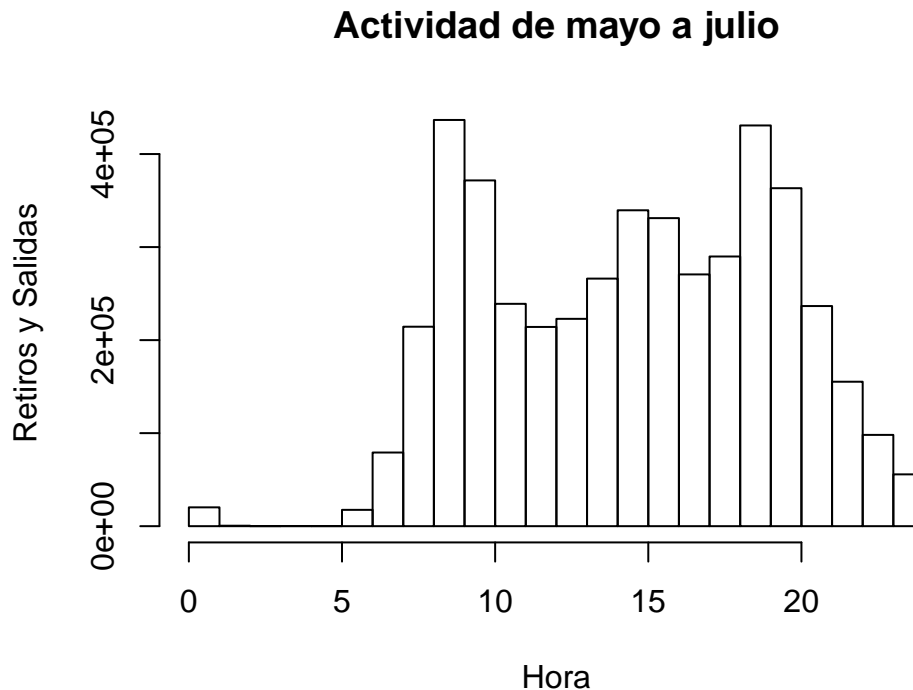
[illegible]

## Sección 2.a: Ecobici (4 horas) Intermedio

En la página de datos abiertos de Ecobici (<https://www.ecobici.cdmx.gob.mx/es/informacion-del-servicio/open-data>) baja los datos de movilidad de los últimos 3 meses y contesta las siguientes preguntas:

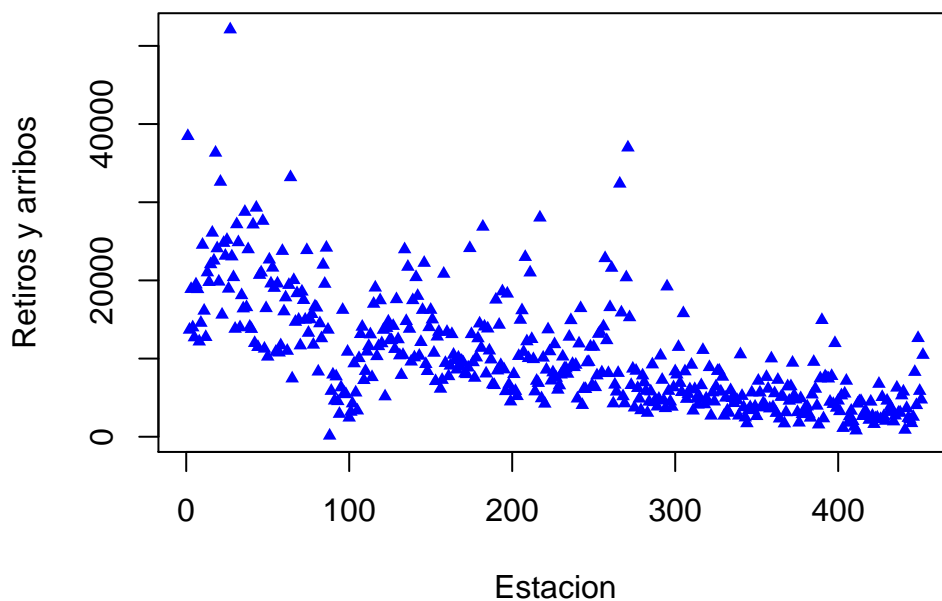
1. ¿En qué horarios hay mayor afluencia y en qué estaciones? Da una breve descripción de por qué crees que es así.

Durante esta sección utilizaremos los registros de Ecobici de mayo a julio de 2016. Los valores de agosto y septiembre fueron omitidos intencionalmente ya que hay un error generalizado en ellos. En estos meses, la hora de retiro a lo largo de un día contiene únicamente valores en minutos y segundos. Es probable que esto sea un error de formato y que sólo sea necesario reescalar estos valores de un intervalo de 0 a 60 min a un intervalo de 0 a 24 horas, pero debería ser confirmado antes de hacer uso de estos datos.



*Figura 1: Uso por horario.*

La figura 1 muestra dos claros incrementos de afluencia, el primero alrededor de las 8 y el segundo alrededor de las 19 hrs. El incremento de afluencia se debe a que Ecobici es usado para ir y regresar del trabajo. Existe un incremento menor alrededor de las 14 hrs. Éste se debe probablemente al uso de Ecobici para trasladarse de la escuela a casa y al movimiento durante el tiempo de comida en las oficinas de la ciudad.

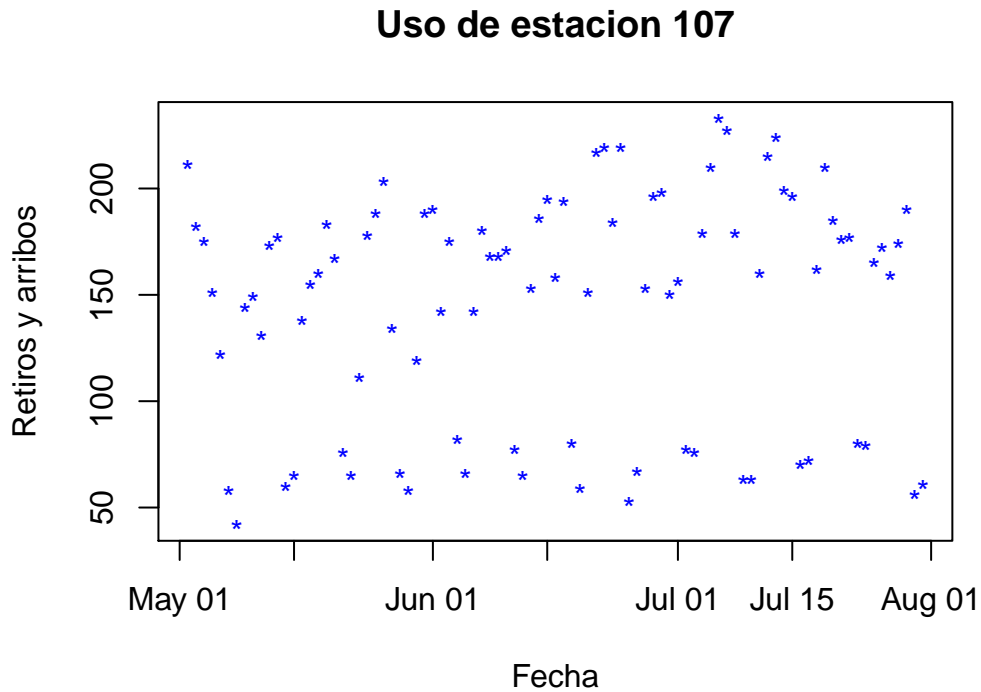


*Figura 2: Uso por estacion.*

En cuanto al uso por estación, la figura 2 muestra una mayor afluencia en las estaciones con numeración menor. Probablemente éstas fueron las primeras estaciones y están distribuidas en áreas que fueron prioritarias al inicio del programa. Estas áreas deben tener mayor interés, tráfico peatonal y costumbre de uso que el resto de las estaciones.

2. A partir de un análisis temporal:
  - a. ¿En qué estaciones puedes observar una tendencia de uso a la alta?

- b. ¿Puedes categorizar las estaciones con base en su tendencia de uso?  
c. Demuestra tus conclusiones gráficamente

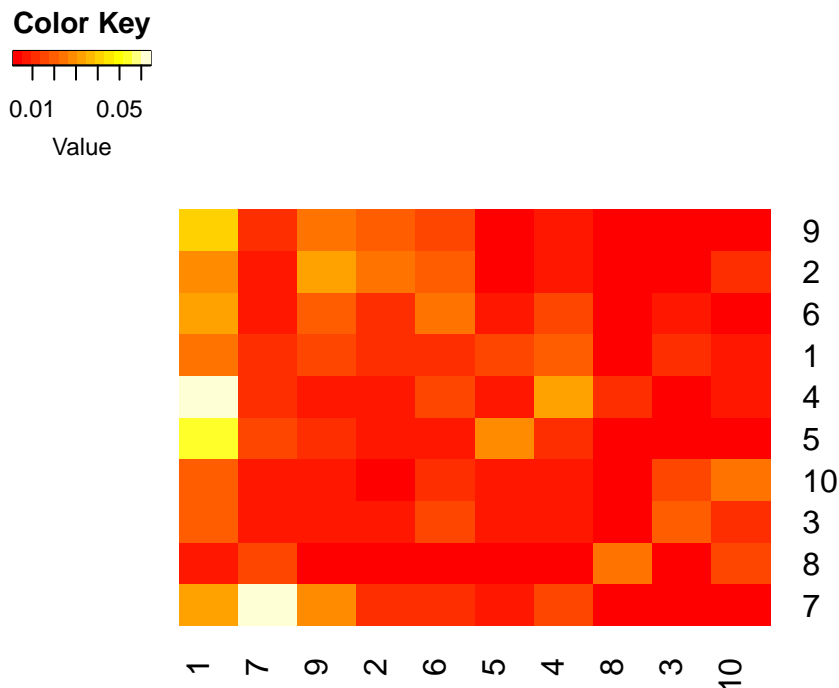


*Figura 3: Uso de la estación número 107.*

Hacer regresión lineal con el tiempo como variable de predicción y el uso como variable de respuesta es útil para determinar la tendencia de uso a lo largo de los tres meses analizados. La regresión no es adecuada para predicción, el modelo no es lineal, pero servirá como medida de la tendencia general a lo largo de los tres meses. Al hacer regresiones de este tipo para cada una de las estaciones podemos comparar tendencias a través del coeficiente de la variable de predicción. Coeficientes negativos indican una tendencia a la baja y coeficientes positivos indican una tendencia a la alta. Los resultados en los meses analizados muestran poca variación general. La estación 107, mostrada en la figura 3, tiene la mayor tendencias a la alta. La figura 3 muestra una gran decrecimiento del uso durante los fines de semana, ésta es una característica común en muchas de las estaciones.

En cuanto a la tendencia a lo largo de un día ordinario, esta característica será analizada en la pregunta 4 de esta sección. En este caso sí existen diferencias claras que causarán la separación de estaciones en los diferentes clusters.

3. Por cada estación de Ecobici, identifica cómo están correlacionadas las entradas-salidas entre las otras estaciones (Hint: Puedes usar un heatmap para mostrar la correlación o matrices de origen destino).

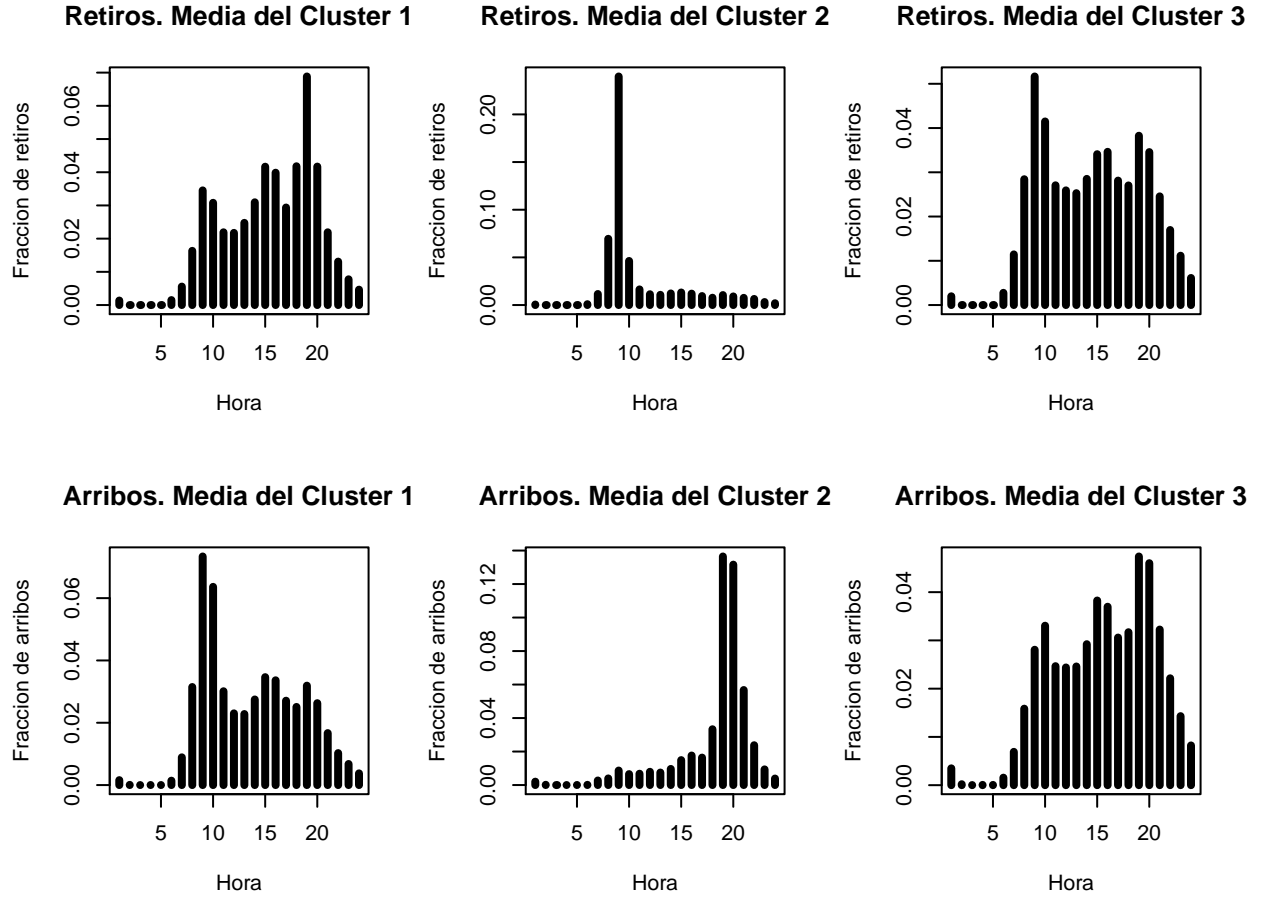


*Figura 4: Heatmap del retiro/arribo en las primeras 10 estaciones.*

Para describir la relación entre estaciones, utilizaremos un heatmap. Los renglones denotan estaciones de retiro y las columnas representan estaciones de arribo. Los valores del heatmap representan la proporción de trayectos dirigidos a cada estación, dada una estación de retiro. La suma de cada renglón es 1. En la figura 4 sólo se muestran los valores de arribo y retiro para las primeras 10 estaciones, mostrar todas haría la gráfica excesivamente densa.

4. Usa un método de aprendizaje no supervisado para encontrar “perfiles de uso” de las estaciones. Lo que debes de hacer es categorizar a las estaciones en diferentes grupos a partir de su comportamiento de entradas y salidas. Explica qué método usaste y por qué. De los grupos que encuentres describe las características que puedes inferir de estos a partir de lo descubierto en el inciso anterior.

Cada estación será caracterizada en este ejercicio como un vector de 48 entradas, éstas representan la distribución de retiros por hora y la distribución de arribos por hora. La distribución de arribos y retiros durante un día es intuitivamente una buena descripción de cada estación. Ambas distribuciones serán normalizadas, por lo que la suma de los valores para cada estación será de dos. Al hacer un análisis visual de las distribuciones para cada estación es claro que existen tres clases generales de estaciones. Algunas estaciones tienen una gran cantidad de retiros por la mañana y una gran cantidad de arribos en la tarde. El segundo tipo de estaciones tiene una dinámica opuesta, una gran cantidad de arribos por la mañana y una gran cantidad de retiros por la tarde. La última de las clases tiene incrementos de uso alrededor de los mismos momentos pero la cantidad de arribos y retiros son similares.



*Figura 5: Punto Medio de cada uno de los tres clusters.*

La figura 5 muestra la media de clusters obtenidos a través de k means con tres clusters. Estas medias muestran claramente la dinámica mencionada antes. El primer cluster, arribos por la mañana y retiros en la tarde, contiene 146 elementos y proviene probablemente de estaciones en areas comerciales. El segundo cluster, retiros por la mañana y arribos en la tarde, es el más pequeño, con 10 elementos, y coincide con estaciones en zonas residenciales. El tercer cluster es el más grande, con 296 elementos, y contiene estaciones que deben pertenecer a zonas mixtas de espacios comerciales y residenciales.