# IMT2118: Análisis del precio de las viviendas en el condado de Manhattan durante Mayo del 2024 y Abril del 2025.

Matías V. Esteban C.

# Índice

# 1. Introducción

## 1.1. Contexto

El condado de Manhattan es conocido por ser la zona residencial más costosa de toda la ciudad de New York, contando con un aproximado de 8,6 millones de habitantes, y, según la página *tinsa* en 2013, también es conocido por ser una de las zonas de tiendas más caras y exclusivas de todo el mundo. Un ejemplo concreto de esto ocurrió con la empresa *Zara*, que gastó una suma total de 324 millones de dólares por un local de este distrito.

Las razones del precio de sus residencias son variadas, pero destaca por sobre todo en ser un lugar lujoso y turístico, y no es para menos, puesto que «alberga varias atracciones turísticas famosas, como el **Empire State Building**, el **Chrysler Building**, el **Rockefeller Center**, el **World Trade Center**, el **puente de Brooklyn** y **Central Park**».

## 1.2. Descripción del problema

Manhattan es uno de los lugares más caros para vivir, por ende, la meta de este proyecto consiste principalmente en poder facilitar la localización de los lugares potencialmente más efectivos para poder proliferar adecuadamente en esta ciudad, según los intereses de un potencial interesado en adquirir una vivienda. Se intenta responder a la pregunta **¿En qué zonas es más probable encontrar propiedades en venta que sean favorables para un futuro comprador, considerando factores económicos, índices delictivos, cercanía a servicios de transporte, de salud y educación, entre otros?**
Esto lo hacemos basándonos en ejemplos de ventas ocurridas entre Mayo del 2024 y Abril del 2025. La idea principal es hacer un análisis geoespacial de los datos de ventas ocurridas en ese periodo, para posteriormente realizar un modelo predictivo del precio estimado según ciertos parámetros.

**¿Qué consideramos para evaluar qué lugares son los más favorables?**

a. **Educación:** Para estudiantes, es importante localizar la distancia a la que se encuentran los centros educativos con respecto a su vivienda.

b. **Seguridad:** Se ha considerado la cantidad de tiroteos/incidentes cercanos a una vivienda como una métrica indicativa de la seguridad. Además, saber que viviendas disponen de menos riesgo de inundaciones provocadas por huracanes y lluvias es una consideración importante para la seguridad, por eso consideramos estos dos factores como relevantes en el análisis.

c. **Sanidad:** La distancia hacia los centros médicos/hospitales es muy importante, de eso puede llegar a depender la vida.

d. **Vegetación:** Identificar la ubicación de las áreas verdes cercanas es muy importante para la salud integral del ser humano. Ayuda a la calidad del aire y la regulación de las temperaturas elevadas.

f. **Precio:** Los compradores podrán visualizar fácilmente las ubicaciones de viviendas vendidas según su rango de precios.

# 2. Objetivos del proyecto

## 2.1. Objetivo general

El objetivo principal, como la pregunta del inicio lo plantea, es responder a la necesidad que tiene un potencial comprador de viviendas; saber qué precio de viviendas ronda cierto sector de la ciudad, qué tanta seguridad hay ahí, qué tan cercana está la propiedad de lugares de importancia como hospitales y universidades, entre otros factores. En resumen, el objetivo general es dar una visión amplia en todos estos factores, no sólo focalizarse en algún aspecto de ellos, para así ayudar a potenciales compradores de viviendas a tener una visión panorámica general de las ventas ocurridas en el periodo de estudio.

### 2.2. Objetivos específicos

Para los objetivos específicos, se muestran diferentes visualizaciones estáticas en todo el proyecto, que ayudan a ver la relevancia de los factores considerados. A raíz de esto, los objetivos específicos que disponemos corresponden a responder las siguientes preguntas claves para este trabajo:

a. **¿Cómo es la vegetación y el riesgo de inundación en la zona de Manhattan considerando que es el condado con mayor población de New York?**

b. **¿Cuál es la distancia mínima de las viviendas a servicios básicos de salud, educación y transporte, en Manhattan?**

c. **¿Cómo se distribuyen espacialmente las zonas que tuvieron viviendas vendidas en Manhattan? ¿Existen lugares más caros que el resto? ¿Cómo se ve afectado el precio de la vivienda dado los parámetros consdierados en el proyecto?**

## 3. Descripción de los datos

### 3.1. Datos vectoriales con sus enlaces específicos

a. **MTA Subway Entrances and Exits**: Este conjunto de datos contiene un total de 2121 registros y se trata de una lista de las entradas y salidas de las estaciones de metro de New York. No se aplicó un filtro para obtener únicamente los registros que estén dentro del AOI de Manhattan, porque tiene sentido buscar entradas cercanas aledañas a Manhattan. Este dataset está en este enlace.

b. **NYPD Shooting Incident Data (Historic)**: Este dataset cuenta con un total de 29.745 registros y es una lista de todos los tiroteos registrados ubicados en New York a partir del 2006. Se ha filtrado por la fecha de interés estudiada. Este dataset está en este enlace.

c. **Hospitales**: Este dataset cuenta con un total de 79 registros y es una lista de los centros médicos que estan situados en New York, tales como: centros médicos, hospitales, clínicas, entre otros. Este dataset está en este enlace.

d. **College University**: Este dataset cuenta con un total de 78 registros y es una lista de las ubicaciones de los sistemas de educación de New York. Este dataset está en este enlace.

e. **Property rolling sales data**: Este dataset cuenta con un total de 17700 registros (aproximadamente) y es una lista de todas las ventas historicas realizadas en New York. Se aplicó un filtro para obtener únicamente los registros que estén dentro del AOI de interés, y, además, se hizo uso de la API de Open Cage Data para poder buscar las ubicaciones exactas de esas propiedades que se han vendido (Geocodificación), pero se ha tomado una muestra pequeña menor a 2mil registros por limitaciones en el proceso. Este dataset está en este enlace.

f. **Shapefile USA**: Este dato es un archivo geoespacial que contiene representaciones digitales de elementos geográficos. Nosotros elegimos United States como nuestro país de referencia y descomprimiendo el .zip, eso nos dará los archivos necesarios para el proyecto. Luego cargamos el Shapefile de USA y filtramos el AOI en Manhattan para tener nuestra área de interés. Se puede encontrar este Shapefile en este enlace.

### 3.2. Datos raster

a. **Sentinel-2**: Según la página de Google Earth Engine, «es una misión de imágenes multiespectrales de alta resolución y ancho de franja que respalda los estudios de monitoreo terrestre de Copernicus, incluido el monitoreo de vegetación, la cubierta del suelo y el agua, así como la observación de las vías navegables interiores y las áreas costeras». Imágenes en este enlace.

b. **Sentinel-1**: Según la propia página de Google Earth Engine, «proporciona datos de un instrumento de radar de apertura sintética (SAR) de banda C y doble polarización a 5.405 GHz (banda C). Esta colección incluye las escenas de detección de rango terrestre (GRD) de S1, procesadas

con Sentinel-1 Toolbox para generar un producto calibrado y ortocorregido. La colección se actualiza todos los días. Los recursos nuevos se transfieren en un plazo de dos días después de que están disponibles.». Se pueden solicitar las imágenes en este enlace.

# 4. Desarrollo

## 4.1. Procesamientos

a. **ETL**: La práctica de ETL o «Extract, Transform, Load» fue necesario a la hora de realizar el código. Este proceso consiste en extraer los datos (principalmente de los enlaces mencionados en la sección de «Descripción de datos». Luego hay que transformarlos, esta etapa consiste en una limpieza principalmente, se eliminan las filas duplicadas, nos aseguramos de que cada fila tenga su tipo de variable adecuada y descartamos las filas que no nos interesan. Finalmente, el paso de cargar consiste en guardar estos datos en un nuevo dataset. Para más información, se puede visitar el siguiente enlace.

b. **Uniones y proyecciones**: Este procesamiento fue vital para poder encontrar las ubicaciones de las viviendas que se han vendido del dataset de `rollingsales_manhattan.xlsx`, se hicieron uniones de datos vectoriales y raster, proyecciones en diferentes CRS a los datasets. Las ubicaciones fueron dadas por la API de Open Cage Data.

c. **Filtramiento por condición y limpieza**: Este tipo de filtramiento se usó en prácticamente todos los datasets correspondiente a este proyecto, pues se tuvo que remover cualquier tipo de dato que no correspondiese a Manhattan, también se limpiaron datos no relevantes.

d. **Segmentación de hexágonos**: Este algoritmo se trata de una división de un área en celdas hexagonales iguales para facilitar el análisis espacial y la visualización de datos geográficos. Fue bastante efectivo para ver las distintas celdas con su precio aproximado, mostrando cuáles serían los lugares que cuestan mayor cantidad monetaria, entre otras representaciones de los parámetros en juego. Para la representación hexagonal, se ha utilizado la librería h3 de Uber.

## 4.2. Tipos de análisis

a. **Análisis espacial**: Puesto que se usan imágenes de Sentinels 1 para detectar áreas inundadas en el área geográfica de Manhattan y, además, se usan imágenes de Sentinels 2 para poder detectar diversos índices espectrales, tales como: el NDBI y NDVI.

b. **Análisis temporales**: En el caso de las inundaciones, realizamos la comparativa entre distintas fechas concretas donde hubieron anegamientos registrados (2023-2024).

c. **Análisis estadístico**: Se ha realizado el análisis de Moran para poder ver la autocorrelación de los datos relacionados a tiroteos en institutos educativos.

d. **Análisis predictivo**: Se ha tomado el último modelo de clasificación de la clase de Deep Learning para datos SAR, específicamente el modelo de clasificación **Random Forest**, pero aplicando el algoritmo a un modelo de predicción, tomando como base a los datos de ventas del dataset `gdf_properties` para el entrenamiento. Se puede ver un ejemplo de uso análogo del modelo aquí.

## 4.3. Modelos implementados

a. **DBSCAN**: El algoritmo de DBSCAN o «Density-Based Spatial Clustering of Applications with Noise» es un modelo de agrupamiento no supervizado, se ha utilizado para agrupar los tiroteos en instituciones educativas entre sí para poder observar la distancia a la que están cada una.

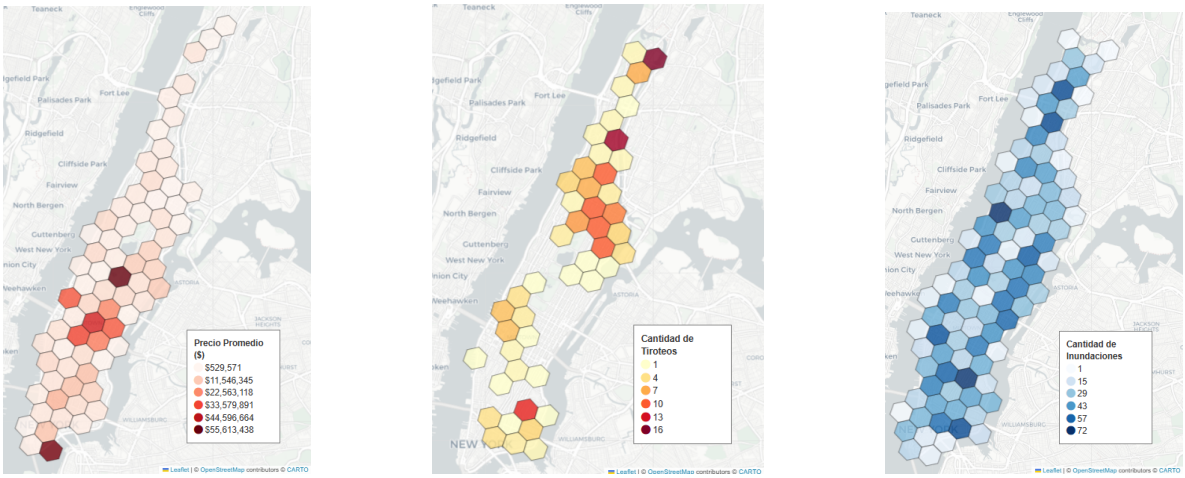
b. **Búsqueda de caminos de Dijkstra**: Este algoritmo, característico de la teoría de grafos, se encarga de determinar el camino más corto entre un nodo origen y un nodo destino dentro de una red, el cálculo se realiza todo de una vez.

c. **Random Forests**: Es un modelo de aprendizaje automático de tipo predictivo que combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste. Es totalmente ideal para utilizarlo en clasificación como en la regresión.

# 5. Análisis final

## 5.1. Resultados

Tenemos una comparación de los parámetros de precio, tiroteos en instituciones educativas, e inundaciones de las zonas agrupadas en hexágonos, lo que nos permite ver su distribución espacial en las visualizaciones. Posteriormente, en el modelo predictivo, esto nos ayudará a predecir el precio aproximado de las viviendas.




(a) Precios de la zona (b) Tiroteos en la zona (c) Zonas con inundaciones

Figura 1: Comparación de eventos.

Otro resultado interesante ocurrió cuando se decidió plantear una hipótesis nula, esto es, suponer que hay una distribución aleatoria de puntos y luego comprobar estadísticamente que no es así, para ello usamos el método de Monte Carlo, simulando una distribución de puntos aleatoria miles de veces, para finalmente comparar los resultados obtenidos con nuestra distribución de puntos inicial, la idea es tomar el algoritmo KNN para tomar un promedio de los precios de viviendas agrupadas según el mismo ZIP CODE, para poder demostrar que existe un agrupamiento, donde viviendas con precio similar están cerca de viviendas con precio similar, y no simplemente se distribuyen espacialmente las viviendas con precios de venta aleatorios, sino que el sector realmente importa.

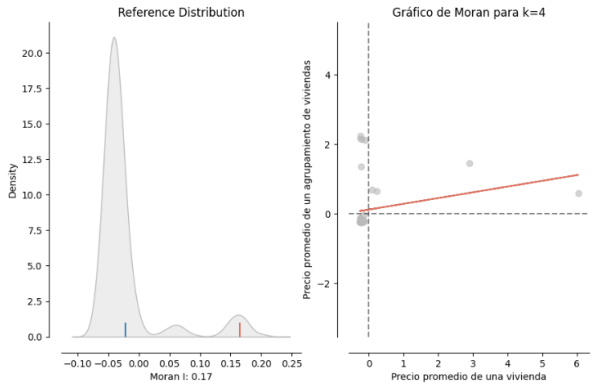



Figura 2: Análisis I de Moran.

| Parámetro | Valor | Interpretación |
|---|---|---|
| Vecinos ($k$) | 4 | — |
| Valor de Moran's I | 0.1655 | Autocorrelación positiva |
| Valor p | 0.0308 | Significativo (p menor a 0.05) |

Cuadro 1: Análisis de autocorrelación espacial de precios por ZIP Code.

Una vez obtenidos los resultados estadísticamente representativos, concluimos que nuestra hipótesis nula no es cierta, así que demostramos que si hay agrupamiento espacial (clustering) de los puntos con sus atributos de precio, así que el precio no es aleatoriamente distribuido espacialmente en Manhattan.

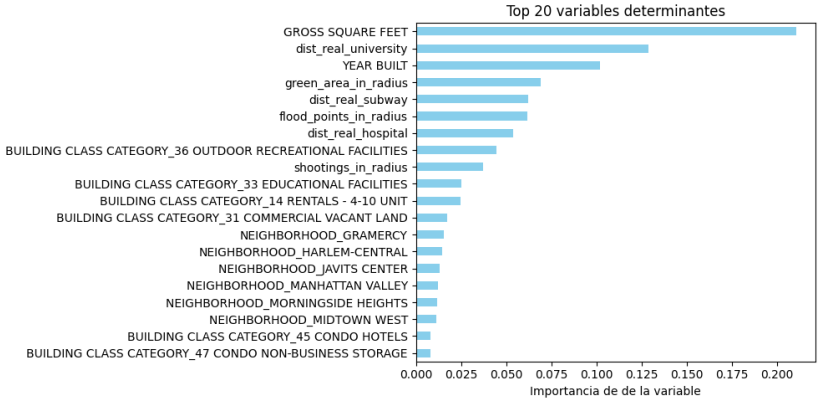



Figura 3: Resultados del modelo predictivo.

Podemos ver las variables más determinantes en el precio de cada vivienda, que según el modelo es el espacio de la propiedad. Además, hay una visualización interactiva donde se pueden variar los parámetros, y la visualización se actualizará con propiedades vendidas en el rango de precio dada por el modelo. Finalmente, se agregó una visualización dinámica para ver la influencia de los parámetros en el modelo predictivo.

## 5.2. Conclusiones

Si bien los resultados son prometedores, existen diversas áreas en las que el proyecto puede perfeccionarse, entre las que destacan:

a. **Las distancias incluyen calles cerradas por infraestructuras de la ciudad:** Si bien el algoritmo Dijkstra captura una distancia relativamente precisa, no se consideran calles cerradas en ciertos horarios por diferentes motivos.

b. **Para lograr mejores resultados para el proyecto, sería necesario disponer de un conjunto de datos más amplio y equilibrado:** La mediana y el promedio aritmético del precio de venta de las viviendas es de aproximadamente \$10,000 y \$1,800, respectivamente. El modelo entrega una predicción con un coeficiente de determinación de aproximadamente 44 % ($R^2$) y un error absoluto medio (MAE) de alrededor de \$1.85 millones.

   Esto indica que, si bien el modelo capta cierta estructura, aún presenta una alta dispersión en sus predicciones. Ha pesar de que se limitado la muestra del entrenamiento al percentil 95, para unos mejores resultados es necesario usar un modelo predictivo más potente, especialmente porque en Manhattan hay una alta variación de precios entre los extremos, pero principalmente se requiere de una colección mayor de datos, ya que sólo estamos utilizando datos geocodificados de propiedades vendidas en una cantidad menor a 2mil, lo que es bastante deficiente para que el modelo pueda capturar altas variaciones de precios según los parámetros dados, hay que tomar en cuenta que el 70 % de estos datos se utilizan en el entrenamiento, mientras que el 30 % restante se utiliza para probar los resultados del entrenamiento.

Más allá de las limitaciones, el proceso realizado permitió un acercamiento fructífero al análisis de datos Geoespaciales, consolidando aprendizajes que serán fundamentales para el futuro.

# 6. Referencias


1. Tinsa. (2023). *Manhattan, la isla más cara del mundo*. https://www.tinsa.es/blog/curiosidades/manhattan-la-isla-mas-cara-del-mundo/

2. Partir a Nueva York. (2023). *Manhattan: información del distrito más famoso de Nueva York*. https://www.partir-a-new-york.com/es/distrito-de-nueva-york/manhattan-es

3. GIS and Beers. (2021). *Cálculo del índice NDBI para análisis urbanísticos*. https://www.gisandbeers.com/calculo-indice-ndbi-analisis-urbanisticos/

4. J. Leal-Villamil. (2022). *Análisis comparativo de asertividad para tres índices de zonas construidas aplicados a ciudades colombianas*. Revista de Ingeniería, Universidad Pedagógica y Tecnológica de Colombia (UPTC). https://revistas.uptc.edu.co/index.php/ingenieria_sogamoso/article/view/15018/12232

5. Google Earth Engine. (2023). *COPERNICUS S2 SR HARMONIZED – Google Earth Engine Dataset Catalog*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED#bands

6. EPSG.io. (2023). *EPSG Geodetic Parameter Dataset*. https://epsg.io/

7. Esri. (2023). *How Spatial Autocorrelation (Moran's I) works*. https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

8. Esri. (2023). *What is a z-score? What is a p-value?* https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

9. Scikit-learn developers. (2024). *Scikit-learn: Machine learning in Python*. https://scikit-learn.org

10. GeoPandas. (2024). *GeoDataFrame.dissolve*. https://geopandas.org/en/stable/docs/reference/api/geopandas.GeoDataFrame.dissolve.html

11. PySAL. (2024a). *KNN – libpysal.weights.KNN*. https://pysal.org/libpysal/generated/libpysal.weights.KNN.html#libpysal.weights.KNN

12. PySAL. (2024). *API Reference – libpysal*. https://pysal.org/libpysal/api.html

13. OSMnx. (2024). *User Reference*. https://osmnx.readthedocs.io/en/stable/user-reference.html

14. NetworkX. (2024). *multi_source_dijkstra_path_length – NetworkX*. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.shortest_paths.weighted.multi_source_dijkstra_path_length.html

15. H3. (2024). *H3 Geoindexing System*. https://h3geo.org/

16. GeoPandas. (2024). *GeoSeries.buffer*. https://geopandas.org/en/stable/docs/reference/api/geopandas.GeoSeries.buffer.html

17. La Nación. (2023, septiembre 29). *Las intensas lluvias provocan inundaciones repentinas en Nueva York e interrupciones de servicios*. https://www.lanacion.com.ar/estados-unidos/las-intensas-lluvias-provocan-inundaciones-repentinas-en-nueva-york-e-interrupciones-de-servicios-nid29092023/

18. Spectrum Noticias NY. (2024, agosto 6). *Ha comenzado la lluvia: Advierten de posibles inundaciones. Advertencia de viaje por tormentas eléctricas*. https://spectrumnoticias.com/ny/nyc/noticias/2024/08/06/ha-comenzado-la-lluvia--advierten-de-posibles-inundaciones--advertencia-de-viaje-por-tormentas-electricas

19. Scikit-learn developers. (2024). *sklearn.ensemble.RandomForestRegressor*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

20. GeeksforGeeks. (2023). *Random Forest Regression in Python*. https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/

21. Kim, A. (2021, marzo 23). *Random Forest Regression - Clearly Explained!*. YouTube. https://www.youtube.com/watch?v=YUsx5ZNlYWc

22. StatQuest with Josh Starmer. (2018, diciembre 4). *Random Forests Part 1 – Building, using and evaluating*. YouTube. https://youtu.be/AYICIq5jnhU

23. StatQuest with Josh Starmer. (2018, diciembre 4). *Random Forests Part 2 – Out-of-bag error, feature importance and more!*. YouTube. https://youtu.be/kFwe2ZZU7yw?t=718

24. Moonbooks. (n.d.). How to clear the output in a Jupyter Notebook cell after each for loop iteration. https://en.moonbooks.org/Articles/How-to-clear-the-output-in-a-Jupyter-Notebook-cell-after-each-for-loop-iteration-/

25. IPython. (n.d.). IPython.display — IPython 9.0.2 documentation. https://ipython.readthedocs.io/en/9.0.2/api/generated/IPython.display.html

26. Akyürek, B. (2020, junio 10). Map Visualization with Folium. https://medium.com/datasciencearth/map-visualization-with-folium-d1403771717