ANALISIS DEL MERCADO INMOBILIARIO DE MELBOURNE

Juan Martín Rival

ENTREGA FINAL

CODERHOUSE

CODERHOUSE

EL CASO: MERCADO INMOBILIARIO DE MELBOURNE

OBJETIVO

Analizaremos el comportamiento de la oferta inmobiliaria en la Ciudad de Melbourne, Australia.
Se buscará describir las características de la oferta, identificar las principales variables que influyen en el precio, y generar un modelo predictivo que permita predecir el precio de publicación de un área urbana determinada.
De forma adicional, se busca poner a prueba la tesis de Topalov (1987), sobre la incidencia de la distancia al centro urbano como variable explicativa del valor del precio del suelo urbano.

Este trabajo esta orientado a contribuir a la comprensión del comportamiento del mercado inmobiliario, y espera poder aportar herramientas a especialistas en urbanismo, responsables de políticas habitacionales/inmobiliarias y tomadores de decisiones en procesos de desarrollo inmobiliario y/o grandes actores en el mercado inmobiliario.

CODERHOUSE

HIPÓTESIS DE TRABAJO

PRINCIPALES HIPOTESIS

La hipótesis central es que el precio de la propiedad esta determinado principalmente por una interacción entre el tamaño de las propiedades (Cantidad de metros cuadrados construidos y/o totales del lote) y la distancia promedio del centro de la ciudad.

HIPOTESIS SECUNDARIAS

Indagar el efecto que tienen sobre el precio:
La cantidad y tipo de ambientes en una vivienda (Cuartos, baños y cocheras).
La antigüedad de la vivienda.
Evaluar si el tipo de propiedad tiene algún efecto significativo.
Y tratar de determinar si hay otras variables o interacciones entre variables que tienen algún tipo de efecto sobre el precio.

CODERHOUSE

DATASET

Dataset de **publicaciones inmobiliarias** en la ciudad de **Melbourne, Australia, publicadas entre** 2016 y 2017, relevado por Tony Pinos, **scrapeadas** de la plataforma *Domain.com.au.* Consultado en Kaggle

Descripción del dataset

- **13580 casos**
- **21 Características/ Features**
- Característica a predecir: ***Price*** → Precio
- Principales variables independientes:
  - *Landsize* → Tamaño del lote
  - *BuildingArea* → Superficie Construida
  - *Rooms* → Cantidad de cuartos
  - *Bathroom* → Cantidad de baños
  - *Type* → Tipo de vivienda
  - *Suburb* → Barrio/Suburbio
  - *Lattitude* → Latitud
  - *Longitude* → Longitud
  - *YearBuilt* → Año de construcción
  - *Distance* → Distancia al centro comercial

CODERHOUSE

OTROS INSUMOS

AIRBNB MELBOURNE

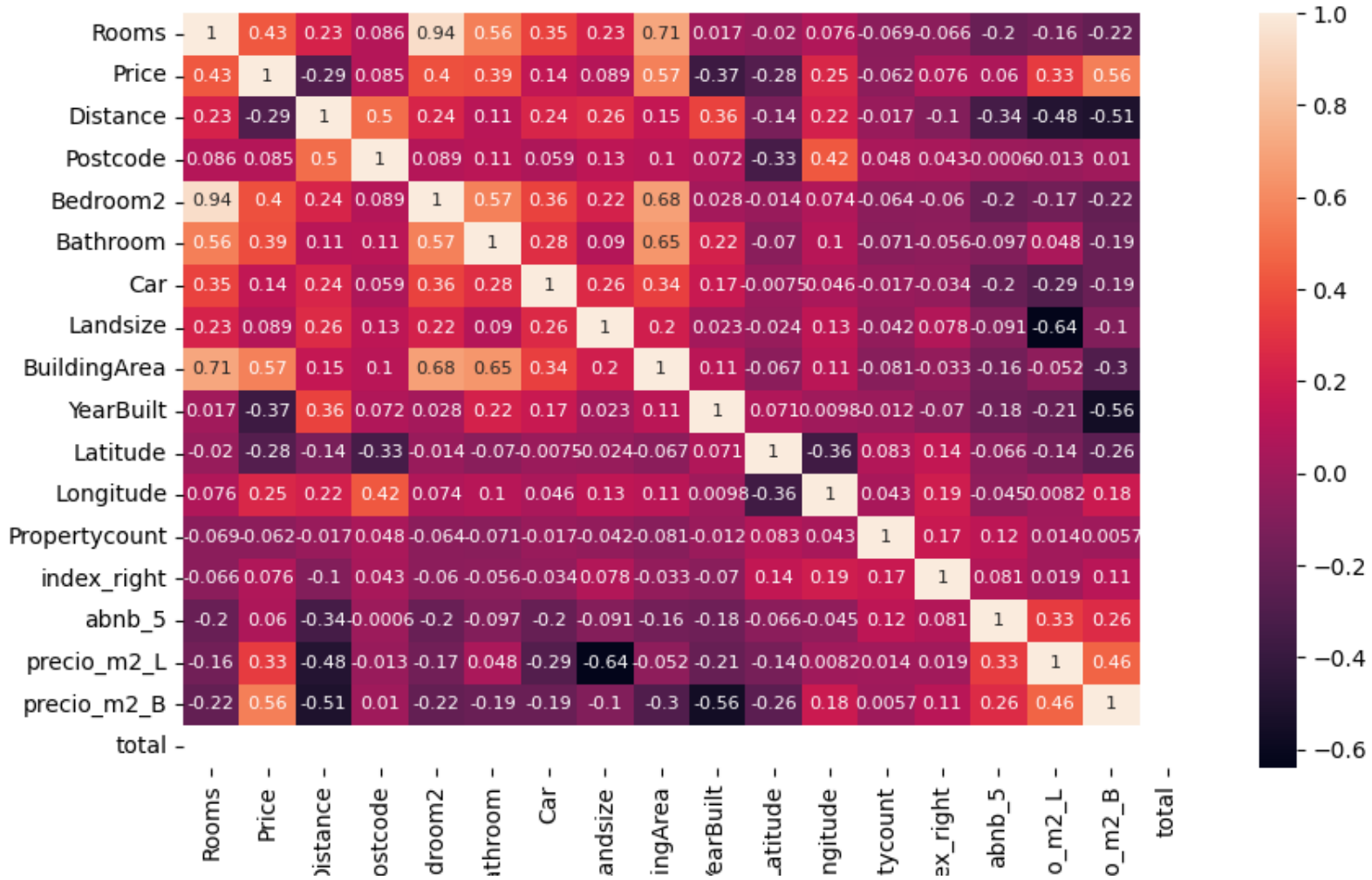
Dataset de publicaciones de **alquileres temporales en airbnb** para la ciudad de **Melbourne**, publicado en http://insideairbnb.com/melbourne/ (diciembre 2023).

- **25087 Publicaciones**
- **Fecha de publicación: 29/12/2023**
- **Geojson con límites de barrios de Melbourne**

Utilizado para enriquecer el dataset de análisis, calculando la concentración de ofertas de airbnb alrededor de la vivienda publicada (Bufffers de 500 mts).
Ver Notebook anexo: Airbnb Melbourne.ipynb

CODERHOUSE

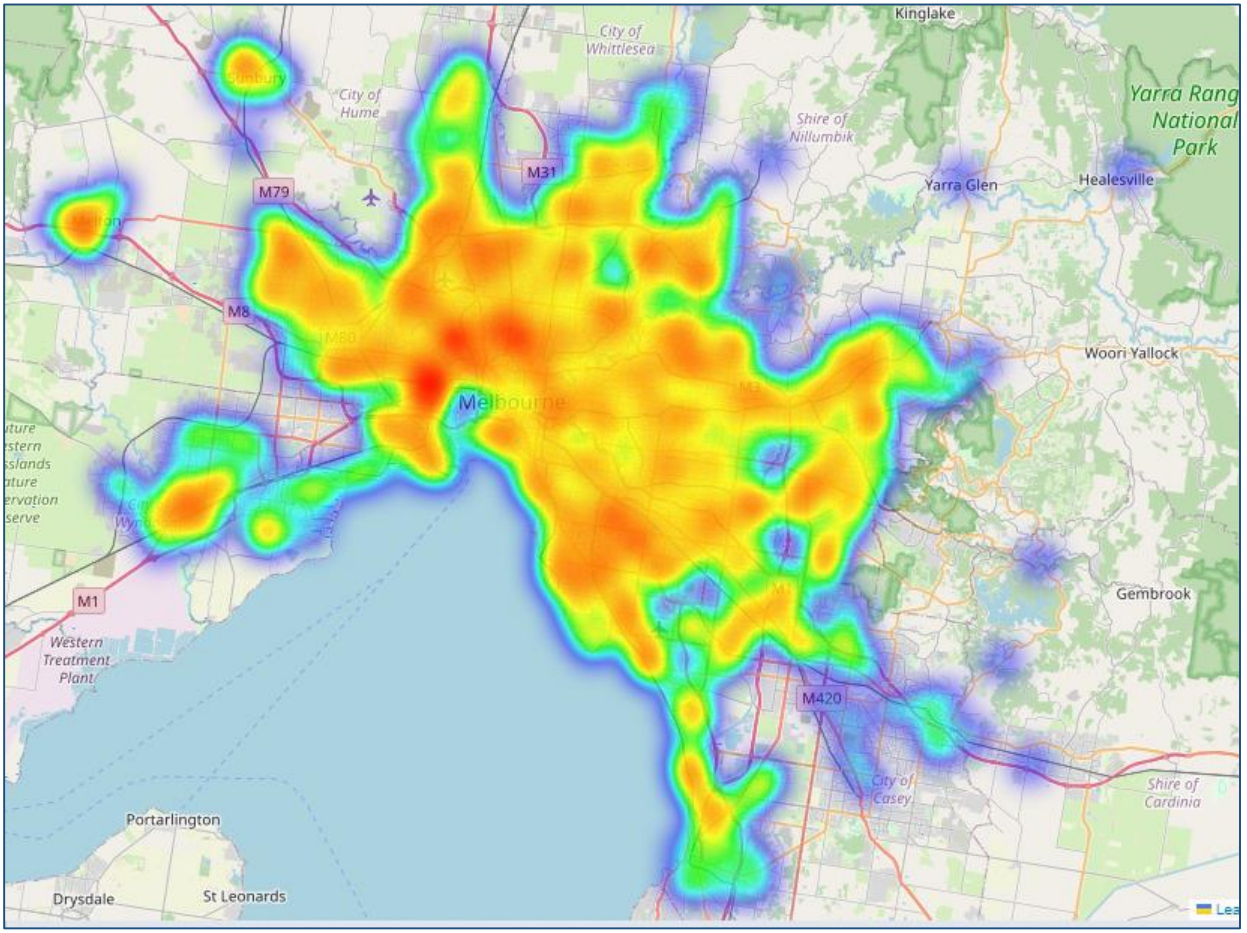
INTERACCION ENTRE ATRIBUTOS



Si bien no hay ninguna independiente variable que tenga una alta correlación con la variable de precio, en una primera instancia se observa cierta correlación positiva con la superficie construida y con la cantidad de cuartos.

DISTRIBUCIÓN GEOGRÁFICA DE LA OFERTA INMOBILIARIA

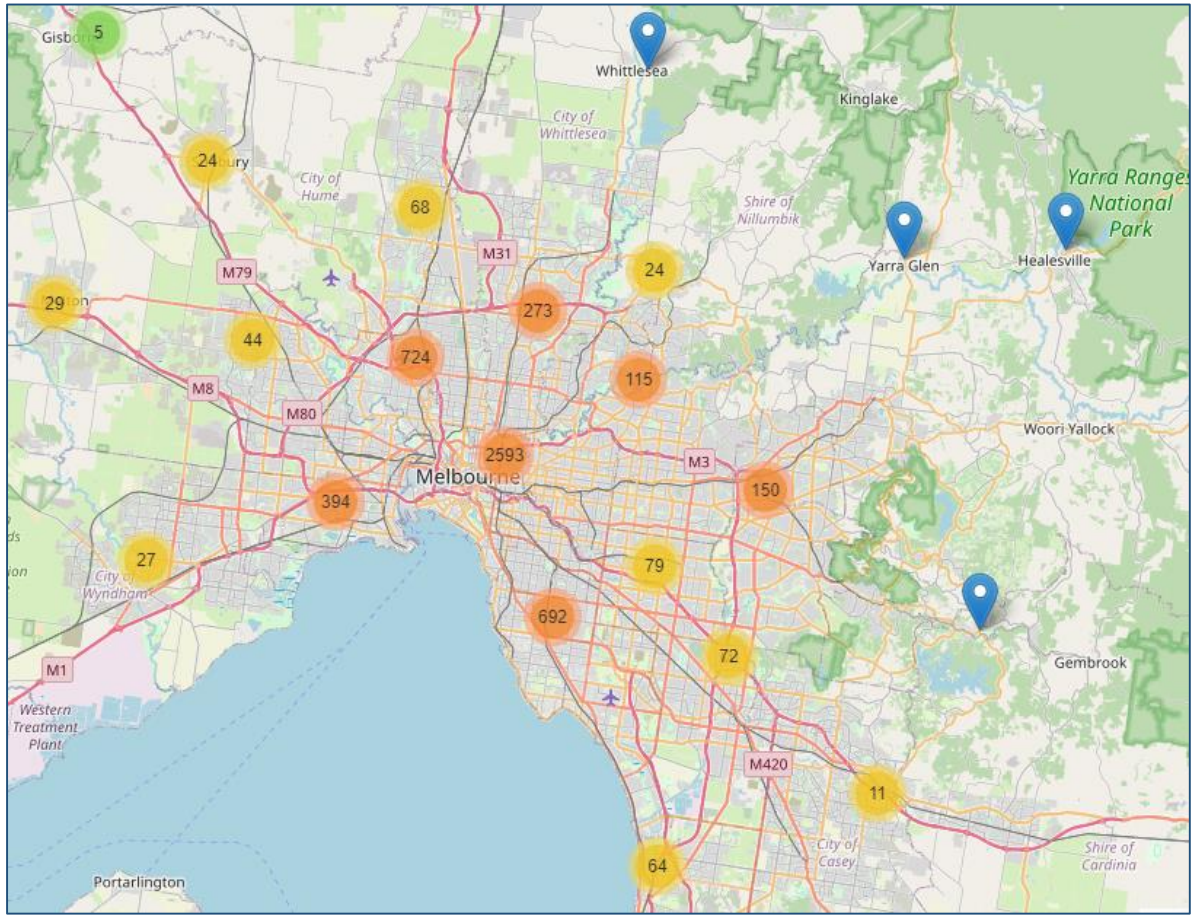
MAPA DE CALOR



Ver en mapa

mapa_calor.html

NODOS SEGÚN BARRIO/ SUBURBIO



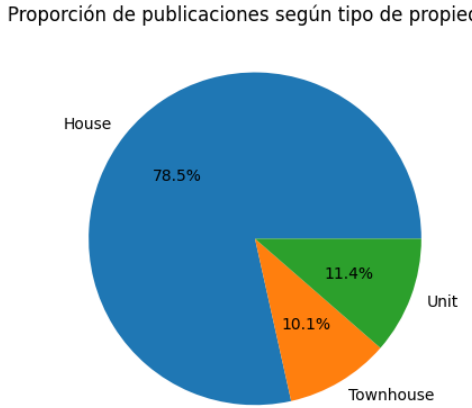
Ver en mapa

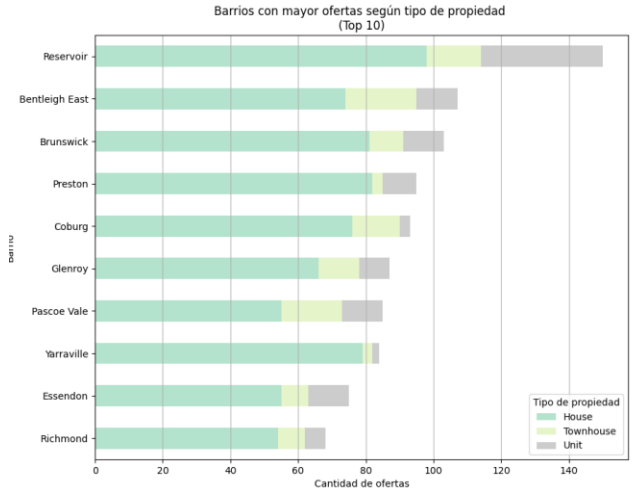
mapa_nodos.html

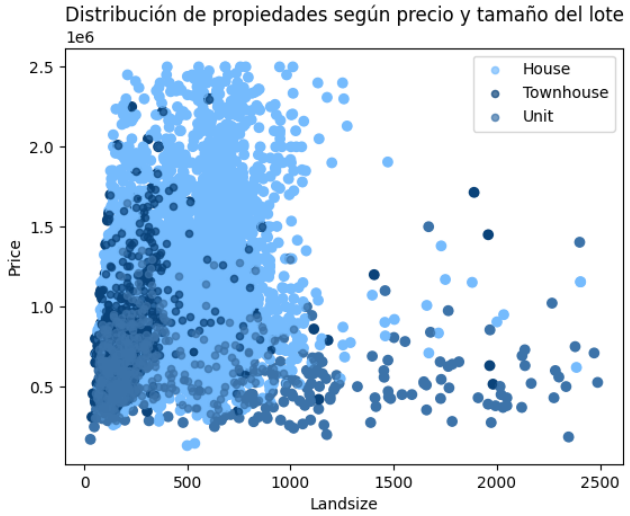
De forma esperable, la mayor concentración de la oferta se nuclea en el centro de la ciudad y la Bahía, y se expande progresivamente hacia los suburbios en forma radial

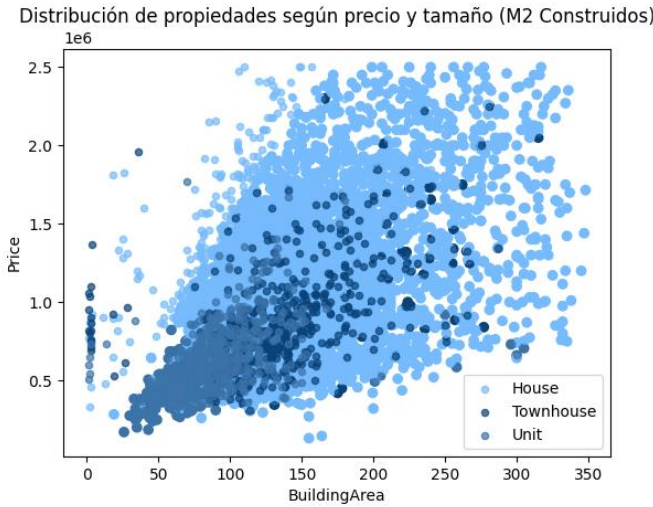
CODERHOUSE

Tipo de vivienda





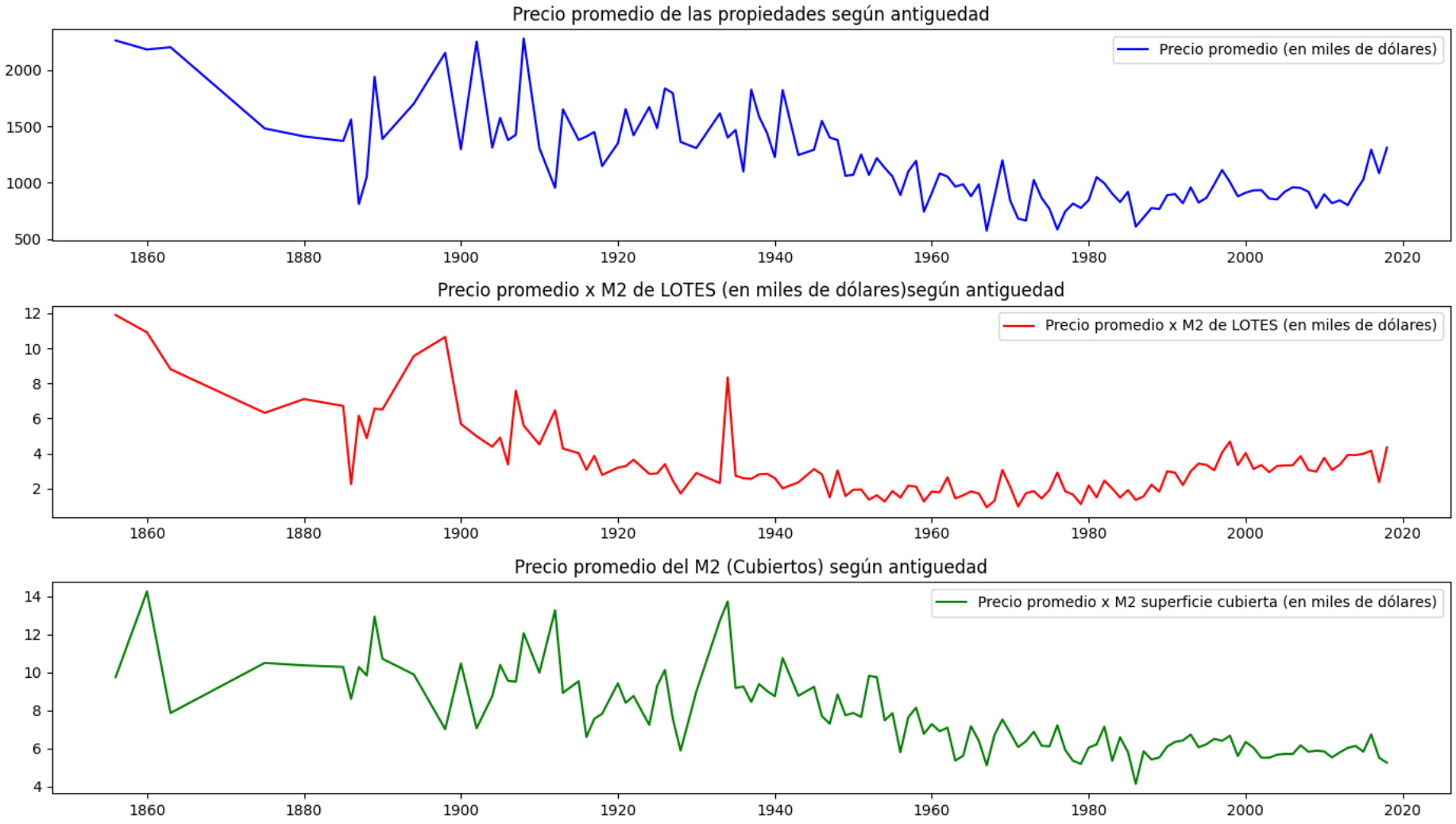




- La mayoría de las publicaciones (casi 8 de cada 10) son “Casas”.
- Los barrios que concentran la mayor cantidad de ofertas publicadas son también los que tienen una mayor proporción relativa de otros tipos de viviendas que no son casas (*Units* o PHs y *Townhouses* o departamentos)
- No se observa una relación clara entre el tamaño del lote o terreno y el precio
- En cambio, al considerar el tamaño de la vivienda (área construida) si se insinúa cierta correlación positiva con el precio.

CODERHOUSE

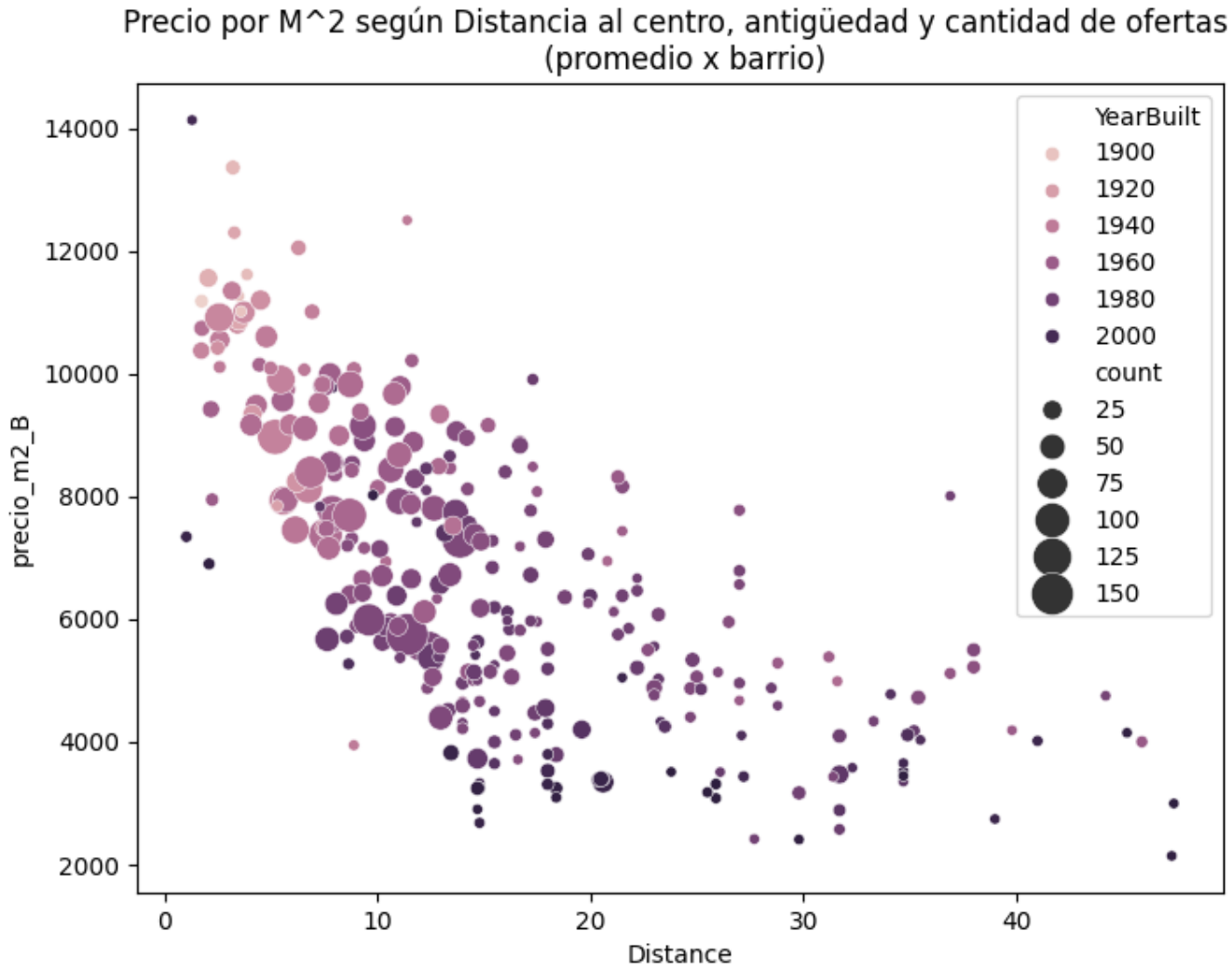
INTERACCION ENTRE ATRIBUTOS: Antigüedad vs Precio



Si bien hay muchos picos, se observa una **tendencia decreciente,** según la cual de a que **a medida que las propiedades son mas nuevas, el precio por metro cuadrado tiende a reducirse.**

CODERHOUSE

INTERACCION ENTRE ATRIBUTOS:
Distancia al centro, antigüedad y precio



Al agruparlo por barrio y comparar las medias vemos existe una **correlación inversa entre el precio y la distancia al centro**. Se observa que cuanto más se aleja cae el precio y la cantidad de ofertas, pero a la vez son viviendas más nuevas. Esto coincide con lo observado en los mapas y la antigüedad: A medida que crece el casco urbano se expande sobre áreas rurales, permitiendo viviendas de mayor tamaño, pero de menor precio por Metro cuadrado construido, y a la vez, con una menor densidad o concentración de viviendas.

CODERHOUSE

PRINCIPALES EMERGENTES DEL EDA

- ✓ La distribución de la oferta tiene una **forma radial** desde el centro de la ciudad hacia la periferia.
- ✓ Casi el **80%** de la oferta son **casas**, aunque en el centro de la ciudad hay mayor diversidad en la tipología.
- ✓ La **distancia al centro comercial** de la ciudad tiene una **relación negativa con el precio promedio**.
- ✓ Se encuentra una **relación positiva entre el tamaño de la vivienda y el precio** de la misma, y parece haber una relación entre el precio y el tamaño del lote. Esto podría explicarse por una relación espuria, donde lo que estamos observando es en realidad el efecto de la distancia.
- ✓ La cantidad de cuartos esta muy asociado al tamaño de la vivienda, y por lo tanto al precio de la misma.
- ✓ La **antigüedad de las viviendas muestra cierta correlación con el precio**, pero sobre todo con la distancia al centro. Esto puede explicar su efecto sobre el precio.

CODERHOUSE

INGENIERIA DE ATRIBUTOS

En el proceso de enriquecimiento de la base de datos y de ingeniería de atributos se realizaron dos conjuntos de operaciones:

1. Se elaboraron **atributos de tipo geográfico**, que dieran cuenta del entorno. Esto se hizo utilizando información incluida en el propio dataset, así como incorporando información geográfica externa al mismo (Límites barriales y concentración de oferta de Airbnb).

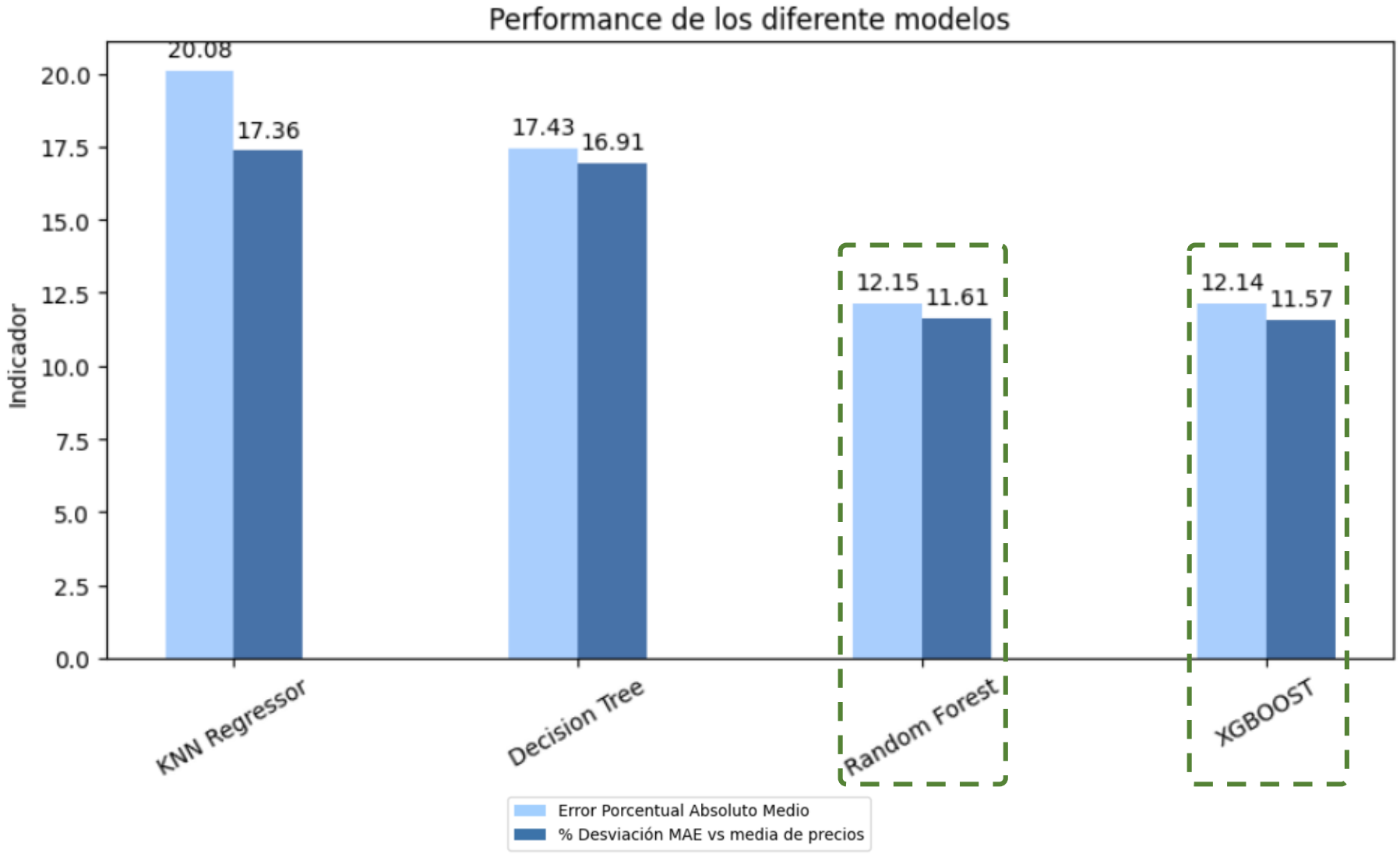
   ***Nota:** La proyección geográfica utilizada fue EPSG:4203, correspondiente al departamento de Victoria, Australia.*

2. Se calcularon **nuevos atributos intrínsecos al dataset**: Las más relevantes fueron el *Factor de Ocupación total (FOT)*, es decir, la relación existente entre el tamaño del lote y la superficie construida, y *Precio promedio del M2 por barrio*.

Durante el proceso de análisis se consideraron **otros atributos, pero fueron descartados por no aportar mejoras al modelo** (Los principales fueron clusters de concentración de ofertas mediante K Means y mediante DBSCAN, así como clusters de densidad (DBSCAN) de concentración de publicaciones de Airbnb. En este último caso se terminó optando por buffers de 500 metros.

CODERHOUSE

COMPARATIVA DE MODELOS



CODERHOUSE

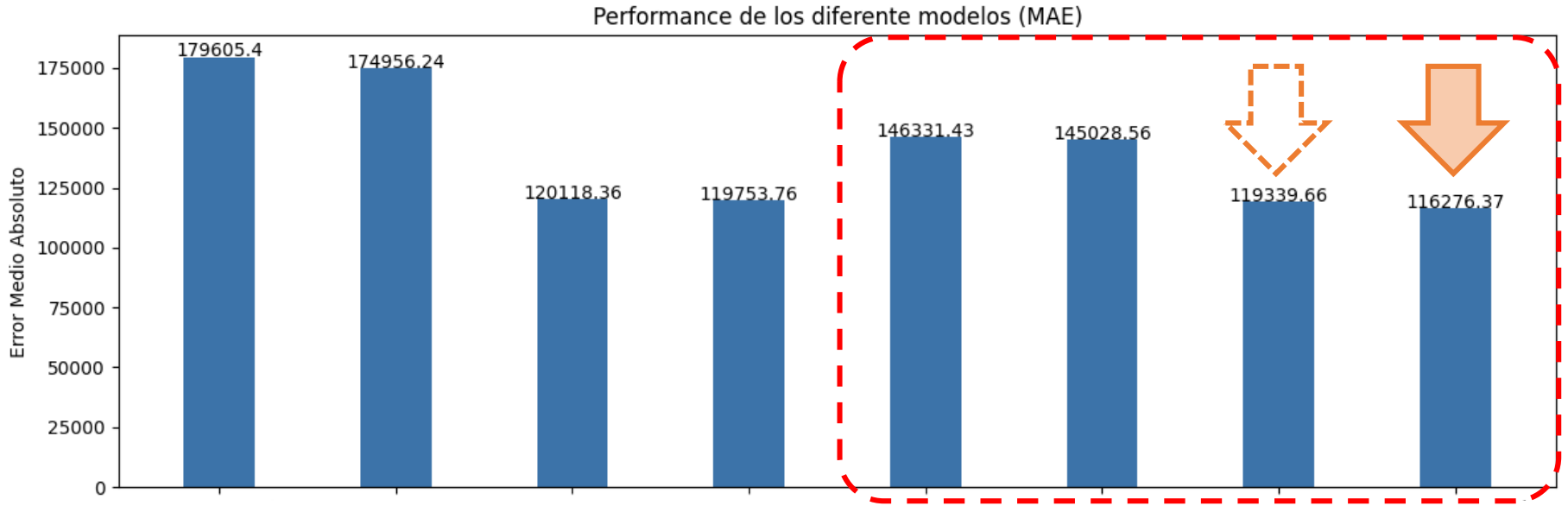
SELECCIÓN DEL MODELO

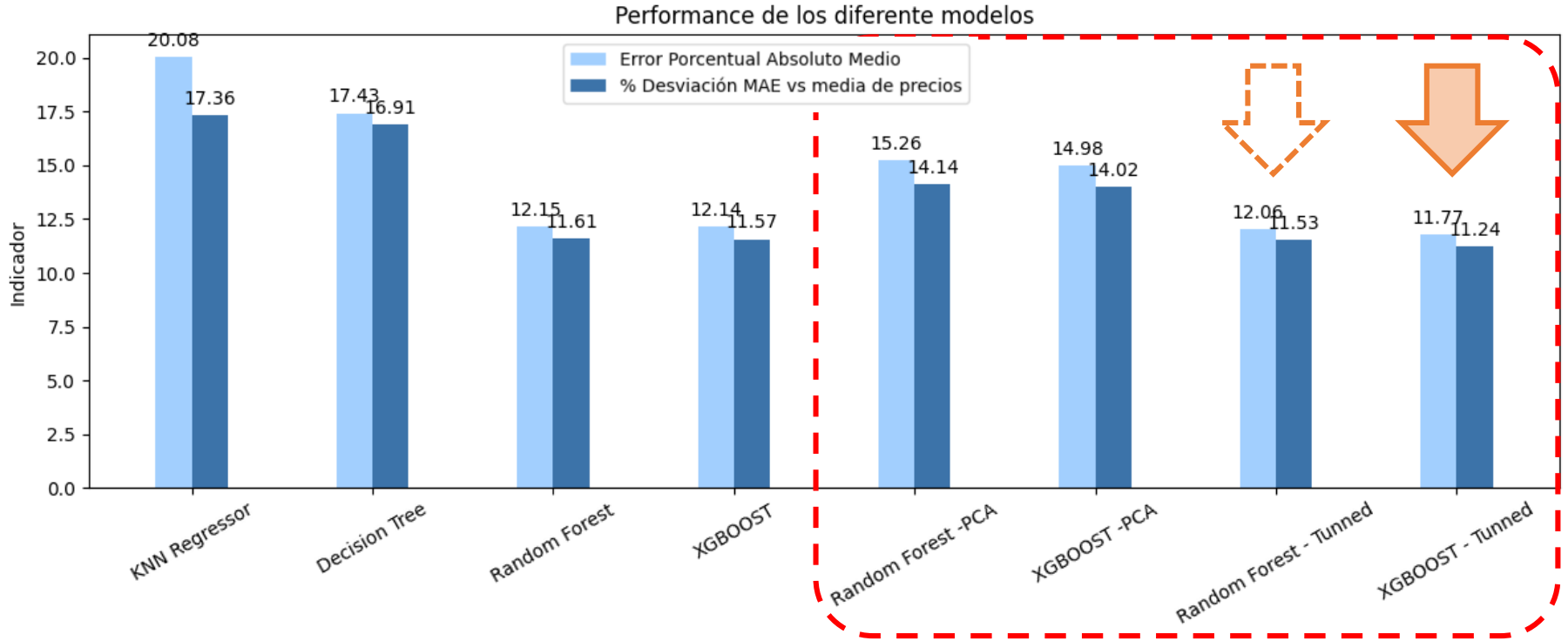
	Modelo	Características	N vars	MAE	MSE	R2	MAPE	Desviación vs media (%)
M1	**KNN Regressor**	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	179605,4	62067668909	0,698	20,084	17,36
M1	**Decision Tree**	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	174956,2	62899474427	0,694	17,432	16,91
M1	**Random Forest**	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	**120118,4**	**62899474427**	**0,867**	**12,153**	**11,61**
M1	**XGBOOST**	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	**119753,8**	**27963388219**	**0,864**	**12,139**	**11,57**

Luego de una primera iteración, los modelos que muestran las mejores métricas son el **Random Forest y XGBOOST**, ambos con valores muy cercanos. Selecciono **ambos para optimizar**.

CODERHOUSE

OPTIMIZACIÓN DEL MODELO: PRINCIPALES INDICADORES





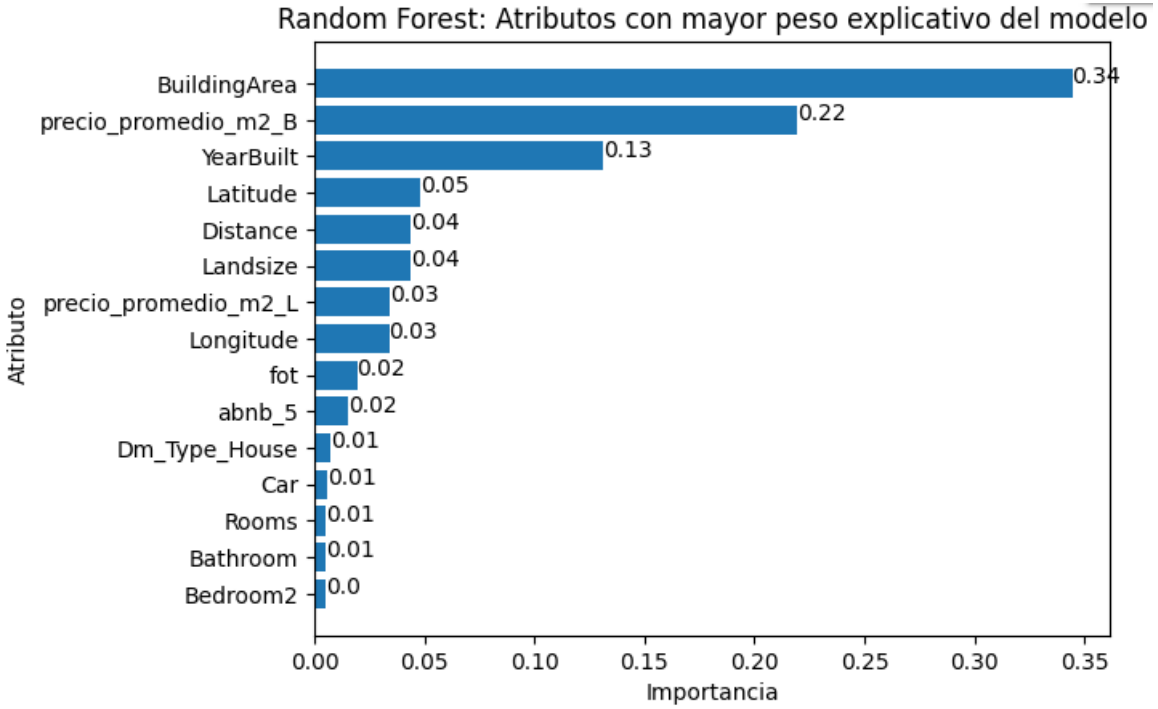
CODERHOUSE

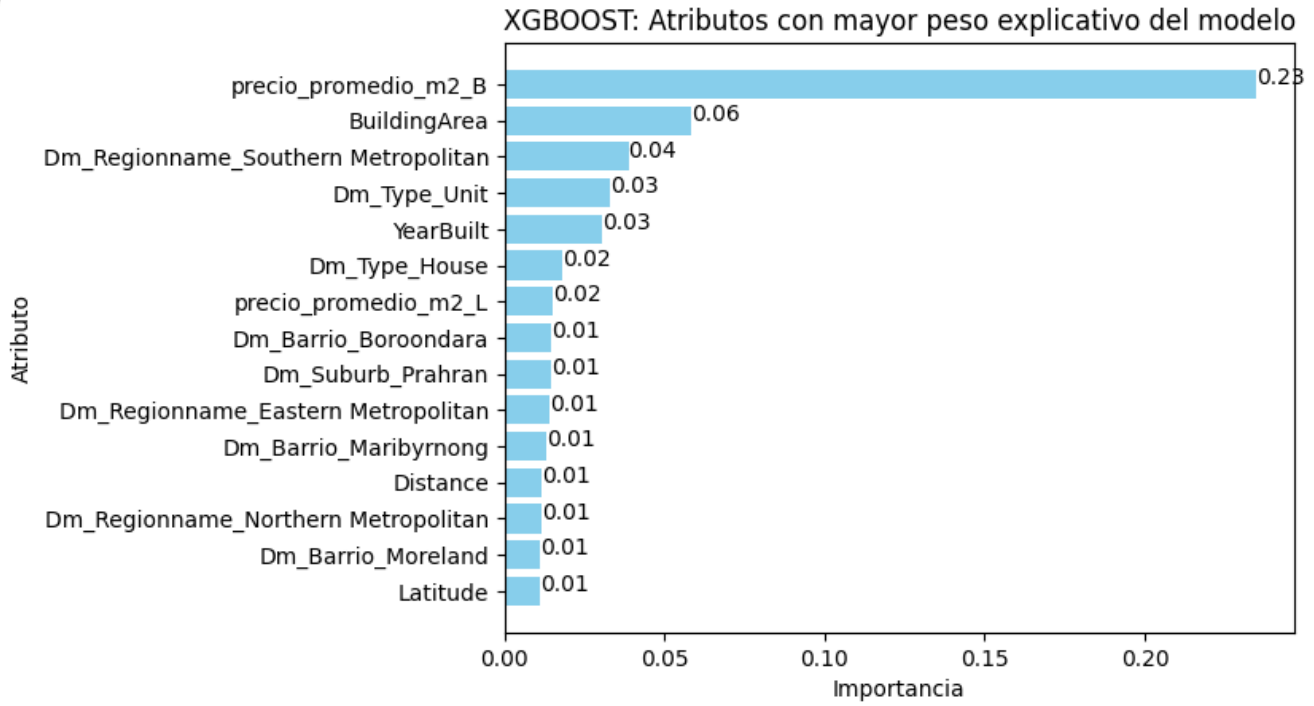
OPTIMIZACIÓN DEL MODELO: PRINCIPALES INDICADORES

	Modelo	Características	N vars	MAE	MSE	R2	MAPE	Desviación vs media (%)
M1	KNN Regressor	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	179605,4	62067668909	0,698	20,084	17,36
M1	Decision Tree	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	174956,2	62899474427	0,694	17,432	16,91
M1	Random Forest	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	120118,4	62899474427	0,867	12,153	11,61
M1	XGBOOST	Modelo Básico. Incluye todas las variables propias + atributos geo y nuevos atributos.	18	119753,8	27963388219	0,864	12,139	11,57
M2	**Random Forest - PCA**	PCA	18	146331,4	40347061730	0,803	15,263	14,14
M2	**XGBOOST -PCA**	PCA	18	145028,6	39361056212	0,808	14,982	14,02
M3	**Random Forest - Tunned**	**Crosvalidation y parámetros optimizados con Halving Grid Search**	**18**	**119339,7**	**39361056212**	**0,869**	**12,058**	**11,53**
M3	**XGBOOST - Tunned**	**Crosvalidation y parámetros optimizados con Halving Grid Search**	**18**	**116276,4**	**25531232962**	**0,876**	**11,773**	**11,24**

Al incorporar reducción de dimensionalidades mediante Análisis de Componentes Principales (PCA), empeora la performance del modelo e incrementa su tiempo de procesamiento, por lo cual fue descartado.
Ambos modelos optimizados muestran una leve mejora con respecto a su versión original, alcanzando niveles muy similares entre ellos.

OPTIMIZACIÓN DEL MODELO: PRINCIPALES ATRIBUTOS INTERVINIENTES





En el modelo de **Random Forest Optimizado**, la **superficie construida** es el atributo con mayor peso explicativo del modelo (34%), seguido por el precio promedio por barrio del Metro Cuadrado (22%), el año de construcción/antigüedad (13%), Latitud (5%), Distancia (4%) y tamaño del lote (4%). Es decir, **entre los principales 6 atributos concentran el 82% del peso explicativo del modelo**.

Por el contrario, en el modelo de **XGBOOST** el **peso explicativo se encuentra mucho más distribuido**, alcanzado para sus principales 6 atributos un 40% de del peso explicativo acumulado. En este modelo, precio promedio por barrio del Metro Cuadrado es la principal variable explicativa (23%), seguida por la superficie construida (6%) y la pertenencia a la región Southern Metropolitan (4%), que es el centro comercial de la ciudad.

Cabe indagar y profundizar más en esta diferencia observada, y en **la posibilidad de generar modelos específicos con un distinto conjunto de atributos**, para tratar de mejorar sus performance.

CODERHOUSE

PRINCIPALES RESULTADOS (I)

- El modelo con mejor peformance es XGBOOST, alcanzando -una vez optimizados sus hiperparámetros y cross validation mediante - un Error Porcentual Absoluto Medio (MAPE) de 11.77, y con una bondad de ajuste (R2) de 0.88. El Error Medio Absoluto (MAE) es de 116276.37, que en relación a la media de precios del conjunto de testeo representa un error de 11,24%. Es decir, que **las predicciones alcanzadas tienen un error de aproximadamente 11% en el valor de las viviendas.**

- El modelo con mejor peformance es XGBOOST, alcanzando -una vez optimizados sus hiperparáemtros y cross validation mediante - un Error Porcentual Absoluto Medio (MAPE) de 11.77, y con una bondad de ajuste (R2) de 0.88. El Error Medio Absoluto (MAE) es de 116276.37, que en relación a la media de precios del conjunto de testeo representa un error de 11,24%. Es decir, que **las predicciones alcanzadas tienen un error de aproximadamente 11% en el valor de las viviendas.**

- En dicho modelo, el precio promedio por barrio del Metro Cuadrado construido es el atributo con mayor peso explicativo, alcanzando una importancia de 0.23, seguido por la superficie total construida (0.06) y perteneciente a la región de Southern Metropolitan (0.04). El tipo de propiedad (0.03 para "Unit" y 0.02 para "House"), y la antigüedad (0.03) también tienen un rol relevante.

CODERHOUSE

PRINCIPALES RESULTADOS (II)

- El Análisis de Componentes Principales (PCA) no implicó mejoras en la performance del modelo, por lo cual fue descartado.

- El hecho de que la Latitud aparezca entre los atributos con mayor peso y no así la longitud, nos indica que es más determinante en el precio el eje Norte-Sur que el Este-Oeste.

- La diferencia en el peso explicativo de los atributos y su distribución entre los dos modelos optimizados sugiere una posible vía para mejorar los modelos **generando conjuntos de atributos específicos para cada uno**.

- En contra de la hipótesis planteada al principio, la distancia al centro no parece ser por si misma una variable con suficiente peso explicativo. Pareciera estar subsumido su efecto con las dimensiones de superficie y de precio promedio por barrio.