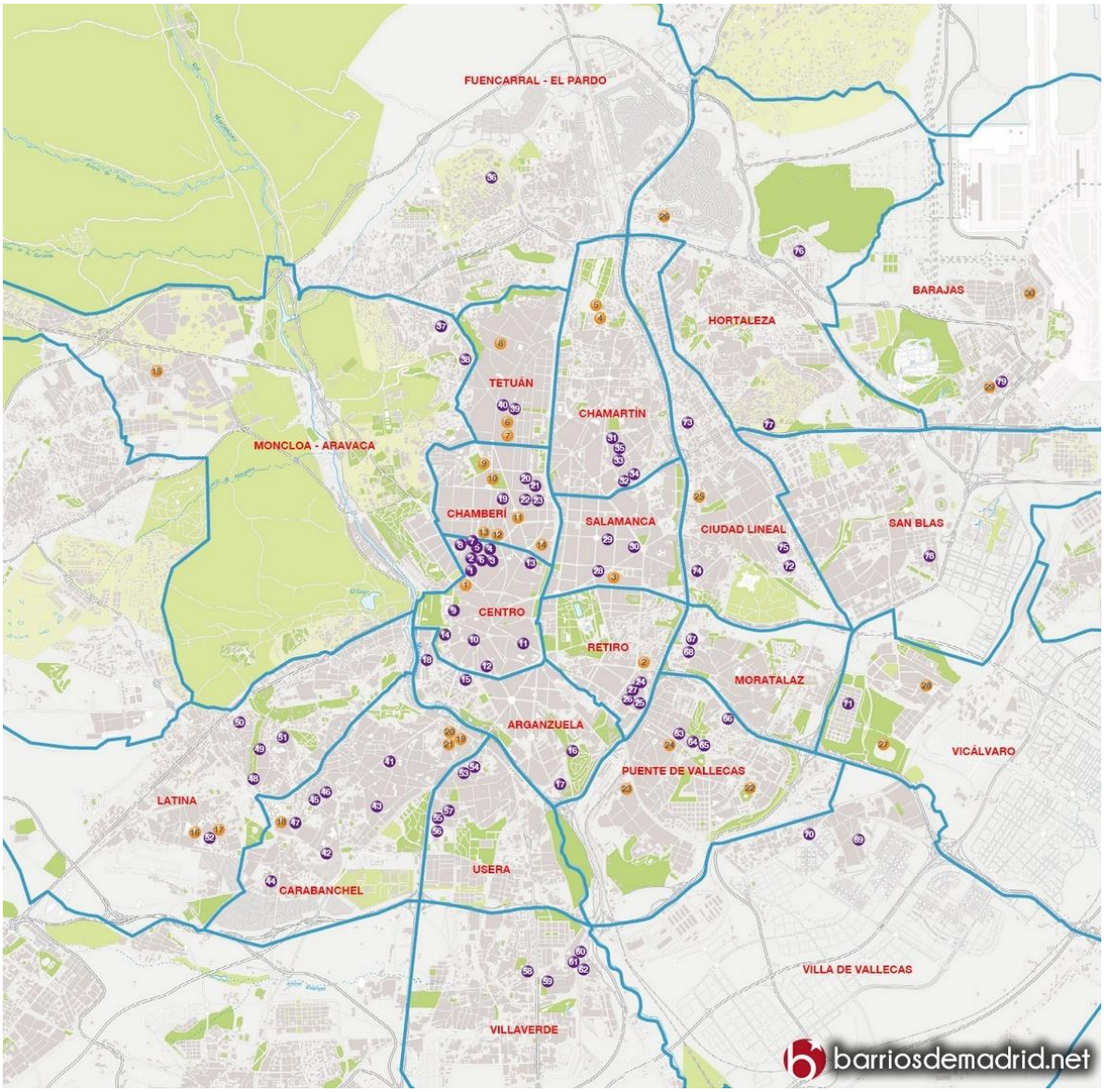


INDIVIDUAL DELIVERY: VIVIR EN MADRID

24/08/2020 A 31/08/2020

@CRISTIDATAS

DATA SCIENCE

THE BRIDGE

VISIÓN GENERAL:

Este proyecto trata de averiguar de qué forma ha afectado la pandemia del coronavirus en la evolución del precio de las casas en Madrid capital y sus 21 distritos, para lo que analizaré tanto los precios de venta como los de alquiler.

Comenzaré con una visión general del estado de los precios en la actualidad (en julio de 2020) en toda España, comparando tanto las 17 comunidades autónomas como las 50 capitales de provincia. Descubriendo qué posición ocupa Madrid dentro de los rangos de precios del resto de las capitales de España.

A continuación, analizaré la evolución de los precios desde 2007 (en plena burbuja, justo antes de la crisis inmobiliaria) hasta la actualidad (julio 2020). Me centraré primero en una comparativa entre España, Madrid provincia (y CAM) y Madrid ciudad con la ciudad que he considerado más parecida a Madrid de toda España: Barcelona (con evidentes diferencias) y también con Barcelona provincia para igualar los parámetros.

En este análisis me centraré en distintos períodos. Primero una visión general desde 2007, luego desde enero de 2019 para ver cuál era la tendencia anual y cómo afectó la llegada de la pandemia y el estado de alarma. Y por último otro tramo desde enero hasta julio de este año para ver en profundidad cuál ha sido la evolución tanto de los precios de la venta como los del alquiler.

A continuación, haré los mismos análisis que con España, Madrid y Barcelona, pero esta vez centrándome en un ámbito geográfico más local: los 21 distritos de Madrid, desde 2007, tanto en lo referente a la venta como al alquiler.

Todos los gráficos se muestran por pantalla y luego se guardan en la carpeta correspondiente a su operación (venta o alquiler) con un nombre descriptivo, tanto en .png como en .html. Todo el proyecto se ha hecho con funciones que están en los diferentes módulos y que automatizan todas las tareas al llamarlas desde main.ipynb.

OBJETIVOS:

Mi objetivo ha sido llegar a la opción A, aunque hoy (31 de agosto de 2020, intentaré llegar a la opción A+ haciendo Pull Request.

A lo largo de este documento respondo a todas las cuestiones de las opciones.

ESPECIFICACIONES:

. SOFTWARE:

El sistema operativo ha sido Windows 10. En cuanto al código, es necesario VS Code u otro lector de Jupyter Notebook. El lenguaje es Python y las librerías necesarias son Pandas y Plotly, que se importan en el momento de ejecutar el archivo main.ipynb. En el mismo archivo se encuentran los otros programas que ha sido necesario instalar: Xlrd (para trabajar con datos de Excel), Psutil (necesario para exportar gráficos de plotly), orca y Kaleido (finalmente usé Kaleido para la exportación de gráficos). Sólo importé y usé Matplotlib y Seaborn para las matrices de correlación. También es aconsejable un lector de pdf para poder ver la documentación incluida.

. HARDWARE:

Como requisito mínimo se ha testado en un Asus con un procesador Intel i5 y 8 GB de RAM, aunque el proyecto se ha realizado con un Acer i7.

. REQUERIMIENTOS:

Los datos se han obtenido de la página web del portal inmobiliario Idealista. Están guardados en 4 ficheros (libros) de MS Excel situados en la carpeta src del proyecto. Cada vez que el archivo main se ejecuta hace un llamamiento a estos ficheros y trabaja con ellos.

Ésta ha sido la única forma de obtener los datos más actualizados (julio 2020). Este proyecto queda abierto a la actualización de estos datos más adelante (cuando se publiquen los datos de agosto u otros meses) y quizás a la incorporación de otros datos muy interesantes como la firma de hipotecas, que es un dato obtenido del

INE pero que no se ha incorporado porque actualmente sólo hay datos hasta el primer trimestre del año, con lo que realmente no se llega a ver el efecto del coronavirus en la concesión de préstamos hipotecarios.

PASOS:

I. INVESTIGACIÓN DEL CONTEXTO:

El conocimiento de esta materia viene dado sobre todo por búsquedas personales en portales inmobiliarios detectando en la práctica la poca variación de los precios durante la crisis del coronavirus, al menos hasta ahora.

Durante el proceso de documentación para el estudio he consultado al Ministerios de Transportes, Movilidad y Agenda Urbana (Fomento) y al Instituto Nacional de Estadística.

II. OBTENCIÓN DE DATOS:

Dado que la principal fuente en España sobre estadísticas inmobiliarias es precisamente el portal Idealista (tanto las inmobiliarias, los medios de comunicación y los organismos oficiales lo citan como fuente directa), mi fuente de datos también ha sido Idealista. No tanto la búsqueda individual como los informes que realizan.

Durante la búsqueda de datos he obtenido también datos de Airbnb y del INE. No he llegado a incorporar ninguno de éstos porque ambos estaban actualizados justo hasta antes de que empezara la pandemia. Este proyecto queda abierto a incorporarlos, especialmente los del INE que se referían a la firma de hipotecas y que no actualiza los datos del segundo trimestre (que es el relevante) hasta septiembre.

III. DATA WRANGLING:

Mi primera idea fue la API de Idealista, después de una ardua tarea para acreditarme (no sólo hay que tener token sino una identificación para conseguir el token que además hay que codificar en base64, una especie de token doble), durante el breve espacio de tiempo en el que conseguí el acceso comprobé que los datos obtenidos no servían para este estudio porque el resultado es una búsqueda de inmuebles individuales (conseguía 450 en cada petición) y con una limitación de peticiones que imposibilitaban hacer el pretendido estudio a través de meses y años.

En el mismo portal elaboran una serie de informes sobre venta y alquiler que eran los que realmente me eran útiles. Después de ponerme en contacto con Idealista no conseguí ningún enlace o fichero con estos datos y el webscrapping resultó imposible por las barreras de seguridad que pone la propia web. Así que la única manera ha sido ir copiando manualmente cada informe en varias hojas de varios libros Excel que luego he recuperado, concatenado y tratado mediante pandas. Para este proceso he utilizado la siguiente función válida para distintos dataframes.

Función hacer_df() de mining_tb que recibe como parámetro un archivo de Excel, hace un dataframe con sólo la primera hoja y sólo las 2 primeras columnas (fecha y precio por m²) poniendo como índice la primera y, luego, mediante un for va recorriendo el resto de las hojas concatenando la segunda columna de cada una (precio) al primer dataframe. La función devuelve un dataframe con las fechas como índice y los precios de cada ámbito geográfico como columnas.

IV. MINERÍA Y LIMPIEZA DE DATOS:

Algunas columnas de los dataframes tienen algunos valores ausentes (los valores más antiguos, dependen del momento en que se empezaron a recabar datos en algunos distritos de Madrid). En este caso no los he querido convertir a cero porque comprobé que podía pasar el resto de la columna a float y trabajar con los datos (con int no funcionaba) y de este modo al hacer los gráficos de línea se veía comenzar cada una en el momento en el que había datos (sin partir de cero y tener una subida repentina) y no tuve que prescindir de filas incompletas pero válidas para los gráficos.

El principal obstáculo al obtener los datos fue que los valores del precio por m² eran cadenas de texto con caracteres alfanuméricos (el punto para los miles y también "€/m2"). Para limpiar estos caracteres usé esta función que me sirvió para todos los dataframes.

Función hacer_float() de mining_tb que toma como parámetro un dataframe, hace una copia, elimina los caracteres no numéricos, el 2 de m² y si hay alguna coma (para decimales) la cambia por un punto (todo esto con Regex) para después convertir las cadenas numéricas a float. La función devuelve un dataframe con valores tipo float.

También he hecho otra función que selecciona las columnas a analizar y que me ha servido para todos los análisis de todos los dataframes.

Función seleccion_columnas() de mining_tb para seleccionar el ámbito (columnas) a analizar. La función toma como parámetros un dataframe y la posición de dos columnas para seleccionar un rango. La función devuelve la lista de esas columnas seleccionadas.

V. VISUALIZACIÓN Y GUARDADO:

He tratado de hacer funciones que me automatizaran todos los procesos (tratamiento y que fueran lo más flexibles posible para tratar datos. Todos los gráficos son interactivos y se guardan en sus correspondientes carpetas con nombres descriptivos.

Cada una de ellas sirve tanto para venta como para alquiler y todas sirven para distintos ámbitos geográficos.

A continuación, detallo las que he usado para analizar el momento actual (julio 2020).

Función act_pie() de visualization_tb para ver el estado reciente de los precios en distintos ámbitos geográficos. La función toma como parámetros el dataframe de precios recientes, el ámbito geográfico (AUTONOMÍA O CIUDADES) y la operación (venta o alquiler) y devuelve un gráfico pie interactivo con la distribución de los valores correspondientes, primero lo muestra por pantalla y luego lo guarda en la carpeta correspondiente a su operación.

Función hist_v_2020() de visualization_tb para ver la distribución de los precios según distintos ámbitos geográficos. La función toma como parámetro en dataframe, la operación (venta o alquiler) y el ámbito (AUTONOMÍA o CIUDADES) y devuelve un histograma interactivo de 5 bines que guarda en la carpera correspondiente a la operación.

A continuación, detallo las que he usado para analizar la evolución de los precios desde 2007 en distintos períodos.

Función evolucion() de visualization_tb para ver la evolución a lo largo de los años o meses del precio por m² de la vivienda en distintos ámbitos geográficos. La función toma como parámetros el dataframe a analizar, la operación que se quiera analizar ("venta" o "alquiler") y dos listas: una con los rangos de columnas o ámbitos geográficos a analizar y otra con los años de inicio del análisis hasta la actualidad. En los gráficos aparece sombreado el período del estado de alarma

(desde el 14 de marzo al 21 de junio de 2020). Primero muestra la imagen de cada gráfico en modo interactivo y después guarda cada gráfico con su propio nombre descriptivo en la carpeta correspondiente a la operación dentro de plots en resources.

Función histo07() de visualization_tb para ver la frecuencia de precios desde 2007 en España, Barcelona provincia, Barcelona ciudad, Madrid provincia y Madrid ciudad. La función toma como parámetros el dataframe el tipo de operación (venta o alquiler) y el ámbito geográfico y devuelve varios histogramas interactivos de 5 bines que guarda en la carpera correspondiente a la operación.

Función reciente() de visualization_tb que compara el último valor del precio de la vivienda (julio de 2020) con el mínimo y máximo históricos desde 2007 o desde que se tengan datos. La función toma como parámetros el dataframe a analizar, la lista de listas de las columnas a comparar (ámbito geográfico) y el tipo de operación ("alquiler" o "venta").

HIPÓTESIS:

La hipótesis de partida es que la crisis provocada por el coronavirus, aunque en principio podría pensarse que supondría un derrumbamiento de los precios de las viviendas, en realidad no ha tenido grandes efectos, sobre todo en las ventas y tampoco demasiado en los alquileres.

OPCIÓN C

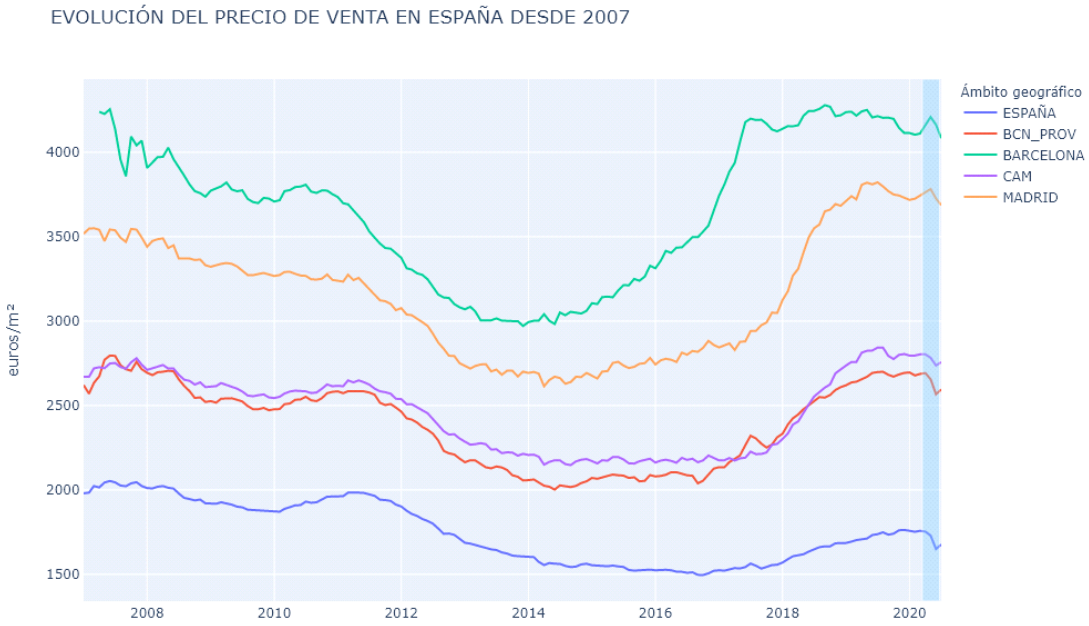
1. Cada paso y código está documentado tanto en el archivo main.ipynb como en los módulos correspondientes.

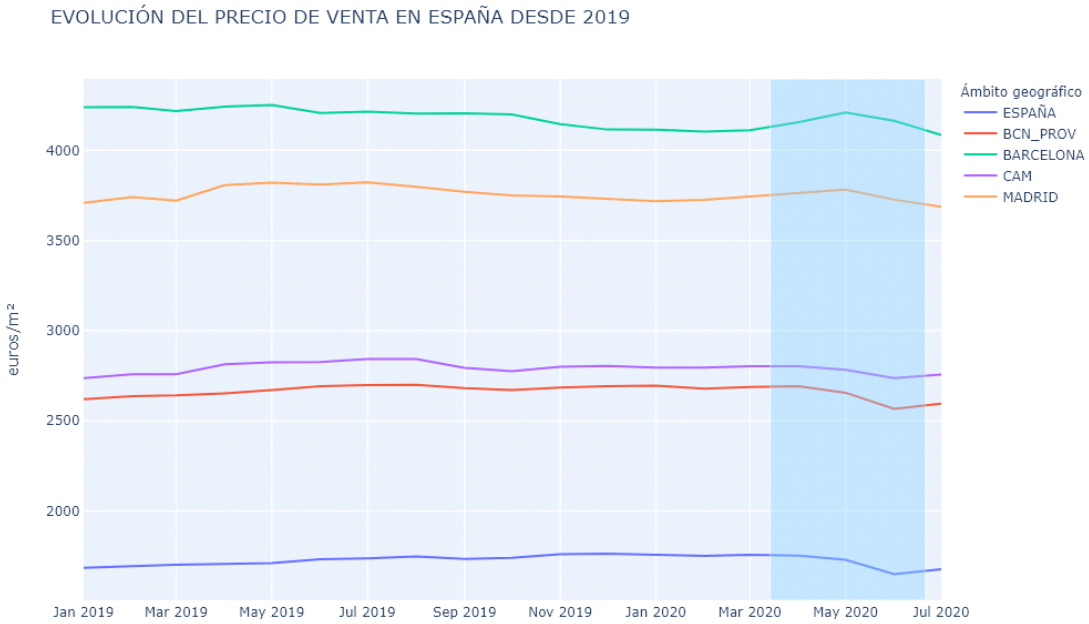
2. Datos recolectados en archivos Excel con los datos más recientes. No es posible hacer actualizaciones automáticas porque la url no admite webscrapping (de momento).

3. Ha sido necesario limpiar los datos, no en el sentido de eliminar o cambiar valores sino en el sentido de cambiar el tipo de datos (de string a float) y de limpiar los caracteres alfanuméricos previamente.

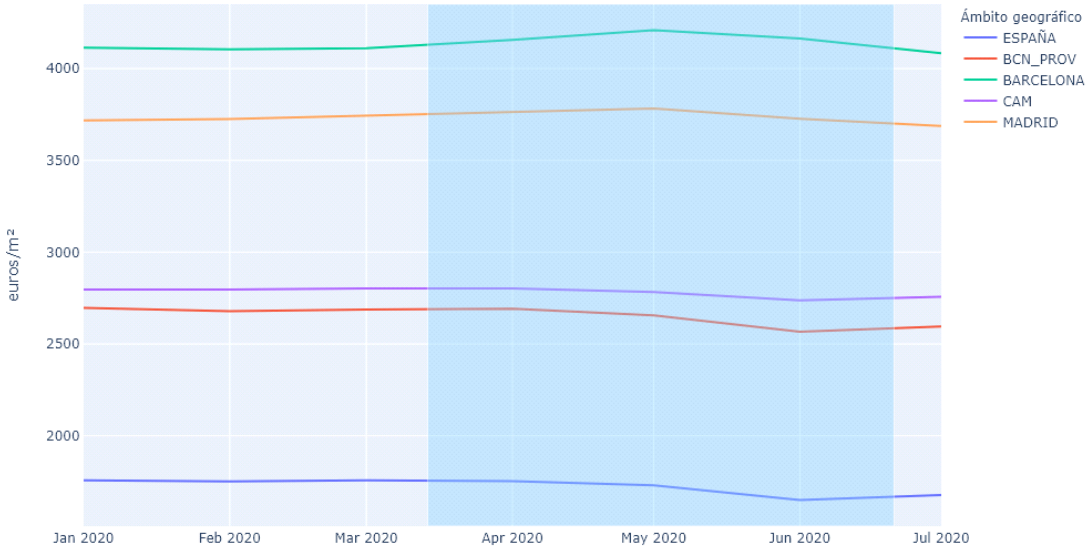
4. He creado una API que retorna un json con los datos limpios.

5. Tendencias. En cada gráfico, sombreado en azul se encuentra marcado el estado de alarma decretado en España, desde el 14 de marzo al 21 de junio de 2020.

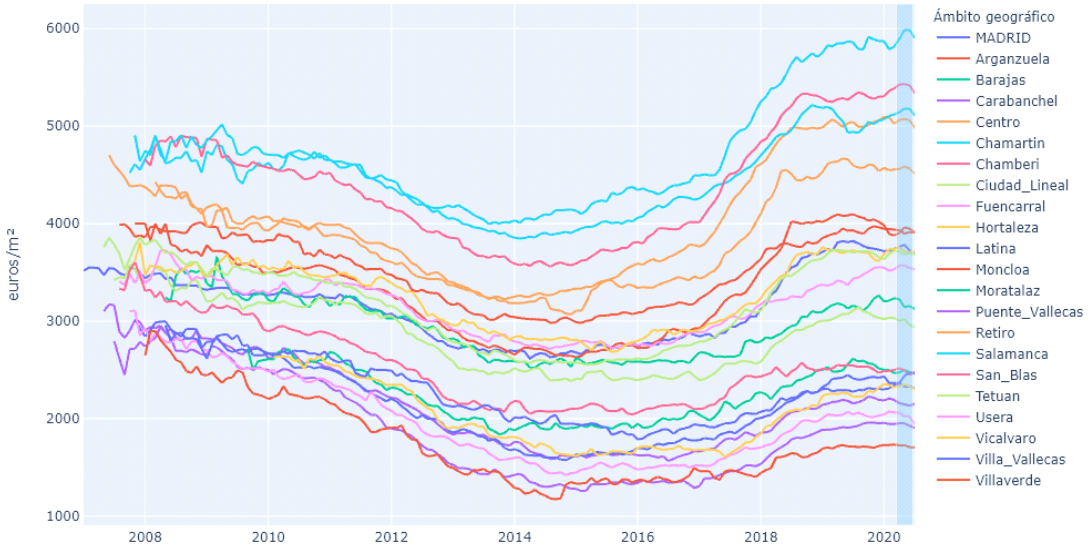




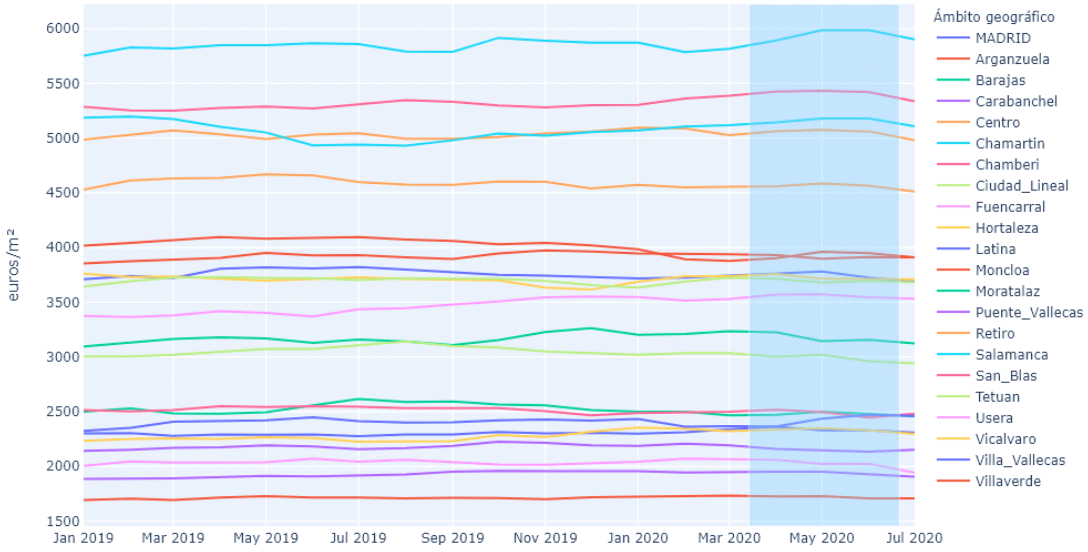
EVOLUCIÓN DEL PRECIO DE VENTA EN ESPAÑA DESDE 2020



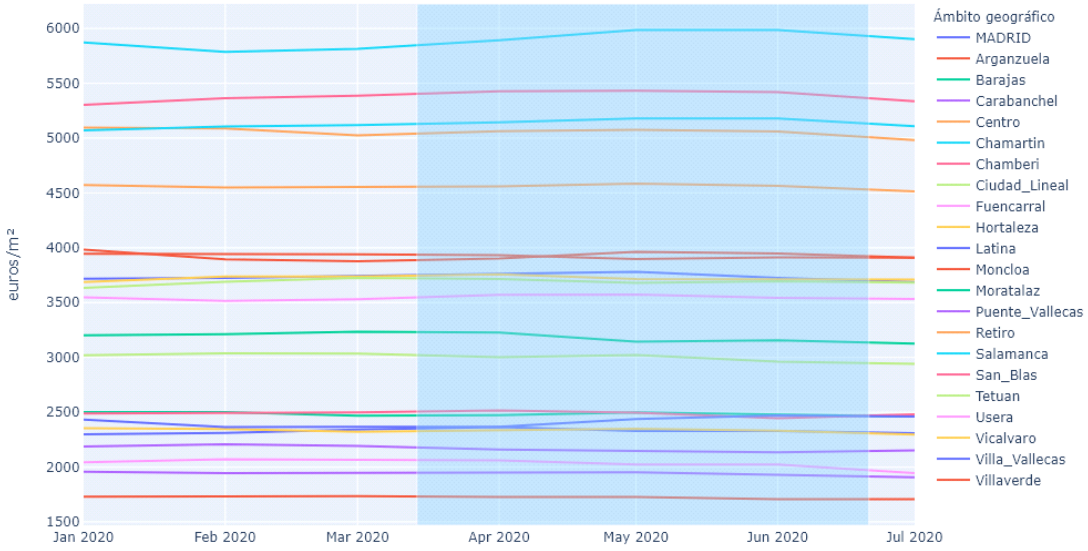
EVOLUCIÓN DEL PRECIO DE VENTA EN MADRID DESDE 2007



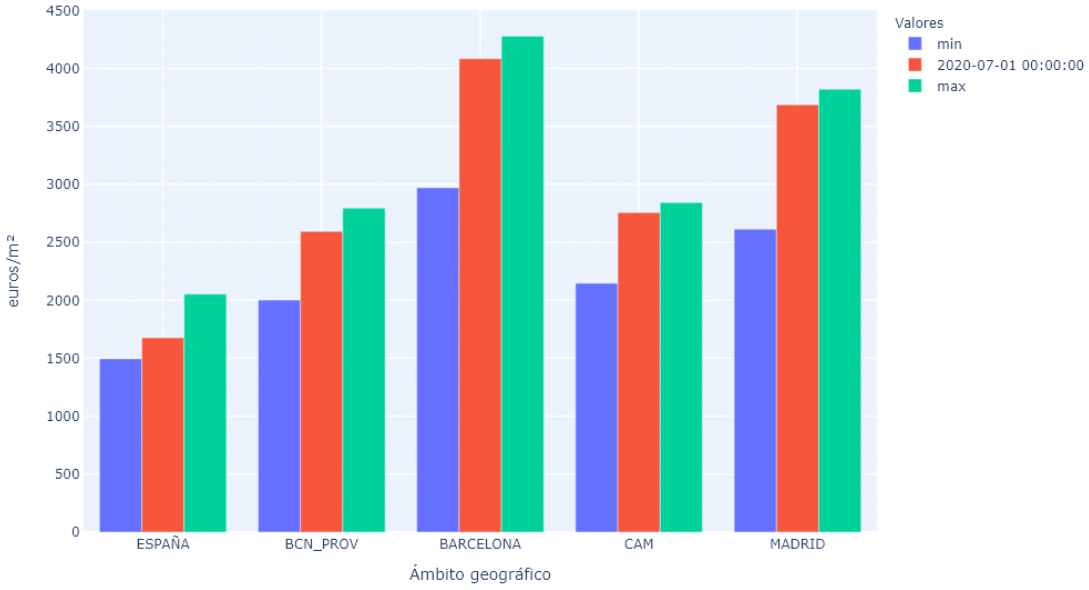
EVOLUCIÓN DEL PRECIO DE VENTA EN MADRID DESDE 2019



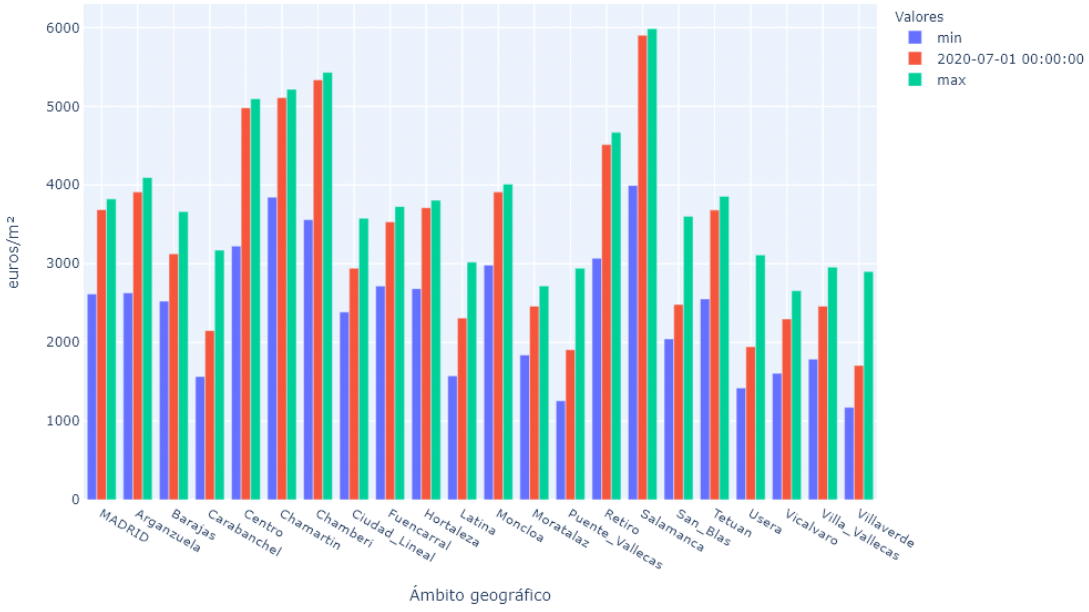
EVOLUCIÓN DEL PRECIO DE VENTA EN MADRID DESDE 2020



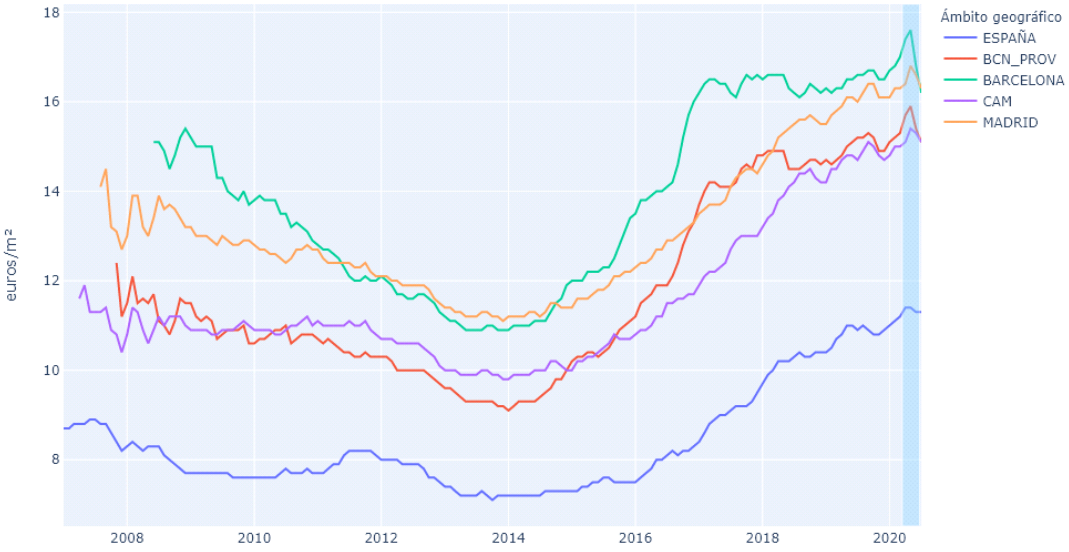
COMPARACIÓN RECIENTE CON EL MÍNIMO Y MÁXIMO HISTÓRICOS DE LA SERIE: VENTA EN ESPAÑA



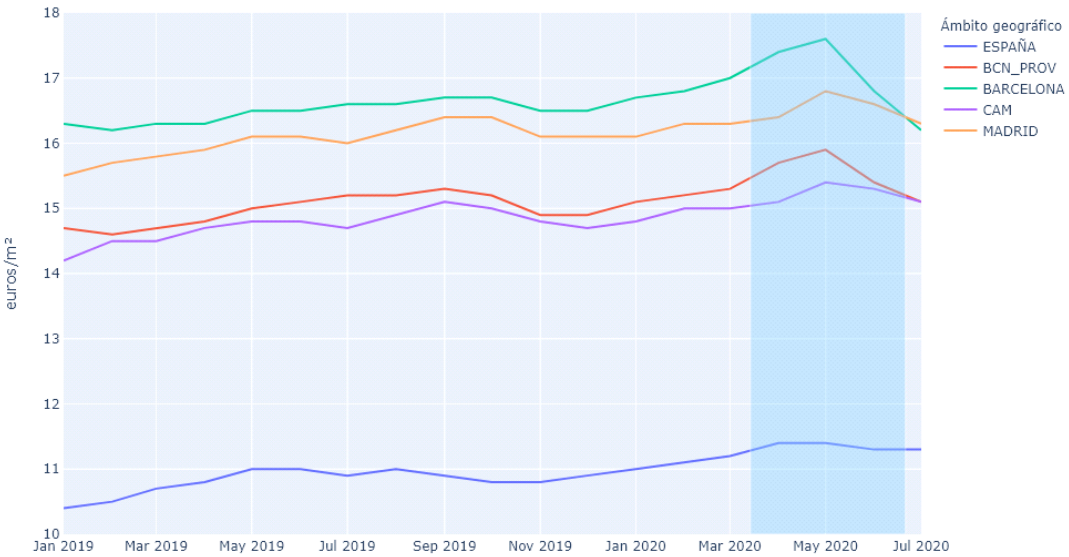
COMPARACIÓN RECIENTE CON EL MÍNIMO Y MÁXIMO HISTÓRICOS DE LA SERIE: VENTA EN MADRID



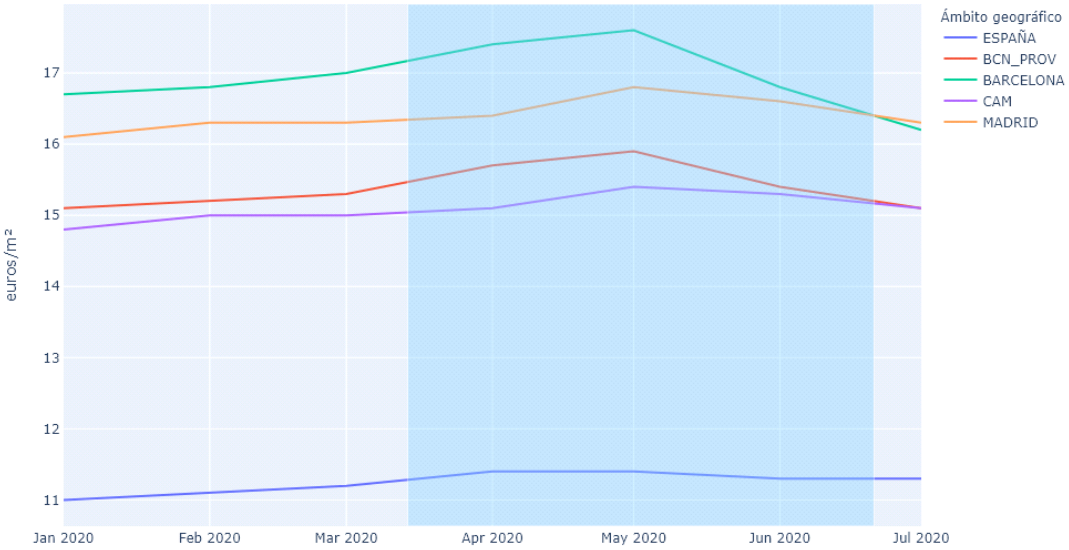
EVOLUCIÓN DEL PRECIO DE ALQUILER EN ESPAÑA DESDE 2007



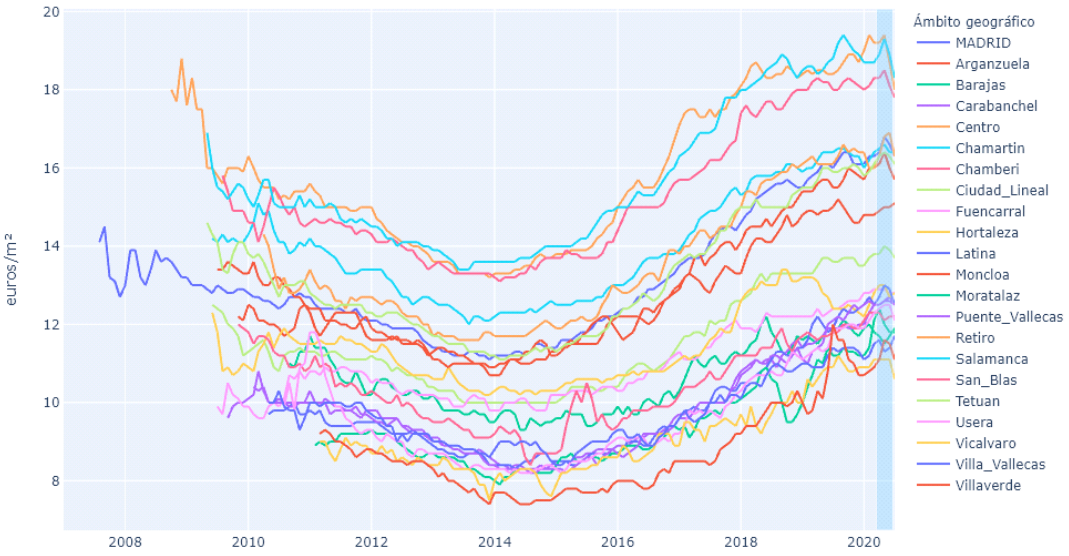
EVOLUCIÓN DEL PRECIO DE ALQUILER EN ESPAÑA DESDE 2019



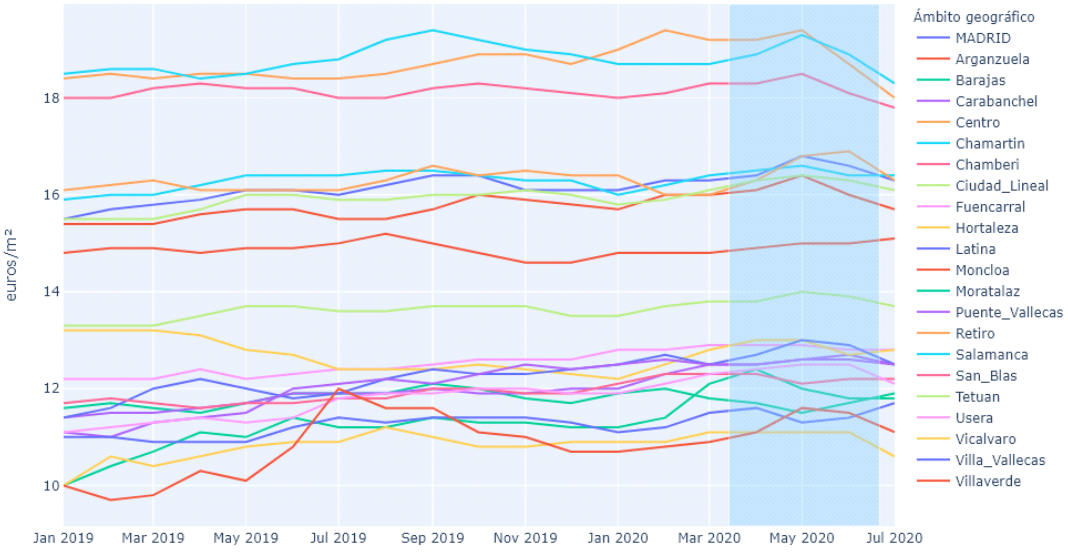
EVOLUCIÓN DEL PRECIO DE ALQUILER EN ESPAÑA DESDE 2020



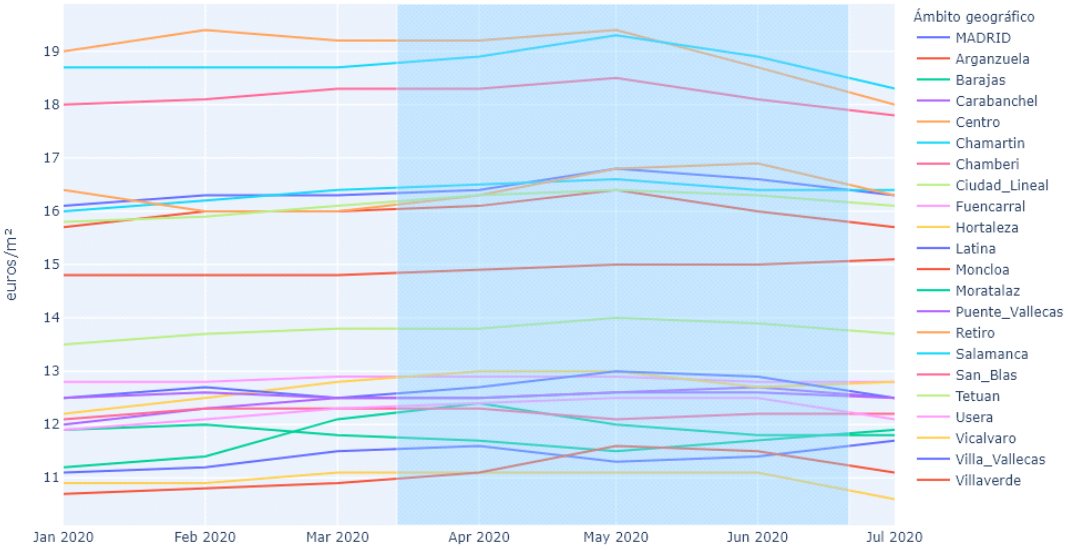
EVOLUCIÓN DEL PRECIO DE ALQUILER EN MADRID DESDE 2007



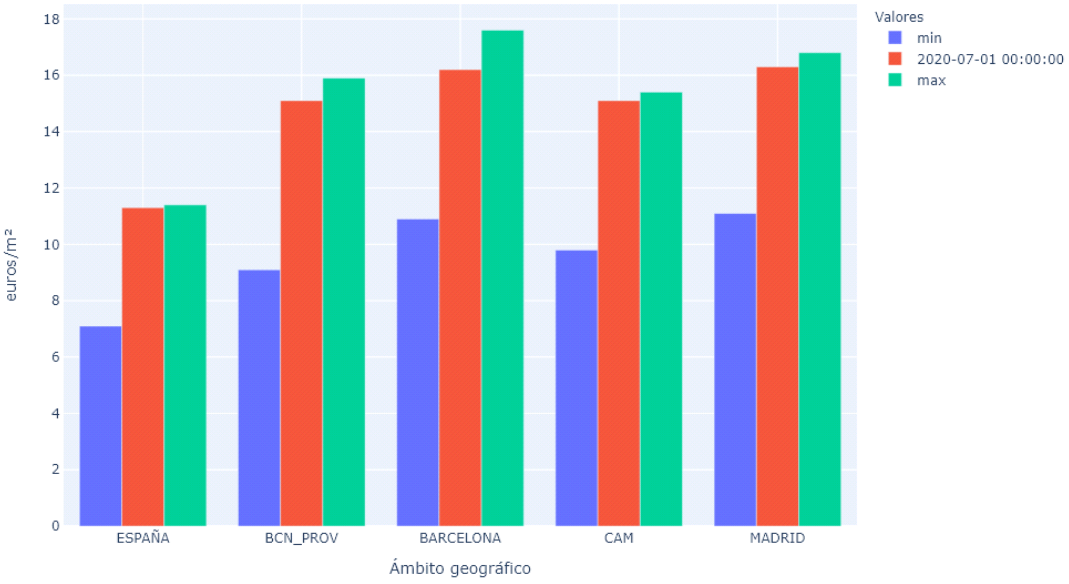
EVOLUCIÓN DEL PRECIO DE ALQUILER EN MADRID DESDE 2019



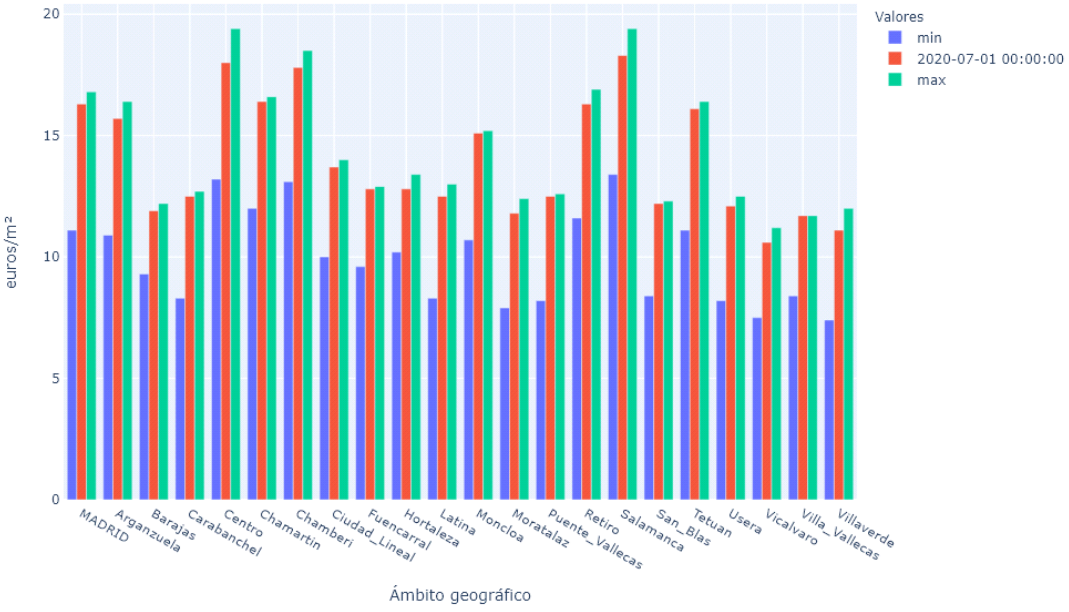
EVOLUCIÓN DEL PRECIO DE ALQUILER EN MADRID DESDE 2020



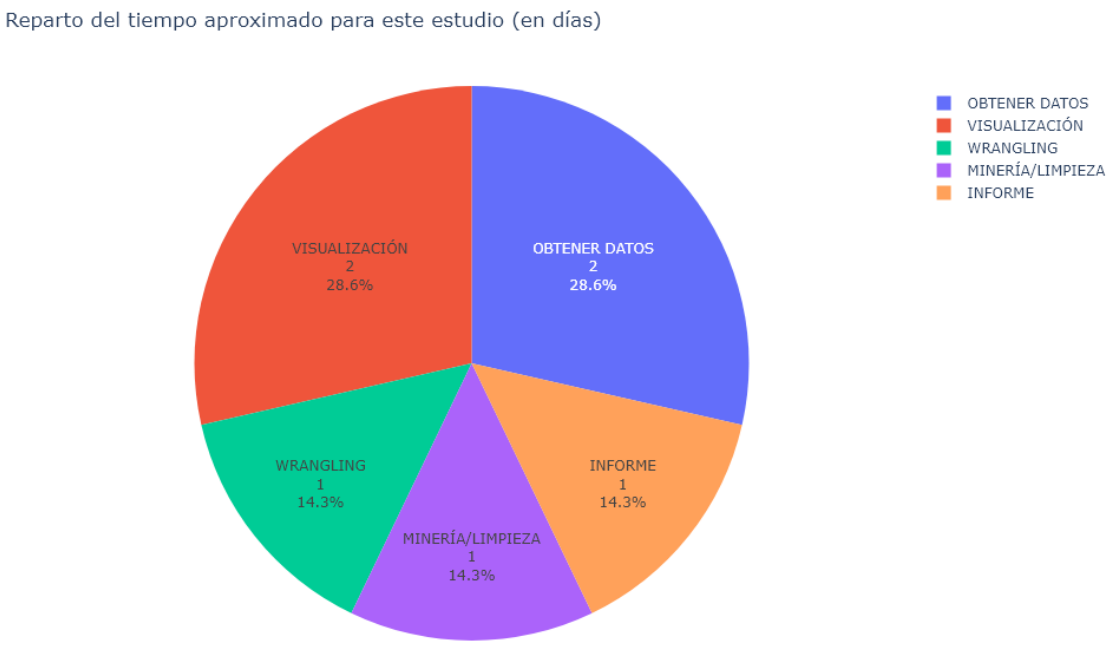
COMPARACIÓN RECIENTE CON EL MÍNIMO Y MÁXIMO HISTÓRICOS DE LA SERIE: ALQUILER EN ESPAÑA



COMPARACIÓN RECIENTE CON EL MÍNIMO Y MÁXIMO HISTÓRICOS DE LA SERIE: ALQUILER EN MADRID



6. Gráfico con el tiempo dedicado a cada parte de este estudio (7 días).



7 Preguntas.

7.a. ¿Ha sido posible demostrar la hipótesis? ¿Por qué?

Ha sido posible demostrar la hipótesis no sólo de que no ha habido un derrumbe de los precios de la vivienda, sino que en un primer momento incluso hubo hasta un crecimiento (en mayo, más acusado en el alquiler, pero también en las ventas) que posteriormente o se ha estabilizado o ha bajado hasta valores similares al inicio de la pandemia.

7.b. Conclusiones del estudio.

A pesar de la crisis universal generada por la pandemia, el mercado de la vivienda en España sigue su curso, al menos por ahora.

Los precios tanto en venta como en alquiler seguían una curva estable ligeramente ascendente poco antes del coronavirus. Con el estado de alarma los precios no sólo no bajaron, sino que iniciaron una aceleración ascendente que alcanzó su máximo en el mes de mayo. Después se inició un leve descenso que actualmente

(julio 2020) se encuentra aproximadamente en los mismos niveles que al inicio de la pandemia.

7.c. ¿Qué cambiaría si necesito hacer otro proyecto EDA?

En este caso me he dado cuenta de que habría necesitado otro tipo de datos relacionados con el tema para sacar unas conclusiones más exactas y detalladas. Sigo pendiente de que se publiquen los datos relativos a las hipotecas del segundo trimestre de 2020 y sé que habría sido muy importante contar con los datos económicos de cada distrito o ciudad en concreto. Este proyecto queda abierto a nuevas actualizaciones posteriores.

7.d.¿Qué he aprendido haciendo este proyecto?

La próxima vez no perderé el tiempo tratando de hacer cosas que luego no me van a aportar nada (como acceder a la API de Idealista y que los datos que obtenga no sean valiosos), intentaré comprobar mejor antes su utilidad.

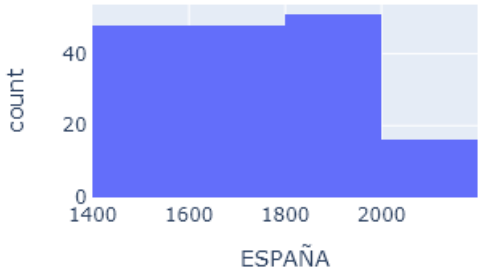
Utilizar distintas fuentes de datos desde un principio para obtener relaciones entre cada una de ellas.

OPCIÓN B

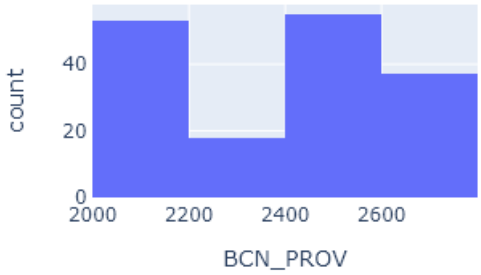
1. Histograma de cada columna, con bines=5 ¿Cómo son los rangos?

En los histogramas de precios desde 2007 se ve la distribución de los rangos de precios para los distintos ámbitos (España, Barcelona y Madrid). Viendo una frecuencia más alta de precios altos en las dos ciudades y muy por detrás las provincias, aunque siempre muy por encima de la media nacional. Esto en cuanto a venta, porque en cuanto a alquiler el elevado precio de los pisos en Barcelona y mucho más acusado que en Madrid, estando ambas también por encima de la media nacional.

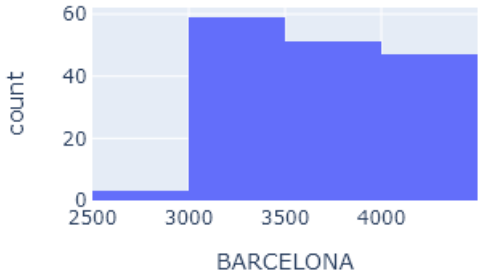
Precios venta desde 2007 en venta



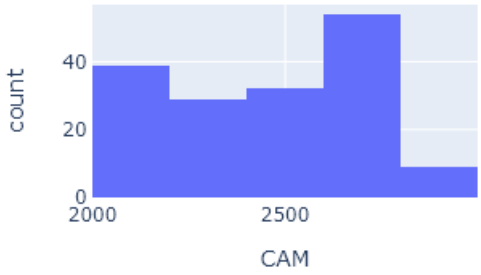
Precios venta desde 2007 en venta



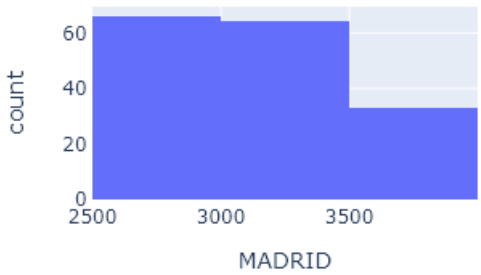
Precios venta desde 2007 en venta



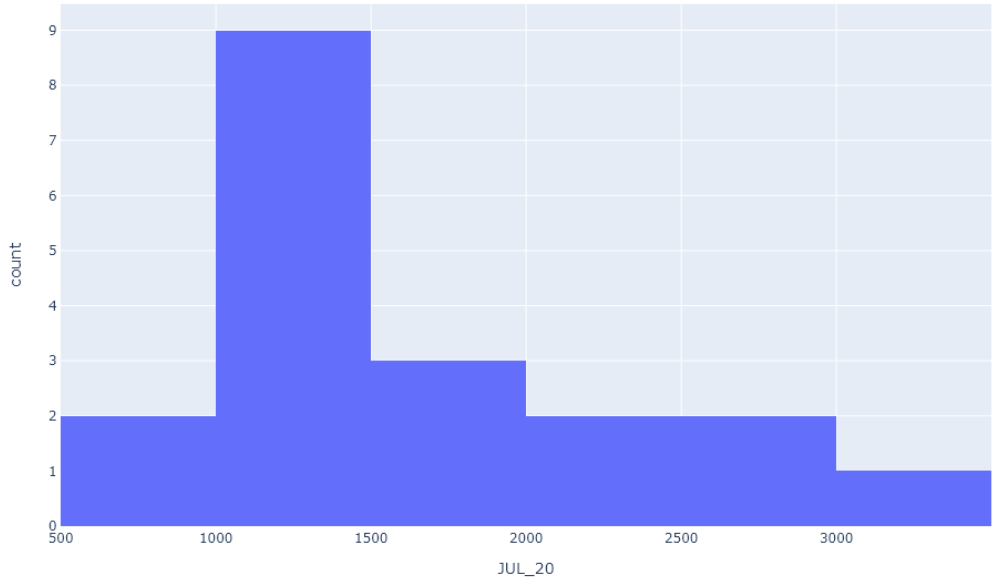
Precios venta desde 2007 en venta



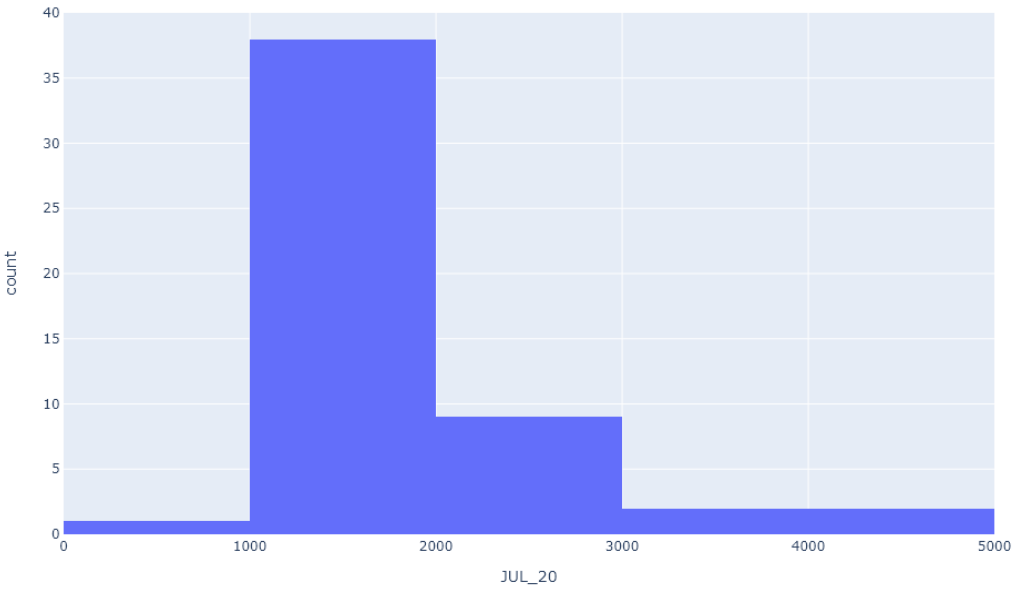
Precios venta desde 2007 en venta



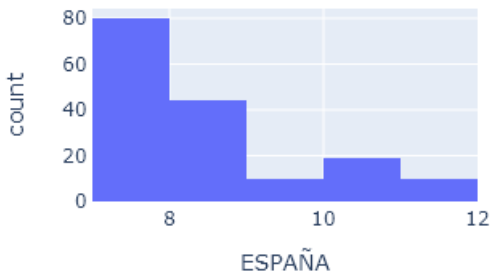
Distribución en julio de 2020 de los precios de venta por AUTONOMIA



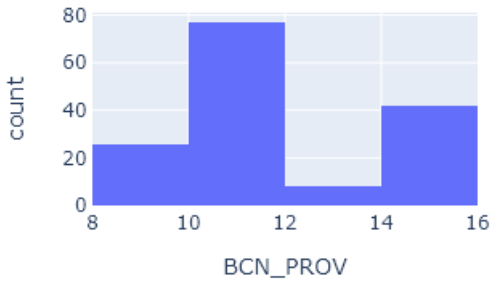
Distribución en julio de 2020 de los precios de venta por CIUDAD



Precios alquiler desde 2007 en alquiler

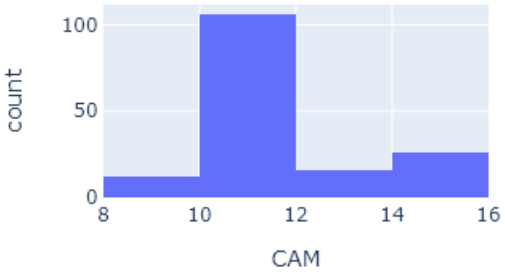


Precios alquiler desde 2007 en alquiler

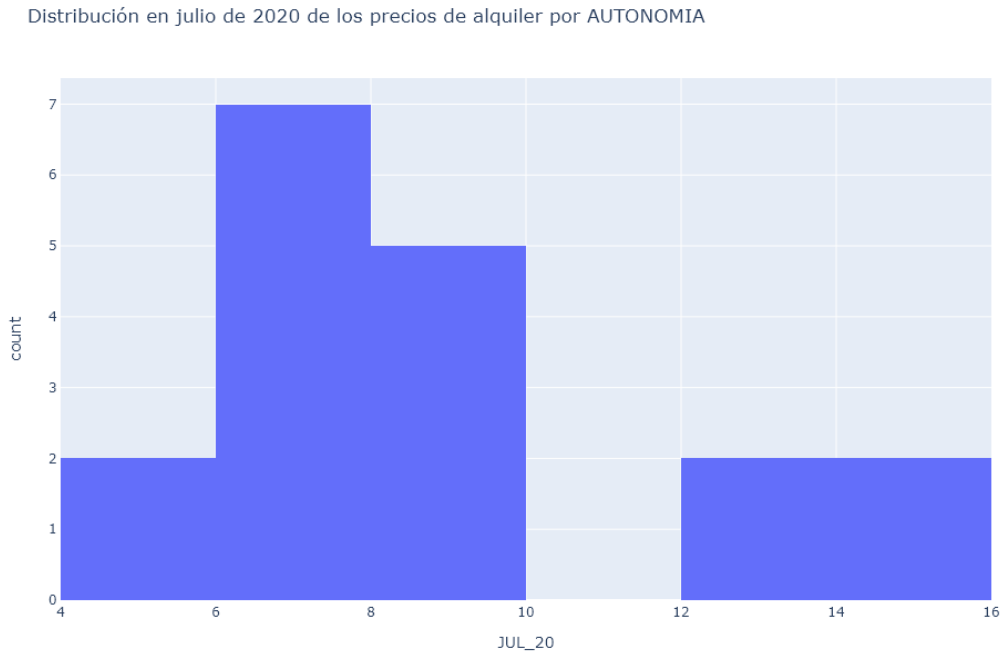


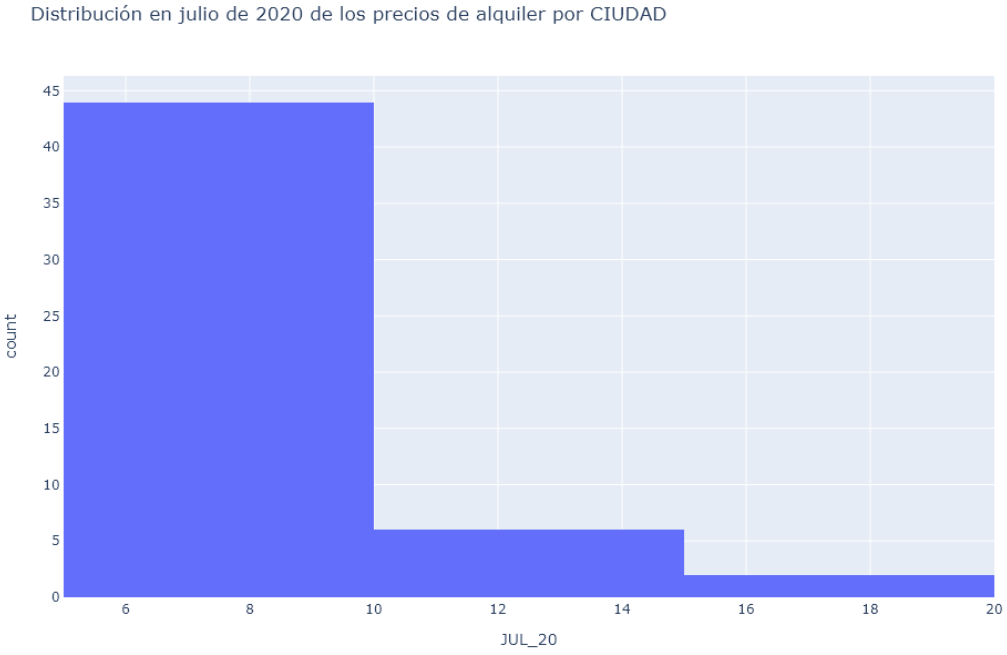
Precios alquiler desde 2007 en alquiler

Precios alquiler desde 2007 en alquiler

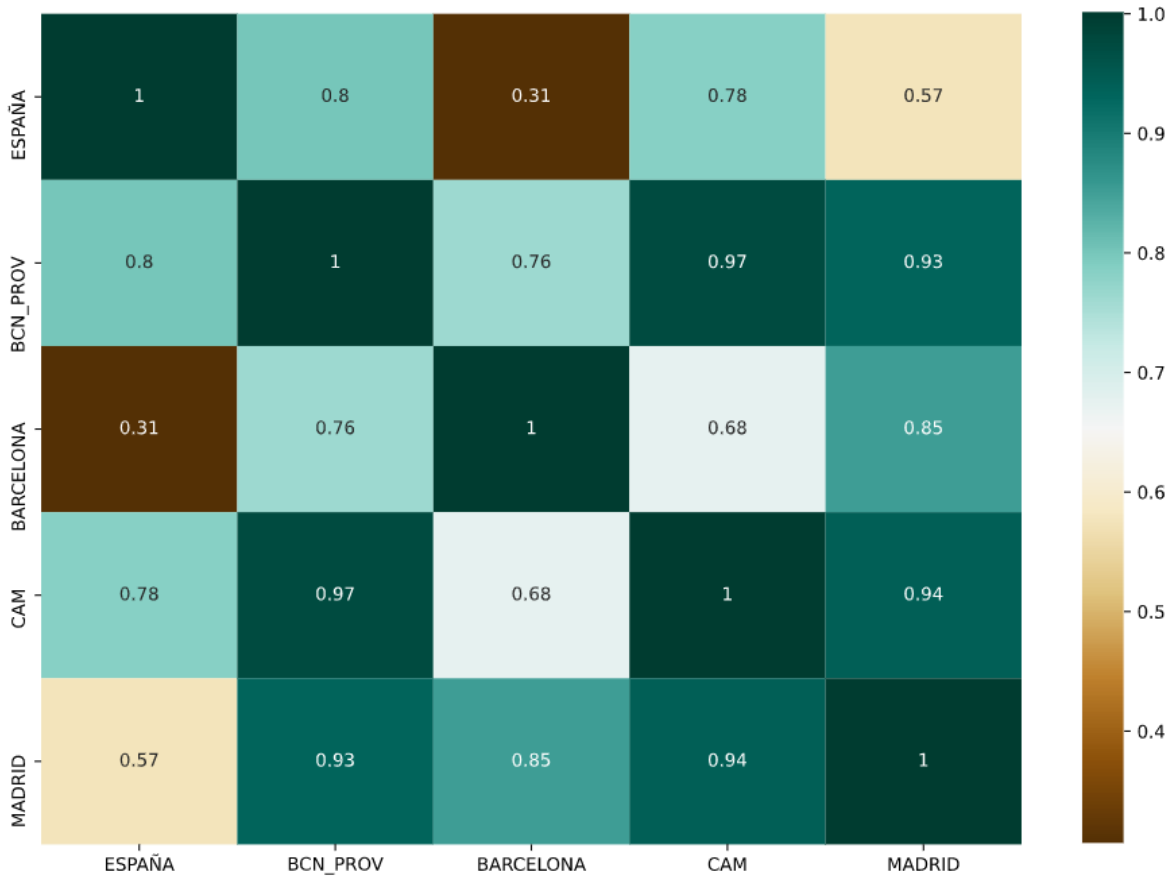


Precios alquiler desde 2007 en alquiler

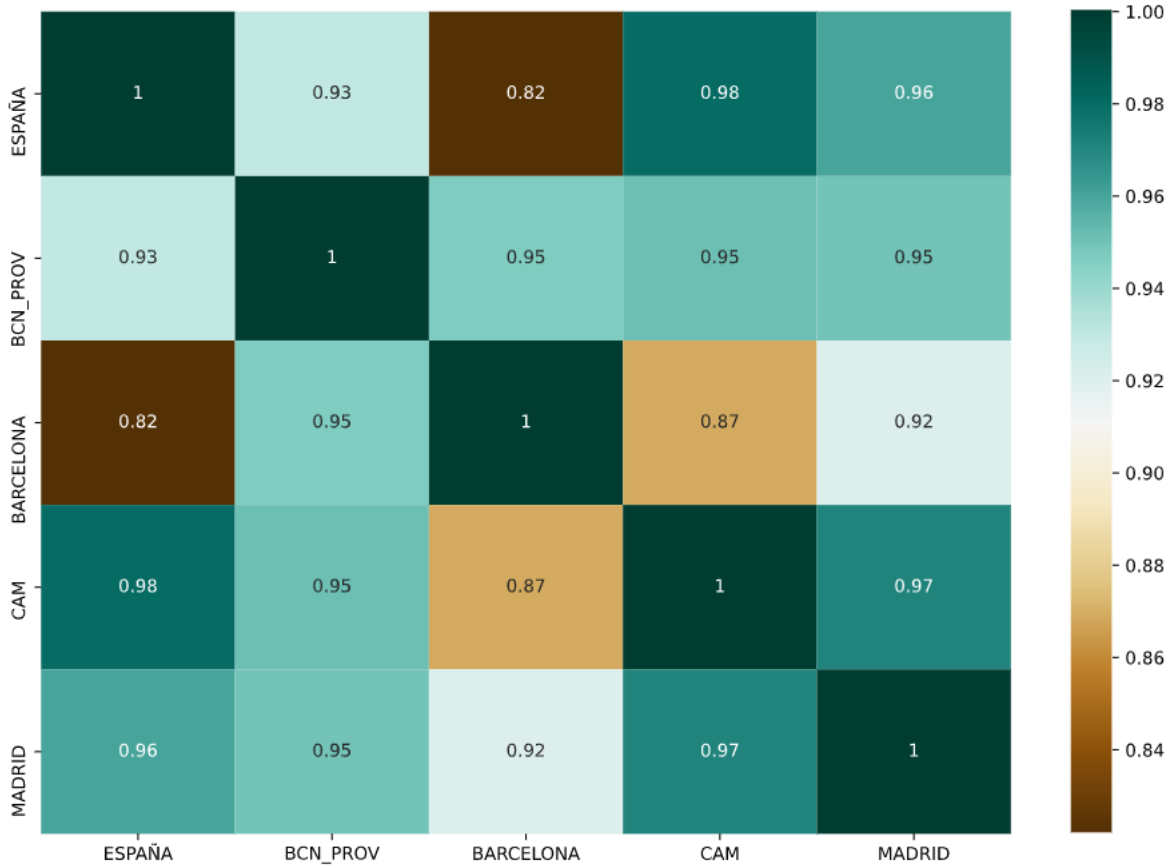




2. Matriz de correlación. ¿Cuáles son las columnas con mayor correlación?



En el caso de la venta, compruebo que Barcelona ciudad destaca en cuanto a poco relacionada con España.



En cuanto al alquiler, todos los ámbitos parecen estar relacionados proporcionalmente.

3. Usar funciones Matplotlib en vez de pandas. Pues no he usado pandas, pero sólo he usado Matplotlib con las matrices de correlación (y Seaborn), sobre todo he usado funciones de Plotly porque a pesar de tener mayor complejidad me ha permitido crear gráficos interactivos que he documentado en el apartado de visualización.

OPCIÓN A

1. Cada gráfico está guardado en su propio archivo dentro de la carpeta correspondiente a su operación (venta o alquiler) y con un nombre descriptivo, tanto en .png como en .html.

2. He usado módulos para cada funcionalidad. El archivo main.ipynb no tiene ningún loop o función (algún método muy básico y puntual como renombrar una variable).

3. En vez de Matplotlib o Seaborn he usado Plotly porque, como comenté en el apartado B, a pesar de tener mayor complejidad me ha permitido crear gráficos interactivos que he documentado en el apartado de visualización (Matplotlib y Seaborn sólo en las matrices de correlación).

4. Preguntas.

4.a. ¿Hay outliers o datos extraños?

No hay outliers los valores trazan curvas que crecen y decrecen a ritmos diferentes, pero dentro de cierta lógica y armonía.

No hay datos extraños. Como he comentado al explicar la minería de datos, hay columnas con datos ausentes (las correspondientes a los distritos de Madrid) porque se comenzó a recopilar datos en fecha posteriores a 2007. No he borrado las filas ni he convertido los valores a cero para mejorar los gráficos, viendo que convirtiendo los valores a float podía trabajar con ellos.

4.b. ¿Cuáles son las columnas con más valores repetidos? No es que haya valores repetidos, sino que cada columna (autonomía, ciudad, distrito...) tiene sus propias tendencias en determinadas épocas (sus propios rangos de precios que suelen mantener la misma diferencia respecto a las otras columnas).

OPCIÓN A+

Esta noche (31 de agosto de 2020) intentaré hacer un Pull Request del proyecto. Espero que me haya dado tiempo.