

Resultados prueba técnica

Juan Camilo Salgado Ramírez

Agosto 2022

Pregunta 1: Variables que afectan el precio por m² de una vivienda

Objetivo y metodología

La idea en esta pregunta es entender qué factores de una vivienda afectan su precio.

Para resolver este reto seguí los siguientes pasos:

1. **Búsqueda de información adicional:** incluí datos de criminalidad alrededor de cada vivienda.
2. **Limpieza de información:** ejecuté diferentes procesos para asegurar que la información tenga la calidad óptima para estimar modelos.
3. **Análisis descriptivo:** correlaciones simples para entender relaciones entre los features y el precio_m2.
4. **Estimación de modelos:** estimé una regresión lineal simple (inferencia) y un modelo de XGboosting (predicción).

Búsqueda de información adicional

Existe mucha información que complementaría el análisis y permitiría entender qué factores influyen en el precio por m² de una vivienda. Por ejemplo, la cantidad de amenities (parques, centros comerciales, restaurantes, etc) cercanas a la vivienda, o inclusive la distancia de la vivienda a sitios de trabajo podrían afectar su precio. Por cuestiones de tiempo me limité a incluir información oficial de criminalidad geolocalizada, reportada por el Gobierno de la CDMX, para entender qué tantos crímenes se cometan cerca a una vivienda.

idCarpeta	Año_inicio	Mes_inicio	FechaInicio	Delito	Categoría	Sexo	Edad	TipoPersona	CalidadJuridica	competencia	Año_hecho	Mes_hecho	FechaHecho	H
8828678.0	2021.0	Enero	01/01/2021	PORACION DE ARMA DE FUEGO	DELITO DE BAJO IMPACTO	NaN	NaN	MORAL	OFENDIDO	INCOMPETENCIA	2021.0	Enero	01/01/2021	
8828681.0	2021.0	Enero	01/01/2021	ROBO DE ACCESORIOS DE AUTO	DELITO DE BAJO IMPACTO	NaN	NaN	MORAL	VICTIMA	FUERO COMUN	2021.0	Enero	01/01/2021	

Decidí incorporar esta información de la siguiente manera:

1. Clasifiqué el tipo de crimen en 3 grupos: crimen grave, crimen no grave y robo.
2. Calculé la totalidad de incidentes de cada una de estas categorías en un radio de 200 metros alrededor de cada vivienda.

Búsqueda de información adicional

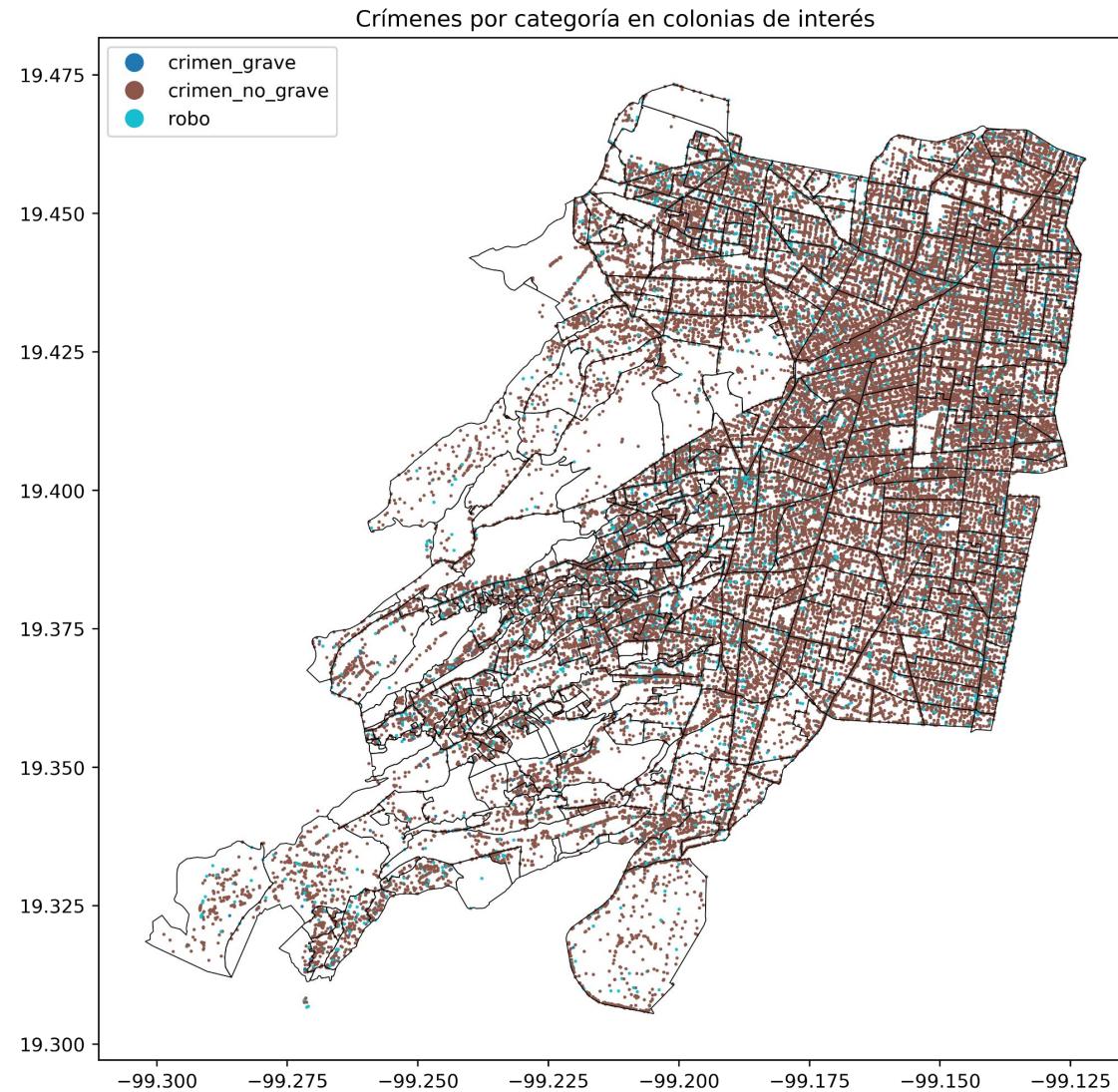
DELITO DE BAJO IMPACTO	69308
ROBO A TRANSEUNTE EN VÍA PÚBLICA CON Y SIN VIOLENCIA	4083
ROBO DE VEHÍCULO CON Y SIN VIOLENCIA	1676
HECHO NO DELICTIVO	1175
ROBO A NEGOCIO CON VIOLENCIA	848
ROBO A REPARTIDOR CON Y SIN VIOLENCIA	580
ROBO A PASAJERO A BORDO DEL METRO CON Y SIN VIOLENCIA	503
VIOLACIÓN	445
HOMICIDIO DOLOSO	266
LESIONES DOLOSAS POR DISPARO DE ARMA DE FUEGO	215
ROBO A PASAJERO A BORDO DE TAXI CON VIOLENCIA	127
ROBO A CASA HABITACIÓN CON VIOLENCIA	107
ROBO A CUENTAHABIENTE SALIENDO DEL CAJERO CON VIOLENCIA	97
ROBO A PASAJERO A BORDO DE MICROBUS CON Y SIN VIOLENCIA	71
ROBO A TRANSPORTISTA CON Y SIN VIOLENCIA	8
VIOLACIÃ<U+0093>N	6
ROBO DE VEHÃCULO CON Y SIN VIOLENCIA	2
SECUESTRO	1

Name: Categoría, dtype: int64



crimen_no_grave	70483
robo	8102
crimen_grave	933

Búsqueda de información adicional



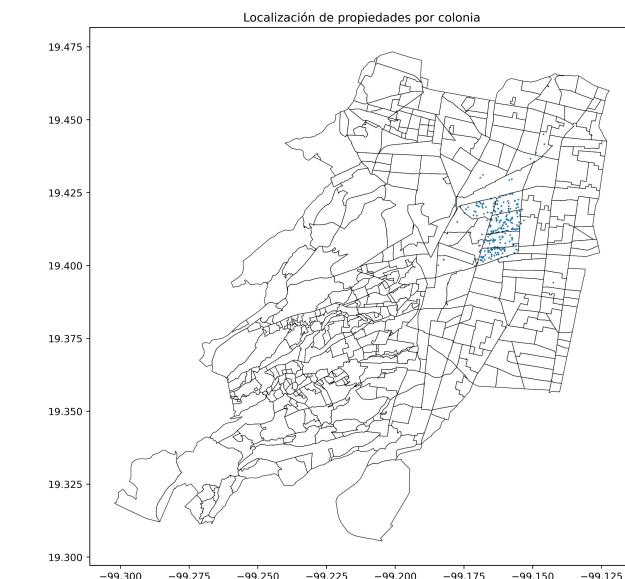
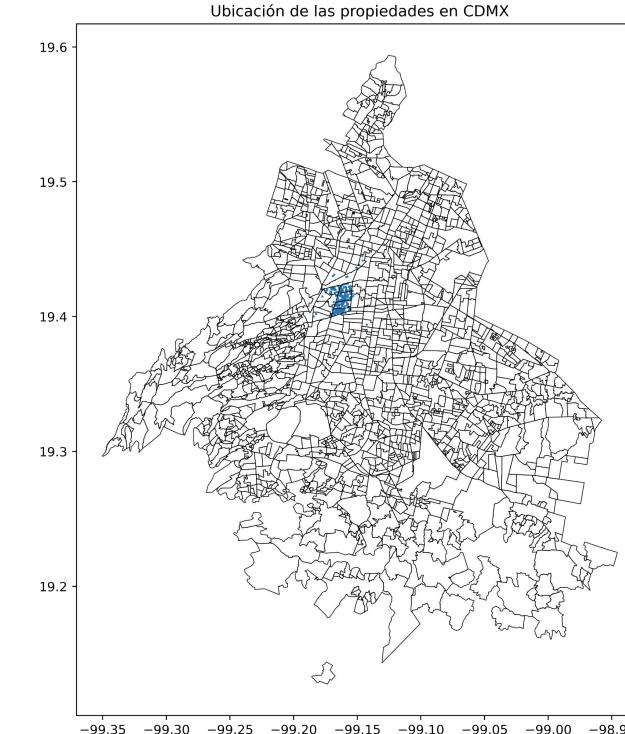
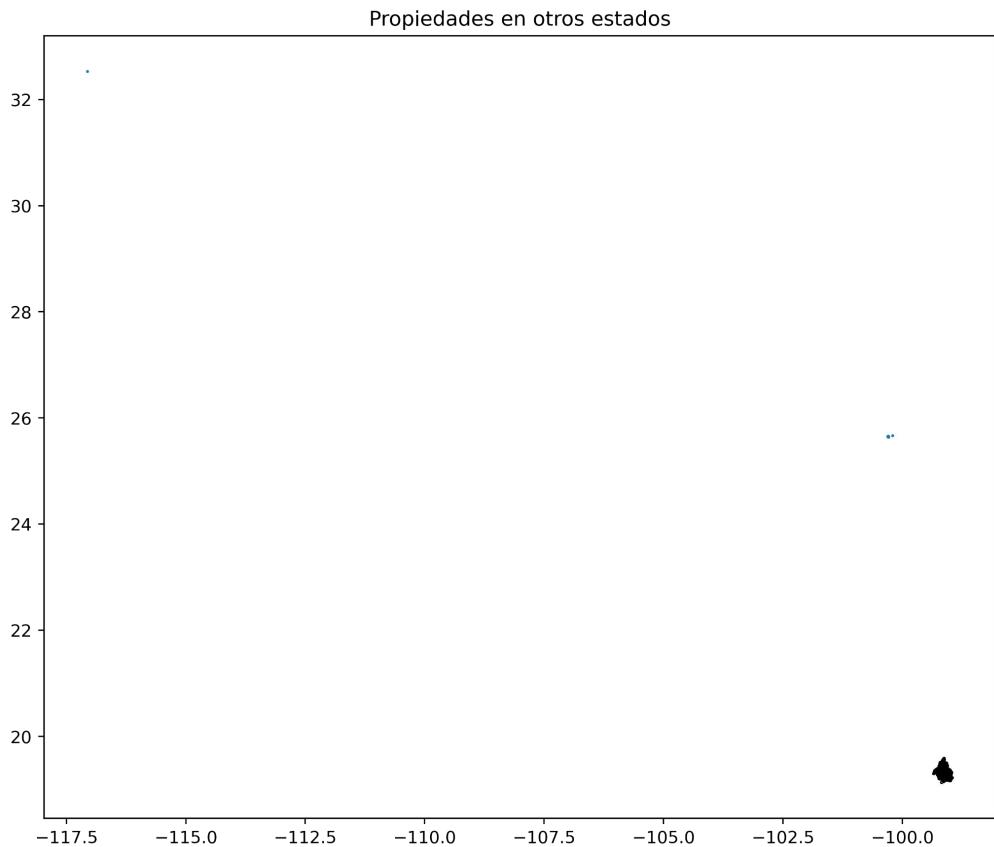
Limpieza de información

Desarrollé los siguientes pasos para limpiar los datos:

1. Crucé la localización de cada vivienda con la capa de colonias de CDMX y eliminé datos de otros estados.
2. Eliminé datos con más de un año de publicación, porque los precios de 2 años atrás son muy distintos a los actuales. Por lo tanto usar esos valores incluiría ruido al análisis.
3. Eliminé el 1% más alto de las variables baños y área de la vivienda, porque se veían muy raros.

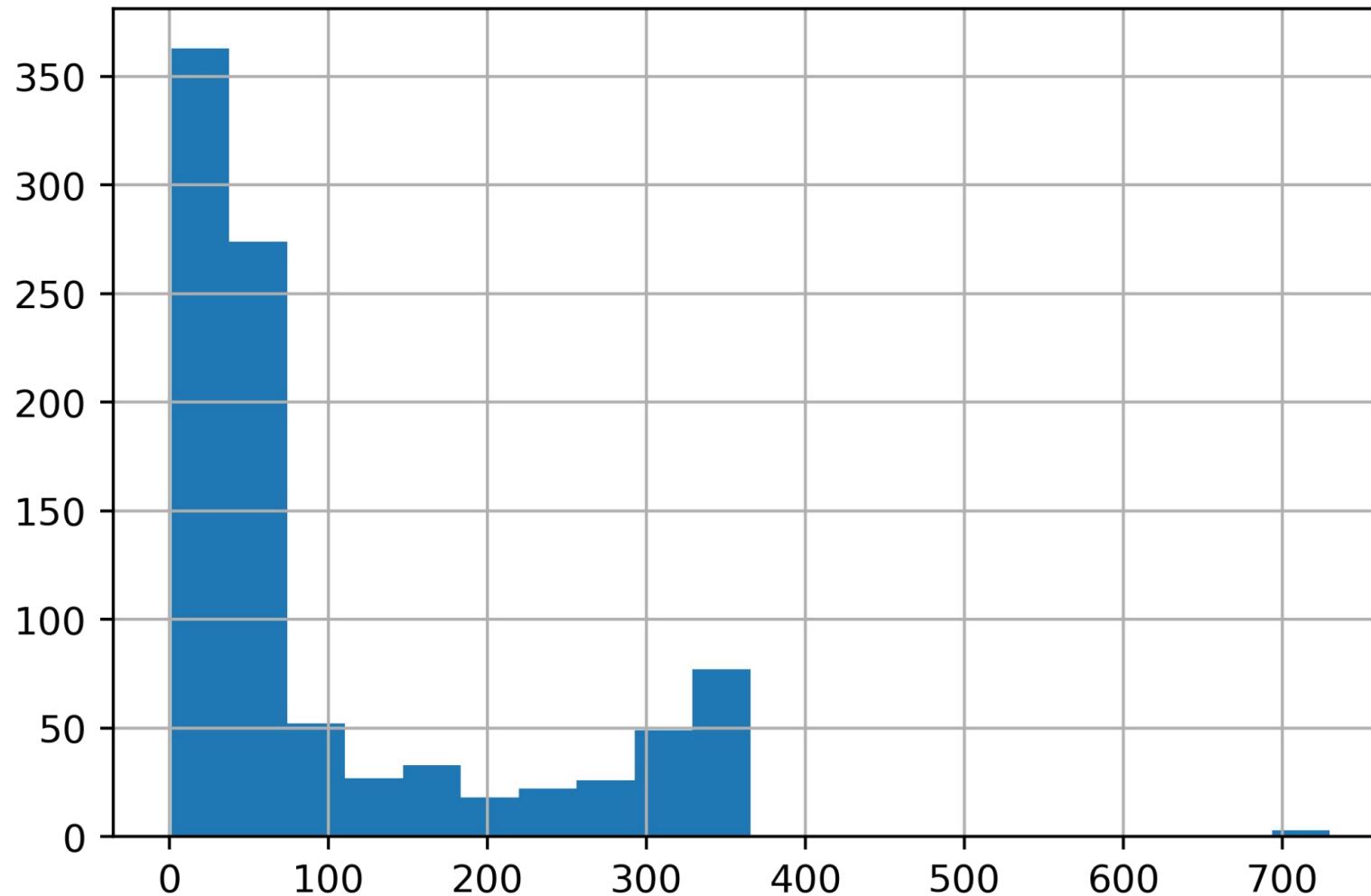
Adicionalmente consideré que para la variable amenities los valores nulos se podían interpretar como un cero.

Limpieza de información



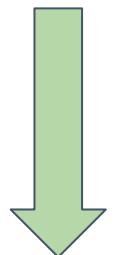
Limpieza de información

Distribución tiempo de publicación en sitio



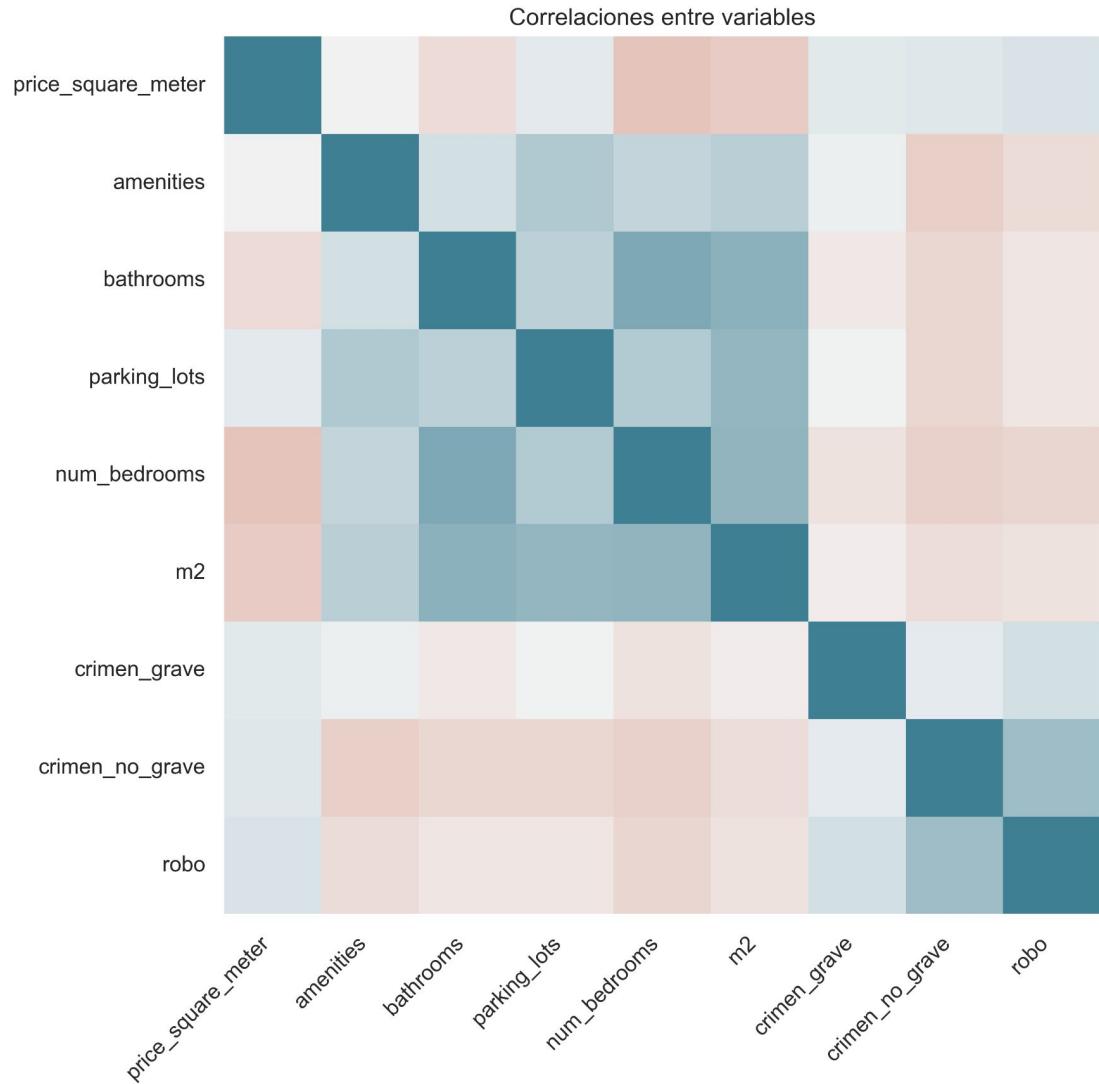
Limpieza de información

	price_square_meter	amenities	bathrooms	parking_lots	num_bedrooms	m2	crimen_grave	crimen_no_grave	robo
count	941.000000	941.000000	941.000000	941.000000	941.000000	941.000000	941.000000	941.000000	941.000000
mean	61419.038109	2.077577	1.977683	1.331562	2.096706	112.473974	0.111583	33.524973	3.202976
std	19200.329466	2.208142	0.897118	0.479974	0.613847	236.271807	0.431824	12.607692	2.383059
min	582.246879	0.000000	1.000000	1.000000	1.000000	31.000000	0.000000	0.000000	0.000000
25%	53033.707865	0.000000	2.000000	1.000000	2.000000	75.000000	0.000000	23.000000	2.000000
50%	61837.096774	2.000000	2.000000	1.000000	2.000000	93.000000	0.000000	33.000000	3.000000
75%	70866.141732	4.000000	2.000000	2.000000	2.000000	129.000000	0.000000	41.000000	4.000000
max	146524.373626	8.000000	23.000000	3.000000	4.000000	7210.000000	3.000000	82.000000	26.000000



	price_square_meter	amenities	bathrooms	parking_lots	num_bedrooms	m2	crimen_grave	crimen_no_grave	robo
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	61658.452097	2.092593	1.922658	1.331155	2.076253	102.075174	0.113290	33.674292	3.204793
std	19093.588096	2.187286	0.527859	0.480059	0.605916	41.820958	0.435875	12.627177	2.395356
min	4712.041885	0.000000	1.000000	1.000000	1.000000	31.000000	0.000000	0.000000	0.000000
25%	53333.333333	0.000000	2.000000	1.000000	2.000000	75.000000	0.000000	24.000000	2.000000
50%	61851.134021	2.000000	2.000000	1.000000	2.000000	93.000000	0.000000	33.000000	3.000000
75%	70824.685664	4.000000	2.000000	2.000000	2.000000	127.000000	0.000000	41.750000	4.000000
max	146524.373626	7.000000	3.000000	3.000000	4.000000	270.000000	3.000000	82.000000	26.000000

Análisis descriptivo



- Las variables más correlacionadas con el precio por metro cuadrado de una vivienda son el número de baños, número de habitaciones y el área de la vivienda. Cada una de estas variables se relaciona negativamente.
- Las variables adicionales que incluí en el análisis tienen una correlación leve y positiva con el precio por m² de la vivienda.

Estimación de modelos

OLS Regression Results

Dep. Variable:	price_square_meter	R-squared:	0.182			
Model:	OLS	Adj. R-squared:	0.175			
Method:	Least Squares	F-statistic:	25.35			
Date:	Sat, 06 Aug 2022	Prob (F-statistic):	1.74e-35			
Time:	16:15:50	Log-Likelihood:	-10258.			
No. Observations:	918	AIC:	2.053e+04			
Df Residuals:	909	BIC:	2.058e+04			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	6.58e+04	3308.884	19.887	0.000	5.93e+04	7.23e+04
amenities	465.6494	289.868	1.606	0.109	-103.239	1034.538
bathrooms	5409.1897	1503.317	3.598	0.000	2458.814	8359.566
parking_lots	1.161e+04	1444.539	8.039	0.000	8777.884	1.44e+04
num_bedrooms	-9663.6735	1298.350	-7.443	0.000	-1.22e+04	-7115.562
m2	-146.0647	19.054	-7.666	0.000	-183.460	-108.670
crimen_grave	1496.3716	1339.418	1.117	0.264	-1132.340	4125.084
crimen_no_grave	75.0009	52.684	1.424	0.155	-28.396	178.397
robo	405.4822	274.620	1.477	0.140	-133.481	944.445
Omnibus:	67.549	Durbin-Watson:	1.551			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	300.923			
Skew:	0.124	Prob(JB):	4.52e-66			
Kurtosis:	5.794	Cond. No.	684.			

- Las variables significativas son el número de baños (+), número de garajes (+), habitaciones (-) y el área de la vivienda (-).
- Las variables relacionadas con la criminalidad cercana a una vivienda no son relevantes estadísticamente.

Estimación de modelos

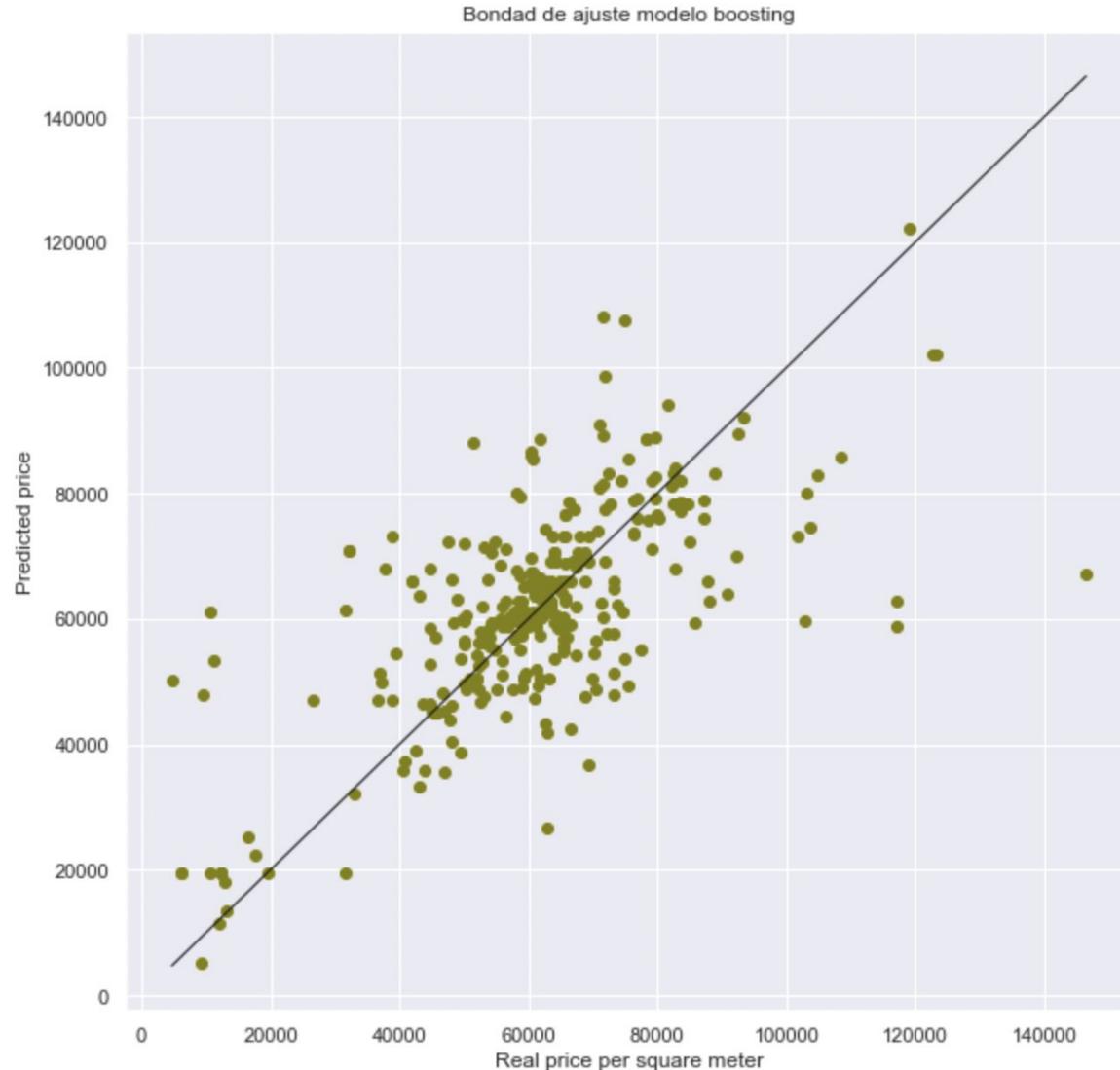
OLS Regression Results

Dep. Variable:	price_square_meter_log	R-squared:	0.171			
Model:	OLS	Adj. R-squared:	0.164			
Method:	Least Squares	F-statistic:	23.48			
Date:	Sat, 06 Aug 2022	Prob (F-statistic):	6.82e-33			
Time:	16:15:50	Log-Likelihood:	-457.54			
No. Observations:	918	AIC:	933.1			
Df Residuals:	909	BIC:	976.5			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	10.9626	0.076	143.525	0.000	10.813	11.113
amenities	0.0324	0.007	4.842	0.000	0.019	0.046
bathrooms	0.1523	0.035	4.388	0.000	0.084	0.220
parking_lots	0.2344	0.033	7.030	0.000	0.169	0.300
num_bedrooms	-0.2413	0.030	-8.050	0.000	-0.300	-0.182
m2	-0.0028	0.000	-6.293	0.000	-0.004	-0.002
crimen_grave	-0.0289	0.031	-0.933	0.351	-0.090	0.032
crimen_no_grave	0.0019	0.001	1.583	0.114	-0.000	0.004
robo	0.0141	0.006	2.224	0.026	0.002	0.027
Omnibus:	417.576	Durbin-Watson:	1.399			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2312.546			
Skew:	-2.044	Prob(JB):	0.00			
Kurtosis:	9.614	Cond. No.	684.			

- Para entender la magnitud en la que las variables significativas afectan el precio por metro cuadrado de una vivienda, estimé una regresión pero con la variable de precio_m2 en logaritmo, lo cual pone las cifras en términos porcentuales.
- Incrementar el número de baños en 1 incrementa el precio por m2 de una vivienda en 15%, un garaje adicional incrementa el precio_m2 en 23%, una habitación adicional disminuye el precio_m2 en 24% y un metro cuadrado adicional disminuye el precio en 0.3%.

Estimación de modelos

MAPE XG: 25.38 %



- Estimé un modelo de XGboosting con estas mismas variables para entender qué tan bien lograríamos predecir el precio por m².
- El modelo logra capturar los patrones que hacen que una vivienda cara sea cara, o visceversa.
- Sin embargo el ajuste no es tan bueno, porque el MAPE es de 25%, lo cual indica que en promedio el modelo se equivoca un 25%, lo cual es mucho para una variable tan sensible como el precio de una vivienda.

Conclusiones

Los datos parecen indicar que las variables que más se relacionan con el precio de una vivienda son: número de baños, número de garajes, número de habitaciones y área del inmueble.

Las variables de criminalidad parecen no estar tan relacionadas con el precio del inmueble

Para incrementar la precisión de un potencial modelo predictivo se podrían incluir variables adicionales como los diferentes tipos de amenities que tiene un inmueble (ej: gimnasio, piscina, etc) o la distancia a sitios de trabajo.

Pregunta 2: Principales temas en reviews de restaurantes

Objetivo y metodología

La idea en esta pregunta es lograr caracterizar los principales temas en una muestra de reviews de restaurantes.

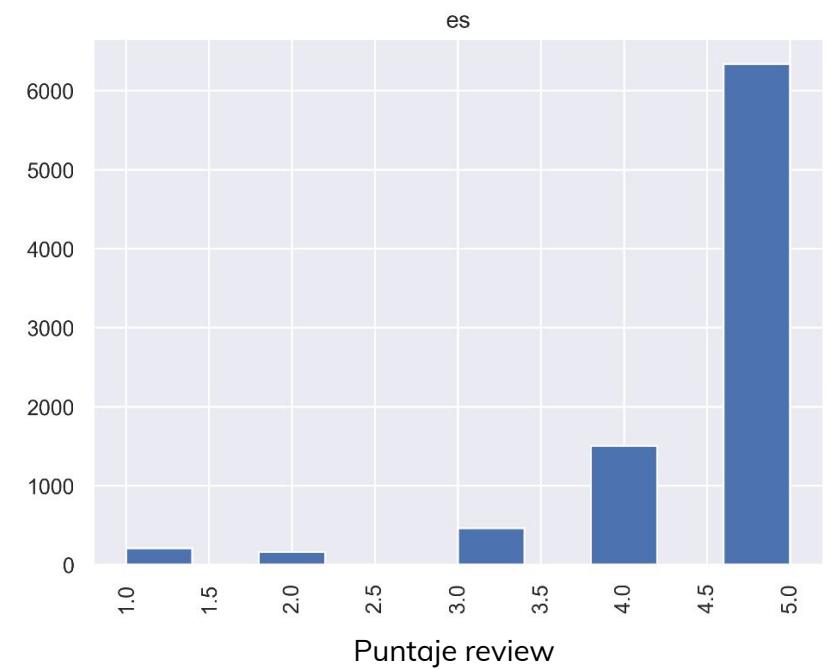
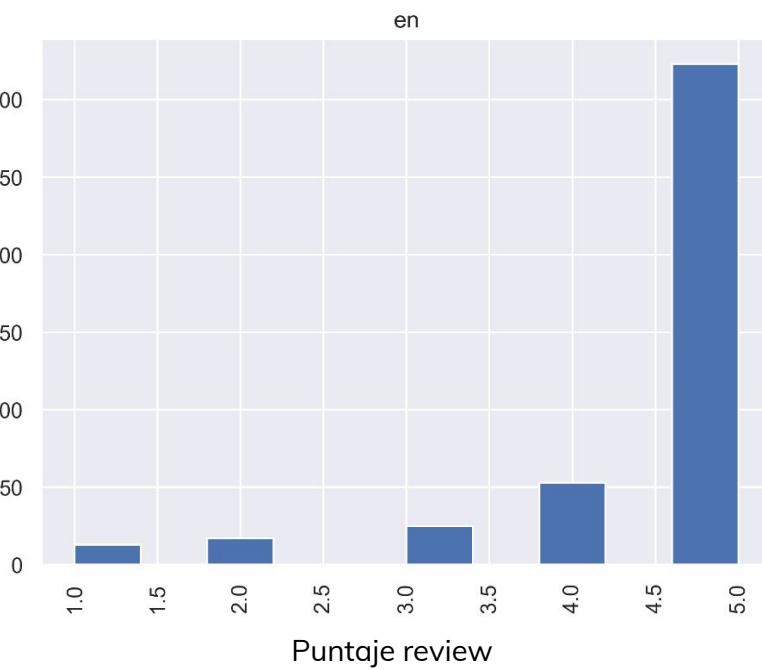
Para resolver este reto seguí los siguientes pasos:

1. **Limpieza de información:** ejecuté diferentes procesos para asegurar que la información tenga la calidad óptima para estimar modelos.
2. **Análisis descriptivo:** entender principales palabras en las reviews
3. **LDA topic modelling:** estimé los 10 principales temas en los que se hablan en las reviews.

Limpieza de información

En general la información estaba bastante limpia. Simplemente me limité a los reviews hechos en español o inglés porque son los que concentraban la mayor cantidad de comentarios:

es	8677
en	431
pt	23
sl	11
fr	6
pl	4
gl	3
id	2
sv	2
ko	2
ar	2
ca	2
ja	2
de	2
zh	2
hi	2
zh-Hant	2
it	2
fi	1
tr	1
ro	1
ru	1
ku	1
cs	1



Análisis descriptivo

Estas son las palabras más comunes usadas en las reviews de cada idioma:

Español



Inglés



Análisis descriptivo

Partiendo por calificaciones bajas (1 o 2) o altas (4 o 5) de las reviews otorgadas:



LDA topic modelling

Desarrollé un modelo estándar de LDA para caracterizar los tópicos principales en las reviews en español y en inglés.

Para realizar este modelo se debe primero realizar los siguientes pasos:

1. Tokenizar las palabras y eliminar stopwords.
2. Extraer la palabra raíz de cada palabra.
3. Entender la frecuencia con que cada palabra aparece en todas las reviews.

Finalmente extraje los 10 principales tópicos de las reviews en español y el inglés, usando las reviews previamente procesadas.

LDA topic modelling

Topic: 0
Words: 0.052*"buen" + 0.023*"lug" + 0.022*"com" + 0.022*"delici" + 0.018*"tac" + 0.013*"mejor" + 0.012*"preci" + 0.012*"servici" + 0.012*"bien" + 0.011*"sup"

Topic: 1
Words: 0.048*"buen" + 0.040*"lug" + 0.033*"com" + 0.013*"excelent" + 0.013*"varied" + 0.011*"atencion" + 0.011*"recomend" + 0.010 *"gran" + 0.009*"sol" + 0.009*"opcion"

Topic: 2
Words: 0.087*"ric" + 0.051*"servici" + 0.045*"buen" + 0.040*"com" + 0.038*"excelent" + 0.017*"atencion" + 0.012*"lug" + 0.010*s i" + 0.010*"amabl" + 0.010*"aliment"

Topic: 3
Words: 0.042*"com" + 0.032*"buen" + 0.029*"ric" + 0.029*"lug" + 0.025*"... " + 0.023*"preci" + 0.021*"mejor" + 0.019*"sabor" + 0.0 17*"atencion" + 0.016*"excelent"

Topic: 4
Words: 0.055*"buen" + 0.053*"atencion" + 0.050*"excelent" + 0.041*"delici" + 0.038*"com" + 0.032*"lug" + 0.017*"ric" + 0.014*"rec omend" + 0.014*"sabor" + 0.014*"preci"

Topic: 5
Words: 0.035*"lug" + 0.024*"buen" + 0.021*"bien" + 0.015*"sabor" + 0.012*"ric" + 0.012*"servici" + 0.011*"much" + 0.010*"calid" + 0.009*"com" + 0.009*"gust"

Topic: 6
Words: 0.042*"buen" + 0.027*"delici" + 0.026*"tac" + 0.024*"lug" + 0.022*"com" + 0.022*"sabor" + 0.021*"mejor" + 0.016*"excelent" + 0.016*"servici" + 0.015*"varied"

Topic: 7
Words: 0.072*"buen" + 0.066*"excelent" + 0.059*"com" + 0.046*"servici" + 0.026*"lug" + 0.024*"preci" + 0.020*"ric" + 0.012*"delic i" + 0.012*"sup" + 0.010*"atencion"

Topic: 8
Words: 0.027*"excelent" + 0.023*"lug" + 0.020*"recomend" + 0.016*"com" + 0.015*"mejor" + 0.015*"encant" + 0.014*"tort" + 0.013*p reci" + 0.013*"ric" + 0.012*"buen"

Topic: 9
Words: 0.051*"com" + 0.033*"lug" + 0.025*"buen" + 0.020*"ric" + 0.019*"delici" + 0.019*"excelent" + 0.014*"ambient" + 0.014*"agra d" + 0.013*"preci" + 0.011*"sabor"

LDA topic modelling

Topic: 0

Words: 0.024*"food" + 0.014*"realli" + 0.013*"also" + 0.012*"great" + 0.012*"tri" + 0.011*"restaur" + 0.009*"place" + 0.009*"us"
+ 0.008*"everyth" + 0.008*"servic"

Topic: 1

Words: 0.018*"delici" + 0.016*"food" + 0.015*"super" + 0.012*"restaur" + 0.011*"would" + 0.010*"definit" + 0.010*"nice" + 0.009*"good"
+ 0.008*"menu" + 0.007*"staff"

Topic: 2

Words: 0.027*"food" + 0.024*"servic" + 0.022*"good" + 0.019*"great" + 0.019*"nice" + 0.015*"place" + 0.015*"restaur" + 0.010*"like"
+ 0.008*"meal" + 0.008*"tradit"

Topic: 3

Words: 0.027*"food" + 0.017*"place" + 0.014*"good" + 0.013*"\$" + 0.012*"great" + 0.010*"us" + 0.008*"delici" + 0.008*"servic" +
0.008*"nice" + 0.008*"excel"

Topic: 4

Words: 0.030*"servic" + 0.025*"food" + 0.016*"excel" + 0.015*"great" + 0.012*"amaz" + 0.012*"restaur" + 0.011*"dish" + 0.009*"good"
+ 0.009*"tast" + 0.008*"menu"

Topic: 5

Words: 0.027*"food" + 0.019*"great" + 0.014*"servic" + 0.013*"place" + 0.012*"restaur" + 0.011*"citi" + 0.011*"mexico" + 0.010*"really"
+ 0.009*"best" + 0.009*"one"

Topic: 6

Words: 0.023*"food" + 0.013*"good" + 0.012*"dish" + 0.012*"us" + 0.011*"realli" + 0.010*"restaur" + 0.010*"servic" + 0.009*":"
+ 0.008*"love" + 0.007*"experi"

Topic: 7

Words: 0.020*"menu" + 0.016*"food" + 0.014*"place" + 0.013*"servic" + 0.011*"tast" + 0.011*"n't" + 0.011*"great" + 0.010*"recomme
nd" + 0.010*"experi" + 0.008*"also"

Topic: 8

Words: 0.032*"food" + 0.019*"servic" + 0.017*"amaz" + 0.016*"experi" + 0.015*"place" + 0.015*"great" + 0.014*"time" + 0.013*"exce
l" + 0.012*"delici" + 0.011*"restaur"

Topic: 9

Words: 0.022*"food" + 0.018*"restaur" + 0.017*"great" + 0.013*"mexico" + 0.011*"servic" + 0.010*"citi" + 0.010*"experi" + 0.009*"go"
+ 0.008*"dish" + 0.007*"would"

LDA topic modelling

Estos son los principales temas identificados por el modelo:

Temas en español	Temas en inglés
Lugar	Food
Variedad comida	Food, menu
Servicio	Service
Precio	Place
La atención	Service
Sabor	Service
Sabor	Dish
Comida	Menu
Lugar	Place
Ambiente	Food

Conclusiones

Tanto en español como en inglés parece haber unos patrones respecto a los temas más relevantes para los clientes al momento de dejar una reseña. Particularmente la calidad y la variedad de la comida, el servicio y el ambiente del restaurante parecen ser las más importantes para los comensales.

Sin embargo una diferencia sutil entre ambos tipos de clientes es que las personas que dejan reviews en inglés parecen no preocuparse por el precio de la comida, mientras que las personas que dejan reviews en español si. Esto podría obedecer a que, por temas de conversión de la moneda y salarios más altos, los clientes extranjeros no se ven tan afectados por precios altos en una moneda distinta a la de ellos porque su poder adquisitivo es mayor. Ellos simplemente van a disfrutar una experiencia sensorial.