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.

AGEB	Bebés	AGEB	Bebés	AGEB	Bebés	AGEB	Bebés	AGEB	Bebés	AGEB	Bebés
0012	11.3	0027	16.8	0031	28.8	0046	53.8	0050	35.1	0065	47.6
0084	27.0	0099	40.4	0101	11.0	0116	39.5	0135	55.9	014A	37.4
0169	28.7	0173	43.2	0188	4.7	0192	5.1	0205	3.7	021A	4.5
0224	3.9	0239	8.3	0243	82.1	0258	5.3	0262	3.5	0277	3.3
0281	2.1	0296	22.4	0309	32.2	0313	59.8	0328	76.1	0332	47.38
0347	36.0	0351	42.3	0370	35.4	0385	35.3	0417	14.2	0421	57.7
0440	50.2	0455	29.6	0474	108.4	0489	40.9	0493	41.3	0506	33.2
0510	20.2	0525	16.0	053A	44.7	0544	22.1	0559	13.2	0563	86.1
0578	56.7	0597	35.6	060A	21.4	0614	13.7	0629	17.7	0633	12.9
0648	8.6	0667	45.2	0671	16.3	0686	15.0	0690	24.7	0703	12.6
0718	6.2	0722	13.4	0737	4.4	0741	29.0	0756	9.9	0760	5.4
0775	9.7	078A	7.4	0794	4.6	0807	12.9	0811	5.9	0826	4.6
0830	5.4	0845	9.3	0864	59.2	0898	19.3	0900	22.6	0915	18.3
092A	18.9	0949	6.4	0953	11.9	0968	22.7	0972	15.8	0987	9.2
0991	17.7	1006	17.9	1010	18.2	1025	3.4	103A	15.8	1044	18.0
1059	15.1	1063	17.8	1078	51.0	110A	23.2	1114	8.3	1129	8.2
1133	11.6	1148	33.2	1152	16.3	1171	103.3	1186	22.5	1190	21.1
1203	1.2	1218	12.2	1222	30.3	1237	28.0	1241	26.9	1260	26.8
1275	21.4	128A	16.2	1294	25.6	1307	16.7	1330	66.1	1345	73.9
135A	64.3	1364	55.3	1379	7.7	1434	48.7	1453	55.7	1468	24.9
1472	43.5	1487	42.5	1519	44.9	1523	37.0	1542	4.3	1557	58.8
1561	24.7	1580	61.9	1595	58.9	1627	52.6	1631	48.1	1646	51.8
1650	49.2	1665	23.6	167A	66.2	1684	43.2	1699	60.2	1716	66.8
1720	27.2	1735	77.4	174A	80.9	1754	29.7	1769	66.0	1773	38.4
1788	40.2	1792	27.9	1805	36.4	181A	50.3	1824	42.4	1839	44.7
1843	70.1	1858	80.4	1881	20.2	1913	32.1	1928	16.0	1932	25.9
1947	38.1	1951	49.8	1966	26.5	1970	39.4	1985	45.2	199A	32.9
2004	0.0	2019	29.8	2023	41.0	2038	35.3	2042	36.3	2057	40.5
2061	39.6	2076	24.0	2080	18.5	2095	44.0	2108	45.6	2112	92.6
2131	13.8	2146	23.0	2150	31.4	2165	34.1	2184	20.7	2199	46.6
2201	28.0	2216	35.6	2220	34.4	2235	41.0	224A	28.2	2254	35.9
2269	13.3	2273	8.8	2288	47.9	2292	39.0	2305	36.3	231A	25.0
2324	26.28	2339	17.5	2343	55.4	2358	14.2	2362	22.0	2377	19.9
2381	3.2										

Sección 2.a: Ecobici (4 horas) Intermedio

En la pgina de datos abiertos de Ecobici (`https://www.ecobici.cdmx.gob.mx/es/informacion-del-servicio/open-data`) baja los datos de movilidad de los ltimos 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 unicamente valores en minutos y segundos. Es probable que ésto 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 esto debería ser confirmado antes de hacer uso de estos datos.

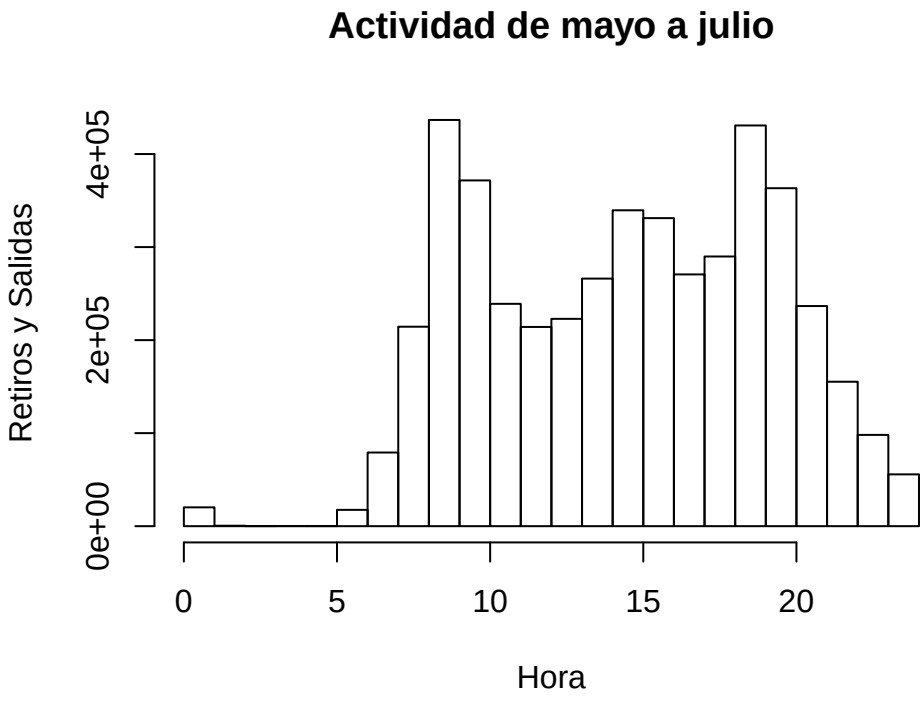

Figura 1: Uso por horario.

La figura 1 muestra dos claros incrementos de afluencia, el primero alrededor de las 8 y el segundo aredeor 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.

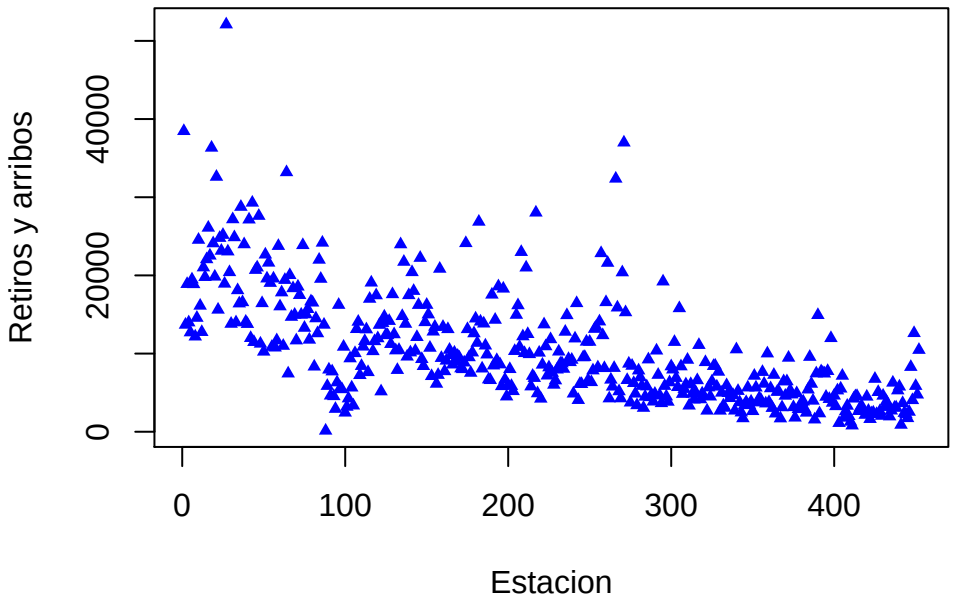

Figura 2: Uso por estacion.

En cuanto al uso por estación, la figura 2 muestra una mayor afluencia en las estaciones con numeración menor. Probablemente estas fueron las primeras estaciones y están distribuidas en areas 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 grficamente

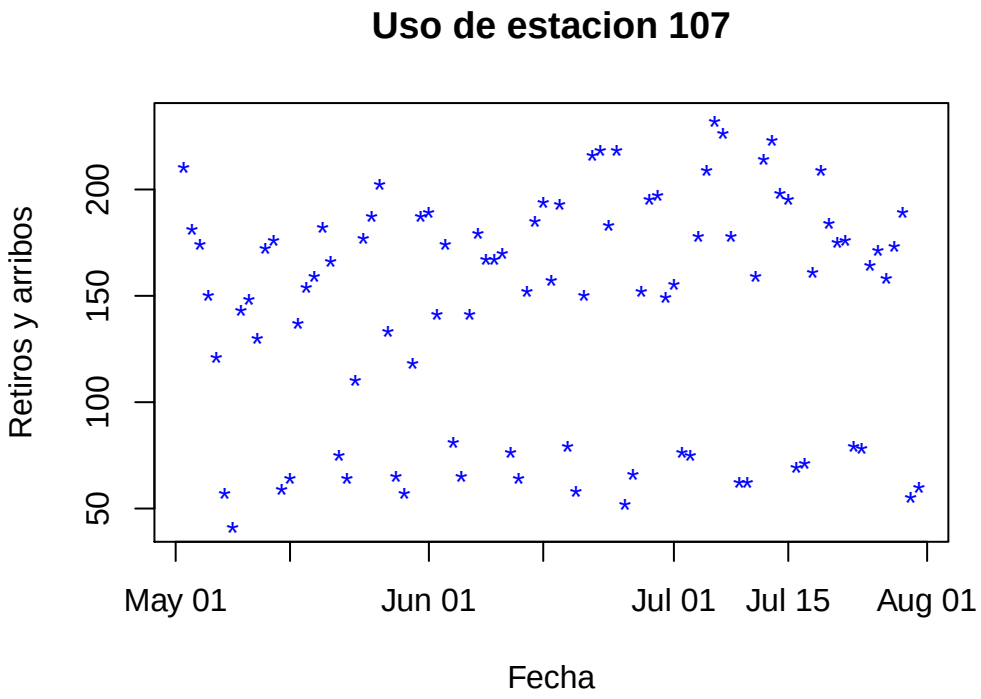

Figura 3: Uso de la estación número 107.

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 variacion general. La estació 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 oridnario, 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).

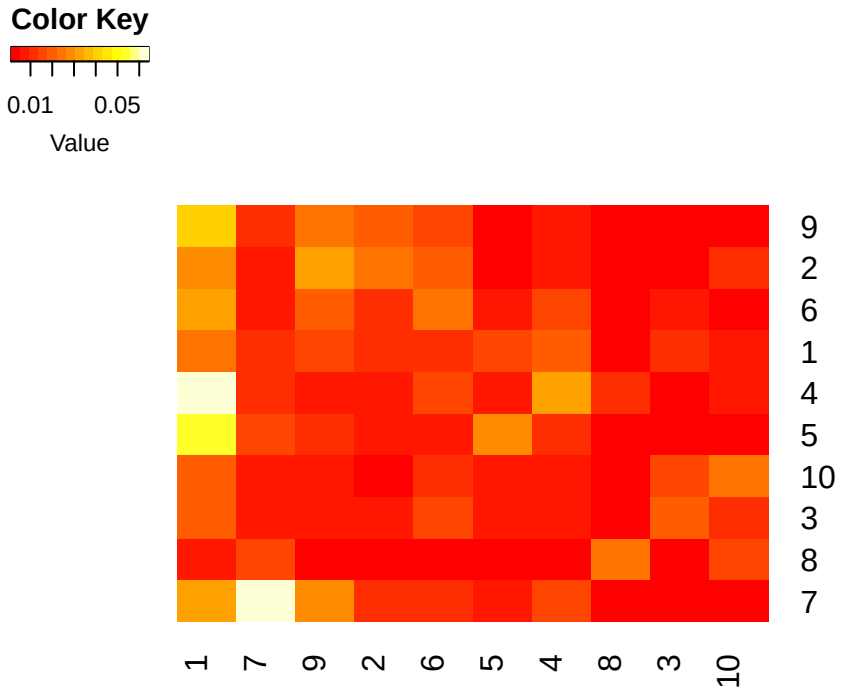

Figura 4: Heatmap del retiro/arribo en las primeras 10 estaciones.

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 encontraste 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 estació 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.

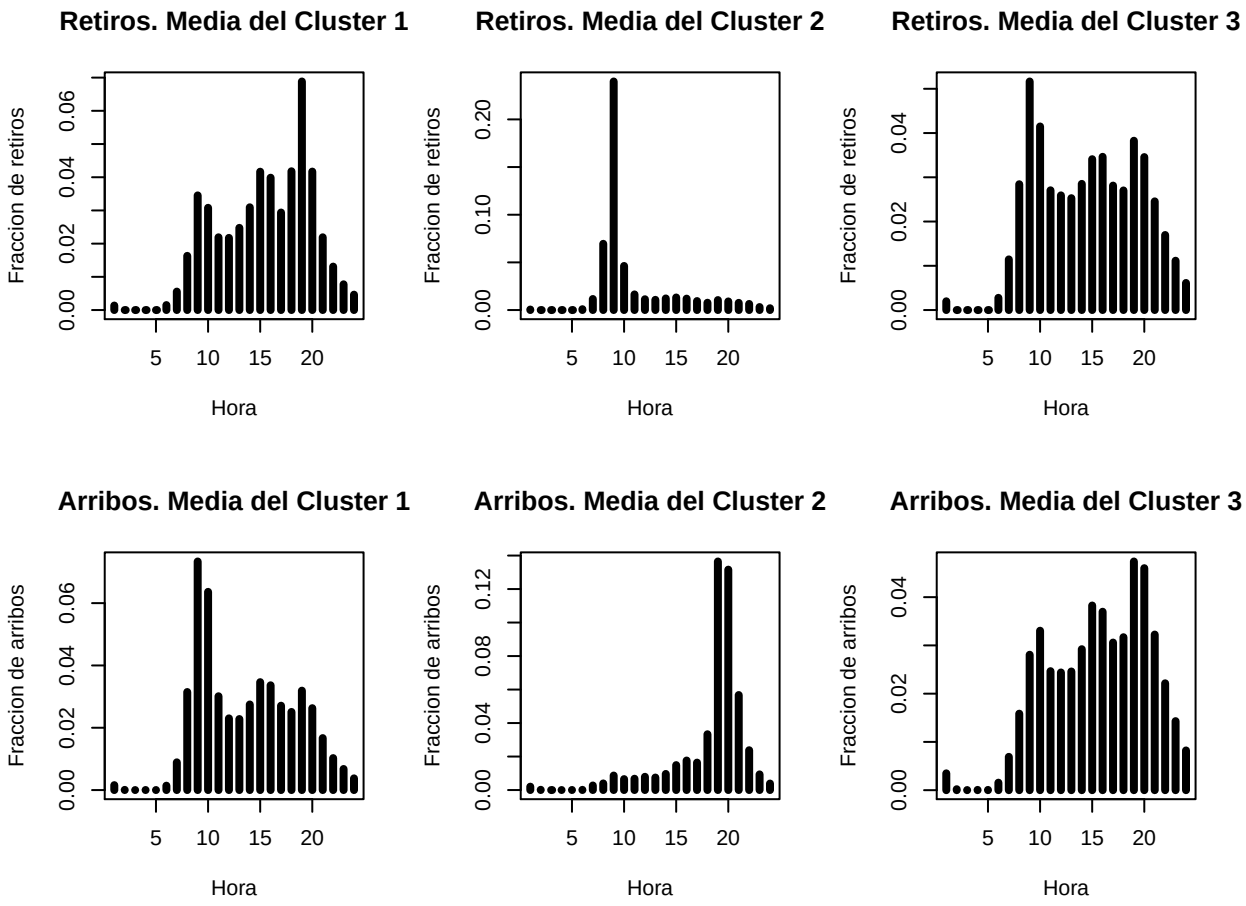

Figura 5: Punto Medio de cada uno de los tres clusters.

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 148 elementos y proviene probablemente de estaciones en areas comerciales. El segundo cluster, retiros por la maana 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 247 elementos, y contiene estaciones que deben pertenecer a zonas mixtas de espacios comerciales y residenciales.