Prueba Técnica OPI

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. En lista tus fuentes y presenta los resultados.

   Se utilizaron datos por AGEB y manzana urbana (http://www.inegi.org.mx/sistemas/consulta_resultados/ageb_urb2010.aspx?c=28111&s=est) para estimar a los bebés de 0 a 6 meses de edad que habitan en el 2017 en cada AGEB de Álvaro Obregón . Para esto utilizamos solo los datos de las edades de 0 a 2 años, de 3 a 5 años, de 12 a 14 años y de 15 a 17 años y las copiamos en el archivo "finalbebes.xlsx". En dicho archivo se considera una proyección de dichos datos al 2017 (suponiendo que el número de menores de edad en 2010 sigue vivo al 2017). De acuerdo a esta proyección, el número de personas que en 2010 tenían de 0 a 2 años se presenta como el número de personas de 7 a 9 años y así sucesivamente hasta llegar a los menores del 2010 que en 2017 ya tienen 24 años. Finalmente se realiza una regresión lineal simple para los datos de cada AGEB. Con esta regresión se estima la cantidad de niños de 0 a 2 años vivos en el 2017 y dicha cifra se divide entre 4 para estimar los bebés de 0 a 6 años.

   Los resultados de los bebés de 0 a 6 años por AGEB se muestran en la última columna de dicho archivo. Por ejemplo, en el primer renglón, que contiene los datos totales se presenta una aproximación de 7119 bebés de 0 a 6 meses.

   Nota: Solo la última cifra (resultado) se presenta redondeada.

2. 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:

   Nota: La razón por la que decidí abordar este problema en vez del de Profeco es porque estaba teniendo muchos problemas para abrir el archivo por su peso (20GB). Este problema lo comencé resolviendo solo con los datos de diciembre. Al terminar se me ocurrió la "brillante" idea de tratar de procesar los datos de los tres meses en el mismo libro de excel y de un momento a otro mi computadora se trabó y ya no logré volver a abrir el archivo sobre el que estaba trabajando. No me da tiempo de rehacer todo, así que el problema 1 y 2 los resolveré platicando lo que pensaba hacer y lo que creo que hubiera sucedido.

Por otro lado, los problemas 3, 4 y 5 ya estaban resueltos, pero solo para los datos de diciembre, así que solo presentaré dichos resultados.

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

   La estrategia para resolver este problema es crear una tabla dinámica en excel para facilitar el conteo de entradas de las bicis, ya sea por hora o por estación. Luego, obtener la afluencia sumando todas las entradas y salidas de los tres meses (por hora o por estación) luego reordenar los datos por afluencia y en base a eso contestar esta pregunta. En alguna parte del archivo "Libro(Autoguardado)" debería aparecer la afluencia (creo que por hora) solo para las bicis de diciembre.

   Me imagino que las horas con más afluencia deberían ser las horas de entradas y salidas de las personas a los trabajos o escuela, pero con aún más afluencia a las horas de salida ya que las personas tienen menos prisa y podrían optar por este transporte, en lugar de alguno más rápido.

   También me imagino que las estaciones con más afluencia deberían ser las cercanas a las ciclovías y a parques.

2. A partir de un análisis temporal:

   1) ¿En qué estaciones puedes observar una tendencia de uso a la alta?

      Este problema no lo pude resolver debido a falta de espacio en mi computadora y mala distribución del espacio en mi archivo (falta de experiencia con archivos muy grandes). Sin embargo, yo hubiera pensado que cerca de centros comerciales se podría esperar una tendencia de uso a la alta, debido a la cercanía con de estos meses con las fechas navideñas y de fin de año.

   2) ¿Puedes categorizar las estaciones con base en su tendencia de uso?

      La idea aquí era crear un vector de dimensión 2 para cada estación. La primera entrada del vector sería la diferencia de afluencias entre noviembre y octubre y la segunda entre noviembre y diciembre. Después, se podría utilizar el método de k-medias para categorizar las estaciones (por clusters).

   3) Demuestra tus conclusiones gráficamente

      Como aún no estoy familiarizada con métodos para escoger el valor k, por cuestiones de tiempo hubiera graficado todos los vectores como puntos en el espacio XY y luego buscado los

clusters “ a ojo” para decidir el valor de k y entonces poder utilizar el método de k-medias (posiblemente confirmando mis capacidades visuales ;P).

3. Por cada estación de Ecobici, identifica cómo están correlacionadas las entradas-salidas entre las otras estaciones

   Para este problema primero se crea la matriz de salidas y llegadas de diciembre. Esta es la matriz que en cada entrada ij tiene el número de bicicletas que fueron de la estación i a la estación j en el mes de diciembre. Esta matriz por suerte fue rescatada y adjunta con el nombre de “Libro1.csv”.

   Se puede pensar esta matriz como n vectores de dimensión n, donde n es el número de estaciones. Cada vector representa a una estación de salida y las entradas de dicho vector representan las estaciones a las que se dirigieron las bicis que salieron de dicho vector. En este sentido, cada estación es un vector aleatorio de dimensión n y por lo tanto tiene sentido calcularle su matriz de correlaciones.

   La matriz de correlaciones se calculó utilizando R en el archivo “Rclustersinterconecciones”.

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.

   Se utilizó un método de dendograma para categorizar, ya que debido a la alta dimensionalidad de los datos era la forma más visual de observar comportamientos similares entre las estaciones.

   A partir de la matriz de correlaciones, se creó un dendograma en el archivo llamado “PruebaHC” en el dendograma se considera a los elementos de la matriz de correlaciones como vectores y se juntan en ramas con distancias cercanas. Ahora, si se deseara categorizar a las estaciones en diferentes grupos utilizando dicho dendograma, bastaría con hacer un corte a cierta altura y los descendientes una misma rama resultante (de dicho corte) representarían a los elementos de un mismo cluster.

   Observando el archivo “PruebaHC”, parece prudente separar a las estaciones en solo 4 clusters, ya que la longitud de las ramas a la altura en la que se encuentran solo cuatro de ellas es muy grande.

En este sentido uno podría suponer que los "parientes" (desentiendes de una misma rama, resultante del corte), están muy cerca entre si, ya que tienen un "comportamiento similar". En este sentido, "comportamiento similar" significa que son estaciones que tienden a comunicarse con el mismo grupo de estaciones y por lo tanto uno puede pensar que se encuentran geográficamente cerca.

Además se puede pensar que las bicicletas solo cambian de estación a estación de un mismo cluster, ya que tienen comportamientos de dirección similares, yo pensaría que si se pusieran bicis de un mismo color por cada cluster, tardarían un buen tiempo en revolverse.

Antes de realizar el siguiente inciso, pensé equivocadamente que las cuatro ramas corresponderían a zonas muy marcadas de la CDMX (como norte, sur, este u oeste).

Nota: Para ver el archivo "PruebaHC" y observar las estaciones se requiere mucho zoom, pero se puede leer.

5. BONUS: En el sitio de Ecobici te puedes registrar para obtener URLs que regresan información sobre cada estación (Número de Slots, Latitud, Longitud). Usa la información de algunas estaciones para explicar el comportamiento de la relaciones que encontraste en la pregunta 3. Explica cómo los atributos geográficos te pueden ayudar a entender las relaciones.

   Como ya se mencionaba en la pregunta anterior, el hecho de que las estaciones tengan comportamiento parecido debe significar que estén geográficamente cerca, ya que la alta correlación significa que las bicis que salen de estaciones altamente correlacionadas tienden a desembocar en los mismos tipos de estaciones.

   Para este problema se utilizaron las coordenadas para localizar las estaciones que corresponden a cada cluster. Por lo tanto tenemos 4 mapas de estaciones(un mapa para cada cluster de estaciones). Los mapas se realizaron utilizando la página http://www.darrinward.com/.

   Las siguientes son las páginas resultantes con las coordenadas de los elementos de cada cluster:

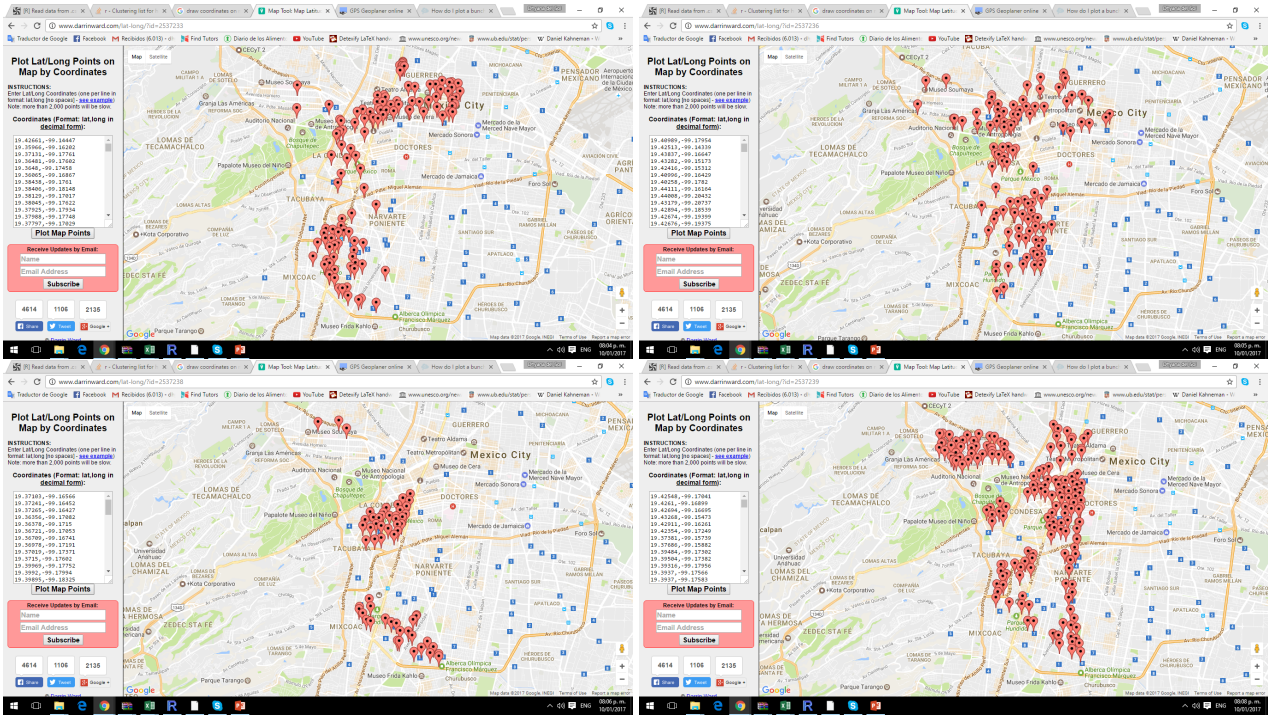
   http://www.darrinward.com/lat-long/?id=2537238

   http://www.darrinward.com/lat-long/?id=2537239

   http://www.darrinward.com/lat-long/?id=2537236

   http://www.darrinward.com/lat-long/?id=2537233

Las siguientes son capturas de pantalla de dichos mapas: Nota: Debajo de los mapas vienen observaciones de las gráficas y conclusiones finales.



En el tercer mapa se pueden observar con mucha claridad tres "montoncitos" de estaciones, esto quiere decir que en una buena clasificación de clusters cada montoncito debería pertenecer a un cluster distinto. En este sentido, la observación geográfica nos está dando más información que la observación del dendograma. En los demás mapas también se puede observar, aunque con menos claridad, una necesidad de separar en más clusters. Todo esto sugiere fuertemente que se requerirían muchos más clusters, tal vez unos 15 en lugar de solo 4.

Conclusión: Aunque se comienzan a apreciar diferencias geográficas entre los 4 clusters propuestos, se nota mucho más una necesidad de replantear el problema dividiendo esta vez en más clusters. Para poder sugerir clusters con características más marcadas posiblemente se necesitaría buscar un algoritmo apropiado que nos diera una mejor sugerencia de cuántos clusters tomar, o simplemente volver a intentar todo con 15 clusters esta vez y ver qué sucede.

Yo pienso que con más clusters, quedarían muy marcadas las características de cada categoría, ya que con solo 4 clusters ya se empiezan a apreciar los "montoncitos geográficos de estaciones".