FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico N° 1: Análisis Exploratorio de Datos ZonaProp

Organización de Datos [75.06 / 95.58]
Segundo Cuatrimestre 2019

Grupo N° 43

Nombre	Padrón
Peyrano Diego Javier	94254
Escobar Benitez Maria Soledad	97877
Luciano Ortiz	100323
Fernando Moreno	96683

Contenido

Introducción

Objetivo del Trabajo Práctico

El objetivo de este trabajo práctico es aplicar las herramientas vistas en clase para realizar un análisis exploratorio de datos del registro histórico de publicaciones de ZonaProp. El set de datos incluye el registro de todas las publicaciones realizadas entre los años 2012 y 2016 en México.

La idea principal es estudiar los datos y ver qué cosas interesantes se pueden encontrar sobre los datos para luego volcar los mismos en este informe.

Sobre la Empresa

ZonaProp es el mayor portal de compra y venta de inmuebles.

El portal cuenta con la mayor variedad de casas, departamentos, oficinas comerciales y más.

Cuenta con una página web y una aplicación. ZonaProp permite buscar inmuebles filtrando las publicaciones de acuerdo a los requisitos de cada usuario, pudiendo seleccionar si desea comprar, alquilar, vender y además indicar el tipo de inmueble. También puede filtrarse de acuerdo a la ubicación, como ciudad, barrio, etc., o de acuerdo a la cantidad de ambientes y lo más importante, de acuerdo al precio, permitiendo así que cada usuario sólo vea las publicaciones acerca de inmuebles a los que puede acceder de acuerdo al rango de valores que puede permitirse.

Sobre los Datos

El dataset consta de propiedades en venta en México entre los años 2012 y 2016, valuadas en pesos mexicanos. El archivo train.csv tiene 240K filas y 22 columnas.

- id: Un id numérico para identificar la propiedad.
- titulo: El título de la propiedad publicada.
- descripcion: La descripción de la propiedad publicada.
- direccion: La dirección de la propiedad.
- ciudad: La ciudad de la propiedad.
- provincia: La provincia donde está localizada la propiedad.
- lat: Latitud.
- lng: Longitud.
- tipodepropiedad: El tipo de propiedad (Casa, departamento, etc).
- metrostotales: Metros totales de la propiedad.
- metroscubiertos: Metros cubiertos de la propiedad.
- antiguedad: Antigüedad de la propiedad.
- habitaciones: Cantidad de habitaciones.
- garages: Cantidad de garages.
- banos: Cantidad de baños.

- fecha: Fecha de publicación.
- gimnasio: Si el edificio o la propiedad tiene un gimnasio.
- usosmultiples: Si el edificio o la propiedad tiene un SUM.
- piscina: Si el edificio o la propiedad tiene una piscina.
- escuelascercanas: Si la propiedad tiene escuelas cerca.
- centroscomercialescercanos: Si la propiedad tiene centros comerciales cerca.
- precio: Valor de publicación de la propiedad en pesos mexicanos.

SECCION I - Análisis introductorio

Antes de comenzar con el análisis de los datos para cada set de datos específico nos tomamos el tiempo de reconocer los datos, observar tipos de datos, información que proveen, cantidad de memoria que ocupan y demás detalles que se describirán en esta sección.

1.0 - Validez de los Datos

A fines prácticos la validación de datos se refiere a la manipulación de los tipos correspondientes a cada uno de los atributos de archivo csv sobre el cual se realizan los procedimientos.

En este caso utilizando pandas, el package o biblioteca de funciones en Python del paquete instalado e importado al entorno permite la transformación en tiempo de ejecución y lectura del archivo raíz.

Alrededor de 24000 elementos fueron procesados para cada una de las categorías transformando en su mayoría por necesidad, datos del tipo flotante a enteros. Esto no solo permite una utilización menor de memoria, sino que también las operaciones y el análisis numérico mantiene un acercamiento mas exacto con los valores establecidos. Por ejemplo, sería innecesario disponer de datos como 'usos múltiples' bajo un formato de reales cuando su funcionalidad es discreta. Con excepción de las coordenadas y fechas, todos los demás expresan cantidades discretas y enteras.

En cuanto a los precios se hizo una transformación a enteros pese a la posibilidad de dejarlo lógicamente en flotante. Sabemos que los precios pueden manejarse como numero reales, sin embargo en este casos se pudo observar con notoriedad que cada uno de ellos no tenía números decimales, por lo tanto, salvo para alguna operación posterior que requiera este formato, se manejaron simplemente como enteros.

1.1 Tratamiento de los datos en memoria

La conversión de elementos es necesaria para aquellos casos en los que el set de datos tiene dimensiones demasiado elevadas. Este no es dicho caso, pero conociendo las categorías, y teniendo predisposición de las herramientas necesarias, es sabio interpretar y codificar los datos para su futuro proceso. Sin importar el tamaño, el manejo correcto de los datos predispone un uso mas dósil y al mismo tiempo más escalable en memoria.

Un ejemplo de ello es claramente como los tiempos de ejecución de los notebooks son reducidos (tanto los dispuestos en la nube como procesos locales) gracias al tamaño de los datos desde el punto de vista de los distintos tipos de constantes.

Se redujo el uso de memoria de 45MB a 27MB aproximadamente y los tipos resultantes son finalmente los siguientes:

dtypes: category(1), datetime64[ns](1), float64(2), int32(14), object(5)
memory usage: 27.7+ MB

Casos especiales fueron el manejo de ambas coordenadas (tanto latitudes como longitudes) en números flotantes (por tener la exactitud de un número real) y “Category”, en el uso de los tipos de propiedades. Un solo caso de fechas y por último 5 objetos (en estos casos todos strings).

Para todos los demás restantes la conversión a enteros simplifico principalmente la memoria definitiva utilizada del set.

Sección II - Análisis Cartográficos

Esta sección corresponde a ubicar principalmente sobre el territorio continental mexicano la correspondencia de los datos arrojados por el set. El eje principal de cada uno de los análisis de esta sección está comprometido principalmente con las coordenadas.

De manera anticipada y tras la validación de los datos se puede observar que tenemos elementos Nan que se reemplazaron por elementos enteros igual a 0.

En este caso se eliminaron anticipadamente aquellas filas que no tenían una referencia clara con respecto a que o donde apuntaba en lo que respecta a datos geográficos.

2.0 – Análisis Global

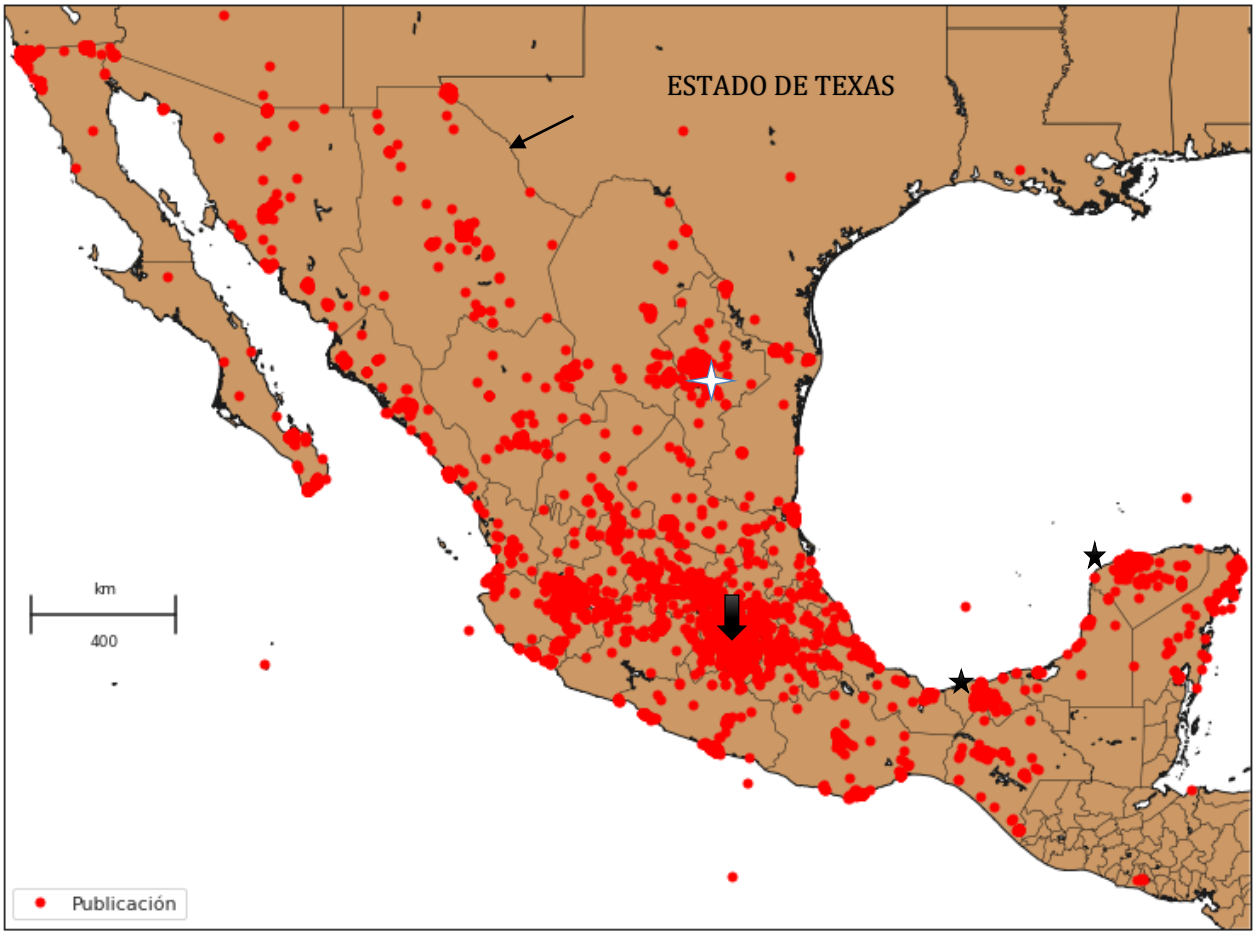
Arrojando las ubicaciones sobre el globo gracias a los datos de coordenadas se puede observar lo siguiente:

Tras el plot anterior se puede mostrar que una amplia cantidad de ubicaciones presentan coordenadas que se encuentran tanto dentro como fuera del país de México. Y no solamente ello, sino que también hay publicaciones cuyas coordenadas no coinciden con territorio continental o islas, sino que se encuentran en el medio del mar u océanos. Un análisis simple podría indicar una cantidad bastante similar en el continente europeo, Latinoamérica y Asia Oriental. Con respecto a Oceanía, Europa oriental y Medio Oriente prácticamente no presentan publicaciones. Solamente África y Arabia Saudita se destaca por apenas una o dos publicaciones a simple vista.

Esto amerita analizar en detalle la cantidad exacta de publicaciones en el exterior de México el cual es muy reducido pero definitivamente los datos arrojan dichas ubicaciones, ya sea por error o verdadera publicación.

2.1 - Publicaciones en territorio mexicano

Publicaciones en territorio Continental Mexicano



La ciudad de México (refiriéndose a la capital de dicho país y marcado con una flecha hacia abajo) y los estados en los alrededores acaparan la mayor cantidad de publicaciones. Estos pueden verse principalmente en la zona media del país. Por otro lado, las zonas abarcadas por aquellos estados aledaños tanto en el sur o el norte presentan mayoritariamente ubicaciones sobre sus respectivas capitales.

Vuelve a incrementarse también en las fronteras al sur una elevada cantidad mientras que en la frontera norte es inferior en comparación. No obstante, cabe destacar que la frontera norte abarca una zona perimetral mucho mayor en comparación con el sur. Sobre todo, puede verse que a lo largo de la frontera con Texas (señalado con una flecha) están muy cercanas al límite. Por otro lado, en el sur se presentan más allá de aquellas dentro de los límites algunas acumulaciones alrededor de las capitales de dichos estados (marcados con estrellas negras). Y por último otra acumulación en Nuevo León (marcado con una flecha blanca).

2.2 - Publicaciones más populares

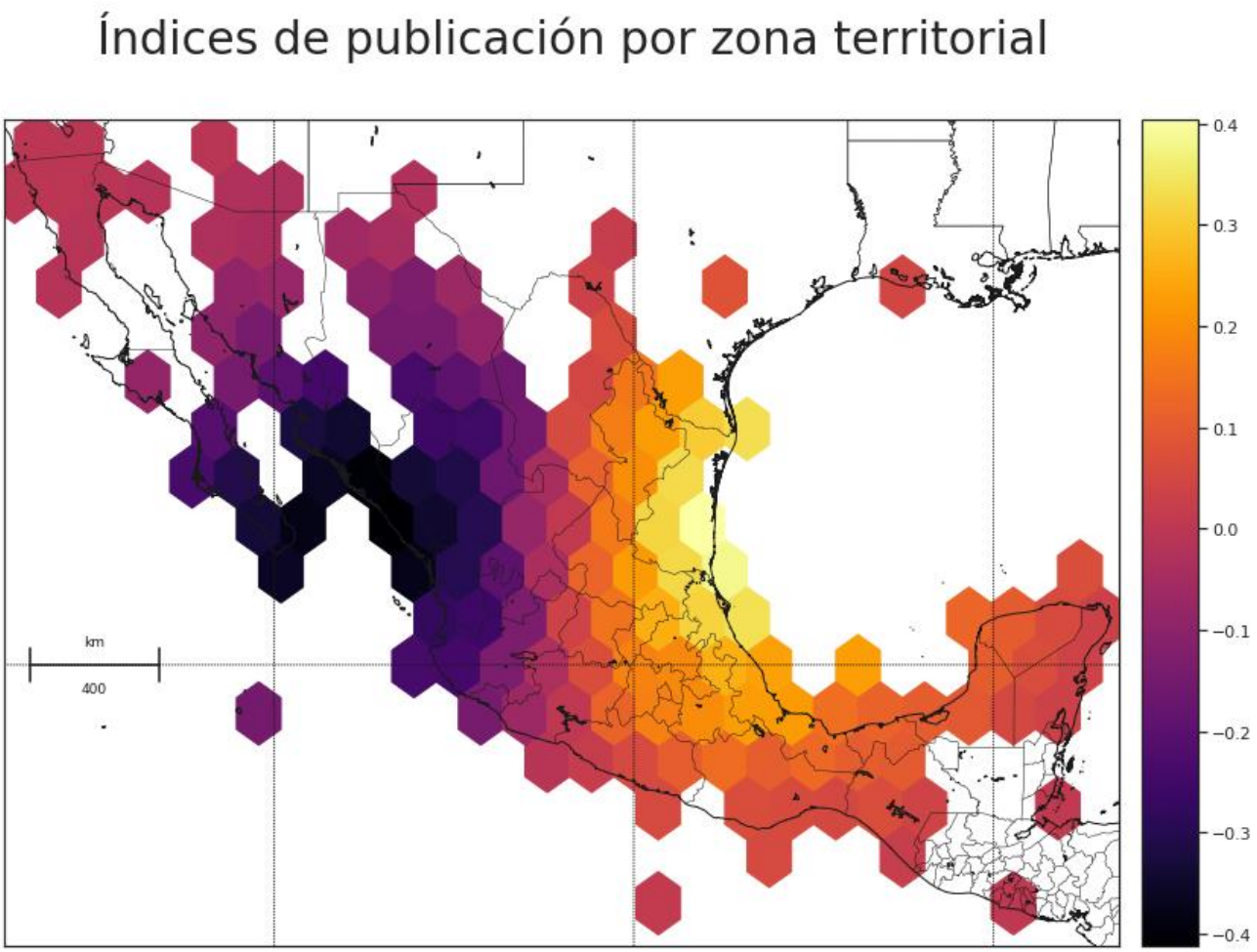
Eligiendo como tipos de publicaciones según el tipo de inmueble, tomamos las variacioens mas comunes o al menos las que presentan mayor cantidades empezando por la mas abundante (casas) hasta la menos abundante (Comercios) .

Comenzando por las casas, es la publicaion mas común y se la puede encontrar en todo tipo de lugares, desde periferias, centros,capitales y zonas rurales como urbanas.

La situación es muy similar para las Apartamentos pero de manera mucho mas aisladas entre sí. Por otro lado los terrenos son mas abundantes particularmente en la zona central y sur del país y no tanto en la zona Norte. Lo mismo sucede para los condominios exeptuando que estos mantienen de manera predilecta ubicaiones en la capital del paíz y zonas aledañas.

Por otro lado los comercios es muy usual y casi en general que se encuentren principalmente en las zonas de ciudades más importantes o capitales de cada estado. Esto mantiene una logica natural en donde los comercios practicamente marcan la tendencia de las cudades mas importantes y allá donde cubren dichas ciudades, estas publicaioens suelen agruparse y estar muy juntas entre sí.

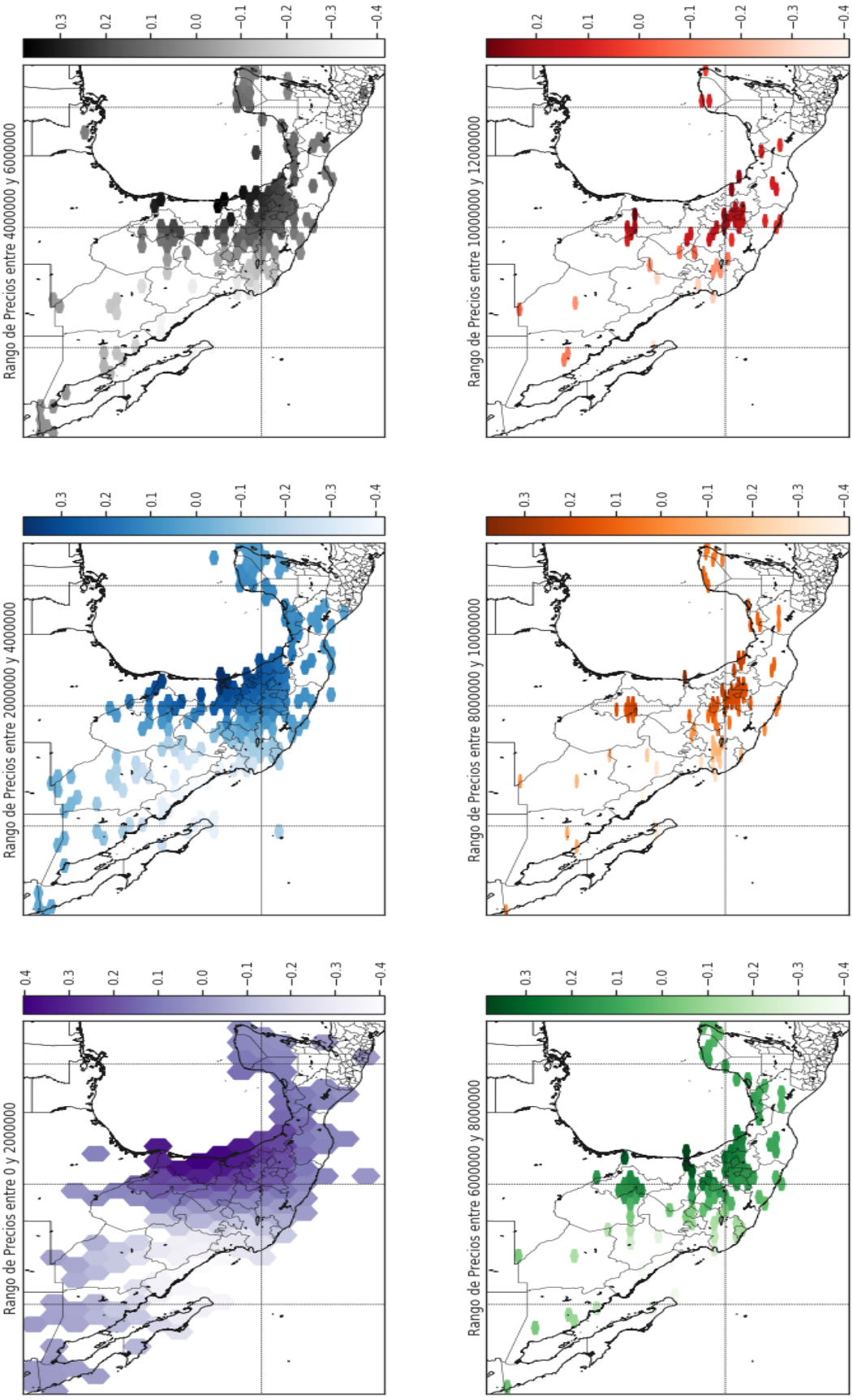
2.3 - Índices de publicaciones por zona territorial



En coincidencia con otros análisis este tipo de visualización arroja la tendencia por zona de acuerdo a la cantidad de publicaciones que se cubren a la largo del territorio mexicano. Se puede ver claramente como la zona media acapara la mayor cantidad de publicaciones contra la costa este del país. Este tipo de visualización a diferencia del posicionamiento simple de coordenadas establece por una zona o área determinada un patrón de cantidad de publicaciones por metros cuadrados procesando las coordenadas de la zona cubierta. Mostrando clara presencia de mayores publicaciones al sur y un poco menos hacia el norte, recordando que la región o superficie de esta última es mucho mayor que la anterior.

Cada región (o zona hexagonal) cubre alrededor de casi más de 150km^2

2.4 – Índices de publicaciones de acuerdo al Precio



En primer lugar elegimos disponer del mapa de manera horizontal ya que constan de 6 subplots cuya disposición muestran claramente la evolución del heatmap en la zona territorial mexicana, mantenerlos de manera vertical no contribuye a tener un panorama general más amplio.

Hablando de lo que nos muestra, podemos ver que, en la secuencia, viendo como el precio escala de menor a mayor en cada mapa, las publicaciones van acentuándose cada vez más en el centro del país.

Podemos advertir claramente que las publicaciones más baratas se encuentran más al norte y las más caras hacia el centro del país. Las tendencias indican notablemente esta convergencia de la misma manera para la cantidad de publicaciones como la relación con sus precios.

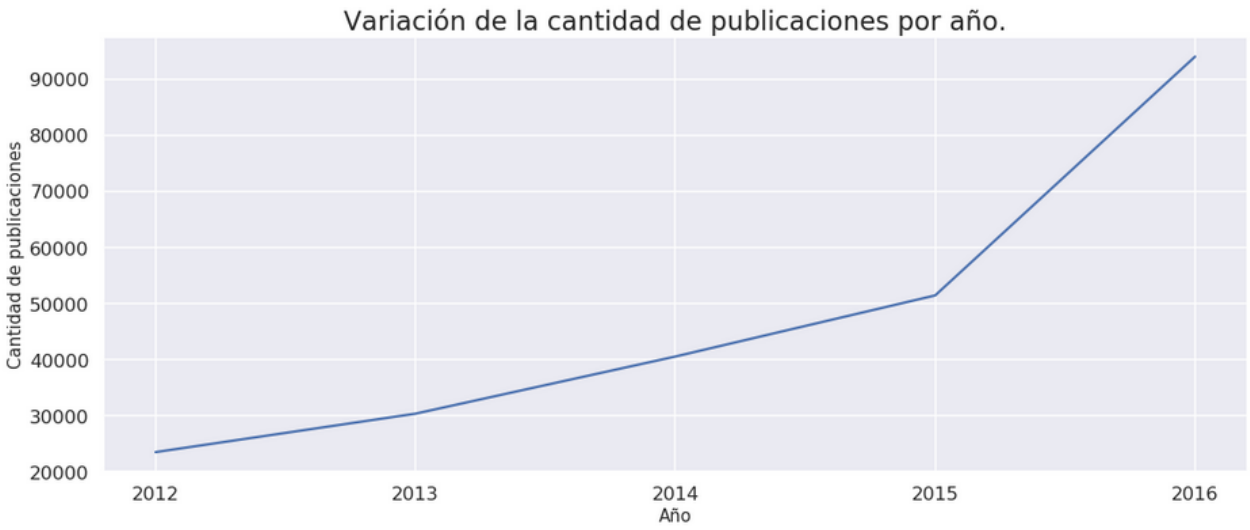
Sección III - Análisis Temporales

Los datos analizados corresponden a un intervalo de tiempo entre 2012 y 2016 inclusive. El objetivo principal de esta sección es encontrar que detalles importantes se pueden destacar a lo largo de este periodo que estén involucrados con variaciones en los precios u otras características de los inmuebles detallados en el tiempo.

3.0 - Variación de la cantidad de publicaciones a través del tiempo

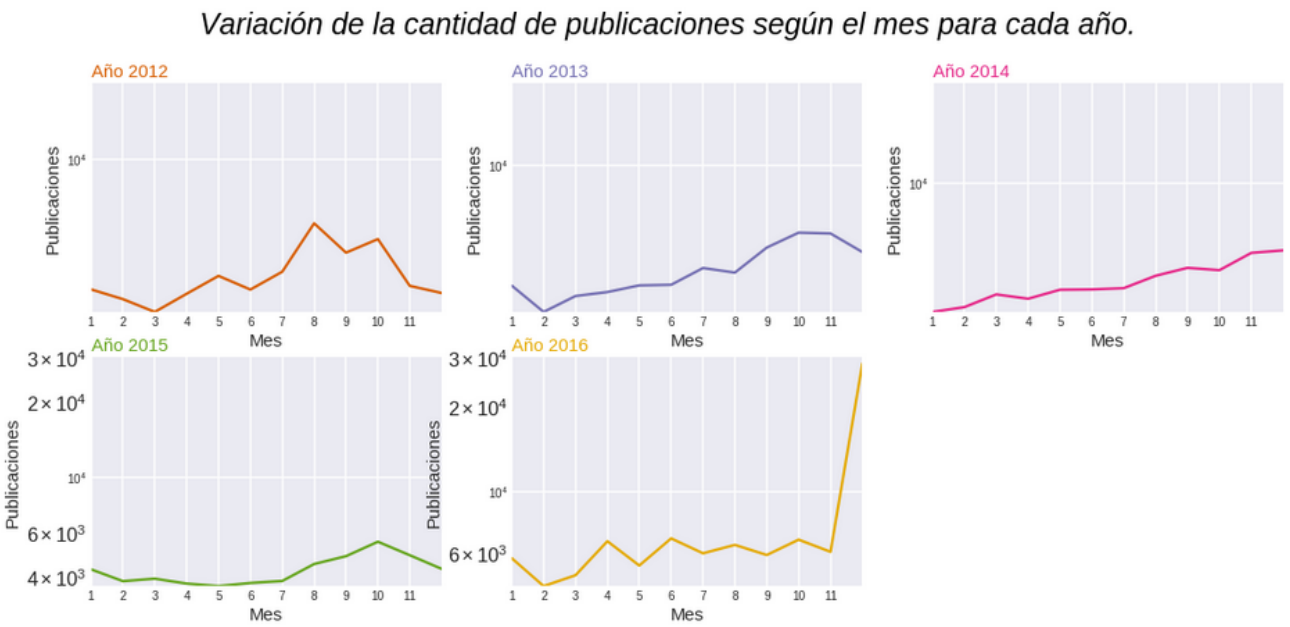
Se observa que la variación de la cantidad de publicaciones realizadas en cada año dispuesto es un simple grafico de línea, el cual sirve para detallar a grandes rasgos como varían la cantidad de publicaciones de acuerdo a cada año registrado.

Como se dijo anteriormente el periodo corresponde al 2012-2016 y la tendencia en el mismo con respecto a la cantidad de publicaciones por año de los inmuebles ha ido notablemente en aumento, No solo ello, sino que además luego del año 2015 puede verse una pendiente aun más favorable al crecimiento. Zonaprop demuestra entonces que además de presentar un aumento en sus operaciones también a dominado el mercado inmobiliario alcanzando sus mayores índices de aumento a lo largo del 2015.



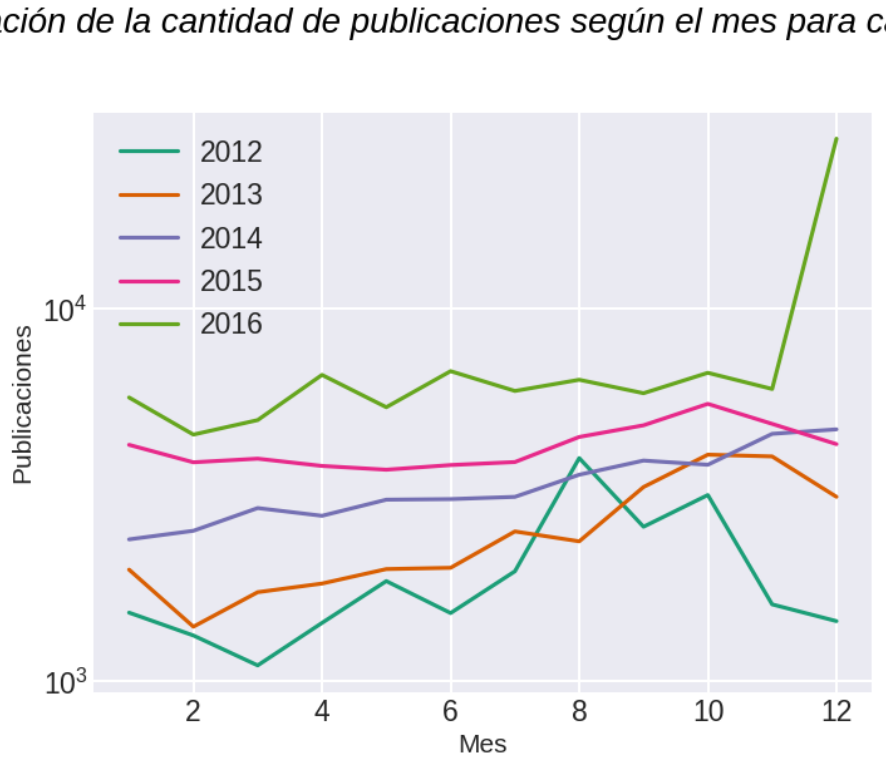
3.1 - Variación de publicaciones según los meses para cada año

Debido al resultado del gráfico anterior se explora un poco más en detalle el comportamiento de la cantidad de publicaciones para cada año, ahora profundizando un poco más en la cantidad de publicaciones realizadas por mes para cada uno de los años.



Este gráfico permite ver las tendencias de la variación de la cantidad de publicaciones en más detalle. La mayoría de los casos, durante la segunda mitad del año, incrementan las publicaciones un poco más de lo normal, con respecto a los primeros meses, y luego los últimos 2 o 3 meses de cada año las publicaciones comienzan a disminuir, es por esto que puede verse que alrededor de los meses de enero y febrero las publicaciones siguen disminuyendo. Se puede insinuar que el último cuatrimestre es el de mayor variación con respecto a un aumento.

Es posible que esta clase de detalles muestra una incidencia en otras características de las publicaciones a lo largo del informe, tales como el precio u cualidades temporales adicionales.



Esta visualización muestra en conjunto la misma variación mencionada anteriormente

3.2 - Variación del precio a lo largo del tiempo

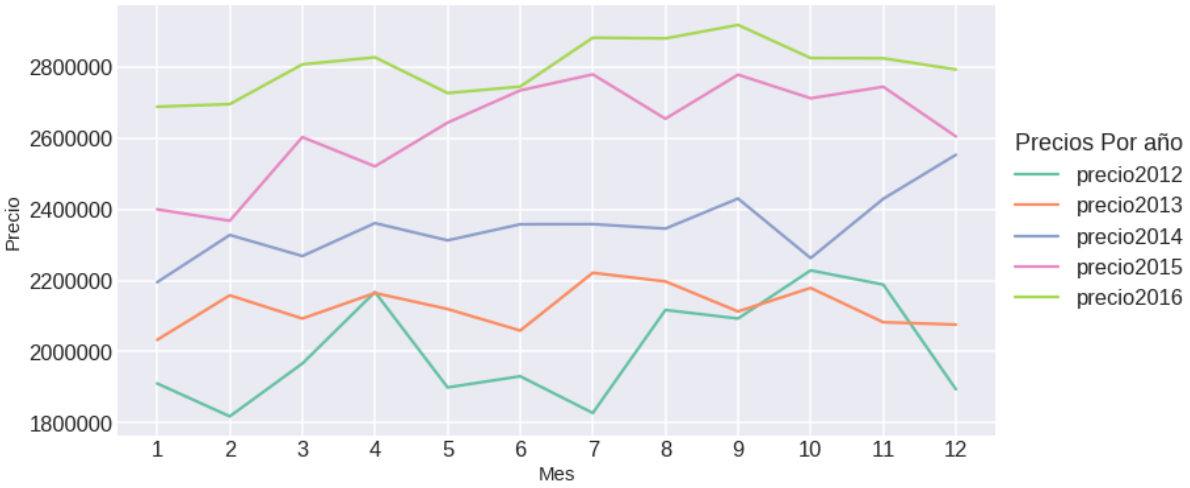
Para tener una primera visión a grandes rasgos lo primero es observar como varía el precio promedio de propiedades por año y se visualiza en un gráfico de líneas. Al igual que en la sección anterior, se obtiene un gráfico con tendencia creciente a lo largo de los años, el mismo se puede observar a continuación.

Como se menciona, en la sección anterior vimos que la cantidad de publicaciones iba en aumento a lo largo del tiempo, y aquí observamos el mismo comportamiento sobre la variable precio.

Ya que las publicaciones aumentaron a lo largo del tiempo, resulta lógico ver que el precio varíe un poco, ya que se tienen más publicaciones, más datos y mayor cantidad de valores para los precios, aunque la fuerte tendencia creciente podría pensarse que es causada por algún tipo de devaluación en la moneda local, pero eso no puede asegurarse. Por ahora sólo se menciona como una posible hipótesis o simplemente una conjetura apresurada, mientras que se elige preferentemente analizar los datos aún más para encontrar respuestas a este comportamiento.

Para ver un poco más en detalle, el gráfico que está a continuación muestra claramente como cada año los precios son más altos. También algo parecido ocurre con la variación en la cantidad de publicaciones. Durante los meses intermedios, los precios aumentan y durante el último mes bajan un poco.

Variación del precio promedio de las propiedades según el mes para cada año.



Además, son pocos los casos en donde los precios vuelven a coincidir con los de algún año anterior. Claramente van en aumento, ¿tendrá esto que ver con algún aumento en la oferta de algún tipo de propiedad en particular, la cual resulta particularmente más costosa? Es sólo una pregunta que sirve de referencia para esta instancia, pero se intenta responder con el próximo análisis.

Algo que resulta particularmente interesante, es como el factor tiempo resultará útil a la hora de predecir el precio, pues cuando la fecha es más “alta”, mayor es el precio promedio, por lo que será un buen feature para analizar y probar en los modelos predictivos a desarrollar.

3.3 - Publicaciones por año para los tipos de propiedades más populares

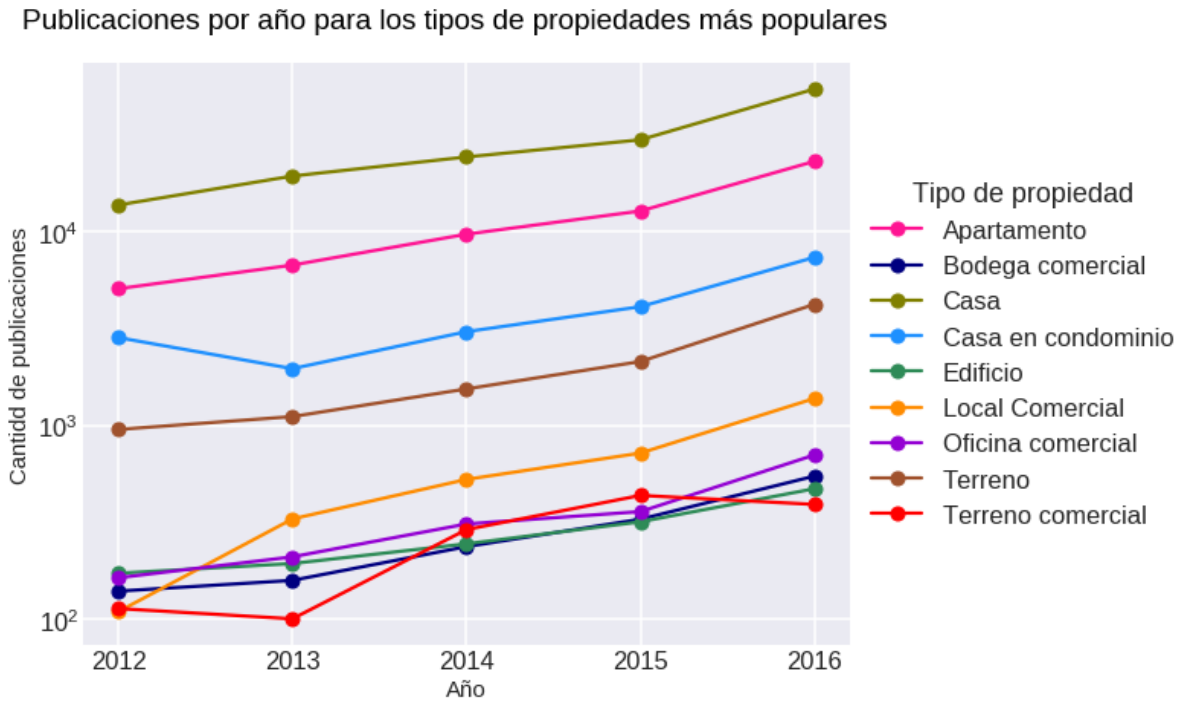
¿Cómo varía la oferta de propiedades según su tipo a lo largo del tiempo? ¿Tendrá esto algo que ver con el comportamiento de la variación de precios?

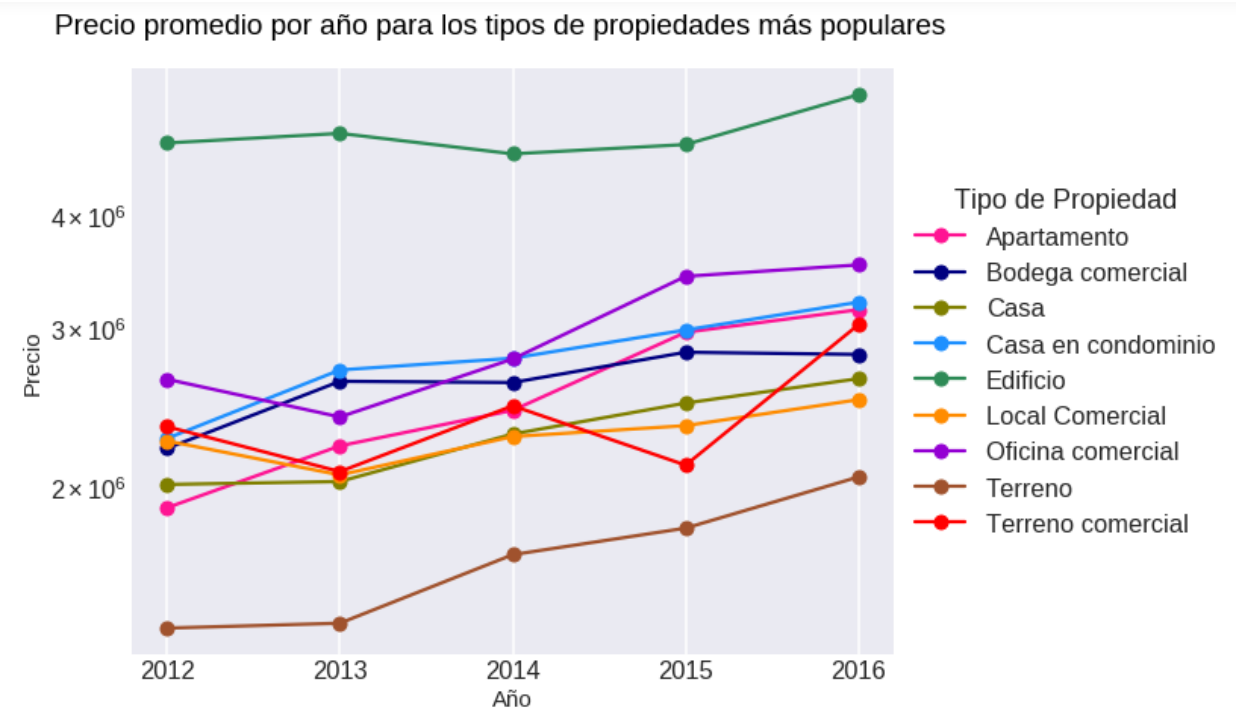
La idea de este análisis es ver cómo varían las publicaciones, por año, para los tipos de propiedades más populares. Para identificar a los tipos de propiedades más populares se decide filtrar a los tipos que tengan más de 100 publicaciones por año.

Primero se realiza un conteo de la cantidad de publicaciones registradas para cada tipo de propiedad y luego se filtran las más populares, luego esos datos son volcados en un gráfico para visualizar su comportamiento.

Como es interesante encontrar una relación con el precio, se realiza un análisis similar sobre dicha variable. Agrupando las propiedades por tipo se toma un promedio de los precios registrados para cada tipo de propiedad, luego se filtran los tipos que anteriormente se clasificaron como populares.

A continuación, se encuentran ambos gráficos, con el propósito de observar detenidamente el comportamiento que describe acerca de los diferentes tipos de propiedad.





Aclaración: en ambos gráficos el color para cada tipo de propiedad es el mismo.

Se observa cómo el gráfico que muestra la cantidad de publicaciones tiene las líneas más dispersas, sin embargo, el gráfico sobre los precios tiene una "acumulación" de líneas en un rango de precio, entre ~2 millones y poco más de 3 millones de pesos mexicanos, además este comportamiento se mantiene a lo largo del tiempo sin importar como varíen los valores de los precios.

Se puede deducir que la mayoría de las propiedades se mantienen dentro de cierto rango de precios, en el cual casualmente se encuentran la mayoría de las propiedades con más publicaciones, es decir, las propiedades populares se mueven dentro del mismo rango de precios.

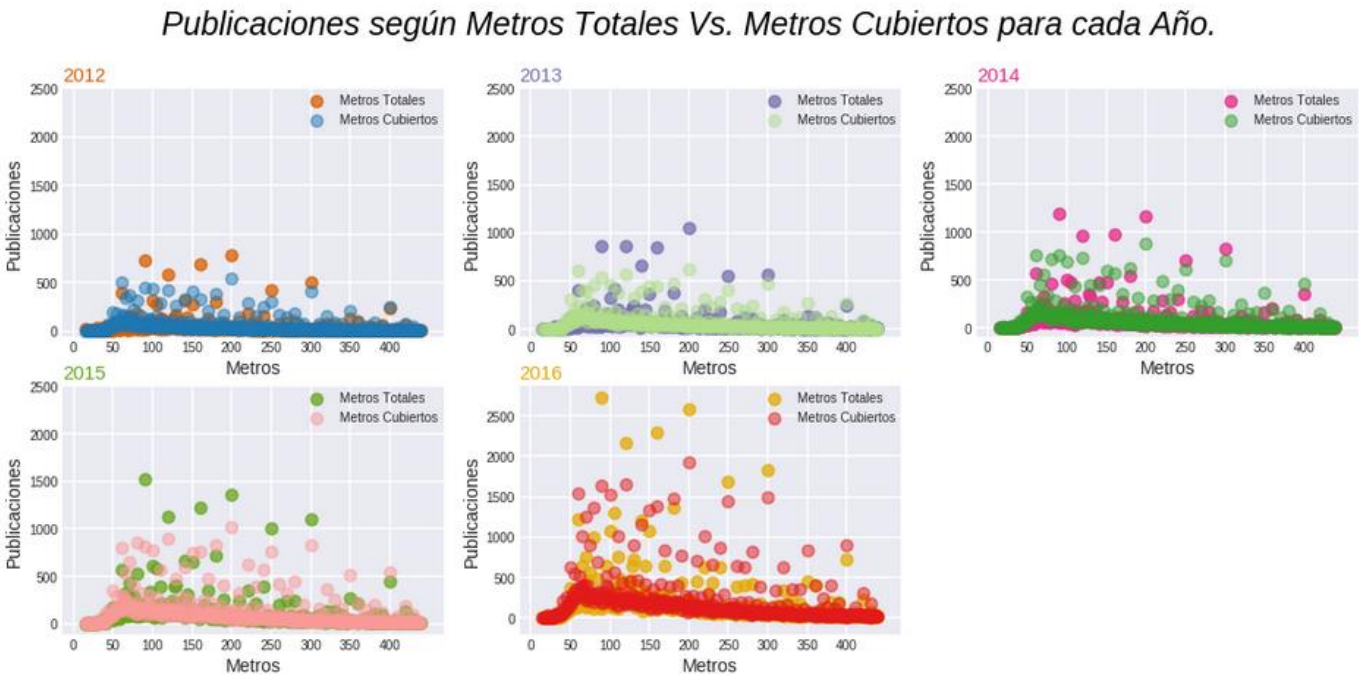
Otra curiosidad para observar es como el tipo de propiedad "Casa" es el más popular, con una cantidad que escala por encima de las 10 mil publicaciones anuales y vemos también que su precio no registra fluctuaciones demasiado bruscas a través de dicho periodo y el mismo se mantiene dentro de los tipos accesibles. De todos ellos, el tipo "Terreno" es el más accesible en cuanto a precio. Era esperable, pues se trata en su mayoría de lotes vacíos sin ningún tipo de construcción, donde el valor se le asigna sólo dependiendo de la zona en la que se encuentre ubicado.

La contracara de todo esto es el tipo "Edificio", el cual resulta ser el más caro de todos, superando los 4 millones de pesos mexicanos, y casualmente, no es uno de los más populares en publicaciones, ubicándose apenas por encima de las 100 publicaciones anuales.

3.4 - Metros Cubiertos vs. Metros Totales de Propiedades a través del Tiempo

Es de interés ver cómo se comportan a través del tiempo. Ver si hay algo interesante sobre ellos, si hay algunos tamaños más populares, si a través del tiempo las propiedades tienen más o menos metros de superficie, ya sea cubierta o total, y luego, ver si esto nos puede dar información sobre el precio.

Se continúa viendo como varía la cantidad de publicaciones con respecto a los metros de las propiedades.



En el grafico se muestra, como se amplió anteriormente, que la cantidad de publicaciones aumenta a través del tiempo, lo cual no es nada nuevo para nosotros, por lo que nos enfocamos en el comportamiento de los metros cubiertos y totales.

En general se observa que el comportamiento de los metros totales y metros cubiertos es parecido, se distribuyen entre los mismos valores para los metros, aunque vemos que los metros totales son un poco más populares que los metros cubiertos en cuanto a número de publicaciones.

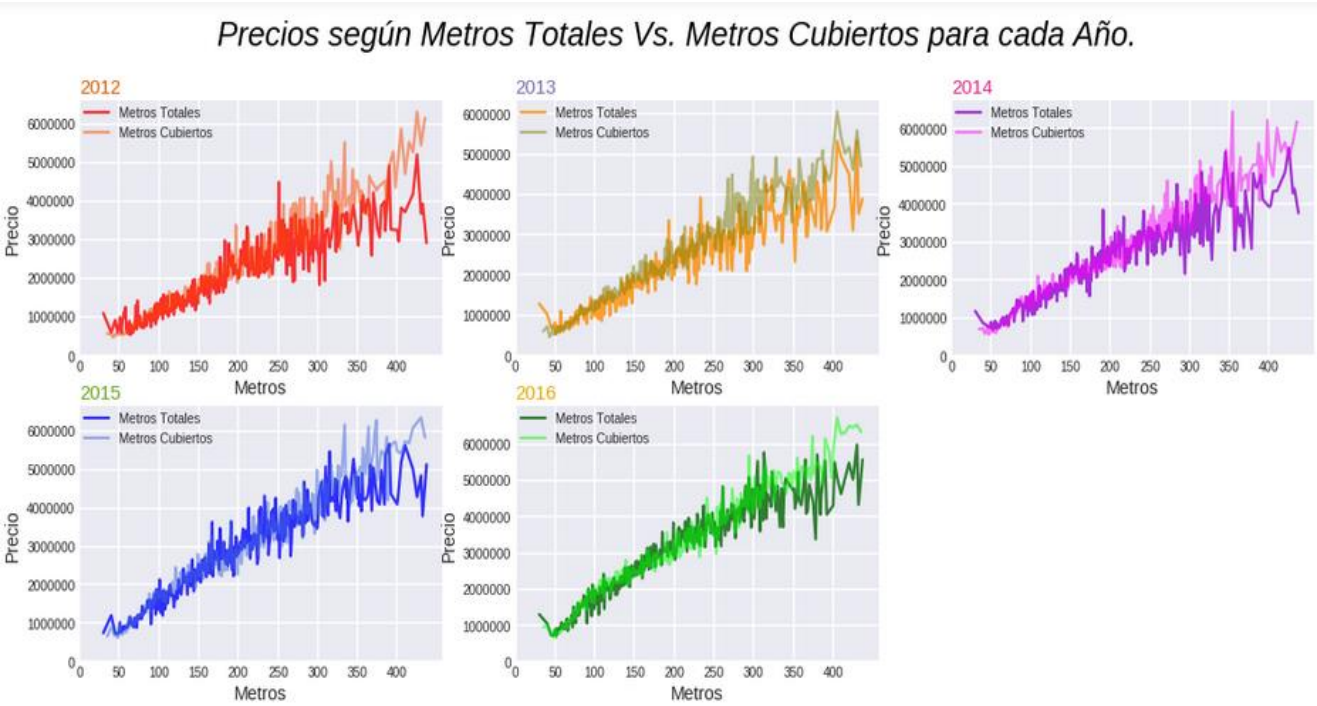
En todos los años se destaca un pequeño pico creciente entre las propiedades que abarcan entre 50 y 100 metros, por lo que podría decirse que estos tamaños para las propiedades son los más populares. También se ve un aglutinamiento de puntos cercanos al eje x, esto nos indica que a medida que la cantidad de publicaciones disminuye, la viabilidad en los metros aumenta, ya sean cubiertos o totales.

En general vemos que los valores para los metros cubiertos tienen un comportamiento bastante uniforme en la mayoría de los gráficos, a diferencia de los metros totales que suelen registrar puntos de popularidad más altas, en general entre los 100 y 200 metros.

Por último, en el gráfico correspondiente al año 2016 se puede observar que las popularidades aumentan abruptamente en comparación al resto de los años y como se analizó con anterioridad esto se debe a la suba de la cantidad de publicaciones registradas para ese año.

Se observa que ahora los metros entre 50 y 300 recibieron una gran cantidad de publicaciones, por lo que se podría decir que se han vuelto más populares.

Intentemos ver ahora cómo se relaciona esto con los precios.



En el gráfico se observa que los metros totales y cubiertos se comportan de una manera bastante razonable, cuando menos metros dispone el precio es menor y a medida que los metros aumentan el precio lo hace también. Las variaciones de precios entre metros cubiertos y metros totales no son demasiado significativas en la mayoría de los casos. También se observa el aumento de los precios a lo largo del tiempo, aunque no es demasiado evidente en estos gráficos como lo han sido en otros casos analizados anteriormente.

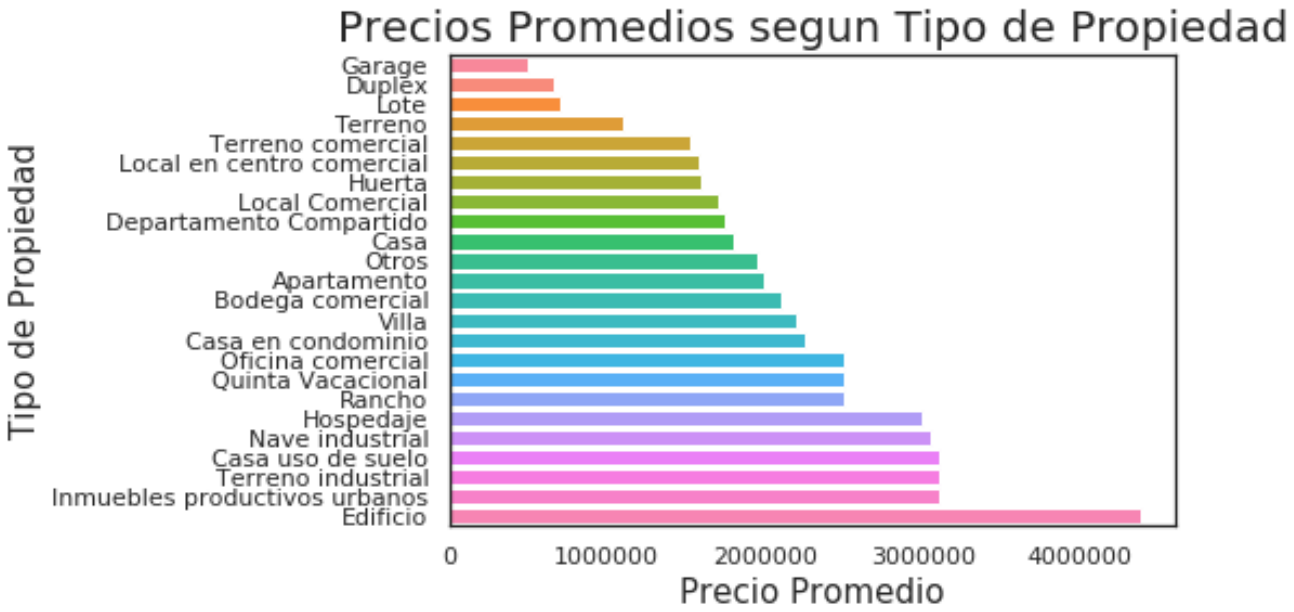
Por lo observado se puede concluir que lo interesante de estas variables es que podrían llegar a funcionar muy bien como features a la hora de probar modelos para la predicción de precios, puesto que se comportan de manera similar.

SECCION IV - Análisis sobre Precios

A continuación, se procede a analizar el precio de las propiedades según las distintas características de estas.

En esta sección será de utilidad estudiar en profundidad todo lo relacionado al precio puesto que nos servirá para sentar las bases para un proceso de análisis predictivo.

4.0 - Tipo de Propiedades VS Precio



Ya se menciono en análisis anteriores la manera en la que se disponen los precios entre tipos de inmuebles o propiedades más populares, mostrando su tendencia temporal. Sin embargo, en esta sección se analizará en más detalle aquellos parámetros de precios en total a lo largo del periodo completo.

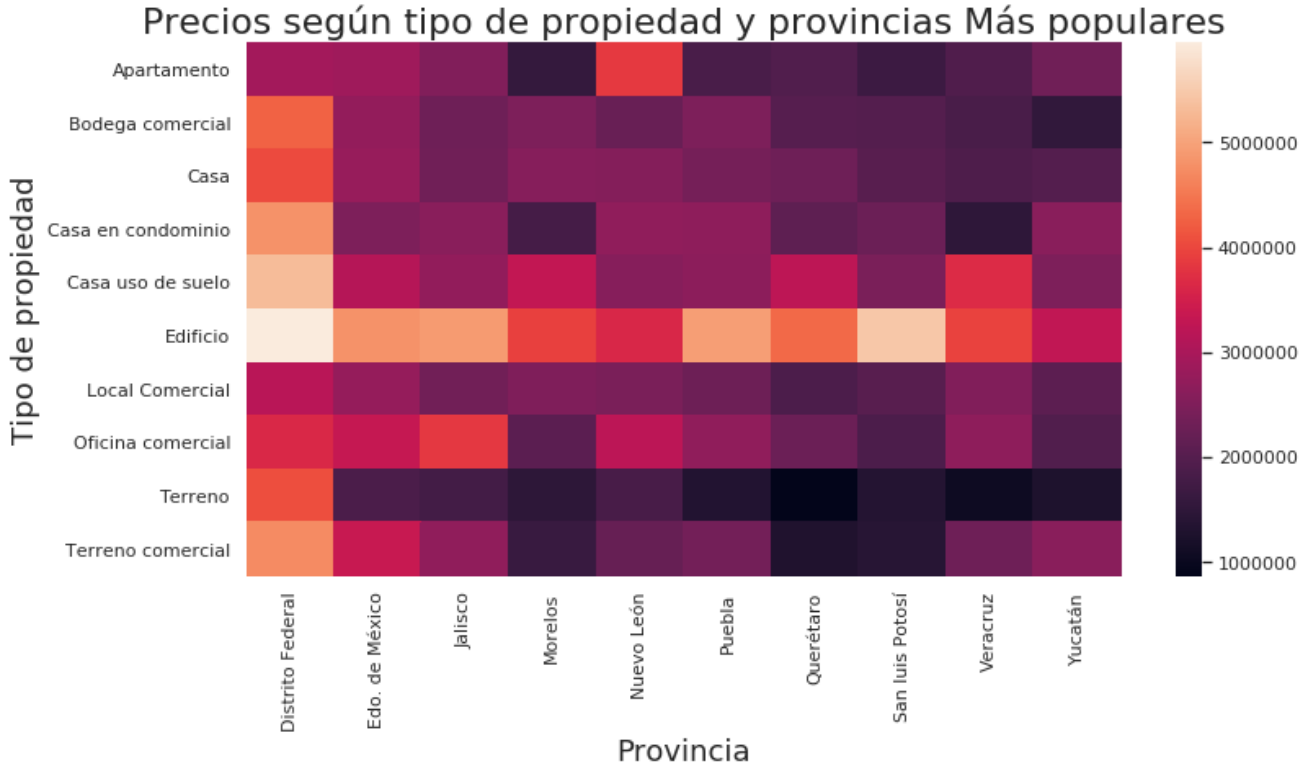
Con respecto a la última visualización claramente la variación de los precios son bastantes razonables de acuerdo al tipo de propiedad.

De todas maneras, es muy notable que las casas son en promedio más baratas que los departamentos y los terrenos más caros corresponden con zonas industriales.

Cabe aclarar que hay pocas publicaciones de “Garage” y “Duplex”. Si bien los que corresponden con “Garage” con pocas es bastante lógicas que sean de las publicaciones más baratas. Por otro lado, la poca cantidad de dúplex publicados quizás estén forzando un dato que no arroje la realidad acerca de su propio precio. Hay que tener cuidado con esta clase de datos.

Se decidió publicar esta sección porque si bien podría cometerse algún error para alguna clase de propiedad como lo es un dúplex los demás tipos de ambientes parecen tener sentido

4.1 - Tipo de Propiedades vs popularidad por provincias y precios



Esta clase de visualizaciones arrojan datos interesantes acerca de cuáles son aquellos estados que disponen en promedio de los mejores precios de acuerdo al tipo de propiedad. Sin embargo, aún con estos datos no podemos inducir ninguna conclusión especifica porque debemos recordar que tenemos un lapso de publicaciones de 5 años, devaluación, y algunos estados no tienen tantas publicaciones como otras. A simple vista podríamos decir que en Morelos se encuentran los apartamentos más baratos, pero quizás sea a razón de que encontremos solo un par de departamentos. Es entonces que no sería tan representativo como debería o se desea.

Igualmente se puede continuar arrojando lo siguiente: Los terrenos son más baratos en Querétaro y Veracruz, y los edificios más caros notablemente en el Distrito Federal, acaparando no solo ese sino otros tipos de propiedades con los precios más caros.

En lo que respecta a edificios se detalla notoriamente que sea cual sea el estado representa el tipo de propiedad más caro alcanzando picos en San Luis Potosí y el anterior mencionado Distrito Federal.

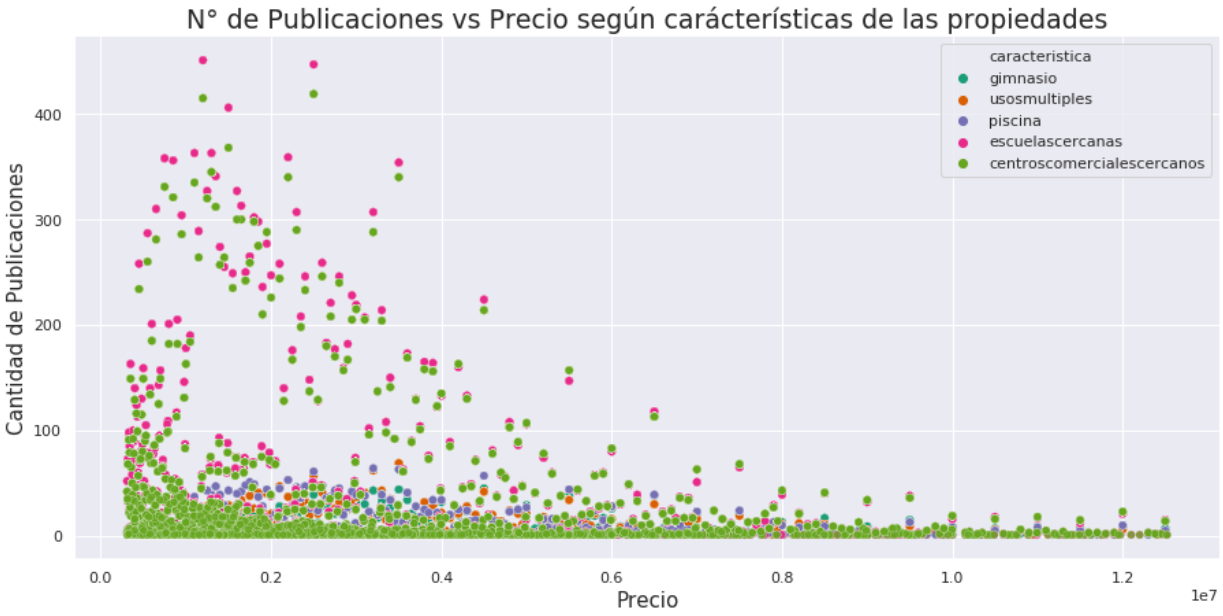
Lo mismo se podría decir para los terrenos en sentido contrario como el tipo de propiedad más barato en todos los estados que mantienen más publicaciones en contraposición con el precio promedio que presenta a nivel nacional. Sin embargo, vuelve a alcanzar una elevación notable del precio promedio en el Distrito Federal, el lugar con mayor demanda de propiedades y mayor cantidad de Publicaciones.

4.2 – Comportamiento según las características

Observando el set de datos vemos que hay varias columnas que indican si las propiedades tienen o no ciertas características, las mismas son “gimnasio”, “usosmultiples”, “piscina”, “escuelascercanas” y “centroscomercialescercanos”. Las mismas indican con 1 o 0 si la casa posee o no alguna de estas cualidades.

Como aclaración cabe destacar que para este análisis sólo nos interesan las propiedades que posean datos sobre estos campos, por lo que desestimamos las propiedades con campos nulos.

Se observa cómo se comportan las publicaciones y el precio de las propiedades de acuerdo a la posesión o no de alguna de estas características, veamos:



Vemos como las propiedades con centros comerciales cercanos y con escuelas cercanas tienen un comportamiento bastante parecido, esto también se debe a que sean las más populares, pues se observa la variabilidad de la cantidad de publicaciones sobre las mismas, estos colores abarcan casi todo el gráfico. Pero además son cualidades importantes a la hora de decidir sobre la adquisición de una propiedad, suponiendo que una familia con niños quiere mudarse a un nuevo hogar le interesará tener escuelas cercanas como también que es mucho más cómodo tener centros comerciales cercanos pues eso supone más accesibilidad, ya sean mercados, lugares de recreación, servicios, etc. Hay baja cantidad de publicaciones con elevada variación de precios.

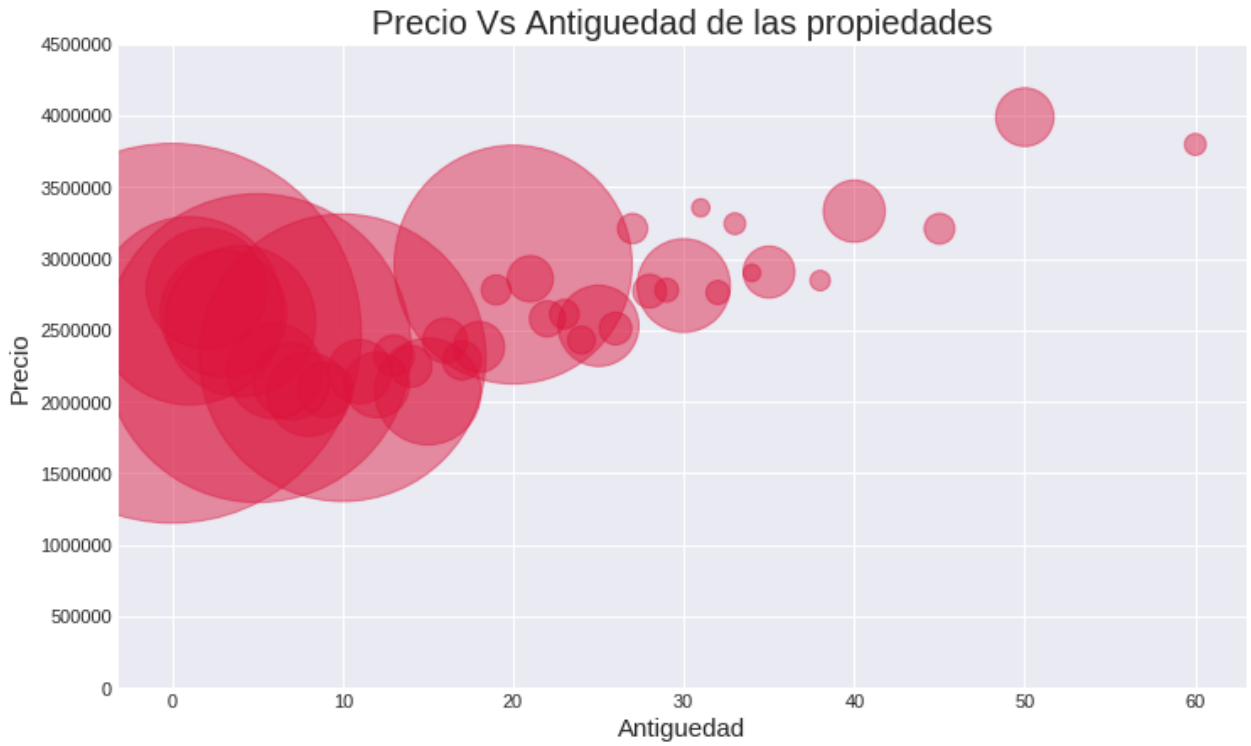
A medida que la cantidad de publicaciones aumenta, el precio comienza a disminuir, por lo que parecen ser mucho más populares las publicaciones de propiedades que posean estas características y se encuentren a un precio más accesible.

4.3 – Relación entre Precio y Antigüedad

Existe un campo que nos indica la antigüedad de las propiedades, el mismo se representa con un valor numérico, que indica los años de antigüedad, que va desde 0 años para las propiedades nuevas hasta 80 años para las más viejas, que es el valor máximo registrado.

Es interesante observar como se comporta el precio en relación a esta variable. Se buscan respuestas a: ¿qué propiedades son más caras? ¿serán las nuevas o las más antiguas?

Se cuenta la cantidad de publicaciones por cada valor de antigüedad, se realiza teniendo en cuenta el filtrado de los valores que tienen pocas publicaciones, para evitar caer en "la ecuación más peligrosa de la historia", luego volcamos nuestros datos en una visualización para poder entenderlos mejor, veamos el resultado.



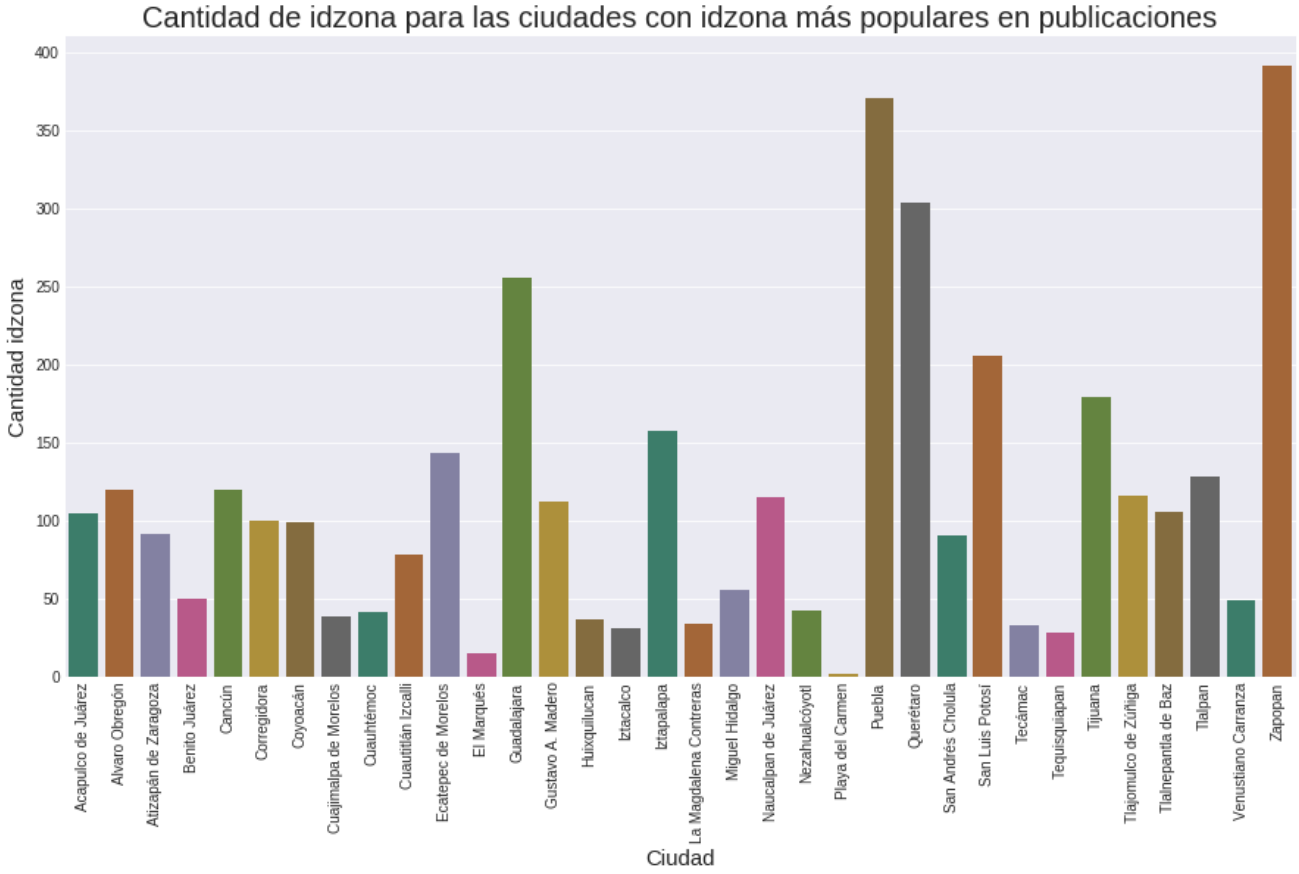
Las propiedades con una antigüedad entre 0 y 10 años tienen una gran cantidad de publicaciones registradas y precios variados que se encuentran entre 1 y 4 millones de pesos mexicanos. Pero si se observa bien, se tiene un aglutinamiento de propiedades que se encuentran por debajo de los 20 años de antigüedad, cuyo precio se encuentra entre 1 y 3 millones.

Si se traslada el análisis a propiedades que se encuentran entre los 15 años hasta casi los 30 años de antigüedad aproximadamente, el precio varía menos, debido a que tiene menor cantidad de publicaciones y se encuentra entre un rango de 2 a casi 4 millones de pesos mexicanos.

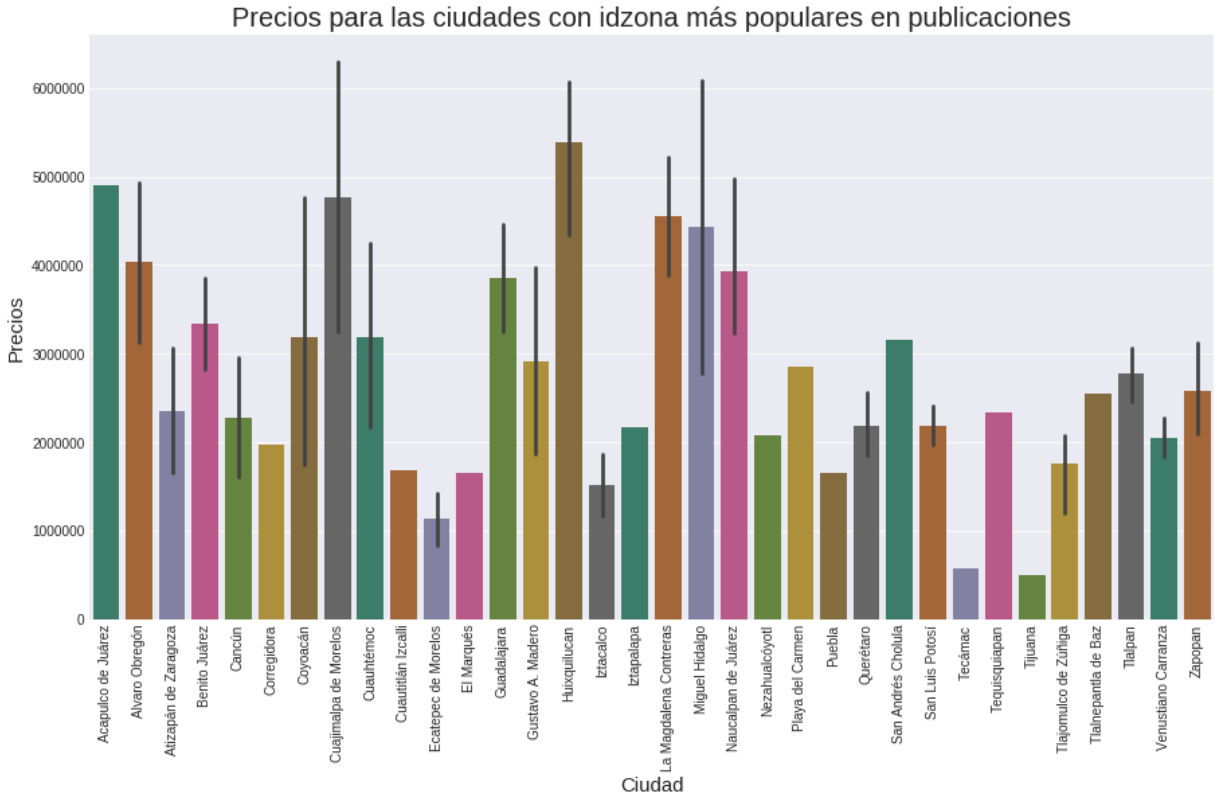
A medida que se superan los 30 años de antigüedad el precio aumenta, sin bajar de los 2.5 millones aproximadamente. También hay propiedades que rondan los 50 años de antigüedad cuyos precios son elevados.

También tenemos que el número de publicaciones disminuye, por lo que las casas con mayor antigüedad son menos populares entre las publicaciones.

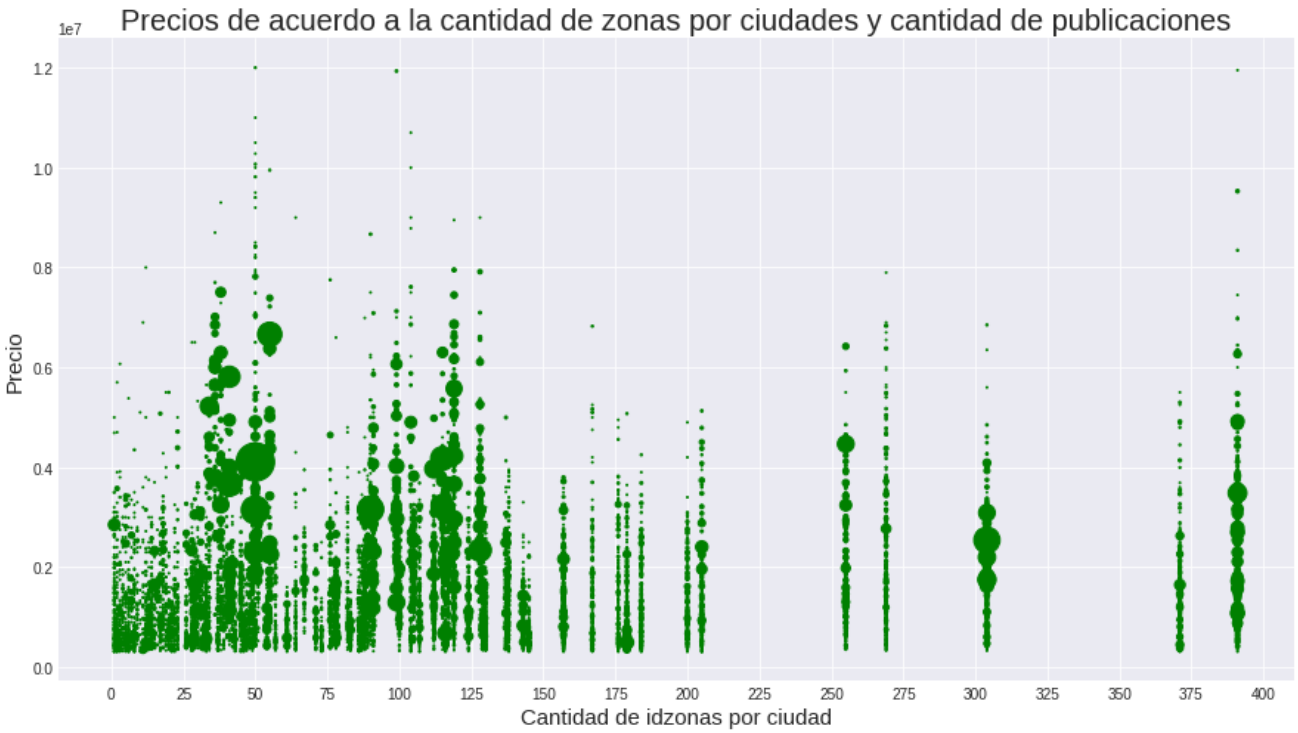
4.5 – Relación entre precios y Id de zona



Veamos ahora qué sucede con el precio para estas ciudades.



A simple vista parece que las ciudades con mayor cantidad de id zonas no escalan precios tan altos. Es probable que tenga que ver con la cantidad de publicaciones y la variabilidad de los datos dado el volumen de publicaciones.



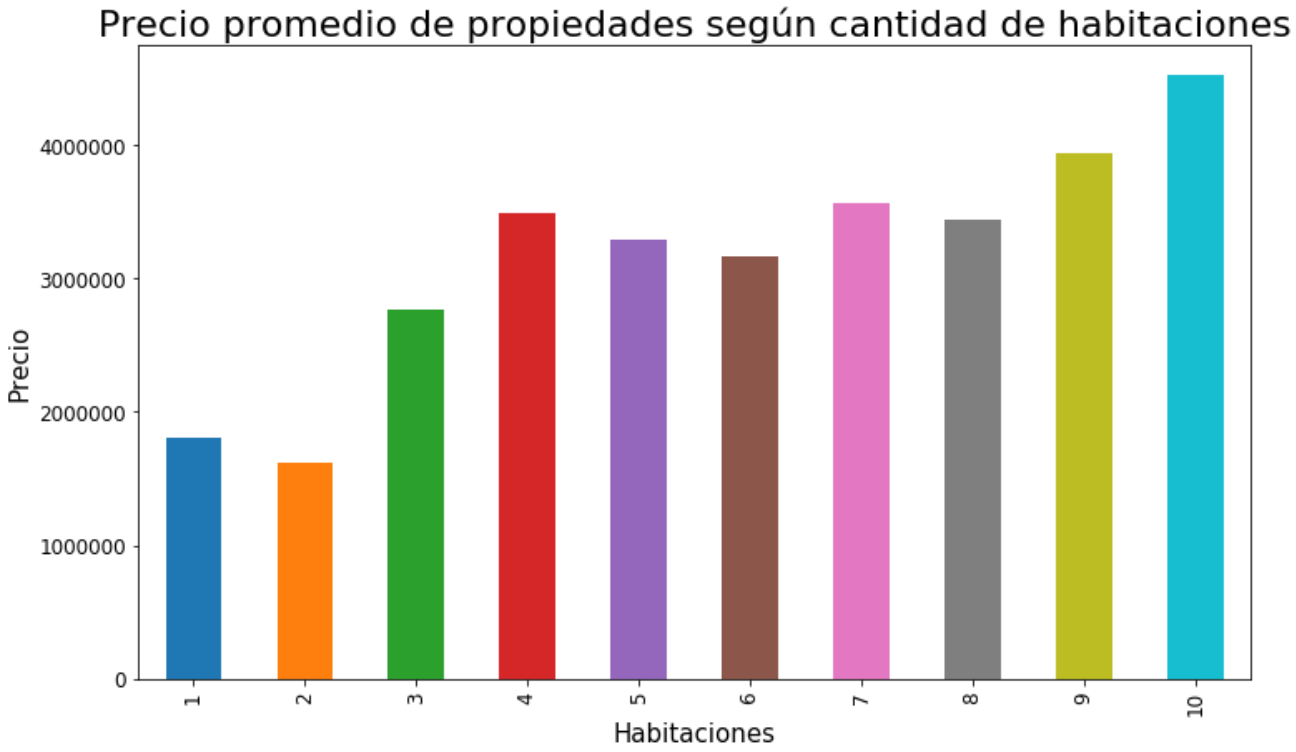
No parece mostrarnos nada interesante sobre los datos. La cantidad de publicaciones según la cantidad de idzonas difiere en valores para cantidades de idzona por ciudad.

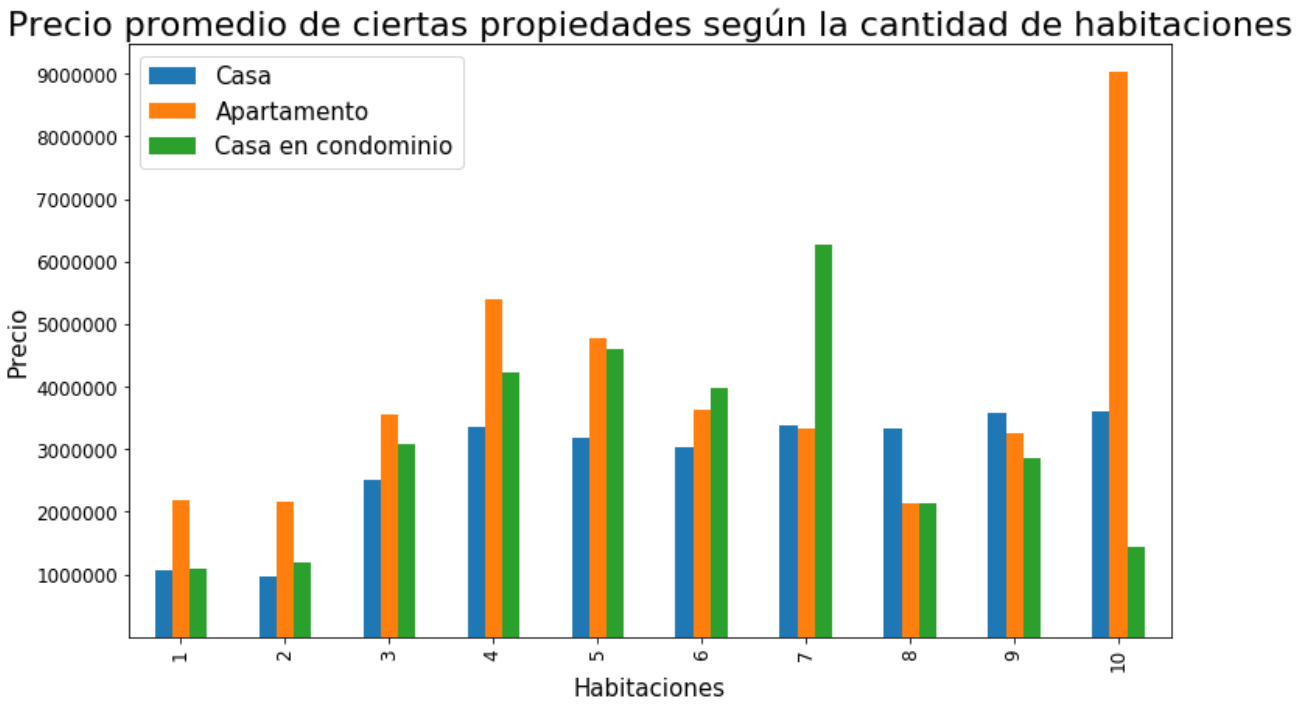
No se observa una relación clara con el precio, pero sí se observa que tiene mayor cantidad de publicaciones en precios debajo de los 8 millones, y más aún debajo de los 4 millones. Esto surgió en otros análisis específicos, por lo que no es considerado importante para arrojar nuevos datos.

4.6 - Relación entre precio y los distintos ambientes

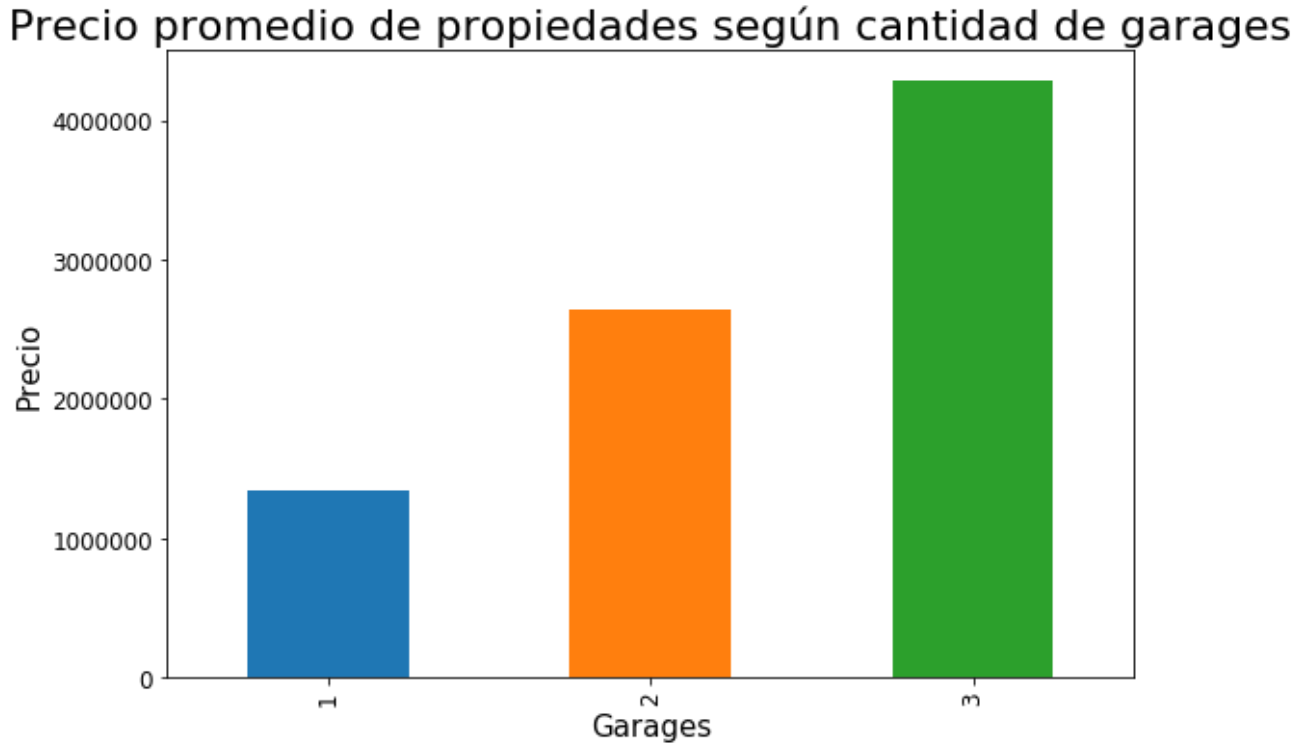
Con el objetivo de comparar los precios de las propiedades con ciertas características particulares, especialmente algunas en las que se observó que había una correlación positiva en la matriz de correlaciones que se encontraba en el notebook "metaDataNavent". Así, se comparan los promedios de los precios de las propiedades según los distintos valores posibles que podían tomar las variables consideradas. El análisis, por un lado, para todos los tipos de propiedades y por otro comparando tres tipos de propiedades específicos (los tres con mayor cantidad de publicaciones) que son Casa, Apartamento y Casa en condominio.

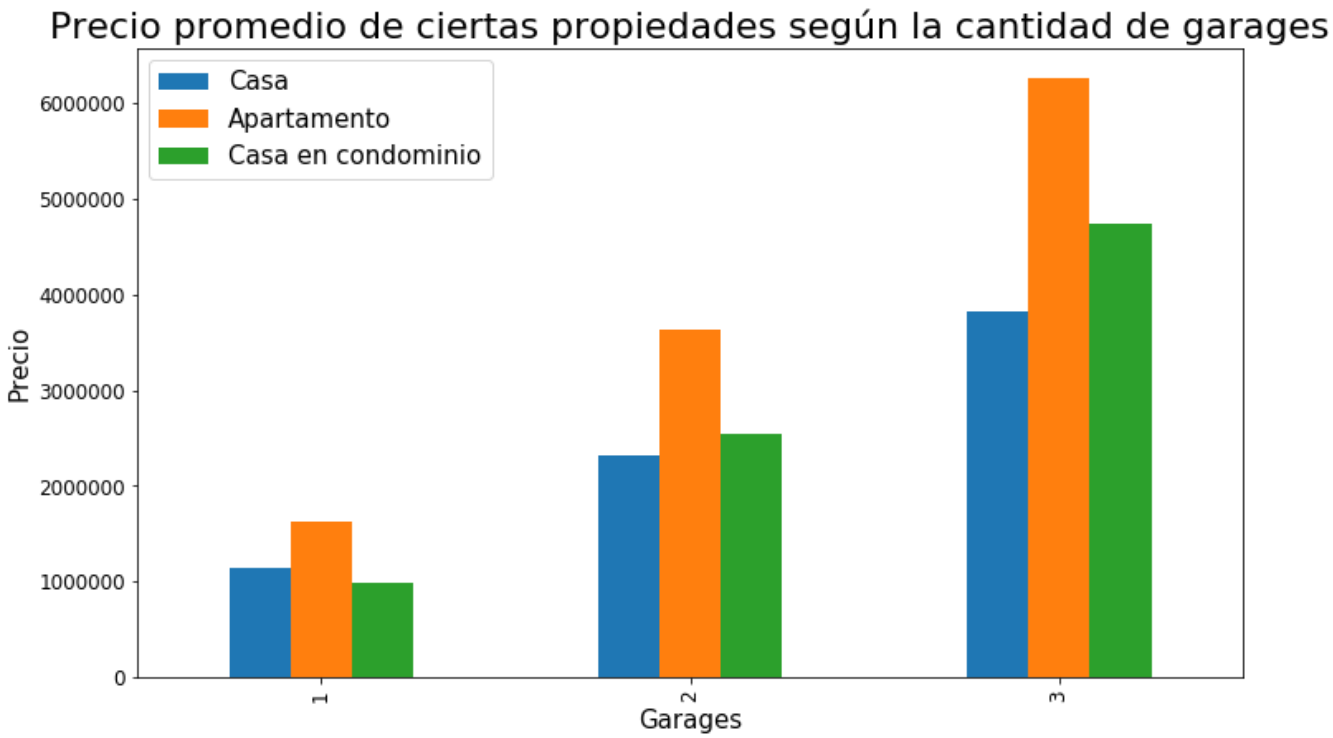
Primero visualizamos como cambia el precio según la cantidad de habitaciones.



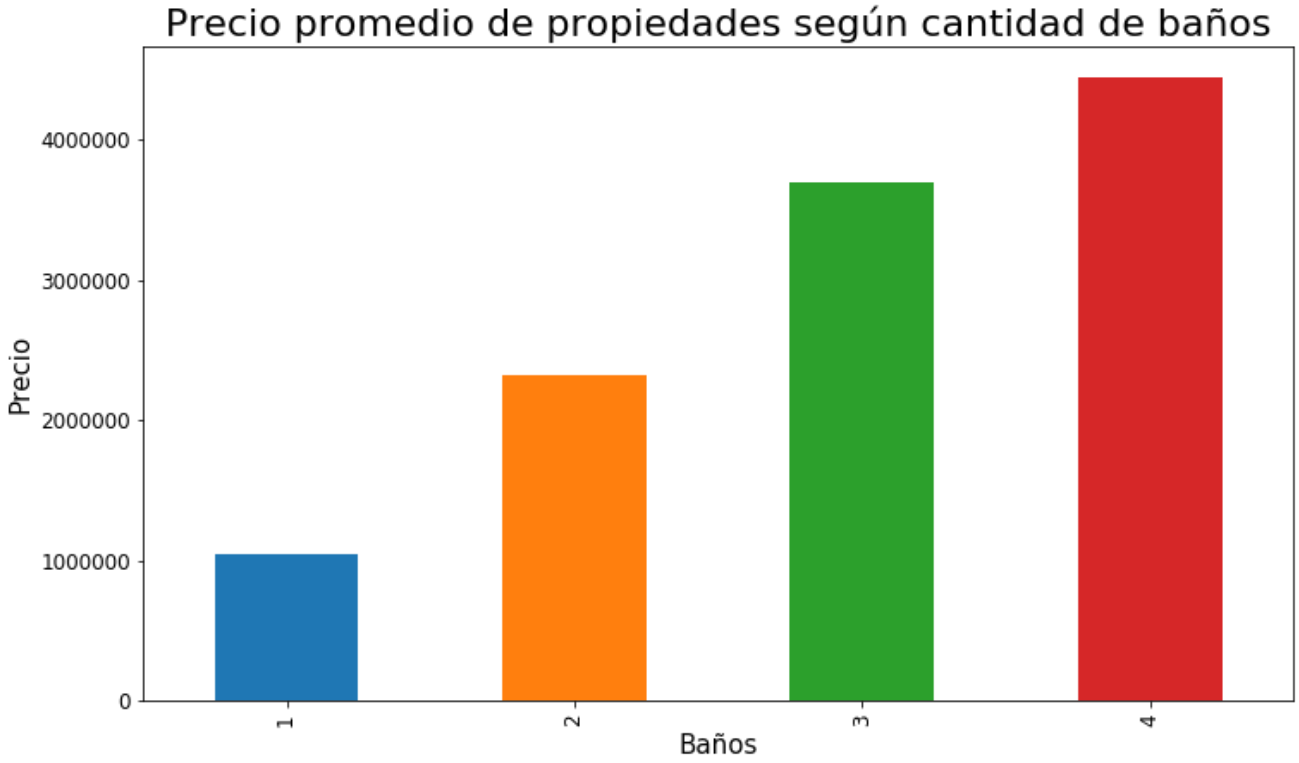


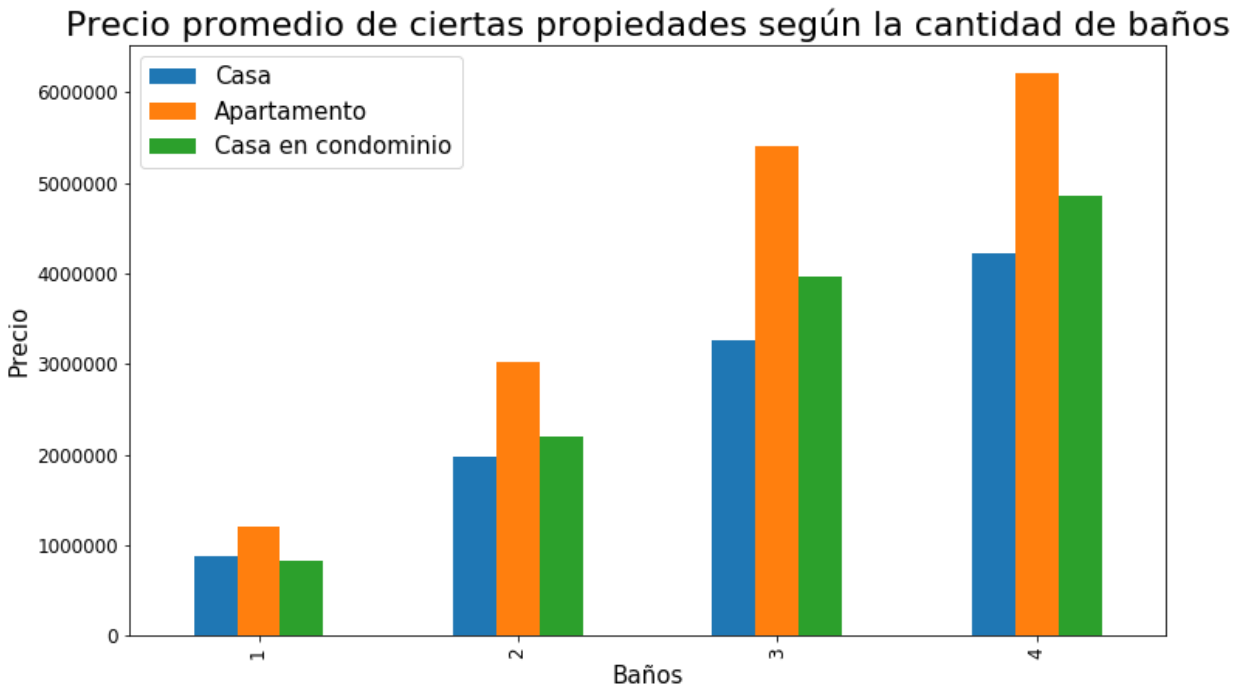
En estos gráficos vemos en general un aumento de precio cuantas más habitaciones hay, lo cual sería lógico, aunque se observan fluctuaciones como una disminución entre 4 y 6, lo que muestra que este aspecto por sí solo no determina el precio de una propiedad. Para las comparaciones entre los 3 tipos más comunes, vemos que las casas mantienen cierta estabilidad. En apartamentos hay un pico muy elevado para la cantidad máxima de habitaciones, y las casas en condominio tienen una disminución en el final.



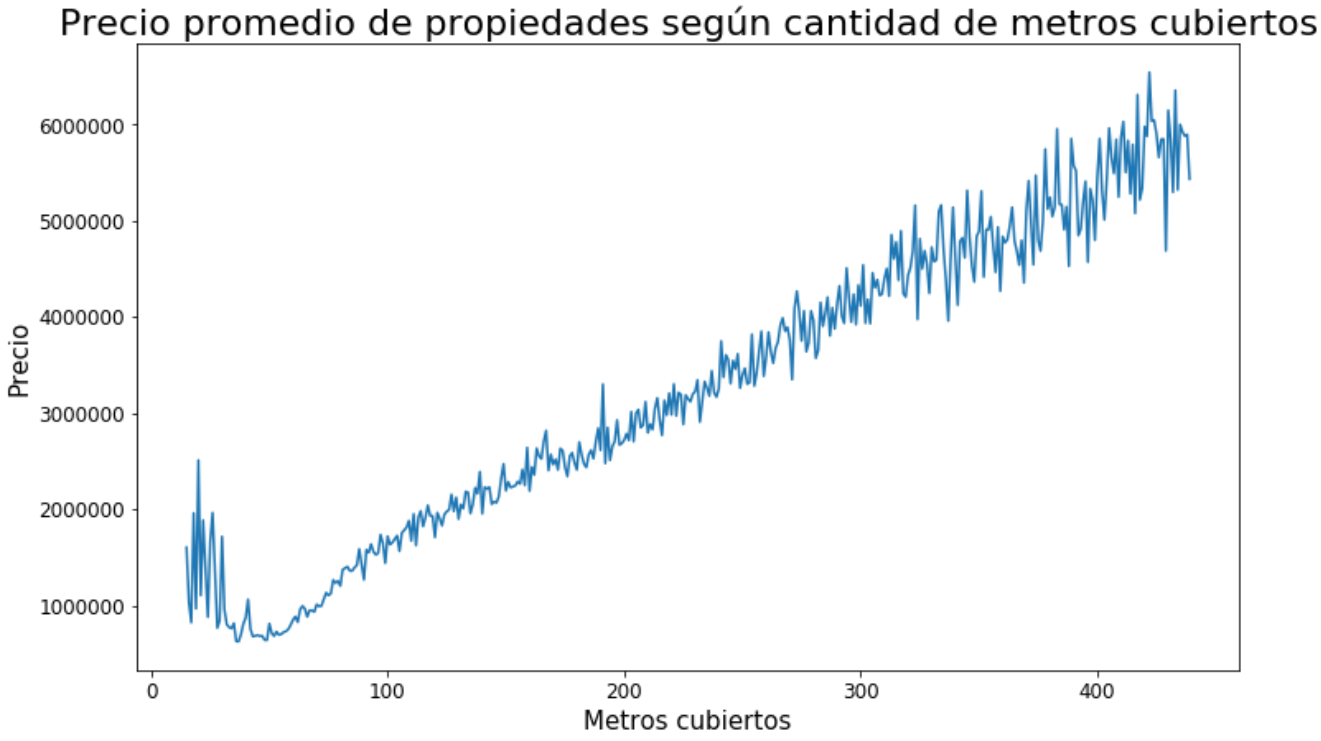


Se observa un notable aumento de los precios conforme más cantidad de garages, cuya relación también se mantiene individualmente para los tipos particulares de propiedades; especialmente se ve un gran aumento en los apartamentos.

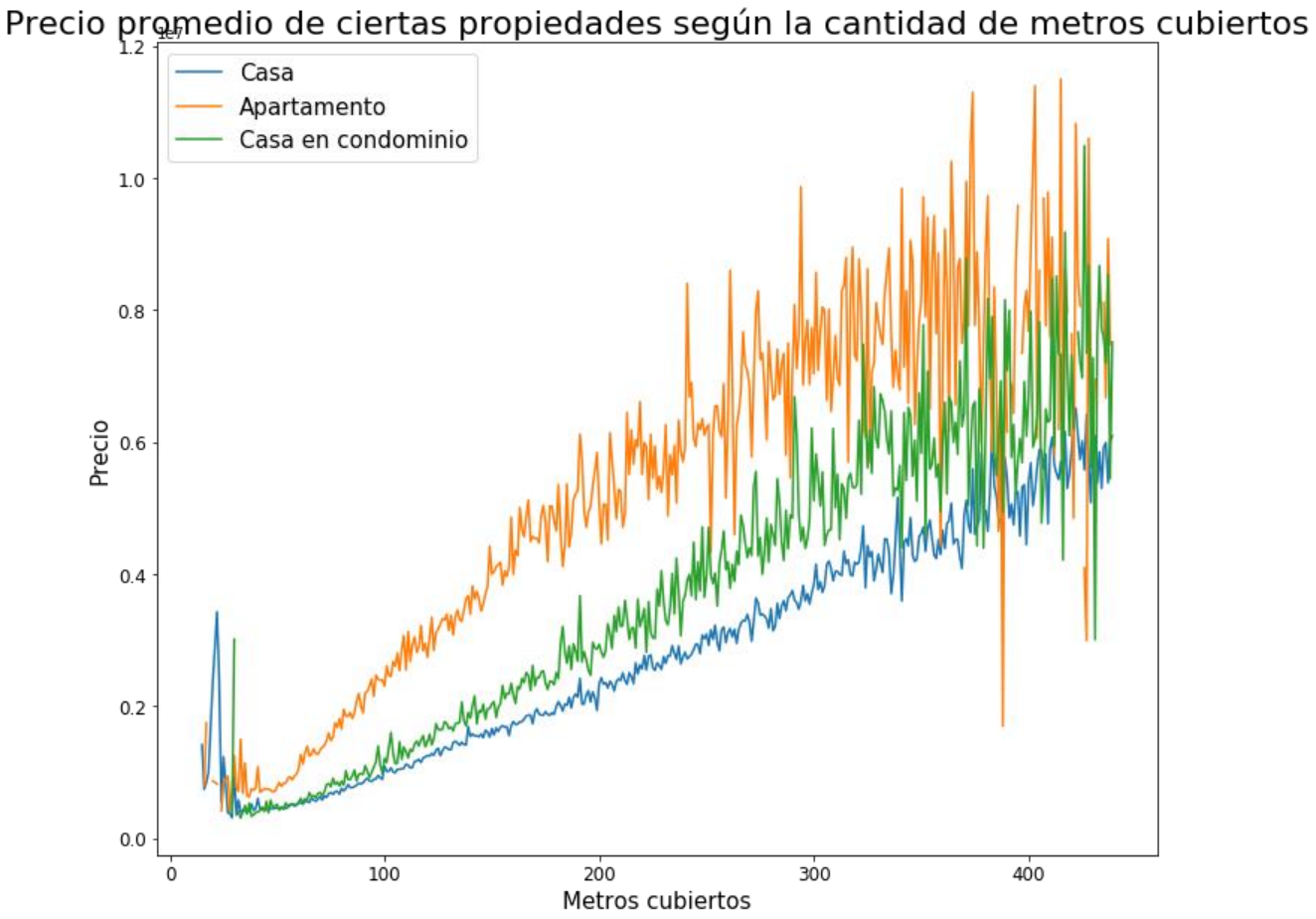


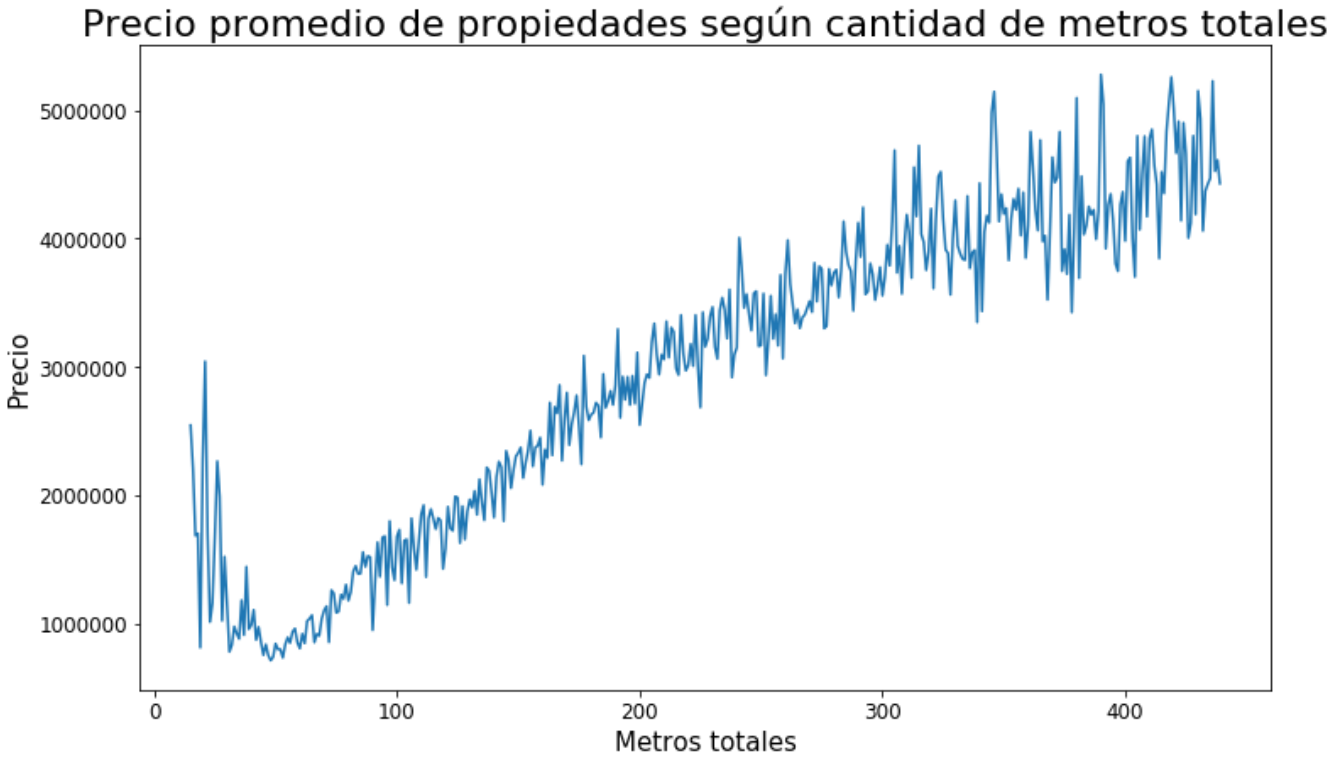


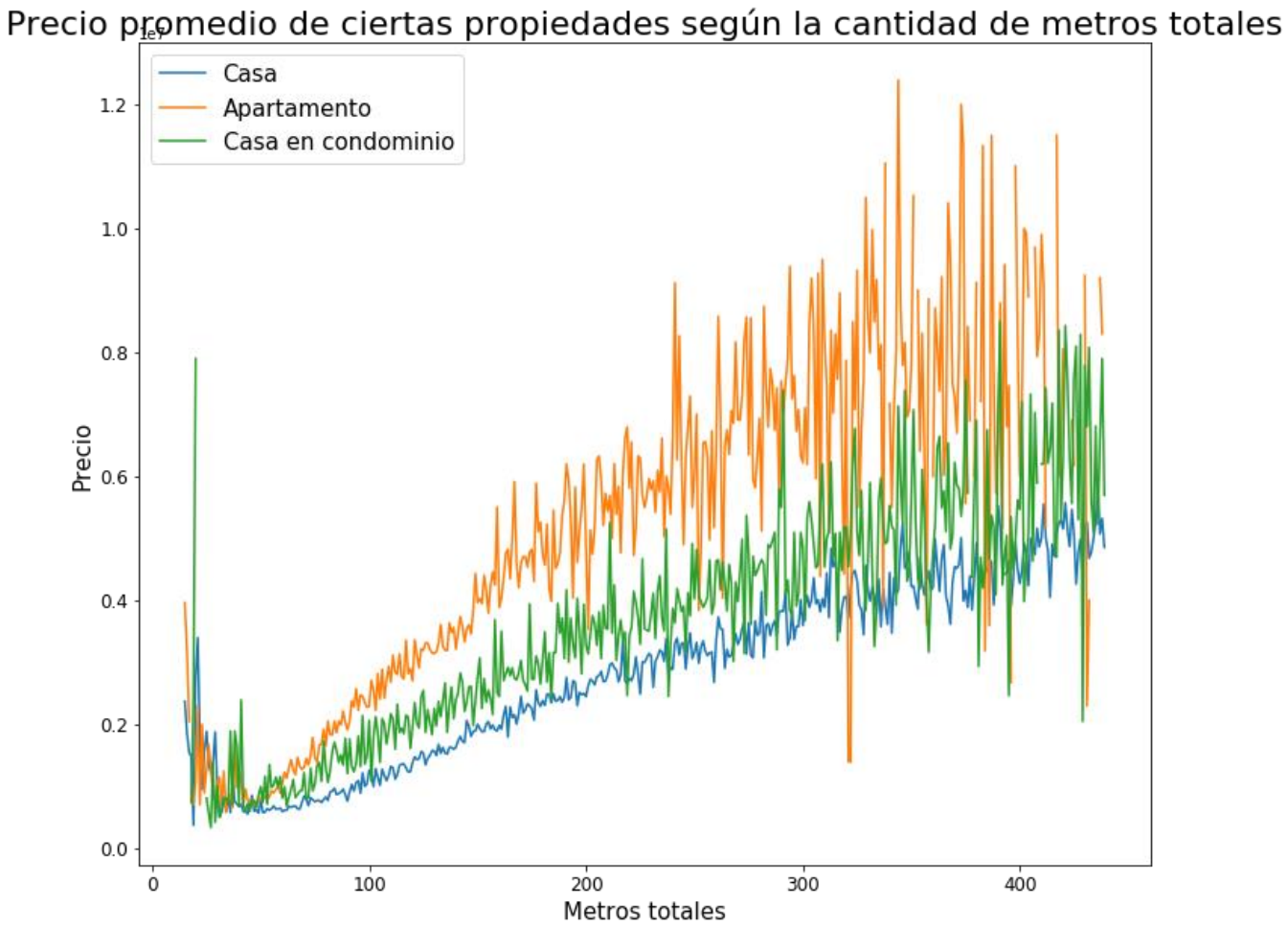
Para la cantidad de baños, la tendencia es la misma que con la cantidad de garages, lo que muestra que estas dos características son más específicas y revelan más sobre la clase de propiedad sobre la que se está tratando.



D

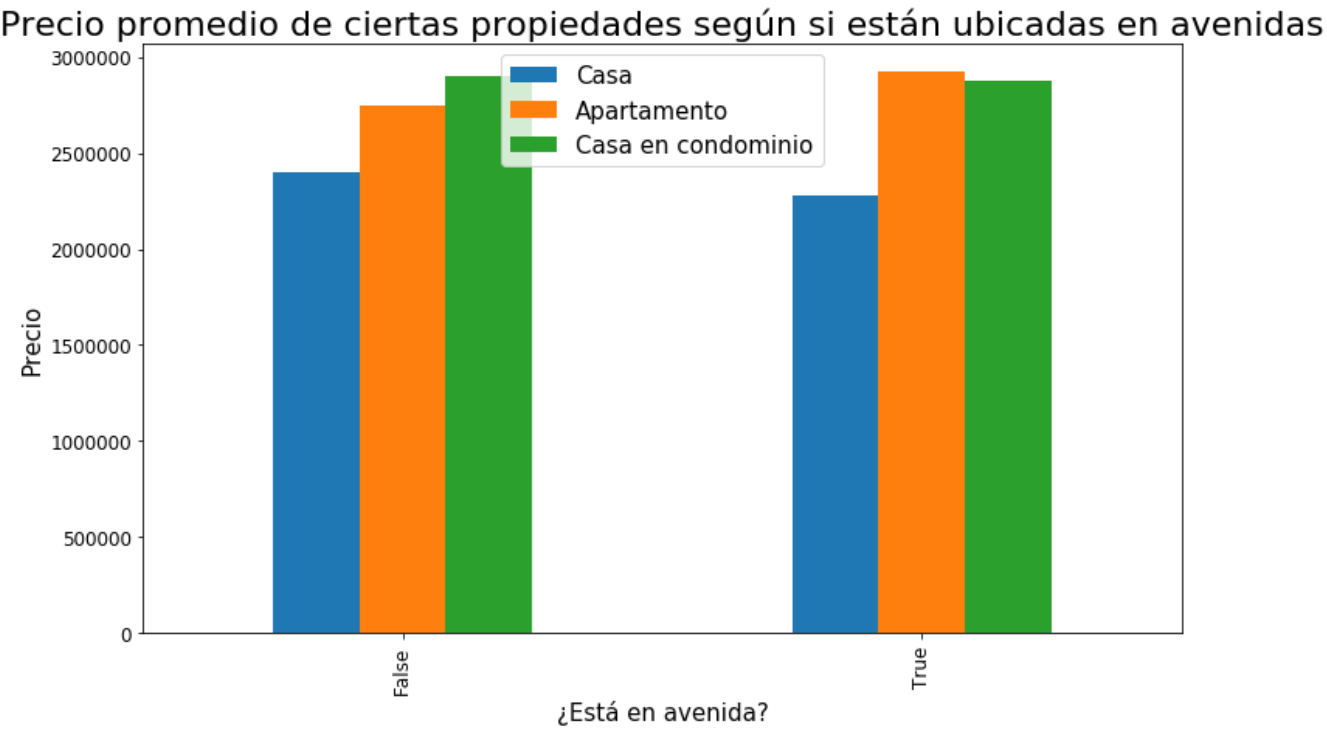

Precio promedio de ciertas propiedades según la cantidad de metros cubiertos
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Precio
Casa
Apartamento
Casa en condominio
0
100
200
300
400
Metros cubiertos


Precio promedio de propiedades según cantidad de metros totales
5000000
4000000
3000000
2000000
1000000
Precio
0
100
200
300
400
Metros totales



En el caso de metros totales y cubiertos, las tendencias son muy similares, con un poco más de ruido en los gráficos de metros totales. Para el caso general hay un pequeño ruido en el comienzo y luego aumenta el precio a medida que los metros se incrementan. Tiene una tendencia lineal de crecimiento y aumenta el ruido sobre el final.

Para las comparaciones por tipo de propiedad, el ruido es mayor en el principio (sobre todo en las casas en condominio para metros totales), y luego, también hay una tendencia lineal de crecimiento, pero en relación con las casas en condominio, y sobre todo para los apartamentos, la cantidad de ruido sobre el final es bastante grande. Todo esto muestra que la cantidad de metros influye razonablemente en el precio, sobre todo los metros cubiertos, pero sobre los extremos hay otros aspectos que pueden influir y marcar una diferencia.



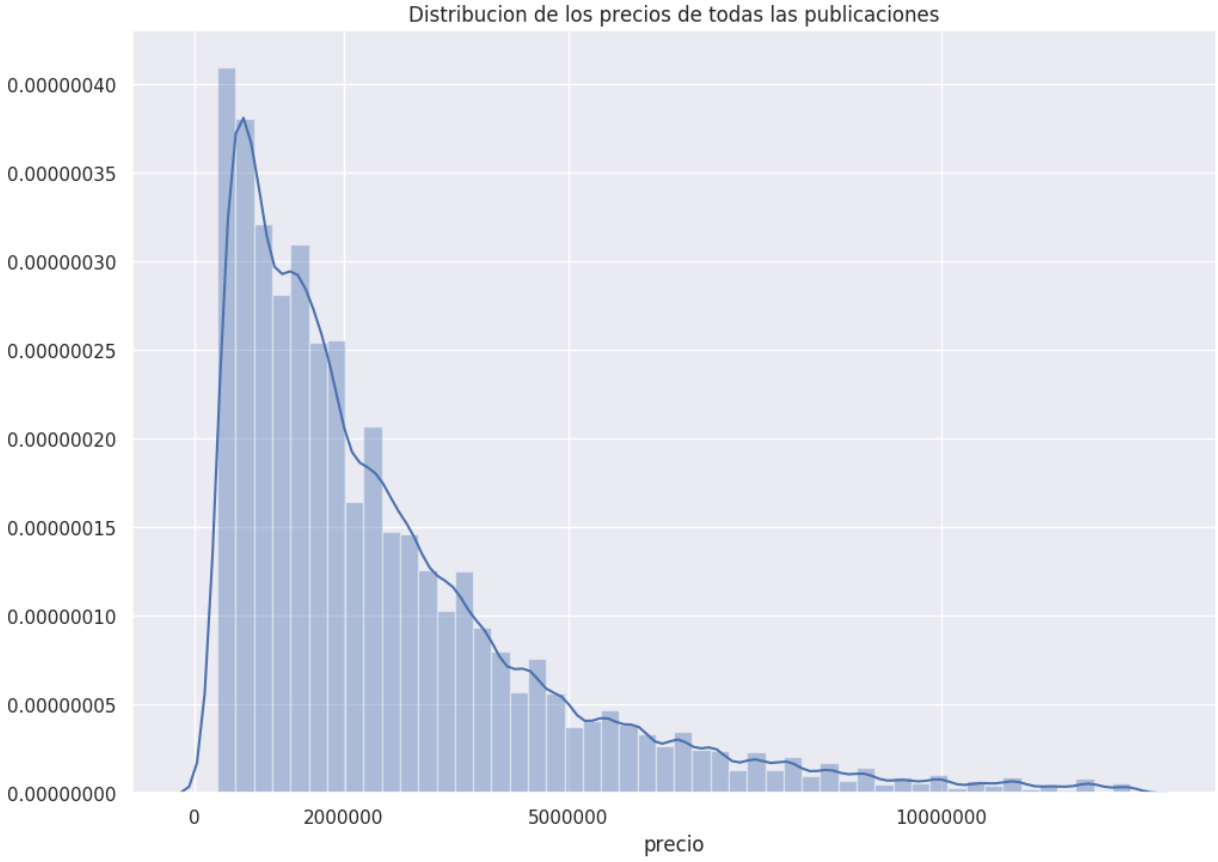
Finalmente, se hizo este análisis sobre los precios en función de si las propiedades estaban ubicadas o no en avenidas, aprovechando la columna con el dato de la dirección, y suponiendo a priori que las que efectivamente están sobre avenidas son más caras que las que no. El resultado general fue que el promedio de precio de las que si están ubicadas en avenidas es levemente superior de aquellas que no. Sin embargo, en los análisis particulares se ven distintas cosas. Para las casas en avenidas hay una disminución, para los apartamentos un aumento y para las casas en condominio es prácticamente igual. De esta manera, se ve que el factor de la ubicación puede influir, pero no será determinante sobre el precio final.

SECCIÓN V - Análisis sobre Títulos y Descripciones

Esta parte del análisis exploratorio se basa en una función implementada en python que, dada cierta Serie de pandas de tipo string, cuenta todas las palabras aparecidas en toda la Serie y las veces que apareció en total. Con esto puedo listar todas las palabras que aparecieron en, por ejemplo, la columna "Descripcion" de nuestro set de datos, y además tener la cantidad de veces que apareció cada una.

5.0 - Análisis por grupos de precios

Se procede a observar la distribución de los precios de todas las publicaciones

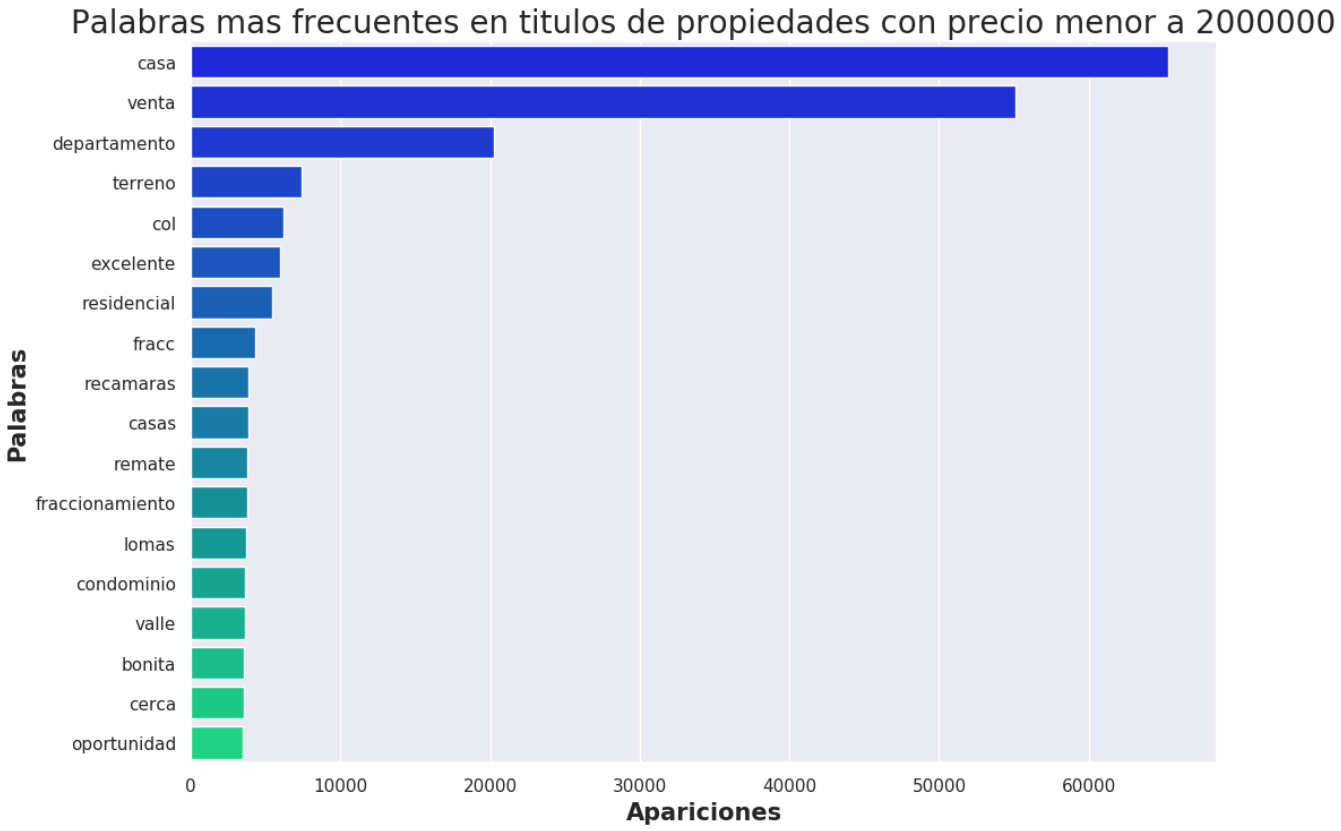


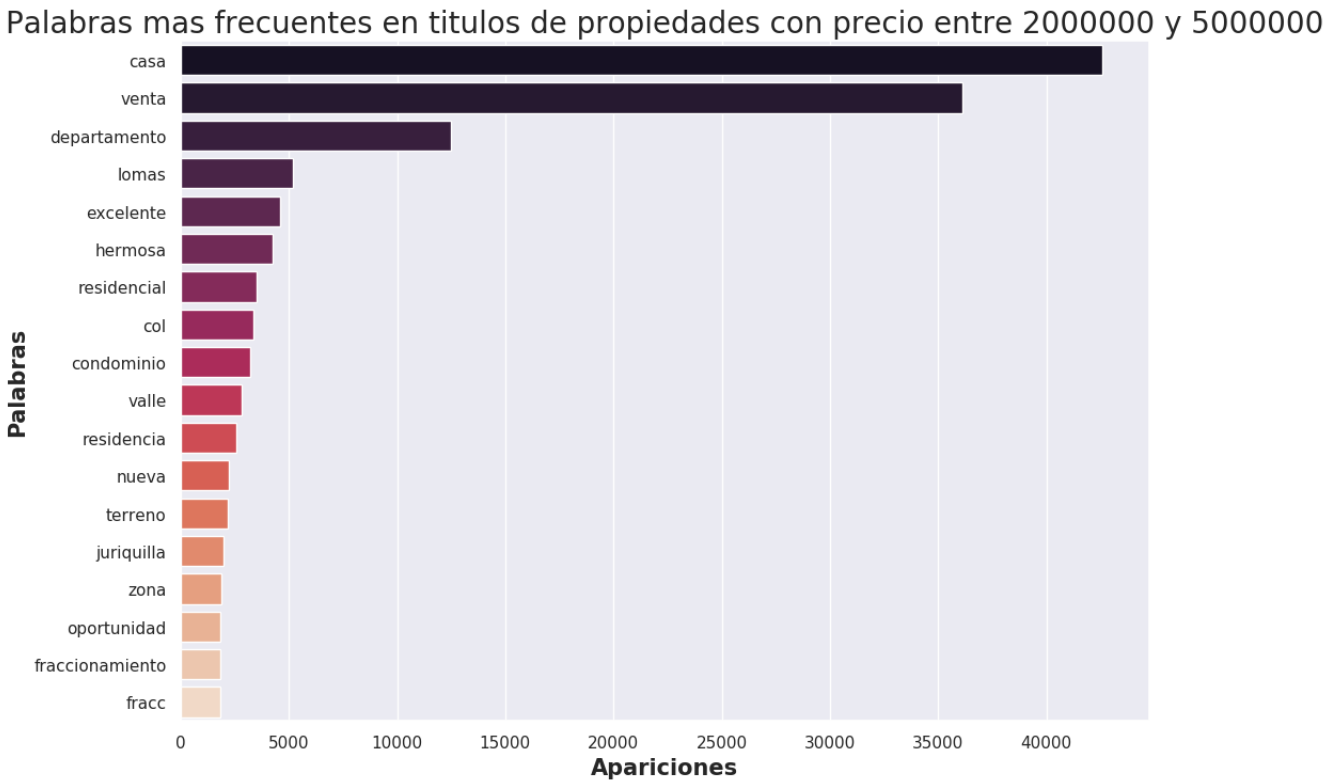
En esta parte del análisis se divide arbitrariamente las publicaciones en 3 grupos:

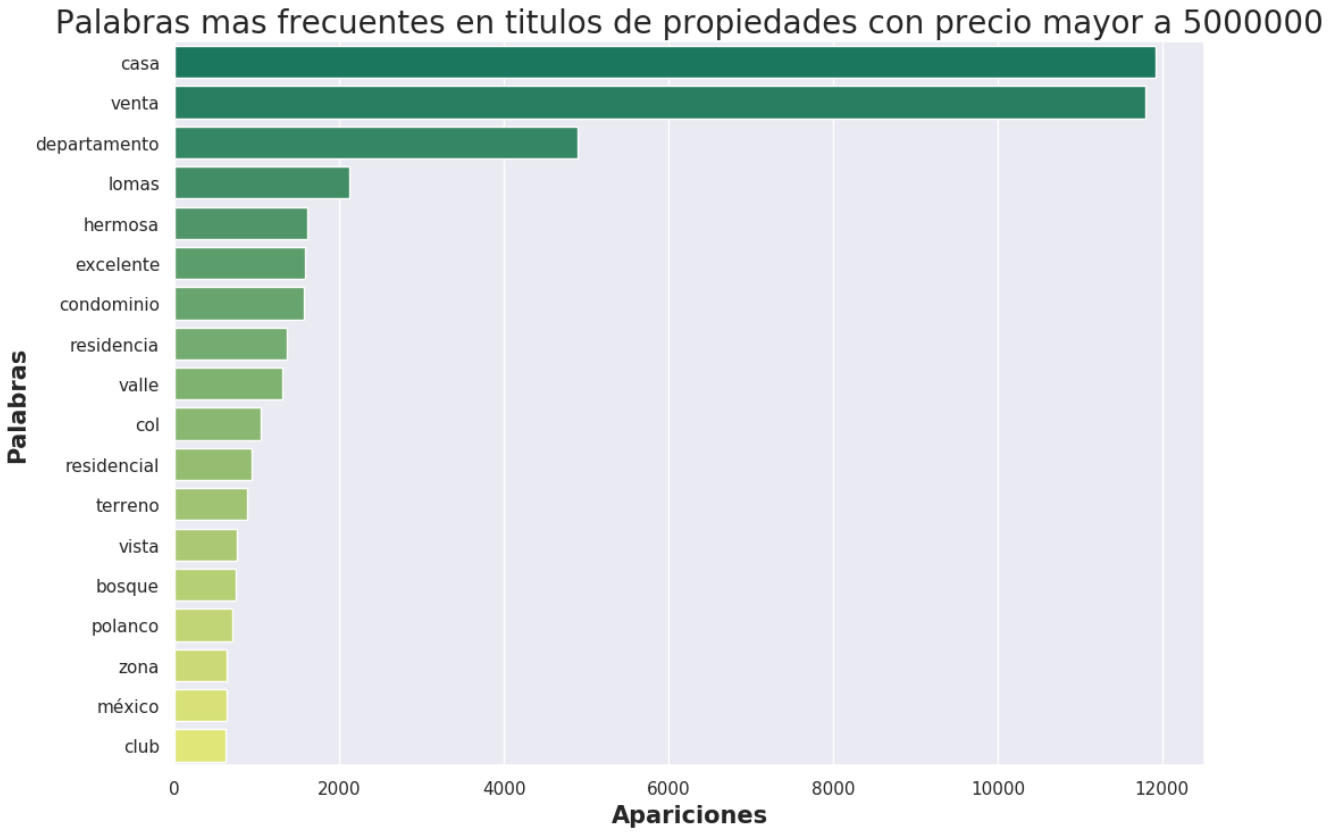
Grupo 1: Precio menor a 2000000.
Grupo 2: Precio entre 2000000 y 5000000.
Grupo 3: Precio mayor a 5000000.

5.1 – Palabras más populares por grupos según precios

Contando todas las palabras que aparecieron en cada grupo de precios, se puede plotear las que estuvieron en el top18 de cantidad de apariciones y sacarse conclusiones interesantes.







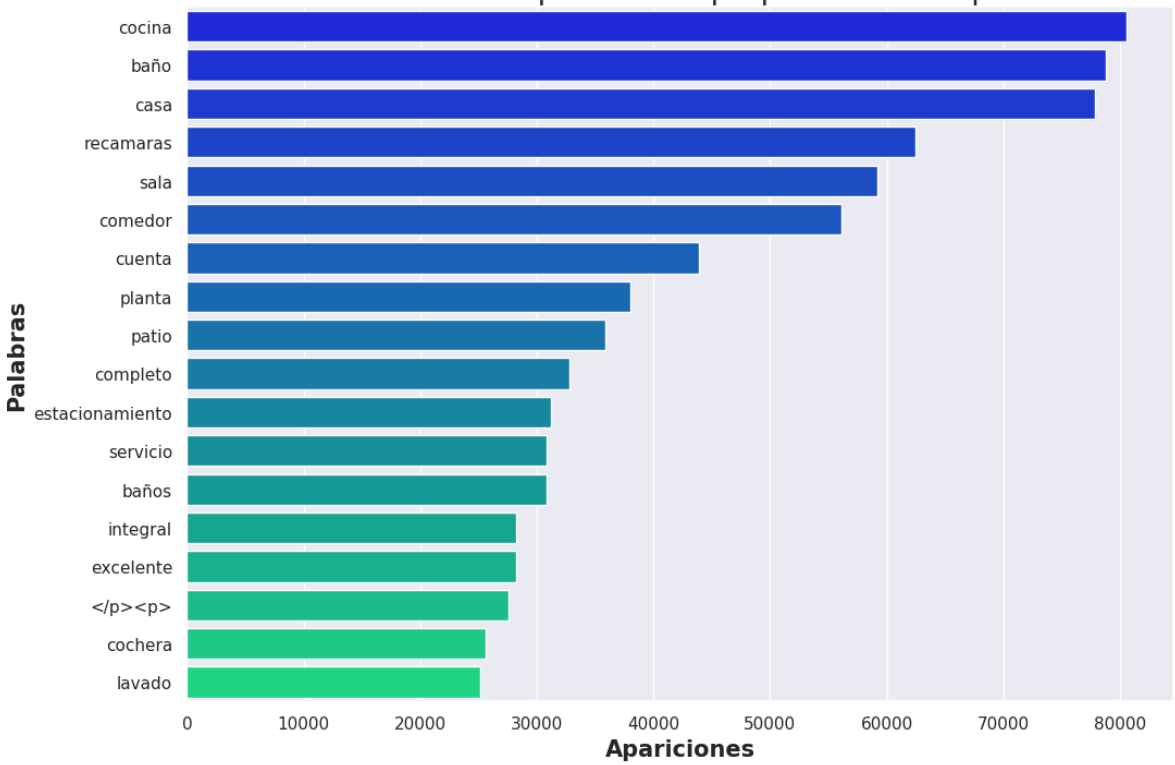
Las palabras “casa”, “venta” y “departamento” aparecen con elevada frecuencia en todos los grupos. Con este método no se ven diferencias demasiado significativas entre las palabras de los títulos.

5.2 – Palabras más importantes

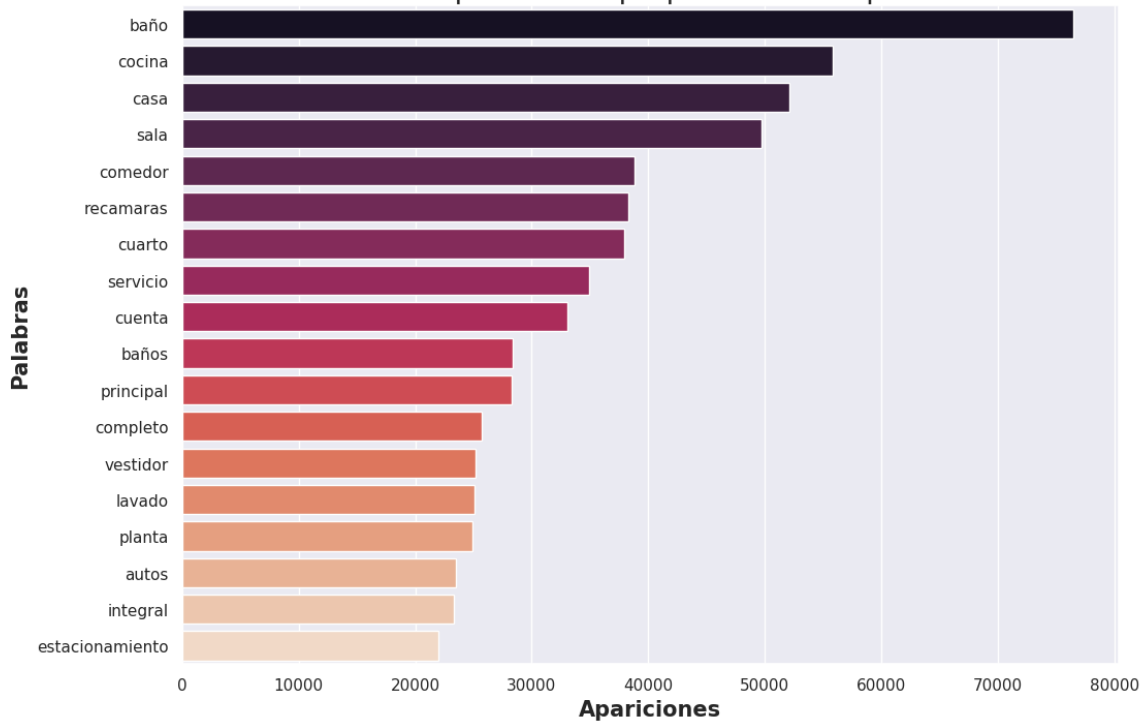
¿Cuales son las palabras que más aparecen en las descripciones? ¿Cambian según el grupo?

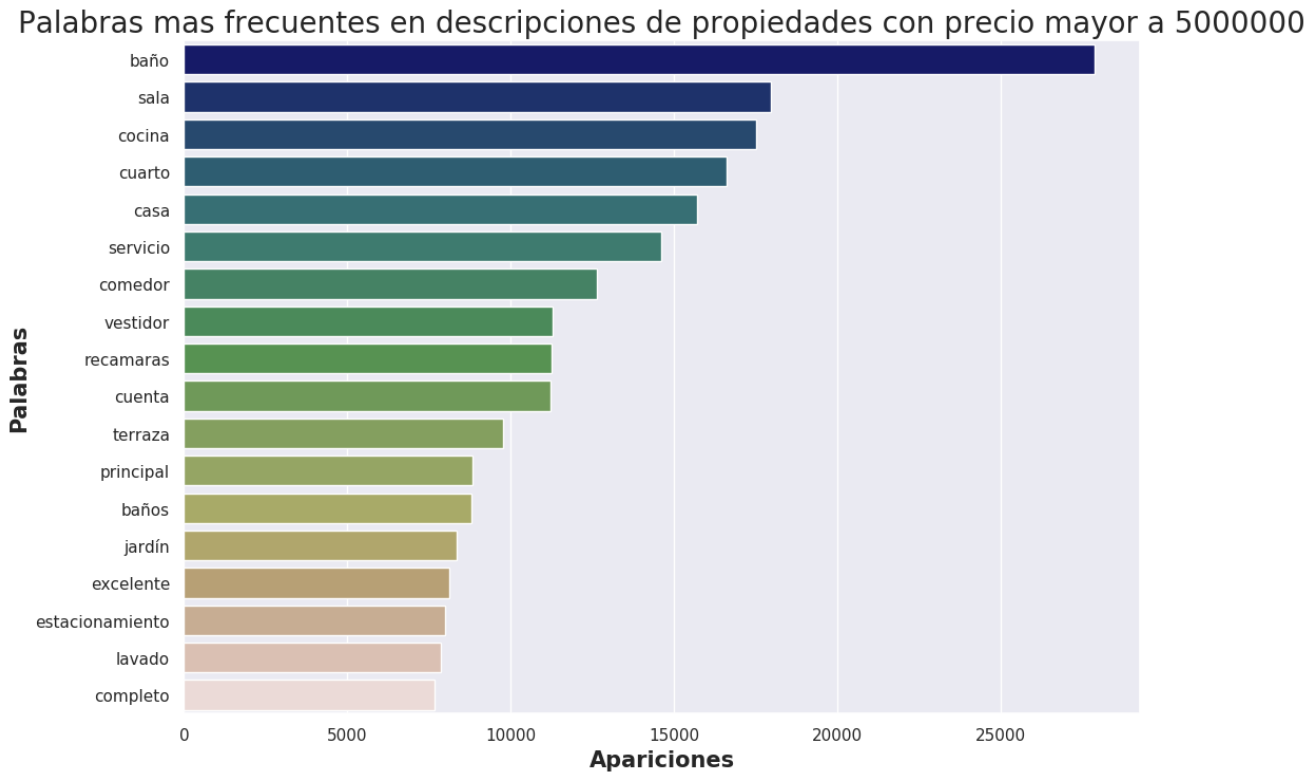
Haciendo el mismo análisis que el anterior pero para descripciones, se intenta buscar diferencias interesantes

Palabras mas frecuentes en descripciones de propiedades con precio menor a 2000000



Palabras mas frecuentes en descripciones de propiedades con precio entre 2000000 y 5000000





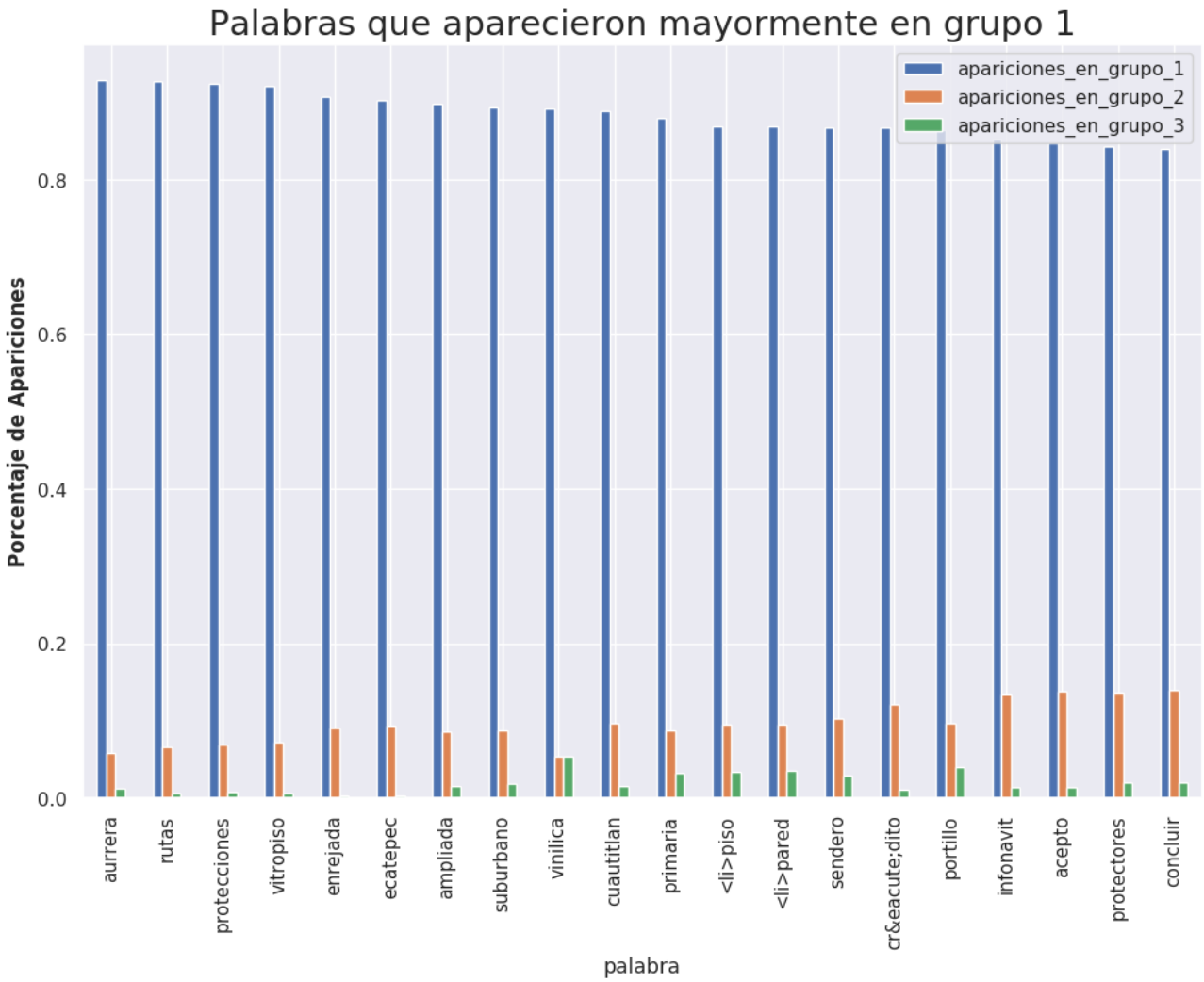
Se puede observar que:

- La palabra 'baño' aparece más o menos en casi todas las propiedades sin importar el precio.
- Ídem para las palabras 'excelente' 'baños' 'servicio' 'cuenta' y 'lavado'.
- Las palabras 'cocina' y 'casa' aparecen mucho más en propiedades de precio menor a 2000000
- La palabra 'jardín' y 'terraza' aparece más seguido en propiedades de precio mayor a 5000000.
- Las palabras 'patio' 'cochera' solo en las de precio menor a 2000000.
- Las palabras 'vestidor' y 'cuarto' no aparecen significativamente en las de precio menor a 2000000 y si en el resto.
- La palabra 'estacionamiento' ' no tiene importancia en las de precio mayor a 5000000.

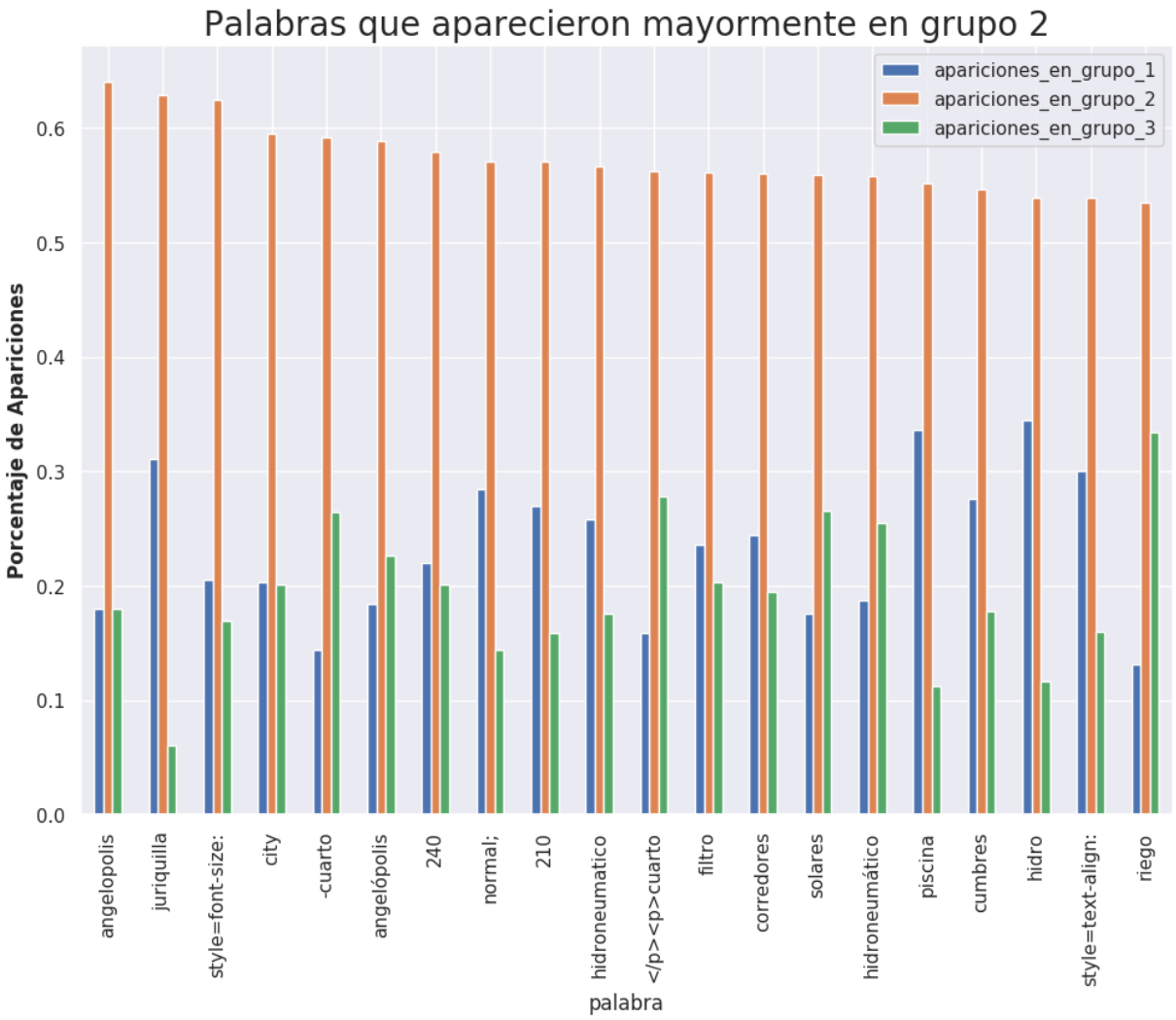
5.3 – Palabras características por grupos

Este análisis se centra en las palabras que aparecieron en las descripciones más de 500 veces en total.

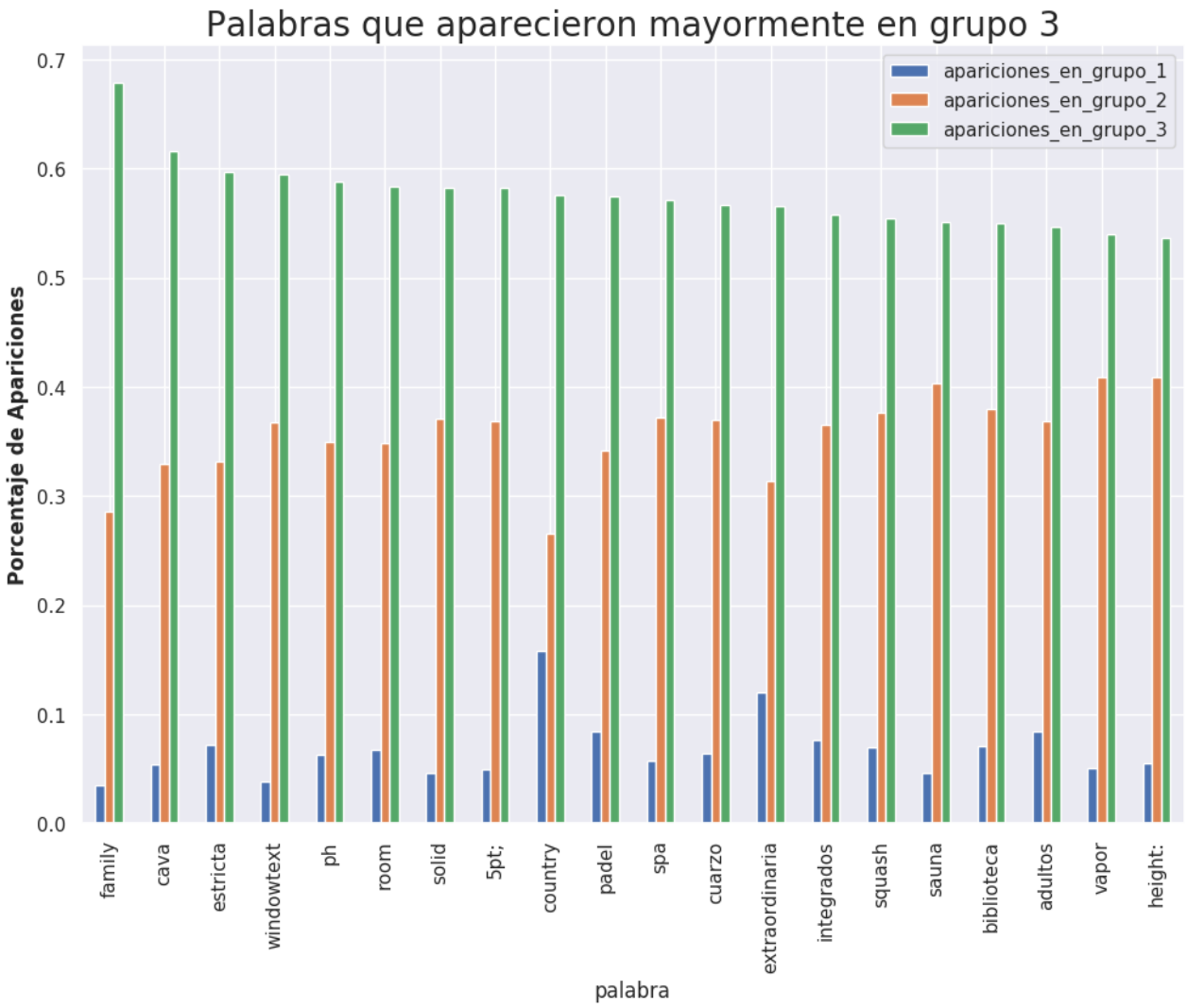
Para esas palabras se calcula cual fue el porcentaje de veces que aparecieron en cada grupo y realizar un gráfico de cuáles fueron las que tuvieron mayor porcentaje por cada grupo.



Palabras como 'vitropiso' 'suelo' 'ampliado' 'suburbano' aparecen significativamente más veces en el grupo 1'



En el grupo 2 se encuentran barrios como 'juriquilla' y 'angelopolis' ademas de 'hidroneumatico' (un tipo de tanque de agua) 'cumbres' 'riego' y 'filtro



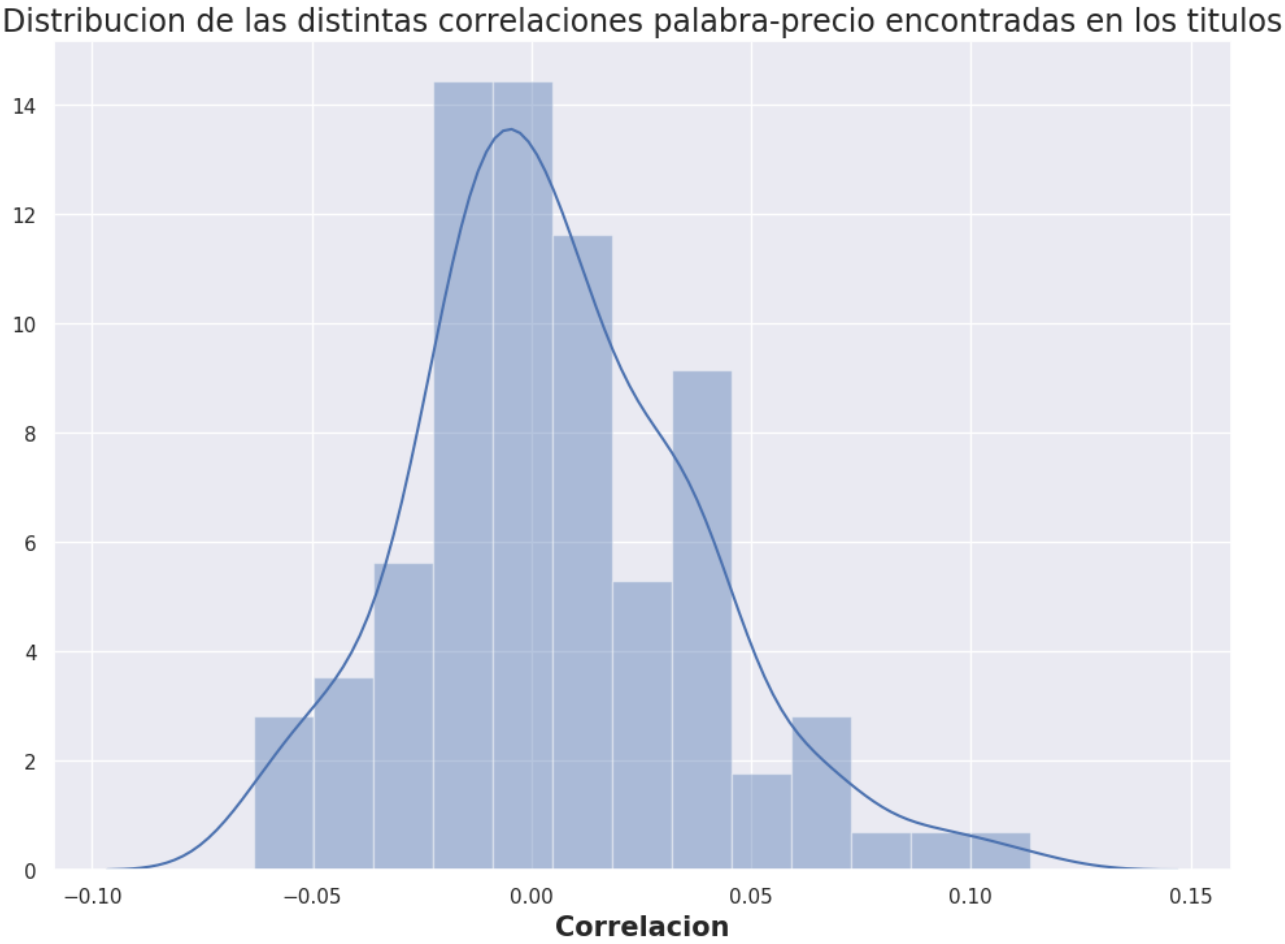
En el grupo 3 encontramos palabras referidas a lujos como 'family' 'cava' 'ph' 'spa' 'cuarzo' 'sauna', 'vapor.', entre otras.

Las palabras varían significativamente entre grupos. Estos gráficos nos muestran que hay palabras que, cuando aparecen, suelen aparecer en grupos de propiedades con determinados precios. Las palabras que contiene la descripción me pueden ayudar a determinar el grupo de precios al que corresponde una propiedad.

5.4 - Correlación entre las palabras del título de la publicación y su precio

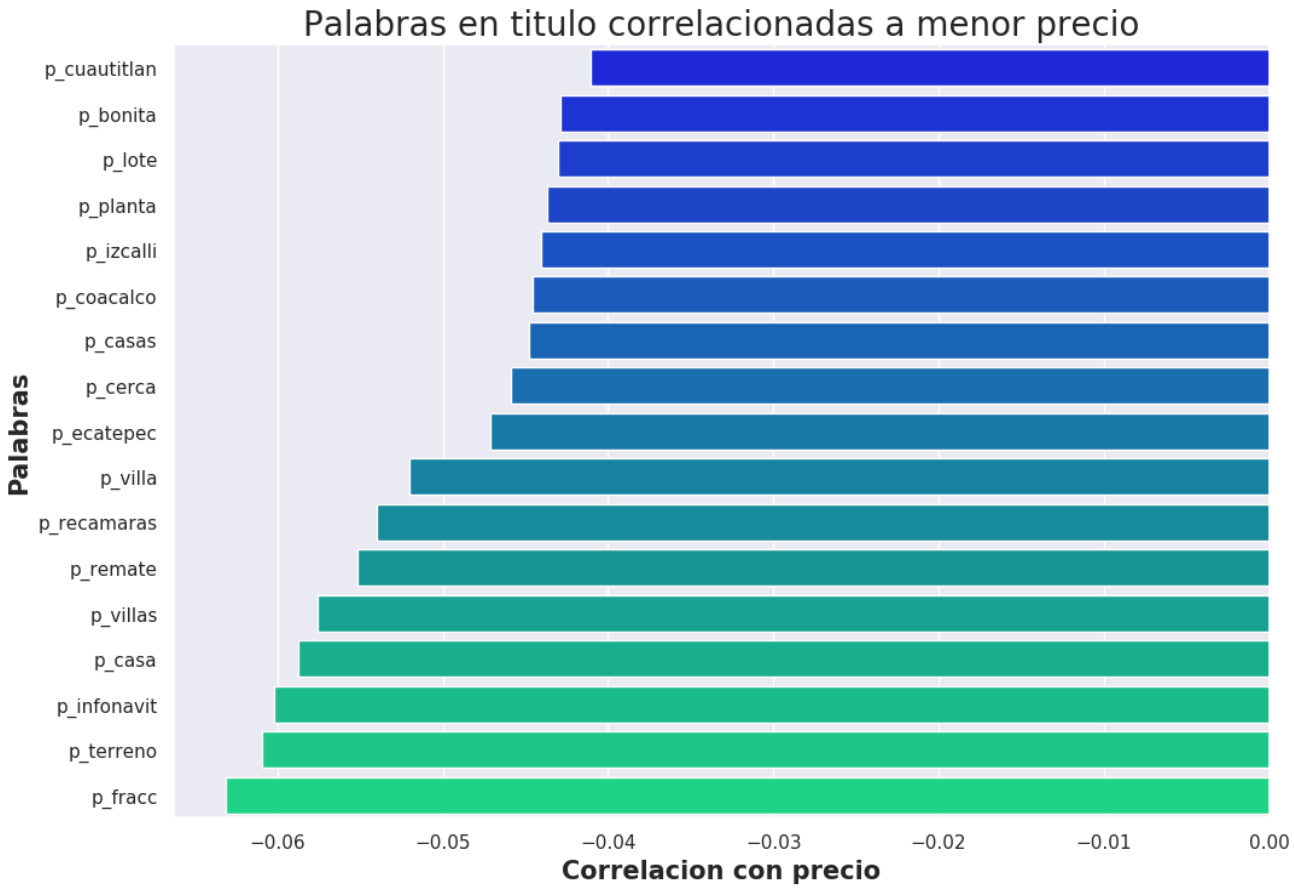
Para este análisis se desarrolla una lista con las palabras que aparecieron más de 800 veces entre todos los títulos.

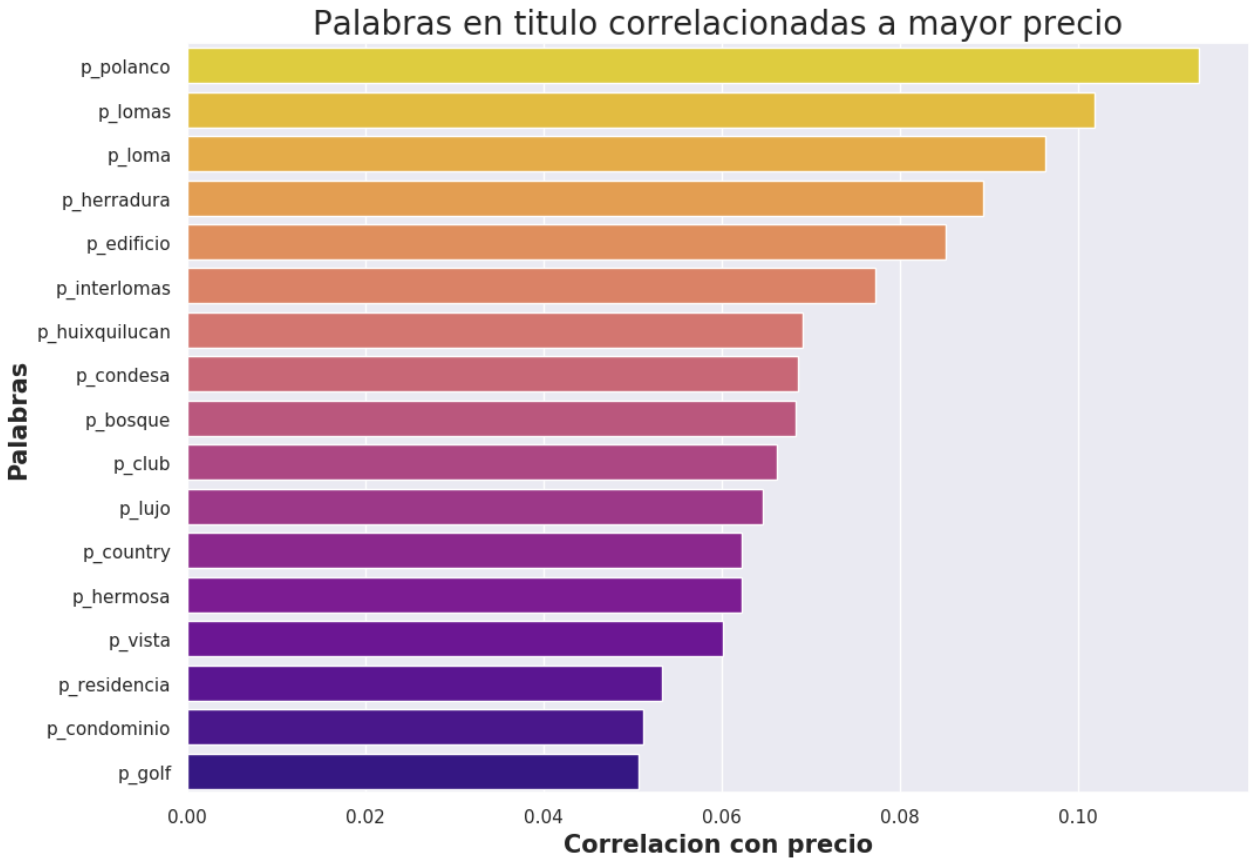
En el dataframe con los datos "train.csv" se agrega una columna por cada palabra de interés, con un 1 si esa publicación tiene esa palabra en el titulo y 0 en caso contrario. Luego se busca la correlación entre cada una de esas columnas y la columna precio y se procede a analizar los resultados.



Las correlaciones no son muy fuertes pero pueden aportar algunos detalles.

Son de interés las palabras de mayor valor absoluto de correlación. Cuanta mayor correlación, quiere decir que a más apariciones hubo mayor precio, y cuanto menor correlación (correlación negativa), quiere decir que hubo más a menor precio.


Palabras en titulo correlacionadas a menor precio
p_cuautitlan
p_bonita
p_lote
p_planta
p_izcalli
p_coacalco
p_casas
p_cerca
p_ecatepec
p_villa
p_recamaras
p_remate
p_villas
p_casa
p_infonavit
p_terreno
p_fracc
Palabras
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.00
Correlacion con precio

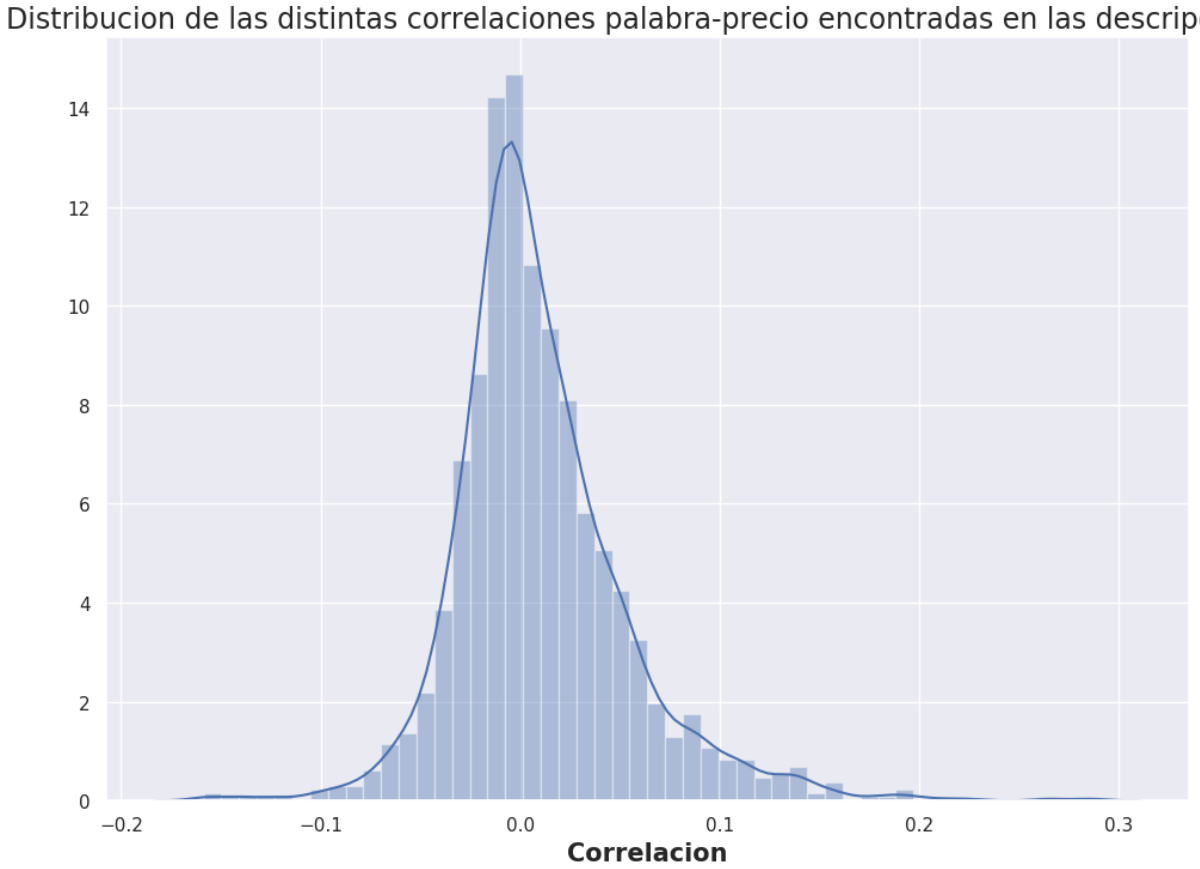

Palabras en titulo correlacionadas a mayor precio
p_polanco
p_lomas
p_loma
p_herradura
p_edificio
p_interlomas
p_huixquilucan
p_condesa
p_bosque
p_club
p_lujo
p_country
p_hermosa
p_vista
p_residencia
p_condominio
p_golf
Palabras
0.00
0.02
0.04
0.06
0.08
0.10
Correlacion con precio

Palabras como “terreno”, “infonavit”, “villa”, “remate” y “fracc” (al parecer refiere a fraccionamiento, un tipo de localidad en México) traen correlacionado menor precio

Luego palabras como “Polanco” (un barrio considerado por muchos el mejor ubicado en México) y “loma”, “country”, “club”, “lujo”, y “edificio” traen correlacionadas mayor precio como uno esperaría.

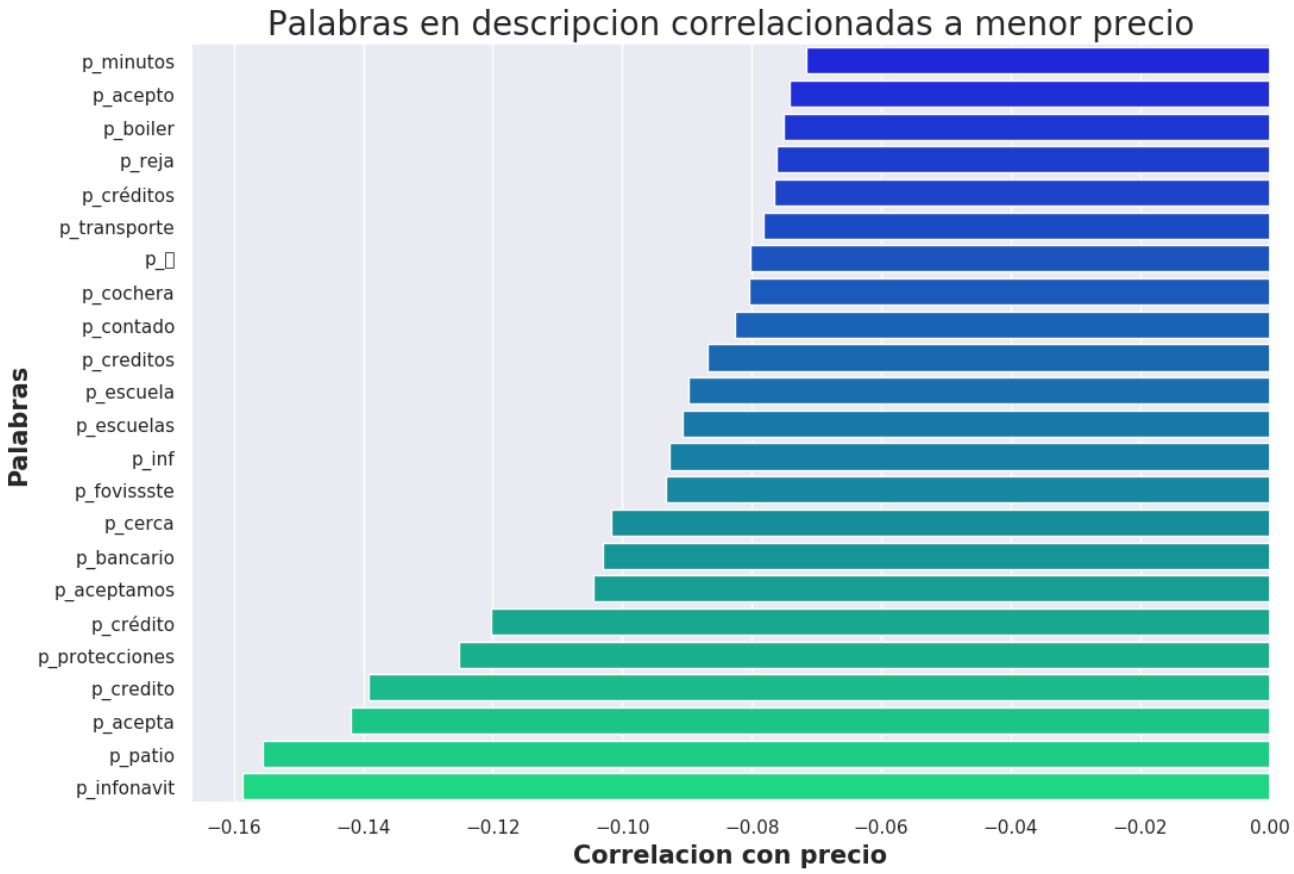
5.5 - Correlación entre las palabras de la descripción de la publicación y su precio

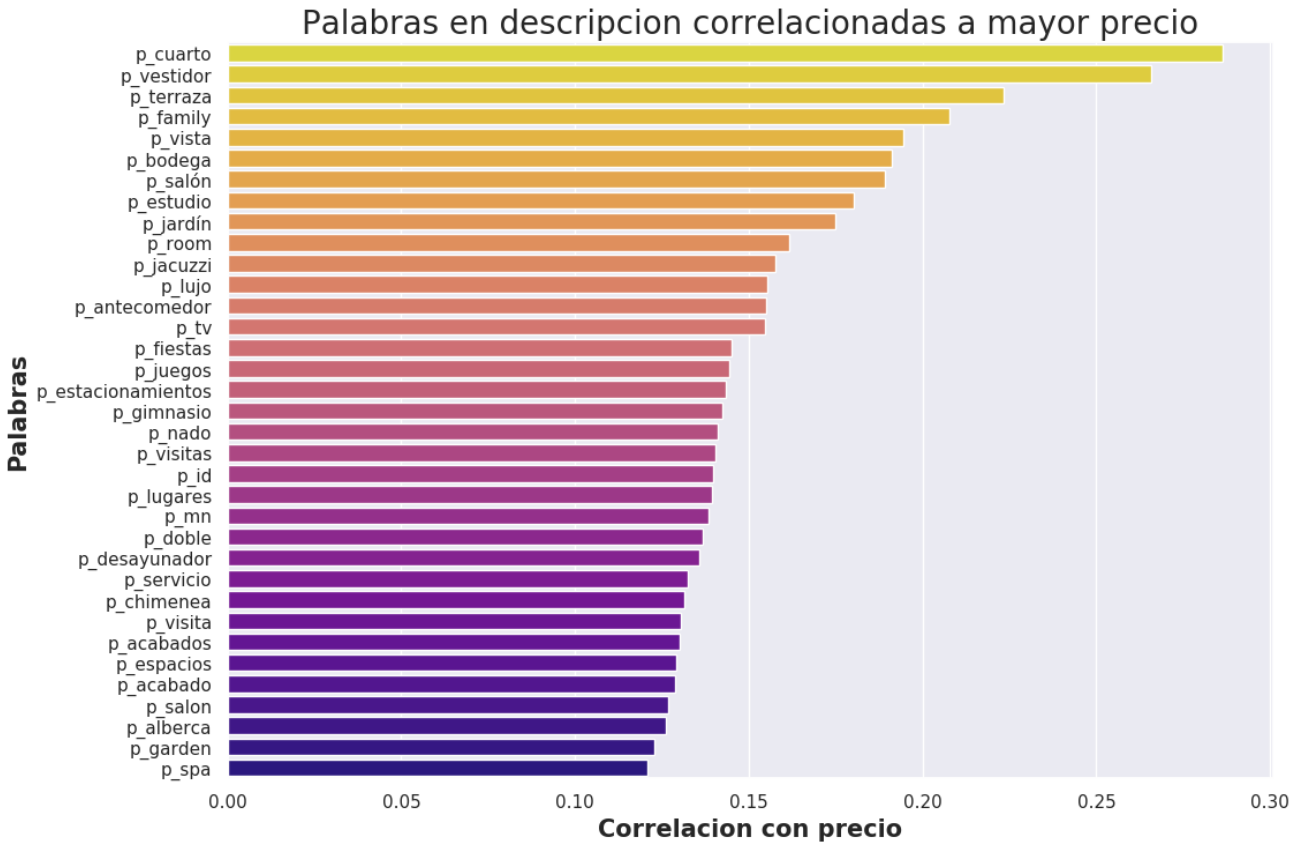
EL análisis es el mismo que el anterior pero para las descripciones y se obtiene lo siguiente:



Estas correlaciones se distribuyen de una manera más amplia que los títulos, quiere decir que hay correlaciones un poco más significativas.

A continuación, se observa cuáles son las palabras en las descripciones con más interés:


Palabras en descripcion correlacionadas a menor precio
p_minutos
p_acepto
p_boiler
p_reja
p_créditos
p_transporte
p_
p_cochera
p_contado
p_creditos
p_escuela
p_escuelas
p_inf
p_fovissste
p_cerca
p_bancario
p_aceptamos
p_crédito
p_protecciones
p_credito
p_acepta
p_patio
p_infonavit
Palabras
−0.16
−0.14
−0.12
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
Correlacion con precio


Palabras en descripcion correlacionadas a mayor precio
p_cuarto
p_vestidor
p_terraza
p_family
p_vista
p_bodega
p_salón
p_estudio
p_jardín
p_room
p_jacuzzi
p_lujo
p_antecomedor
p_tv
p_fiestas
p_juegos
p_estacionamientos
p_gimnasio
p_nado
p_visitas
p_id
p_lugares
p_mn
p_doble
p_desayunador
p_servicio
p_chimenea
p_visita
p_acabados
p_espacios
p_acabado
p_salon
p_alberca
p_garden
p_spa
Palabras
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Correlacion con precio

Las palabras en las descripciones se correlacionan con el precio más significativamente que las de los títulos. Esto puede ayudar bastante a la hora de predecir precios. Además, al observar detenidamente las palabras, estas tienen parecen tener sentido.

En las de menor precio aparece 'Infonavit', que es un Instituto del Fondo Nacional de la Vivienda para los Trabajadores, se asume que va a proponer precios accesibles para las viviendas. También aparece la palabra 'fovissste', un fondo encargado de otorgar créditos para vivienda a los trabajadores del Estado.

Luego en “palabras en descripciones correlacionadas a mayor precio” encontramos a palabras como 'terraza' 'family' 'lujo' 'chimenea' 'garden' etc, de las cuales se espera que aparezcan en publicaciones con precio altos.

SECCION VI Conclusiones Finales

6.0 - Conclusiones finales en base a los datos arrojados

Finalizado el análisis exploratorio se concluye listando ciertos descubrimientos que resultaron de mayor interés y que fueron tomados en cuenta a la hora de crear features para la predicción del precio en la competencia de Kaggle.

Publicaciones. Se puede observar como la cantidad de publicaciones va en aumento a lo largo del tiempo, lo cual nos señala como ZonaProp fue ganando bastante popularidad desde el año 2012 hasta el año 2016.

Precios. Al igual que las publicaciones se observa una tendencia creciente en los promedios de los precios a lo largo de los años registrados. Esto particularmente resulta útil a la hora de crear features para la predicción de este, puesto que podríamos estimar el rango en el que puede caer el precio de acuerdo a la fecha de una publicación.

Ubicación vs. Precio. En los análisis cartográficos se muestra que las publicaciones asociadas a propiedades más baratas se encuentran más al norte y las asociadas con las más caras hacia el centro del país. Esto también supone una buena práctica para utilizarse como un buen feature y realizar predicciones de la variación de los precios.

Tipo de propiedad vs. Precio. Los tipos de propiedades más populares se mueven dentro de un mismo rango de precios. También tenemos tipos que tienen diferencias muy notorias de precios, como por ejemplo los “edificios” que son muy costosos y los terrenos con precios bajos, por lo que será útil tomar en cuenta estos datos si lo que nos interesa es predecir el precio.

Antigüedad vs. Precio. Otro aspecto interesante es el comportamiento del precio en relación a la antigüedad de las propiedades. Los precios en promedio son muy variados para propiedades que están por debajo de los 30 años de antigüedad, luego a medida que la antigüedad aumenta, por encima de esa cifra, los precios se elevan.

Atributos de los inmuebles vs. Precio. Se llaman "atributos" a los diferentes campos hallados en el set de datos que nos señalan si la casa posee ciertas características o no, como la cantidad de habitaciones, baños, garages, metros cubiertos y metros totales. En la mayoría de los casos notamos cómo el precio va en aumento a medida que estos atributos aumentan, por lo que resulta bastante interesante para futuros features.

Palabras en títulos y descripciones. Realizando divisiones en grupos 1, 2 y 3, de acuerdo al rango en el que caen los precios de las propiedades publicadas, se puede notar ciertos patrones en las palabras para estos 3 grupos. Palabras más frecuentes en unos o en otros, por lo que se concluye que existen ciertas palabras que ayudan a determinar el grupo de precios en el que se encuentra una propiedad.

SECCION VII Información del Grupo

6.0 – Información del grupo

Enlace del trabajo realizado en repositorio GitHub :

https://github.com/luctiz/Grupo43-TP1

Código QR al mismo enlace:

Escanear para acceder al enlace.