Examen OPI - Data Science

MCC. Benjamin Perea Medina

January 6, 2017

1 Nacimientos

Para resolver este problema yo analice los últimos 5 censos de población, de ellos extrajé el número de bebes (entre 0 y 1 años) que viven en la GAM. Con esta información hice un modelo lineal (porque la tendendencia así lo mostraba, ver figura 1). Posteriormente con este modelo estime el número de bebes de entre 0 y 1 años que viven en la GAM en enero del 2017. Después investigué el número de nacimientos por mes, esto lo hice porque los censos generalmente se llevan a cabo a principios o a mediados de ao, de esto depende la edad del bebe y el número de bebes con una edad menor o igual a 6 meses. Con esta información, construí una tabla de frecuencias relativas y acumuladas, para ver la distribución de edades por mes. Finalmente multipliqué el resltado del modelo lineal con la distribución acumulada y obtuve el resultado que se puede ver si se ejecuta el script **niños6meses.r**.

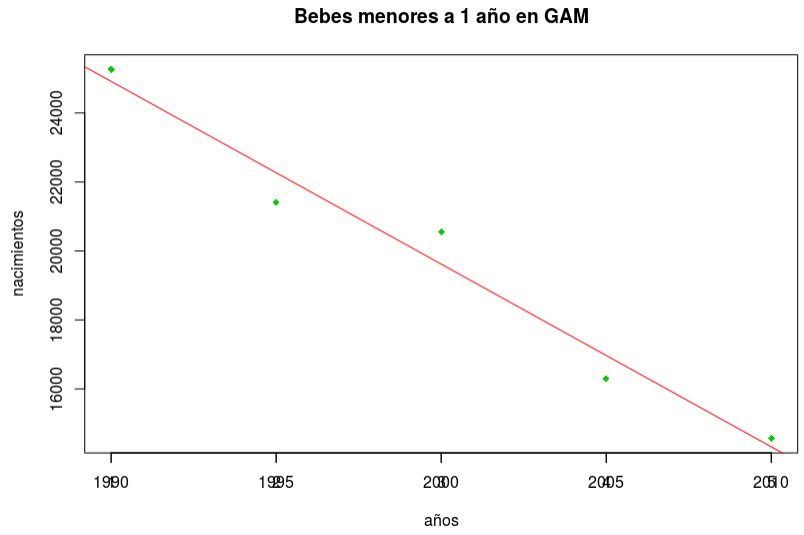

Figure 1: Modelo de Regresión lineal

2 2.A, Ecobici

1. Los horarios en los que hay mayor afluencia son de 8:00 a 10:00, 14:00 a 16:00 y de 18:00 a 20:00, esto se debe principalmente a los horarios de entrada, comida y salida de los trabajos. Las 5 estaciones que reguistran mayor uso promedio son 27 (Londres-Sevilla), 271 (Londres y Sevilla), 18 (Reforma-RioGuadalquivir), 1(Bruno Trave-Avenida México- Coyoacan) y 21 (Reforma-Lieja).

2. Análisis temporal

   - En las estaciones en las que se observa una tendencia de uso alta son, 27,271,18,1,21,41,61,217,36 y 15.
   - Estas estaciones se pueden categorizar en base a su uso promedio y despues establecer clasificarlas con base en la distribución de los cuantiles.
   - Un ejemplo grafíco de esta clasificación esta dada en la figura 2

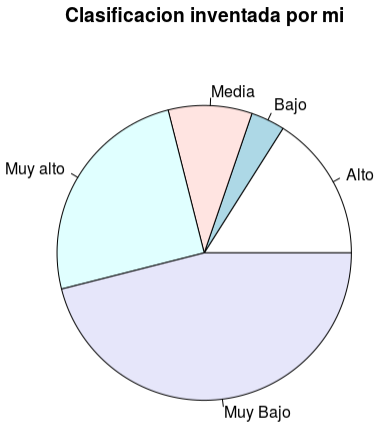

Figure 2: Pie de clasificación mia

3. Para dar solución a este problema, construí una matriz de origen destino entre las estaciones, esta se puede ver en el scrip **bici.r**.

4. Me base en dos modelos para ver los perfiles de uso de las estaciones.

   - En el primer caso me base en un cluster jerárquico, porque estos algoritmos de clustering se basan en el uso de alguna métrica para organizar la clasificación. Esto lo hice porque quería ver si la métrica euclidiana o la de manhattan influian

en los resultados, es decir si la distancia en el uso que a acada estacion se le da, facilita o no la clasificación, ver figuras 3 y 4. Sin embargo, esto no resulto en un cambio significativo en los resultados.

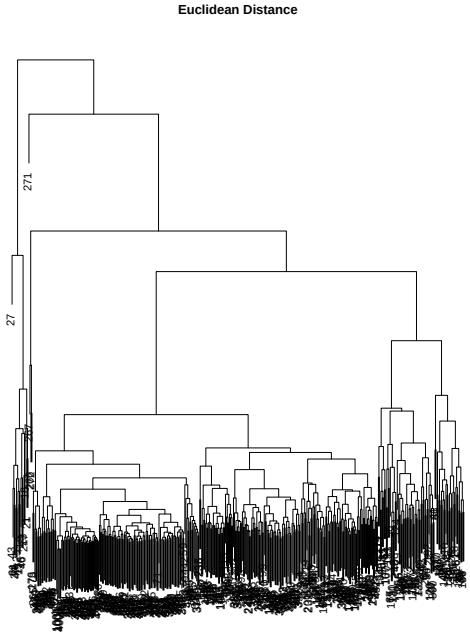

Figure 3: Cluster jerárquico con métrica euclidiana

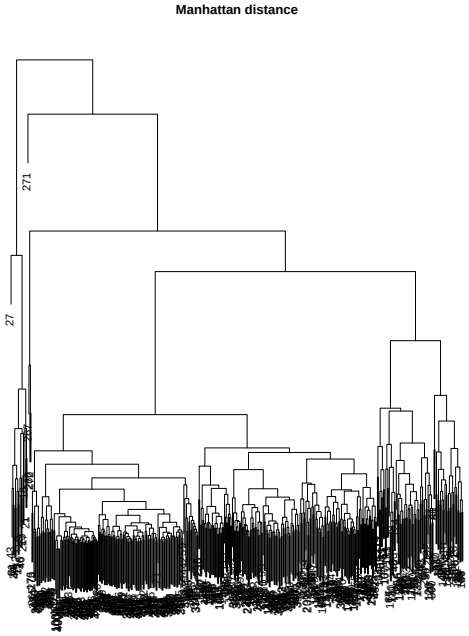

Figure 4: Cluster jerárquico con métrica manhattan

- En segundo lugar utilice el algoritmo de clusterin k-means con k=6, usé este algoritmo porque es basado en la idea de que cada cluster tiene una distribución multinormal y solo difieren de las medias de cada uno. Con este cluster pretendo construir una clasificación que se base en un numero fijo de grupos y que me permita identificarlos facilmente, ver las figuras 5 y 6.

  Con lo anterior podemos apreciar que el algoritmo de K-means es mejor para perfilar cada estación en cuanto a su uso. Tambien se puede apreciar que la mayoria de la estaciones presentan un uso medio-bajo, por lo cual se recomendaria replantear futuras rutas.

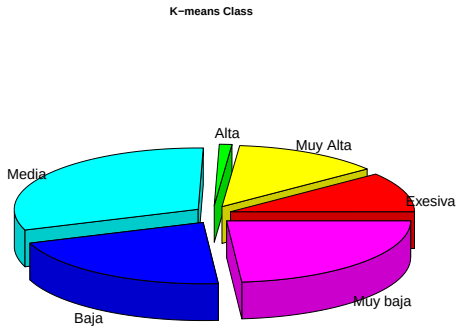

Figure 5: Pi chart calsificación por k-means

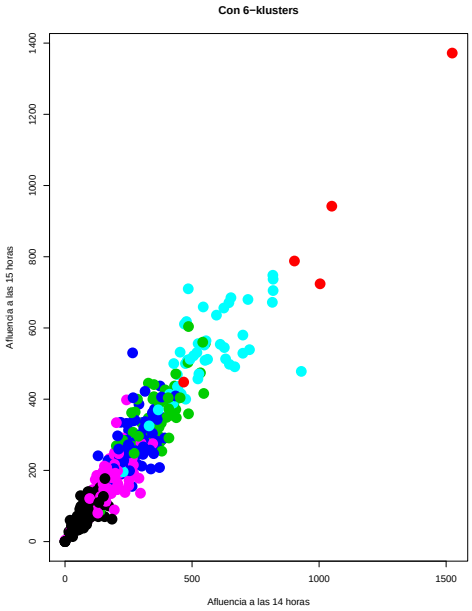

Figure 6: Distribución de las clases

BONUS Con base en la información geográfica y los datos, puedo concluir que el uso de la ecobici depende de tres aspectos, el número de oficinas en la zona, la falta de lugares de estacionamiento y la cerania de transportes públicos a las estaciones, ya que estos aspectos facilitan la movilidad de los ususarios