1. Introducción

El mercado inmobiliario es un sector dinámico y complejo donde múltiples factores influyen en los precios de las propiedades. En contextos urbanos como Nueva York, la ubicación geográfica, las características físicas de las propiedades y los contextos socioeconómicos locales desempeñan roles cruciales. Sin embargo, los modelos predictivos tradicionales tienden a ignorar la espacialidad intrínseca de los datos (entendida como que los valores observados en los datos están influenciados no solo por las variables del modelo, sino también por la posición espacial en la que se encuentran) [1][2], lo que puede limitar su capacidad para capturar patrones locales. En este trabajo, se propone mejorar un modelo de regresión lineal propuesto dentro del foro Kaggle sobre el dataset de New York Housing Market mediante la incorporación de nuevas variables demográficas y de localización basado en el modelo hedónico, complementado con un enfoque de Regresión Geográficamente Ponderada, esto por que se plantea como problema que los modelos propuesto en el foro al no considerar la dependencia espacial intrínseca no logran predecir y explicar correctamente la plusvalía. Por lo tanto, como objetivo general se busca evaluar la mejora en la capacidad predictiva de un modelo hedónico mediante la incorporación de otras dimensiones y el uso de técnicas de modelación espacial con el planteamiento de la siguiente hipótesis:

- **Hipótesis nula (Ho):**

La capacidad predictiva de un modelo hedónico vista desde el R2, AIC Y BIC no mejora al incorporar la espacialidad intrínseca de los datos mediante un modelo geográficamente ponderado.

- **Hipótesis alternativa (Ha):**

La capacidad predictiva de un modelo hedónico vista desde el R2, AIC Y BIC mejora al incorporar la espacialidad intrínseca de los datos mediante un modelo geográficamente ponderado.

2. Marco Teórico

2.1. Modelo hedónico

Es una técnica econométrica utilizada para explicar el precio de un bien en función de sus características. Este modelo parte del supuesto de que el precio de un bien compuesto puede descomponerse en el valor de sus atributos individuales. En el contexto inmobiliario, el modelo hedónico [3] permite descomponer el precio de una propiedad en función de distintas dimensiones bajo un enfoque multidimensional considerando no sólo características propias de un bien sino que también de su entorno. Las dimensiones se pueden resumir en:

- **Infraestructura:** Incluye atributos físicos de la vivienda como superficie total, superficie construida, cantidad de habitaciones y baños, materiales de construcción, etc. Estas características influyen directamente en la percepción de calidad y funcionalidad del bien.

- **Geodemografía:** Considera factores poblacionales y socioeconómicos en relación a una georreferenciación, como densidad de población, nivel de ingresos promedio, tasa de actividad laboral, etc. Estas características aportan información sobre el contexto social y económico que puede aumentar o disminuir el valor percibido de un bien.
- **Localización:** Incluye la proximidad a servicios especializados como educación, salud, recreación, etc; y exposición factores medioambientales como ruido, emisiones, etc. Estas características afectan directamente la conveniencia y atractivo de un bien.

Matemáticamente el modelo hedónico se representa como:

$$P = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \text{ ; (1)}$$

Donde:

- P: Precio del bien.
- $X_1, X_2, \ldots, X_n$: Conjunto de características del bien.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes que reflejan el impacto de cada característica en el precio.
- ε: Error.

2.2. Econometría y autocorrelación espacial

Es un campo especializado de la econometría que se enfoca en analizar y modelar datos que presentan dependencia espacial. Esto significa que las observaciones no son independientes entre sí, sino que están correlacionadas con base en su proximidad geográfica. De esta manera los modelos econométricos espaciales abordan problemas como la autocorrelación espacial.

Una manera es a través del Índice de Moran [4][5] que describe la tendencia de los valores de una variable a ser similares o diferentes en función de su localización geográfica, para ello cuantifica la intensidad y dirección de la autocorrelación en un conjunto de datos según indica la siguiente ecuación:

$$I = \frac{N}{\Sigma_i \Sigma_j w_{ij}} * \frac{\Sigma_i \Sigma_j w_{ij}(x_i - \underline{x})(x_j - \underline{x})}{\Sigma_i (x_i - \underline{x})^2} \text{ ; (2)}$$

Donde:

- I: Índice de Moran.
- N: Número total de observaciones.
- w_{ij}: matriz de pesos espaciales (proximidad entre observaciones).
- x_i: Valor de la variable para la observación i.
- $\underline{x}$: Media de la variable.

Los valores del índice van entre -1 y +1:

- Valores cercanos a +1 indican autocorrelación positiva (valores similares de agrupan).
- Valores cercanos a -1 reflejan autocorrelación negativa (valores disímiles se agrupan).

- Valores cercanos a 0 indican ausencia de autocorrelación.

2.3. Regresión Geográficamente Ponderada (GWR)

Es una técnica de modelado espacial que permite capturar la variación local en las relaciones entre las variables dependientes e independientes. A diferencia de los modelos globales, GWR [6][7] ajusta una ecuación de regresión para cada punto en el espacio, ponderando las observaciones cercanas de acuerdo con su distancia, para ello calcula un ancho de banda que define el radio de influencia espacial alrededor de cada punto geográfico. Específicamente, controla cómo las observaciones vecinas afectan la estimación de los coeficientes locales. Un ancho de banda grande genera modelos más globales, ya que incluye más observaciones en el cálculo de los coeficientes locales. Un ancho de banda pequeño genera modelos más locales, ya que utiliza un número limitado de observaciones cercanas.

La forma general del GWR es:

$$y_i = \beta_0(u_i, v_i) + \Sigma_{k=1}^{K} \beta_k(u_i, v_i) x_{ki} + \varepsilon_i \text{ ; (3)}$$

Donde:

- (u_i, v_i): Coordenadas espaciales en el punto i.
- $\beta_k(u_i, v_i)$: Coeficientes específicos para el punto i, que varían en el espacio.
- x_{ki}: Valores de las variables explicativas para i.

3. Metodología

3.1. Carga y Análisis Preliminar de Datos

La carga de datos se realiza conectando directamente a Kaggle, en específico el dataset de New York Housing Market, este incluye en 4800 filas atributos como precio (PRICE), tamaño de la propiedad (PROPERTYSQFT), número de habitaciones (BEDS), número de baños (BATH), tipo de vivienda (TYPE), par de coordenadas (LAT, LON), entre otros. En esta etapa, se hace un análisis exploratorio inicial para entender la estructura de los datos y detectar valores inusuales o inconsistencias. Se generan estadísticas descriptivas y gráficos de distribución para identificar posibles valores extremos. Entre los patrones observados, se encontraron registros con un número desproporcionadamente alto de baños y habitaciones, que no corresponden a viviendas típicas, sino a edificios o condominios multifamiliares. Estas observaciones fueron marcadas para su eliminación en etapas posteriores. Además, se verificaron valores faltantes y se analizaron posibles correlaciones entre variables principales.

3.2. Enriquecimiento de Datos

Para mejorar el modelo predictivo bajo lo que plantea teóricamente el modelo hedónico se incorporan dimensiones adicionales basadas en datos geográficos y socioeconómicos. En primer lugar, se transforma el conjunto de datos en una capa de entidad geográfica, utilizando las coordenadas de latitud y longitud para crear puntos espaciales. Esto permite realizar operaciones espaciales y añadir información externa relevante. Los puntos fueron configurados bajo el sistema de referencia espacial métrico (EPSG:3857).

Como base, el dataset trae características de la dimensión de infraestructura, para integrar la dimensión geodemográfica se considera la American Community Survey (ACS) 2020 a través de la API del Censo de EE.UU [8]. Las variables a considerar son la población total (total_population), el ingreso promedio (median_income), y la distribución etaria, clasificando a la población en grupos económicamente activos e inactivos para obtener una tasa de población activa (tpa). La integración de esta información se realiza mediante una unión espacial, considerando un radio de 500 metros alrededor de cada vivienda para capturar las características del vecindario inmediato.

En cuanto a la dimensión de localización, mediante la API de Open Street Map [9] se calcula la distancia euclidiana desde cada vivienda hasta puntos de interés clave como Central Park (d_cp) y Times Square (d_ts), añadiendo las variables de distancia.

3.3. Limpieza de Datos

Tras el análisis exploratorio inicial y el enriquecimiento, se realiza una limpieza que consiste en la eliminación de registros con valores atípicos en las variables de número de baños (mayores a 4) y habitaciones (mayores a 7) para garantizar que los datos representen viviendas unifamiliares y no edificios. Así mismo se realiza un ajuste de datos por rango intercuartílico utilizando el Método Winsor solamente a la dimensión de infraestructura.

3.4. Transformación de Variables

Dado que el modelo predictivo constituye a una regresión lineal con variable objetivo precio, bajo los supuestos de un modelo de este tipo, se evalúan diversas transformaciones para que su distribución sea lo más cercana una normal, por lo tanto se prueban transformaciones del tipo logaritmo, raíz cuadrada, recíproco y cuadrado. Mediante análisis por distribución por histograma y gráficos QQ se selecciona la mejor transformación [10].
Las variables explicativas numéricas también son transformadas para mejorar su correlación con la variable objetivo. Y para evitar sesgos por diferencias en magnitud entre variables, todas son escaladas mediante el método de Min-Max Scaling [11], que normaliza los valores en un rango de 0 a 1. Este paso asegura que las variables con mayores magnitudes no dominen los cálculos del modelo.

En cuanto a las variables categóricas, se realiza una agrupación para mejorar su representatividad. Las sublocalidades (SUBLOCALITY) se agrupan en localidades mayores (LOC) para reducir la granularidad y aumentar la robustez estadística de las categorías. Además, los tipos de vivienda son clasificados en cuatro grupos principales: (i) Vivienda Unifamiliar que representa propiedades diseñadas para uso exclusivo de una familia, como casas o casas móviles; (ii) Vivienda Multifamiliar como propiedades con múltiples unidades habitacionales, como dúplex y casas adosadas; (iii) Propiedades en Condominio o Cooperativa como viviendas bajo asociaciones como condominios o cooperativas; (iv) y otros como terrenos y propiedades atípicas o que no calzan con otra categoría. Posteriormente, estas categorías fueron codificadas mediante un One-Hot Encoder (dejando una fuera), generando variables binarias para cada tipo de vivienda. Este enfoque permite que el modelo capture de manera efectiva las diferencias entre categorías sin introducir un orden implícito.

3.5. Base Maestra y Modelos Probados

El resultado de las etapas anteriores permite conformar una base maestra que considera todas las variables iniciales más las nuevas obtenidas ya sea desde el enriquecimiento, limpieza y transformación. Con esta nueva base es posible la prueba de diferentes modelos de regresión lineal: (i) Modelo original de Kaggle; (ii) Modelo original con tratamiento de datos; (ii) Modelo hedónico.

3.6. Análisis de Residuos y Autocorrelación Espacial

Para todos los modelos e independiente de sus resultados se realiza un análisis de residuos de cada modelo para evaluar su distribución (detectar problemas de heterocedasticidad) y existencia de patrones espaciales no capturados. Si bien se asume que la la variable objetivo de precio de la vivienda presenta autocorrelación espacial, se espera que si un modelo no es capaz de capturar la espacialidad intrínseca esta se ve reflejada en sus residuos (si no es capturada se ve un patrón y el índice de moran es positivo, caso contrario los residuos son de carácter aleatorio en el espacio).

3.7. Modelo GWR

En caso de existir dependencia espacial en los residuos, se implementa un modelo de Regresión Geográficamente Ponderada (GWR) considerando la selección de un ancho de banda, y calibración del modelo que permite el cálculo de coeficientes locales para cada observación, permitiendo analizar variaciones espaciales en las relaciones entre las variables explicativas y el precio además de buscar una mejora en la capacidad predictiva. Si esto ocurre se dice que la

incorporación de la espacialidad intrínseca mejora un modelo, siendo reflejado también en la distribución aleatoria de sus residuos.

3.8. Métricas de análisis

Independiente del tipo del modelo, con el fin de comparar se utilizarán las métricas de (i) coeficiente de determinación R2, que mide la proporción de la variabilidad en la variable dependiente que es explicada por las variables independientes del modelo; (ii) Criterio de Información de Akaike (AIC) como media que busca evaluar el modelo estadístico considerando tanto la bondad de ajuste como el número de parámetros que utiliza el modelo para evitar el sobreajuste; (iii) Criterio de Información Bayesiano (BIC) que mide la bondad de ajuste del modelo con una penalización por el número de parámetros utilizados, pero introduce un término que crece más rápido con el número de parámetros, lo que lo hace más estricto contra modelos con mayor número de parámetros.

3.9. Alcance

Cabe destacar que el proceso metodológico empleado no incluye la partición de datos ni la validación cruzada, ya que se trata de un análisis exploratorio. El objetivo principal es determinar si la inclusión de la espacialidad intrínseca mejora la capacidad predictiva del modelo. Por lo tanto, la relevancia de esta mejora se evalúa directamente sobre el conjunto completo de datos. Si se observan mejoras significativas, la partición de datos y la validación cruzada podrían considerarse como etapas adicionales en futuras metodologías.

4. Resultados y análisis

4.1. Enriquecimiento de datos

Como integración de las dimensiones geodemográfica se consideró mantener ingreso medio (mi_avg) y tasa de población activa medida como la razón entre población económicamente activa e inactiva. Y localización como la distancia desde cada punto observado hacia Central Park y Time Square.

Figura 1: Representación de censo 2020 a nivel de manzana (block). Fuente: Census Gov 2020.

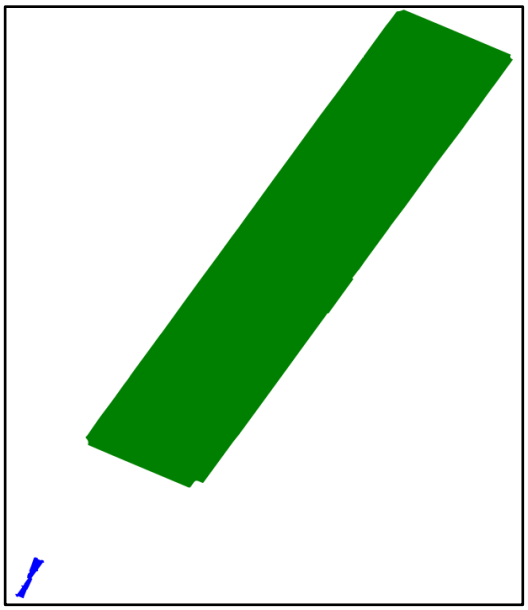

Figura 2: Representación de Central Park y Time Square a nivel poligonal. Fuente: OSM.

4.2. Análisis preliminar

El conjunto de datos numéricos y categóricos se resume en las Tablas 1 y 2. Entre las variables numéricas, destacan las distancias euclidianas a puntos clave como Central Park (d_cp) y Times Square (d_ts), con valores que oscilan entre ~19 metros y ~49 km. Esto sugiere que las propiedades están distribuidas tanto en áreas céntricas como en zonas periféricas. La media de estas distancias (14.9 km y 16.1 km, respectivamente) indica que muchas viviendas están relativamente alejadas de estos puntos de interés. En cuanto a las variables demográficas, el ingreso promedio por área (mi_avg) presenta un rango amplio, desde $17,305 hasta $250,001, reflejando disparidades económicas significativas entre las zonas. La tasa de población activa (tpa) también muestra variaciones notables, con valores que van desde 0 hasta más de 5,000 por unidad geográfica, sugiriendo la necesidad de ajustes para manejar esta heterogeneidad. Las características físicas de las propiedades, como su tamaño (PROPERTYSQFT), presentan valores extremos que van desde 230 ft² hasta más de 65,000 ft². Esto incluye tanto viviendas unifamiliares como propiedades comerciales o terrenos. Por otro lado, el número de habitaciones (BEDS) y baños (BATH) muestra valores máximos de 50, lo cual sugiere la presencia de edificios que posteriormente fueron eliminados del dataset.

En cuanto a la variable objetivo, el precio (PRICE), presenta una distribución altamente sesgada, con valores que van desde $2,494 hasta más de $2.1 mil millones. La mediana de $825,000 indica que la mayoría de las propiedades tienen precios moderados, pero los valores atípicos elevan significativamente la media.

Entre las variables categóricas, el tipo de vivienda (TYPE) destaca como una característica clave, con 13 categorías únicas. Las más comunes son "Co-op for sale" (30.2%) y "Condo for sale" (9.5%). Estas categorías, al ser heterogéneas, pueden requerir una reagrupación para simplificar el análisis. Las variables geográficas, como LOCALITY y SUBLOCALITY, ofrecen niveles de granularidad diferentes, con LOCALITY concentrando el 52.2% de las observaciones en "New York", mientras que SUBLOCALITY aporta mayor detalle con 21 valores únicos.

Tabla 1: Resumen de variables numéricas.

Variable	Media	Mínimo	1er Cuartil	Mediana	3er Cuartil	Máximo
d_cp	14927.3	19.3	6253.7	14386.7	21824.4	49098.8
d_ts	16065.8	123.4	7542.1	16895.2	22420.2	47428.8
mi_avg	96442.5	17305.1	65794.1	86513.6	119434.4	250001.0
tpa	279.4	0.0	155.9	204.0	292.3	5314.6
PROPERTYSQFT	2184.2	230.0	1200.0	2184.2	2184.2	65535.0
PRICE	2356940.2	2494.0	499000.0	825000.0	1495000.0	2147483647.0
BEDS	3.4	1.0	2.0	3.0	4.0	50.0
BATH	2.4	0.0	1.0	2.0	3.0	50.0

Tabla 2: Resumen de variables categóricas.

Variable	Valores Únicos	Valor Modal	Frecuencia Modal (%)
SUBLOCALITY	21	New York	21.2
LOCALITY	11	New York	52.2
TYPE	13	Co-op for sale	30.2

4.3. Datos post procesamiento y transformación

La mejor transformación para la variable objetivo es del tipo logaritmo según se muestra en la Figura 3.

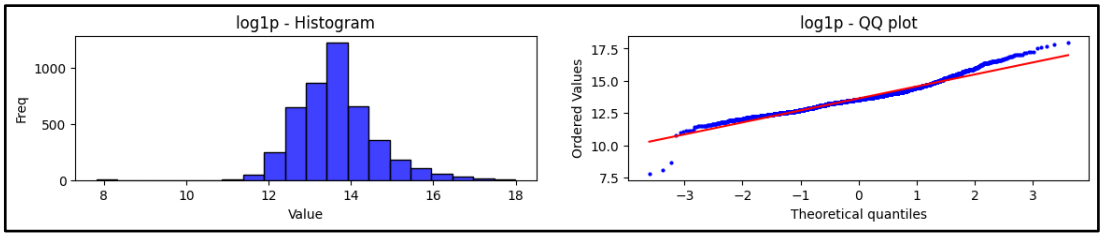

Figura 3: Distribución de variable objetivo de precio de vivienda (PRICE) tras transformación logarítmica.

La Figura 4 representa las distribuciones de las variables después de aplicar las transformaciones para mejorar su correlación con la variable objetivo. Las variables PROPERTYSQFT, BEDS, pi_avg, mi_avg y d_ts se mantuvieron idénticas, mientras que en BATH, pa_avg, d_cp se aplicó una transformación logarítmica (log1p).

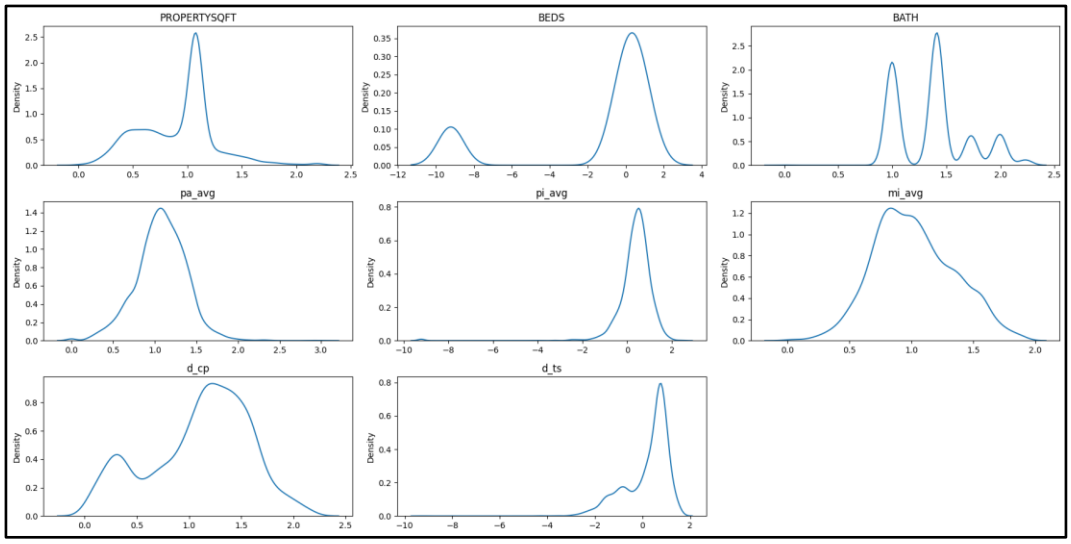

Figura 4: Distribución de variables explicativas tras procesamiento y transformaciones.

En la Figura 5 se observa que espacialmente la variable objetivo presenta claramente un patrón espacial alcanzando sus valores máximos en el sector norponiente (cercano a puntos de localización de atracción); lo mismo ocurre con el nivel de ingresos. En cuanto a la población económicamente activa e inactiva se ve que existe una tendencia a un recíproco, es decir, donde hay predominancia de población activa, existe carencia de inactiva, a excepción de los sectores centro y sur poniente, esto sugiere que la tasa población activa presentará valores mayores en el sector centro nor poniente. Las distancias muestran que radialmente aumentan en la medida que se alejan de los puntos de atracción; y por último la dimensión de infraestructura es más bien homogénea o aleatoria espacialmente pero con predominancia en valores menores.

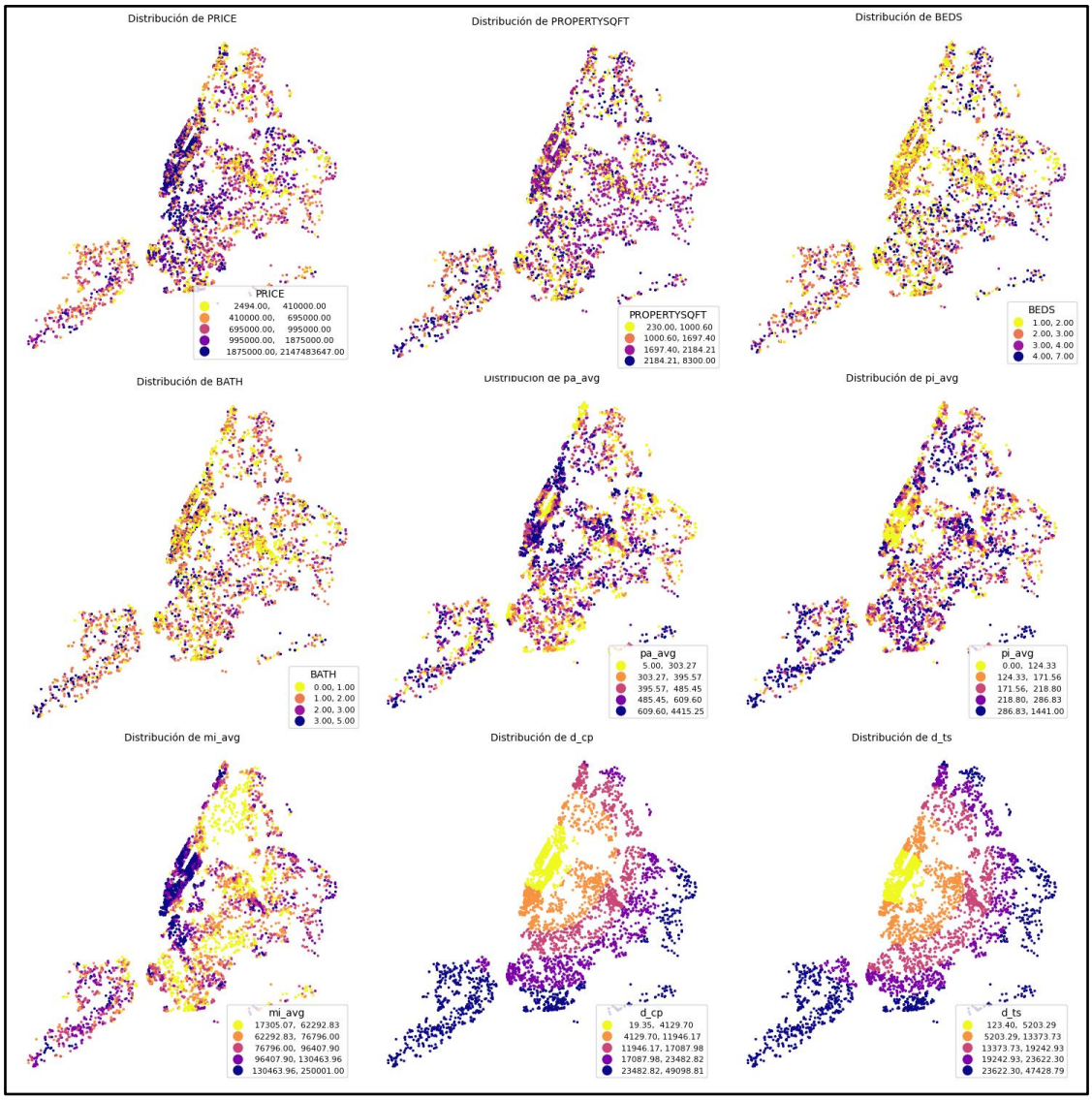

Figura 5: Distribución espacial de variables.

La Figura 6 muestra las distribuciones de las variables después de aplicar el escalado. Este proceso asegura que las variables, independientemente de sus magnitudes originales, sean comparables entre sí, lo cual es esencial en modelos sensibles a escalas, como la regresión lineal y los análisis espaciales. La variable PROPERTYSQFT mantiene un máximo cercano a 0.5, reflejando que la mayoría de las propiedades tienen tamaños medios. BEDS y BATH, muestran valores máximos en valores bajos, lo que representa la predominancia de propiedades más pequeñas en el conjunto de datos. CAT ORDINAL es un campo formado por el ordenamiento del tipo de vivienda, que posteriormente no fue utilizado. La población activa (pa_avg) e inactiva (pi_avg), presentan distribuciones ajustadas y compactas tras el escalado. Por su parte, mi_avg, muestra una distribución mayormente uniforme, pero con máximo cercano a 0.4. Las distancias a Central Park (d_cp) y Times Square (d_ts), muestra que la mayoría de las observaciones están lejanas a estos puntos de atracción.

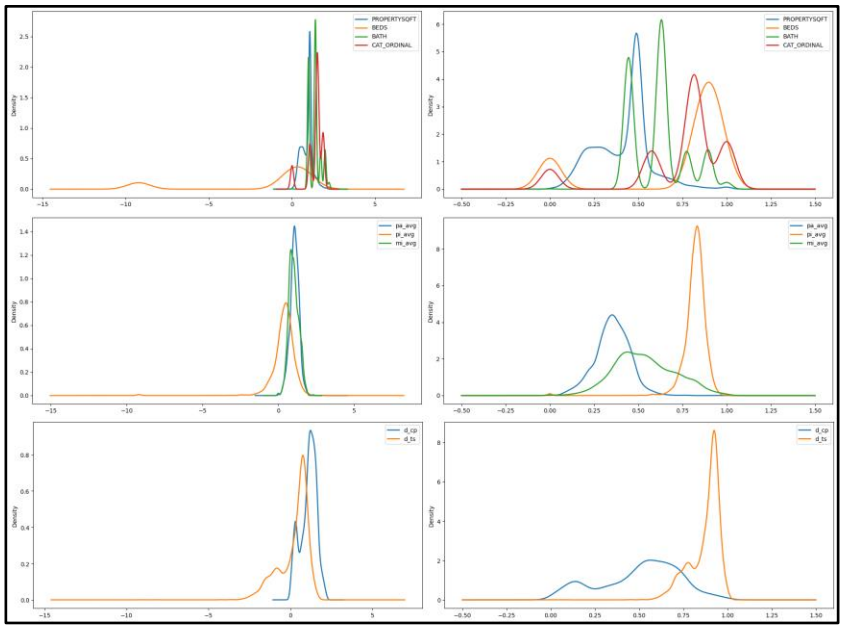

Figura 6: Distribución de variables explicativas tras proceso de escalado por método Min-Max.

Las variables categóricas quedaron de la siguiente manera:

La agrupación de SUBLOCALITY en cinco grandes áreas geográficas hacia un campo LOC, reduce la granularidad inicial de 21 sub localidades a una estructura más manejable. Manhattan (NY), Queens (QC) y Brooklyn (KC) concentran la mayor cantidad de propiedades, representando el 79% del total. Staten Island (RC) y The Bronx (BC) tienen menor representación, lo cual es consistente con su densidad habitacional más baja.

Tabla 3: Resumen de agrupación para conformación de campo LOC.

Categoría	**Descripción**	**Total de Registros**
NY	Manhattan, conocida por ser el centro financiero y cultural de Nueva York.	1204
QC	Queens, un área diversa en términos culturales y demográficos.	1191
KC	Brooklyn, popular por su estilo de vida moderno y propiedades residenciales.	1083
RC	Staten Island, caracterizada por áreas residenciales y menor densidad poblacional.	522
BC	The Bronx, una mezcla de áreas urbanas y residenciales.	449

La reorganización de los tipos de vivienda en un campo CAT (CC, VU, VM, Otros) reduce la complejidad inicial de 13 valores únicos, facilitando la interpretación. La categoría dominante es CC (Condominios y cooperativas) con más de la mitad de los registros (52%), seguida de las viviendas unifamiliares (VU) con 21% y multifamiliares (VM) con 17%. Las propiedades en la categoría "Otros" representan el 9% y son las menos comunes.

Tabla 4: Resumen de agrupación para conformación de campo CAT.

Categoría	Descripción	Total de Registros
CC	Condominios y cooperativas, propiedades bajo asociaciones (e.g., condo, co-op).	2316
VU	Viviendas unifamiliares, diseñadas para uso exclusivo de una familia.	960
VM	Viviendas multifamiliares, como dúplex y casas adosadas.	772
Otros	Terrenos y propiedades atípicas o especiales (e.g., terrenos en venta).	401

4.4. Modelos de regresión

La Tabla 5 muestra los modelos de regresión lineal probados, el primero corresponde a la fórmula original donde se considera el detalle de cada localidad y las coordenadas de forma independiente, este presenta la peor capacidad predictiva con bajo valor R2 de 0.017 y valores de AIC y BIC mayor (179300 y 179500). Como mejora de este modelo se propone el reemplazo de la localidad explicada (campo LOC) mejorando así su capacidad predictiva (y reducción de error) debido al aumento de muestra de datos por cada categoría reflejado en su AIC a 8939 y del BIC a 8997. Finalmente el modelo hedónico propuesto considera agregar la cantidad de baños, y tipo de vivienda como dimensión de infraestructura, tasa de población activa e ingresos como dimensión geodemográfica y distancia Central Park y Time Square como dimensión de localización. Las coordenadas y localidad se omiten dado que posteriormente esta información será considerada al capturar la espacialidad intrínseca. Este modelo mejora logrando explicar el 64.1% de la variabilidad en los precios de las propiedades. Sus valores de AIC (7597) y BIC (7667), son los más bajos entre los modelos evaluados. Estos indicadores confirman que, pese a la complejidad adicional del modelo, el ajuste mejora sin incurrir en penalizaciones excesivas, y que si bien la mayor parte del modelo se podría explicar por características como la superficie, el agregar otras dimensiones permiten una mejor explicación de la plusvalía de una vivienda. De igual Manera se debe tener en consideración que el Condition Number de todos los modelos es muy alto (17300) indicando que existen problemas de multicolinealidad, afectando a la inflación y sesgo de la capacidad predictiva, como mejora se podría considerar un análisis más detallado para mejorar esto, ya sea eliminando variables o transformarlas (por ejemplo, agrupándolas, análisis de componentes principales, regularización, etc.).

Tabla 5: Resumen de modelos de regresión lineal entrenados.

Modelo	**Fórmula**	**R2**	**Ajuste General (AIC/BIC)**
Modelo Original	PRICE ~ TYPE + BEDS + PROPERTYSQFT + LATITUDE + LONGITUDE + LOCALITY	0.017	AIC: 179300, BIC: 179500
Modelo Mejorado (LOC)	PRICE ~ LOC + BEDS + PROPERTYSQFT + LATITUDE + LONGITUDE	0.514	AIC: 8939, BIC: 8997
Modelo Propuesto	PRICE ~ BATH + BEDS + PROPERTYSQFT + tpa + mi_avg + d_ts + d_cp + CAT_VU + CAT_VM + CAT_Otros	0.641	AIC: 7597, BIC: 7667

Por su parte la Tabla 6 muestra el detalle de los coeficientes estimados para el modelo propuesto. El intercepto, con un coeficiente de 13.499 representa el valor esperado del precio cuando todas las variables independientes son iguales a cero (este podría interpretarse como valor base una vivienda o bien sugiere que aún faltan otras variables que expliquen el precio como costos de construcción u otra información como medioambiental u otros servicios especializados). Entre las variables físicas, vemos que todas son estadísticamente significativas y de signo positivo indicando que por cada unidad adicional, se ve aumentada la plusvalía del bien; las viviendas de tipo multifamiliar aumentan más el precio, no así la categoría de otros que si bien su signo es negativo, es una variable no significativa. Para la dimensión geodemográfica se ve que los niveles de ingreso son significativos, no así la tasa de población activa, sin embargo se seguirá manteniendo en el modelo; ambos coeficientes son de signo positivo lo que sugiere que en la medida que en un barrio reside población con mayor ingresos y o mayor fuerza de trabajo aumenta la plusvalía. En cuanto a la dimensión de localización ambas son significativas y de signo negativo lo que significa que a mayor distancia de los puntos de atracción, los precios tienden a disminuir, esto confirma que la proximidad a áreas de interés central tiene un efecto positivo en los valores inmobiliarios.

En resumen, los signos de los coeficientes reflejan relaciones coherentes con la variable objetivo (PRICE). Las variables como BATH, BEDS, PROPERTYSQFT, mi_avg, y la proximidad a Times Square tienen un impacto positivo significativo en los precios. Este análisis refuerza la importancia de incorporar múltiples dimensiones en el modelo para capturar los factores determinantes del valor de las propiedades.

Tabla 6: Resumen de coeficientes del modelo de regresión propuesto.

Coefficient	**Value**	**Std Err**	**T-Value**	**P-Value**
Intercept	13.499	0.224	60.272	0.000
BATH	2.472	0.076	32.737	0.000
BEDS	0.275	0.030	9.266	0.000
PROPERTYSQFT	1.263	0.067	18.992	0.000

tpa	7.87E-04	3.46e-05	0.023	0.982
mi_avg	1.262	0.061	20.779	0.000
d_ts	-3.182	0.271	-11.734	0.000
d_cp	-0.142	0.090	-1.579	0.114
CAT_VM	0.222	0.030	7.388	0.000
CAT_VU	0.161	0.027	5.974	0.000
CAT_otros	-0.052	0.033	-1.593	0.111

4.5. Análisis de autocorrelación

Si bien el modelo propuesto ya mejora los resultados, es necesario analizar si existe dependencia especial de la variable objetivo y los residuos. La Figura 7 sugiere que el modelo es robusto en cuanto a que no presenta problemas de heterocedasticidad, sin embargo, según la Figura 8 no es robusto en cuanto a la captura de la espacialidad intrínseca, es decir, es capaz de predecir precios, pero es propenso a tener mayor error para ciertos sectores en específico (según muestra la concentración de valores altos y bajos en el mapa temático). El índice de moran sobre indica tanto para la variable objetivo y residuos (Figura 9 y 10 respectivamente con 50 vecinos cercanos) que existe autocorrelación espacial positiva, por lo tanto se justifica la utilización de un modelo espacial, lo cual podría mejorar los resultados.

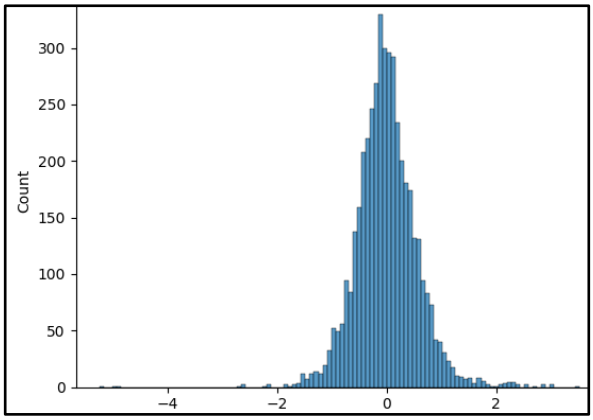

Figura 7: Distribución de residuos del modelo propuesto.

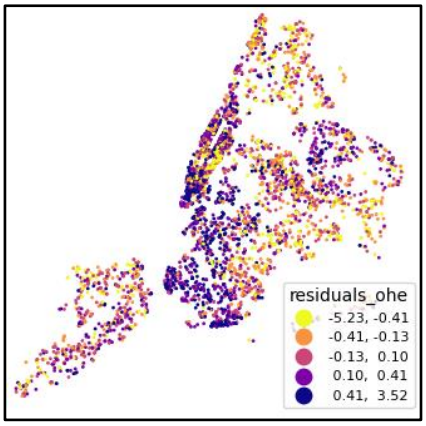

Figura 8: Distribución espacial de residuos del modelo propuesto.

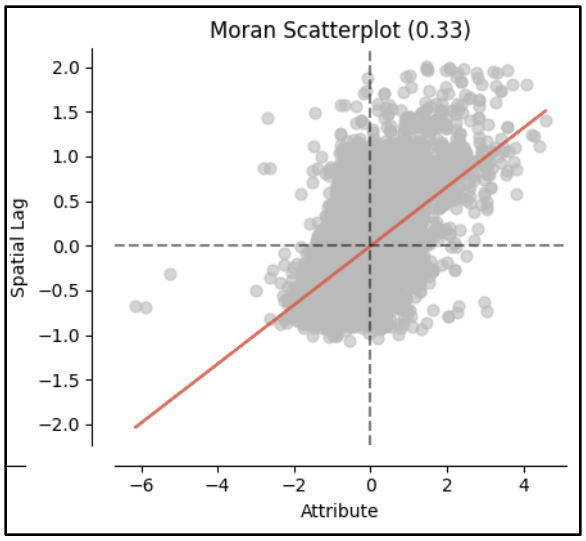

Figura 9: Índice de Moran sobre variable objetivo.

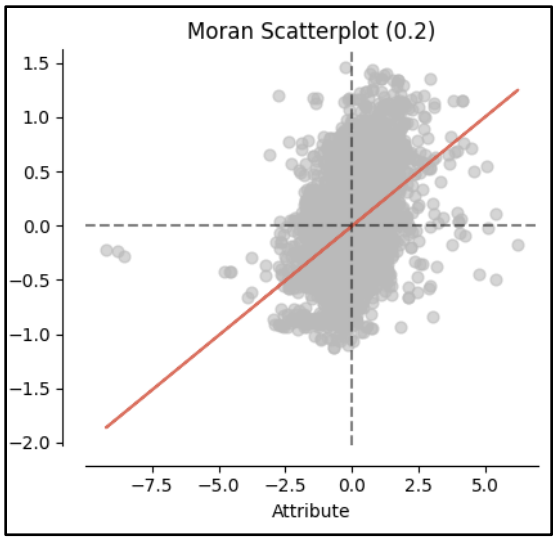

Figura 10: Índice de Moran sobre residuos del modelo.

4.6. Modelo GWR

El modelo GWR mejora los resultados del modelo propuesto al capturar la espacialidad intrínseca variabilidad en las relaciones entre las variables explicativas y la variable objetivo (PRICE). Esto se refleja en un incremento del R2 de 0.641 (modelo global) a 0.746, lo que indica que el modelo captura el 74.6% de la variabilidad de los precios de las propiedades. Esta mejora se atribuye a la capacidad del GWR para ajustar coeficientes localmente dado un ancho de banda de 1133 (en general estima que para cada valor observado requiere de dicha cantidad de vecinos cercanos, el aumentar la variabilidad espacial de los dato, es decir, mayor datos podría disminuir este ancho de banda), considerando las dependencias espaciales intrínsecas de los datos. Así mismo, los valores ajustados de AIC (6213.612) y BIC (6807.091) son menores a los del modelo global.

En cuanto a los coeficientes, las variables BATH, BEDS, y mi_avg mantienen un signo positivo en la mayoría de las áreas, indicando su efecto positivo general en la plusvalía. Sin embargo, otras variables, como PROPERTYSQFT, d_cp, y d_ts, presentan signos que varían entre positivo y negativo en diferentes zonas. Esto refleja que el impacto de estas variables depende del contexto local. Por ejemplo, estar más lejos de un punto céntrico como Times Square podría estar relacionado con precios más altos en zonas periféricas más exclusivas.

La variabilidad de los coeficientes de tipo de vivienda (CAT_VM, CAT_VU, y CAT_Otros) también destaca la influencia del contexto local en la valoración de las propiedades. Por ejemplo, en ciertas áreas, las viviendas multifamiliares (CAT_VM) tienen un impacto positivo mayor, mientras que en otras su contribución es menor o incluso negativa.

En general, el GWR permite identificar patrones locales que el modelo global no puede capturar, ofreciendo una interpretación más rica y detallada de los datos. Esto es especialmente valioso

en contextos como el mercado inmobiliario, donde las dinámicas espaciales son complejas y heterogéneas. Los ajustes de AIC (6213.612) y BIC (6807.091) inferiores a los del modelo global refuerzan la mejora en términos de ajuste y eficiencia estadística.

Tabla 7: Resumen de coeficientes locales del modelo GWR.

R2	0.746						
AIC (6213.612) **- BIC** (6807.091)							
Variable	**Media**	**STD**	**Mínimo**	**Mediana**	**Máximo**	**Signo Global**	**Signo GWR**
Intercepto	22.629	23.521	-10.667	12.452	102.780	+	-/+
BEDS	0.230	0.074	0.024	0.230	0.419	+	+
BATH	2.246	0.888	0.493	2.153	3.864	+	+
PROPERTYSQFT	1.147	0.625	-0.008	1.227	2.163	+	-/+
mi_avg	0.809	0.342	0.187	0.746	1.595	+	+
tpa	-0.000	0.000	-0.001	0.000	0.000	+	-
d_cp	3.002	9.830	-11.664	0.139	36.532	-	-/+
d_ts	-14.936	32.161	-125.203	-1.434	31.308	-	-/+
CAT_VM	0.408	0.212	-0.061	0.435	0.823	+	-/+
CAT_VU	0.332	0.166	-0.204	0.342	0.635	+	-/+
CAT_Otros	0.024	0.207	-0.314	-0.043	0.425	-	-/+

El análisis de autocorrelación espacial muestra que los resultados obtenidos por el modelo GWR logra capturar la espacialidad sobre los residuos según se muestra en la Figura 11, sin embargo se debe considerar que esto depende también de la cantidad de vecinos cercanos que se utilizan al momento de calcular el Índice de Moran, a mayor cantidad de vecinos se tiene mayor información del entorno, por lo tanto, se captura mejor la espacialidad. El propósito es ver si reduce o no su valor a un estado aleatorio; Sobre 950 vecinos ya se obtiene esto y considerando el mismo caso comparativo con 50 vecinos cercanos, ya se ve una mejora de 0.2 (modelo inicial) a 0.05. Esto sugiere que aún faltan variables que pudiesen explicar el entorno, por ejemplo, acceso a servicios especializados (salud, educación, transporte), así como mayor detalle de características de vivienda como material de construcción, superficie (útil, terreno).

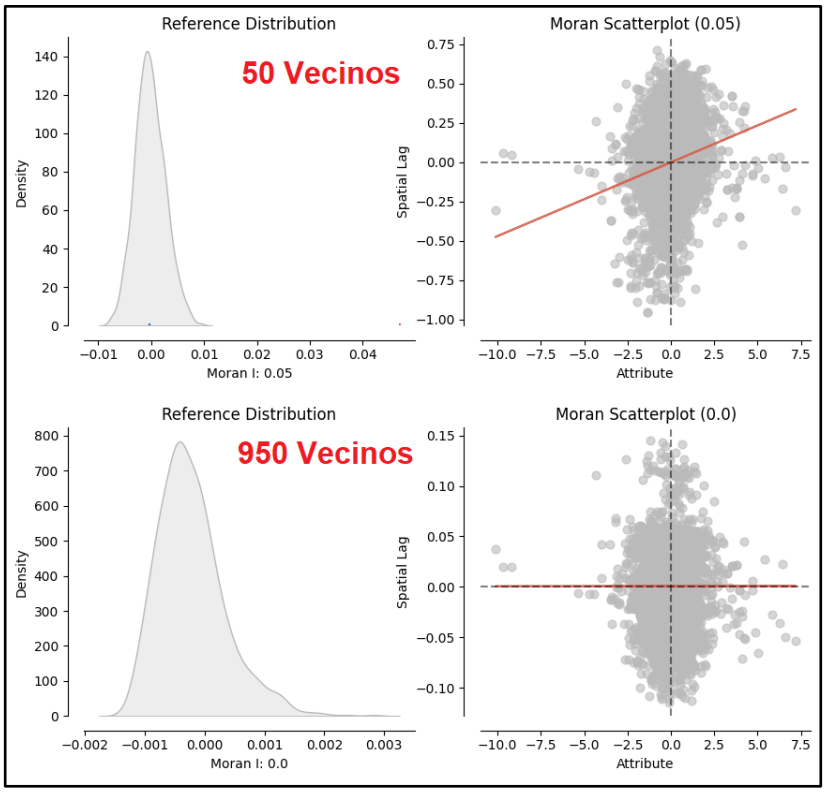

Figura 11: Índice de Moran estimado para residuos de GWR.

Por otra parte, la Figura 12 muestra la distribución de los coeficientes locales, de esta manera permiten explicar cómo cada variable explica la plusvalía según la localización. En general las variables de dimensión de infraestructura aumentan su valor en la medida que se encuentra cercano al Central Park y Time Square, zonas son consideradas las más exclusivas y demandadas, donde cada unidad adicional de lujo o espacio incrementa significativamente el valor percibido de una propiedad, debido al acceso a servicios premium, turismo, entretenimiento y espacios verdes [12][13]. Sin embargo, el impacto de las habitaciones es menor en estas áreas centrales, esto podría explicarse porque Manhattan, es una zona de alta densidad urbana donde las propiedades suelen ser más compactas. En este contexto, los compradores priorizan la ubicación por encima del tamaño, y las propiedades con más habitaciones tienden a estar fuera de estas áreas céntricas, donde el espacio es más asequible. Para las categorías de vivienda, las propiedades unifamiliares y multifamiliares muestran un mayor impacto positivo en los precios cerca de los principales atractivos. Esto se debe a su escasez en áreas densamente urbanizadas, lo que las convierte en bienes exclusivos y altamente demandados. Por el contrario, los terrenos y propiedades categorizadas como "Otros" presentan un impacto positivo en los precios a medida que se alejan. Esto puede explicarse porque, en zonas más alejadas, la disponibilidad de terrenos es mayor y su valor está asociado a su potencial de desarrollo, ya sea residencial o comercial [12][14]. En cuanto a la dimensión geodemográfica el ingreso promedio por área tiene un impacto positivo hacia el sur de Times Square y al noreste de Central Park, hacia noreste de Central Park existen zonas como el Upper East Side, reconocidas por su exclusividad, altos ingresos y acceso privilegiado a servicios de lujo y atractivos culturales como museos. Por otro lado, hacia el sur de Times Square, sectores como Chelsea y Greenwich Village combinan altos

ingresos con un entorno cultural y de entretenimiento dinámico, lo que eleva el valor de las propiedades [10]. Y para la dimensión de localización, los coeficientes relacionados con la distancia a Central Park alcanzan sus máximos hacia el sur y suroeste de Times Square, la demanda inmobiliaria está más influenciada por la vida urbana dinámica y el acceso a zonas comerciales y de entretenimiento como Broadway y Midtown [12][14]. Por otro lado, la distancia a Times Square presenta sus máximos en la misma zona. Por último se observa que los residuos tienen una distribución aleatoria similar a la de un ruido y ya no se definen patrones claros.