

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Bases de Datos no Relacionales

Primer Proyecto

ECOBICIS

La Banda Gangrena

Integrantes:

Carlos Octavio Ordaz Bernal - 158525

Amanda Velasco Gallardo - 154415

Paola Mejia Domenzain - 157093

José Sánchez Aguilar - 156190

Fecha de entrega del proyecto:

16 de octubre de 2018

Índice

1. Introducción	3
2. Solución	4
2.1. Demanda del servicio	4
2.1.1. Cantidad de viajes por hora del día	4
2.1.2. Cantidad de usuarios por edad	5
2.1.3. Cantidad de viajes por edad y género	5
2.2. Perfil de viajes y usuarios	5
2.3. Mapas de visualización	5
2.3.1. Concentración cicloestaciones	5
2.3.2. Cicloestaciones multimedia	5
2.3.3. Cluster de estaciones y número de bicicletas	6
2.3.4. Devolución de Bicicletas	6
3. Características de la solución	6
4. Obtención y almacenamiento de los datos	6
5. Estructura de las tablas de la BD	7
6. Resultados	8
6.1. Demanda del servicio y clasificación de usuarios	8
6.1.1. Edad de los usuarios	8
6.1.2. Demanda del servicio por día del mes de enero	8
6.1.3. Demanda del servicio por hora del día	10
6.1.4. Viajes realizados por grupo etario y género	11
6.2. Perfil de viajes y usuarios	12
6.3. Mapas de visualización	15
6.3.1. Concentración cicloestaciones	15
6.3.2. Cicloestaciones multimedia	18
6.3.3. Cluster de estaciones y número de bicicletas	20
6.3.4. Devolución de Bicicletas	22
7. Conclusiones	24

1. Introducción

ECOBICI es un sistema de bicicletas públicas de la Ciudad de México que permite a los usuarios registrados tomar una bicicleta de cualquier cicloestación y devolverla en la más cercana a su destino.[1] Sin embargo, existen los siguientes aspectos que son relevantes de ser estudiados para mejorar el sistema:

1. Conocer la demanda de ECOBICIS por día del mes y por hora del día para asegurar que haya disponibilidad en días y horas clave.
2. Estudiar el uso que distintos grupos etarios hacen del programa para intentar volverlo atractivo y práctico a más grupos de la población.
3. A los usuarios puede parecerles importante conocer el perfil de los demás usuarios así como la afluencia a cada cicloestación con la finalidad de elegir la estación que más les convenga.
4. Visualizar la concentración y localización de las cicloestaciones en la Ciudad de México
5. Visualizar las cicloestaciones multimedia que cuentan con bicicletas eléctricas.
6. Visualizar el número de bicicletas disponibles por cicloestación o el número de lugares disponibles para regresar bicicletas.

2. Solución

Para analizar los aspectos relevantes del sistema de ECOBICIS, se elaboró un sistema de análisis de datos. La arquitectura del sistema se muestra en la figura 1

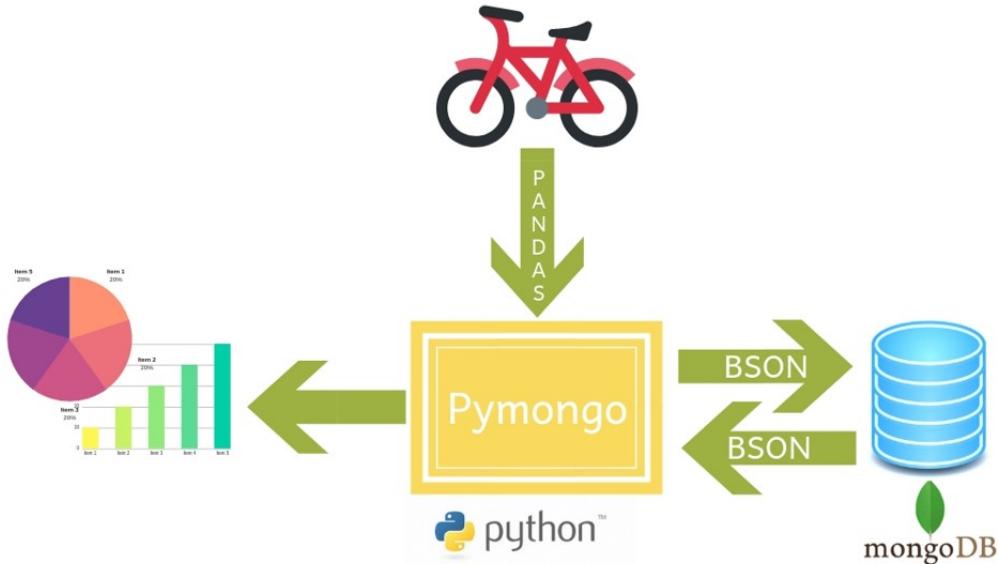


Figura 1: Diagrama mostrando la arquitectura del sistema

La bicicleta en la figura 1 representa los datos obtenidos del laboratorio de datos de la Ciudad de México.[2] Estos datos fueron importados a Python con la librería Pandas. A continuación, la librería PyMongo permitió establecer conexión con MongoDB. Se crearon colecciones en MongoDB para almacenar los datos y realizar consultas. Finalmente, se utilizaron las librerías Folium y Matplotlib así como el software R para visualizar la información resultante de las consultas a las colecciones.

2.1. Demanda del servicio

Con la extensiones NumPy y Matplotlib fue posible graficar el comportamiento de la red ECOBICI.

2.1.1. Cantidad de viajes por hora del día

Para entender la dinámica del servicio se obtuvo la cantidad de viajes realizados en un mes y se graficó contra las horas de servicio (5:00 - 00:30

del día siguiente).

2.1.2. Cantidad de usuarios por edad

Se clasificó a los usuarios en grupos de 5 años para analizar el uso que cada grupo de la población hace del sistema y por qué. Esto se graficó en un histograma para grupos que van de los 18 a los 80 años.

2.1.3. Cantidad de viajes por edad y género

Se clasificó a los usuarios por su grupo de edad y género. Se graficó la cantidad de viajes que realizaron durante un mes contra el grupo etario y género.

2.2. Perfil de viajes y usuarios

Se tomaron datos de características clave de los usuarios de cada cicloestación así como de los viajes que realizan, se agruparon por colonia y se graficaron utilizando caras de Chernoff. Con estos resultados se creó un mapa que muestra el perfil de viajes y usuarios por zona.

2.3. Mapas de visualización

Con la librería Folium fue posible crear visualizaciones geográficas de los datos para estudiar aspectos relevantes del sistema de ECOBICIS.

2.3.1. Concentración cicloestaciones

Para que los futuros usuarios conozcan la ubicación de las cicloestaciones y su concentración dentro de la Ciudad de México, se creó un mapa de calor que resalta de amarillo las zonas de la ciudad con mayor concentración y en azul aquellas con menor concentración.

2.3.2. Cicloestaciones multimedia

En un mapa de la Ciudad de México, utilizando las coordenadas de las cicloestaciones multimedia, se colocaron marcadores verdes que al darles clic, muestran el nombre de la estación. De esta forma los usuarios pueden visualizar solo las cicloestaciones con bicicletas eléctricas.

2.3.3. Cluster de estaciones y número de bicicletas

Se creó un mapa filtrando las estaciones con un número de bicicletas mayor a 20. Asimismo, para una visualización más amigable de las cicloestaciones, se agruparon por zonas con un número que indica el número de estaciones en la zona.

2.3.4. Devolución de Bicicletas

Se creó un mapa filtrando las estaciones con un número de lugares disponibles para regresar bicicletas mayor a 20. Asimismo, para una visualización más amigable de las cicloestaciones, se colocó un marcador cuyo tamaño es proporcional al número de lugares disponibles para regresar bicicletas. Es decir, mientras más lugares disponibles, mayor será el marcador.

3. Características de la solución

Para visualizar los mapas se utilizó la librería “Folium” en Python. Folium permite crear mapas de calor, agregar marcadores a mapas y crear clusters de marcadores.

Para graficar las caras de Chernoff se utilizó la librería “aplpack” en R y éstas fueron agregadas a un mapa con Photoshop.

Para realizar histogramas y gráficas de barras y de dispersión se utilizó la librería Matplotlib de Python.

4. Obtención y almacenamiento de los datos

Los datos se obtuvieron del laboratorio de datos de la Ciudad de México en formato CSV. Utilizando la librería JSON en Python, se cambió el formato a JSON eliminando la columna índice. A continuación, se utilizaron colecciones en Mongo para almacenar los datos y hacer consultas.

5. Estructura de las tablas de la BD

Los datos de la colección “bicis” tienen la siguiente estructura:

```
{'Genero_Usuario': 'M',
'Edad_Usuario': '46',
'_id': ObjectId('9rt536c222c724e3c287a3314'),
'Bici': 9990,
'Ciclo_Estacion_Retiro': 150,
'Fecha_Retiro': '01/08/2018',
'Hora_Retiro': '00:00:13',
'Ciclo_Estacion_Arribo': 179,
'Fecha_Arribo': 01/08/2018,
'Hora_Arribo': '00:23:38'}
```

Los datos en la colección “estaciones” siguen la siguiente estructura:

```
{'district': 'CUA',
'status': 'OPN',
'_id': ObjectId('5bb546c22a724e3c287a3314'),
'stationType': 'BIKE,TPV',
'slots': 27,
'lat': 19.433296,
'lon': -99.168051,
'addressNumber': 'S/N',
'bikes': 0,
'nearbyStations': '2,3,85',
'zip': 6500.0,
'address': '001 - Río Balsas-Río Sena',
'name': '1 RIO BALSAS-RIO SENA',
'id': 1}
```

6. Resultados

6.1. Demanda del servicio y clasificación de usuarios

6.1.1. Edad de los usuarios

Una de las formas de caracterizar a la población es analizando los atributos de la muestra seleccionada. En este caso, estudiamos la edad de las personas que hicieron uso del sistema ECOBICI durante el mes de enero. Para ello definimos la siguiente consulta que, utilizando la librería `matplotlib.pyplot` grafica el resultado:

```
#Histograma de las edades de los usuarios
edades = []
for doc in coll.find({}, {"_id":0,"Edad_Usuario":1}):
    edades.append(doc["Edad_Usuario"])
plt.figure()
plt.hist(edades,bins=10,range=(18,80))
plt.title("Histograma de edades",size=15)
plt.xlabel("Edad en años")
plt.ylabel("Cantidad de personas")
```

La gráfica obtenida se presenta en la siguiente figura. Podemos observar que, la mayor cantidad de usuarios tienen una edad entre 26 y 30 años. Asimismo, vemos que son menos los usuarios que tienen una edad mayor a 60 años, y no hay usuarios de menos de 18 años debido a que el sistema exige ser mayor de edad.

6.1.2. Demanda del servicio por día del mes de enero

Es de interés conocer los días en que mayor cantidad de personas hicieron uso de las bicicletas del sistema ECOBICI durante el mes de enero que es nuestro mes de estudio de acuerdo con los datos obtenidos. Haciendo uso de la librería `matplotlib.pyplot` para graficar el resultado, se definió la siguiente consulta:

```
fechaRetiro = []
cantidad = []
for i in range(1,32):
    if i < 10:
        fecha = "0"+str(i)+"/01/2018"
    else:
        fecha = str(i)+"/01/2018"
```

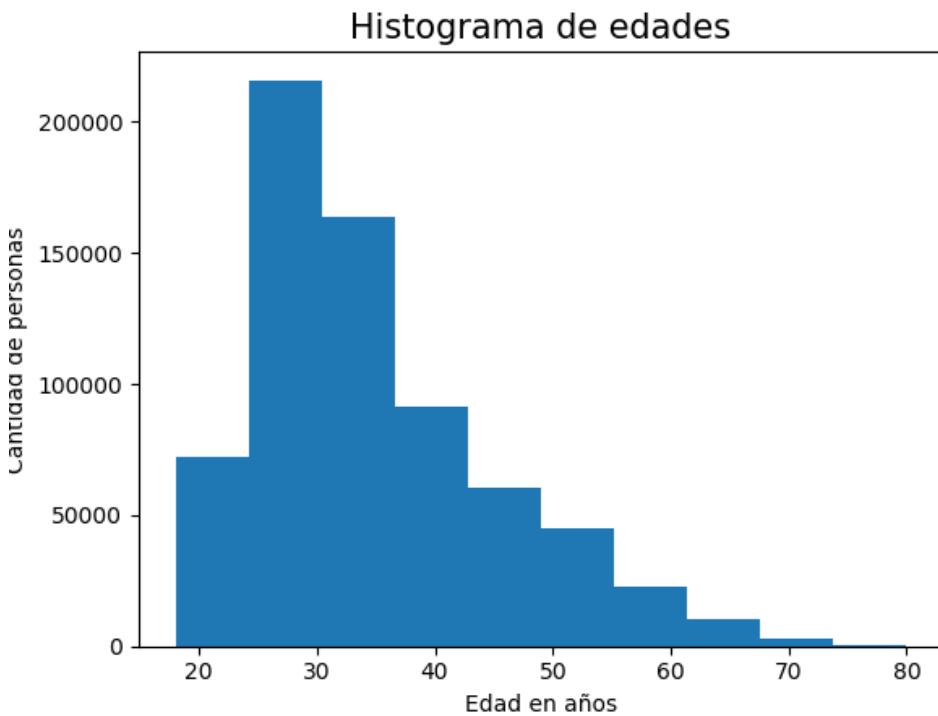


Figura 2: Cantidad de usuarios al día durante el mes de enero de 2018.

```

fechaRetiro.append(fecha)
cantidad.append(coll.find({"Fecha_Retiro":fecha},...
    ...{"id":0,"Bici":1}).count())
plt.figure()
plt.plot(fechaRetiro,cantidad,'bo',fechaRetiro,cantidad,'b')
plt.title("Cantidad de usuarios al día",size=15)
plt.xlabel("Fecha")
plt.xticks(rotation=90)
plt.ylabel("Cantidad de usuarios")
plt.grid(True)

```

La gráfica obtenida se presenta en la siguiente figura. Podemos observar que, la mayor cantidad de usuarios se registró en los días 9, 10, 11 y 23 de enero de 2018. Siendo el 10 de ese mismo mes el día donde se presenta la cantidad máxima.

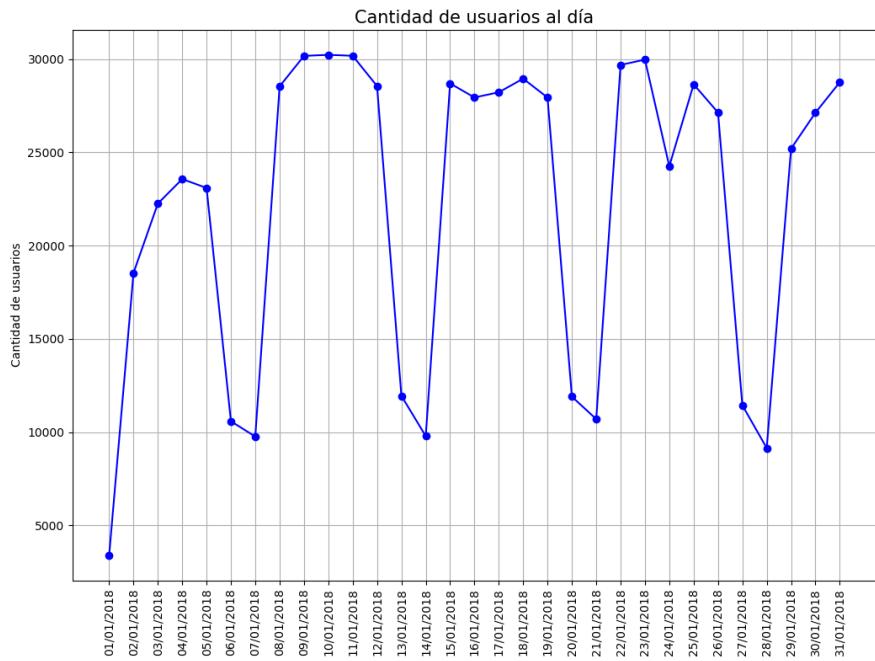


Figura 3: Cantidad de usuarios al día durante el mes de enero de 2018.

6.1.3. Demanda del servicio por hora del día

Para la obtención de los datos se recurrió al uso de la propiedad *group* de MongoDB que permite agrupar documentos en una colección y realizar funciones de agregación.

```
db.bicis.aggregate(
  [
    { '$group':
      {'_id':
        {'hour': {$convert: { input: {$arrayElemAt:[{$split: [
          '$depHour', ':']}, 0]}, to: "int" }}},
        'trips': { '$sum': 1 }
      }
    },
    {$sort:{'_id.hour':1}}
  ]
)
```

La información obtenida en la gráfica nos arrojó que, muy similar al tránsito de vehículos motorizados, el sistema de ECOBICI tiene horas 'punta' de

demandas a las 8am, 2pm y 6pm.

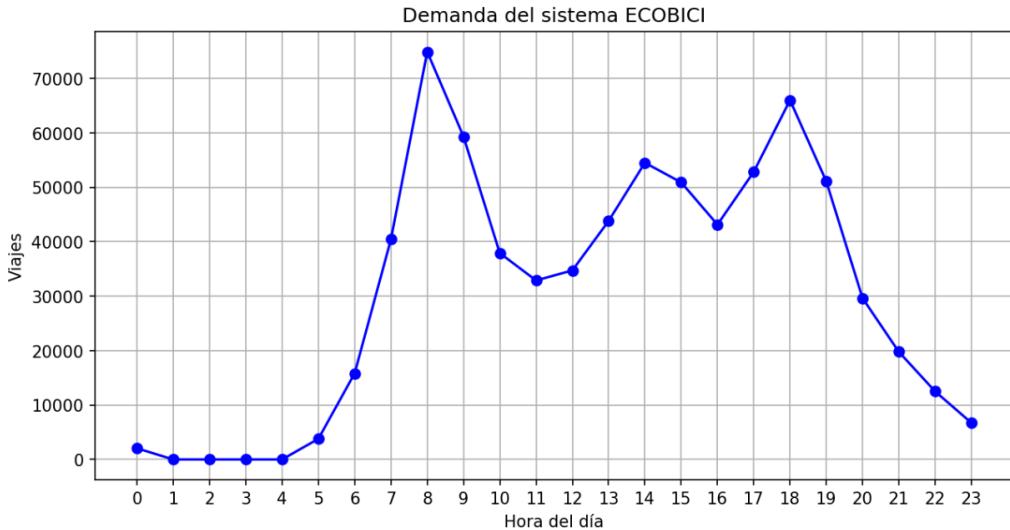


Figura 4: Demanda del servicio en el mes de agosto 2018

6.1.4. Viajes realizados por grupo etario y género

Para la obtención de los datos se recurrió al uso de la propiedad *bucket* de MongoDB la cual permite categorizar la información en varios grupos de documentos llamados *buckets*

```
db.bicis.aggregate([
  {
    $bucket:{
      groupBy : "$userAge",
      boundaries:[16,31,46,61,76],
      default:"other",
      output :{
        "total" : {$sum : 1},
        "male" : {$sum : {$cond: { if: { $eq: [ "$gender", "M" ] }, then: 1, else: 0 }}} ,
        "female" : {$sum : {$cond: { if: { $eq: [ "$gender", "F" ] }, then: 1, else: 0 }}} }
      }
    },
    {$sort:{'_id.hour':1}}
  }
])
```

Los gráficos obtenidos nos arrojaron que el 75 % de los usuarios son hombres mientras que el 25 % restante son mujeres. También se observó que los grupos de usuarios más activos son los de jóvenes y adultos en edad laboral

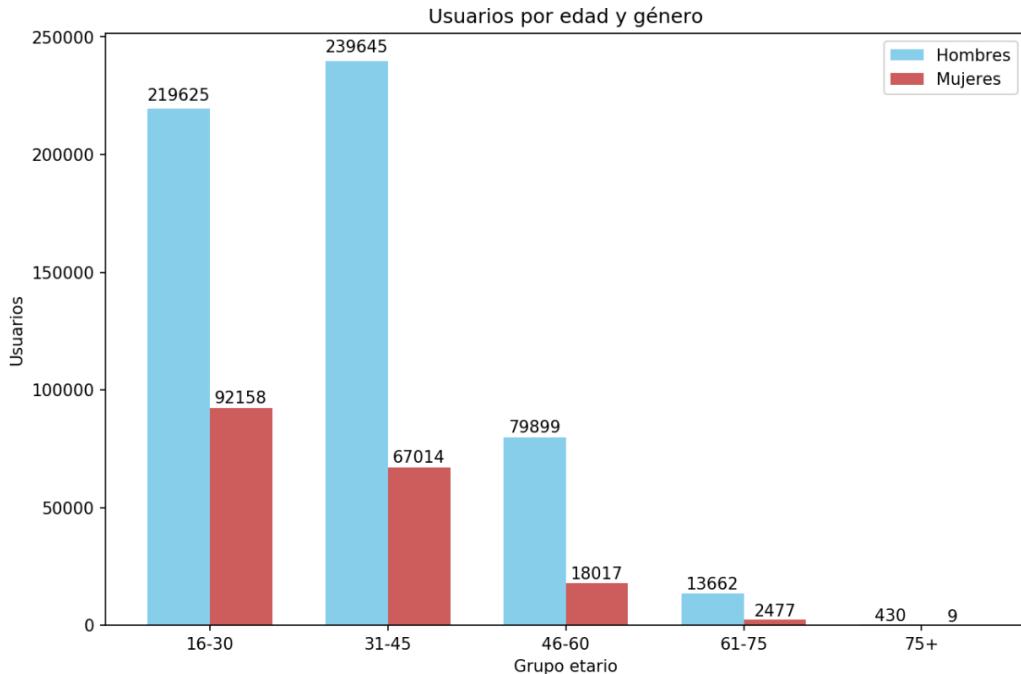


Figura 5: Demanda del servicio en el mes de agosto 2018

6.2. Perfil de viajes y usuarios

Se eligieron los siguientes parámetros para realizar un perfil del usuario y de los viajes que realiza:

- Edad promedio de los usuarios
- Clasificación por género de usuarios hombres y mujeres (valores negativos indican hombres, positivos indican mujeres)
- Cantidad promedio de viajes al día
- Longitud promedio de los viajes
- Hora del día con mayor demanda

Los perfiles se deseaban obtener por colonia, por lo que se realizaron consultas que acumularan los valores para códigos postales comunes. Para ello, primero se hizo un mapeo de cada cicloestación al código postal al que pertenece creando un diccionario con la siguiente consulta:

```
for doc in collEst.find
  ({},
  {"_id":0,"id":1, "zip":1}) :
```

Utilizando este diccionario se realizaron las siguientes consultas para obtener los datos de perfilado:

```
for doc in collBic.find(
  {"Ciclo_Estacion_Retiro":clave},
  {"Edad_Usuario":1, "Genero_Usuario":1, "Ciclo_Estacion_Arribo":1}):
  collBic.find(
    {"Hora_Retiro":{"$gte":menor,"$lt":mayor}, "Ciclo_Estacion_Retiro":clave},
    {"_id":1}).count()
```

Además, la distancia recorrida se calculó usando la fórmula del harvesine:

$$d = 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

Figura 6: Fórmula de la distancia de Haversine

donde Φ son las latitudes y λ son las longitudes.

Los resultados obtenidos se enviaron a un archivo csv para ser usado posteriormente en R para graficar. Una muestra de dichos resultados se presenta a continuación. Cabe mencionar que la columna de Colonia se rellenó manualmente pues los datos no eran parte de la base original.

CP	Colonia	Distancia	Horas	Genero	Edad	Viajes
3020	Narvarte_Poniente	60.98711751	2	-33632	35.3536588	41.77020891
3100	Del_Valle_Centro	64.70935813	1	-30188	32.57796217	48.63308308

Cuadro 1: Resultados extraídos del archivo csv

Las caras de Chernoff se graficaron con la función `faces()` en R. Se elaboró un poster en Photoshop para sobreponer cada cara generada a su posición correspondiente en el mapa de la Ciudad de México.

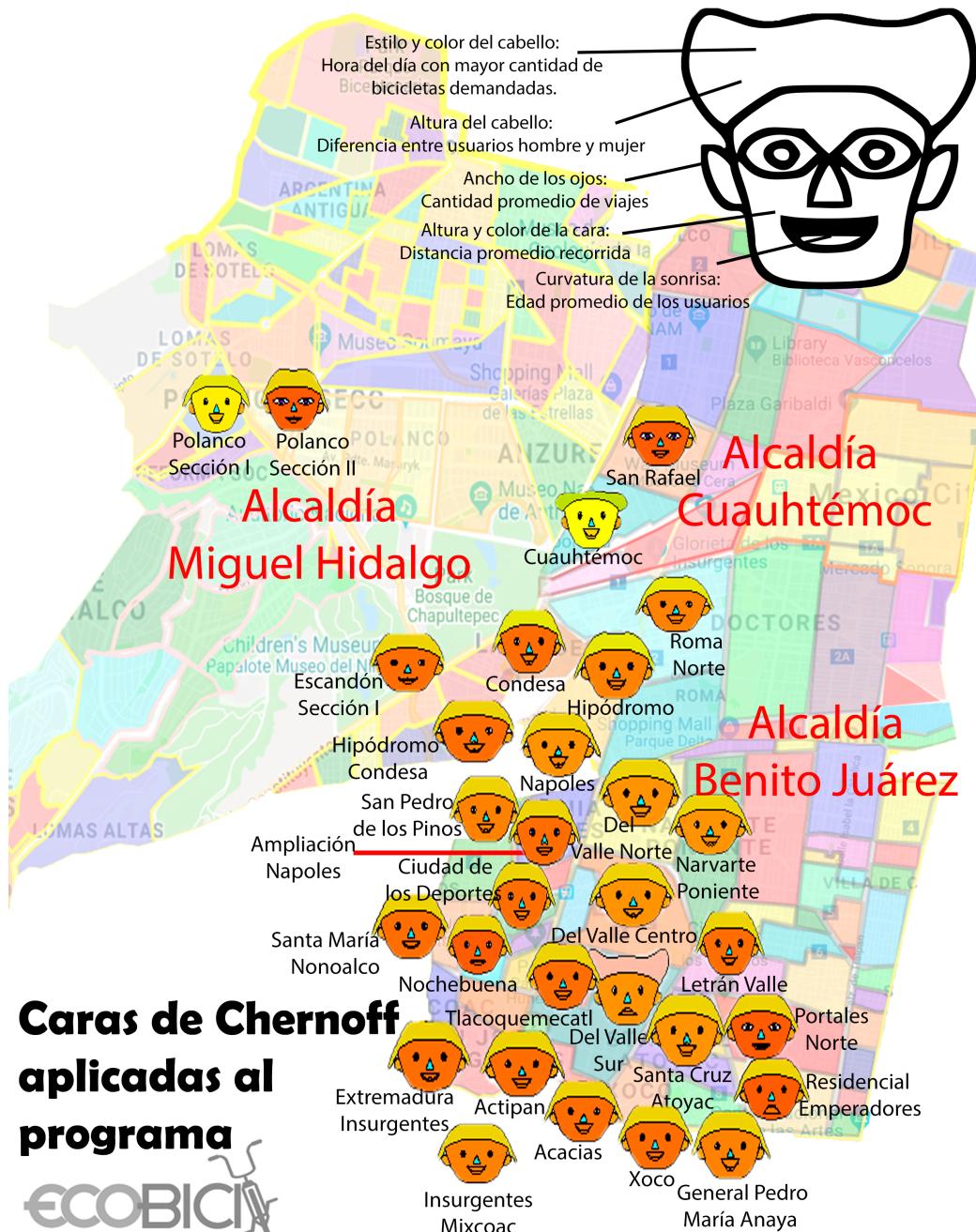


Figura 7:

Se puede observar que los perfiles de los usuarios de las cicloestaciones suelen ser bastante homogéneos aunque con ciertas excepciones notorias:

- En Polanco Sección I y Cuahtémoc las distancias recorridas por viaje son considerablemente mayores mientras que las menores están en Polanco Sección II.
- En Cuahtémoc y en especial en Del Valle Sur las horas de más alta demanda suelen ser más tarde.
- Si bien en todas las colonias hay más usuarios hombres que mujeres, en Del Valle Sur la diferencia es mucho mayor.
- En Polanco Sección II, San Rafael y Portales Norte la cantidad promedio de viajes es mucho mayor.
- En Residencial Emperadores y Del Valle Sur la edad promedio de los usuarios está muy por debajo del resto.

6.3. Mapas de visualización

6.3.1. Concentración cicloestaciones

La siguientes consultas se utilizaron para obtener los datos para construir los mapas:

```
min_lat = db.bici.find_one(sort=[("lat", 1)])["lat"]
max_lat = db.bici.find_one(sort=[("lat", -1)])["lat"]
min_lon = db.bici.find_one(sort=[("lon", 1)])["lon"]
max_lon = db.bici.find_one(sort=[("lon", -1)])["lon"]
```

Los resultados de las consultas arrojan las coordenadas máximas y mínimas de las cicloestaciones. Los resultados fueron los siguientes:

```
min_lat = 19.35827
max_lat = 19.444031
min_lon = -99.207808
max_lon = -99.130918
```

A continuación, se calculó el punto medio para crear el mapa con la siguiente instrucción:

```
cdmx_calor = folium.Map(location=[cen_lat,cen_lon],zoom_start = 12)
```

El mapa de calor de la figura 8 se elaboró consultando todos los elementos de la colección con la siguiente consulta:

```
for doc in coll.find():
    lon = doc["lon"]
    lat = doc["lat"]
    print(lon,lat)
```

Dando como resultado las coordenadas de las cicloestaciones que se pueden ver en la figura 8.

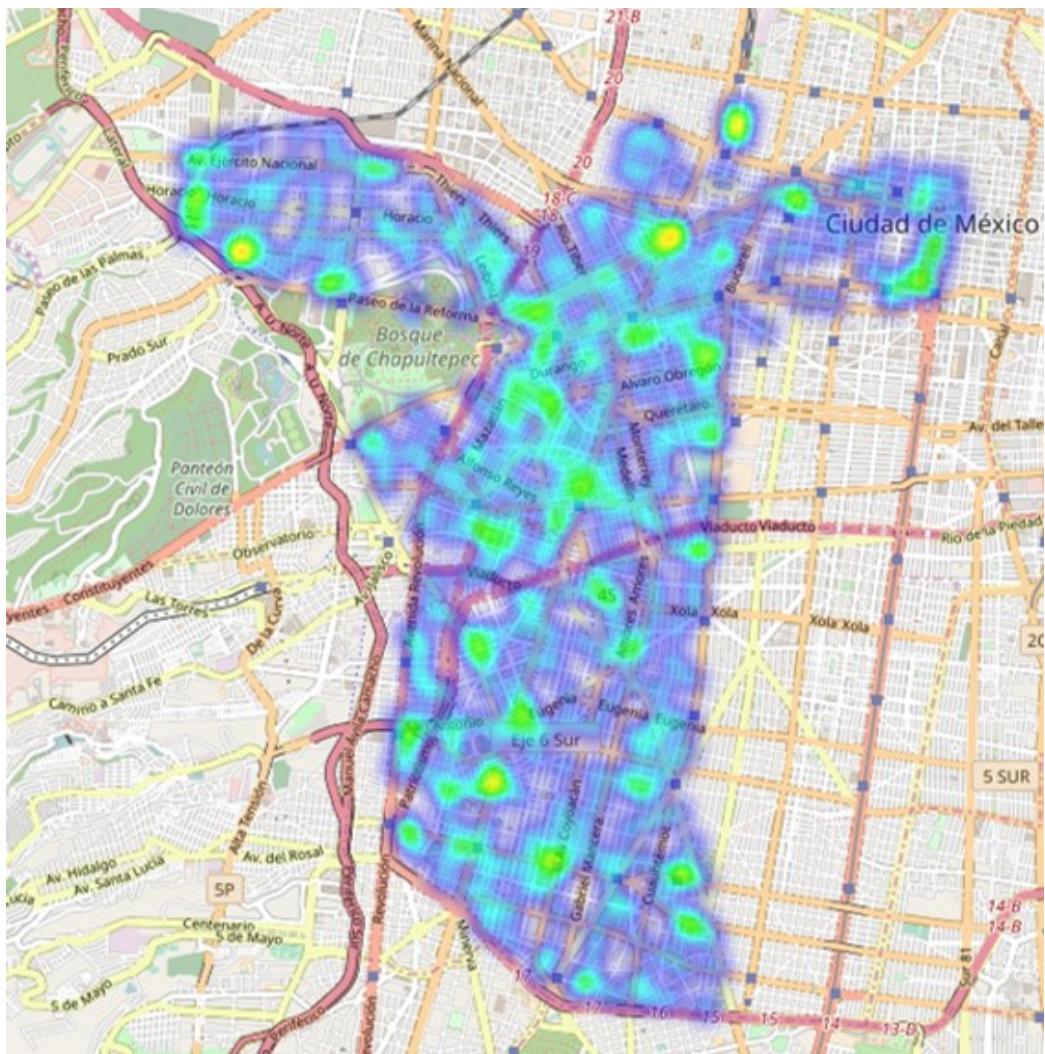


Figura 8: Mapa de calor

6.3.2. Cicloestaciones multimedia

La siguiente consulta se utilizó para obtener los datos:

```
for doc in coll.find
({'stationType' : {'$regex' : ".*TPV.*"}},
{'stationType':1, 'name':1, 'lat':1,'lon':1, '_id':0}):
```

Una muestra de los resultados de la consulta se muestra a continuación:

```
{'stationType': 'BIKE,TPV', 'lon': -99.168051, 'lat': 19.433296,
'name': '1 RIO BALSAS-RIO SENA'}
{'stationType': 'BIKE,TPV', 'lon': -99.158668, 'lat': 19.431655,
'name': '3 REFORMA-INSURGENTES'}
{'stationType': 'BIKE,TPV', 'lon': -99.154752, 'lat': 19.433321,
'name': '10 REFORMA-RAMIREZ'}
```

La figura 9 muestra los pinos colocados en las coordenadas obtenidas de la consulta.

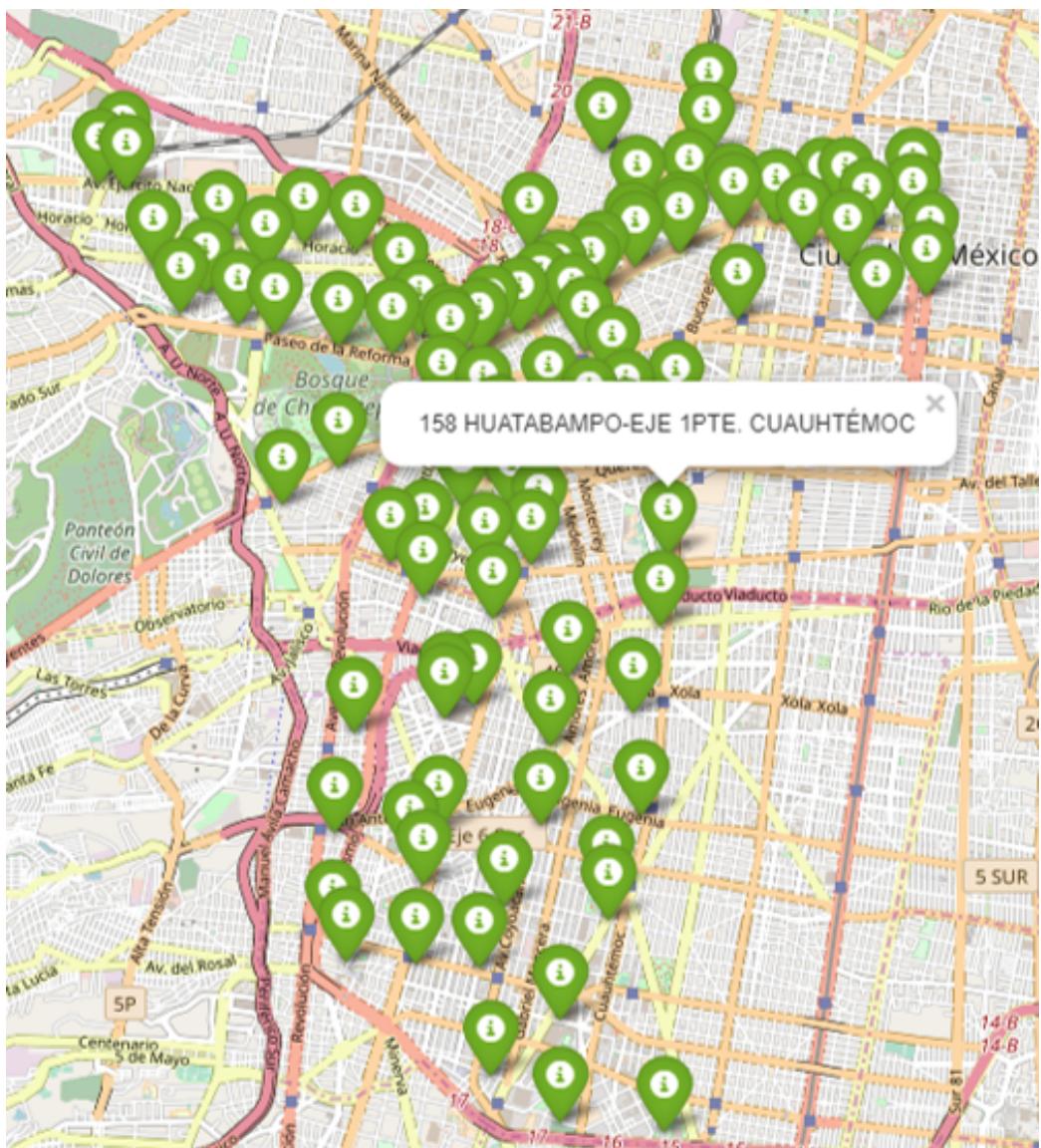


Figura 9: Pines

6.3.3. Cluster de estaciones y número de bicicletas

La siguiente consulta se utilizó para obtener el número de estaciones con más de 20 bicicletas.

```
resp = db.bici.count({"bikes": {"$gte": 20}})
```

A continuación, se usó la siguiente consulta para obtener la información de las 41 estaciones con 20 bicicletas o más:

```
for doc in coll.find({"bikes": {"$gte": 20}},  
{'bikes':1, 'name':1, 'lat':1,'lon':1, '_id':0}):
```

Una muestra de los resultados de la consulta se muestra a continuación:

```
{'lon': -99.154752, 'lat': 19.433321,  
'name': '10 REFORMA-RAMIREZ', 'bikes': 34}  
{'lon': -99.169164, 'lat': 19.42653,  
'name': '16 REFORMA-RIO GUADALQUIVIR', 'bikes': 23}  
{'lon': -99.162614, 'lat': 19.429115,  
'name': '27 REFORMA-HAVRE', 'bikes': 20}
```

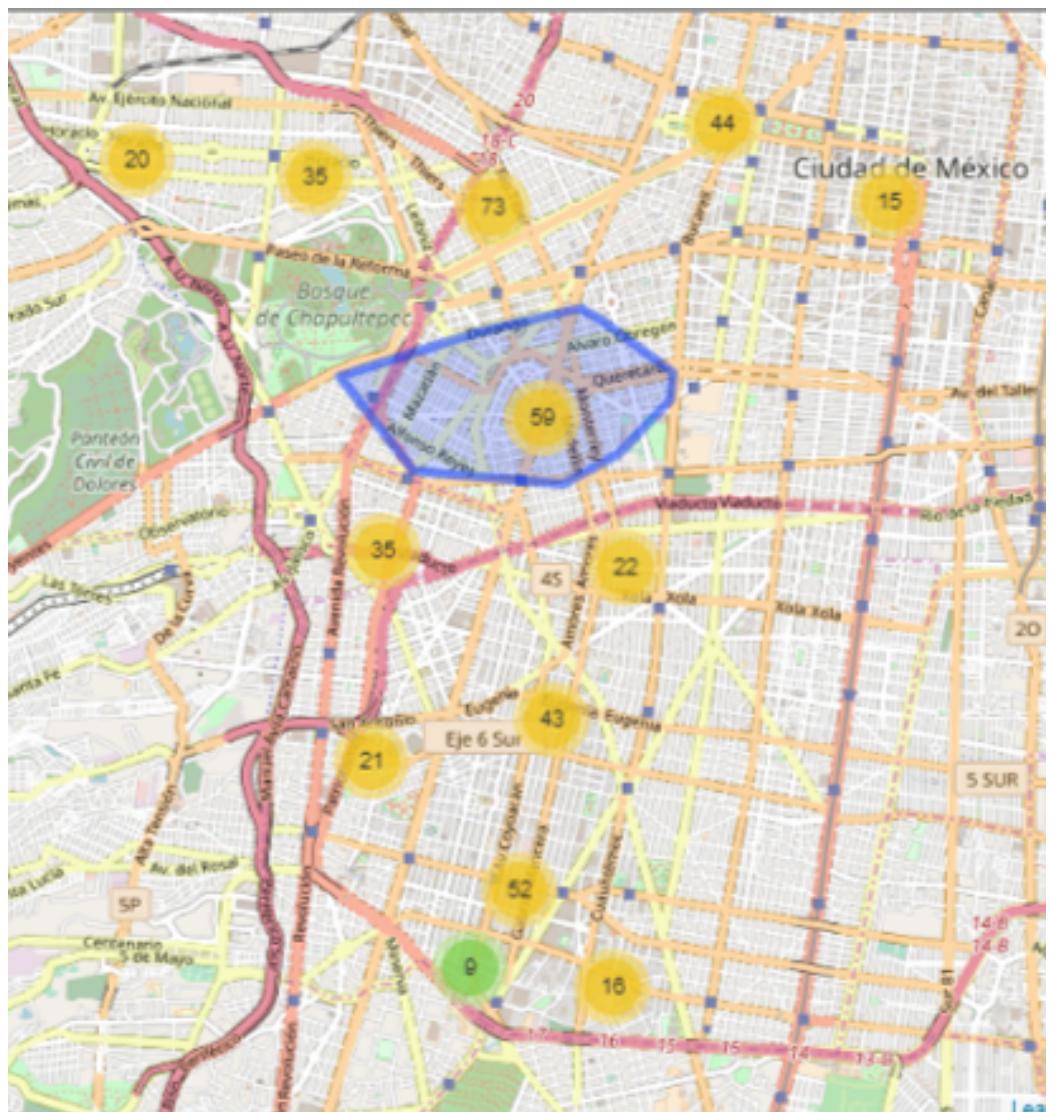


Figura 10:

6.3.4. Devolución de Bicicletas

La siguiente consulta se utilizó para obtener los datos:

```
for doc in coll.find
  {"slots": {"$gte": 20}},
  {'slots': 1, 'name': 1, 'lat': 1, 'lon': 1, '_id': 0} :
```

Una muestra de los resultados de la consulta se presenta a continuación:

```
{'lon': -99.168051, 'lat': 19.433296,
'name': '1 RIO BALSAS-RIO SENA', 'slots': 27}
{'lon': -99.158668, 'lat': 19.431655,
'name': '3 REFORMA-INSURGENTES', 'slots': 32}
{'lon': -99.175166, 'lat': 19.425468,
'name': '7 RIO ELBA-RIO LERMA', 'slots': 24}
```

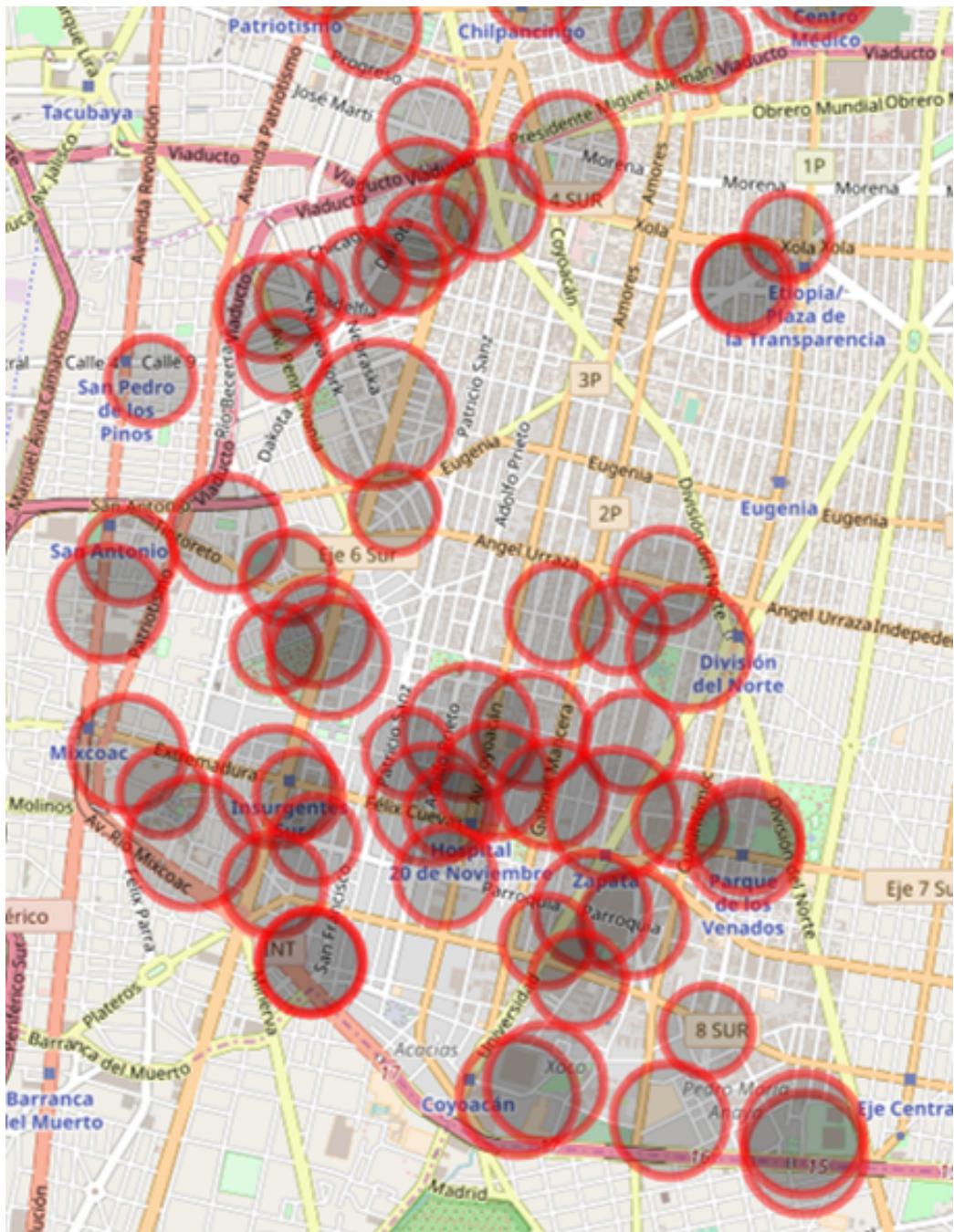


Figura 11:

7. Conclusiones

Fue posible elaborar un sistema para analizar los datos de ECOBICI y visualizar la información a través de mapas y gráficas para resolver dificultades de los usuarios o futuros usuarios.

La gran cantidad de datos fue almacenada en una colección en MongoDB como documentos de JSON. Esto permite que el sistema pueda escalar y en un futuro se puedan visualizar y analizar los datos de futuras estaciones de ECOBICI dentro de la Ciudad de México o incluso dentro del país.

Los resultados nos permiten hacer sugerencias sobre las acciones concretas que se deben tomar en cuenta si se desea extender la cobertura y alcance del programa. Por ejemplo: podemos observar en el histograma de las edades de los usuarios que se presenta un sesgo a la derecha, una de las sugerencias que podemos realizar es ampliar la cobertura del programa, por medio de campañas de conciencia ambiental o haciendo más flexibles los requisitos de inscripción, para que más sectores de la población puedan acceder a éste. También, se pueden instalar más estaciones en las diferentes alcaldías con el propósito de aumentar el número de personas que utilizan bicicleta en vez de automóviles.

Sin embargo, es importante resaltar que la base de datos con la que se trabajó tenía datos faltantes, insuficientes o inconsistentes en algunos campos. En particular, se encontraron varias cicloestaciones cuyos números de identificación no correspondían a ningún código postal o de las cuales no se tenía registro en la colección de estaciones. Afortunadamente estos casos fueron mínimos y no impidieron llevar a cabo un análisis efectivo. También se encontraron datos falsos en lo que respecta a las edades de los usuarios puesto que en dos casos se obtuvieron edades promedio menores a los 18 años, lo cual es claramente incongruente.

Referencias

- [1] G. de la Ciudad de México, “¿qué es ecobici?” 2018, [Accedido el 15-10-2018]. [Online]. Available: <https://www.ecobici.cdmx.gob.mx/es/informacion-del-servicio/que-es-ecobici>
- [2] L. para la Ciudad, “Ecobici,” 2016, [Accedido el 15-10-2018]. [Online]. Available: <https://github.com/LabPLC/api.labcd.mx/wiki/Ecobici>