UNIVERSIDAD
COMPLUTENSE
MADRID

Grado en Matemáticas y Estadística

Big Data. Práctica 3

Plan de mejora del servicio BiciMAD con Pyspark

Marina Pescador

Miguel Zabaleta

Índice

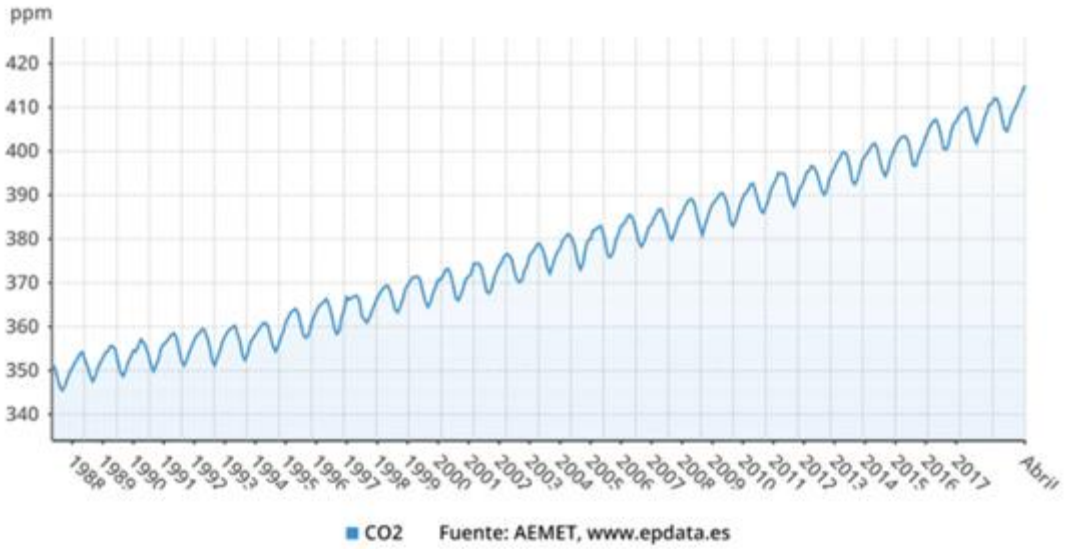
1. Introducción

1.1 El problema del medio ambiente

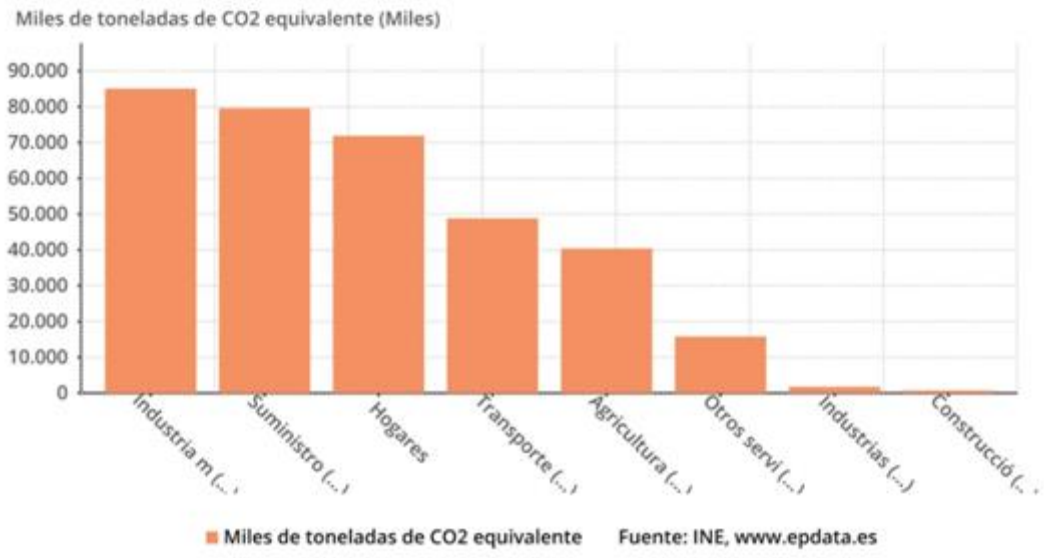
La sostenibilidad del medio ambiente es y sin duda será en el futuro un tema crucial para garantizar el bienestar de las personas y del planeta. Uno de los ocho objetivos de desarrollo del milenio está enfocado precisamente a esto, garantizar la sostenibilidad del medio ambiente incorporando los principios del desarrollo sostenible en las políticas y los programas nacionales. Es bien sabido que uno de los problemas del medio ambiente más grave es el de la **contaminación del aire**, que supone un deterioro en la salud respiratoria y cardiovascular de todos nosotros y de los animales.

En un estudio publicado en nuevatribuna, se estimó que el uso del **transporte público** evita la emisión de hasta 5 millones de toneladas de CO2 al año, por tanto, fomentar y optimizar los planes dirigidos al transporte público es una muy buena estrategia de cara a mejorar la calidad de vida en nuestro planeta.

Analicemos más de cerca el estado en el que nos encontramos. A continuación se presenta una gráfica de la evolución de CO2 en la atmósfera:

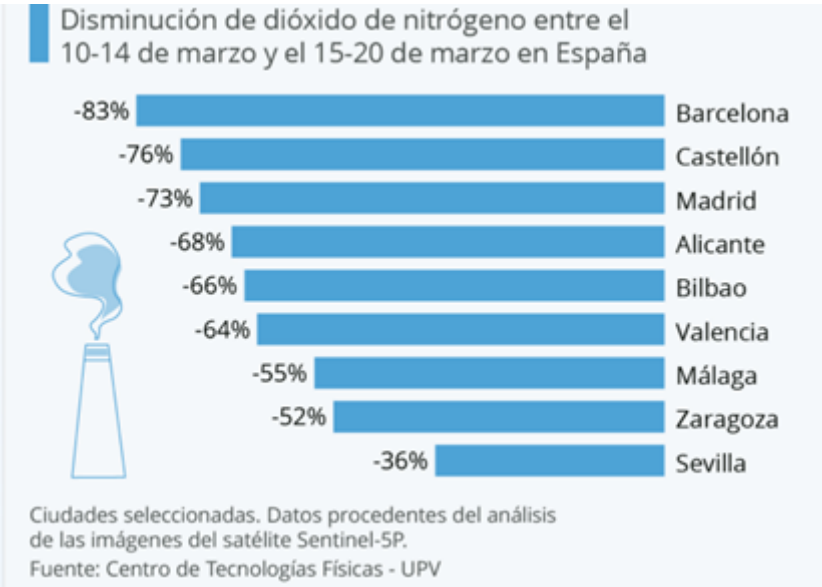


Como se puede ver, el crecimiento no ha parado en los últimos años.



Como se aprecia en este gráfico, el transporte es uno de los sectores que más GEI (gases de efecto invernadero) emitieron en 2017 en España.

Por otro lado, uno podría decir que los datos actuales indican otra cosa, ya que la aparición del virus COVID-19 y el confinamiento que trajo han supuesto una reducción de la contaminación del aire en todo el país, como puede verse en esta gráfica.



Sin embargo, es razonable pensar que conforme se vuelva a la normalidad, la emisión de gases de efecto invernadero volverán a los niveles que se tenían antes. Por tanto, los proyectos que fomenten la mejora del medio ambiente siguen siendo de gran importancia y es necesario lograr el mayor impacto posible con ellos.

1.2 MADRID 360

Uno de los proyectos en Madrid centrales para promover este cambio es **MADRID 360**, la estrategia que planteó se planteó en Madrid a finales de 2019 para cumplir con los objetivos de calidad del aire de la Unión Europea. Las características más destacables en relación con nuestro proyecto son las siguientes:

- Se prevé que las iniciativas rebajaran los óxidos de nitrógeno (NOx) un 15 % más que el anterior plan anticontaminación.
- Además, no habrá calderas de carbón en la ciudad a partir del 1 de enero de 2022 y se pretende eliminar el 50 % de las que funcionan por gasóleo en ocho años. Se destinarán 50 millones de euros en ayudas para alcanzar estos objetivos hasta 2023.
- Se apuesta por la bicicleta, la moto y otros transportes alternativos. La bicicleta es uno de los medios más limpios para moverse por la capital y, por ello, el Ayuntamiento de Madrid continuará con el proceso de **expansión de BiciMAD** dentro y fuera de la M-30 y creará nuevos carriles bici y ciclocarriles que tendrán que ser sometidos a estudio.

Como viene indicado, uno de los focos del plan MADRID 360 es continuar con la ampliación del proyecto BiciMAD. Por tanto, cualquier mejora que se diseñe en este proyecto supone una valiosa contribución en fomentar el cumplimiento de los propósitos de MADRID 360.

Si bien el aspecto de expansión de BiciMAD es muy importante, nosotros proponemos un enfoque diferente. Pretendemos comprender la distribución de la demanda de las bicis según la zona (código postal) y la estación en la que nos encontremos (primavera, verano, otoño o invierno), para así poder diseñar una propuesta de mejora del plan, optimizando el coste y permitiendo de esta forma una **expansión eficiente**.

Por tanto, nuestro objetivo con la solución que implementamos es **optimizar** la oferta y demanda de las bicis, dependiendo de los dos aspectos mencionados.

2. Diseño e implementación de la solución

La solución que proponemos para alcanzar nuestro objetivo es diseñar 4 mapas, uno por cada estación. En estos mapas, marcaremos con círculos las distintas zonas del centro de Madrid e incluiremos la intensidad de la demanda de viajes realizados en cada zona.

Las zonas se corresponden con los distintos códigos postales, y la forma de incluir la magnitud de la demanda será por **colores**, siendo el rojo un indicador de la demanda más

alta, hasta el gris, de demanda más baja. Además, añadiremos como título de cada círculo el número de viajes en esa zona, para que se pueda acceder a esta información fácilmente.

Una vez hecho esto, seremos capaces de analizar los mapas y sacar conclusiones de la distribución de la demanda en el centro de Madrid, según cada estación. Finalmente, propondremos un plan de mejora de acuerdo a estas conclusiones, de forma que se optimice el coste de oferta-demanda.

2.1 Detalles importantes de la implementación

A continuación, destacamos los aspectos más importantes a tener en cuenta de cara a diseñar la solución propuesta.

Importancia preprocesamiento de datos

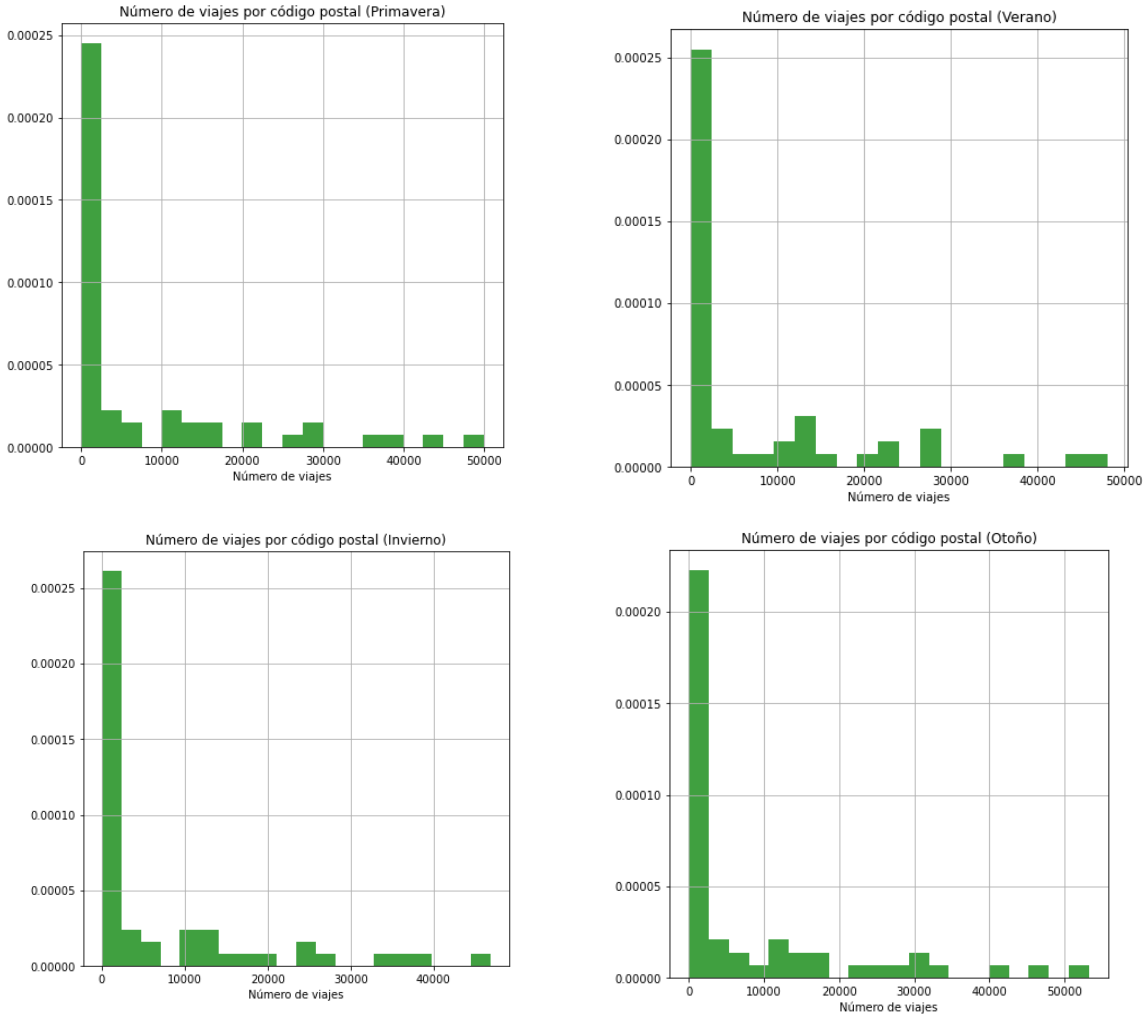
En primer lugar, al estar trabajando con ficheros de datos de gran tamaño (algunos de más de 1GB) y teniendo en cuenta que no se ha trabajado con ellos a parte de la propia recopilación de datos, es claro que el preprocesamiento y limpieza de estos datos cobra mucha importancia.

Tendremos por tanto que hacer mucho hincapié en este punto y asegurarnos de que no haya registros o variables que puedan afectar al diseño de nuestra solución. Para conseguir esto, algunos de los puntos que tendremos que asegurar son:

- Trabajar con registros que tengan código postal, y que sea de Madrid centro.
- Que estos códigos postales estén bien recogidos. Los registros en los que hubiese un error en la codificación serán descartados.
- Descartar variables no comunes a todos los ficheros. Además, de esta forma reducimos el tamaño de los objetos con los que trabajamos, lo que aumenta la velocidad en la ejecución del código.
- A la hora de tratar con los objetos necesarios para crear los mapas, seguramente sea muy costoso computacionalmente trabajar todo el rato con los objetos *json* originales. Para solventar esto, sería buena idea trasladar esta información a un objeto más pequeño, como podría ser un *dataframe* con los datos que nos interesen, y construir desde este objeto las variables/listas necesarias para construir los mapas.

Recopilación de información adicional

Como parte del tratamiento previo de los datos hemos hecho los siguientes histogramas que muestran el porcentaje de viajes que se realizan en cada código postal. El eje x representa el número de viajes y la proporción en la que las zonas (códigos postales) tienen ese número de viajes (variable *count*). Este resultado fue de utilidad para así luego asignar los diferentes colores en los radios.



Habiendo trabajado en otra ocasión con estos datos, sabemos que los ficheros de ciertos meses contienen un campo que indica las coordenadas geográficas del comienzo del viaje. Sin embargo, en nuestro caso la mitad de los ficheros con los que vamos a trabajar carecen de este campo.

Por tanto, para poder dibujar en los mapas los círculos representativos del número de viajes realizados según el código postal, necesitaremos obtener las coordenadas geográficas de los códigos postales con los que vamos a trabajar.

Por supuesto, tendremos que asegurarnos que esta información es verídica y que nos es de utilidad.

Claridad en los mapas

Uno de los motivos por el que decidimos implementar una solución en forma de mapas es porque nos parece una idea muy **atractiva** de cara a presentar de qué forma y por qué deberían mejorar la distribución de la oferta de bicis, según la zona y la estación.

Un mapa permite obtener información muy gráfica de manera sencilla, sin necesidad de recurrir a sofisticados algoritmos estadísticos que puedan generar desconfianza en personas ajenas a estos conceptos.

Es por ello que debemos asegurarnos que en efecto los mapas que diseñamos cumplen esta función, y lo hacen de la mejor manera posible. Para ello, decidimos que una buena solución es representar la intensidad del número de viajes en una zona por **colores**, desde el rojo para las zonas con más número de viajes, hasta el verde para las zonas con menos viajes. También incluiremos el color gris para aquellos códigos postales donde el número de viajes no sea lo suficientemente grande como para prestar especial atención en su demanda.

De esta forma, habremos conseguido obtener **información muy valiosa** de cara a mejorar la implementación del proyecto BICIMAD y, crucialmente, habremos sido capaces de **transmitir** esta información a las personas al mando de este proyecto.

3. Análisis de resultados, conclusiones, plan de mejora propuesto

3.1 Análisis de los resultados

Lo primero que vamos a hacer es analizar los mapas para cada una de las estaciones.

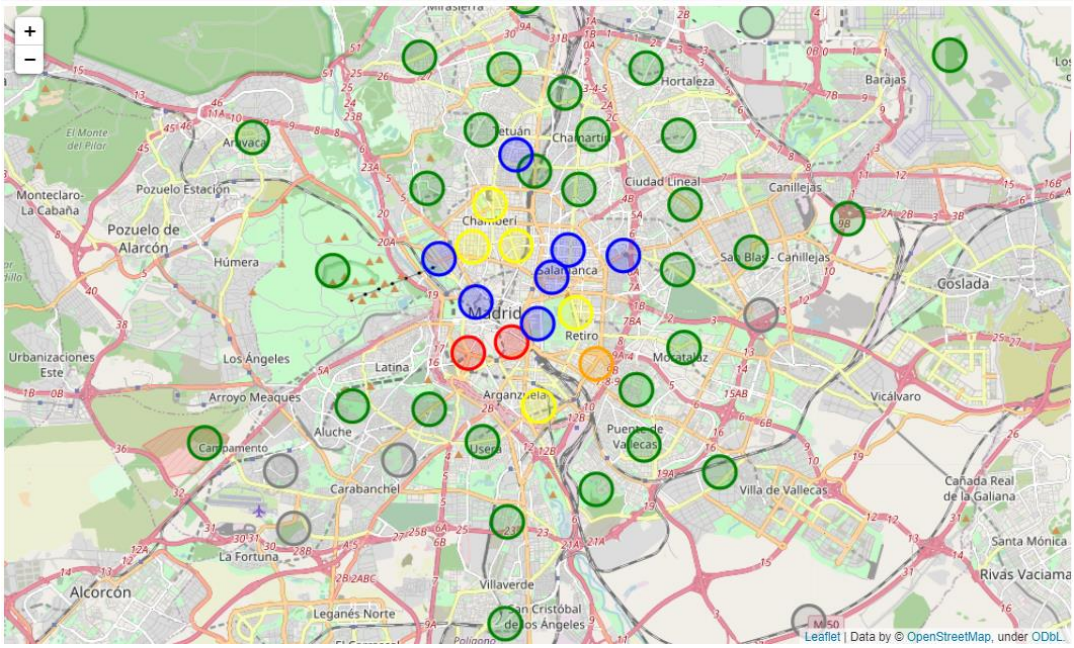
Hemos asignado a cada código postal (*key_code*) un círculo cuyo centro está relativamente centrado en la zona que le corresponde. Estos están coordinados por colores según el siguiente criterio:

- Rojo, la zona con más usuarios (>40000)
- Naranja (>30000)
- Amarillo (>20000)
- Azul (>10000)
- Verde, este es el tipo de zona más común como ya pudimos ver en el histograma y comprobaremos con los mapas (>500)
- Gris, zona límite donde la cantidad de usuarios es mínima (<500)

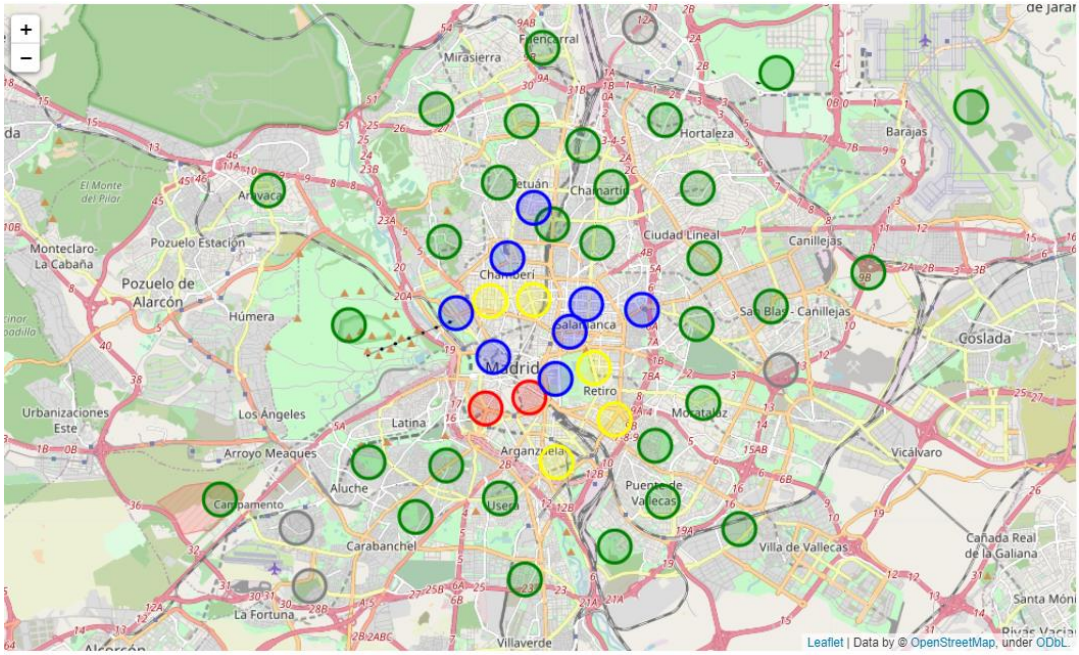
Cabe destacar ciertas características comunes en los datos independientemente de la estación; las zonas más y menos concurridas coinciden en todas las estaciones, aunque el número de usuarios en ellas varía y estas coinciden con las zonas más centrales de la ciudad, como cabría esperar. También podemos observar a primera vista que el invierno es la estación con menos usuarios y el otoño la que tiene más.

A continuación mostramos los mapas creados, los analizaremos y obtendremos conclusiones.

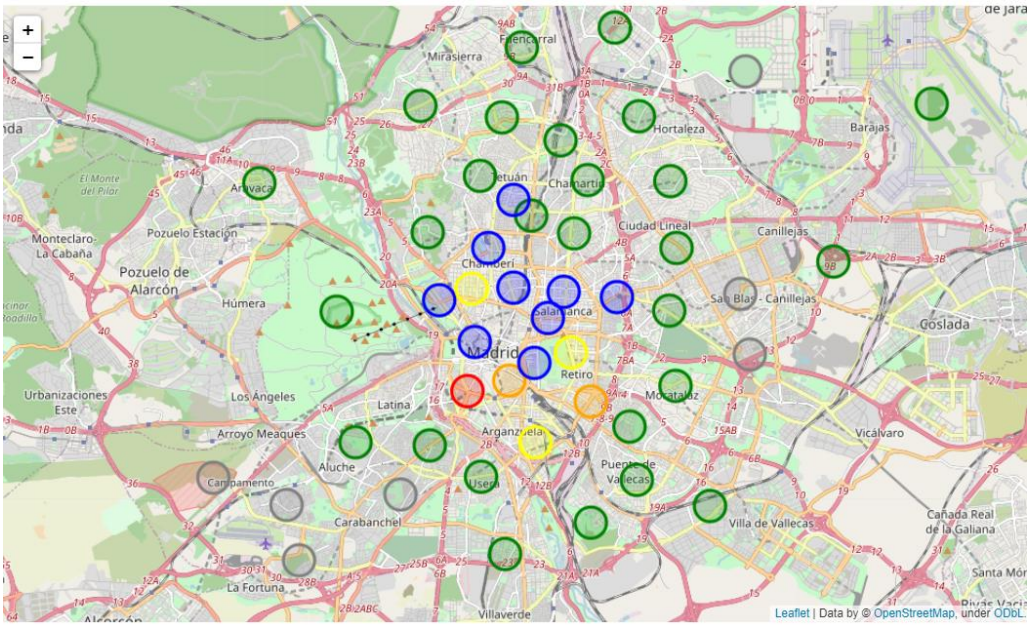
PRIMAVERA



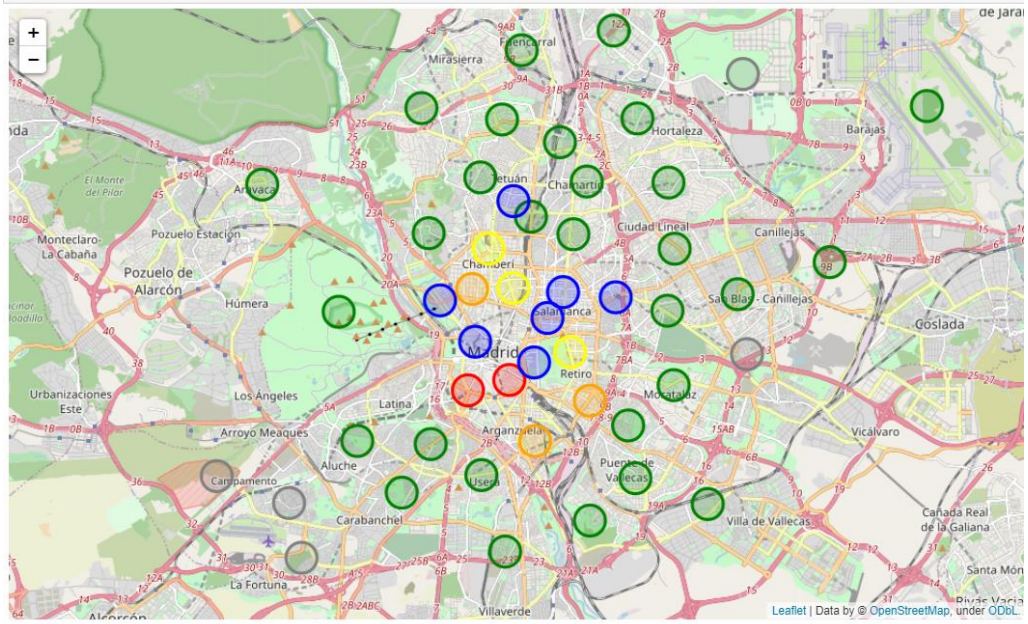
VERANO



OTOÑO



INVIERNO



El número de usuarios por estación es: en primavera 401654, en verano 387026, en otoño 427445 y en invierno: 373375; por lo que otoño es la estación con más usuarios seguida de cerca por la primavera. Esto podría deberse a que durante estas épocas la temperatura no es tan extrema como en verano o invierno y también a que al no corresponder con periodos vacacionales el transporte en general tenga mayor demanda; estos resultados no son sorprendentes y serían difíciles de mejorar.

Es interesante recalcar que hay una menor cantidad de zonas grises en verano, esto indica un uso del servicio en esta zona más recreacional que como medio de transporte (para ir al trabajo etc.).

También observamos que la única zona roja durante todo el año es aquella cuyo código postal es 28005.

3.2 Conclusiones y plan de mejora

De los datos obtenidos se puede conseguir un plan de acción por dos partes:

Por un lado, sobre la **cantidad de bicis mínima** que se debe mantener en cada zona dependiendo de la época del año. Recomendamos prestar una atención especial a las zonas de mayor uso durante todo el año: esto es, el **centro**, que siempre toma colores rojos, naranjas y amarillos.

Al ser un servicio donde no solo se deben tener bicis disponibles para ser utilizadas sino también plazas libres donde poder anclarlas una vez hayan terminado el trayecto, creemos que sería conveniente tener un **mantenimiento diario** en estas zonas mencionadas para que se mantenga un nivel óptimo de bicis/plazas vacías. Este nivel se podría encontrar fácilmente con un pequeño estudio dirigido a esto (por ejemplo, basado en teoría de colas), pero de forma general se recomendaría que la cantidad de bicicletas durante los horarios de más uso estuviese entre el 40% y el 60%.

Este mantenimiento mejoraría la experiencia del usuario ya que evitaría situaciones en las que un usuario no encuentra bicicletas o plazas disponibles en la estación y aseguran la integración del transporte por bicicleta como una opción de uso diario al proporcionar la certeza de saber que no vas a tener tiempos de espera por los problemas anteriores.

También es notable mencionar que al estar todas las zonas relativamente cerca, el transporte de las bicicletas entre estaciones para asegurar una cantidad adecuada **no** sería **costoso** y se podría asegurar que fuese **eficiente** y así no contrarrestar los beneficios tan importantes de este servicio hacia el medio ambiente.

Por otro lado, podemos formular un plan de mejoras a **largo plazo**; este incluiría tanto el mantenimiento de las estaciones y bicis ya disponibles como la introducción de nuevas.

Como hemos podido comprobar, en invierno la demanda es menor que el resto del año, por ello recomendamos que los servicios de mantenimiento se planeen en la medida de lo posible para el **invierno** y si fuese necesario verano, que es la segunda estación con menos demanda. Como es probable que estos servicios se deban repartir durante todo el año, en ese caso el mejor plan sería hacer el mantenimiento de las zonas de menos uso, verdes y grises durante la primavera y el otoño para así poder reservar el mantenimiento de las zonas con mayor demanda para las estaciones con menor uso del servicio.

Para la introducción de nuevos servicios recomendamos añadir tanto plazas de bicicletas como nuevas estaciones en las zonas de más uso. Además, creemos que podría ser interesante añadir infraestructuras como carriles bici que unan estas zonas, ya que si observamos los mapas podemos ver que las zonas de más uso (de rojo y amarillo), aunque todas corresponden al centro están "separadas" por zonas de uso intermedio (azul) correspondientes a los códigos postales: 28014, 28013, 28008, 28001; por lo que sería interesante promover una mejora de la accesibilidad con bicicleta en estas zonas azules y así tener un centro completamente accesible por el servicio de BiciMAD.

Por último, queremos recalcar las zonas grises, si bien éstas corresponden a la periferia, en una ciudad tan poblada como Madrid todas las zonas son lo suficientemente densas como para que este servicio tenga éxito. Por eso, y teniendo en cuenta el plan Madrid 360, un estudio de las **causas de esta falta de uso** es una buena inversión con la consiguiente implementación de una solución; carriles bici y mejor circulación para las bicicletas, mayor promoción del servicio etc.

4. Bibliografía

https://es.wikipedia.org/wiki/Objetivos_de_Desarrollo_del_Milenio

https://nuevatribuna.publico.es/articulo/salud/contaminacion-autobus-equivale-50-coches/20160126135703124751.html

https://elperiodicodelaenergia.com/la-contaminacion-del-aire-en-10-graficos/

https://es.statista.com/grafico/21270/disminucion-de-dioxido-de-nitrogeno-en-espana/

https://blog.emtmadrid.es/2020/02/20/ampliacion-bicimad-2020/

https://www.madrid.es/portales/munimadrid/es/Inicio/Actualidad/Noticias/MADRID-360-la-estrategia-para-cumplir-con-los-objetivos-de-calidad-del-aire-de-la-Union-Europea/?vgnextfmt=default&vgnextoid=3d6c1609d818d610VgnVCM2000001f4a900aRCRD&vgnextchannel=a12149fa40ec9410VgnVCM100000171f5a0aRCRD

https://blog.oxfamintermon.org/los-7-problemas-del-medio-ambiente-mas-graves/