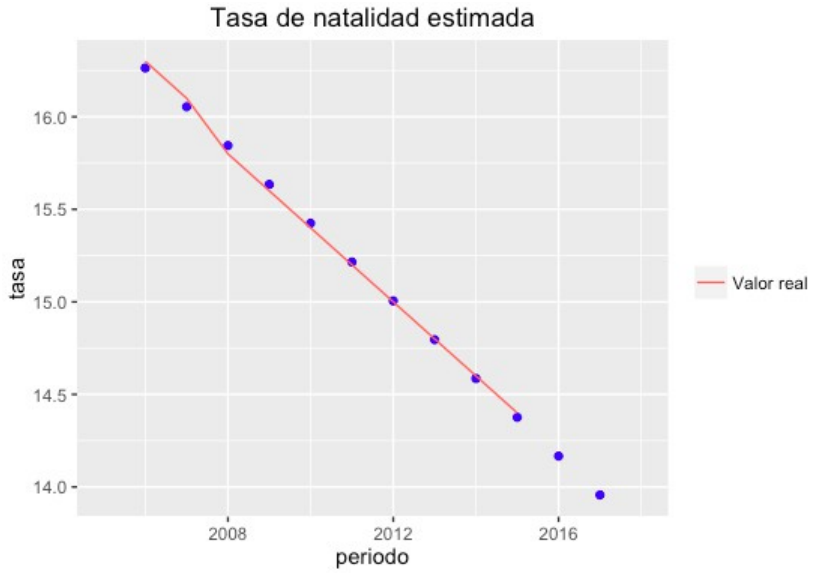
Armando Navarro Rodríguez **Examen OPI**

Sección 1: Bebés (1 hora) Obligatorio.

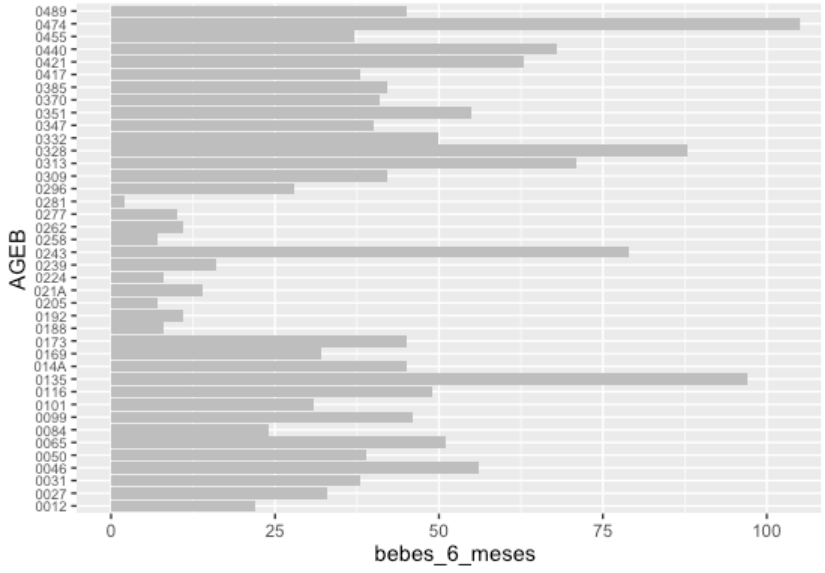
1. Para cada AGEB de la delegación Álvaro Obregón estima cuántos bebés de 0 a 6 meses de edad habitan ahí el día de hoy. Explica tu razonamiento en menos de 300 palabras. Enlista tus fuentes y presenta los resultados. (Hint: revisa el CPV 2010, puede ser útil)

Para la estimación se descargaron los datos correspondientes a los niños (ambos sexos) menores de 2 años que vivían en la delegación Álvaro Obregón en el año 2010 [1]. Se conservó la información de las AGEB y los niños registrados. El 25 % de los registros tienen valores nulos (probablemente no sistemáticos) y se llenó con el valor promedio (7 niños en cada manzana), se agrupó la información por AGEB para obtener el total estimado de niños que vivían en cada AGEB en el año 2010.

Se consultó la tasa bruta de natalidad en la Ciudad de México [2] en los últimos 10 años disponibles (2006 – 2015) y mediante un modelo lineal se estimaron las tasas de los años 2016 y 2017 (predecir más años con este modelo podría ser inadecuado), las predicciones se muestran en la siguiente figura:



Con base en la predicción anterior se estableció una proporción del 90.6 % (tasa del 2017 entre tasa del 2010) de niños menores a 2 años para el 2017. Finalmente, debido a que los niños pudieron nacer en 2015 o en 2016, se determinó que los niños menores a 6 meses son el 24.86 % de ese total (tasa del 2017 entre tasas del 2016 y 2017, dividido entre dos para la mitad del año), se muestran los resultados de las primeras 40 AGEB:



Durante el razonamiento del ejercicio se asumió lo siguiente:

a) La tasa de natalidad de la Ciudad de México es igual a la de la delegación Álvaro Obregón.
b) Los nacimientos durante un mismo año se distribuyen de manera uniforme en cada mes.
c) Los datos se registraron durante el mes de enero (de forma que los niños de 0 a 6 meses nacieron a partir de julio de 2016).

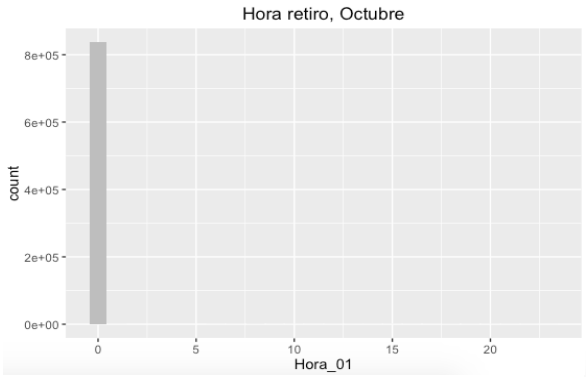
[1] http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/Default.aspx
[2]http://www.beta.inegi.org.mx/temas/natalidad/

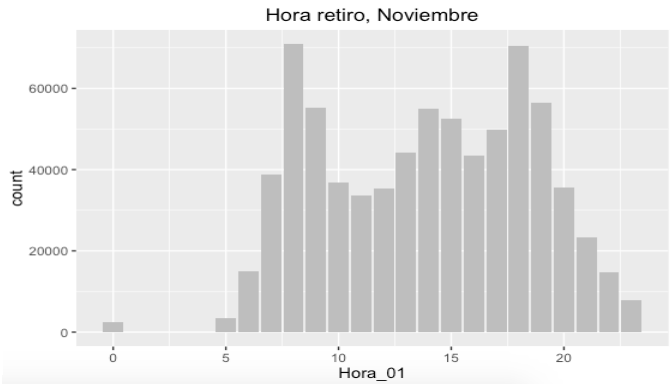
Sección 2.a: Ecobici (4 horas) Intermedio

En la página de datos abiertos de Ecobici, baja los datos de movilidad de los últimos 3 meses y contesta las siguientes preguntas:

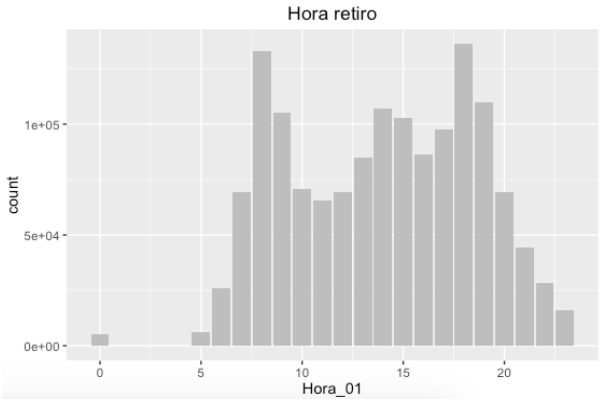
1. ¿En qué horarios hay mayor afluencia y en qué estaciones? Da una breve descripción de por qué crees que es así

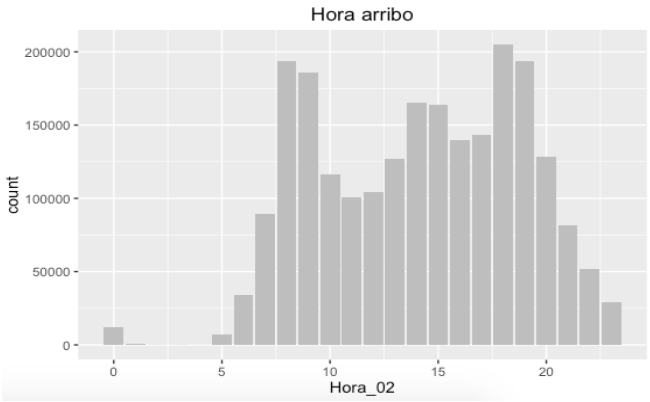
Antes de comenzar con la explicación es importante mencionar que durante el análisis se encontró que la mayoría de los registros de hora de retiro en el mes de octubre ocurren entre las 00 y 01 horas, se consideró esto como erróneo y se descartó esa información:





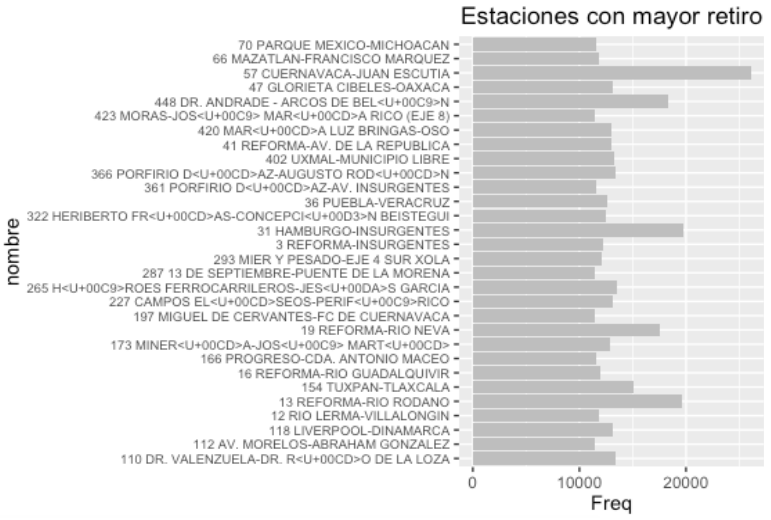
Se agruparon los registros restantes por intervalos de hora y se graficaron los resultados:

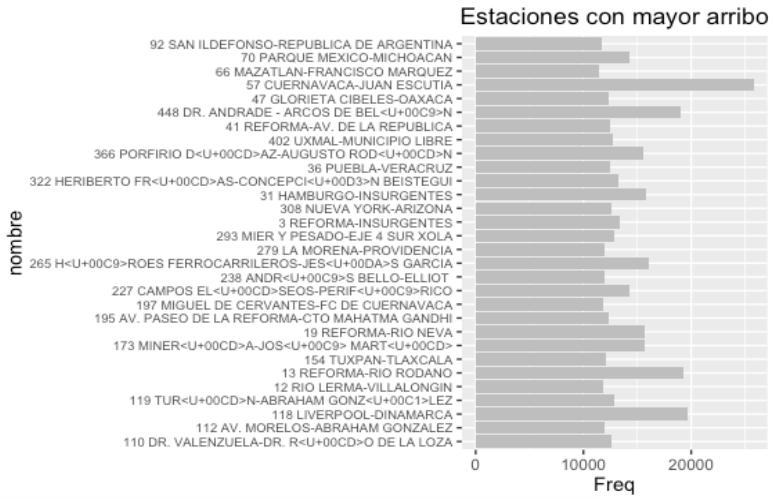




Se observó un comportamiento similar tanto en las horas de retiro como en las de arribo (esto puede sugerir que el tiempo de uso de las bicicletas en general no es prolongado), siendo los horarios de mayor afluencia de 08 a 10 horas y de 18 a 20 horas, existe otro pico importante entre las 14 y 16 horas.

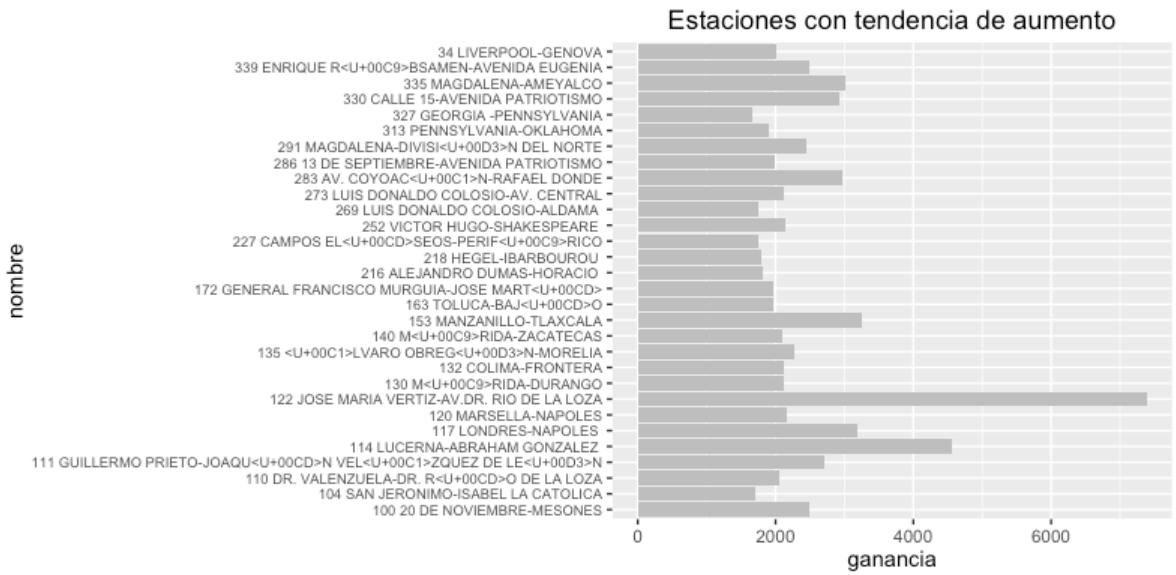
Las estaciones más utilizadas son Cuernavaca - Juan Escutia, Dr. Andrade – Arcos de Belén, Hamburgo – Insurgentes, Reforma – Río Neva y Tuxpan – Tlaxcala y Reforma – Rio Rodano:



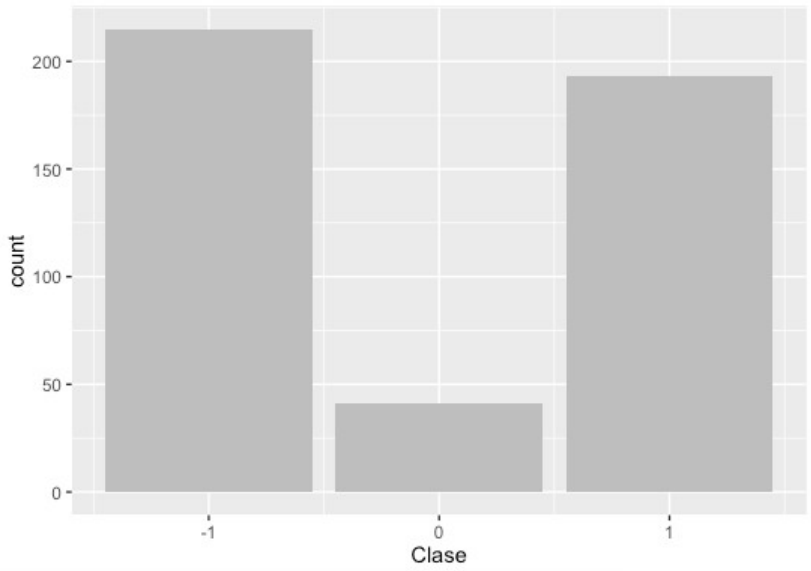


2. A partir de un análisis temporal: ¿En qué estaciones puedes observar una tendencia de uso a la alta? ¿Puedes categorizar las estaciones con base en su tendencia de uso? Demuestra tus conclusiones gráficamente.

Para este ejercicio se consideró el número de registros en cada cicloestación durante los meses de octubre, noviembre y diciembre, se determinó para cada cicloestación la pérdida / ganancia en afluencia en diciembre respecto octubre (otra forma sería comparar los registros mensuales contra los correspondientes en años anteriores). Las estaciones que registran mayor ganancia (y por lo tanto más afluencia) son José María Vertiz – Dr. Río de la Loza, Lucerna – Abrahám Gonzales y Manzanillo – Tlaxcala, como se muestra en la siguiente gráfica:

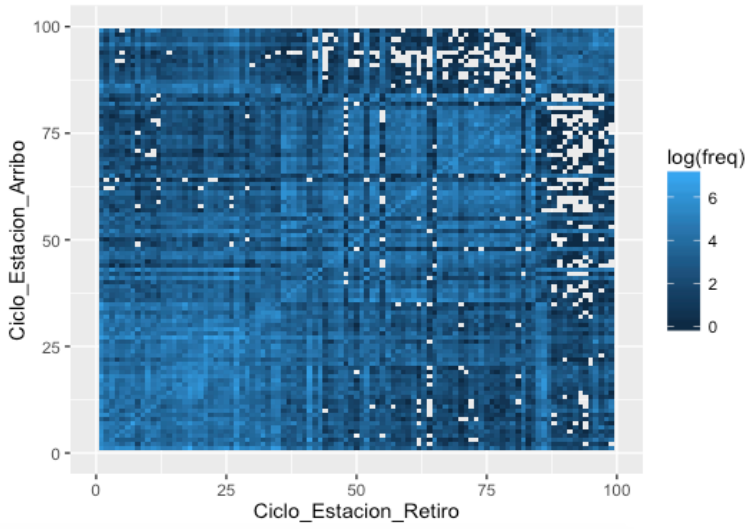


Una forma sencilla de clasificar las cicloestaciones por su ganancia / pérdida en uso es mediante tres clases: La primera (+1) se asigna a aquellas que registraron por lo menos 100 registros más respecto al periodo de referencia (en este caso, octubre), la segunda (-1) se asigna a las que registran pérdidas mayores a 100 registros y la última para las restantes (tienden a ser estables). El histograma de las clases de pertenencia es:



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).

Sería complicado interpretar directamente un heatmap para este ejercicio debido a que existen más de 400 cicloestaciones y el mapa sería muy denso, por lo que se seleccionaron las primeras 100:



Adicionalmente, se aplicó una escala logarítmica respecto al número de registros entre cada par de estaciones (por ejemplo, si hay 100 recorridos entre la estación A y B, en el mapa saldría con un valor de 2), lo anterior permitiría ver con más claridad las regiones donde hay mayor relación. Algo interesante que se puede observar es una diagonal con mayor frecuencia, lo cual sugiere que muchas personas salen y regresan a la misma estación (podrían sólo utilizar la bicicleta para hacer ejercicio, etc.).

Los 10 recorridos más utilizados son:

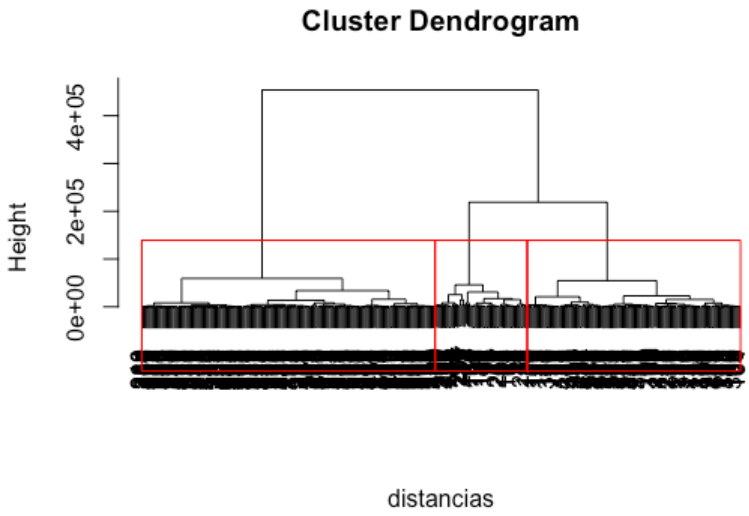
Inicio	**Fin**	**Frecuencia**
Progreso - Cda Antonio Maceo	Porfirio Díaz – Augusto Rodán	1714
Reforma – Rio Rodano	Dr. Andrade – Arcos de Belén	1421
Reforma – Rio Neva	Cuernavaca – Juan Escutia	1133
Romero de Terreros – Gabriel Mancera	La Morena – Providencia	1112
La Morena – Providencia	Romero de Terreros – Gabriel Mancera	1075
La Morena – Providencia	San Francisco – Santa Margarita	1060
Porfirio Díaz – Augusto Rodán	Progreso – Cda Antonio Maceo	1045
Rio Pánuco – Rio Tiber	Claudio Bernard – Dr. Liceaga	916
Rodríguez Saro – Av. Coyoacán	La Morena – Providencia	907
Dr. Andrade – Arcos de Belén	Reforma – Rio Rodano	868

Se puede observar que dentro de estos recorridos, aparecen tres pares ida y vuelta, sin embargo dos de ellos tienen diferencias significativas (Progreso – Cda Antonio Maceo y Porfirio Díaz – Augusto Rodán con 1714 registros de ida y 1045 de vuelta; Reforma – Rio Rodano y Dr. Andrade – Arcos de Belén con 1421 registros de ida y 868 de vuelta), mientras que en el otro son similares (Romero de Terreros – Gabriel Mancera y La Morena – Providencia con 1112 registros de ida y 1075 de vuelta).

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.

Para este ejercicio se construyeron vectores con 8 características correspondientes a cada cicloestación, las primeras cuatro corresponden al número de bicicletas que salen de la estación de 00 a 06, 06 a 12, 12 a 18 y 18 a 00 horas; las últimas cuatro corresponden al número de bicicletas que llegan a la estación de 00 a 06, 06 a 12, 12 a 18 y 18 a 00 horas.

Se decidió utilizar un método de agrupamiento aglomerativo ya que por medio del dendograma se puede (en ocasiones) apreciar el número de clusters que podrían ser adecuados:



Aparentemente, se podría hacer un recorte para tres grupos en una amplia región del dendograma, las medias de los tres grupos son:

```
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==1,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
4197.98551 2309.52174 2769.63768 2149.98551   88.27536 3818.23188 4338.01449 3543.72464
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==2,])
    Frec1     Frec2     Frec3     Frec4     Frec5     Frec6     Frec7     Frec8
903.93182 528.32273 586.24545 419.10455  21.72727 758.85000 854.66364 746.44545
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==3,])
    Frec1     Frec2     Frec3     Frec4     Frec5     Frec6     Frec7     Frec8
2236.3187 1219.0687 1430.0812 1025.4813   56.2625 1802.3312 2218.6312 1754.3125
```

Se puede observar que el primer grupo corresponde a las estaciones de mayor afluencia, el segundo a las de poca afluencia y el tercero a las mediana afluencia. Si aumentamos el número de grupos a seis, obtenemos las siguientes medias:

```
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==1,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
4416.73913 1954.41304 3133.17391 2498.02174   92.08696 4748.00000 4441.60870 2933.00000
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==2,])
    Frec1     Frec2     Frec3     Frec4     Frec5     Frec6     Frec7     Frec8
551.50000 334.86842 336.34211 240.46053  13.76316 366.76316 502.67105 448.07895
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==3,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
1089.93750  630.42361  718.13889  513.38889   25.93056  965.78472 1040.43750  903.91667
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==4,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
2037.14423 1158.80769 1216.20192  866.22115   55.22115 1362.01923 1996.06731 1701.34615
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==5,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
2606.21429 1330.98214 1827.28571 1321.25000   58.19643 2620.05357 2631.96429 1852.67857
> colMeans(super[,c(8,9,7,6,5,4,3,2)][grupos==6,])
     Frec1      Frec2      Frec3      Frec4      Frec5      Frec6      Frec7      Frec8
3760.47826 3019.73913 2042.56522 1453.91304   80.65217 1958.69565 4130.82609 4765.17391
```

El grupo 2 concentra las estaciones con menor afluencia. Los grupos 1 y 6 concentren altas afluencias, sin embargo el grupo 1 tiene más llegadas entre las 06 y 12 horas, mientras que el grupo 6 tiene más llegadas después de las 18 horas. Los grupos restantes tienen afluencias medianas, el grupo 3 tiene una lígera mayor proporción de salidas después de las 18 horas, mientras que la diferencia más notable entre los grupos 4 y 5 es que este último tiene casi el doble de llegadas entre las 6 y 12 horas aún cuando la mayoría de las características son similares.