

Bicimad

ENRIQUE DE ALVEAR , LAURA CANO , LUCÍA ROLDÁN

May 2023

1. Motivación y Descripción del problema

Con objetivo de disminuir el número de accidentes derivados de la interacción de las bicicletas con el tráfico, se quiere construir un carril bici en los trayectos más frecuentados. Con el fin de facilitar el servicio durante el tiempo que dure la construcción, en los recorridos más frecuentados y en hora punta, se reservará un carril del tráfico a la circulación de los ciclos. Por último, de cara a una implementación a largo plazo, se continuará con la construcción del carril bici en las zonas donde residan un mayor número de usuarios de estas biciletas.

Por ello, con objetivo de la resolución de este problema se estudiarán los siguientes datos:

- Los 100 trayectos más frecuentados
- La hora más frecuentada en los trayectos anteriores
- La franja de edad donde más se usan las bicicletas eléctricas.

Cabe destacar que sólo se excluirán los usuarios ocasionales o trabajadores de la empresa en ese estudio.

2. Descripción de los datos

Para abordar el problema usaremos los datos proporcionados por la Comunidad de Madrid de los itinerarios de las bicicletas eléctricas durante junio de 2019.

Los datos se encuentran almacenados en un archivo .json donde cada línea consiste en un diccionario con las siguientes claves e información asociada:

- `_id`: Identificador del movimiento.
- `user_day_code`: Código del usuario. Para una misma fecha, todos los movimientos de un mismo usuario, tendrán el mismo código, con el fin de poder realizar estudios estadísticos de las tendencias diarias de los usuarios.
- `idunplug_station`: Número de la estación de la que se desengancha la bicicleta.
- `idunplug_base`: Número de la base de la que se desengancha la bicicleta.

- `idplug_station`: Número de la estación en la que se engancha la bicicleta.
- `idplug_base`: Número de la base en la que se engancha la bicicleta.
- `unplug_hourTime`: Franja horaria en la que se realiza el desenganche de la bicicleta. Por cuestiones de anonimato, se facilita la hora de inicio del movimiento, sin la información de minutos y segundos. Todos los movimientos iniciados durante misma hora, tendrán el mismo dato de inicio.
- `travel_time`: Tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.
- `user_type`: Número que indica el tipo de usuario que ha realizado el movimiento. Sus posibles valores son:
 - 0: No se ha podido determinar el tipo de usuario
 - 1: Usuario anual (poseedor de un pase anual)
 - 2: Usuario ocasional
 - 3: Trabajador de la empresa
- `ageRange`: Número que indica el rango de edad del usuario que ha realizado el movimiento. Sus posibles valores son:
 - 0: No se ha podido determinar el rango de edad del usuario
 - 1: El usuario tiene entre 0 y 16 años
 - 2: El usuario tiene entre 17 y 18 años
 - 3: El usuario tiene entre 19 y 26 años
 - 4: El usuario tiene entre 27 y 40 años
 - 5: El usuario tiene entre 41 y 65 años
 - 6: El usuario tiene 66 años o más
- `zip_code`: Texto que indica el código postal del usuario que ha realizado el movimiento.

3. Proceso de diseño e implementación de la solución

Para abordar este problema se aplicará el estudio de los datos mediante el archivo `Practica_BiciMad.py` para el mes de Junio 2019. En este se llevan a cabo los siguientes pasos. (Estos pasos se encuentran explicados con detalle en el propio código, a continuación se expone únicamente una idea general) :

- Se accede a los datos y se eliminan aquellos relacionados con los usuarios ocasionales y con los trabajadores de la empresa. También se prescinde de los trayectos en los que la estación de recogida de la bicicleta coincide con la estación de destino, puesto que estos datos no aportarán información acerca de los trayectos más frecuentados.

- Se obtienen los trayectos más realizados a lo largo del mes, de los cuales seleccionamos los 100 mejores. Cabe destacar que este trayecto no diferencia la dirección, sólo el recorrido.
- Por último, de estos 100 trayectos, obtenemos su hora más frecuentada.

4. Resultados

- Los trayectos más frecuentados por los usuarios a lo largo del mes junto a sus frecuencias han sido: [(9, 149), 490], [(49, 135), 434], [(132, 135), 405], [(64, 78), 402], [(149, 169), 380], [(149, 163), 358], [(130, 149), 357], [(157, 163), 322], [(132, 175), 317], [(58, 149), 312], [(26, 175), 310], [(135, 175), 310], [(129, 135), 288], [(57, 135), 284], [(64, 90), 282], [(47, 129), 281], [(52, 135), 280], [(27, 175), 271], [(9, 157), 268], [(118, 163), 266], [(38, 57), 262], [(86, 90), 261], [(1, 175), 260], [(9, 163), 258], [(38, 129), 257], [(129, 175), 253], [(84, 90), 252], [(160, 163), 251], [(42, 135), 249], [(48, 129), 246], [(49, 129), 244], [(57, 129), 243], [(56, 175), 240], [(129, 134), 237], [(83, 115), 236], [(162, 163), 235], [(153, 163), 235], [(130, 157), 229], [(133, 135), 228], [(132, 174), 227], [(149, 168), 226], [(169, 170), 226], [(59, 149), 224], [(45, 135), 223], [(42, 175), 221], [(57, 128), 215], [(79, 90), 213], [(45, 129), 213], [(160, 161), 211], [(26, 57), 210], [(38, 175), 209], [(134, 135), 209], [(30, 157), 206], [(73, 90), 206], [(50, 135), 205], [(19, 163), 204], [(19, 90), 201], [(38, 135), 201], [(30, 163), 198], [(58, 157), 198], [(59, 168), 198], [(25, 175), 198], [(164, 171), 197], [(163, 170), 197], [(145, 161), 195], [(64, 84), 194], [(19, 55), 193], [(64, 79), 192], [(139, 149), 192], [(19, 161), 191], [(59, 163), 191], [(149, 160), 191], [(83, 90), 190], [(118, 168), 190], [(163, 166), 187], [(6, 149), 186], [(129, 133), 186], [(74, 79), 185], [(49, 128), 185], [(78, 90), 185], [(90, 166), 184], [(74, 83), 181], [(49, 55), 181], [(139, 145), 181], [(26, 42), 179], [(65, 84), 179], [(59, 131), 178], [(13, 149), 178], [(57, 136), 177], [(149, 164), 177], [(13, 160), 174], [(139, 154), 173], [(1, 41), 173], [(130, 156), 173], [(149, 161), 172], [(49, 175), 171], [(42, 132), 170], [(42, 57), 169], [(139, 163), 168], [(78, 86), 168]]. Donde cada tupla representa el par de estaciones del trayecto y el número asociado a su frecuencia. Se pueden observar los 10 primeros con más claridad en la siguiente tabla:

	estacion	numerosviajes
0	[9, 149]	490
1	[49, 135]	434
2	[132, 135]	405
3	[64, 78]	402
4	[149, 169]	380
5	[149, 163]	358
6	[130, 149]	357
7	[157, 163]	322
8	[132, 175]	317
9	[58, 149]	312

- Las horas de más tránsito en cada uno de estos trayectos han sido: [((9, 157), (35, '21:00:00.000+0200')), ((42, 132), (22, '00:00:00.000+0200')), ((9, 163), (60, '08:00:00.000+0200')), ((25, 175), (25, '18:00:00.000+0200')), ((13, 149), (20, '19:00:00.000+0200')), ((64, 90), (40, '08:00:00.000+0200')), ((38, 135), (18, '16:00:00.000+0200')), ((129, 134), (32, '20:00:00.000+0200')), ((50, 135), (23, '20:00:00.000+0200')), ((149, 164), (21, '09:00:00.000+0200')), ((160, 161), (29, '16:00:00.000+0200')), ((134, 135), (24, '19:00:00.000+0200')), ((149, 161), (22, '19:00:00.000+0200')), ((145, 161), (48, '08:00:00.000+0200')), ((19, 163), (40, '08:00:00.000+0200')), ((59, 131), (24, '00:00:00.000+0200')), ((64, 84), (38, '08:00:00.000+0200')), ((129, 175), (34, '19:00:00.000+0200')), ((42, 57), (24, '19:00:00.000+0200')), ((73, 90), (35, '09:00:00.000+0200')), ((59, 168), (24, '20:00:00.000+0200')), ((42, 135), (32, '16:00:00.000+0200')), ((149, 160), (22, '19:00:00.000+0200')), ((169, 170), (27, '10:00:00.000+0200')), ((65, 84), (30, '07:00:00.000+0200')), ((157, 163), (55, '14:00:00.000+0200')), ((57, 135), (32, '20:00:00.000+0200')), ((78, 86), (22, '09:00:00.000+0200')), ((45, 129), (25, '18:00:00.000+0200')), ((59, 163), (23, '18:00:00.000+0200')), ((49, 175), (19, '21:00:00.000+0200')), ((19, 55), (37, '09:00:00.000+0200')), ((130, 156), (36, '09:00:00.000+0200')), ((48, 129), (29, '07:00:00.000+0200')), ((130, 157), (29, '19:00:00.000+0200')), ((118, 163), (31, '19:00:00.000+0200')), ((58, 149), (41, '20:00:00.000+0200')), ((42, 175), (34, '17:00:00.000+0200')), ((163, 166), (17, '20:00:00.000+0200')), ((139, 163), (27, '19:00:00.000+0200')), ((49, 135), (45, '14:00:00.000+0200')), ((78, 90), (24, '09:00:00.000+0200')), ((149, 169), (41, '15:00:00.000+0200')), ((135, 175), (39, '08:00:00.000+0200')), ((27, 175), (41, '19:00:00.000+0200')), ((139, 145), (26, '19:00:00.000+0200')), ((1, 41), (20, '18:00:00.000+0200')), ((57, 129), (29, '20:00:00.000+0200')), ((49, 55), (18, '00:00:00.000+0200')), ((57, 128), (29, '21:00:00.000+0200')), ((139, 154), (48, '08:00:00.000+0200')), ((163, 170), (23, '20:00:00.000+0200')), ((26, 175), (36, '19:00:00.000+0200')), ((149, 168), (30, '14:00:00.000+0200')), ((129, 135), (35, '19:00:00.000+0200')), ((49, 129), (31, '19:00:00.000+0200')), ((139, 149), (26, '21:00:00.000+0200')), ((59, 149), (23, '19:00:00.000+0200')), ((64, 78), (58, '07:00:00.000+0200')), ((90, 166), (22, '21:00:00.000+0200')), ((52, 135), (29, '21:00:00.000+0200')), ((19, 90), (20, '16:00:00.000+0200')), ((38, 57), (32, '20:00:00.000+0200')), ((64, 79), (20, '16:00:00.000+0200')), ((130, 149), (69, '18:00:00.000+0200')), ((74, 79), (25, '15:00:00.000+0200')), ((79, 90), (31, '09:00:00.000+0200')), ((26, 57), (27, '17:00:00.000+0200')), ((49, 128), (21, '18:00:00.000+0200')), ((56, 175), (33, '18:00:00.000+0200')), ((58, 157), (25, '19:00:00.000+0200')), ((26, 42), (21, '21:00:00.000+0200')), ((83, 115), (45, '16:00:00.000+0200')), ((9, 149), (89, '18:00:00.000+0200')), ((153, 163), (38, '19:00:00.000+0200')), ((86, 90), (28, '09:00:00.000+0200')), ((1, 175), (37, '19:00:00.000+0200')), ((47, 129), (48, '16:00:00.000+0200')), ((57, 136), (17, '21:00:00.000+0200')), ((30, 157), (39, '09:00:00.000+0200')), ((38, 129), (22, '20:00:00.000+0200')), ((164, 171), (48, '10:00:00.000+0200')), ((132, 135), (48, '19:00:00.000+0200')), ((132, 174), (37, '09:00:00.000+0200')), ((129, 133), (23, '18:00:00.000+0200')), ((84, 90), (34, '08:00:00.000+0200')), ((133, 135), (25, '08:00:00.000+0200')), ((118, 168), (21, '08:00:00.000+0200')), ((149, 163), (43, '20:00:00.000+0200')), ((19, 161), (35, '23:00:00.000+0200')), ((45, 135), (21, '23:00:00.000+0200')), ((83, 90), (21, '22:00:00.000+0200')), ((160, 163), (29, '18:00:00.000+0200')), ((13, 160), (32, '14:00:00.000+0200')), ((6, 149), (24, '19:00:00.000+0200')), ((38, 175), (30, '19:00:00.000+0200')), ((132, 175), (42, '18:00:00.000+0200')), ((74, 83), (22, '21:00:00.000+0200')), ((30, 163), (41, '09:00:00.000+0200')), ((162, 163), (31, '18:00:00.000+0200'))] Cada tupla cuenta con dos subtuplas, la primera de ellas representa el trayecto y la segunda viene determinada por la frecuencia máxima de dicho trayecto y la hora en la que hay

un mayor tránsito de bicicletas en dicho trayecto. Podemos representar estos datos de manera gráfica. Por ejemplo, podemos observar la frecuencia total del trayecto (9,149) durante cada hora del día en el mes de Junio.



- Por último observamos el número de usuarios de los diferentes grupos de edad que usan las bicicletas eléctricas en el mes de Junio.



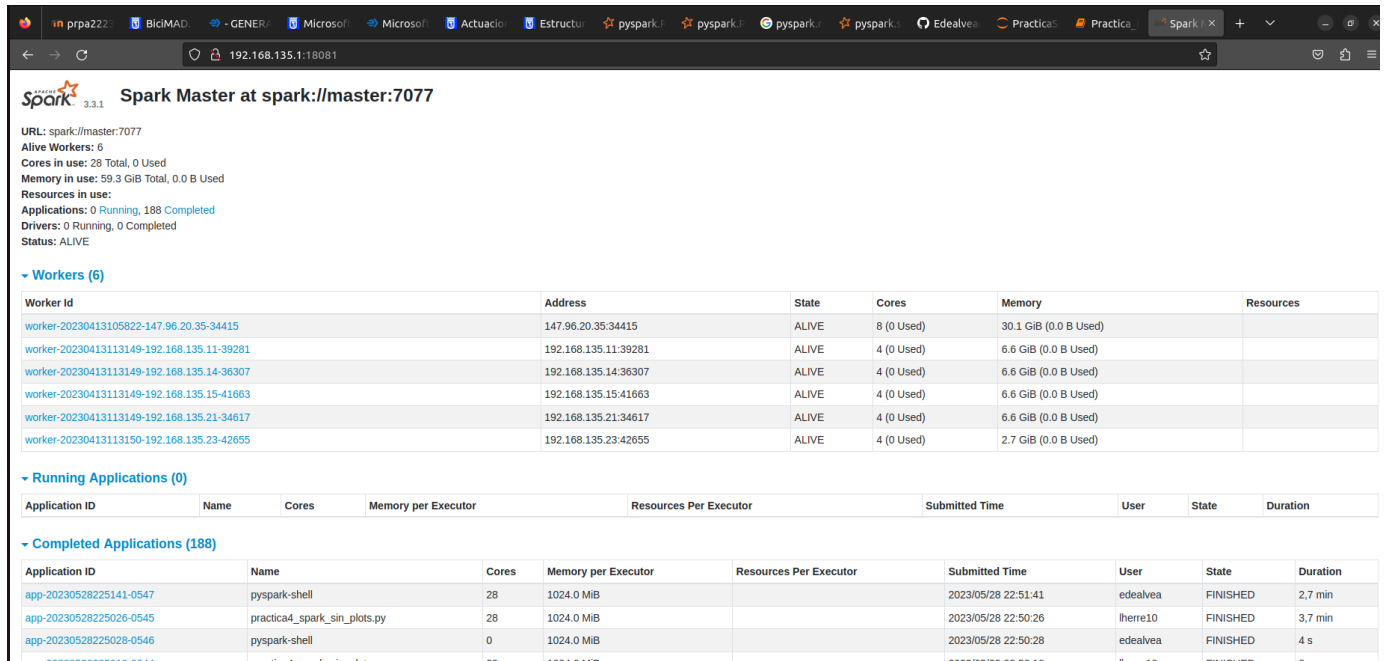
Concluyendo que la población que más usa las bicicletas eléctricas (grupo4) se encuentra en un rango de edad entre 27 y 40 años. (Dejando de lado el grupo 0 que abarca los casos en los cuales no se ha podido determinar el rango de edad).

5. Conclusiones

Como solución final al problema propuesto se sugiere comenzar la construcción del carril bici en los segmentos entre las estaciones 9 y 149 y reservar a las 21h el uso de un carril exclusivo para ciclos. Según el presupuesto se recomienda ir construyendo en el orden indicado en la tabla proporcionada y en caso de que fuera necesario, reservar los carriles en las horas indicadas. Por último, respecto a las zonas donde debe fomentarse la instauración del carril bici, convendría ampliarlo a aquellas donde viva más población de entre 27 y 40 años.

6. Anexo del Cluster

Cabe destacar que se ha ejecutado el archivo con los datos del cluster, se adjunta imagen de su ejecución:



The screenshot shows the Spark Master web interface at `spark://master:7077`. The interface displays the following information:

- URL:** `spark://master:7077`
- Alive Workers:** 6
- Cores in use:** 28 Total, 0 Used
- Memory in use:** 59.3 GiB Total, 0.0 B Used
- Resources in use:**
- Applications:** 0 Running, 188 Completed
- Drivers:** 0 Running, 0 Completed
- Status:** ALIVE

Workers (6)

Worker Id	Address	State	Cores	Memory	Resources
worker-20230413105822-147.96.20.35-34415	147.96.20.35:34415	ALIVE	8 (0 Used)	30.1 GiB (0.0 B Used)	
worker-20230413113149-192.168.135.11-39281	192.168.135.11:39281	ALIVE	4 (0 Used)	6.6 GiB (0.0 B Used)	
worker-20230413113149-192.168.135.14-36307	192.168.135.14:36307	ALIVE	4 (0 Used)	6.6 GiB (0.0 B Used)	
worker-20230413113149-192.168.135.15-41663	192.168.135.15:41663	ALIVE	4 (0 Used)	6.6 GiB (0.0 B Used)	
worker-20230413113149-192.168.135.21-34617	192.168.135.21:34617	ALIVE	4 (0 Used)	6.6 GiB (0.0 B Used)	
worker-20230413113150-192.168.135.23-42655	192.168.135.23:42655	ALIVE	4 (0 Used)	2.7 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (188)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230528225141-0547	pyspark-shell	28	1024.0 MiB		2023/05/28 22:51:41	edealvea	FINISHED	2,7 min
app-20230528225026-0545	practica4_spark_sin_plots.py	28	1024.0 MiB		2023/05/28 22:50:26	lherre10	FINISHED	3,7 min
app-20230528225028-0546	pyspark-shell	0	1024.0 MiB		2023/05/28 22:50:28	edealvea	FINISHED	4 s
app-20230528225010-0544	practica4_spark_sin_plots.py	28	1024.0 MiB		2023/05/28 22:50:10	lherre10	FINISHED	2 s

7. Anexo: Uso de varios ficheros

Por último hemos realizado el mismo estudio utilizando información de varios ficheros mediante el uso del union de rdd (puede verse su implementación en el archivo .py). Aplicando el mismo estudio sobre los datos de los meses de Enero a Junio de 2019 se obtienen los siguientes resultados:

- Las horas de más tránsito en cada uno de estos trayectos han sido:

```

Los trayectos y horas seleccionadas son:
(Formato de los datos -> ((Inicio, Final), (frecuencia_de_viajes, hora_con_mas_afluencia)))
(((149, 161), (20, '17:00:00.000+0100')), ((35, 175), (21, '08:00:00.000+0100')), ((64, 78), (41, '14:00:00.000+0100')), ((108, 113), (23, '14:00:00.000+0100')), ((74, 83), (23, '19:00:00.000+0100')), ((129, 134), (34, '19:00:00.000+0100')), ((26, 175), (29, '19:00:00.000+0100')), ((129, 135), (13, '15:00:00.000+0100')), ((146, 162), (23, '21:00:00.000+0100')), ((57, 67), (26, '07:00:00.000+0100')), ((1, 175), (21, '08:00:00.000+0100')), ((129, 175), (24, '19:00:00.000+0100')), ((132, 175), (14, '14:00:00.000+0100')), ((57, 128), (18, '21:00:00.000+0100')), ((130, 157), (27, '08:00:00.000+0100')), ((9, 149), (38, '18:00:00.000+0100')), ((149, 169), (36, '08:00:00.000+0100')), ((145, 161), (24, '08:00:00.000+0100')), ((129, 133), (21, '19:00:00.000+0100')), ((90, 99), (24, '16:00:00.000+0100')), ((84, 129), (26, '18:00:00.000+0100')), ((30, 163), (18, '09:00:00.000+0100')), ((57, 135), (19, '18:00:00.000+0100')), ((6, 62), (19, '20:00:00.000+0100')), ((139, 145), (18, '19:00:00.000+0100')), ((160, 168), (19, '15:00:00.000+0100')), ((79, 113), (34, '08:00:00.000+0100')), ((64, 79), (25, '09:00:00.000+0100')), ((40, 129), (19, '20:00:00.000+0100')), ((130, 149), (35, '09:00:00.000+0100')), ((56, 175), (12, '22:00:00.000+0100')), ((149, 160), (18, '14:00:00.000+0100')), ((145, 156), (30, '08:00:00.000+0100')), ((64, 83), (11, '20:00:00.000+0100')), ((38, 57), (26, '20:00:00.000+0100')), ((31, 43), (16, '16:00:00.000+0100')), ((75, 103), (24, '19:00:00.000+0100')), ((9, 157), (18, '20:00:00.000+0100')), ((43, 175), (22, '16:00:00.000+0100')), ((49, 129), (17, '20:00:00.000+0100')), ((43, 135), (17, '18:00:00.000+0100')), ((59, 163), (15, '16:00:00.000+0100')), ((169, 170), (29, '08:00:00.000+0100')), ((72, 90), (16, '08:00:00.000+0100')), ((46, 135), (22, '16:00:00.000+0100')), ((162, 163), (31, '15:00:00.000+0100')), ((161, 162), (26, '14:00:00.000+0100')), ((160, 161), (27, '09:00:00.000+0100')), ((148, 149), (27, '19:00:00.000+0100')), ((47, 129), (31, '16:00:00.000+0100')), ((131, 149), (11, '15:00:00.000+0100')), ((153, 163), (26, '14:00:00.000+0100')), ((43, 57), (20, '19:00:00.000+0100')), ((52, 135), (25, '22:00:00.000+0100')), ((6, 149), (12, '20:00:00.000+0100')), ((149, 168), (16, '18:00:00.000+0100')), ((27, 175), (25, '09:00:00.000+0100')), ((84, 166), (34, '11:00:00.000+0100')), ((118, 168), (20, '18:00:00.000+0100')), ((135, 175), (22, '17:00:00.000+0100')), ((19, 55), (32, '09:00:00.000+0100')), ((131, 163), (20, '19:00:00.000+0100')), ((38, 175), (23, '19:00:00.000+0100')), ((95, 168), (18, '16:00:00.000+0100')), ((118, 163), (23, '10:00:00.000+0100')), ((43, 129), (22, '18:00:00.000+0100')), ((19, 161), (30, '16:00:00.000+0100')), ((77, 79), (25, '15:00:00.000+0100')), ((143, 161), (32, '08:00:00.000+0100')), ((9, 163), (55, '08:00:00.000+0100')), ((133, 135), (19, '09:00:00.000+0100')), ((49, 128), (23, '15:00:00.000+0100')), ((132, 135), (31, '14:00:00.000+0100')), ((162, 169), (17, '07:00:00.000+0100')), ((79, 90), (14, '18:00:00.000+0100')), ((130, 161), (33, '09:00:00.000+0100')), ((163, 170), (15, '18:00:00.000+0100')), ((43, 134), (11, '19:00:00.000+0100')), ((13, 48), (40, '12:00:00.000+0100')), ((64, 90), (26, '08:00:00.000+0100')), ((19, 163), (26, '09:00:00.000+0100')), ((57, 129), (13, '19:00:00.000+0100')), ((84, 90), (24, '08:00:00.000+0100')), ((1, 41), (25, '15:00:00.000+0100')), ((48, 129), (23, '07:00:00.000+0100')), ((19, 168), (26, '08:00:00.000+0100')), ((56, 57), (17, '20:00:00.000+0100')), ((149, 163), (25, '09:00:00.000+0100')), ((86, 90), (26, '16:00:00.000+0100')), ((108, 168), (19, '08:00:00.000+0100')), ((53, 83), (20, '20:00:00.000+0100')), ((139, 149), (23, '05:00:00.000+0100')), ((62, 162), (19, '10:00:00.000+0100')), ((49, 135), (12, '19:00:00.000+0100')), ((66, 86), (21, '12:00:00.000+0100')), ((41, 175), (27, '09:00:00.000+0100')), ((4, 160), (29, '08:00:00.000+0100')), ((38, 129), (18, '06:00:00.000+0100')), ((160, 163), (25, '07:00:00.000+0100')), ((142, 161), (27, '09:00:00.000+0100'))]

```

Dónde cada tupla cuenta con dos subtuplas, la primera de ellas representa el trayecto y la segunda viene determinada por la frecuencia máxima de dicho trayecto y la hora en la que hay un mayor tránsito de bicicletas en dicho trayecto.

- El grupo de edad que más hace uso de las bicicletas eléctricas es el grupo 4, es decir, los usuarios entre 27 y 40 años, quienes constituyen en todo el año un total de 77602 usuarios.