Diplomatura Ciencia de Datos,
Inteligencia Artificial y sus Aplicaciones
en Economía y Negocios

Trabajo Final:
Tasación de Inmuebles

Integrantes: BUFFA, Joe; MANZANO, Carolina; MENGHI, Gonzalo;
TORRES YANQUEN, Diego.

TRABAJO FINAL: TASACIÓN DE INMUEBLES

Introducción

El mercado inmobiliario es un sector clave en la economía de cualquier ciudad o región. Tanto compradores como vendedores necesitan información precisa y actualizada sobre los precios de los inmuebles para tomar decisiones informadas. En este sentido, la capacidad de predecir los precios de los inmuebles puede ser muy valiosa.
En este informe se presentará un modelo para predecir los precios de los inmuebles en Córdoba, una ciudad en constante crecimiento que se ha convertido en un mercado inmobiliario cada vez más atractivo, utilizando una base de datos del mes de julio de 2022 proveniente de zonaprop.com.ar en donde distintas personas publicaron inmuebles de su propiedad que se encontraban a la venta.

Descripción del problema y Objetivos del trabajo

A partir de la importancia que implica para los compradores y vendedores de inmuebles conocer el precio de publicación de los mismos para estar informados y aumentar la seguridad en la toma de decisiones, resulta oportuno desarrollar un modelo predictivo que permita la tasación de inmuebles en la Provincia de Córdoba a partir de la base de datos obtenida desde la web dentro de una ventana temporal.
Es en este sentido que el desarrollo del modelo que presentamos persigue algunos objetivos mínimos deseables:

- Nuestra principal motivación es explorar formas de construcción de un modelo preciso para la predicción del precio de los inmuebles, del que resulten estimaciones confiables para compradores, vendedores e interesados en el mercado inmobiliario.
- Pretendemos identificar las características claves de los inmuebles que suponen un mayor impacto en el precio de publicación. Factores como tamaño del inmueble en m2, ubicación, antigüedad y cantidad de ambientes como habitaciones, baños, garajes, etc. resultan - en distinta medida y combinación- influyentes en el precio.
- Adaptar el modelo a distintos tipos de descripciones de inmuebles para que su funcionamiento sostenga la mayor parte de su confiabilidad ante el paso del tiempo.

El desarrollo del modelo mantiene varios objetivos comerciales, y es a partir de esto que inferimos que su puesta en funcionamiento y publicación permitirá a las personas:

- Poder evaluar si el precio de publicación de un inmueble es justo y si se ajusta a sus necesidades y expectativas para poder tomar decisiones más eficientes.
- Obtener negociaciones más efectivas: Estar informados le permitirá a los compradores utilizar esta información para obtener un mejor precio, mientras que a los vendedores les ayudará a establecer un precio competitivo para maximizar sus ganancias.

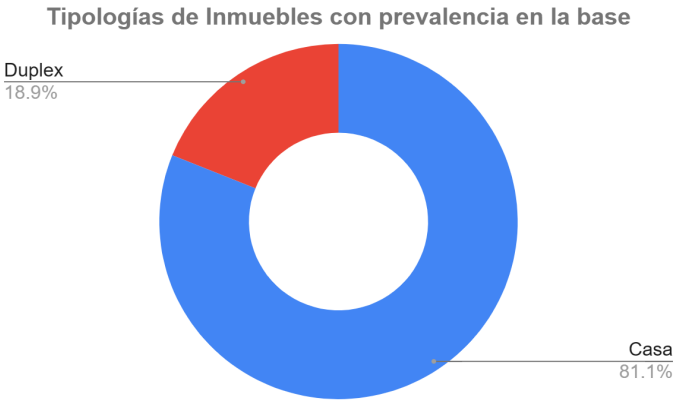
- Poder planificar financieramente a ambas partes, ya que los compradores pueden determinar si el precio publicado se ajusta a su presupuesto y capacidad de endeudamiento, mientras que los vendedores pueden estimar el retorno de su inversión y planificar futuras transacciones.
- Poder identificar oportunidades de inversión y tomar decisiones estratégicas en función de las tendencias de precios.

Análisis exploratorio de los datos y sus conclusiones

La base de datos con la que trabajamos con datos de inmuebles publicados a la venta en julio 2022 en la Provincia de Córdoba cuenta con las características de la siguiente imágen:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7966 entries, 0 to 7965
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   idPost           7966 non-null   int64  
 1   Precio           7966 non-null   int64  
 2   title            7966 non-null   object 
 3   Description      7957 non-null   object 
 4   direccion        7927 non-null   object 
 5   barrio           7927 non-null   object 
 6   m2total          7818 non-null   float64
 7   m2cubierto       7686 non-null   float64
 8   Banos            7703 non-null   float64
 9   Dormitorios      7732 non-null   float64
 10  Ambientes        6881 non-null   float64
 11  Antiguedad       5392 non-null   float64
 12  Estrenar         7966 non-null   int64  
 13  Cochera          5706 non-null   float64
 14  Estado           7966 non-null   int64  
 15  Luminoso         7966 non-null   int64  
 16  Toilette         1705 non-null   float64
 17  coordenadas.lat  7902 non-null   float64
 18  coordenadas.lng  7902 non-null   float64
dtypes: float64(10), int64(5), object(4)
memory usage: 1.2+ MB
```

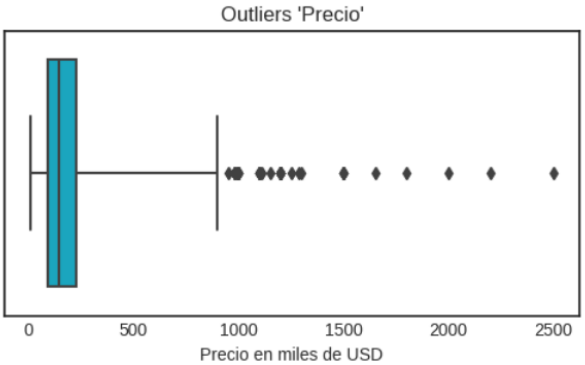
Comenzamos con el **Análisis de las Variables Cualitativas** en donde identificamos y eliminamos 547 valores duplicados de la variable “IdPost” y clasificamos cada publicación por tipología de inmueble publicado (casa, dúplex, departamento, local, terreno, etc) basado en la variable *“title”* donde decidimos mantener sólo aquellas observaciones que correspondan al tipo *“Casa”* o *“Duplex”* y agregarlas a una nueva columna llamada “Type” ya que las mismas abarcan un poco más del 95% de la base de datos.



Por otro lado, caracterizamos cada publicación según las siguientes palabras que consideramos relevantes por la gran frecuencia en la que se mencionan en la variable "Description" y creamos una columna por cada una. Las mismas son: *"Suite", "Pileta", "Patio", "Oportunidad", "Reciclar", "Country" y "Excelente ubicación".*

Al realizar el **Análisis de las Variables Cuantitativas**, nos enfrentamos al gran desafío de lograr que los datos representen valores razonables y lógicos ya que al tratarse de una base de datos creada por personas que poseen criterios y opiniones diferentes al momento de completar la publicación de su propiedad en venta en la página web, la información se torna por momentos abstracta.

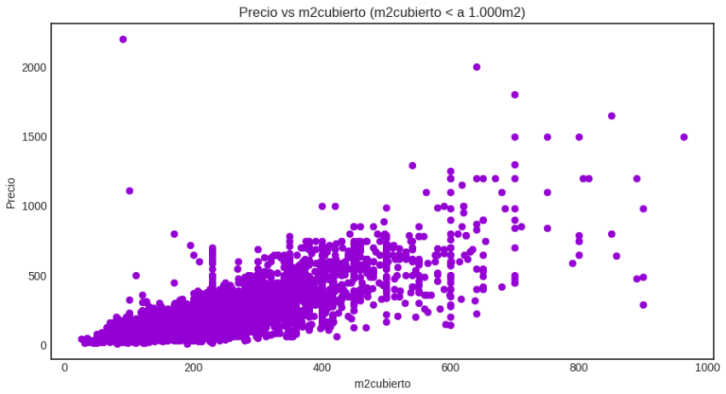
Uno de los casos, por ejemplo, fue en la *variable "Precio"* donde tuvimos que omitir las publicaciones en las que el valor de venta era menor a USD 1.500 ya que consideramos que no tiene sentido alguno.

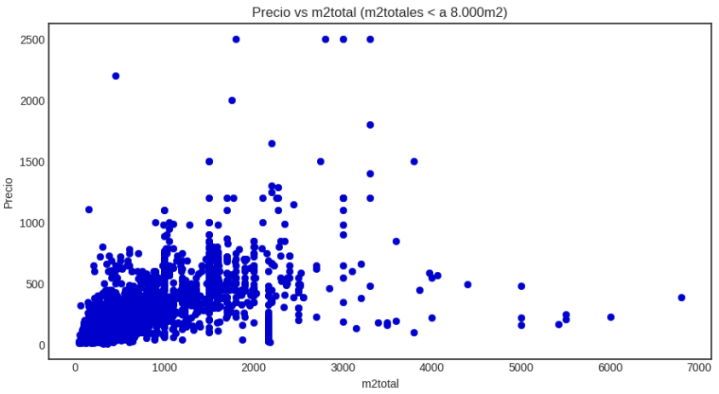



```
Medidas descriptivas de la variable 'Precio'
count    7022.000000
mean      195.914681
std       172.364368
min         9.999000
25%        90.000000
50%       143.000000
75%       230.000000
max      2500.000000
Name: Precio, dtype: float64
```

Descubrimos que en la *variable "m2total" y "m2cubierto"* existían publicaciones con m2 totales y m2 cubiertos menores a 50m2 (los cuales consideramos "valores atípicos"), por lo que decidimos reemplazar los valores no informados por el vendedor por los m2 informados en la variable *"Description"*. Caso contrario, por el promedio de m2totales y m2cubiertos respectivamente de acuerdo al tipo de inmueble que se encontraba vendiendo.

Además, luego de analizar el comportamiento de la variable *"Precio"* con respecto a los *"m2totales"* y *"m2cubiertos"*, concluimos que sería una buena decisión trabajar con aquellas publicaciones en las que los m2cubiertos sean menores a 1.000 m2 y los m2totales sean menores a 8.000 m2, ya que al superar esos valores consideramos estar en presencia de valores atípicos para un inmueble en venta dentro de las tipologías consideradas ("Casas" y "Dúplex").





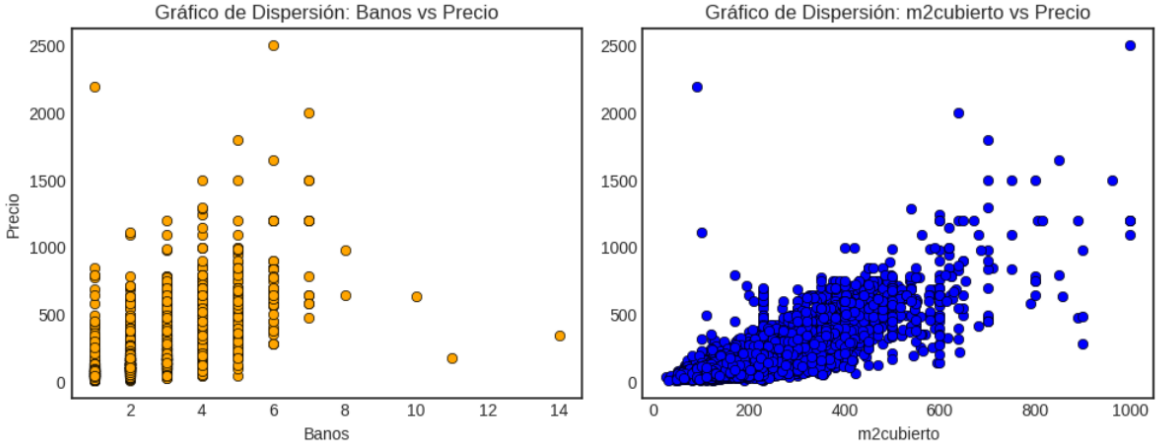
Con el resto de las variables también trabajamos con los valores nulos, como en la variable *"Baños"* donde le asignamos un valor según la información de la variable *"description"*y en caso de no poseer la misma, asignamos el valor uno (1) a la misma; y en "Dormitorios" donde también lo hicimos según lo identificado en la descripción en algunos casos y en el resto, completamos dicho dato con el promedio de la variable. Además, decidimos eliminar aquellas publicaciones que poseen más de seis (6) dormitorios ya que los consideramos "valores atípicos" y no presentan relación con la variable objetivo.
En la variable *"Cochera"* y *"Antigüedad"*, reemplazamos los valores pérdidos por el número cero (0) ya que se supone que si el vendedor del inmueble no completó esa información es porque el lugar no cuenta con garage o el inmueble está a estrenar.
Por último, con respecto a la variable *"Toilette"* y *"Ambientes"*, luego de analizarlas e identificar la cantidad de valores faltantes y a su vez teniendo en cuenta que son campos que se tornan abstractos y a libre interpretación de cada persona que realiza una publicación, decidimos desestimarlas de nuestro análisis.
A partir de estas modificaciones creamos y guardamos una nueva base de datos para poder realizar el modelado de los datos.

Modelo aplicado, Métrica seleccionada, Modelo elegido

Luego del análisis exploratorio y limpieza de los datos, analizamos la correlación de las variables del conjunto de datos y observamos que las que más correlación presentan con la variable objetivo son las siguientes:

- ***"m2cubierto"*** en un 80% de forma positiva
- ***"m2total"*** en un 64% de forma positiva
- ***"Banos"*** en un 63% de forma positiva

Además, observamos una alta correlación entre la variable *"Banos"* y *"m2cubiertos"* donde luego de analizarla concluimos que es algo normal y no implicaría un problema de multicolinealidad; y en *"m2totales"* y *"m2cubiertos"* donde podríamos quedarnos con la variable *"m2cubiertos"* y desestimar los *"m2totales"* ya que manteniendo solamente la primera estaríamos haciendo un buen análisis.



Luego de esto, utilizamos la función **StandardScaler()** de Scikit-Learn a los fines de escalar las variables de nuestro conjunto de datos y probamos los siguientes modelos:

- Análisis de Regresión Lineal Múltiple
- Nearest Neighbors
- Árboles de decisión para problemas de regresión
- Random Forest
- Redes Neuronales

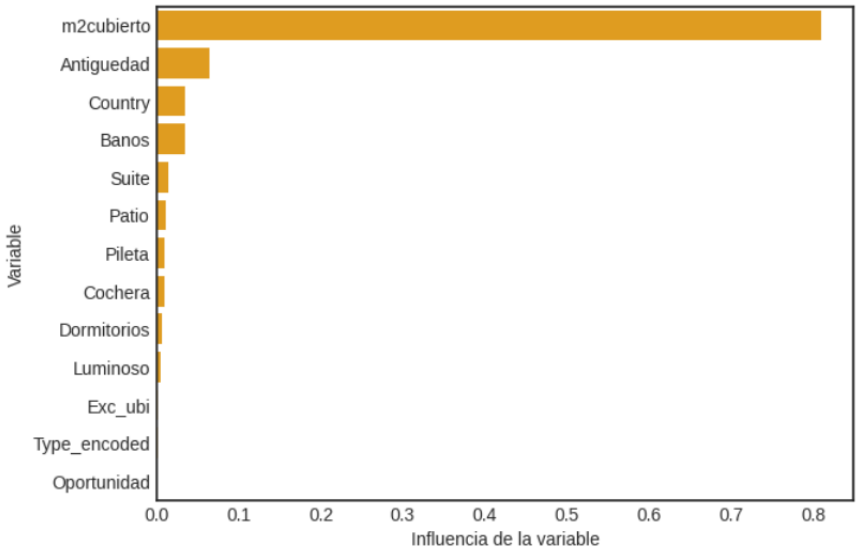
Para cada una de las técnicas de modelización seguimos los siguientes pasos para construir cada uno de los modelos:

- Elegir un algoritmo que implemente la técnica correspondiente
- Buscar una combinación de parámetros eficaz para el algoritmo elegido
- Crear un modelo utilizando los parámetros encontrados
- Entrenar el modelo en el conjunto de datos de entrenamiento
- Probar el modelo en el conjunto de datos de prueba y obtener los resultados

A partir de los resultados obtenidos por cada modelo, presentamos una tabla que muestra las evaluaciones de cada modelo cuando se aplica al conjunto de pruebas X_test:

	Modelo	MAE-Test	MSE-Test	RMSE-Test
0	Ridge Regression	57.721035	10593.832225	102.926344
1	Elastic Net	97.346535	23643.283842	153.763727
2	Nearest Neighbors	57.175323	11491.418886	107.198036
3	Árboles de Decisión	54.340628	10548.140163	102.704139
4	Random Forest	50.617127	9217.708452	96.008898
5	Redes Neuronales	53.054432	9597.764345	97.968180

A la hora de seleccionar el modelo óptimo, es esencial basarnos en métricas de rendimiento adecuadas por lo que decidimos elegir la métrica ***MAE (Error Absoluto Medio)*** ya que buscamos robustez frente a valores atípicos y deseamos conocer la desviación promedio de las predicciones.
Analizando los resultados obtenidos, observamos que el modelo de **"Random Forest"** es el que presenta el menor valor de MAE en 50.6171 lo que significa que, de media, el modelo predecirá un valor mayor o menor real en 50.6171 miles de dólares (USD 50.617,10).
Por otro lado, observamos cuáles son las variables más importantes para este último modelo y concluimos que definitivamente la variable más influyente es la de *"m2cubierto"*



Conclusiones finales

Hay varios aspectos de esta investigación que merecen una mención y nos proponemos extraer lo más significativo de cada etapa de este proyecto. Antes de comenzar el repaso queremos ser categóricos con esta afirmación; recolectamos evidencia y experiencia suficiente para sentirnos satisfechos con el trabajo y los resultados que aquí se presentan, especialmente si consideramos que nuestro objetivo se centró en la exploración de modelos. Y más allá de lo que respecta exclusivamente al tema de investigación nos enfrentamos a un problema cotidiano de la vida socio-económica de cualquier ciudad de Argentina, algo tan esencial como definir el valor monetario de un bien.

Comenzamos el desafío tomando datos de una web pública, es muy probable que al replicar la experiencia podamos permitirnos trabajar a partir de la adquisición de datos subutilizados para agregarle valor a partir de un exhaustivo y cuidadoso proceso de limpieza y posterior EDA. Aprendimos a trabajar en la construcción de un modelo iterativo que explore tantas posibilidades como nuestra creatividad nos lo permita. Agregar valor no pasa desapercibido, nos llevamos la posibilidad de demostrar cuán interesantes pueden volverse los datos que están a los ojos de todos.

Desde un principio notamos que el dataset no escatimaba en problemas en lo que respecta a errores humanos y/o datos que desafían la intuición. Como no siempre es sencillo discriminar si es información útil o trash data entendimos la necesidad de un criterio estadístico para normalizar la muestra. Trabajamos en la limpieza a partir de la supresión de valores duplicados quitando IDs repetidos, de valores atípicos analizando la mediana de las variables y los saltos en el rango intercuartílico, quitamos también variables por estar afectadas por una gran cantidad de valores faltantes y abstractos (como lo fue la variable *“Toilette”* y problemas para encontrar criterios para contar la cantidad de ambientes). Nuestras dificultades en este aspecto quedaron comprobadas al momento del testeo de normalidad y heterocedasticidad de nuestro modelo. Aplicamos a los residuos pruebas de *Jarque-Bera*, el *test omnibus de K2 de D’Agostino* y una prueba adicional de *Kolmogorov-Smirnov*, todo concluye en que la distribución de los errores no se ajusta a una normal, con p-valores muy inferiores a lo estándar, por lo que rechazamos la hipótesis de normalidad.

Confirmar la importancia de la variable *"m2cubiertos"* con datos resultó un hallazgo significativo, a partir del análisis de correlación de variables entendimos rápidamente que se volvería la más importante de nuestro proyecto, avalado más adelante a partir del análisis de influencia de variables. De esta manera, tomando dicha variable junto a: *"Banos", "Dormitorios", "Antiguedad", "Cochera", "Luminoso", "Suite", "Pileta", "Patio", "Oportunidad", "Country", "Exc_ubi", "Type"* logramos explicar poco más del 70% de la variabilidad del precio.

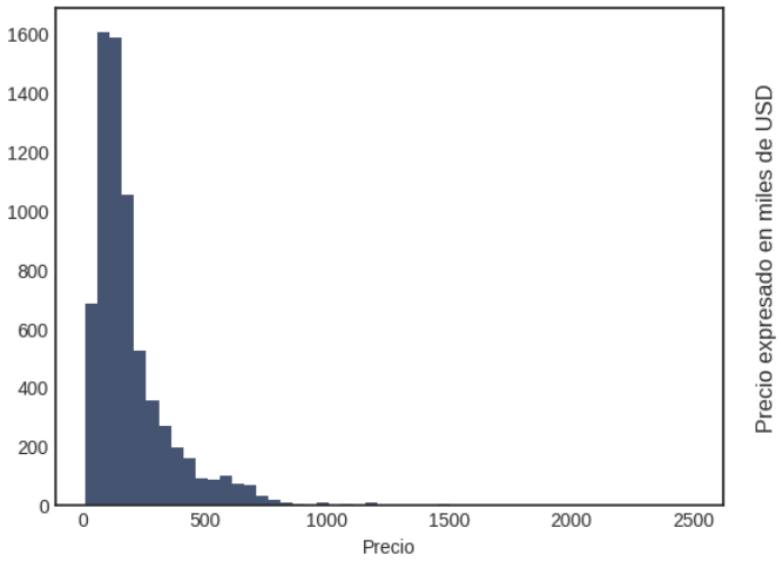
Es buen momento para pensar en nuestra mayor oportunidad de mejora: el análisis geográfico de los inmuebles publicados. Es plausible pensar que además de la cantidad de m2cubiertos de una propiedad, importa también en qué zonas se encuentran ubicados. Si

bien comenzamos a trabajar en la creación de zonas geográficas preferimos dedicar el tiempo a iterar a partir de la información que teníamos disponible.

Es por eso que en la siguiente etapa normalizamos las variables mediante StandardScaler() de Scikit-Learn, lo que nos permitió proceder con una comparativa objetiva de distintos modelos predictivos: desde la Regresión Lineal Múltiple hasta métodos más complejos como Redes Neuronales. En cada uno, se buscó la mejor configuración de parámetros y se evaluó su rendimiento, estableciendo un marco robusto para las fases de entrenamiento y prueba. Los resultados obtenidos nos acercan a un modelo predictivo eficaz. Hemos construido varios modelos predictivos, incluyendo *Random Forest*, *Redes Neuronales*, *Árboles de Decisión*, *Nearest Neighbors*, *Regresión Múltiple Ridge* y *Regresión Múltiple Elastic Net*, y los evaluamos en términos de rendimiento y capacidad predictiva.

Como resultado de este análisis, hemos identificado que el **Modelo "Random Forest"** sobresale en términos de precisión en la predicción de precios de viviendas en Córdoba. Pero también analizamos que los mejores modelos son los que funcionan a partir de métodos de aprendizaje supervisado no lineal. La variación del MAE entre nuestras pruebas nos sugiere que es la mejor para captar complejidades y patrones no lineales en la formación de los precios y evitar en el camino algunos problemas de sobreajuste.

La decisión metodológica que nos llevó a considerar que nuestro modelo es razonable fue que, teniendo un error medio absoluto (MAE) de nuestro modelo seleccionado (Random Forest) de 50,617.1 USD, y considerando que la media del precio de venta en nuestra base de datos es de 195,400.88 USD, la mediana es de 141,750.00 USD, y además el primer cuartil es de 92,000 USD (lo que indica que el 75% de los datos supera este valor), el nivel de error resulta razonable por estar en sintonía con las características del mercado inmobiliario local.



	Precio
count	4874.000000
mean	195.400876
std	168.945723
min	9.999000
25%	92.000000
50%	141.750000
75%	230.000000
max	2500.000000

Elaboración propia en base a los datos proporcionados por Zonajob del año 2022.

No queremos pasar por alto otros aspectos que, al igual que la incorporación de un criterio de análisis geográfico de la variable precio y revisar el rechazo en la hipótesis de normalidad y heterocedasticidad, mejorarían el rendimiento de nuestro modelo. Nos preguntamos por la posibilidad de seguir experimentando, por ejemplo, filtrando el dataset a partir de distintos rangos de precios de venta - menores a la media o mayores a la media- e interpretar a partir de la performance del modelo en cada escenario planteado. Por supuesto, otro punto a considerar está en que trabajamos a partir de una foto temporal de un mes, un interesante camino para recorrer y fortalecer el trabajo es el de ampliar la ventana temporal para que el modelo se vuelva mucho más flexible al discriminar temporadas y movimientos generales del mercado.

Elegimos dejar para el final otro tipo de preguntas - no por eso menos importantes- que tienen que ver con nuestro contexto: ¿cómo influyen en el precio de los inmuebles los factores (algo más) subjetivos como la inestabilidad económica, la desesperación del vendedor o cambio de precios por pagos en efectivo? y lo que nos resulta más apasionante ¿cómo captarlas como variables para nuestro modelo? Notamos algunas tendencias en variables proxy como “oportunidad” o “dueño vende” a las cuales no les asignaremos una importancia directa, pero nos permiten identificar (si las usamos con criterio) otras relaciones más significativas. Nos conformamos temporalmente con una respuesta propositiva: “*itera y triunfarás".*