# EXAMEN FINAL 2024: DESARROLLO DE APLICACIONES PARA LA VISUALIZACIÓN DE DATOS


**CARLES OLUCHA ROYO**

# Contenido

# INTRODUCCIÓN

Para la realización de este análisis se ha utilizado un dataset que contiene los diferentes precios por metro cuadrado clasificándolos por los diferentes barrios de Madrid y atendiendo al momento en el que se tomó la medida y con sus correspondientes variables económicas y sobre las reviews de la misma zona.

Lo primero que hemos realizado antes de hacer las visualizaciones es observar los tipos de datos que se obtenían del mismo y que no hubiera valores nulos. En este caso, se puede observar que no había ningún valor nulo:

```
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   neighbourhood_group  2218 non-null   object 
 1   date                 2218 non-null   object 
 2   m2_price             2218 non-null   float64
 3   inflation            2218 non-null   float64
 4   HICP                 2218 non-null   float64
 5   population_density   2218 non-null   int64  
 6   listings_count       2218 non-null   int64  
 7   minimum_nights       2218 non-null   float64
 8   nigth_price          2218 non-null   float64
 9   availability_365     2218 non-null   float64
 10  listing_reviews      2218 non-null   int64  
 11  number_of_reviews    2218 non-null   float64
 12  reviews_per_month    2218 non-null   float64
 13  hosts_count          2218 non-null   int64  
 14  Private_room         2218 non-null   int64  
 15  Entire_home          2218 non-null   int64  
 16  Hotel_room           2218 non-null   int64  
 17  Shared_room          2218 non-null   int64  
dtypes: float64(8), int64(8), object(2)
```

```
neighbourhood_group    0
date                   0
m2_price               0
inflation              0
HICP                   0
population_density     0
listings_count         0
minimum_nights         0
nigth_price            0
availability_365       0
listing_reviews        0
number_of_reviews      0
reviews_per_month      0
hosts_count            0
Private_room           0
Entire_home            0
Hotel_room             0
Shared_room            0
dtype: int64
```

# VISUALIZACIONES

Lo primero que hemos pensado que era interesante es conocer el precio general del metro cuadrado en Madrid, para hacernos una idea de cómo se distribuye:

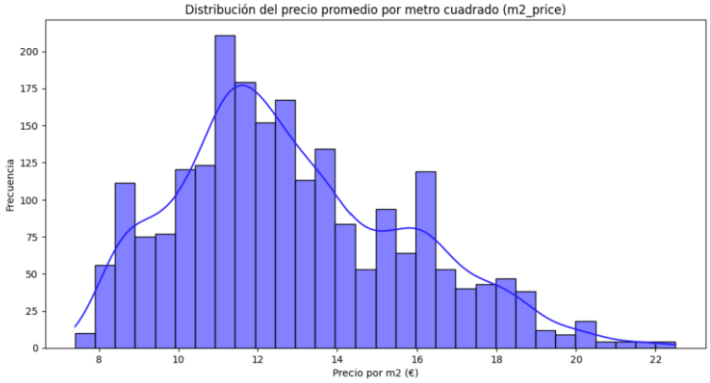
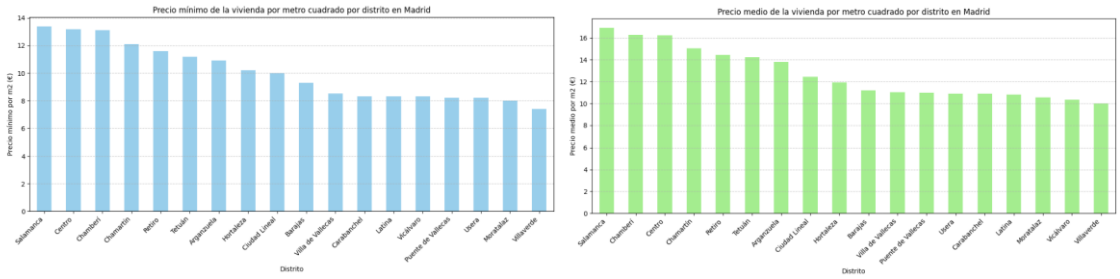



*Ilustración 1: Distribución de los precios*

En este primer gráfico se observa que existe un poco de asimetría a la derecha por lo que la Moda se encuentra más a la izquierda de la mediana de los datos. Esto se traduce

en que hay un gran número de pisos que se encuentran entre 11 y 13€ el metro cuadrado.

Una vez hemos este análisis general de los precios en Madrid, procedemos a observar el precio mínimo y medio por metro cuadrado distribuido por los diferentes barrios de Madrid.




En este caso podemos observar que el barrio con precio medio más alto y precio mínimo más alto en ambos casos es el barrio Salamanca y que por otro lado está Villaverde que es el que tiene menor precio mínimo y medio.

Podemos observar, que, aunque las diferencias entre posición de ranking son muy pequeñas.

Con esto ya sabido, hemos estado observando el precio por barrio con el paso del tiempo en la siguiente visualización

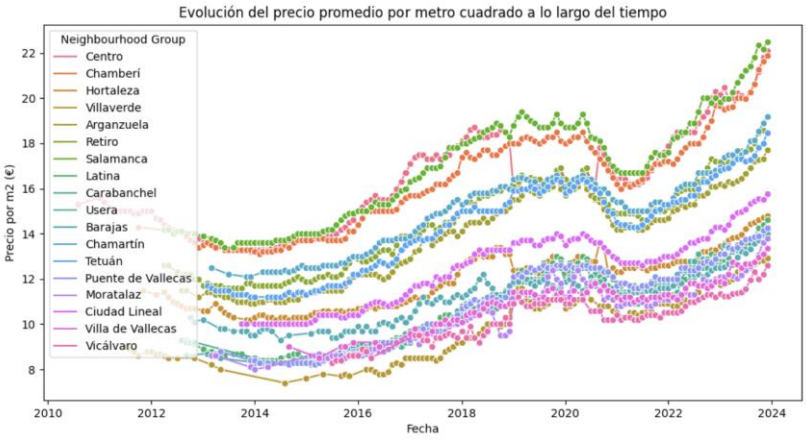
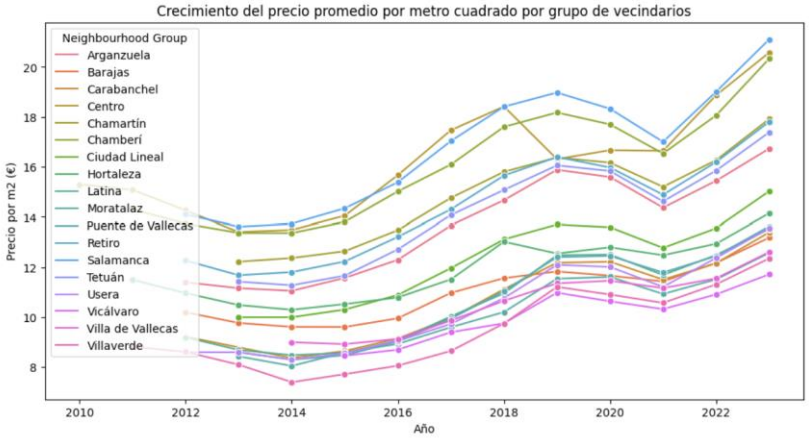
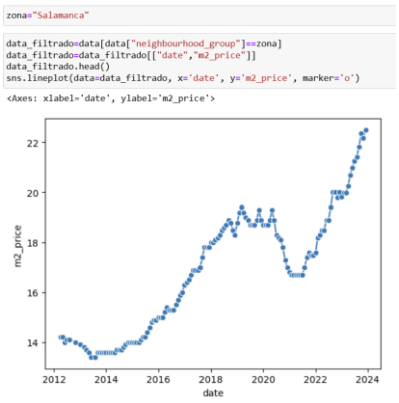



*Ilustración 2: evolución del precio por tiempo y zona*




En esta visualización podemos observar que más o menos todos los barrios han seguido la misma tendencia de forma paralela en sus respectivos precios. Con esto quiero decir que en general han empezado en 2010 a bajar hasta más o menos 2015, para después subir hasta 2020 y debido a la crisis del covid se produzca una pequeña bajada hasta 2021 debido a que la gente se movía menos de sus casas, para finalmente ir subiendo de forma muy rápida hasta 2024.

Como hay muchas líneas en ese gráfico, hemos querido hacer un gráfico en el que pudieras seleccionar un barrio y que saliera solamente ese barrio en concreto:




Como podemos ver, la tendencia mencionada anteriormente sirve para el barrio Salamanca.

Una cosa que se hace muy visible es cuando representamos el precio por noche en los diferentes barrios de Madrid es la crecida tan grande que ha tenido Villaverde en Madrid en 2019, pero después se produjo una caída en 2020.

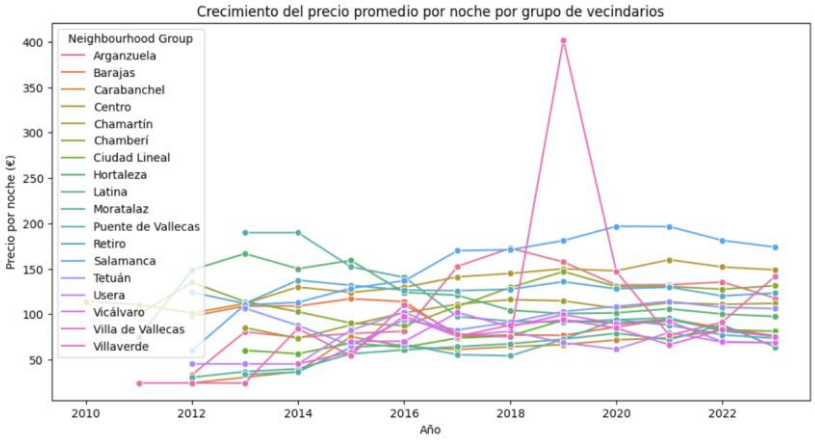
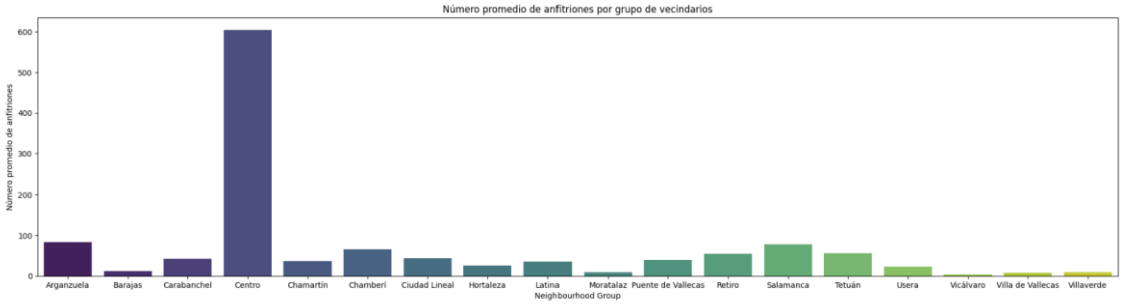


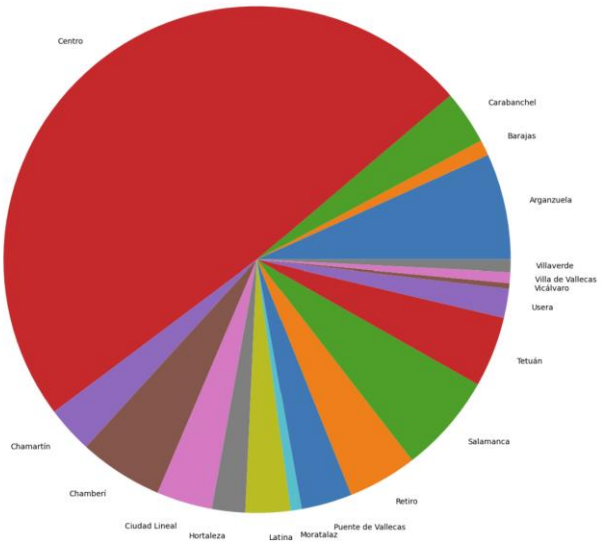

*Ilustración 3: crecimiento del precio por noche*

También considero que es interesante observar en cada vecindario el número total de anfitriones activos en un vecindario porque de esta manera también podemos observar las zonas que tienen una mayor oferta (cosa que se suele venir dado por la demanda).




Como era de esperar donde hay más anfitriones en todo Madrid es en el Centro (toda la zona de Sol).

En este gráfico se puede observar que casi la mitad de los anfitriones se encuentran en la zona de Madrid Centro.



Finalmente, hemos querido hacer un análisis de las variables económicas para ver cómo influye:

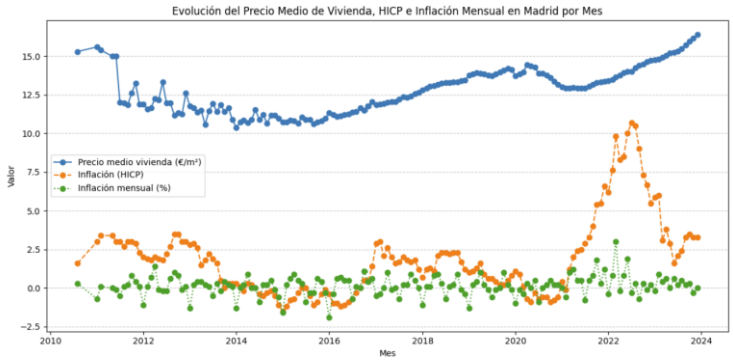



*Ilustración 4: comparación del precio, inflación y HICP*

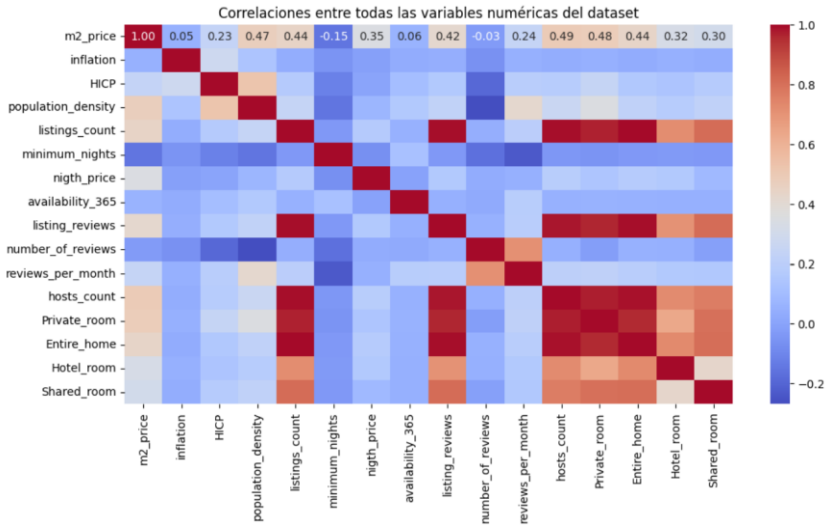
Podemos ver cómo afecta el HICP y la inflación al precio medio por metro cuadrado. Para entender este gráfico, debemos asimilar que el Índice de Precios al Consumo (IPC) y la inflación están estrechamente relacionados, ya que el IPC mide el cambio promedio en los precios de una canasta representativa de bienes y servicios adquiridos por los hogares a lo largo del tiempo, y la inflación se refiere al aumento sostenido en el nivel general de precios, reflejado típicamente en el IPC. Un aumento en el IPC indica que los bienes y servicios son más costosos, lo que implica una pérdida en el poder adquisitivo del dinero. La inflación afecta indirectamente el mercado inmobiliario porque influye en las tasas de interés, la capacidad de ahorro de los consumidores y la percepción de estabilidad económica. En general, cuando la inflación aumenta, los costos de construcción también tienden a subir, lo que puede presionar los precios de las viviendas al alza y a su vez de los de alquiler o turísticos.

Si la inflación es alta y persistente, las tasas de interés suelen incrementarse, encareciendo las hipotecas y disminuyendo la demanda de vivienda, lo que podría estabilizar o reducir los precios. Sin embargo, en mercados con oferta limitada, los precios de la vivienda pueden seguir subiendo a pesar de la inflación. Además, cuando la inflación es moderada y predecible, los bienes raíces son percibidos como una inversión segura frente a la depreciación del dinero, lo que puede aumentar la demanda y, por ende, los precios. En resumen, el IPC y la inflación afectan tanto los costos como la percepción del valor de la vivienda, modulando su precio de manera directa e indirecta.

# MODELOS REALIZADOS

## Regresión Lineal

Una vez hemos realizado todo el análisis de datos, hemos hecho una matriz de correlación para de esta manera observar aquellas variables que mejor se adaptan y entrenan el modelo.

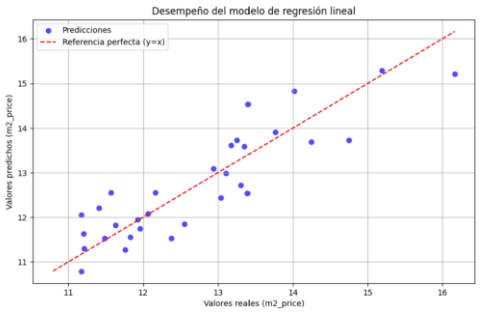



En esta matriz de correlación podemos observar que de las cosas que tienen mayor correlación positiva sobre el precio por metro cuadrado son: population_density y listings_count, listing_revires, host_count, private_room, entire_home.

Por todo ello, lo primero que hemos hecho ha sido un modelo de regresión lineal que contuviera todas las variables del juego de datos para así poder observar aquellas que influyen más o influyen menos.

```
Coeficientes del modelo de regresión lineal:
              Variable  Coeficiente
0                 HICP     0.048465
1   population_density     0.119824
2       listings_count     0.054229
3       minimum_nights    -0.039002
4          nigth_price     0.016377
5     availability_365     0.007202
6      listing_reviews    -0.002844
7    number_of_reviews     0.024240
8    reviews_per_month    -3.010973
9          hosts_count    -0.021392
10        Private_room     0.026603
11         Entire_home    -0.053508
12          Hotel_room    -0.216813
13         Shared_room     0.297948

MSE (Error Cuadrático Medio): 0.343891067434884
R² Score: 0.7804273162720345
```




Podemos ver que las variables que más intervienen en el modelo son “population_density”, o “reviews_per_month” o “Shared_room” entre otras. Por ello, con esta informaicón hemos probado el siguiente modelo:

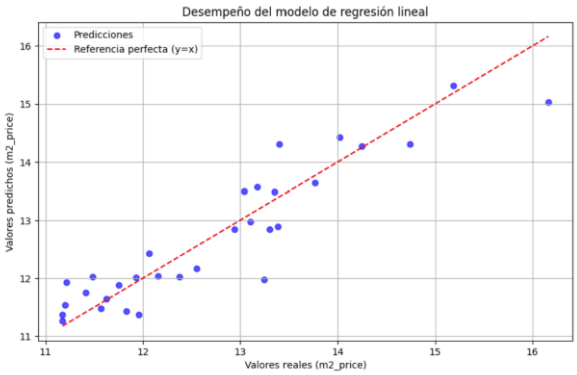
Después, en el que hemos usado las variables que nos parecían útiles para predecir la variable objetivo:

['nigth_price', 'inflation', 'availability_365', 'population_density', 'reviews_per_month']

Con este modelo hemos obtenido el siguiente resultado:

```
Coeficientes del modelo de regresión lineal:
             Variable  Coeficiente
0         nigth_price     0.011255
1           inflation    -0.056762
2    availability_365     0.004667
3  population_density     0.055571
4   reviews_per_month    -0.634367

MSE (Error Cuadrático Medio): 0.22601276694577668
R² Score: 0.8556920068752174
```




Lo cual podemos observar que es un buen modelo para ser capaces de predecir el precio por metro cuadrado medio.

Además, podemos ver que la variable que influye más en el modelo es el número de reviews por mes.

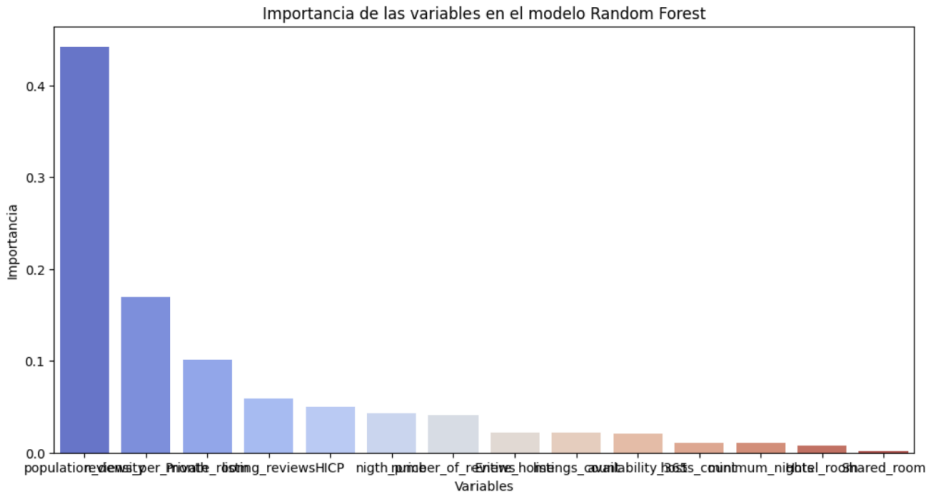
Por todo ello, hemos decidido que este es el mejor modelo de los que hemos probado a la hora de hacer regresión lineal.

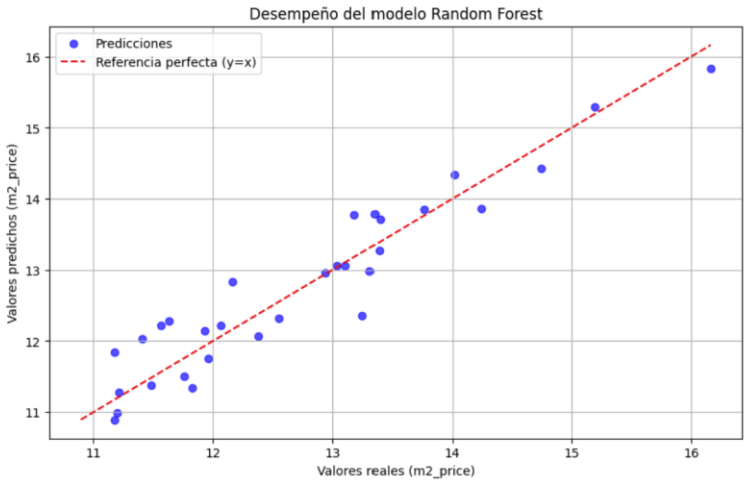
## Random Forest

Hemos seguido un procedimiento parecido aquí en el que hemos probado primero el modelo con todas las variables disponibles en el juego de datos obteniendo este resultado:

```
Importancia de las variables en el modelo Random Forest:
             Variable  Importancia
1   population_density     0.442315
8    reviews_per_month     0.169435
10        Private_room     0.100838
6      listing_reviews     0.059033
0                 HICP     0.050197
4          nigth_price     0.042585
7    number_of_reviews     0.040353
11         Entire_home     0.022042
2       listings_count     0.021475
5     availability_365     0.021130
9          hosts_count     0.010974
3       minimum_nights     0.010183
12          Hotel_room     0.007766
13         Shared_room     0.001673

MSE (Error Cuadrático Medio): 0.15698355856216906
R² Score: 0.8997668025756775
```

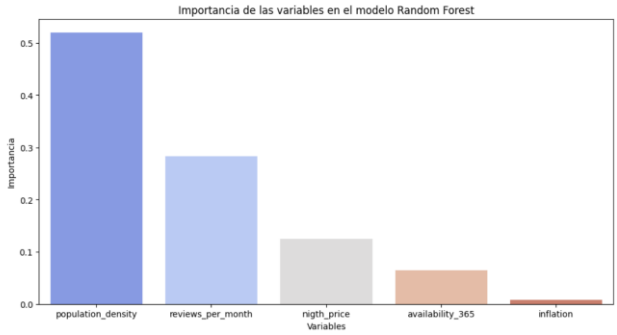


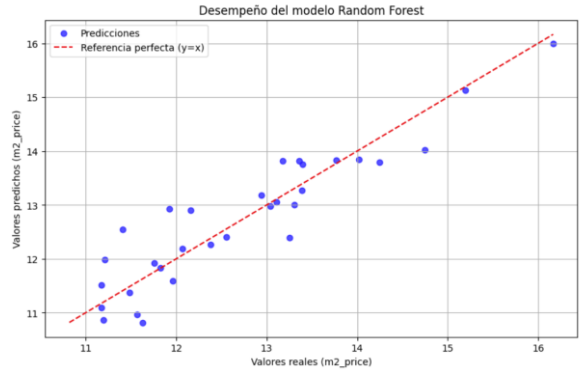




Por ello podemos ver que este modelo tiene una R^2 score mayor todavía que en el anterior y un error cuadrático medio menor, sin embargo hemos usado muchas variables y podemos ver que hay muchas de ellas que son muy poco útiles para el modelo. Por ello vamos a eliminar las variables que menos influyen para crear el modelo último del tipo random forest:

```
Importancia de las variables en el modelo Random Forest:
            Variable  Importancia
3  population_density     0.520313
4   reviews_per_month     0.282915
0         nigth_price     0.124465
2    availability_365     0.064635
1           inflation     0.007672

MSE (Error Cuadrático Medio): 0.23698159028776158
R² Score: 0.8486884693938939
```

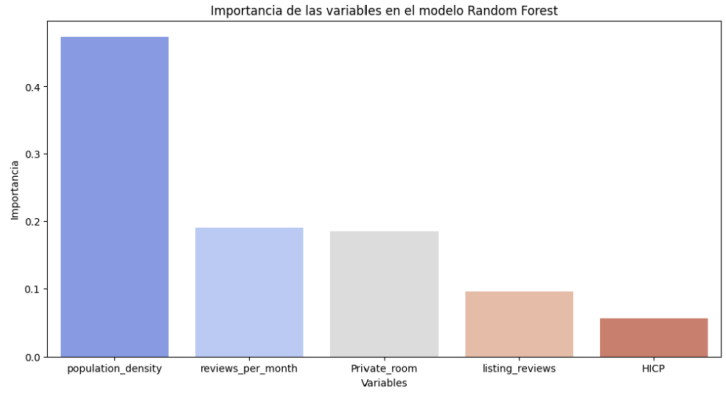


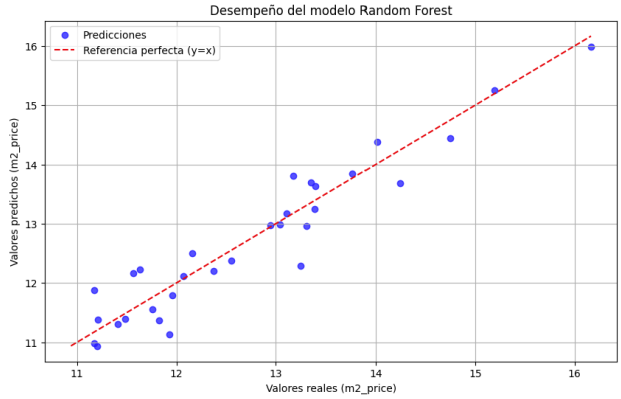




Otro modelo que hemos creado ha sido el siguiente en el que hemos seleccionado otras variables como son:

```
Importancia de las variables en el modelo Random Forest:
              Variable  Importancia
0  population_density     0.474148
1   reviews_per_month     0.190498
2        Private_room     0.184590
3     listing_reviews     0.095341
4                HICP     0.055423

MSE (Error Cuadrático Medio): 0.1513815136065372
R² Score: 0.903343679563056
```




Este modelo es mucho mejor que el anterior, ya que presenta un R^2 mayor y un mse menor reduciendo la complejidad del mismo.

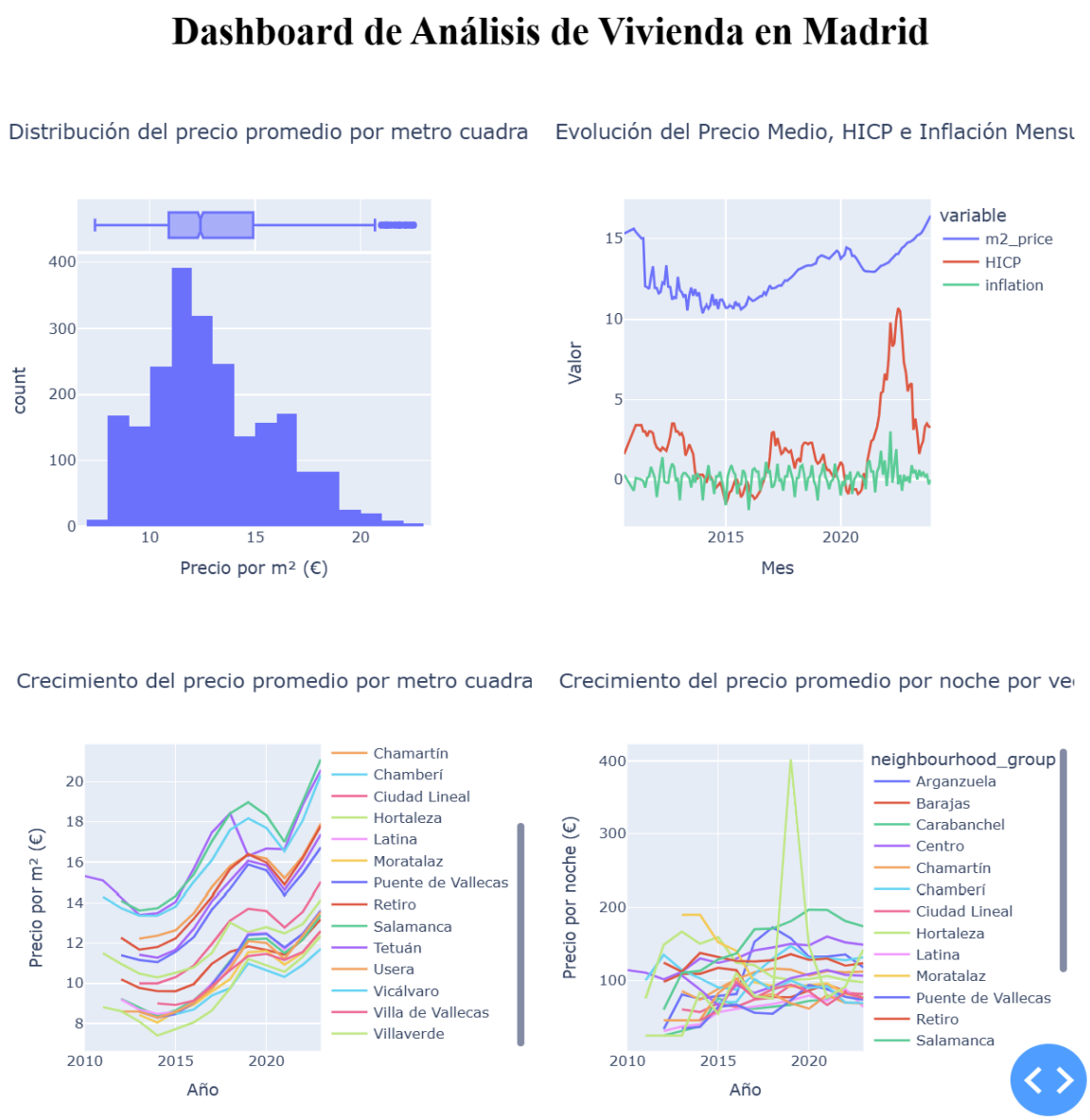
## Conclusión modelo

Por todo lo calculado, si tuviera que escoger un modelo para realizar los modelos utilizaría el último random forest creado ya que siendo muy simple por el número de variables que se necesitan es el que ofrece un mayor R^2 Score con más de un 0,9 y un menor error. Además, nos permite observar de una forma muy sencilla las variables que más influyen a la hora de predecir el precio por metro cuadrado.

Sin embargo, como el enunciado pide que se use regresión lineal múltiple utilizaría el generado con las variables de entrada 'nigth_price', 'inflation', 'availability_365', 'population_density', 'reviews_per_month' ya que es el que presenta menor error y se produce una mayor explicación de la variable output.

# DASHBOARD

Se adjunta una imagen del Dashboard realizado.




Este es el dashboard realizado, en el que he tomado las 4 gráficas que me han parecido más interesantes para explicar los datos que se nos han proporcionado. La explicación de cada gráfica está en el apartado de visualizaciones donde aparecen las mismas gráficas en las ilustraciones Ilustración 1, Ilustración 2, Ilustración 3, Ilustración 4.

# CONCLUSIONES Y RECOMENDACIONES

Por todo ello, si tuviera que realizar recomendaciones al ayuntamiento de Madrid sobre como rebajar los precios del alquiler, lo que haría sería tener en cuenta la densidad de la población de la zona donde quiero reducirlo ya que es de las variables que mayor correlación tiene con la variable de salida con un 0,47.

Además, hemos podido observar que hay zonas que requieren una mayor intervención, ya que en zonas como en el centro de Madrid es donde se encuentran la mayoría de los anfitriones de la ciudad y con diferencia (y aún así es de los precios más caros de la ciudad) y que existe una gran variabilidad de precios en Madrid ya que en el barrio más caro (Salamanca) el precio medio es de 16, mientras que en el más económico es un poco más de 10.

También hemos podido observar cómo ha ido cambiando el precio en el alquiler de la ciudad y que los cambios han sido parecidos como se puede ver en la ilustración de la segunda fila izquierda del dashboard.