

Memoria del Proyecto

Análisis Exploratorio de Datos –exploratory data analysis–

Integrantes

- Roberto Perez Esteban
- Ana Sofía Gomez Ramirez
- Agustín Arganin Castillo

Introducción

Contexto del proyecto

El análisis de datos se ha convertido en una herramienta fundamental para extraer valor a partir de grandes volúmenes de información y apoyar la toma de decisiones basada en evidencia. En el sector inmobiliario, el uso de datos permite identificar patrones de precios, comprender la oferta de vivienda y analizar el impacto de distintas características sobre el valor de los inmuebles.

En este proyecto, se realiza un Análisis Exploratorio de Datos (EDA) utilizando un dataset obtenido de a través de técnicas de scraping al sitio web Idealista, una de las plataformas inmobiliarias más relevantes en España, Italia y Portugal.

Objetivos del proyecto

Objetivo general

El objetivo principal del proyecto es aplicar de forma práctica las herramientas y técnicas aprendidas durante el **bootcamp de Data Science & IA**, incluyendo la carga de datos de archivos en diferentes formatos: `.txt` y `.csv`, limpieza de datos y transformación de los datos, el análisis estadístico descriptivo y la visualización de datos. A través de este análisis se buscamos:

- Evaluar la calidad y estructura de los datos.
- Detectar valores nulos, inconsistencias y outliers.
- Analizar la relación entre el precio y las principales características de los inmuebles.
- Extraer conclusiones iniciales que permitan entender el comportamiento del mercado inmobiliario.

Objetivos específicos

- Analizar la estructura y calidad del conjunto de datos.
- Identificar patrones y relaciones entre variables.
- Detectar valores atípicos y posibles sesgos.
- Extraer conclusiones relevantes para etapas posteriores.

ⓘ Hipótesis

1. ¿La planta influye en el precio por metro cuadrado de la vivienda?
2. ¿El precio de las propiedades en venta en Madrid es más alto en las zonas cercanas al centro de la ciudad que en las zonas periféricas, incluso controlando el tamaño y las características de las propiedades?
3. ¿La ausencia de ascensor reduce significativamente el precio en pisos?

4. ¿El precio de los pisos exteriores aumenta respecto a los interiores?
5. ¿La superficie de los inmuebles depende del distrito al que pertenezca?

Descripción del conjunto de datos

Origen de los datos

- **Fuente del dataset:** los datos se obtuvieron aplicando técnicas de scrapping en distintas épocas del año.
- **Periodo temporal (si aplica):** meses de abril, mayo, junio, octubre y diciembre del año 2023 y los meses de enero, marzo y abril del año 2024.
- **Número de registros:** 29.922
- **Número de variables:** 53

Estructura del dataset

Descripción de las variables incluidas en el conjunto de datos:

| COLUMNA | DESCRIPCIÓN | TIPO DE VARIABLES | IMPORTANCIA |
|--------------|------------------------------------|-------------------|-------------|
| propertyCode | Código único del inmueble | Numérica Discreta | 2 |
| numPhotos | Número de fotos del anuncio | Numérica Discreta | 2 |
| floor | Planta del inmueble | Object | 1 |
| price | Precio total | Numérica Continua | 0 |
| propertyType | Tipo de propiedad | Categórica | 1 |
| operation | Tipo de operación (venta/alquiler) | Categórica | 1 |

| COLUMNA | DESCRIPCIÓN | TIPO DE VARIABLES | IMPORTANCIA |
|------------------------|--------------------------------|-------------------|-------------|
| size | Superficie en mts ² | Numérica Continua | 1 |
| exterior | Si es exterior o interior | Binaria | 2 |
| rooms | Número de habitaciones | Numérica Discreta | 1 |
| bathrooms | Número de baños | Categórica | 1 |
| address | Dirección completa | String | 1 |
| province | Provincia | Categórica | 1 |
| municipality | Municipio | Categórica | 1 |
| district | Distrito | Numérica Discreta | 1 |
| country | País | Categórica | 1 |
| neighborhood | Barrio | Numérica Discreta | 1 |
| latitude | Coordinada latitud | Numérica Discreta | 3 |
| longitude | Coordinada longitud | Numérica Discreta | 3 |
| description | Descripción del anuncio | String | - |
| hasVideo | Si tiene video | Binaria | 3 |
| status | Estado del inmueble | Categórica | 2 |
| newDevelopment | Si es obra nueva | Binaria | 1 |
| hasLift | Si tiene ascensor | Binaria | 1 |
| priceByArea | Precio por mts ² | Numérica Continua | 0 |
| hasPlan | Si tiene plano | Binaria | 1 |
| has3DTour | Si tiene tour 3D | Binaria | 3 |
| has360 | Si tiene vista 360 | Binaria | 3 |
| hasStaging | Si tiene staging virtual | Binaria | 3 |
| topNewDevelopment | Destacado como obra nueva | Categórica | 1 |
| superTopHighlight | Súper destacado | Categórica | 3 |
| newDevelopmentFinished | Obra nueva finalizada | Binaria | 2 |

| COLUMNA | DESCRIPCIÓN | TIPO DE VARIABLES | IMPORTANCIA |
|-------------------------------|--|-------------------|-------------|
| newDevelopment.1 | Duplicado o variable relacionada | Binaria | 2 |
| SubdType | Subtipo del inmueble | Categórica | 2 |
| ex | Variable auxiliar | Numérica Continua | - |
| topPlus | Destacado premium | Binaria | 3 |
| groupDescription | Descripción del grupo o categoría del inmueble | Categórica | - |
| hasParkingSpace | Indica si tiene plaza de parking | Binaria | 2 |
| isParkingSpaceIncludedInPrice | Si el parking está incluido en el precio | Binaria | 3 |
| parkingSpacePrice | Precio del parking si no está incluido | Numérica Continua | - |
| typology | Tipología general del inmueble | Categórica | - |
| subTypology | Subtipología específica | Categórica | - |
| subtitle | Nombre del barrio en Madrid | Categórica | - |
| title | Tipo del piso y su respectiva calle | Categórica | - |
| name_0 | Etiqueta comercial del anuncio | Categórica | - |
| text_0 | Etiqueta comercial del anuncio | Categórica | - |
| name_1 | Tipo de propiedad y lujo | Categórica | - |
| text_1 | Tipo de propiedad y lujo | Categórica | - |

Metodología y herramientas

Metodología aplicada

El desarrollo del EDA siguió las siguientes etapas:

1. Inspección inicial del dataset y unificación del los dataset.
2. Limpieza y preprocesamiento de los datos.
3. Análisis univariado
4. Análisis bivariado y multivariado
5. Test de correlaciones
6. Síntesis de resultados y conclusiones

Herramientas utilizadas

- **Lenguaje:** Python
- **Librerías:**
 - matplotlib
 - numpy
 - pandas
 - scipy
 - seaborn
- **Entorno:**
 - Jupyter Notebook

4. Análisis Exploratorio de Datos

4.1 Análisis de calidad de los datos

- **Carga de datos:**
 - Como son datos que se recuperaron en distintos meses y diferentes años. se realizar las importaciones de los mismos, revisando previamente los datos para luego hacer una unificación final. Esto

generó muchos registros duplicados ya que en cada mes, habían varios archivos y al no conocer exactamente la naturaleza, se realiza la unificación general para luego analizarlos y limpiarlos.

- **Valores nulos y tratamiento aplicado:**
 - Se realizó la imputación de la moda en los registros categóricos.
 - Se aplicó la mediana en variables numéricas continuas.
- **Registros duplicados:**
 - Se eliminaron los registros duplicados dónde todas las columnas coincidían al 100%.
 - Existe una columna denominada `propertyCode` la cual representa el link del anuncio. No sé eliminaron los duplicados filtrando por dicho atributo ya que el mismo anuncio podía estar publicado con diferentes precios. Por falta de tiempo no se realizó un correcto tratamiento del mismo.
- Tipos de datos incorrectos
 - Se normalizaron columnas dónde en su interior se encontraban datos en formato `.json`. Se crearon nuevas columnas con las `key` de las mismas.

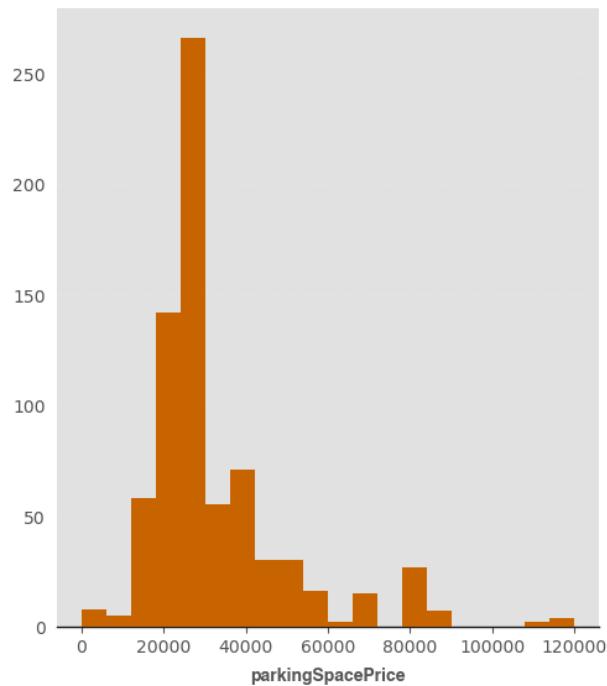
4.2 Análisis univariado

Análisis individual de las variables: Se dejan constancias de algunas gráficas relevantes del análisis de las columnas. No sé ponen todas las imágenes ya que el dataset posee muchas columnas.

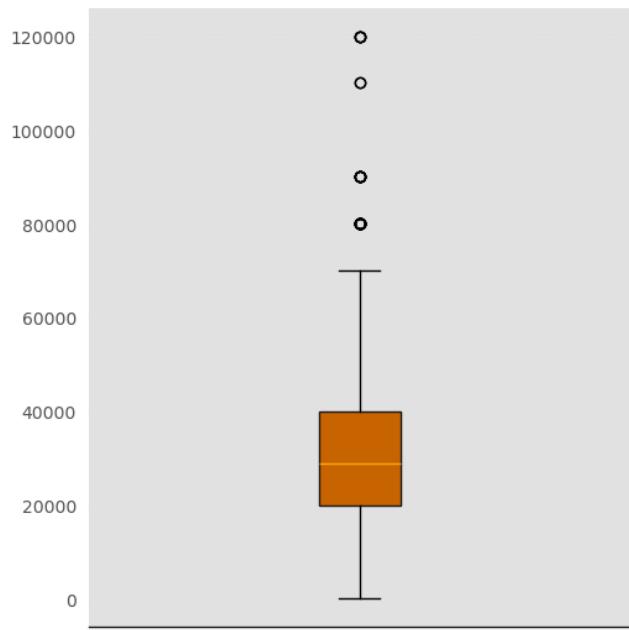
Distribución de variables numéricas

Análisis de parkingSpacePrice

HISTOGRAMA

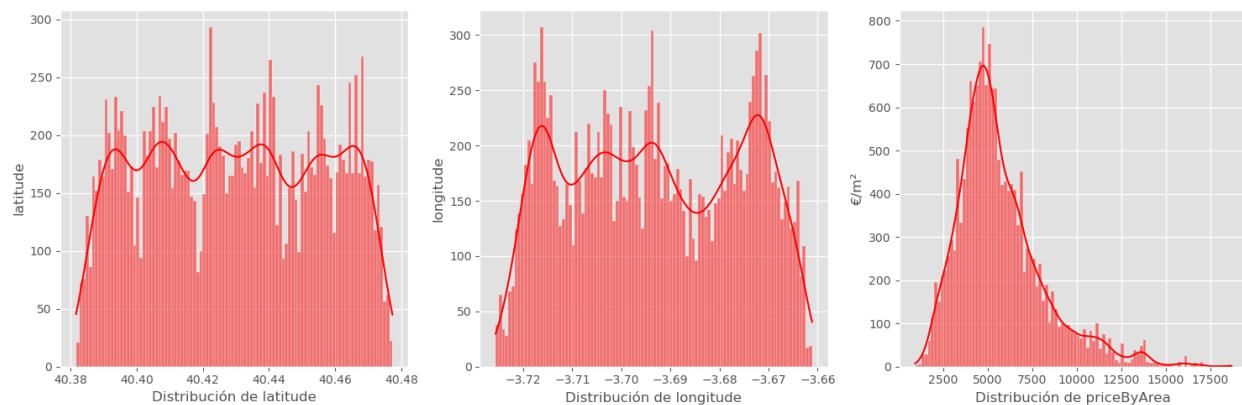
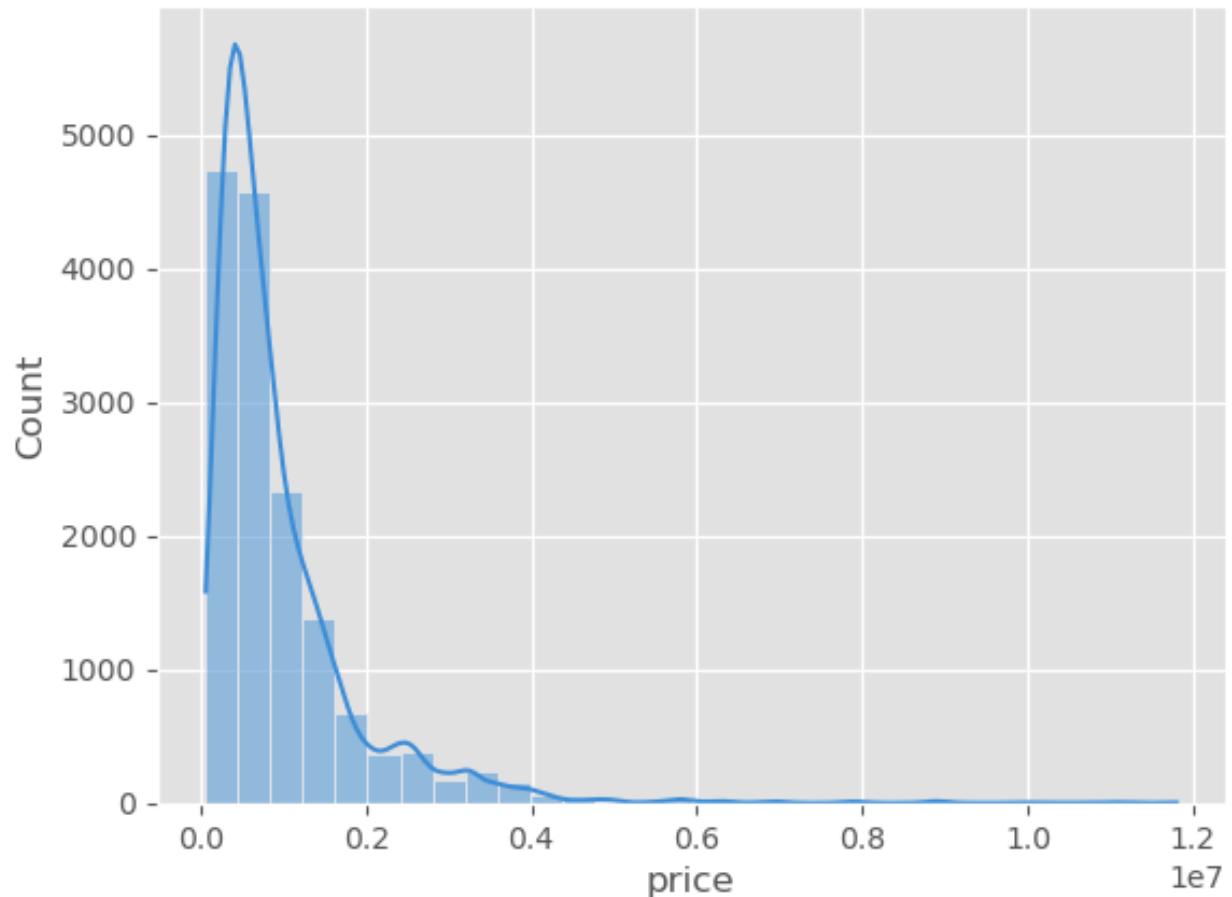


BOXPLOT

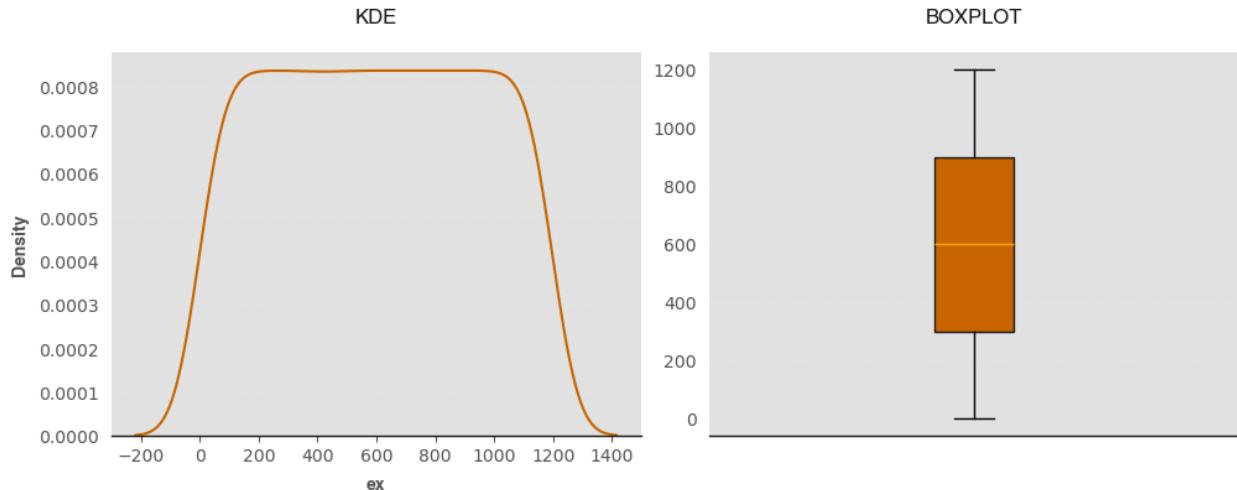


Medidas estadísticas descriptivas

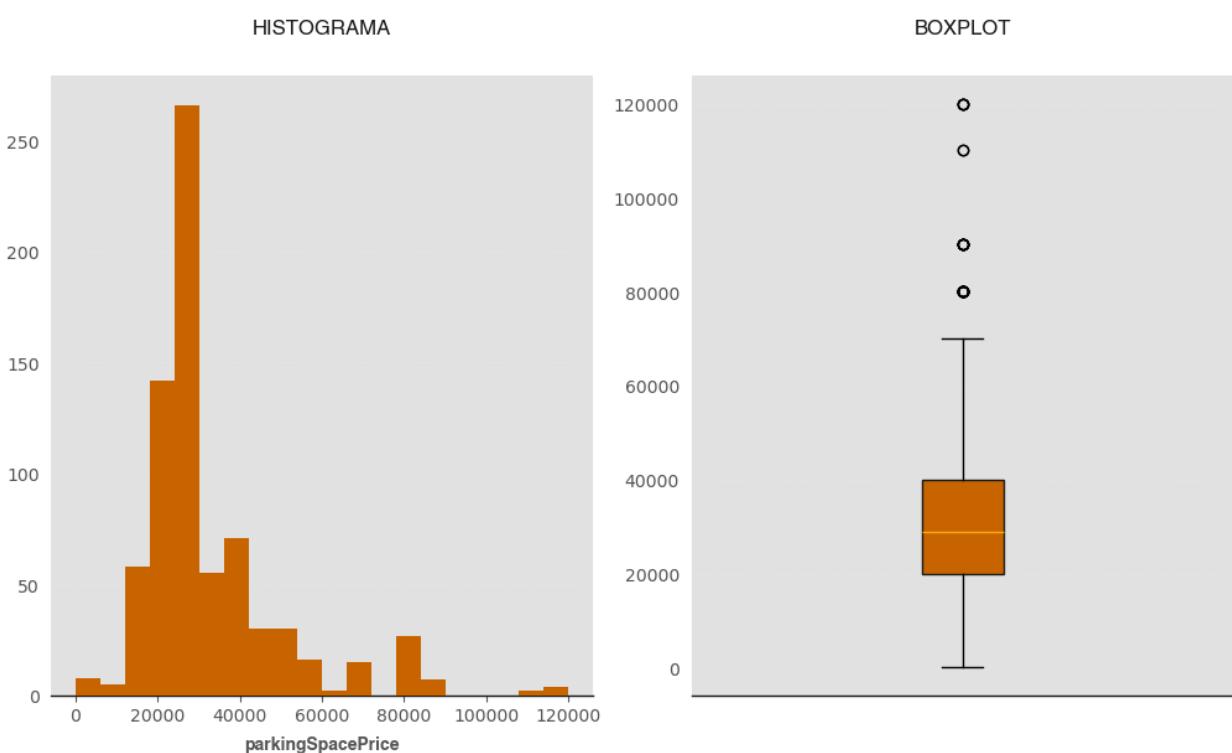
Distribución de PRICE



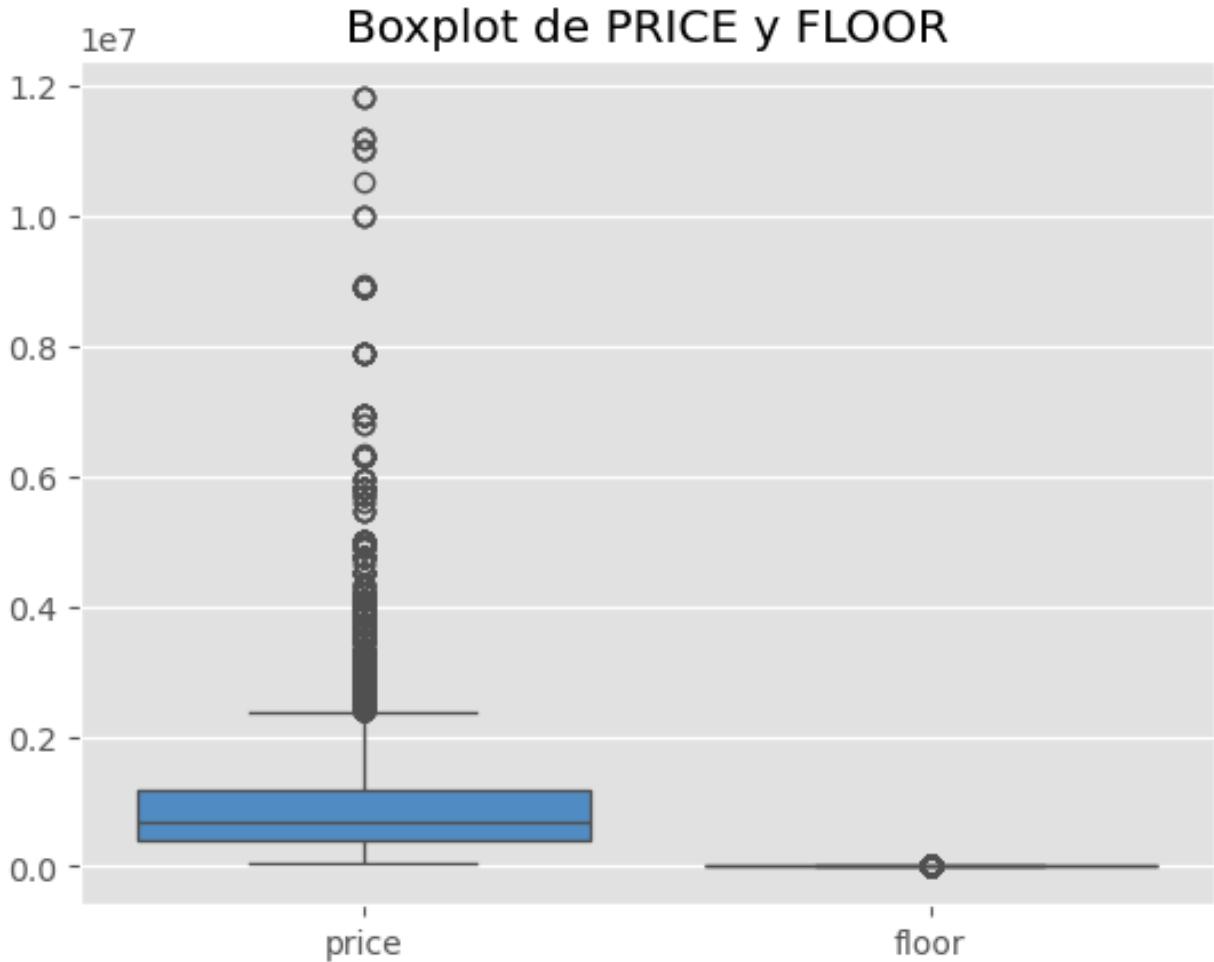
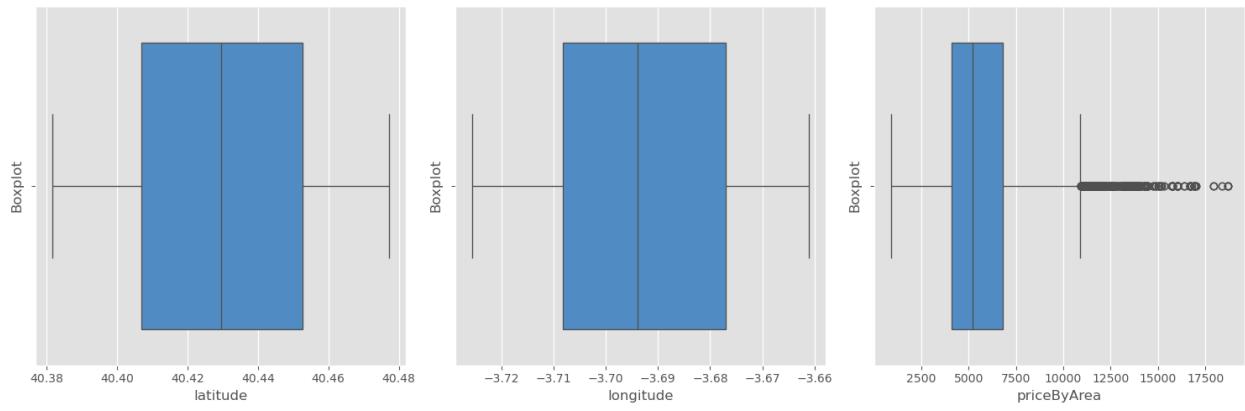
Análisis de ex



Análisis de parkingSpacePrice



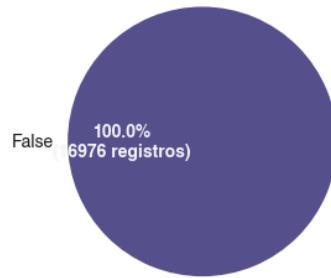
Análisis de outliers



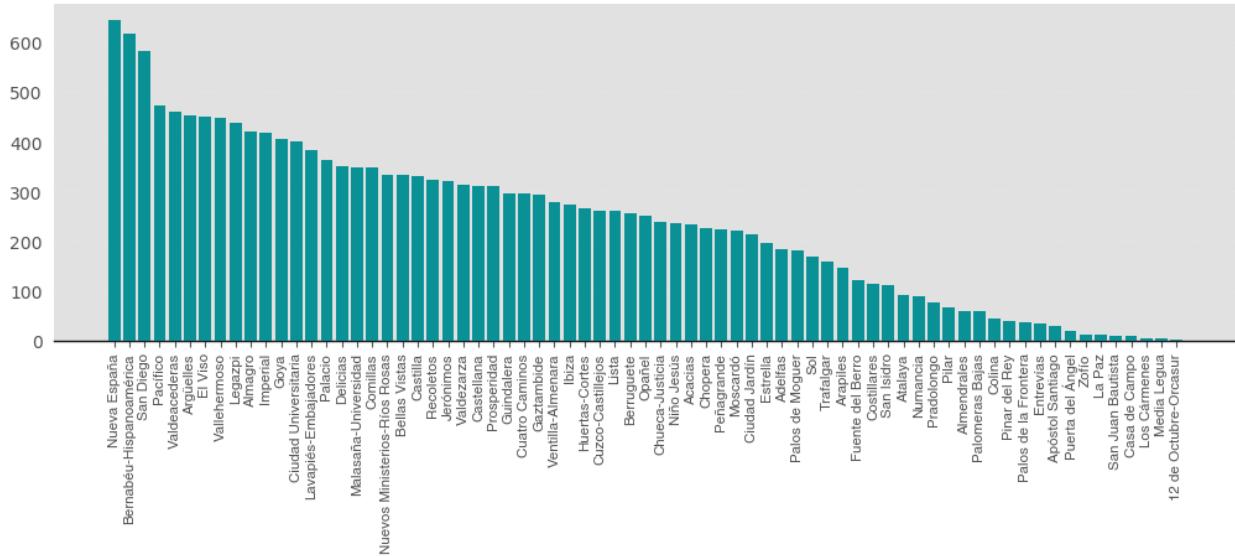
Frecuencia de variables categóricas

Análisis de subtitle

SUBTITLE: VALORES NULOS



SUBTITLE: VALORES DISTINTOS

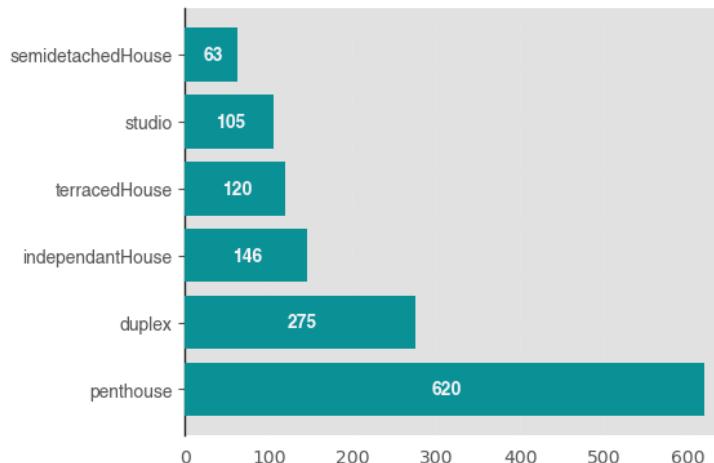


Análisis de SubdType

VALORES NULOS



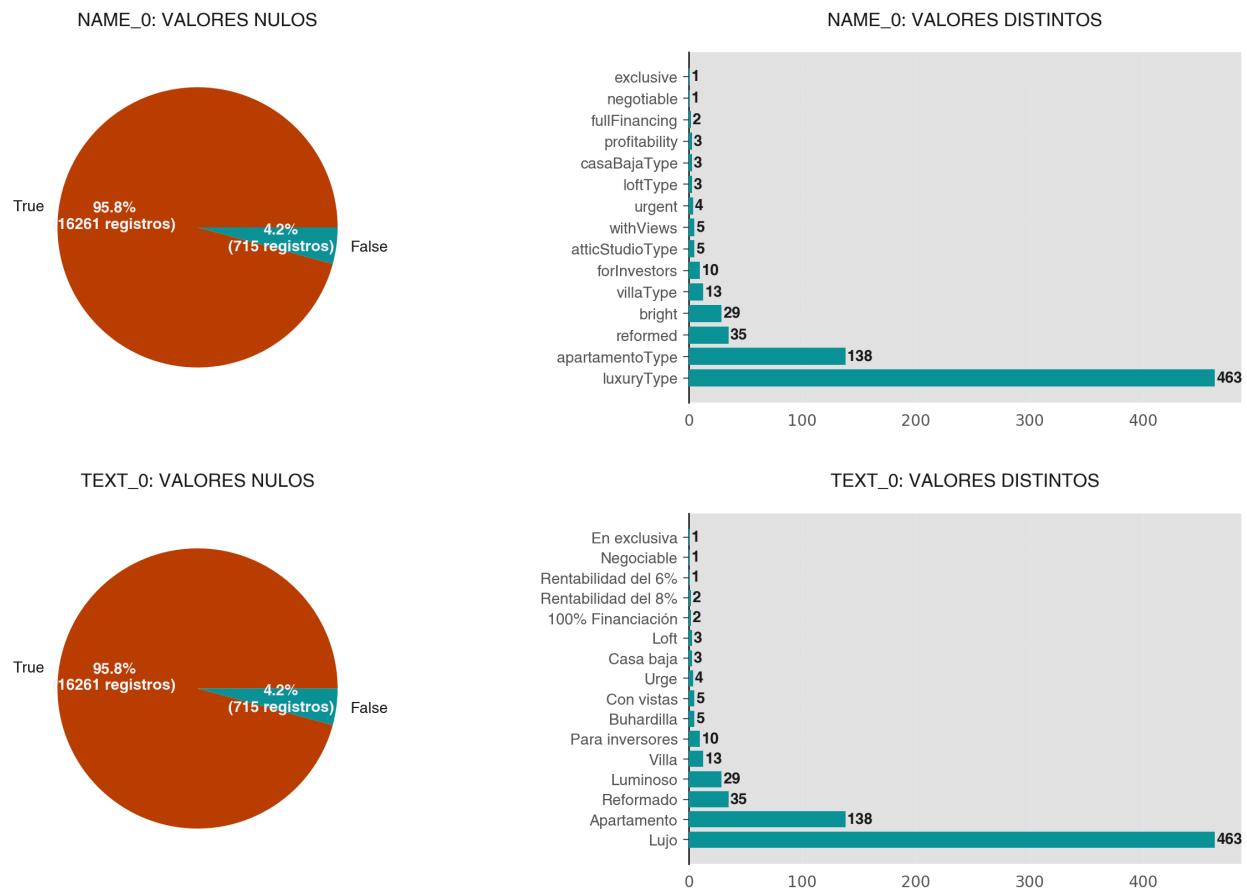
VALORES DISTINTOS



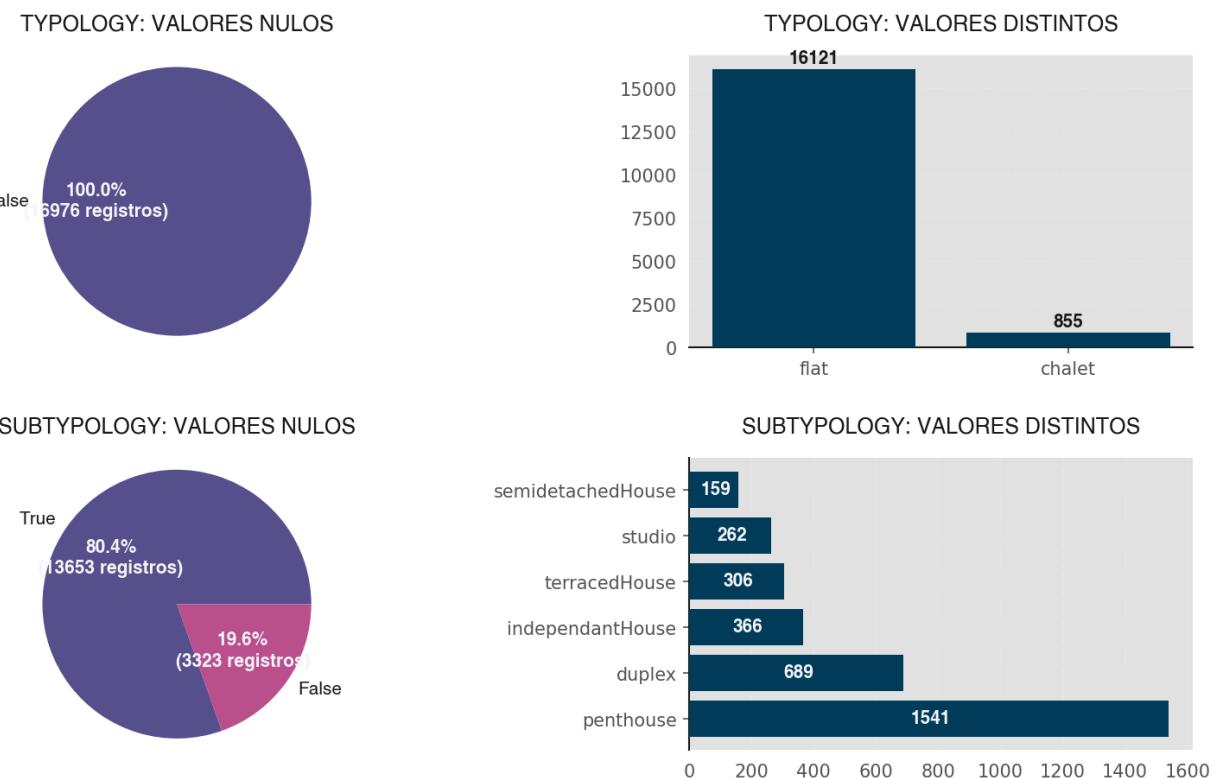
12 de Octubre-Orcasur

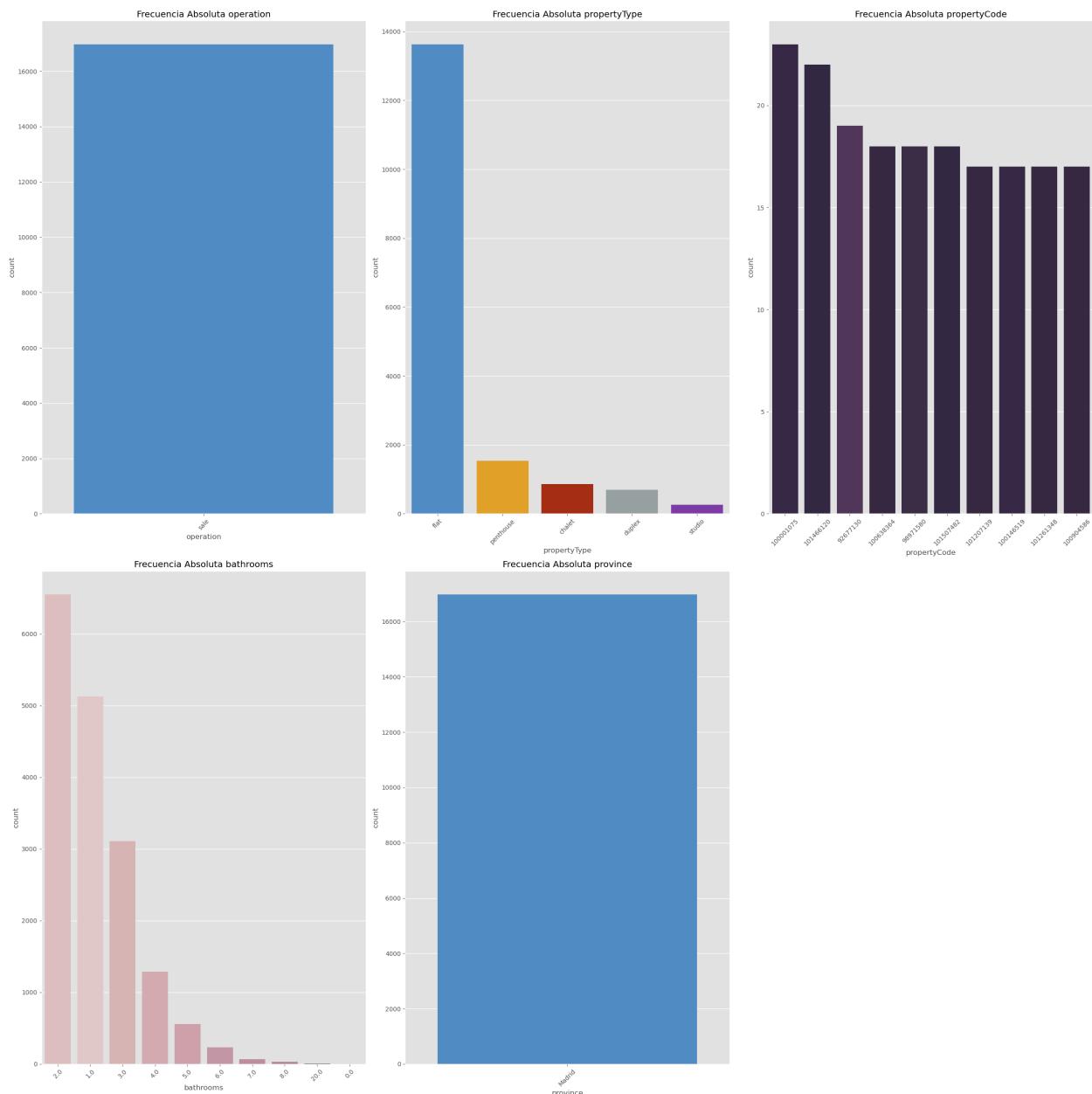
12

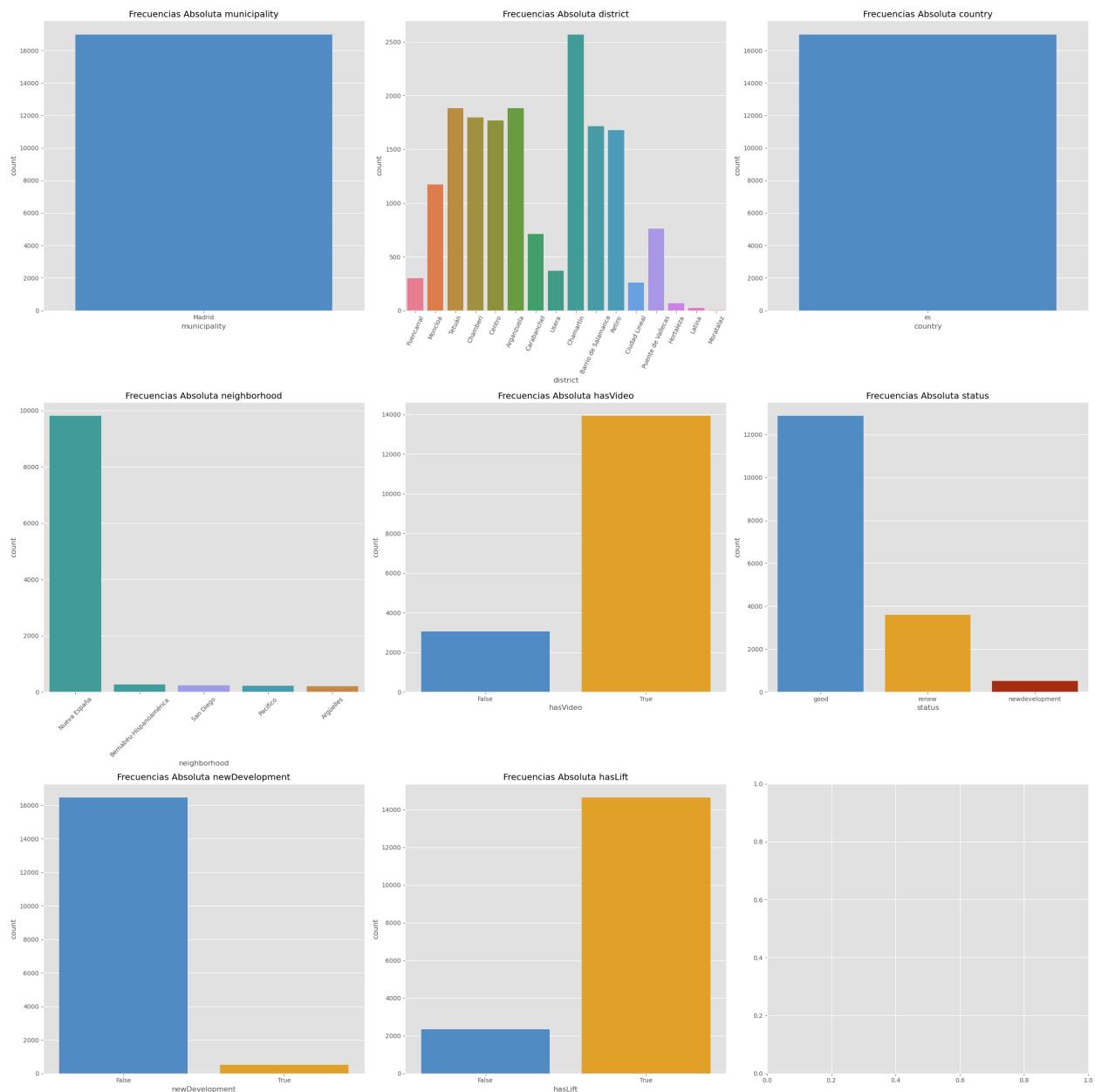
Análisis de name_0 y text_0



Análisis de typology y subTypology

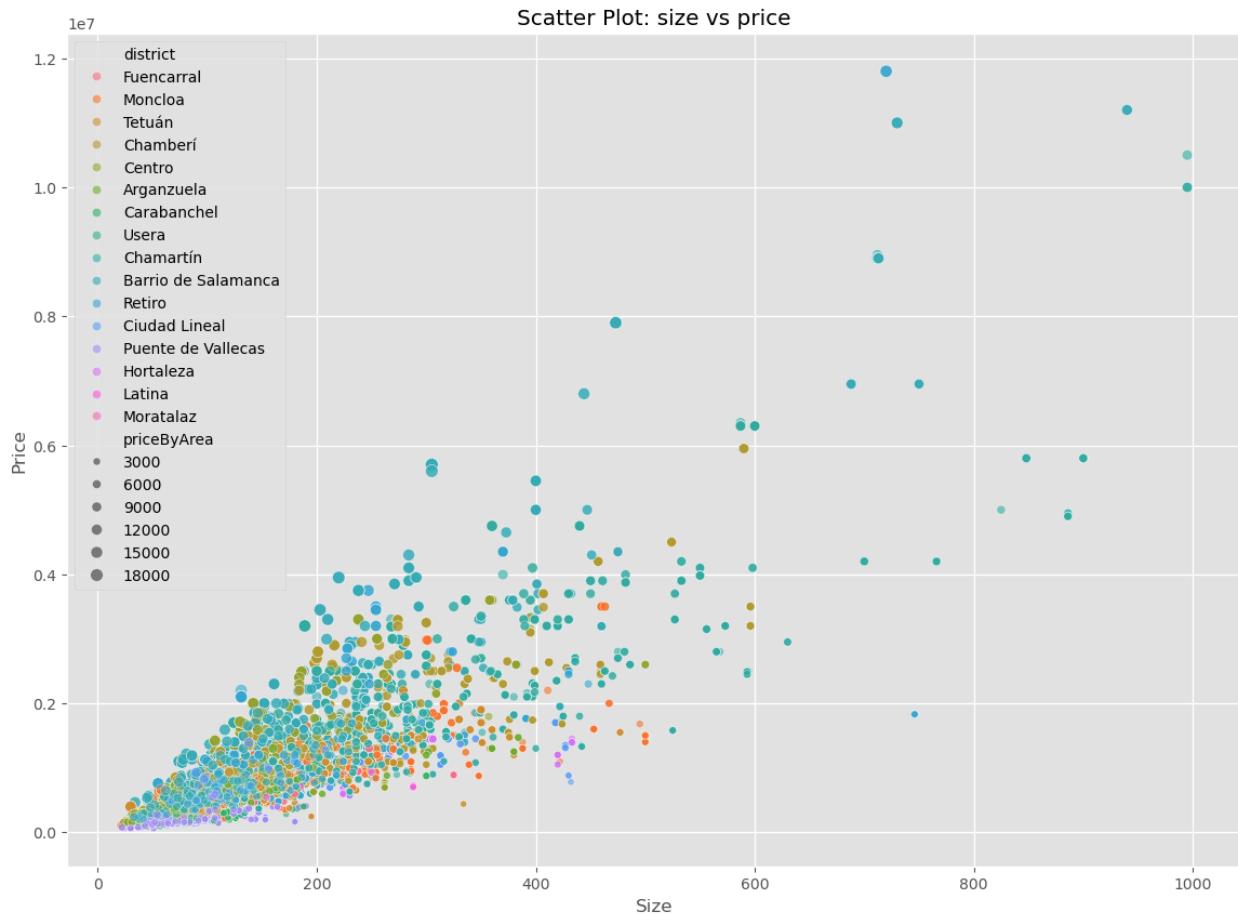
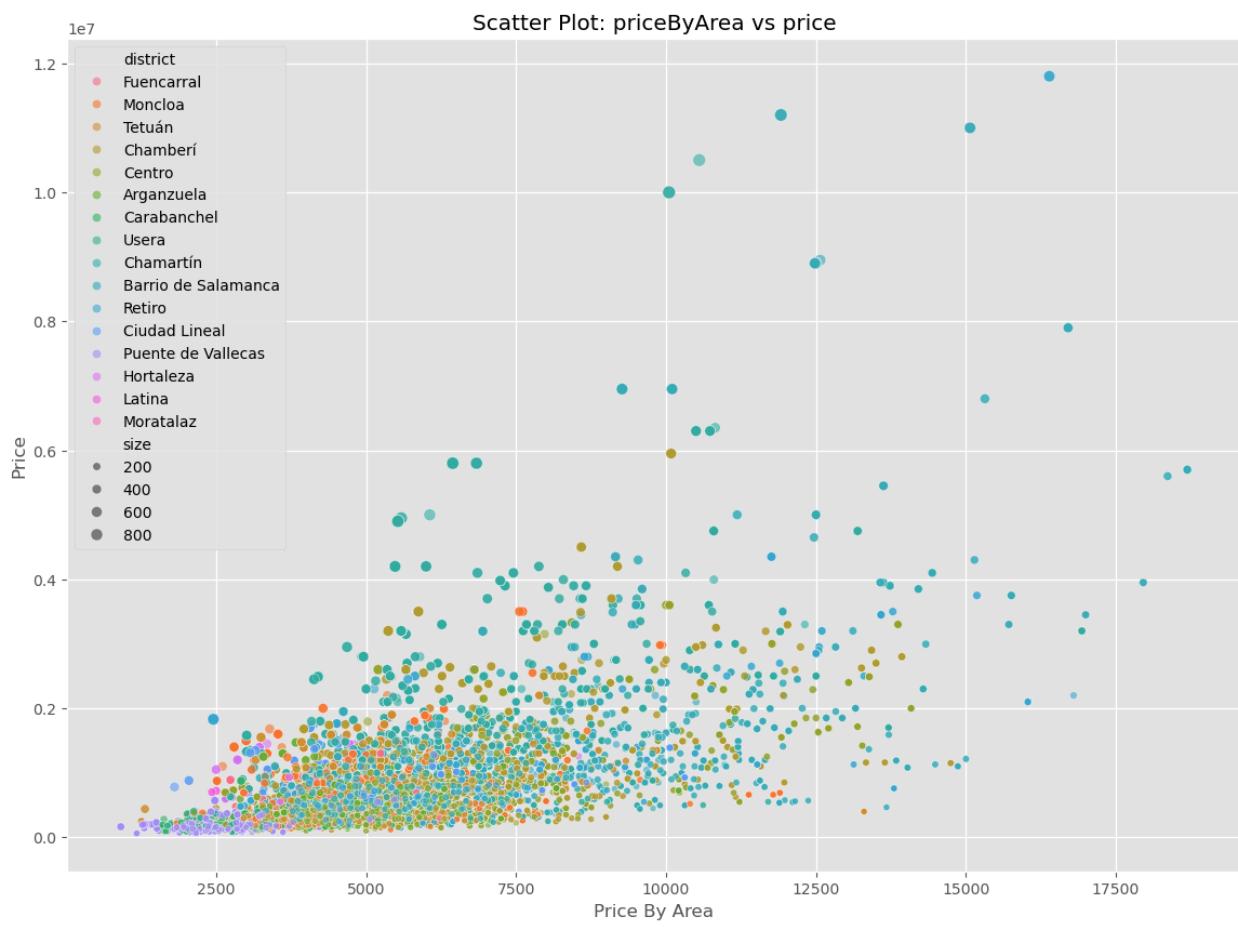






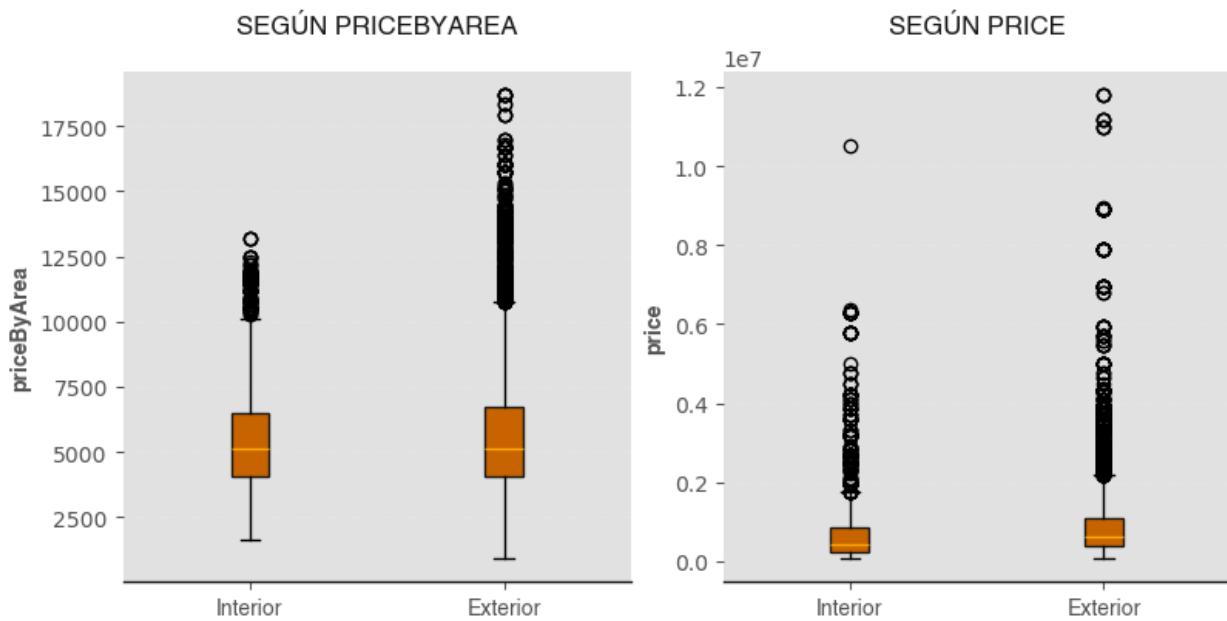
4.3 Análisis bivariado y multivariado

Relación entre variables numéricas

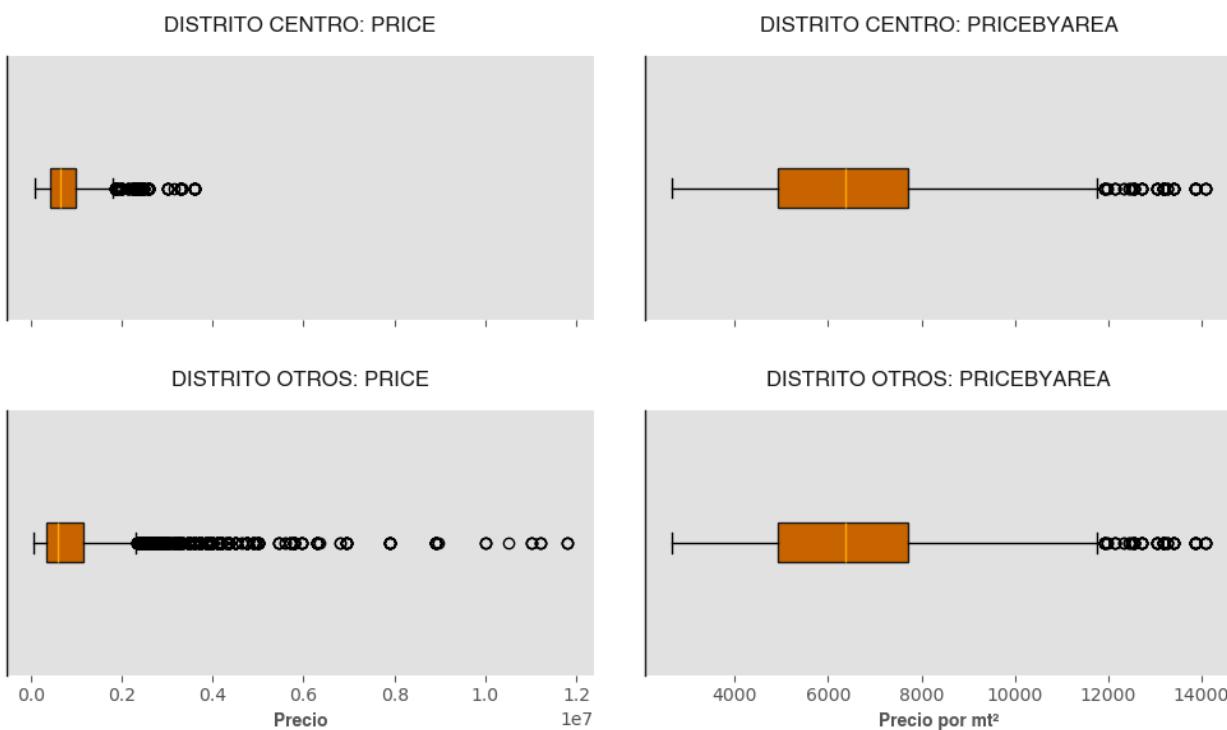


Comparación entre variables categóricas y numéricas

Análisis de exterior e interior

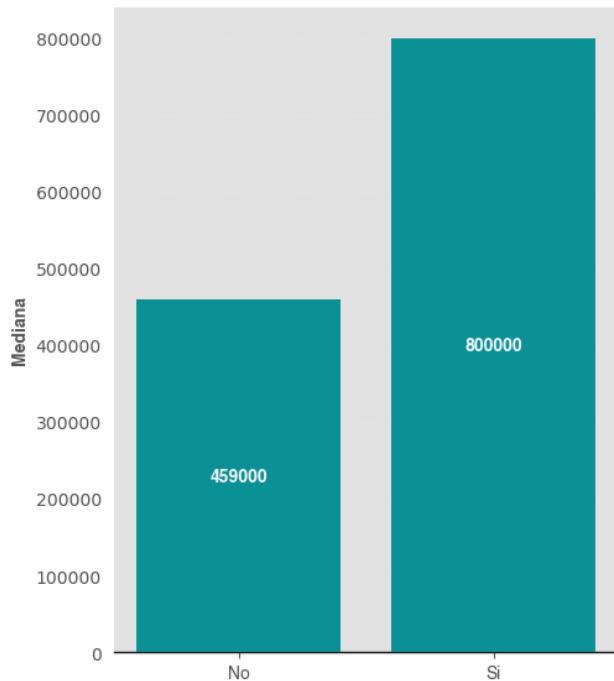


Análisis comparativo: Centro vs Otros

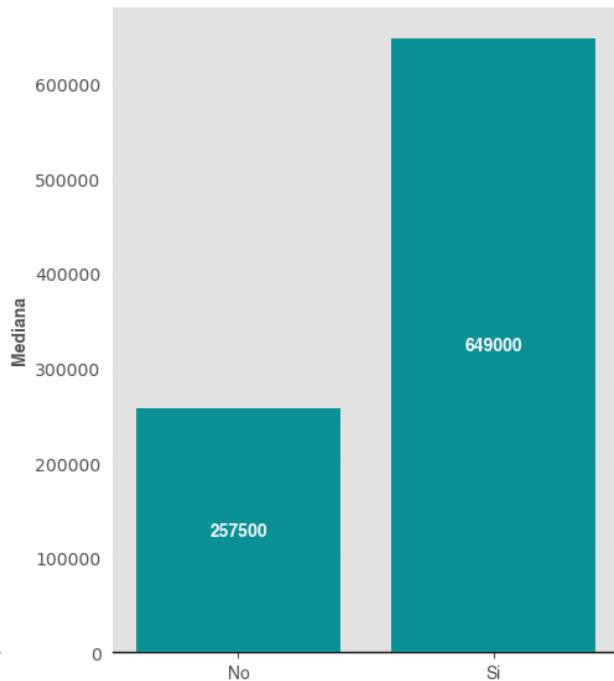


Análisis de hasLift en priceByArea

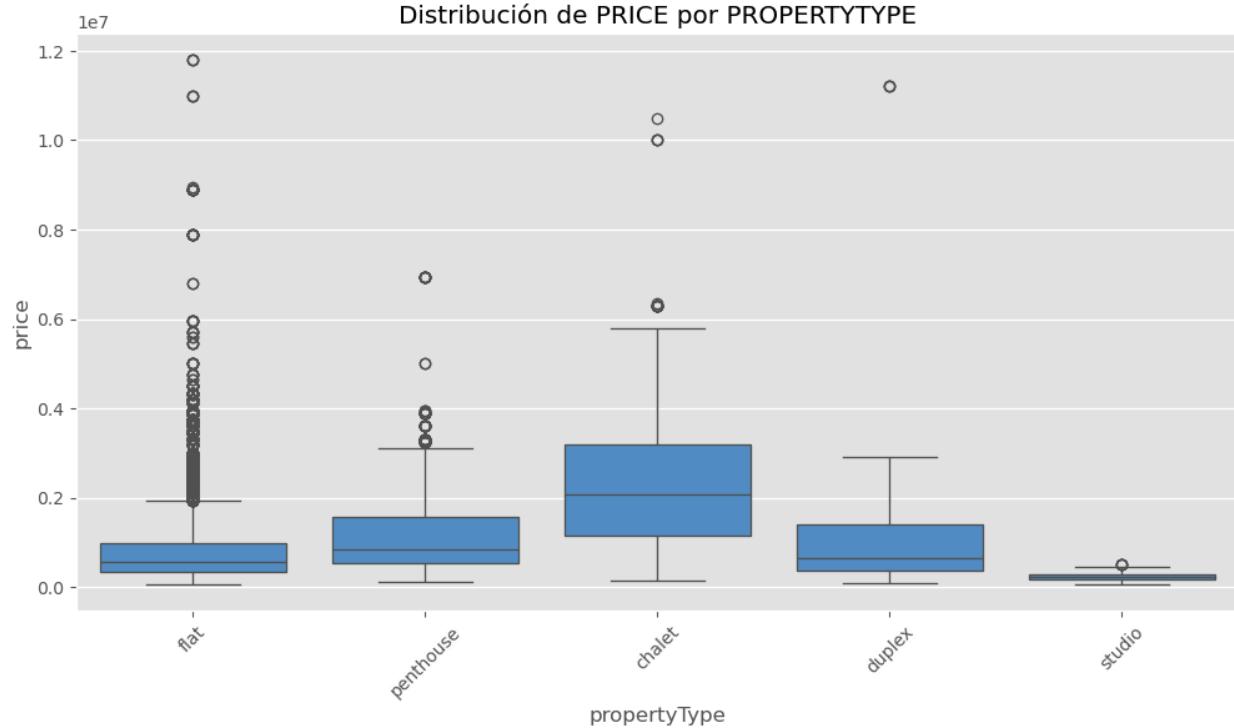
PISOS IGUALES O SUPERIORES AL QUINTO

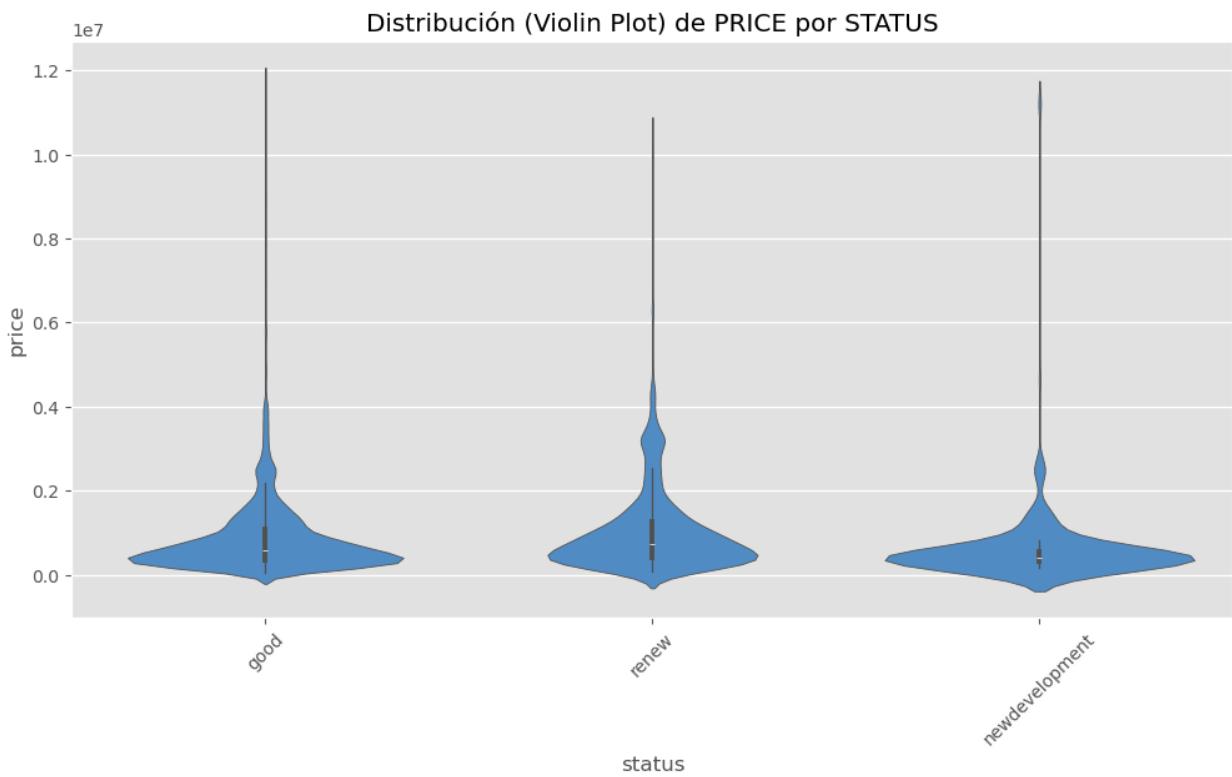
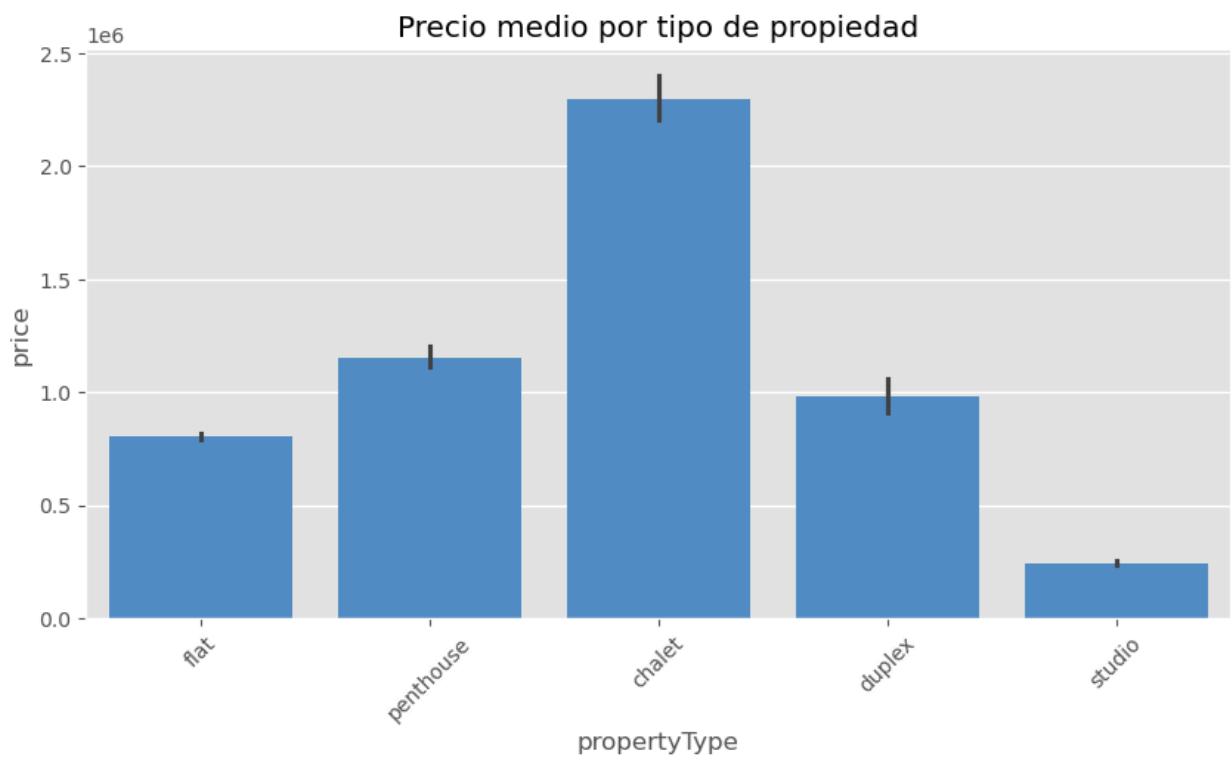


PISOS INFERIORES AL QUINTO

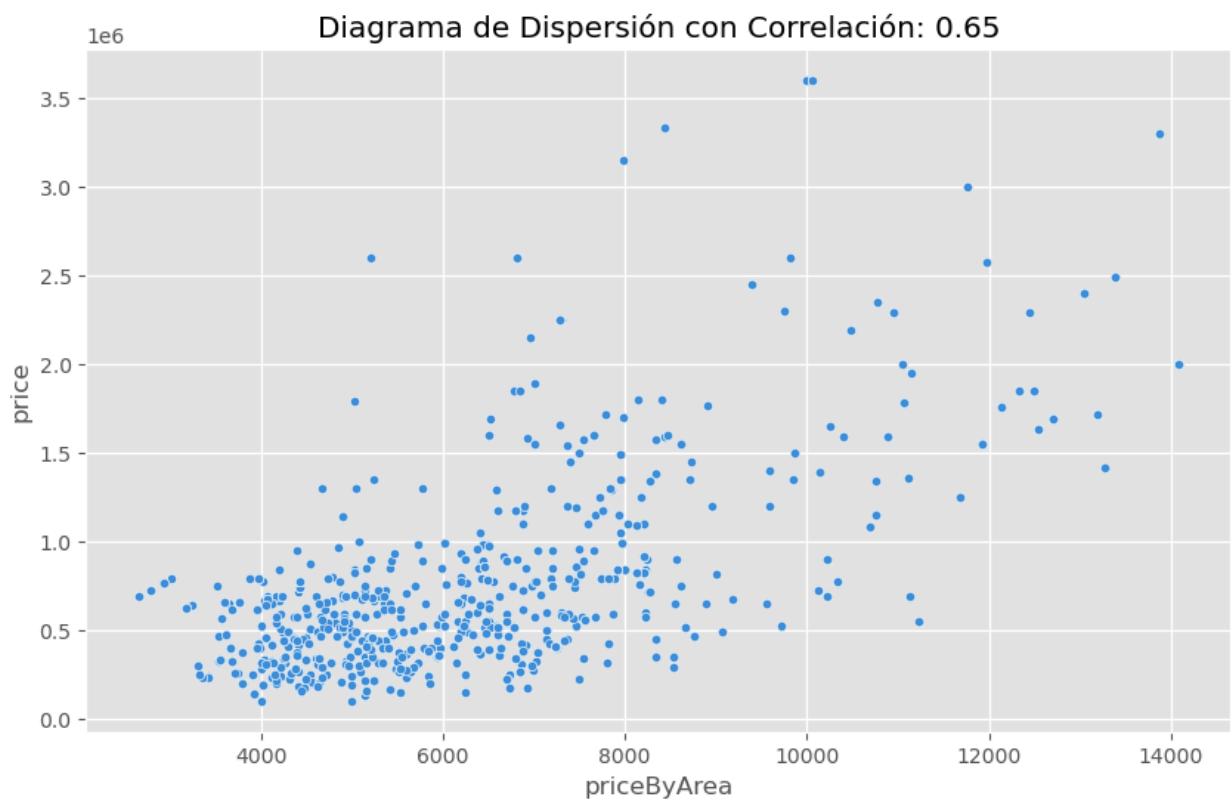
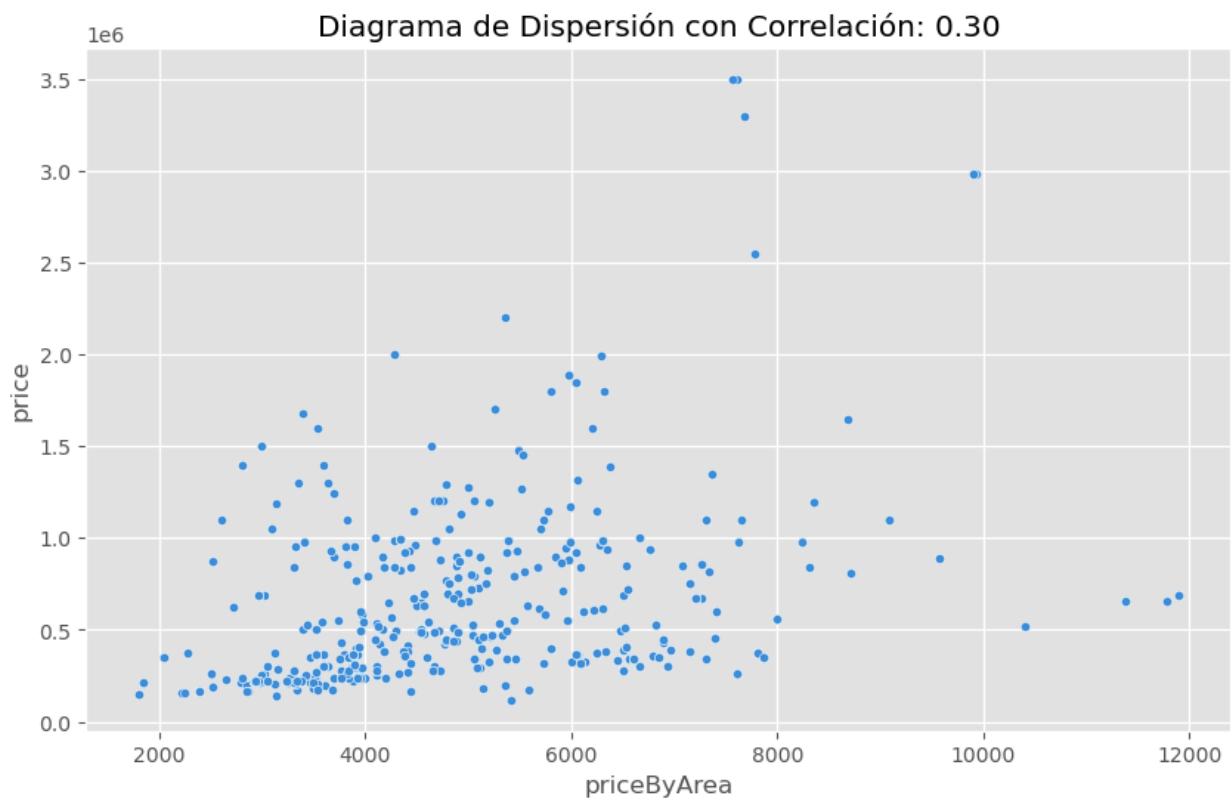


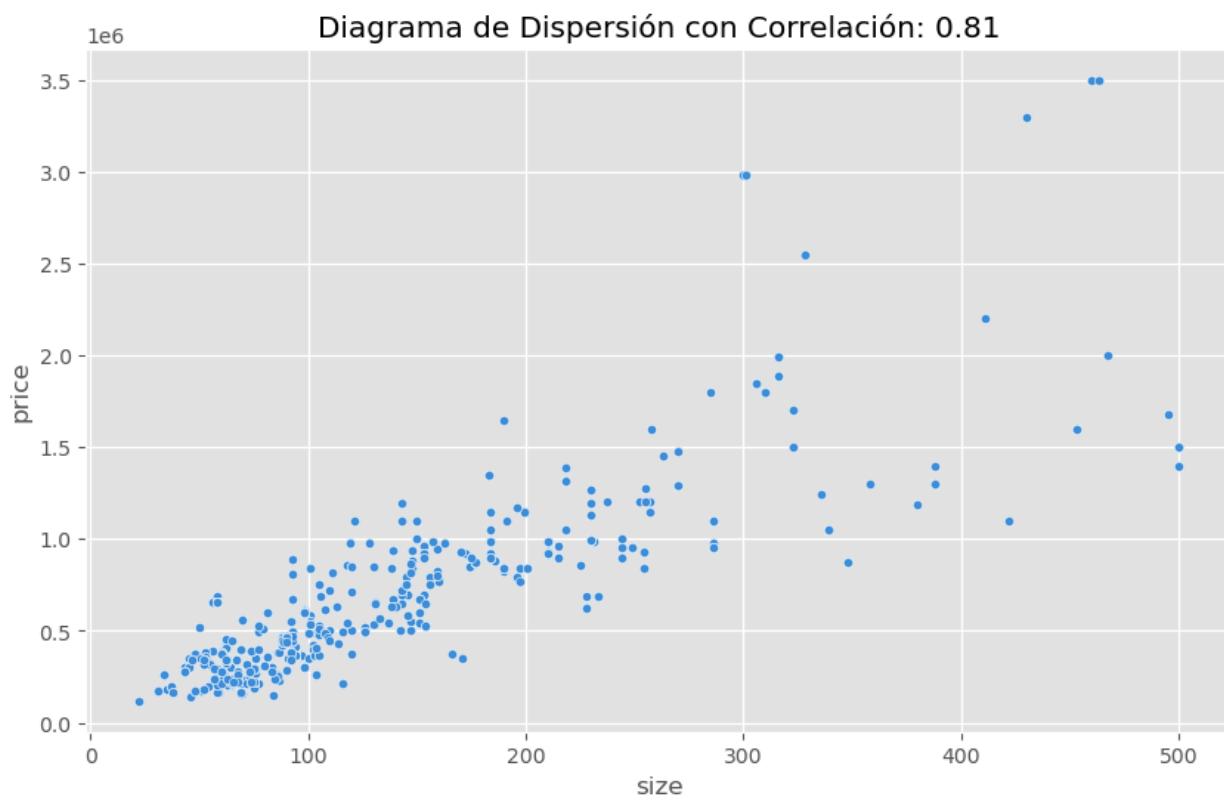
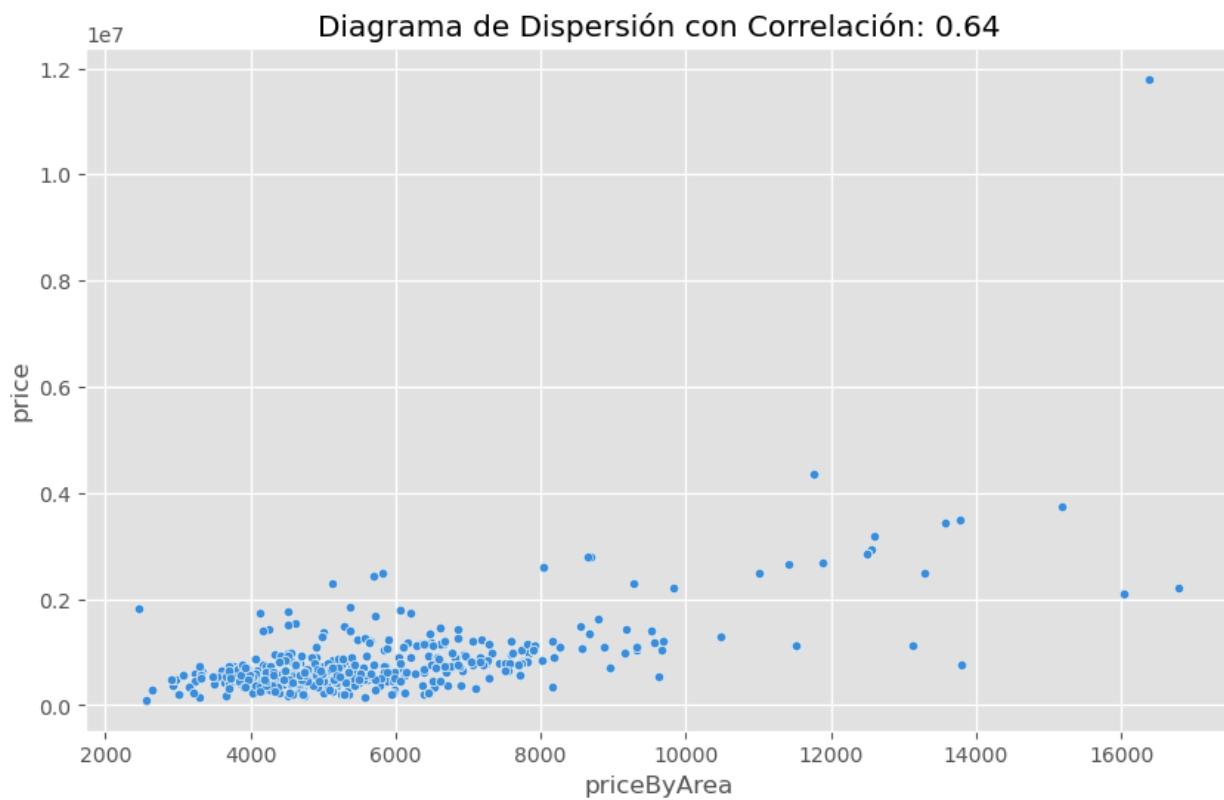
Distribución de PRICE por PROPERTYTYPE





Análisis de correlaciones





Este análisis permitió identificar patrones y relaciones significativas.

5. Resultados y hallazgos principales

Resumen de los insights más relevantes obtenidos durante el análisis:

- **Variables con mayor influencia:** hemos visto que la influencia de la ubicación prima por sobre cualquiera de las otras. La superficie y los commodities del mismo también agregan valor a la propiedad.

7. Conclusiones y trabajo futuro

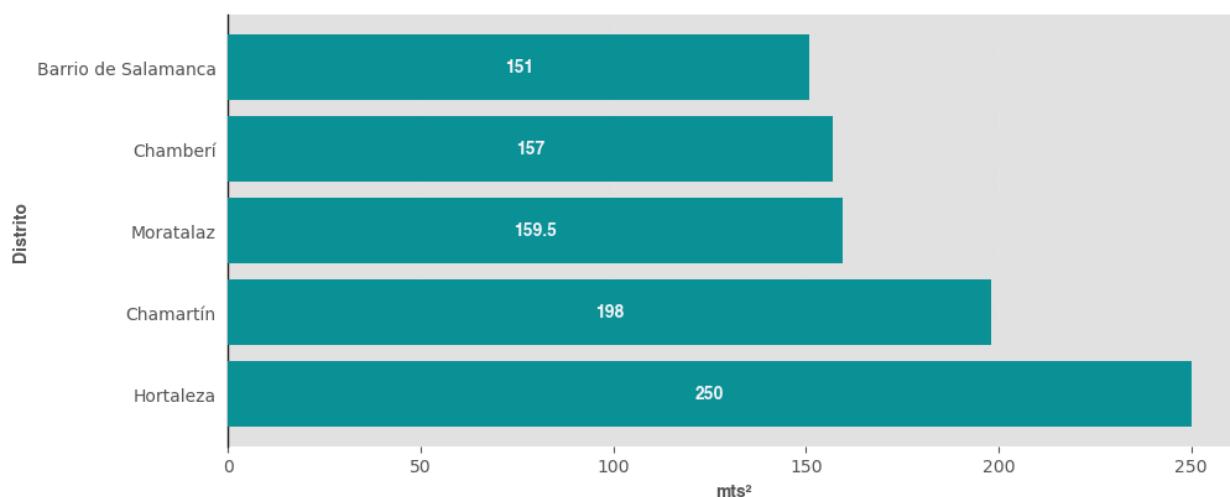
7.1 Conclusiones

Síntesis final de los resultados obtenidos y cumplimiento de los objetivos planteados:

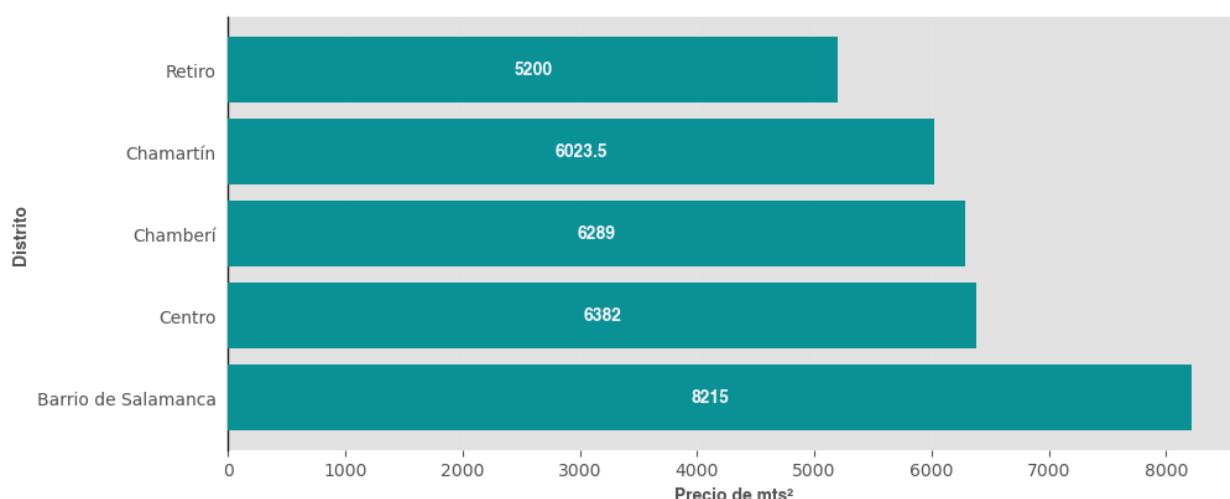
- A partir del quinto piso la mediana del precio por metro cuadrado es un 24% más caro, excluyendo sótano, subsuelo, entrepiso y bajo.
- El precio por metro cuadrado es un 25% más caro en la zona céntrica de Madrid.
- El precio de los pisos que cuentan con ascensor y están arriba del quinto piso son 70% más caros que los que no.
- Los pisos exteriores son 50% más caros que los que interiores.

8. Anexos (opcional)

TOP5: Distrito según mts²



TOP5: Distrito según precio de mts²



TOP5: Distrito según precio de vivienda

