EDA: Análisis de Datos

Estudio de la correlación entre la variación del voto al PP en las elecciones a la asamblea de Madrid de Mayo 2021 y Abril 2019 y la hostelería

Objetivo del proyecto

En la precampaña de las últimas elecciones a la comunidad de madrid se ha polemizado mucho sobre la posible influencia de los bares y restaurantes en los resultados de las elecciones; el propio director del CIS ha expresado su opinión sobre el tema llamando 'tabernarios' a los partidarios de uno de los candidatos de los partidos políticos que se presentaban.

El objetivo de este análisis es demostrar si hay o no una correlación entre el número de bares y restaurantes de los municipios de la comunidad de Madrid y la evolución del voto de esos municipios entre las elecciones de 2019 y 2021.

Contexto del análisis

Dado que el objeto a analizar es la variación del voto en función de los bares para todos los municipios de la Comunidad de Madrid (CAM), el contexto del análisis se sitúa en toda la comunidad autónoma.

Sin embargo, el análisis realizado en el presente documento no ha podido alcanzar el contexto al completo, ya que los datos de locales relacionados con la hostelería en los municipios fuera de la zona metropolitana de la ciudad de Madrid no están disponibles.

Recolección de los datos

Tal y como se ha avanzado en el apartado de contexto del análisis, este EDA necesita de los datos de las votaciones a la asamblea de Madrid de Mayo de 2021 y Abril 2019, así como los datos para obtener los locales de la CAM relacionados con la hostelería.

El conjunto de datos 'votos a la asamblea de Mayo 2021', 'votos a la asamblea de Abril 2019' y 'locales clasificados por actividad' han sido encontrados para el municipio de Madrid; lamentablemente, pese a encontrar los datos 'votos a la asamblea de Mayo 2021' y 'votos a la asamblea de Abril 2019' para los municipios de la CAM fuera del área metropolitana de Madrid, no se ha podido obtener los datos de locales relacionados con la hostelería en los municipios de la CAM.

Origen de los datos

Los datos de este EDA provienen del ayuntamiento de Madrid y del portal de datos abiertos del gobierno [REF1], [REF2] y [REF3].

Periodicidad de actualización de los datos

Los datos de este EDA no tienen una actualización periódica que suponga un riesgo para la estabilidad de las conclusiones del análisis. Dado que este EDA se basa en el análisis de unas votaciones y las licencias de los locales, no se corre el riesgo de que los datos de base se modifiquen, por lo tanto no es necesario descargarlos periodicamente para el análisis.

Limpieza de los datos

Los datos del conjunto de trabajo han tenido que ser procesados para ser explotables, sin embargo, el procesado para cada uno de ellos ha sido distintos

Limpieza de los datos de las votaciones

Los datos de las votaciones han sido limpiados gracias a la herramienta excel ya que dentro del archivo original había encabezados y distribución de los datos en varios niveles y por lo tanto el refinado de los datos con Python resultaba complejo.

Para optimizar el proceso, se eliminaron las cabeceras innecesarias directamente con Excel de forma a generar una archivo que fuese fácilmente interpretable por Python. Una vez en Python, las líneas de los archivos originales se agrupan por barrios, de forma a poder tener esa unidad de referencia disponible en todo el EDA.

Limpieza de los datos de los locales

En este caso, el archivo con los locales y sus actividades si que ha sido filtrado con Python. En un primer acercamiento, se han obviado las columnas innecesarias como la calle o el número de la dirección, así como los datos administrativos de identificación o códigos internos.

Después de este primer subproceso de limpieza inicial de columnas, se han limpiado los datos por filas, ya que no todos los locales del archivo formaban parte del objeto del estudio. Dado que el estudio se enmarca en la hostelería, se han eliminado todos los locales relacionados con otras actividades (a excepción de locales de ocio nocturno y espectáculos, que no estan catalogados dentro de la actividad de la hostelería, y sin embargo sí que están dentro del contexto de este EDA).

Finalmente, una vez eliminadas las filas correspondientes a los locales no relacionadas con la hostelería, se eliminan las columnas relacionadas con las actividades de los locales y se agrupan los resultados por el campo de barrio, generando la tabla final que se utilizará en el EDA, sólo con los campos id_barrio y locales (acumulador con los locales relacionados con la hostelería en el barrio).

Minería de datos

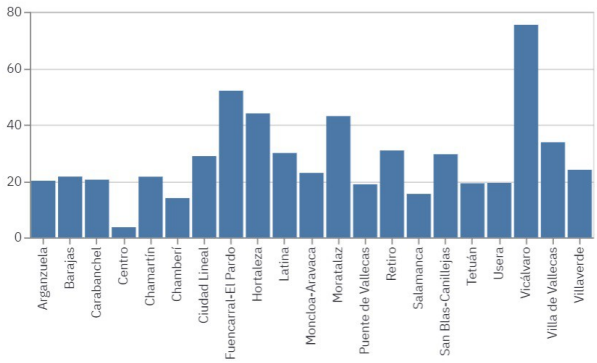
Una vez todas las tablas extraidas de los datos de origen son procesables en Python, se unen para definir el dataset final con el que se va a elaborar este EDA. La prioridad de esta unión de tablas es la adición el acumulador de locales a la tabla del dataset, pero esta no es la única tarea del proceso de data mining.

Además de unir las tablas, se han generado algunas variables 'sintéticas' (o no existentes) a partir de los datos disponibles de las tablas originales, como por ejemplo la variación porcentual del voto o la variación en función del censo de cada barrio, que podrían ser útiles en un estudio avanzado de los datos.

Análisis de los datos

Tendencias de las columnas

Las principales columnas del estudio son D-PP (que representa la variación del voto de cada barrio) y el número de locales relacionados con la hostelería en cada barrio). Si se agrupan estos datos por distritito, se obtienen los siguientes histogramas:

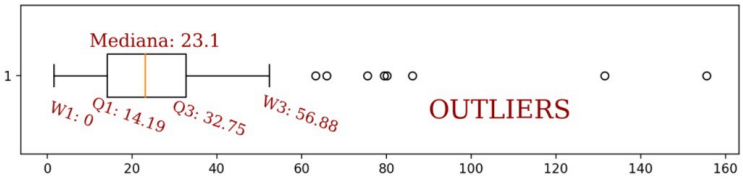


Variación del voto al PP dividida vs el número de locales por distrito

Dado que ambos histogramas no tienen las mismas distribuciones, se puede anticipar de forma intuitiva que no hay correlación entre ambas variables, sin embargo es necesario hacer un análisis más profundo con el fin de determinarlo matemáticamente.

Para ello se calcula la variable (D-PP/Locales) on el fin de tener datos analizables en una única variable. Esta nueva variable tiene esta distribución en un histograma:

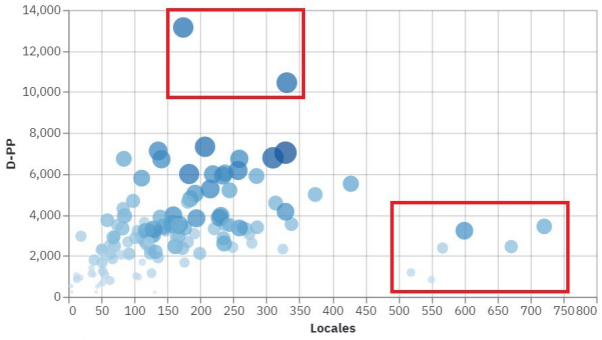
Dado que las distintas barras están muy dispersas con respecto a la media del valor, intuitivamente es posible determinar que se confirma que los datos no están correlacionados, pero es necesario confirmarlo con un análisis más profundo, detalllando el cálculo a nivel de barrio, que es la unidad de agrupación más pequeña que tienen los datos de este EDA.



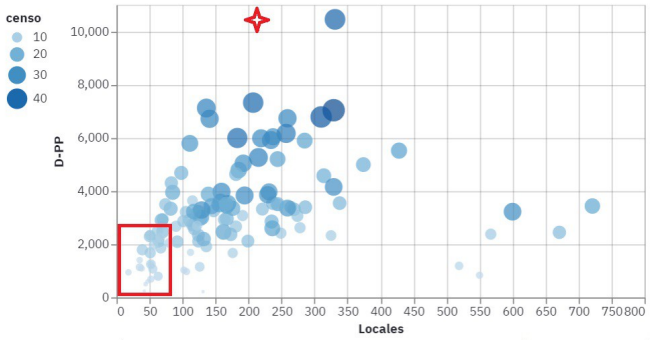
Boxplot: localización de los outliers existentes

Análisis de los datos a nivel de barrio

Antes de analizar todos los datos del nivel de barrio, es interesante hacer un análisis de los mismos con el fin de identificar los outliers si existieran, y si es posible, eliminarlos del dataset con el fin de que no enturbien las conclusiones.



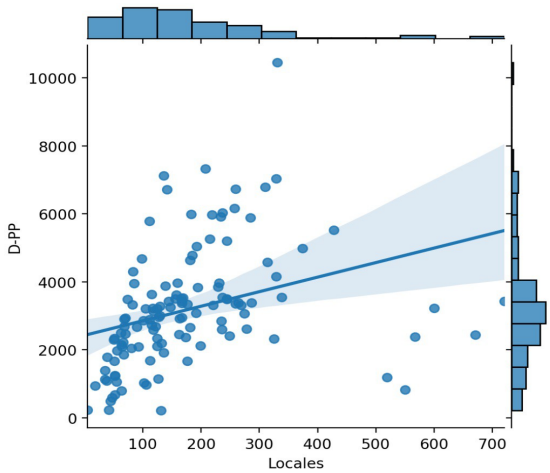
Variación del voto al PP vs locales por barrio (con outliers)



Variación del voto al PP vs locales por barrio (sin outliers)

En este caso, el boxplot muestra que hay una serie de outliers situados por encima del 60, que si se calcula su peso con respecto al total (número de votos en el censo con respecto del censo total) se pueden eliminar de los datos a analizar. Por lo tanto, se puede continuar el análisis eliminando estos outliers, y a continuación se muestran las diferencias entre los datasets con y sin outliers:

El filtrado de outliers es importante ya que los datos que a priori pueden parecer outliers o que corrompan la muestra, en realidad pueden no serlo y por lo tanto ejercer un peso importante en la desviación de los datos y por lo tanto en el cálculo de la correlación final.



En este caso, se ve que lo que podrían ser outliers (enmarcados en rojo), en realidad no lo son, y los outliers se sitúan en otros puntos del gráfico (gráfico izquierdo). Si además se ñaade la recta de regresión a la nube de puntos, se puede deducir el nivel de correlación de los datos.

Conclusión del análisis

La dispersi3 de los datos alrededor de la recta de regresi3 es muy alta, por lo que intuitivamente se puede deducir que no hay correlaci3 entre variables. Adem 疽, el coeficiente de correlaci3 de Pearson es 0.324, que est muy por debajo de un valor que indique una correlaci significativa (0.45), por lo tanto, la variaci3 del voto hacia el PP y el n但ero de locales no est疣correlacionadas.

Por lo tanto, la hipótesis inicial ha sido refutada.

Apartado 9 del nivel C: feedback del EDA

¿Fue posible demostrar la hipótesis?¿Por qué?

Si, fue posible demostrar la hipótesios ya que los datos estaban disponibles, en las conclusiones se ha refutado la hipótesis incicial ya que según los cálculos, no hay una correlación significativa entre la variación de votos hacia el PP y el número de locales de cada barrio de Madrid.

Sin embargo, cabe destacar que esta conclusión no puede tomarse como definitiva ya que la muestra de datos no corresponde con el total de los datos que habría que analizar. Dado que las elecciones a la asamblea corresponden a todo el censo de la CAM, y los datos analizados corresponden sólo con la zona metropolitana de Madrid, hay un margen considerable de incertidumbre

¿Qué se puede concluir del EDA?

Tal y como se ha comentado en el apartado anterior, si bien las conclusiones de los datos de Madrid capital refutan la hipótesis, los datos de los municipios de la CAM pueden confirmarla, y la diferencia en la naturaleza entre los orígenes de los datos (metrópolis vs municipios) puede ser origen de una diferencia notable en las conclusiones. Con lo cual, la conclusión de este estudio debe tomarse como parcial hasta completar el estudio con los datos totales de locales relacionados con la hostelería en la CAM.

¿Qué cambiarías si fuese necesario hacer otro EDA?

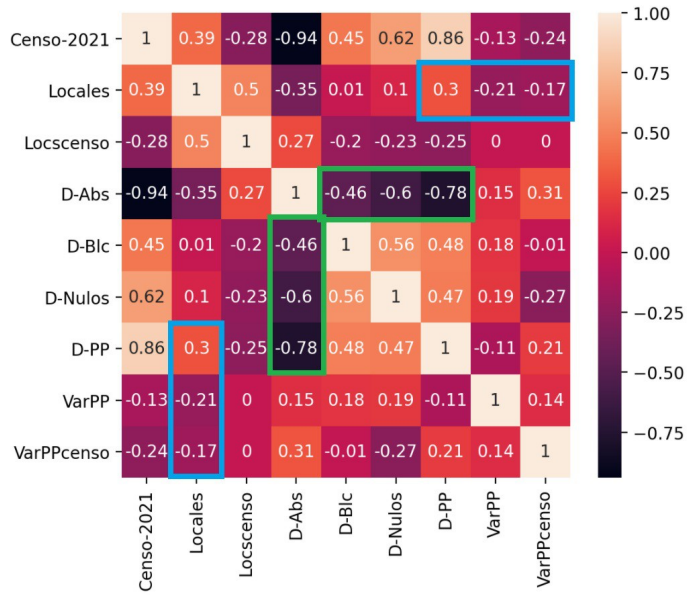
Creo que en conjunto el resultado de este EDA es muy satisfactorio, pero para el próximo EDA, intentaré optimizar el tiempo dedicado a la búsqueda de datos, ya que en este caso he dedicado mucho tiempo a la búsqueda de datos y refinado de datos para la parte del análisis de los datos de los municipios de la CAM, y finalmente no ha sido posible integrarlos en el EDA.

¿Qué aprendiste con este EDA?

En este EDA he aprendido muchas cosas (que conocer las fuentes de datos es fundamental, a hacer datawrangling, a refinar los datos, a crear una página y gráficos interactivos potentes con streamlit y acceder a una base de datos para guardar los datos generados), pero sobretodo que puedo hacer un EDA medianamente completo en menos de 10 días.

Anexo I: Análisis avanzado

Pese a que la hipótesis inicial ha sido refutada con los datos empleados, es posible ampliar los datos a analizar con el fin de averiguar si hay otro tipo de relación con la variación del voto y el número de locales en cada barrio. En este anexo se incluye un análisis de los datos en función de nuevas variables relacionadas con la variación del voto (variación porcentual del voto -VarPP- y variación del voto en función del censo -VarPPcenso-) y la variación del voto con respecto al antiguo ganador de las elecciones en el barrio.



Análisis con variables relacionadas con la variación del voto

Una vez generadas las variables sintéticas, se añaden otras variables como la variación de la abstención, del voto nulo o los votos blancos con el fin de realizar un análisis de amplio espectro con la matriz de correlación para poder identificar correlaciones entre los datos si las hubiera; además, es posible resaltar estos resultados con un mapa de calor, ya que potencia aquellos valores que revelan correlaciones.

Tal y como se puede observa en el mapa de calor, las variables D-PP, VarPP y VarPPcenso (enmarcadas en azul) no tienen ninguna correlación con los locales por barrio (Locales), lo que refuerza la refutación de la hipótesis inicial del EDA.

Sin embargo, es curioso ver cómo se comporta el voto alrededor de la variable de variación de la abstención (marco verde). Según indican estos datos, en combinación con una menor abstención (la media de D-Abs es -1311,5 votos/barrio -el cálculo del dato no se muestra en este análisis-), el sentido del voto se ha polarizado, o bien hacia los votos en blanco o nulos, o hacia el voto al PP.

Explicación del voto al PP en función del voto anterior

Dado que las medidas tomadas tienen un origen político, es posible que el ganador de las anteriores elecciones tenga algún tipo de influencia en el resultado de la variación, por lo tanto se pueden analizar los datos con esa perspectiva para ver si merece la pena hacer un análisis más profundo en esa dirección.

Al observar ambas gráficas, cabe destacar que la dispersión de los datos con respecto al partido que ganó al anteriores elecciones es significativamente diferente en función del partido que las ganó. Las distintas dispersiones con respecto a la recta de regresión provocan una correlación distinta para cada gráfica (PP= 0.576, PSOE= 0.242), y mientras que la correlación de los datos en los barrios donde ganó el PP es alta (casi 0'6), la correlación en los barrios donde ganó el PSOE no llega al 0'25-.

Esta diferencia en los resultados podría evidenciar que en los barrios donde ganó anteriormente el PP los votantes si que han tenido en cuenta las medidas para con el sector de la hostelería, mientras que ese no ha sido un factor de influencia en los barrios donde anteriormente ganó el PSOE.

Anexo II: Referencias

[REF1] Resultados electorales. Asamblea de Madrid 2021

https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Elecciones-y-participacion-ciudadana/Elecciones-/Resultados-electorales-Asamblea-de-Madrid-2021/?vgnextfmt=default&vgnextoid=6d1cb9211cb39710VgnVCM1000001d4a900aRCRD&vgnextchannel=1a47c2338522a210VgnVCM1000000b205a0aRCRD

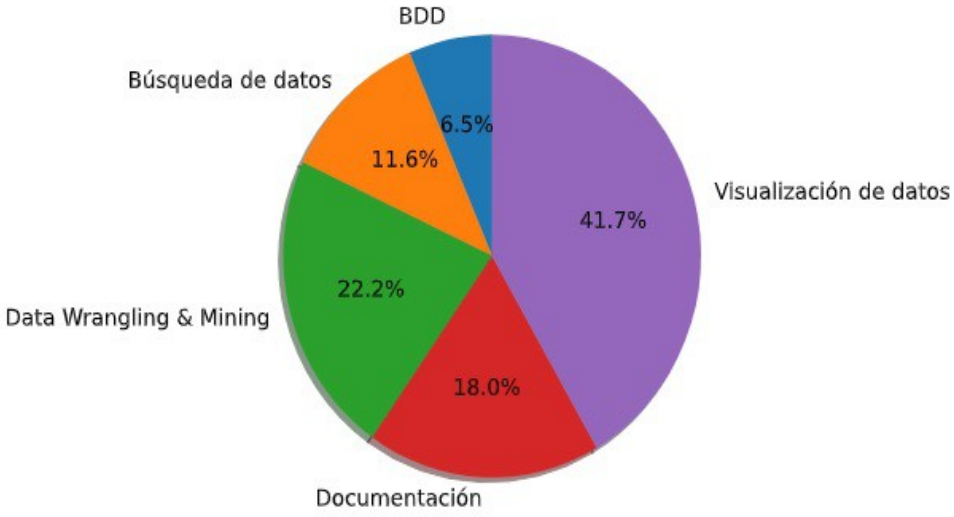
[REF2] Resultados electorales. Asamblea de Madrid 2019

https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Elecciones-y-participacion-ciudadana/Elecciones-/Resultados-electorales-Asamblea-de-Madrid-2019/?vgnextfmt=default&vgnextoid=9d8bb360ceb2b610VgnVCM1000001d4a900aRCRD&vgnextchannel=1a47c2338522a210VgnVCM1000000b205a0aRCRD

[REF3] Censo de locales, sus actividades y terrazas de hostelería y restauración

https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=66665cde99be2410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

Anexo III: Tabla de tiempos del EDA



A continuación se representan los tiempos empleados en cada aspecto del proyecto, que han supuesto un total de 57 horas y 30 minutos:

Ilustración 2: Distribución del tiempo dentro del proyecto

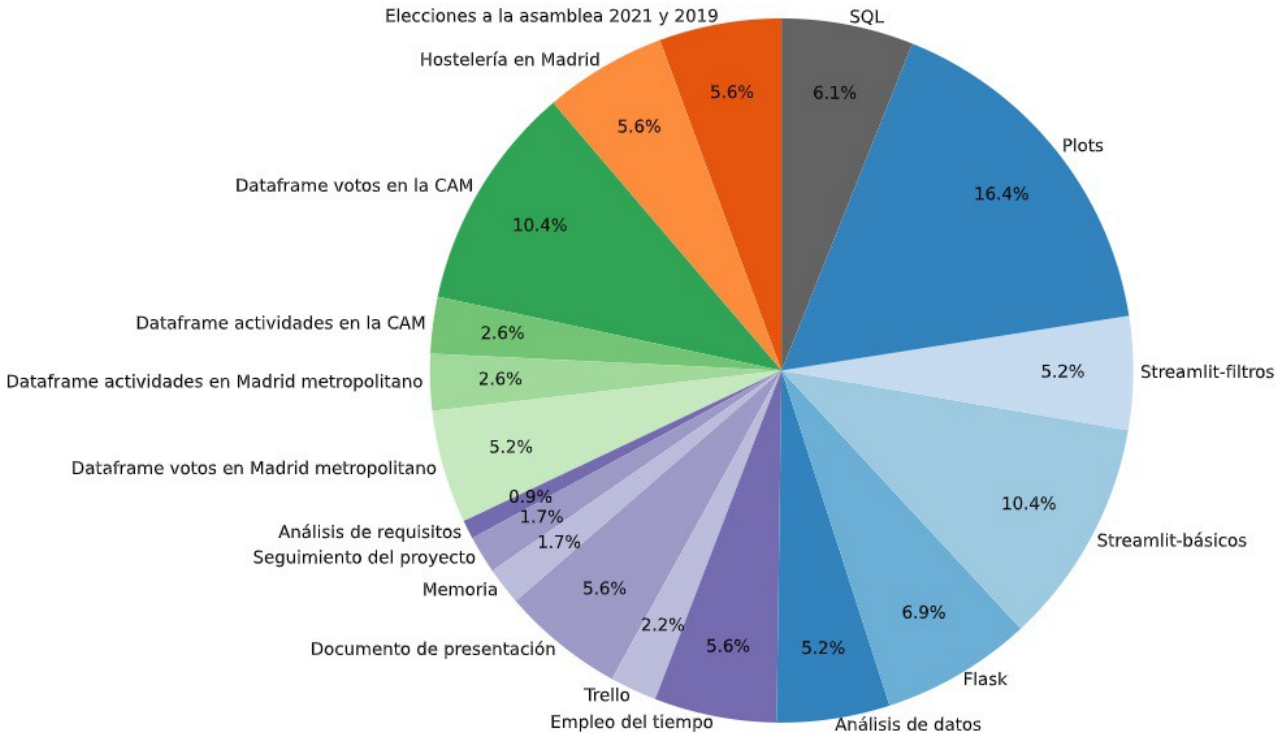


Ilustración 2: Detalle de la distribución de tiempos en el proyecto