



**FACULTAD
DE INGENIERIA**

Universidad de Buenos Aires

Organización de Datos [75.06/95.58]

Trabajo Práctico 1

Análisis de Datos: Publicaciones ZonaProp

2° Cuatrimestre de 2019

Grupo 37 - *Dates*

Alumno	Padrón
Sebastián Ignacio Penna	98752
Miguel Toscano	98385
Ivo Biaus Pardo	98708

Link repositorio GitHub: <https://github.com/sebapenna/OrganizacionDeDatos>

Link Kaggle: <https://www.kaggle.com/toscano98385/kernels>

1. Introducción	3
2. Análisis de Datos	4
2.1. Información General	4
2.1.1. Datos Utilizados	4
2.1.2. Herramientas Utilizadas	4
2.2. Crecimiento ZonaProp	4
2.2.1. Crecimiento Global	5
2.2.2. Crecimiento en cada año	5
2.3. Evolución de Precios	7
2.3.1. Evolución en Peso Mexicano	8
2.3.2. Evolución en Dólar	9
2.3.3. Evolución de la relación Peso Mexicano-Dólar	9
2.4. Tipo de Propiedades	10
2.4.1. Cantidad de Publicaciones según el Tipo	11
2.4.2. Precio según Tipo de Propiedad	12
2.5. Antigüedad	14
2.5.1. Cantidad de publicaciones según Antigüedad	14
2.5.2. Precio según Antigüedad	16
2.5.3. Antigüedad por provincia	17
2.6. Habitaciones	17
2.6.1. Distribución de Habitaciones	18
2.6.2. Precio según Habitaciones	19
2.7. Baños	19
2.7.1. Distribución de Baños	20
2.7.2. Precio según Baños	21
2.8. Garages	21
2.8.1. Distribución de Garages	22
2.8.2. Precio según Garages	23
2.9. Servicios	23
2.9.1. Distribución de Servicios	24
2.9.2. Precio según Servicios	25
2.10. Propiedades Cercanas	25
2.10.1. Distribución según Propiedad Cercana	26
2.10.2. Precio según Propiedades Cercanas	27
2.11. Términos utilizados	27
2.11.1. Título	28
2.11.2. Descripción	29
2.12. Metros Cuadrados	30
2.12.1. Limpieza de datos	30
2.12.1.1. Caso metros totales en NULL	31

2.12.1.2. Caso metros cubiertos en NULL	31
2.12.2. Análisis	32
2.12.3. Metros totales por provincia	35
2.12.4. Precio/Metro por provincia	37
2.12.5. Precio por provincia	37
2.13. Regiones	38
2.13.1. Publicaciones por provincia	39
2.13.2. Cantidad de propiedades por región	40
2.13.3. Distribución de los precios (USD) para cada región.	41
3. Conclusión	42
3.1. ZonaProp y los precios	42
3.2. Tipos principales	42
3.3. Determinación del precio de una propiedad	42
3.4. Ayudar a encontrar una propiedad	43
3.5. Ayudar a crear una buena publicación	43

1. Introducción

El fin de éste informe es extraer información relevante y de utilidad mediante el análisis exploratorio del set de datos compuesto por el registro histórico de publicaciones del sitio web ZonaProp, brindado por Navent. La información con la que se trabajará corresponde particularmente al período comprendido entre 2012 y 2016, para las publicaciones realizadas en México.

A la hora de analizar los datos se decidió poner el enfoque en distintos puntos desde los cuales se podría obtener información relevante:

- Uno intentando evaluar toda la información que sea relevante a Navent, ya sea desde el crecimiento de ZonaProp a lo largo de los años, los distintos tipos de publicaciones que poseen, las distintas regiones que mayor impacto tienen sobre el sitio o la evolución de los precios, entre otros puntos;
- Otro punto desde el cual se trabajó con los datos es pensando en obtener información útil para aquellas personas en el mercado inmobiliario. Ésto abarca todo lo relacionado a la variación de precios en las publicaciones en base a las características de las mismas, así como tener una idea sobre qué tipo de propiedades son más fáciles de obtener dentro de ZonaProp;
- Por último, se buscará brindar información a nuevos clientes de ZonaProp que quieran publicar su propiedad en el sitio. Dentro de ésto se analizará una publicación desde como armar el título, su descripción, hasta cómo definir el precio de la misma en base a sus características;

Analizado cada punto con sus correspondientes visualizaciones e integrando la información obtenida de cada uno esperamos descubrir nuevos aspectos que sean de utilidad a Navent, tanto desde el análisis de los años pasados así como posibles futuras nuevas características del sitio ZonaProp.

2. Análisis de Datos

2.1. Información General

2.1.1. Datos Utilizados

En el desarrollo del trabajo se utilizaron dos sets de datos:

1. **train.csv**: el set de datos principal entregado por Navent, conteniendo las publicaciones de ZonaProp México entre 2012 y 2016;
2. **dollar.csv**: contiene información sobre el precio del dólar en México entre el 12/12/2011 y el 31/01/2017. Mediante éste se convertirá el valor de la propiedad en pesos mexicanos a dólares con el valor del mismo en el día correspondiente a la publicación. Éste set de datos se obtuvo del sitio web <https://mx.investing.com/>.

2.1.2. Herramientas Utilizadas

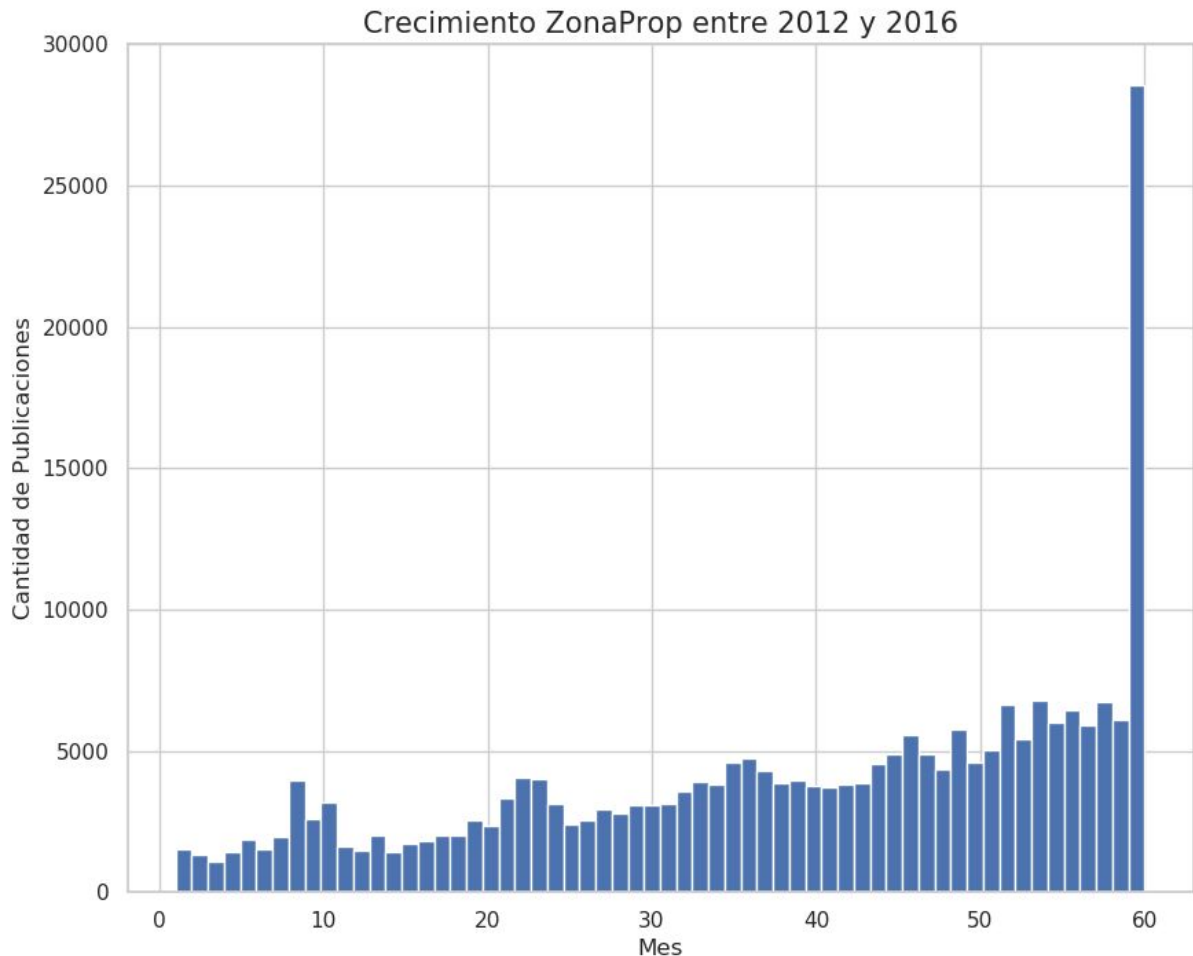
Para el desarrollo del trabajo se trabajó con el lenguaje Python 3, usando principalmente la librería Pandas para el análisis de datos y Matplotlib y Seaborn para la generación de visualizaciones.

Para ciertos problemas particulares del trabajo se utilizaron otras librerías como Wordcloud, Unidecode y Nltk.

2.2. Crecimiento ZonaProp

Dado que el set de datos pertenece a todas las publicaciones del sitio ZonaProp entre 2012 y 2016 un buen punto de partida nos pareció evaluar cómo fue el crecimiento del sitio en términos de cantidad de publicaciones en éste período de tiempo.

2.2.1. Crecimiento Global

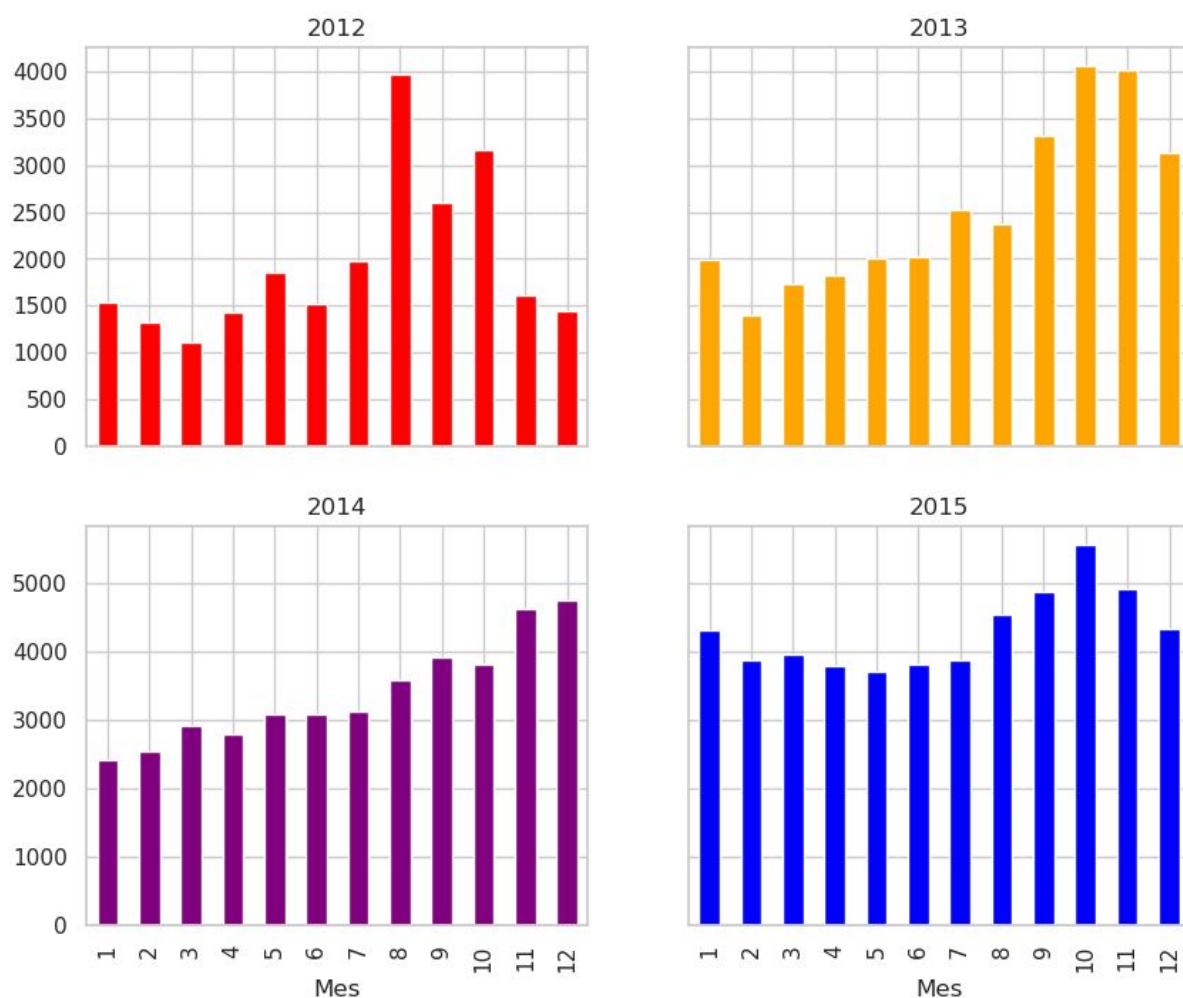


Fácilmente se puede observar una clara tendencia al crecimiento, aunque hay una gran diferencia en el último mes (Diciembre 2016). Previo a éste mes el crecimiento fue paulatino y sin grandes saltos, inclusive teniendo épocas de dos o tres meses de caídas en aquellos meses cercanos a un cambio de año. Sin embargo, en el último mes se tiene un gran incremento en la cantidad de publicaciones dentro del sitio, posiblemente ante alguna campaña publicitaria muy efectiva o una nueva función dentro de ZonaProp que hiciera al sitio mucho más atractivo, información con la cual no contamos.

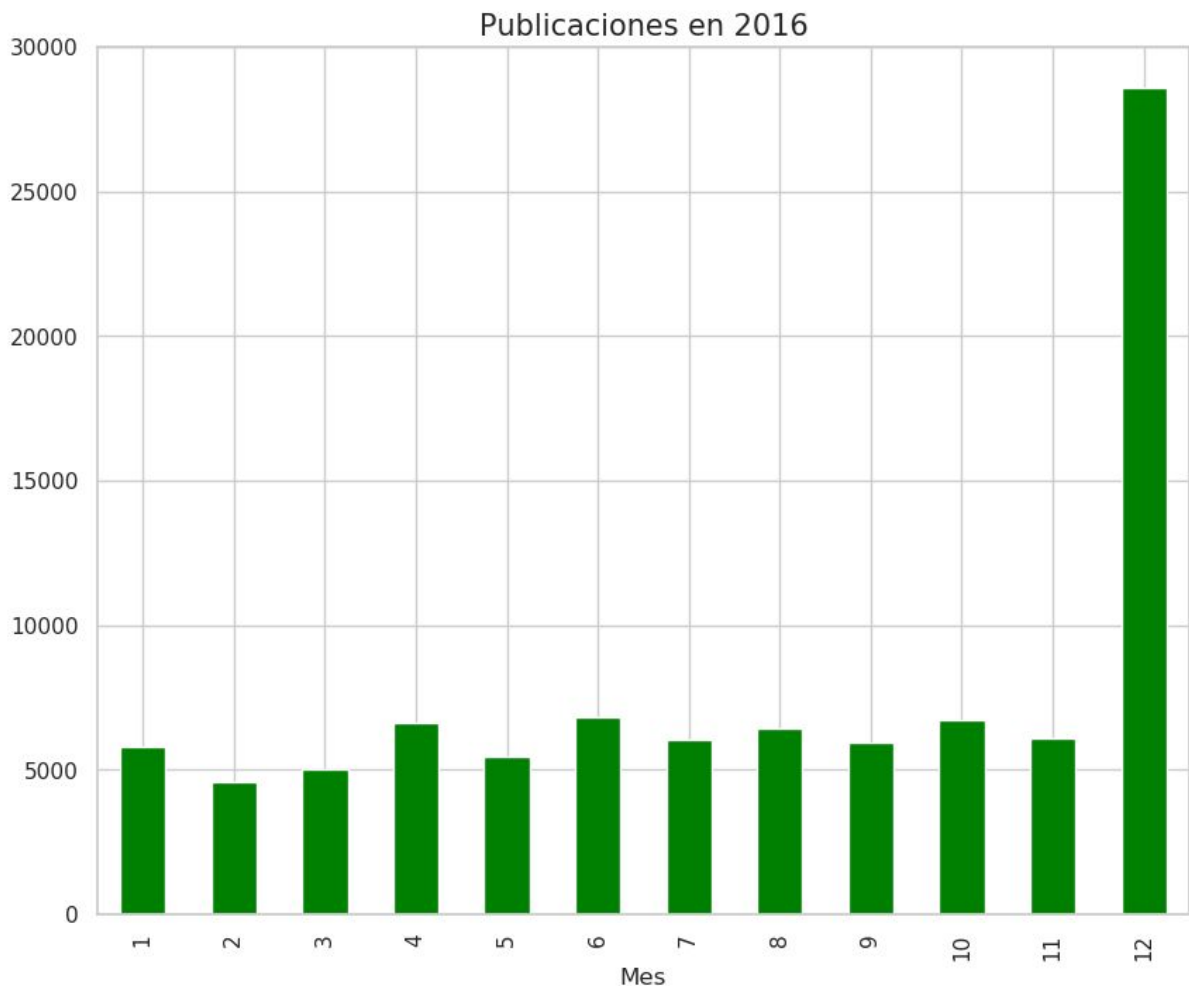
2.2.2. Crecimiento en cada año

Sabiendo como fue el comportamiento del sitio a lo largo de los 5 años nos pareció interesante ver cómo fue el comportamiento en cada año por separado, y así poder observar en detalle que meses presentaban mayor cantidad de publicaciones.

Publicaciones por mes



En los primeros cuatro años (donde el crecimiento fue menor) no hay una tendencia predominante por igual, aunque se puede observar que siempre la mayor cantidad de publicaciones se presentan en la segunda mitad del año, siendo cada vez más tarde, aunque con una caída en lo Noviembre y Diciembre (a excepción del 2014).



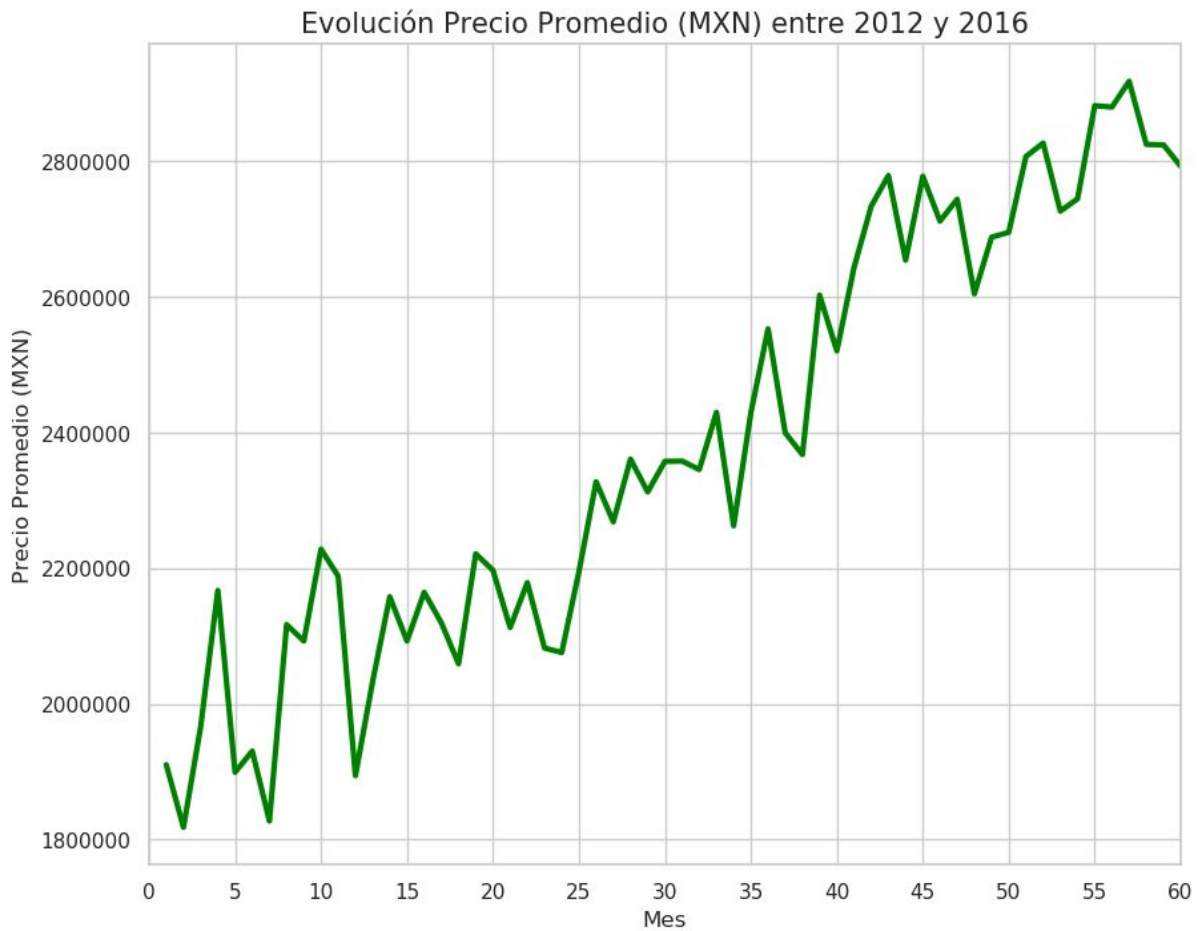
El caso del 2016 se evaluó por separado, debido a su gran crecimiento en Diciembre. Claramente se puede ver que la cantidad de publicaciones tiene valores cercanos a los del año previo, a excepción de éste último mes, e inclusive se ve una mayor estabilidad en la cantidad de publicaciones a lo largo del año, algo que no se presentaba previamente. Cualquiera haya sido la acción o acciones tomadas previo o durante el mes de Diciembre claramente tuvieron un impacto positivo en la cantidad de clientes de ZonaProp.

2.3. Evolución de Precios

Otro punto importante para empezar consideramos era la inflación en el valor de las propiedades a lo largo de los años. Decidimos considerar tanto los valores en pesos mexicanos así como en dólares, dado que es una moneda reconocida como referencia mundialmente y el valor en ésta nos sería mucho más común que en pesos mexicanos.

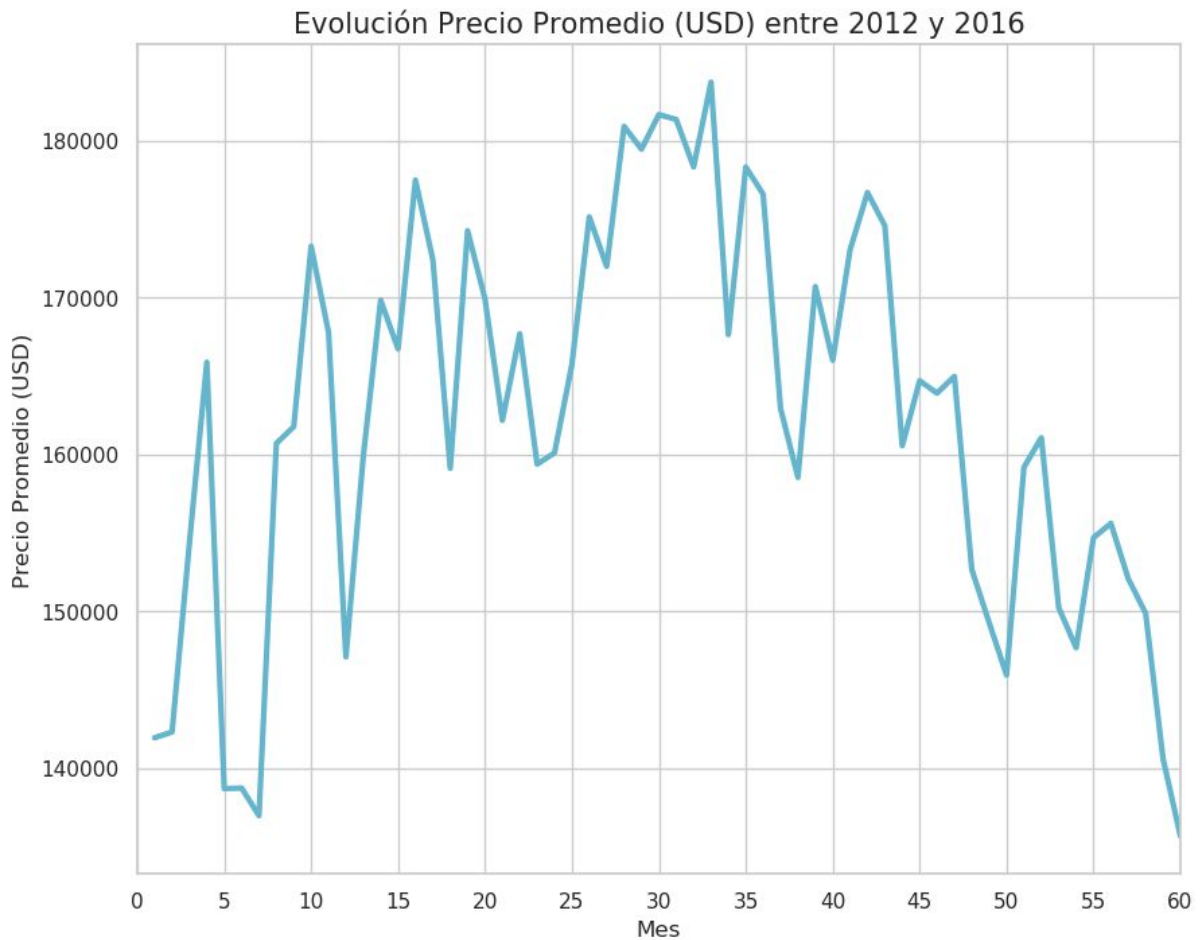
En ambos casos se generó un promedio de los precios en cada cada mes, tomando todas las distintas propiedades que realizaron publicaciones.

2.3.1. Evolución en Peso Mexicano



En estos cuatros años se produjo un aumento cuasi lineal en la moneda local (dentro del mercado de las propiedades), con un crecimiento mucho más acelerado a partir de los meses 25-30.

2.3.2. Evolución en Dólar

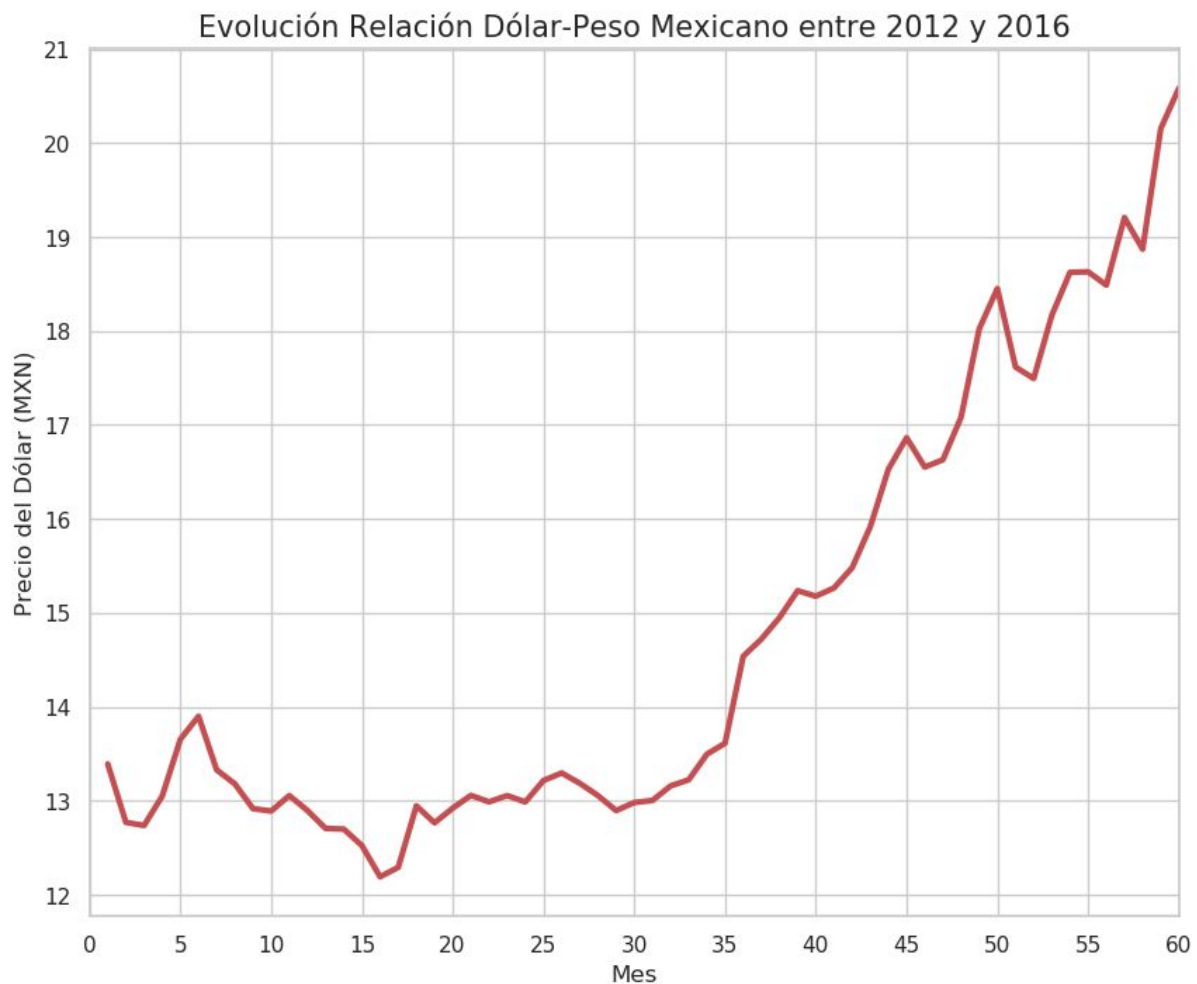


Una vez convertido el valor de todas las propiedades a dólar podemos observar un comportamiento totalmente diferente. Los primeros 30 meses se observa un claro crecimiento en el valor de las propiedades, aunque con ciertos altibajos. Sin embargo, a partir del mes 30 (donde con el peso mexicano se comenzaba a presentar el mayor crecimiento) comienza una caída muy importante en el valor de las propiedades dolarizadas, volviendo a un punto similar al comienzo.

Este gráfico nos da una idea de que muy posiblemente se podría haber producido una importante devaluación en la moneda frente al dólar.

2.3.3. Evolución de la relación Peso Mexicano-Dólar

Como consecuencia de la gran diferencia en las dos visualizaciones previas es importante entender cómo fue variando el cambio entre el peso mexicano y el dólar, lo que puede haber tenido un gran impacto en la definición del precio a las propiedades.



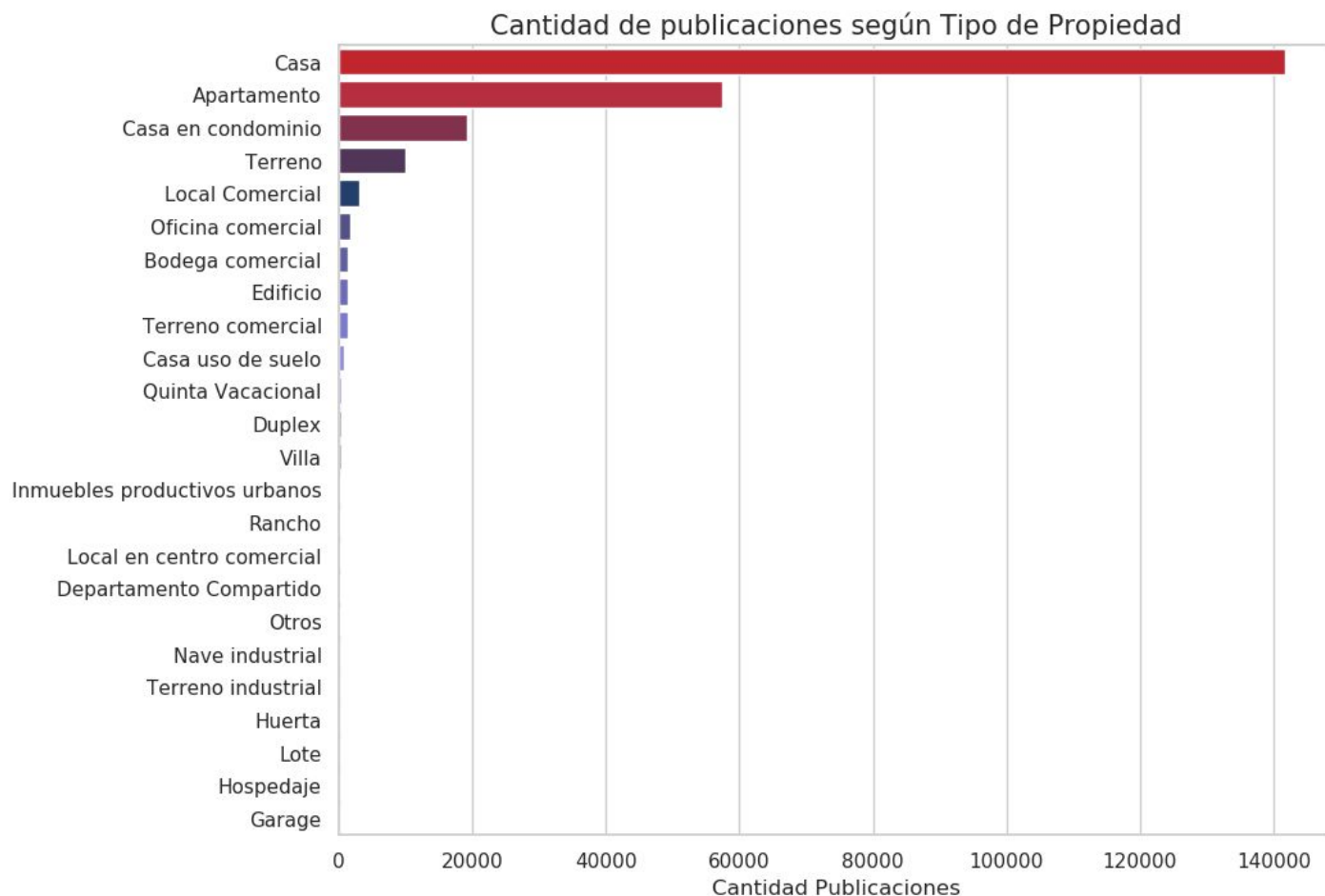
Esta nueva visualización se justifican los comportamientos de las curvas previas: a partir del mes 30 comenzó una importante devaluación de la moneda local creciendo un total de casi 10 pesos en un período de casi tres años. Éste aumento en el cambio de moneda explica porque la diferencia en la variación del precio de las propiedades cuando se modifica la moneda.

Siempre que se obtenga una conclusión sobre el precio ante distintos factores será importante recordar que hubo una gran variación de precios a lo largo de los años estudiados, teniendo un efecto directo al realizar comparaciones.

2.4. Tipo de Propiedades

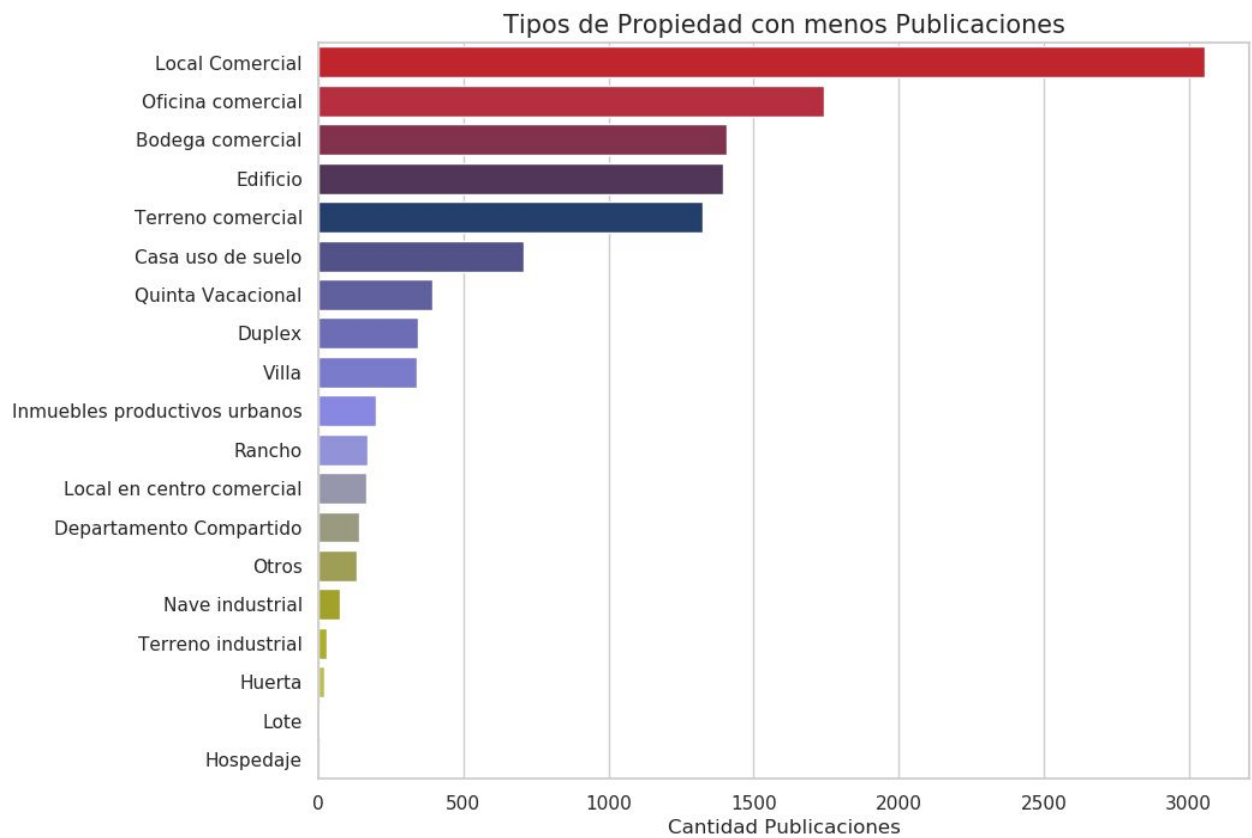
Una vez que comenzamos a trabajar con los datos uno de los primeros puntos que pudimos observar es la gran variedad de tipos de propiedades que se encuentran dentro del sitio. Por lo tanto, se decidió evaluar el impacto de cada una dentro del sitio.

2.4.1. Cantidad de Publicaciones según el Tipo

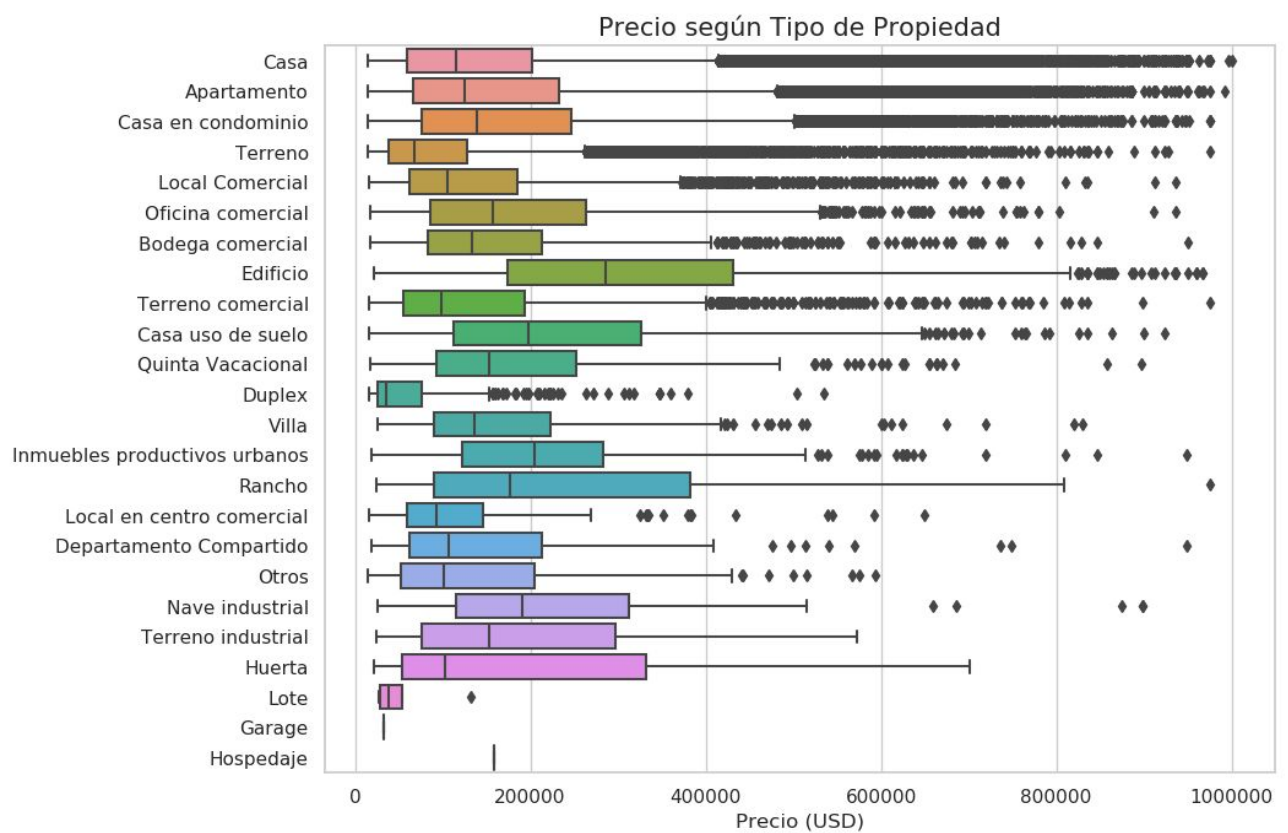


Dentro de las 240000 publicaciones de las cuales se tienen datos más de la mitad pertenecen a casas, claras dominantes dentro de las publicaciones, seguidos por los apartamentos, aunque por una diferencia de más de 80000 publicaciones. El resto de los tipos no alcanzan a representar un ni un 10% de los datos, y si se considerará el tipo 'Casa en condominio' dentro del tipo 'Casa' la diferencia sería aún mayor.

Dado que la gran mayoría de las publicaciones no alcanzan a ser representadas dentro del mismo gráfico que las casas se muestra en detalle la cantidad de publicaciones de éstas.

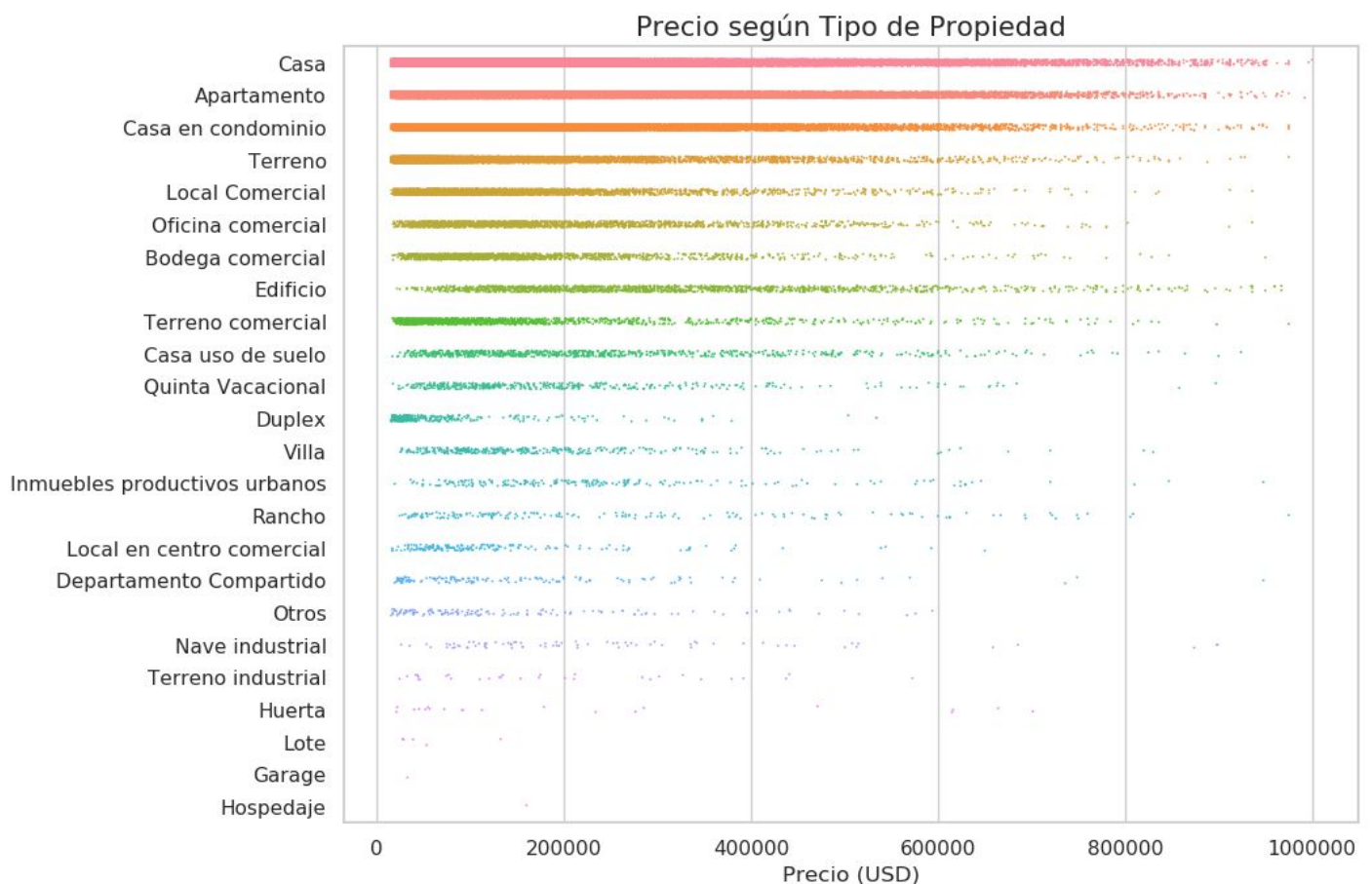


2.4.2. Precio según Tipo de Propiedad



Lo primero que observamos al analizar los precios según el tipo es que la gran mayoría presenta una gran variabilidad en los datos y que aquellos tipos con mayor cantidad de publicaciones muestran muchos outliers. A través de éste plot obtenemos una idea de que en qué rango de valores se encontrarán los precios para cada propiedad, pero se vuelve difícil diferenciar en cuales de ellas encontraremos mayor cantidad de outliers, así como encontrar el rango de precios con mayor densidad claramente.

Para complementar este estudio generamos un scatter plot con jittering, permitiendo entonces determinar a partir de que valores cada propiedad comienza a presentar menos publicaciones, y el cambio en ésta densidad.



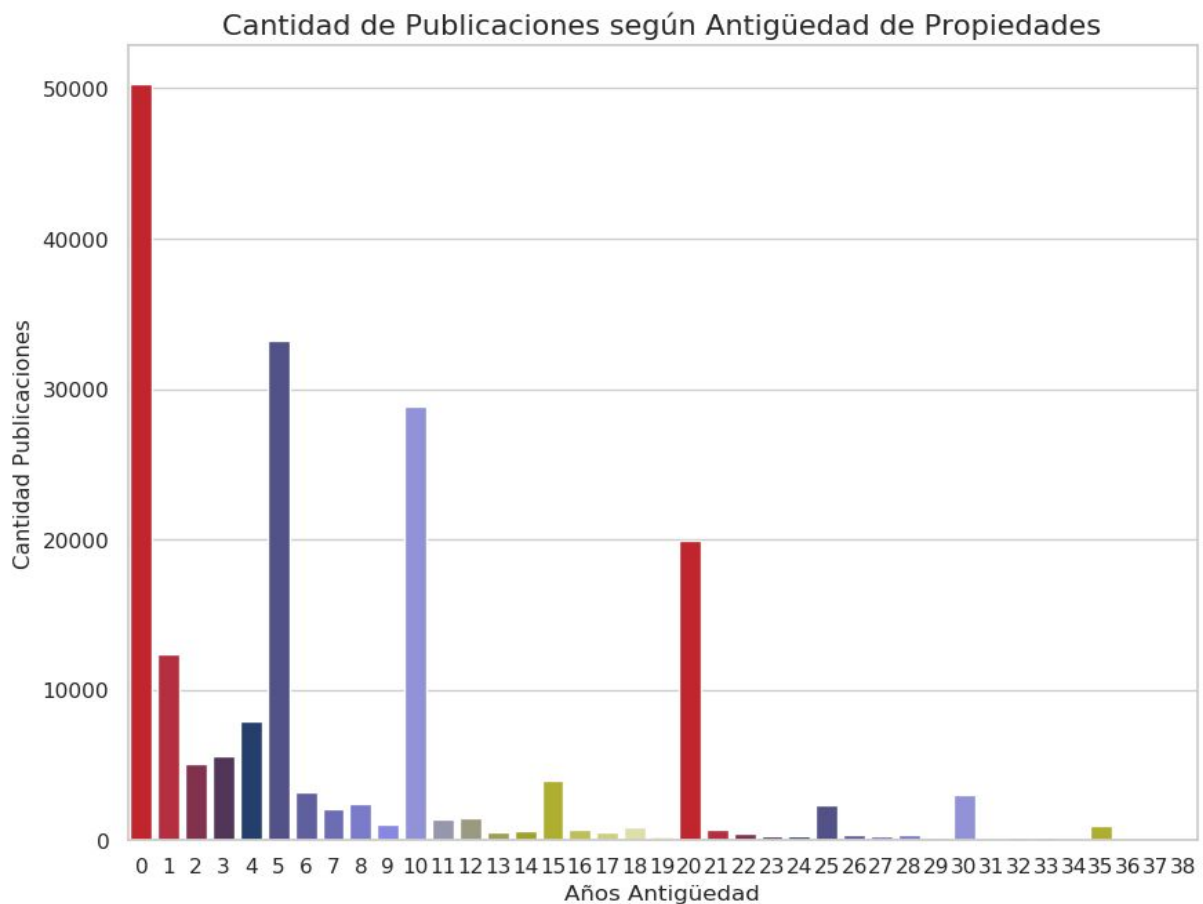
Ahora resulta más sencillo observar que todas las propiedades suelen concentrar sus publicaciones en los valores más bajos, aunque muchas también muestran tener publicaciones con valores altos. Aquellos tipos con mayor cantidad de publicaciones muestran una variabilidad mayor, dado que la densidad de puntos comienza a caer luego de los 600.000 dólares, mientras que las demás caen en valores cercanos a los 400.000, 200.000 dólares o inclusive menos para aquellos tipos con muy pocas publicaciones.

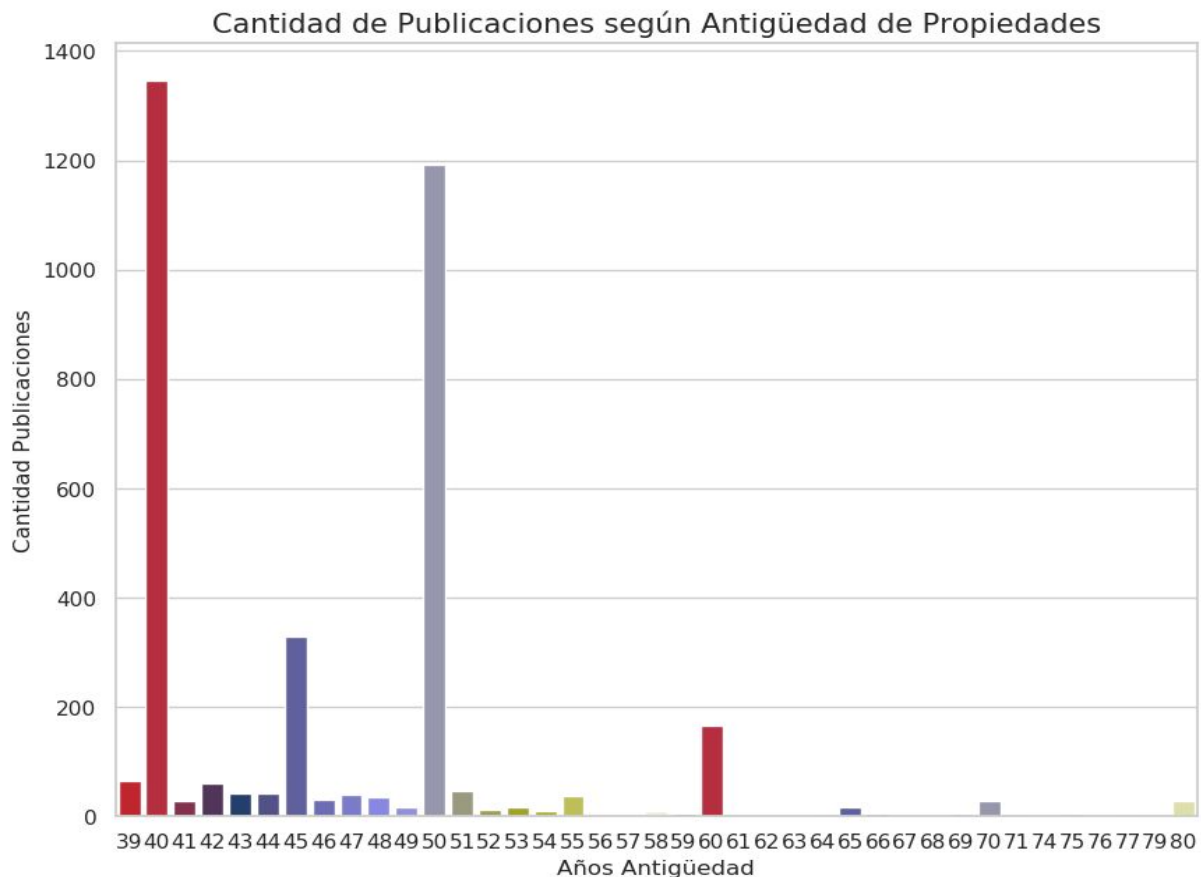
2.5. Antigüedad

Dentro de las distintas características que describen a cada propiedad una que comparten de igual manera todas las propiedades es la antigüedad. Tanto como varían los precios a medida que una propiedad es más antigua, así como cuáles son los valores predominantes en el rango de años son aspectos que se evaluaron.

2.5.1. Cantidad de publicaciones según Antigüedad

El rango de años de antigüedad que presentan las propiedades al estudiar los datos van desde los 0 años (entendemos corresponden a propiedades a estrenar) a los 80 años.

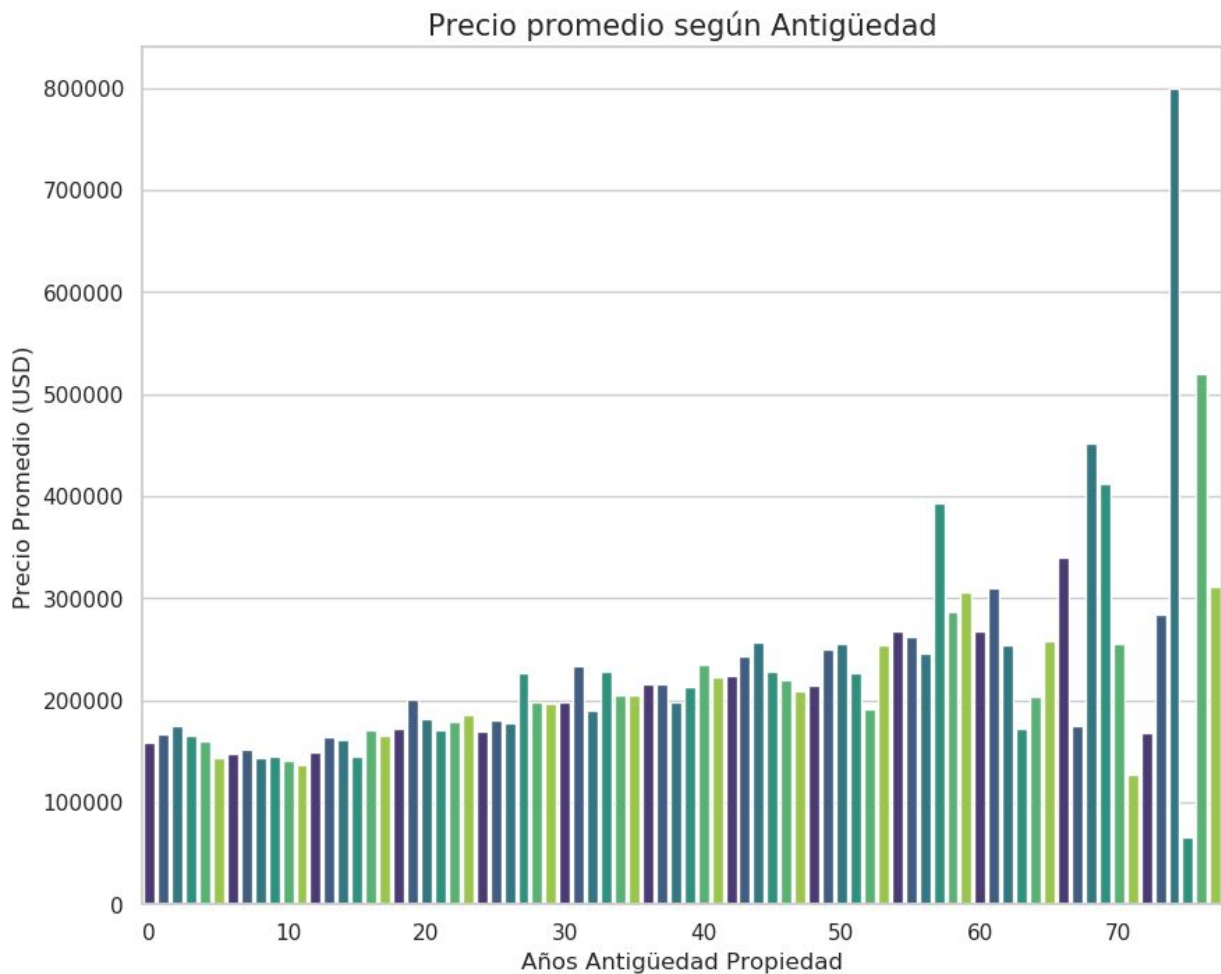




Dentro de ZonaProp se observa que cerca de un 50% de los datos tienen menos de 20 años de antigüedad, siendo mayor la cantidad cuando tienen menos de 10 años, dominando ampliamente aquellas propiedades nuevas (0 años).

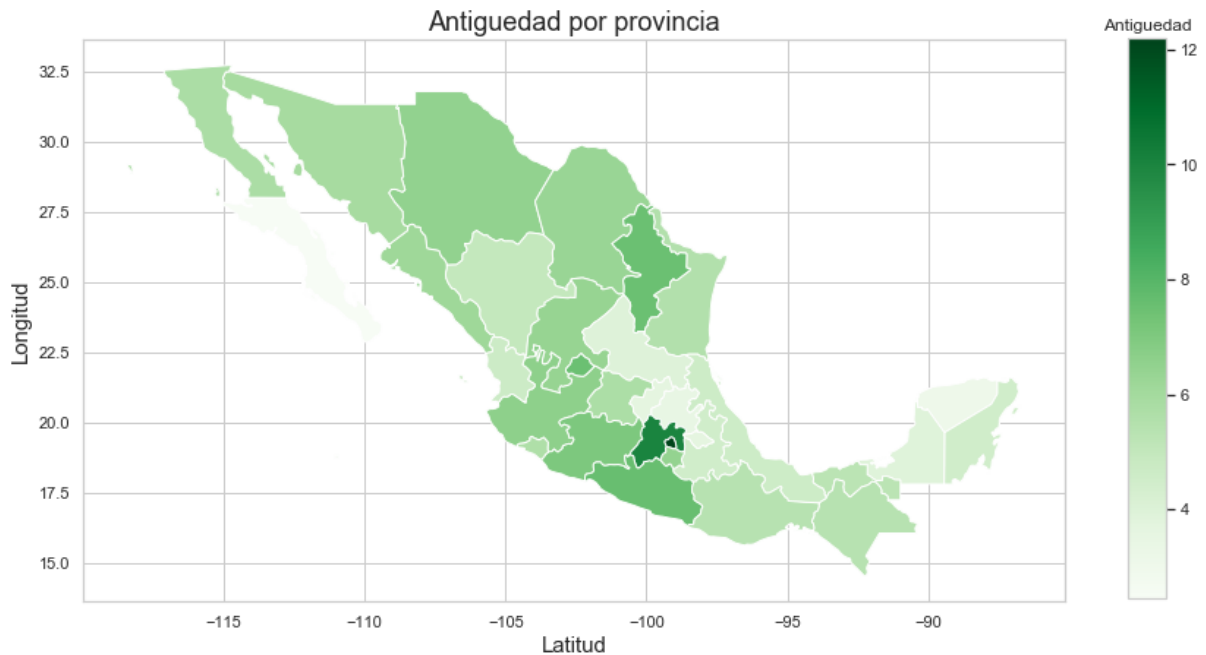
Otro detalle interesante que se observa es que se producen picos en la cantidad de publicaciones para aquellos años múltiplos de 5, independientemente de la cantidad de publicaciones que representan estos valores. No creemos que esto efectivamente represente que las propiedades publicadas suelen tener antigüedades dentro de éstos valores, sino una tendencia de las personas que realizan una publicación a colocar un valor de éste estilo y, como consecuencia, algo más atractivo a la hora de comparar distintas publicaciones.

2.5.2. Precio según Antigüedad



Cuando se evalúa el precio como consecuencia de la antigüedad vemos que a medida que una propiedad suma años su costo aumenta. Sin embargo, es importante tener en cuenta que cuantos más años menos publicaciones se tienen, por lo tanto el promedio del valor estaría claramente dominado por aquellas propiedades con mayor (o menor) valor, cuando las propiedades de menor antigüedad, en especial las de 0, cuenta con una gran cantidad de valores que ajustan éste promedio.

2.5.3. Antigüedad por provincia



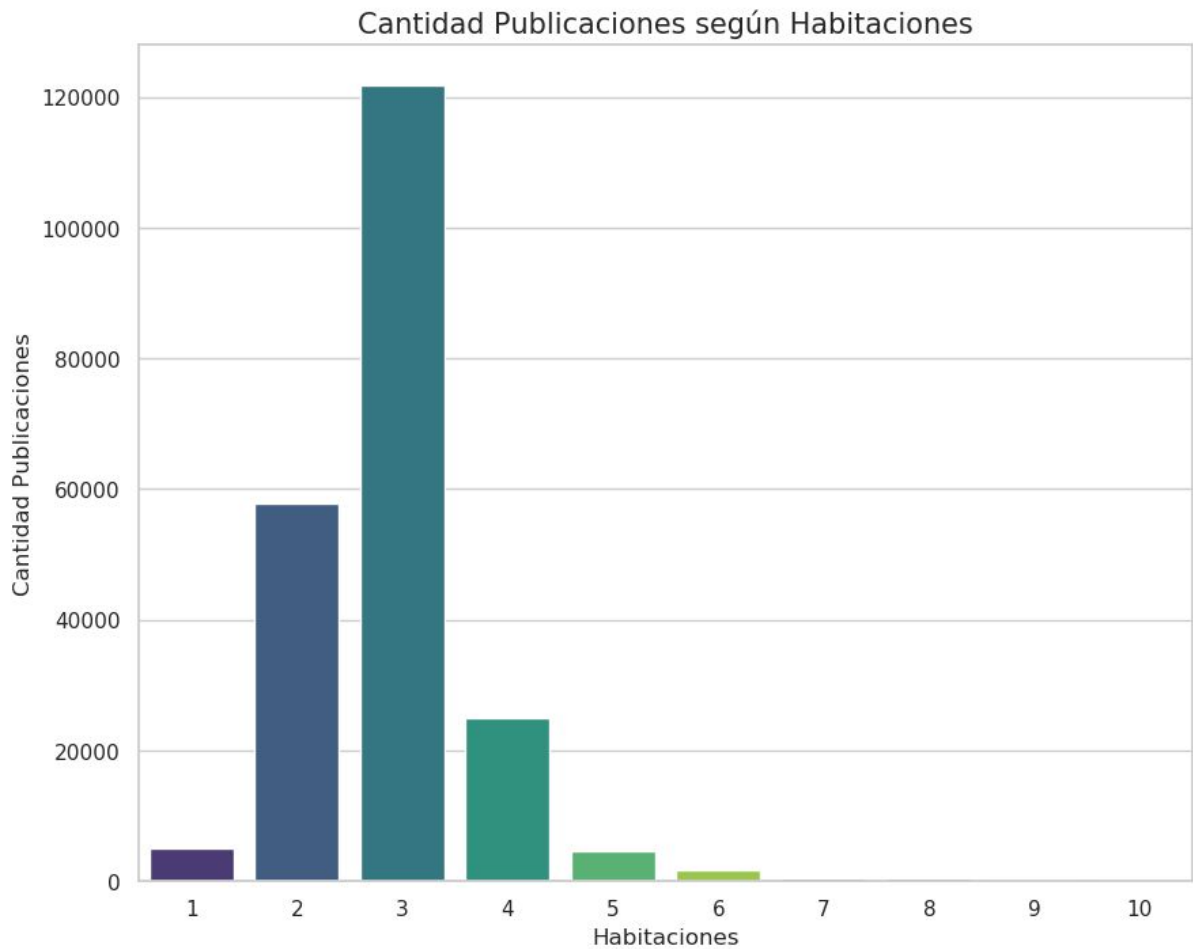
Al hacer un promedio de la antigüedad por provincia, se ve como los valores tienden a pertenecer a los más bajos de la variedad de antigüedades. Es interesante ver cómo las propiedades más antiguas pertenecen mayoritariamente a la capital del país por amplia diferencia.

2.6. Habitaciones

La cantidad de habitaciones que presente una propiedad sin duda debe ser uno de los puntos que más influyen en el precio, por lo que dentro de nuestro de análisis incluimos el estudio de la variación del precio con éstas, así como la distribución a lo largo de las publicaciones. Al evaluar los datos nos encontramos que el rango de valores se encuentra entre 1 y 10 habitaciones.

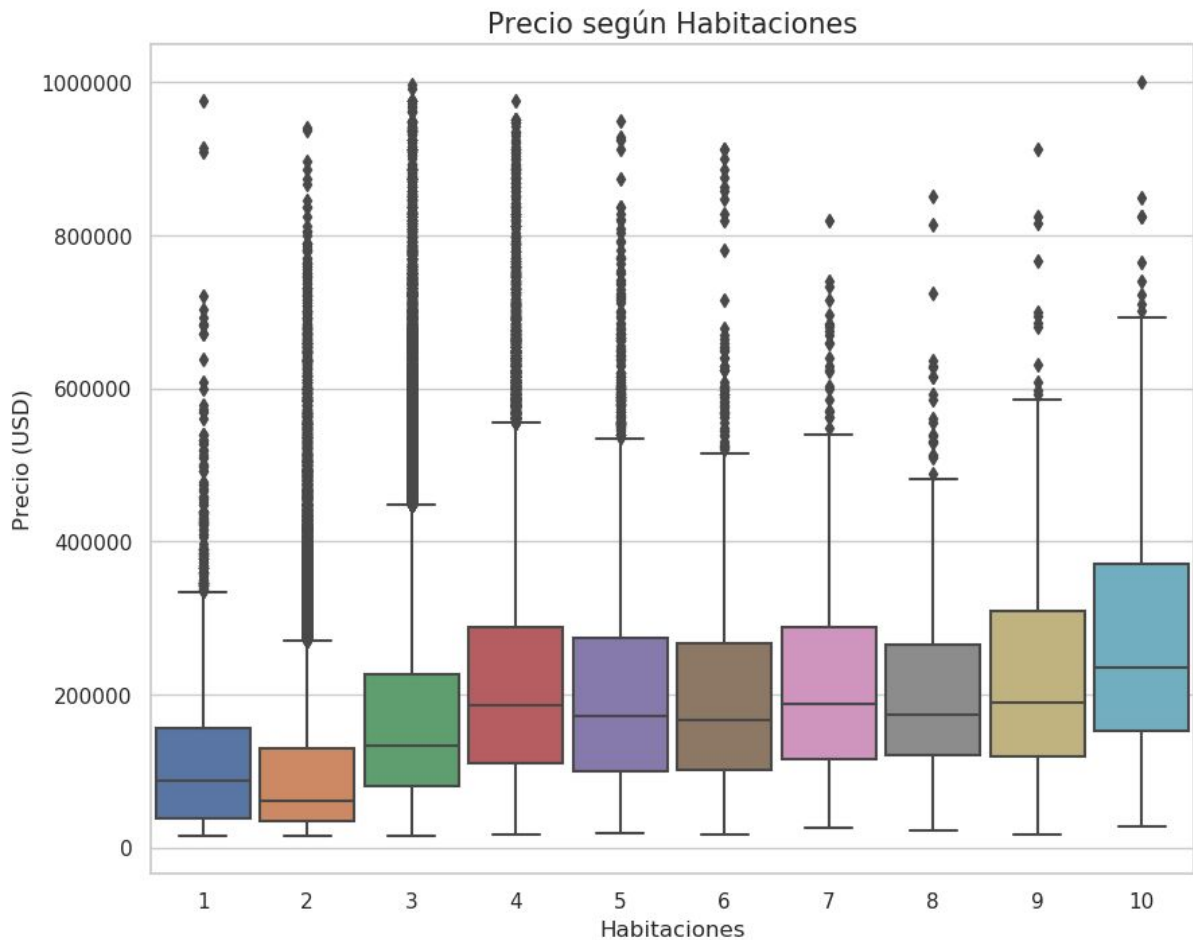
En cuanto a los valores nulos dentro de los datos representan menos de un 10% de los datos, por lo tanto se decidió descartar dichos datos en éste en caso en vez de inferir sus posibles valores.

2.6.1. Distribución de Habitaciones



Cerca del 80% de los datos están concentrados entre 2 y 4 habitaciones (recordar que las casas y apartamentos dominan las publicaciones), mientras que a partir de las 7 habitaciones se dejan de tener valores de impacto en el sitio.

2.6.2. Precio según Habitaciones



En los valores que se observaron como predominantes en la visualización anterior (2, 3 y 4) se observa un claro crecimiento del valor de la propiedad ante el aumento en la cantidad de habitaciones, al mismo tiempo que aumenta la variabilidad en los mismos. Luego de esos valores se alterna entre caídas y subidas en la distribución de los precios así como sus máximos, aunque manteniendo una cierta estabilidad en la variabilidad, creciendo finalmente para el valor máximo de habitaciones.

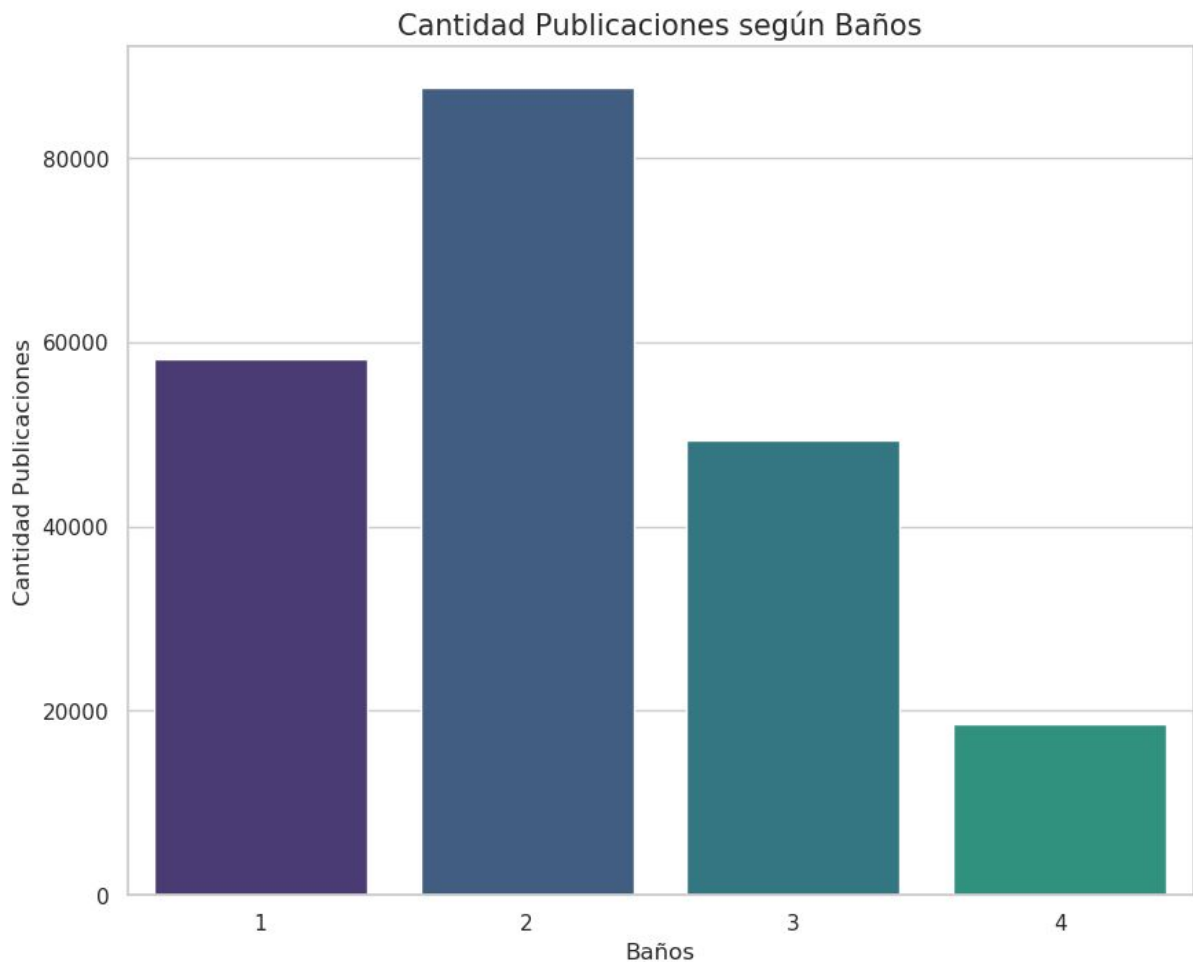
2.7. Baños

Al igual que las habitaciones, otro aspecto de las propiedades que brindan una información sobre su capacidad y posible valor es la cantidad de baños que posee.

Analizando el set de datos las propiedades presentaban una cantidad de baños entre 1 y 4. En éste caso los valores nulos representan cerca de un 10% de los datos pero

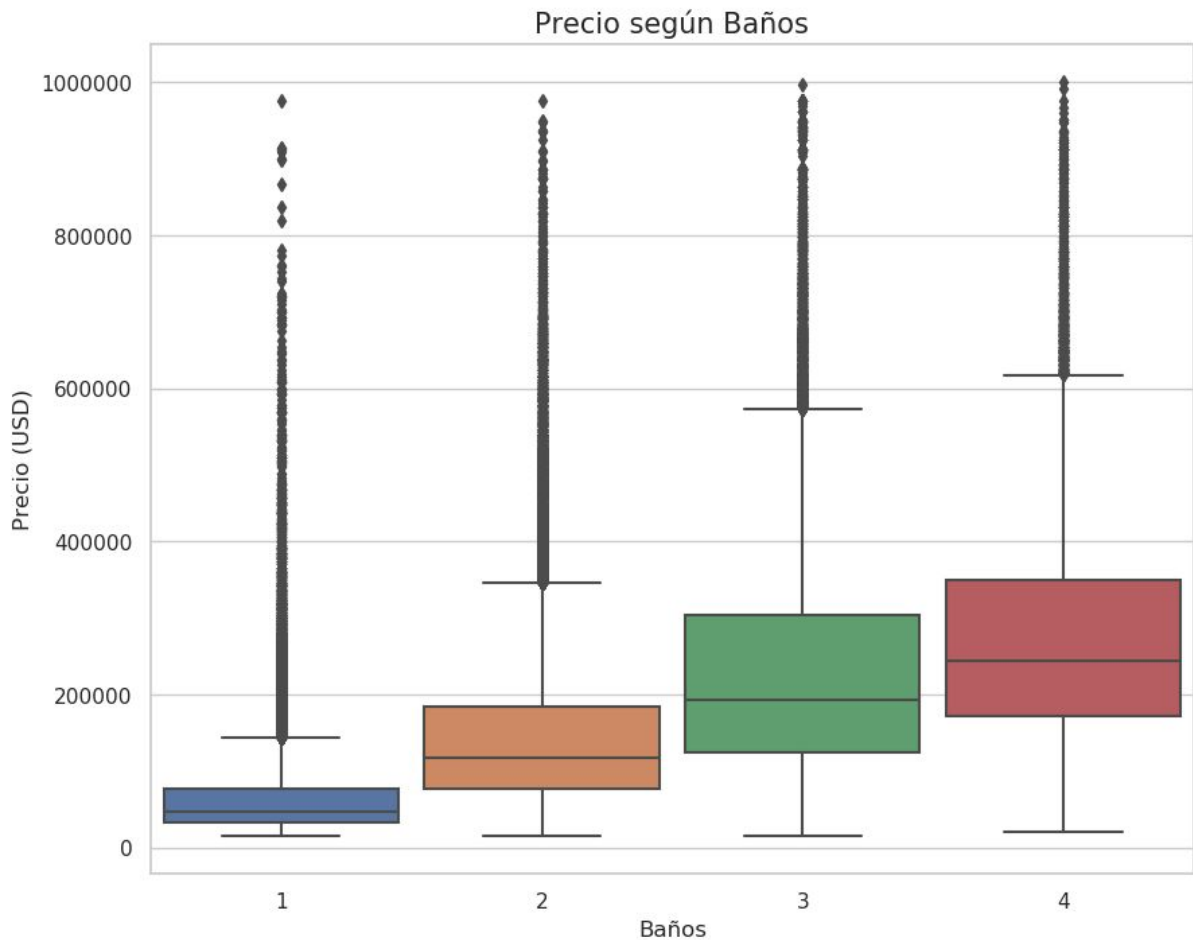
al analizar cada tipo de propiedad no se encontró una tendencia clara para definir estos valores ausentes, por lo tanto se decidió nuevamente descartar dichos datos.

2.7.1. Distribución de Baños



En éste caso todos los posibles valores mostraban una gran cantidad de publicaciones, siendo iguales a aquellos del rango con mayor cantidad de publicaciones en las habitaciones. Cómo era de esperarse, el máximo se da ante una menor cantidad de baños que de habitación, si se estudian en conjunto ambas visualizaciones, lo que nos da una idea de cómo ambas variables en conjunto pueden afectar el precio.

2.7.2. Precio según Baños



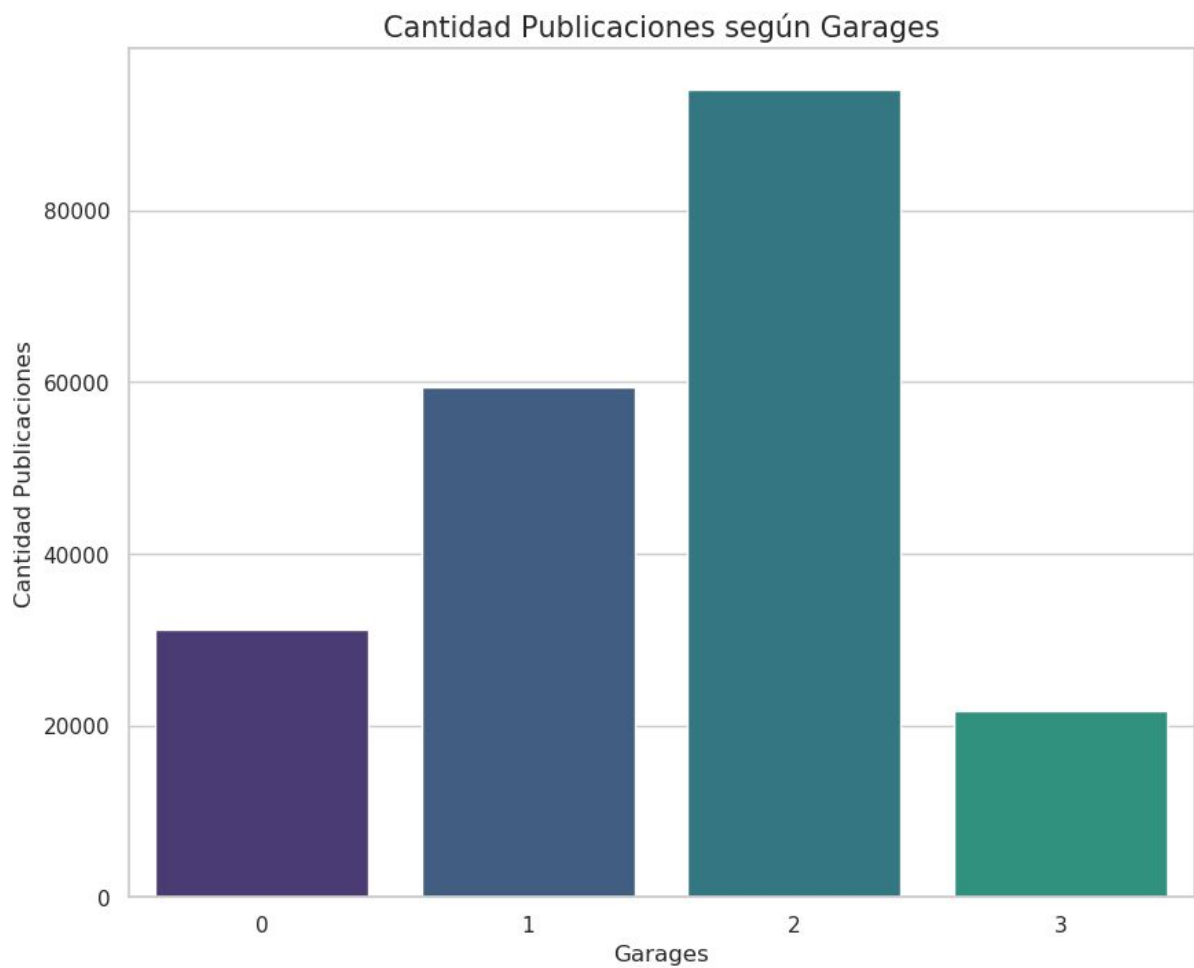
Como se intuyó en el análisis previo, al igual que en el caso de las habitaciones, el precio crece con la cantidad de baños y el rango de éstos en cada caso es muy similar a los valores que tomaban las publicaciones para las diferentes opciones de habitaciones.

2.8. Garages

El tercer y último punto que describe el tamaño y capacidad de una vivienda es la cantidad de garages que posee. En éste caso el rango de valores se encuentra entre 0 y 3, nuevamente comparable a los casos anteriores.

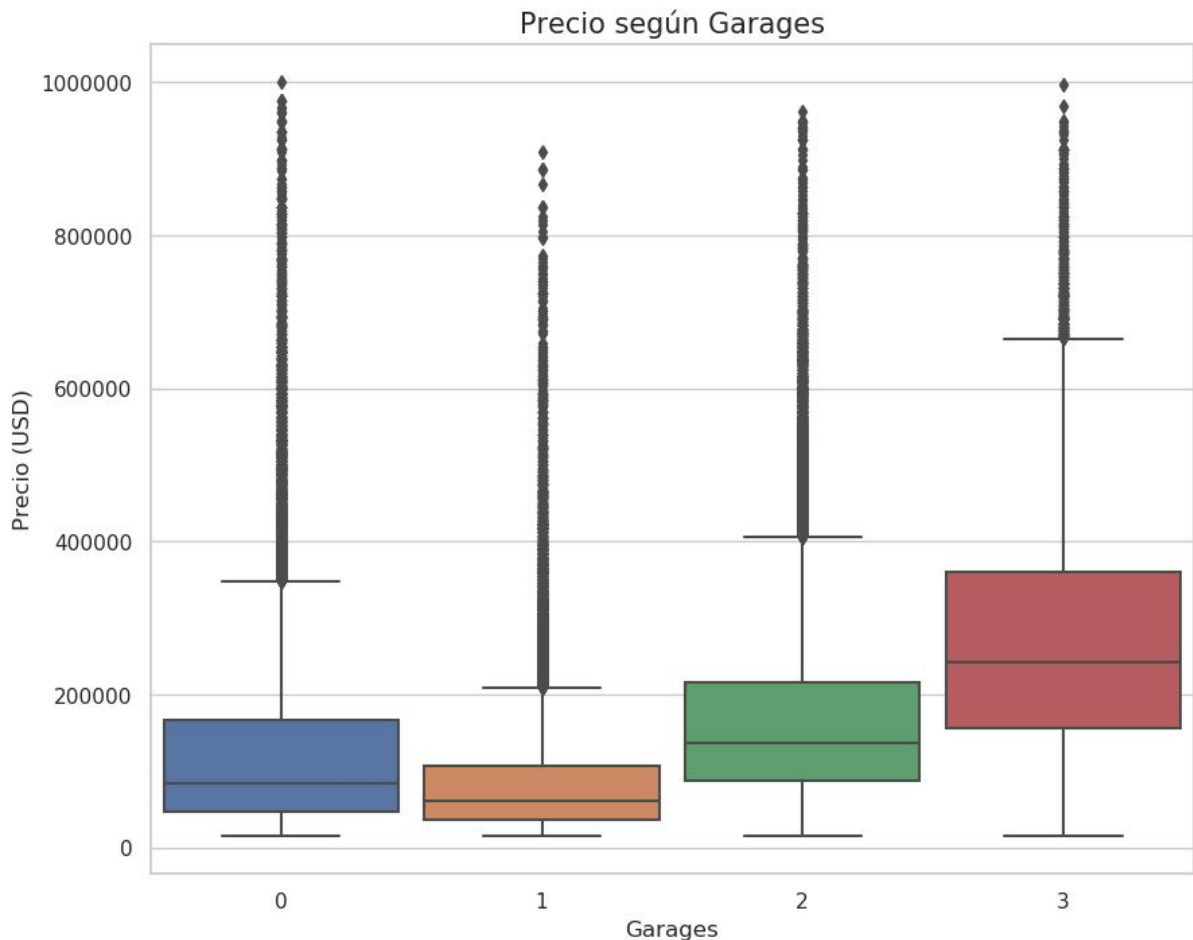
En cuanto a los valores nulos para éste punto se completó con 0 en aquellos tipos de propiedad que mostraran una clara mayoría de éste valor por sobre el resto, y se descartó el resto de datos que no mostraban una clara mayoría en sus valores, siendo un porcentaje menor.

2.8.1. Distribución de Garages



Nuevamente todos los valores presentan una gran cantidad de publicaciones, siendo 2 la cantidad superior. Ésto era de esperarse si se tiene en cuenta el comportamiento de las variables en los dos casos previos.

2.8.2. Precio según Garages



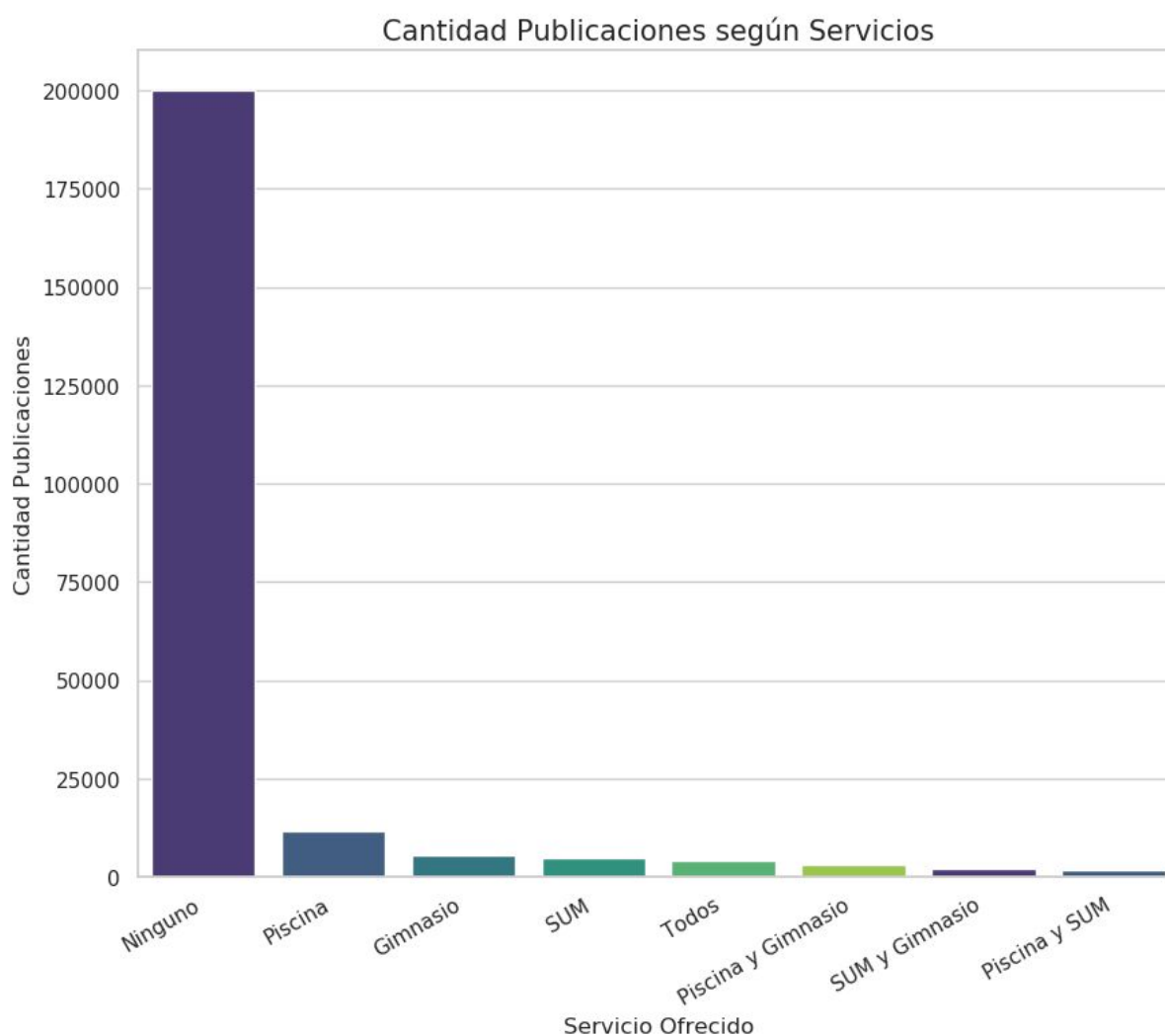
Una vez más se observa que a medida que aumenta la cantidad de la variable aumenta el precio. El hecho de que aquellas propiedades que no tienen garage se encuentran por arriba de las que tienen uno es entendible ya que dentro de éste grupo se encuentran propiedades de otro tipo distinto a casas o apartamentos, cuyos precios se fijan de otra manera.

2.9. Servicios

Algunas características extras de una propiedad incluyen los servicios que ésta ofrecen, teniendo en éste caso datos sobre si tienen o no piscina, gimnasio y salón de usos múltiples (SUM). Considerando que todos estos entretenimientos implican un gran gasto en caso de ser consumidos por fuera de la propiedad, nos parece importante evaluar cual es el impacto que tiene cada uno, así como la combinación entre ellos, a la hora de buscar una propiedad.

Se utilizaron para ello sólo aquellos tipos de propiedad que en algún caso presentarán alguno de éstos servicios, para poder comparar correctamente.

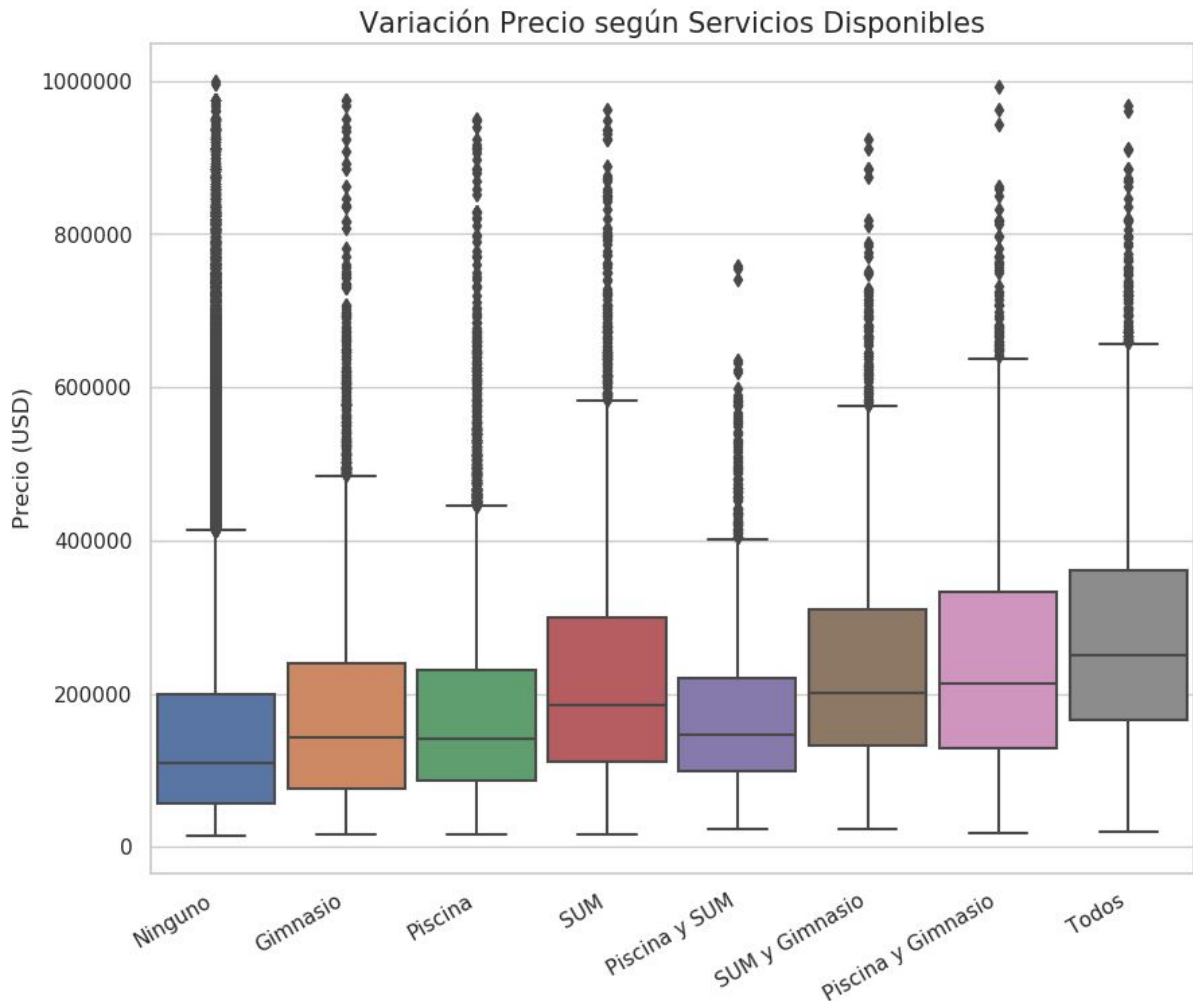
2.9.1. Distribución de Servicios



Más de un 80% de las publicaciones no presentan ningún servicio, y las siguen muy por debajo las distintas opciones por separado, teniendo recién en cuarto lugar las 3 juntas, dejando por debajo combinaciones de a dos.

Recordando que la gran mayoría de las publicaciones corresponden a casas tiene sentido que las mismas no tengan gimnasio ni SUM, pero si aparezca como primera opción que posean una piscina. El resto de los servicios son generalmente ofrecidos por los apartamentos y ésta debe ser la razón por la cual presentan valores más bajos de publicaciones.

2.9.2. Precio según Servicios



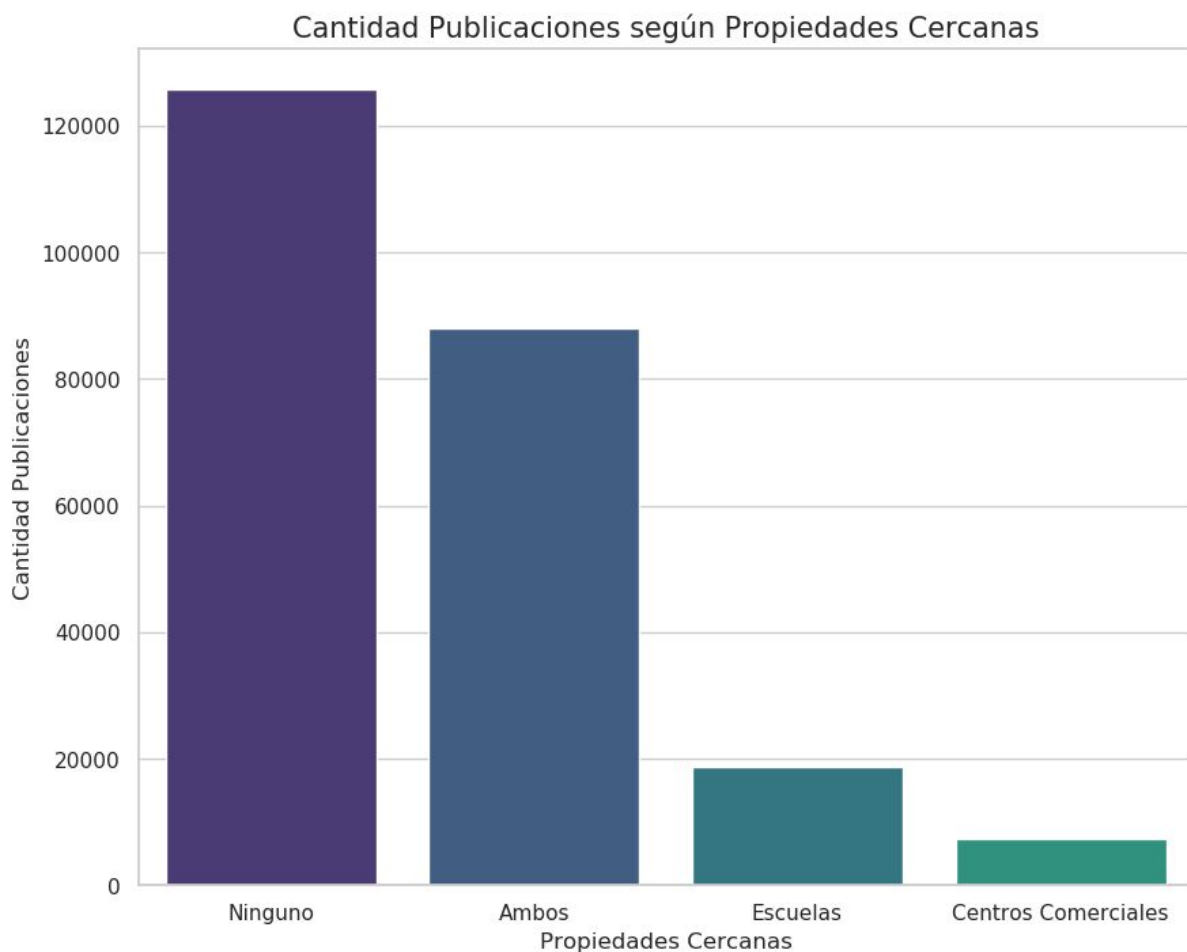
Individualmente, el servicio que más influye en el precio de una propiedad resulta ser el salón de usos múltiples, mientras que el que en más publicaciones aparece, la piscina, tiene una media y valor máximo menor a los otros. Para el resto de los valores es importante tener en mente la cantidad de publicaciones que se tiene para cada uno frente a la distribución de los precios y sus outliers.

2.10. Propiedades Cercanas

Dentro de las publicaciones se indica si tienen o no en sus cercanías tanto una escuela como un centro comercial. En principio nos generó dudas si esta característica de una propiedad podría influir en el precio final, más allá del interés particular de un cliente de ZonaProp de tener cerca alguno de éstos puntos, razón por la cual se generó este análisis.

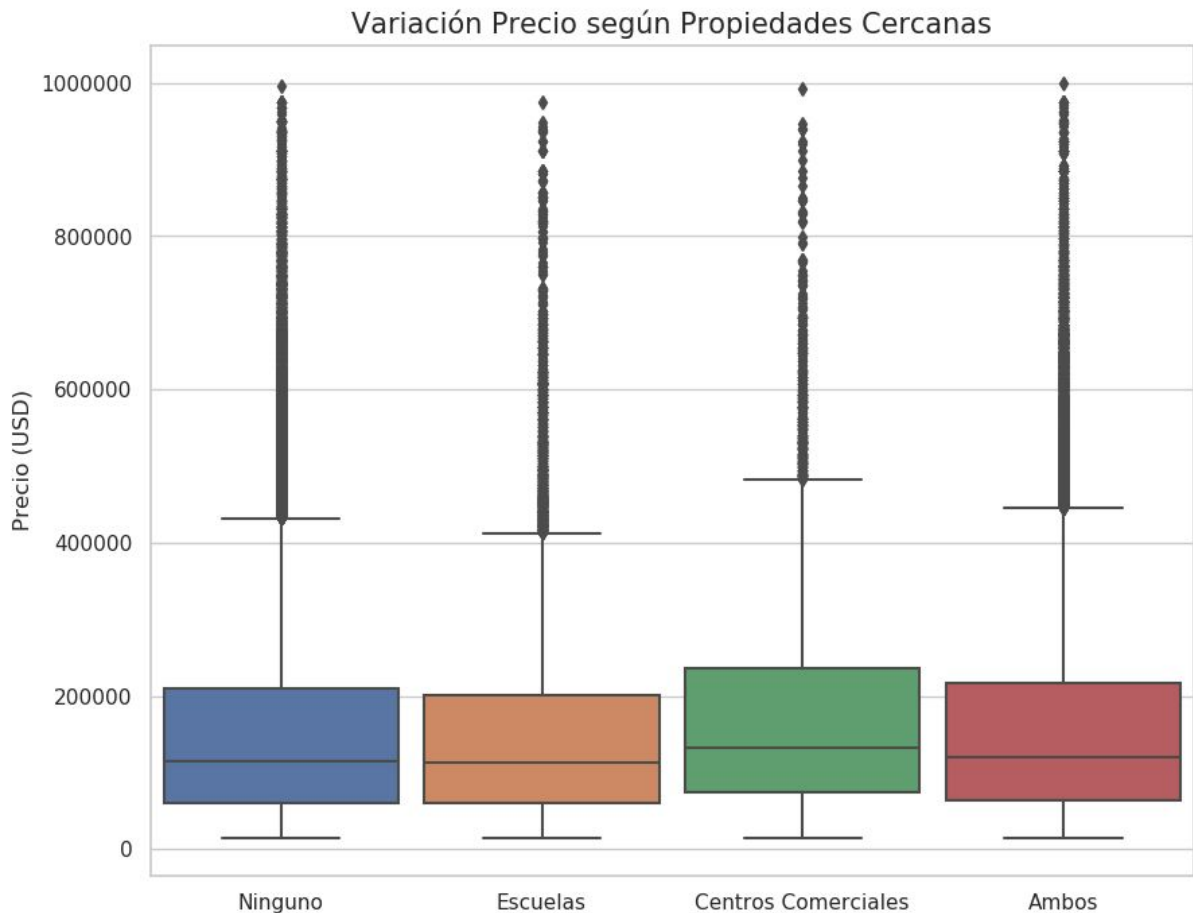
Como en el caso de los servicios se conservaron sólo los datos de aquellos tipos que en alguna publicación presentaran cercanía a alguna de éstas dos propiedades.

2.10.1. Distribución según Propiedad Cercana



Más de la mitad de las publicaciones no se encuentran cercanos a ninguno de ellos, mientras que los restantes presentan generalmente cercanía a ambos. Resta evaluar cómo afectan al precio éstos factores.

2.10.2. Precio según Propiedades Cercanas



En éste caso no se observa una gran variación en el precio según tenga o no centros comerciales y/o escuelas cercanas. Tanto los valores medios, como los percentiles, valores máximos y outliers se mantienen en una línea muy similar para los cuatro casos evaluados.

De ésta manera, a diferencia de las variables previas, a la hora de fijar el precio o buscar una propiedad el hecho de tener o no cerca alguna de estas propiedades no debería ser un factor influyente.

2.11. Términos utilizados

A la hora de tener que realizar una nueva publicación nos pareció interesante observar que palabras suelen compartir tanto los títulos como las descripciones de una propiedad. De ésta manera, se podría ayudar a que un usuario de ZonaProp cree una publicación interesante, atractiva y completa para así facilitar la venta de su propiedad.

2.11.1. Título



Una vez que se atrajo la atención a través del título es importante que la descripción se encuentre completa y detalle todos los puntos de una propiedad. Mediante ésta visualización podemos obtener un concepto sobre qué es importante no olvidar mencionar en nuestra descripción, y así no estar en desventaja frente a otras publicaciones.

En éste caso podemos observar que ya no hay un gran dominio de términos descriptivos sino que se mencionan todas las características que puede poseer la propiedad: los baños, recamaras, sala, comedor, cocina, etc.

Será importante entonces para los usuarios no olvidar mencionar ninguno de éstos puntos (en caso de ser necesario para el tipo de propiedad) para que quién visita su publicación no sienta que tiene pocos detalles sobre la propiedad en cuestión.

2.12. Metros Cuadrados

2.12.1. Limpieza de datos

Para esta sección vamos a validar un poco más los datos, para tener un análisis más preciso. Vamos a intentar salvar la mayor cantidad de valores nulos posibles, separando en 2 casos:

- Las filas que tengan "*metrostotales*" en null, pero con valores en "*metroscubiertos*".
- Las filas que tengan "*metroscubiertos*" es null, pero con valores en "*metrostotales*".

Esto lo hacemos para ver si hay posibilidad de llenar los null de la primera columna con el valor de esa segunda.

" <i>metrostotales</i> " nulos	51467
" <i>metrostotales</i> " nulos, con valores en " <i>metroscubiertos</i> " :	51467

" <i>metroscubiertos</i> " nulos	17400
" <i>metroscubiertos</i> " nulos, con valores en " <i>metrostotales</i> " :	17400

Al comprobar la cantidad filas con valores nulos en la primera columna y valores en la segunda, vemos que hacer un reemplazo es factible para ambos casos.

Procedemos a explicar el criterio usado.

2.12.1.1. Caso metros totales en NULL

Sin importar el tipo de propiedad, si el campo de *metros totales* se encuentra en null pero el campo de *metros cubiertos* tiene datos, se llenara los nulls con estos. Imaginamos que en el caso de propiedades del tipo terrenos, huertas, etc (donde podría predominar los *metros totales* por sobre los *metros cubiertos*) haya sido un error del usuario al ingresar los datos y los metros cubiertos sean los totales, dado que en caso de haber tenido más metros descubiertos es poco probable que no los haya tenido en cuenta, siendo que no sería negativo que la propiedad cuente con un terreno más amplio.

Y en el caso de casas, departamentos, etc, si nada más contaba con *metros cubiertos*, la cantidad de *metros totales* debería ser la misma dado que se le suman 0 metros exteriores.

2.12.1.2. Caso metros cubiertos en NULL

Para los tipos de propiedad como terrenos, huertas etc, en los cuales es más posible que el valor de *metros cubiertos* sea 0, incluso si hay datos en *metros totales*, procedemos a ver la correlación de esos casos, para ver si:

1. Llenar con 0 la columna de "*metroscubiertos*"
2. Si la hipótesis es falsa y la correlación es alta, llenaremos los nulos con el valor de la columna "*metrostotales*"

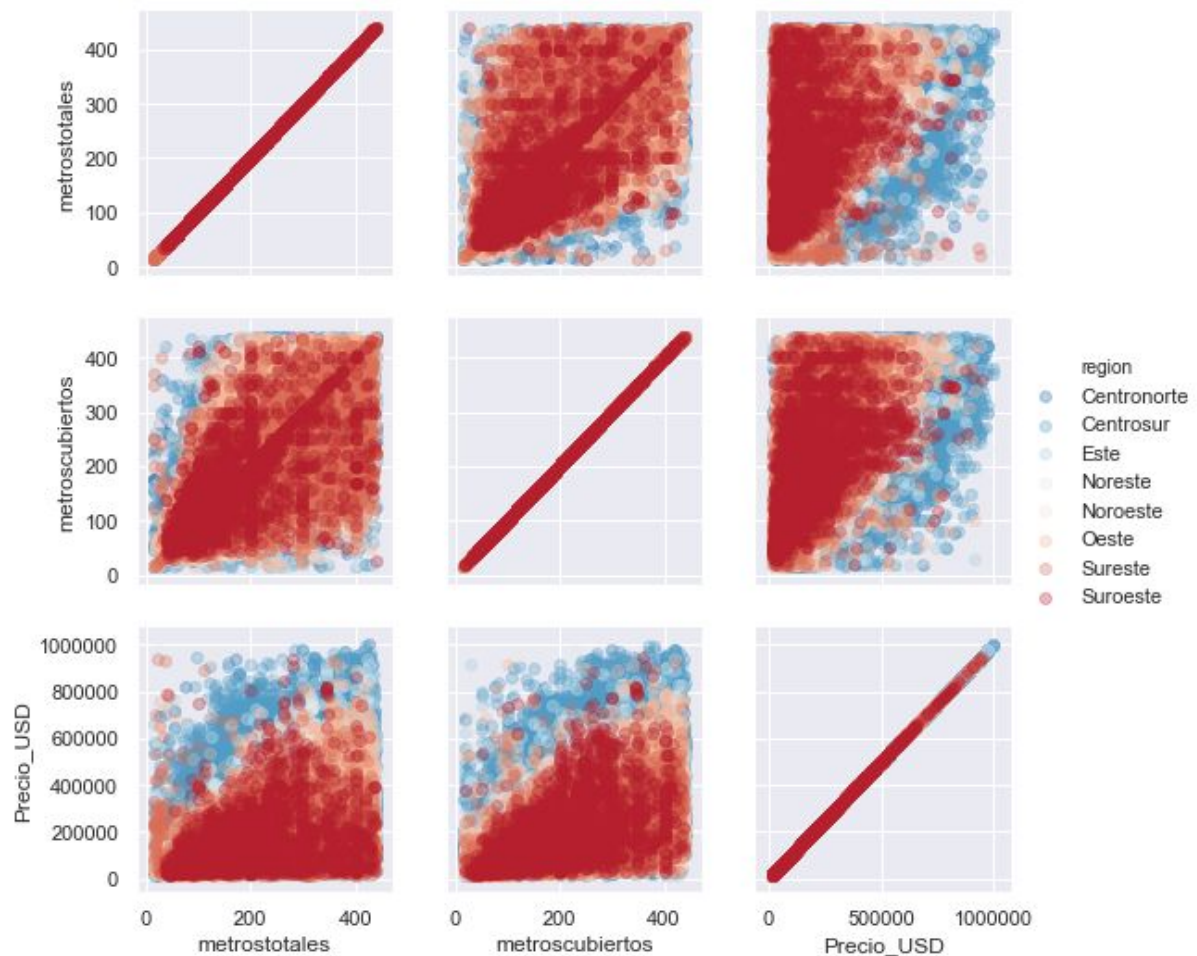
La correlación entre "*metroscubiertos*" y "*metrostotales*" en las propiedades que tienen 0 en "*metroscubiertos*" pero valores en "*metrostotales*" es:

Terreno:	0.79
Terreno comercial:	0.69
Otros:	0.81
Terreno industrial:	0.59
Huerta:	0.99
Garage:	1

Se puede ver que hay una gran correlación entre los *metroscubiertos* y los *metrostotales* incluso en este tipo de propiedades, con excepción de los *Terreno industrial* y *Terreno comercial* que son relativamente bajos en comparación por lo que procedemos a llenar los nulos con el valor de la columna que si tiene datos.

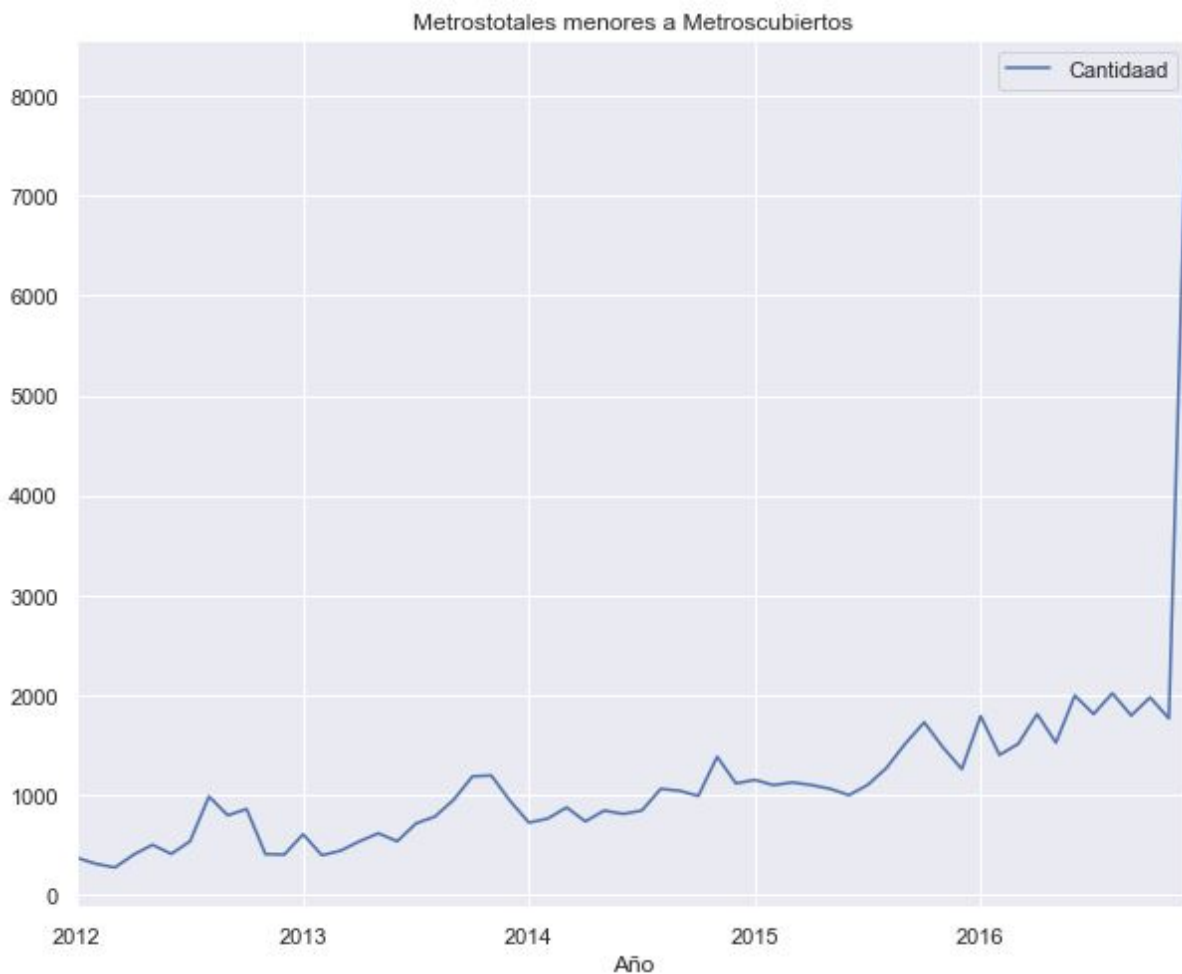
2.12.2. Análisis

Procedemos a ver la relación entre metros cuadrados y el precio en dólares para cada una de las regiones.



Viendo este gráfico rápidamente, nos damos cuenta por ejemplo que hay usuarios que colocaron *metrostotales* menores a *metroscubiertos* en las publicaciones, lo que no tiene sentido si pensamos que *totales* es igual a *cubiertos*+*descubiertos*, por lo que procedemos a eliminar las filas para esta parte del análisis, ya que aplicar que dónde *metrostotales* sea menor a *metroscubiertos*, asignar *metroscubiertos* o la suma de estos 2, podría dar lugar a falsas asunciones, siendo que la cantidad de los datos con este error son 70 mil de 240 mil.

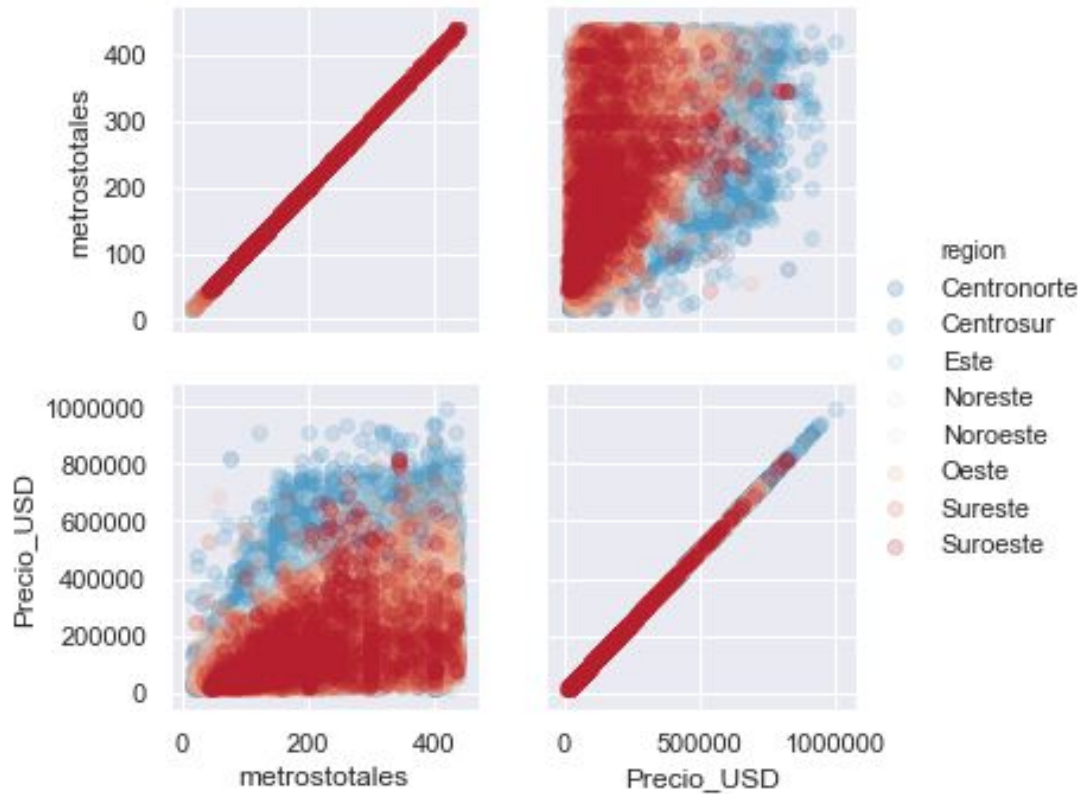
También procedo a ver la evolución de cómo se generaron estos datos mal cargados.



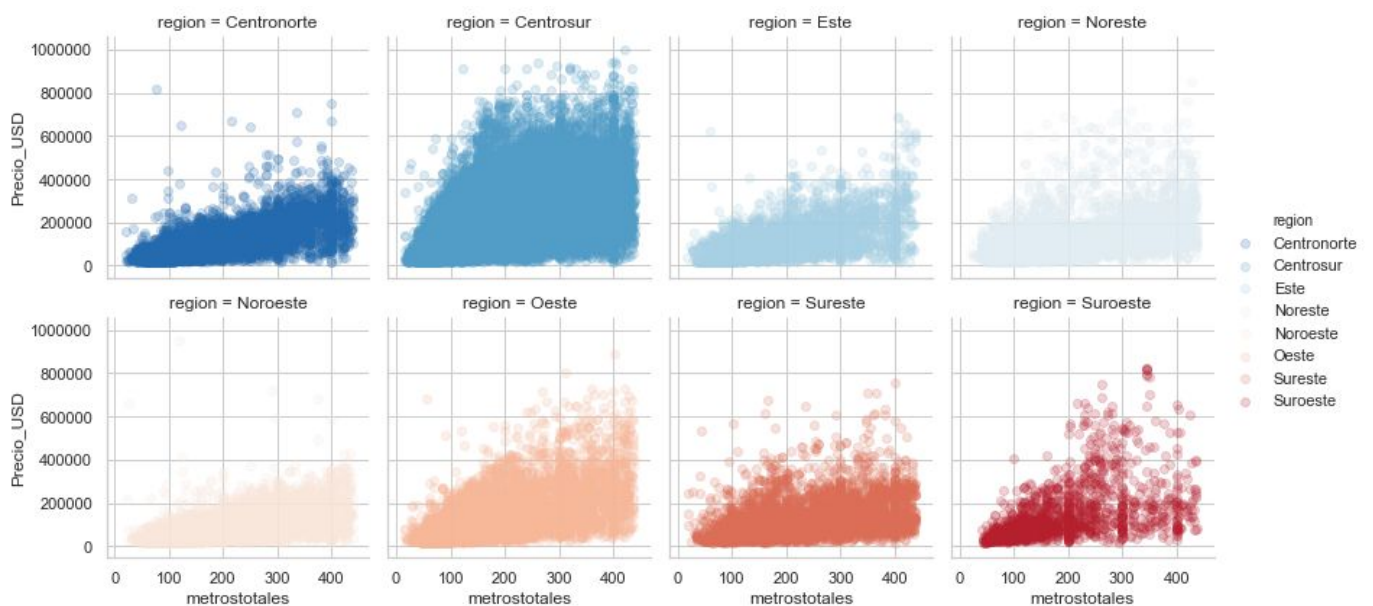
Más allá del crecimiento de los errores con el pasar de los meses (que nadamas siguen la curva de crecimiento de las publicaciones, no es que haya mucho más porcentaje de errores), lo que queremos ver en este set de datos mal cargados, es si la falla fue detenida en algún momento, por alguna validación agregada en el ingreso de los datos si alguien noto la incidencia.

Por lo que se ve, no es el caso y hasta la fecha final a la que llega este set, se siguen generando datos mal cargados, por lo que una validación en los campos que llenan los usuarios podría ser algo bueno a agregar para que en los análisis futuros los datos sean más consistentes.

El gráfico con los datos filtrados , si nos enfocamos en la relación de los *metros* de las propiedades con el *precio (USD)* resulta ser lo siguiente.



Se puede observar que hay una gran diferencia de precios entre las regiones del centro con las demás, pero para mayor claridad procedemos a separar el gráfico en regiones.

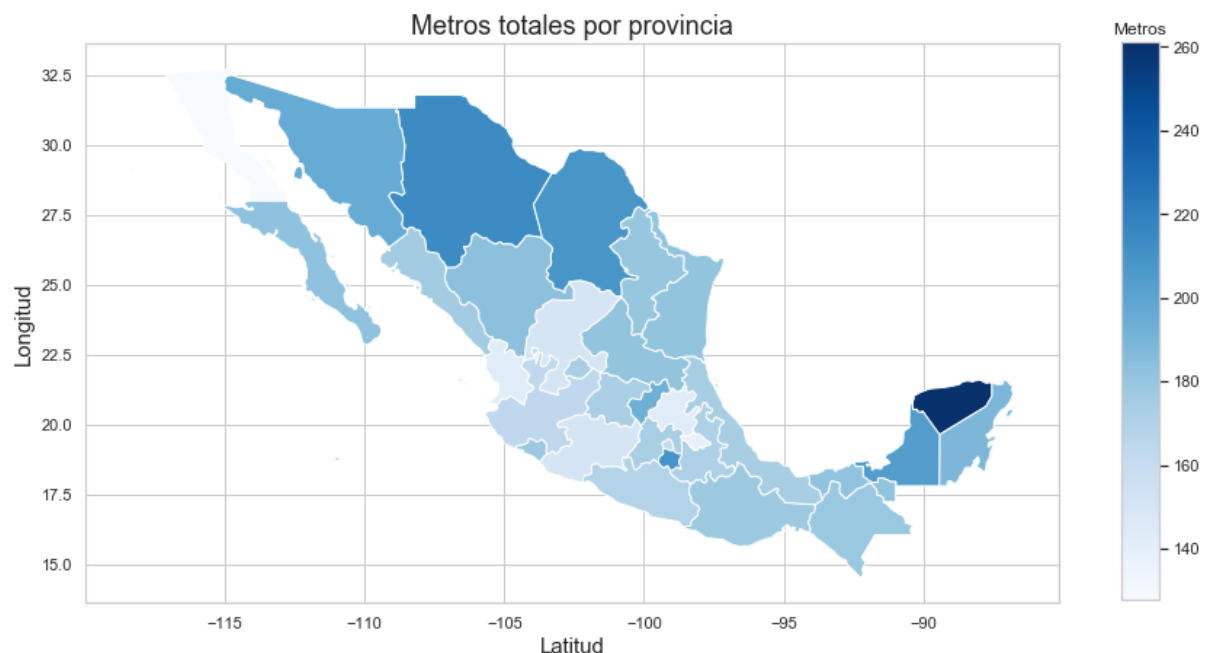


Ahora vemos que la región Centro Norte no se diferencia mucho de la mayoría, sino que la que más se destaca es la región del Centro Sur, que resulta ser la región que contiene a la capital de México, la que tiene los valores más altos por metro cuadrado. También por la densidad de puntos en su gráfico nos podemos dar cuenta de la alta cantidad de publicaciones que corresponden a esta región.

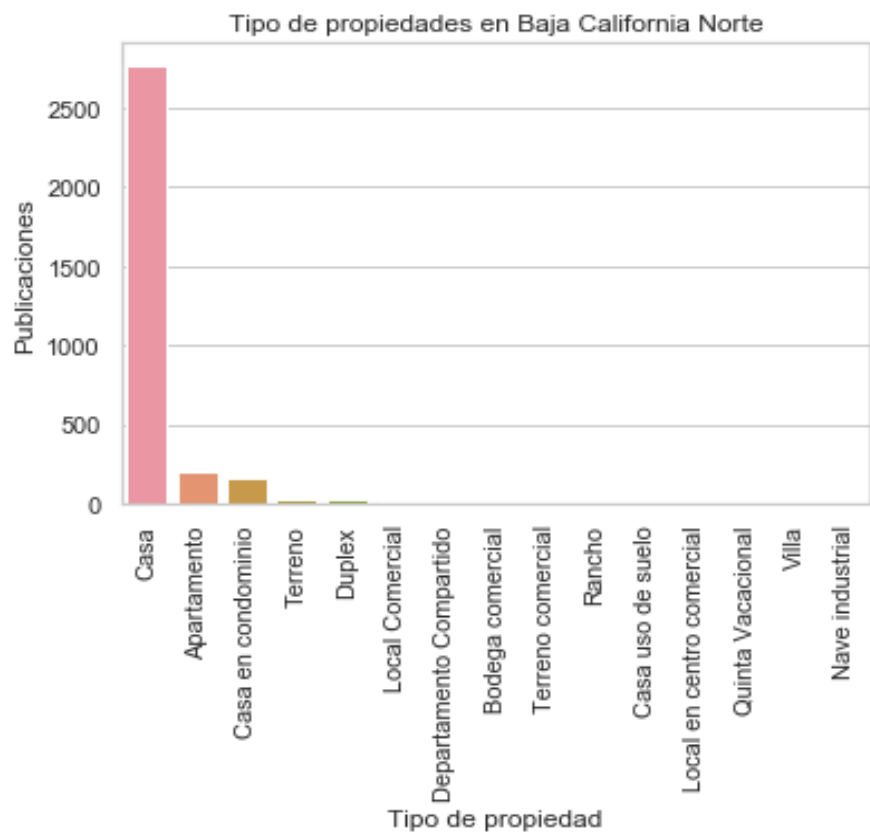
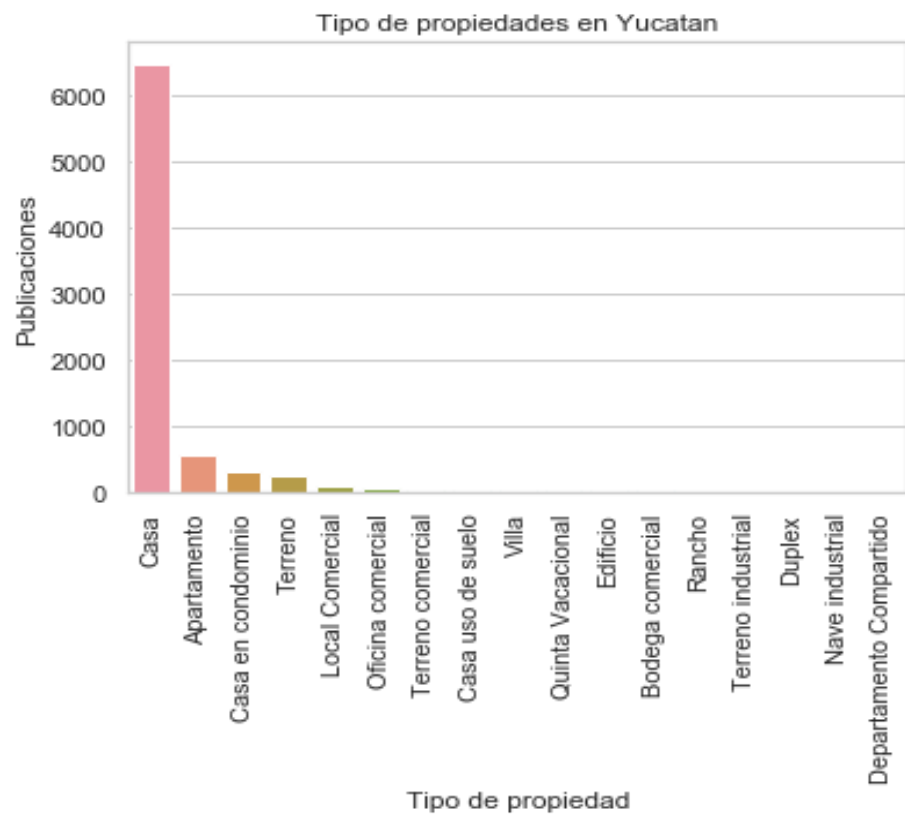
En las regiones que tienen más densidad, se puede ver con más claridad que para las propiedades de 200 o 300 metros cuadrados en adelante, los precios comienzan a estabilizarse, sin subir mucho más para lo que serían propiedades de hasta 450m², algo interesante desde el punto de vista del comprador, ya que por un precio muy similar podría obtener propiedades de hasta el doble de tamaño.

También resalta fácilmente, dada su densidad, la poca cantidad de propiedades mayores a 250m² de la región Suroeste, siendo también de las regiones con menor cantidad de publicaciones, como veremos más adelante.

2.12.3. Metros totales por provincia

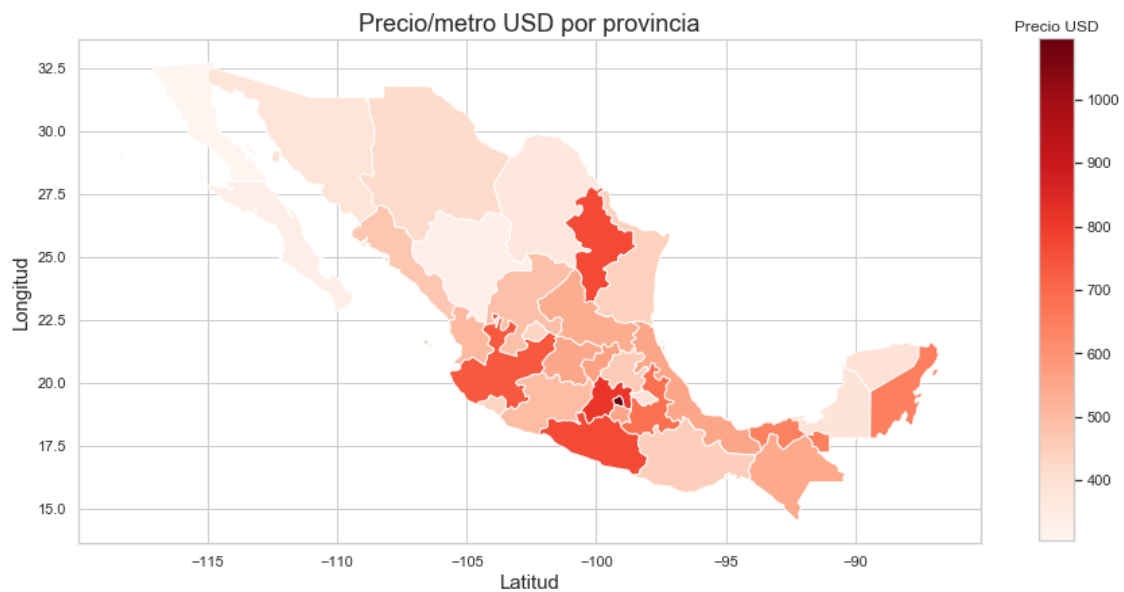


Es interesante ver cómo en este caso las propiedades más grandes pertenecen en general al Norte del País. Un caso particular se da en la provincia de Yucatán, donde es claro como ahí pertenecen las propiedades más grandes por amplia diferencia. También vale mencionar que las propiedades más pequeñas del país pertenecen a Baja California Norte.

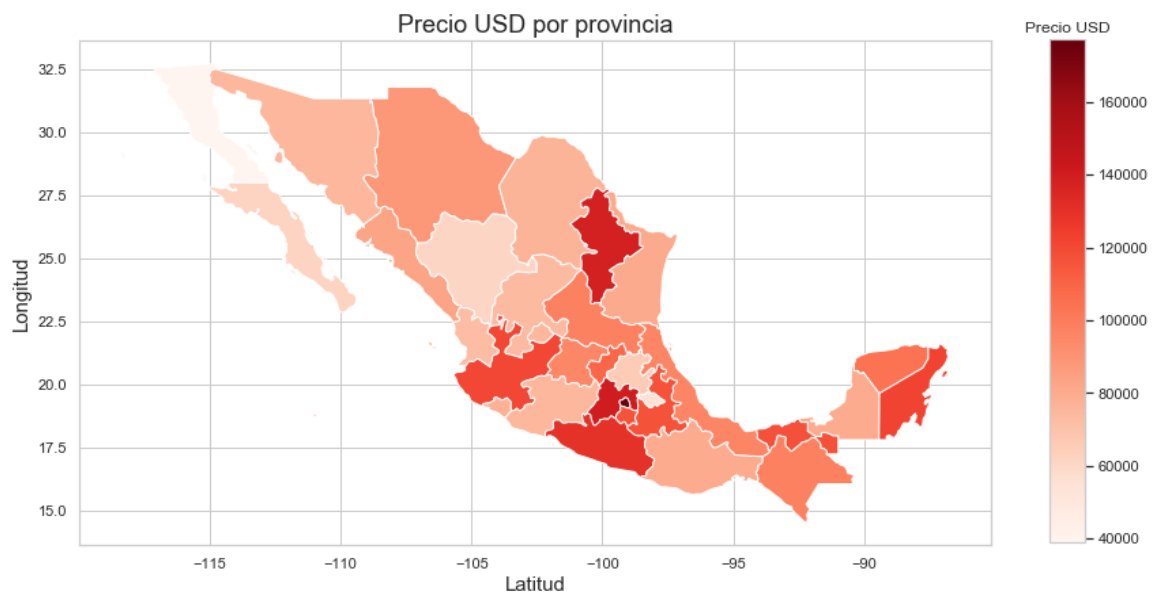


Al analizar que tipo de propiedades predominan en ambos extremos de la cantidad de metros totales, se ve que ambas son zonas donde las viviendas son casas. Este resultado es interesante ya que al ver que las viviendas de una provincia tienen el doble de tamaño que las de otra, uno imaginaría que esto se da como producto de que en una provincia las publicaciones son mayoritariamente casas cuando en la otra son departamentos, el cual no es el caso.

2.12.4. Precio/Metro por provincia



2.12.5. Precio por provincia



Como era de esperarse, las propiedades pertenecientes a la capital son las más caras. Otro aspecto que se puede observar es como el valor del metro cuadrado es menor en el Norte del país. Teniendo en cuenta análisis anteriores se puede decir que por esa zona se pueden obtener las casas con mejor relación tamaño/precio del mercado.

2.13. Regiones

Para esta sección se está tomando que las regiones están conformadas como:



Las regiones están conformadas por los estados de:

Noroeste: Baja California Norte, Baja California Sur, Chihuahua, Durango, Sinaloa, y Sonora.

Noreste: Coahuila, Nuevo León, y Tamaulipas.

Oeste: Colima, Jalisco, Michoacán, y Nayarit.

Este: Hidalgo, Puebla, Tlaxcala, y Veracruz.

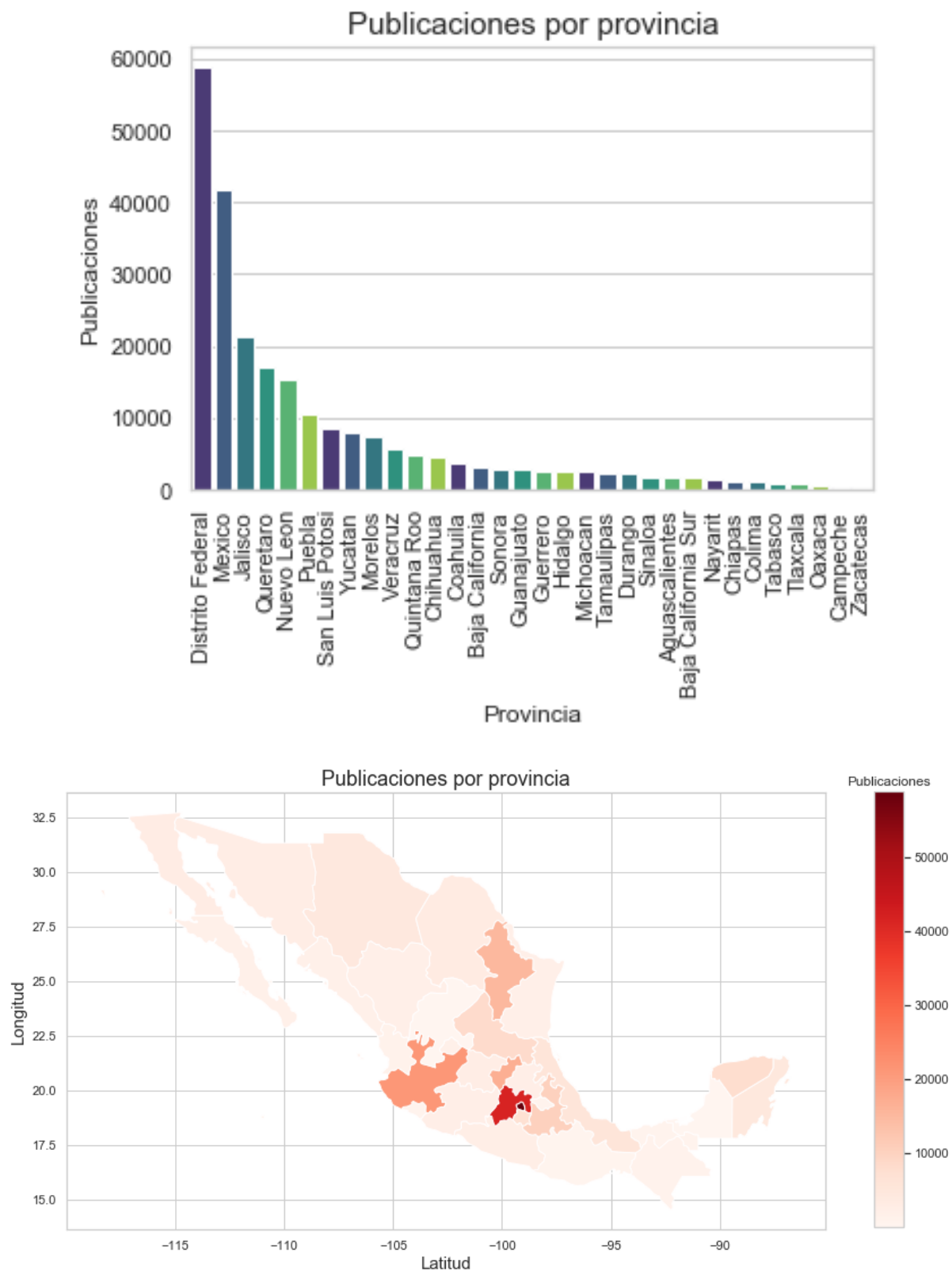
Centronorte: Aguascalientes, Guanajuato, Querétaro, San Luis Potosí, y Zacatecas.

Centrosur: Edo. de México, Distrito Federal, y Morelos.

Suroeste: Chiapas, Guerrero, y Oaxaca.

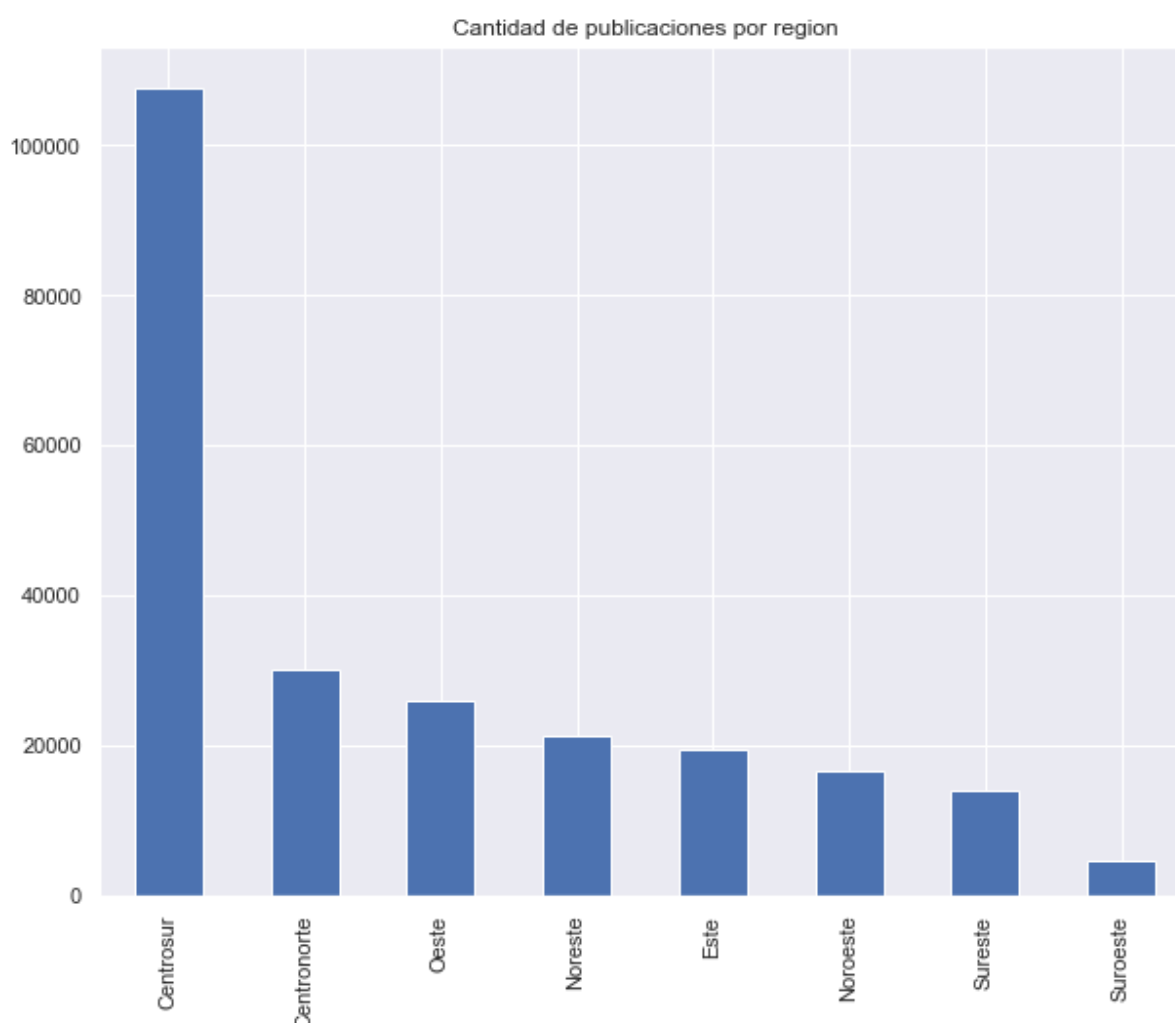
Sureste: Campeche, Quintana Roo, Tabasco, y Yucatán.

2.13.1. Publicaciones por provincia



Ambos gráficos muestran cómo la diferencia entre cantidad de publicaciones por provincia es muy grande, por lo que los datos analizados describen con mayor precisión los comportamientos en aquellas provincias con muchas publicaciones, como el Distrito Federal. En casos como el de Zacatecas, con una cantidad de publicaciones mucho menor, la información obtenida no siempre debería ser considerada como relevante.

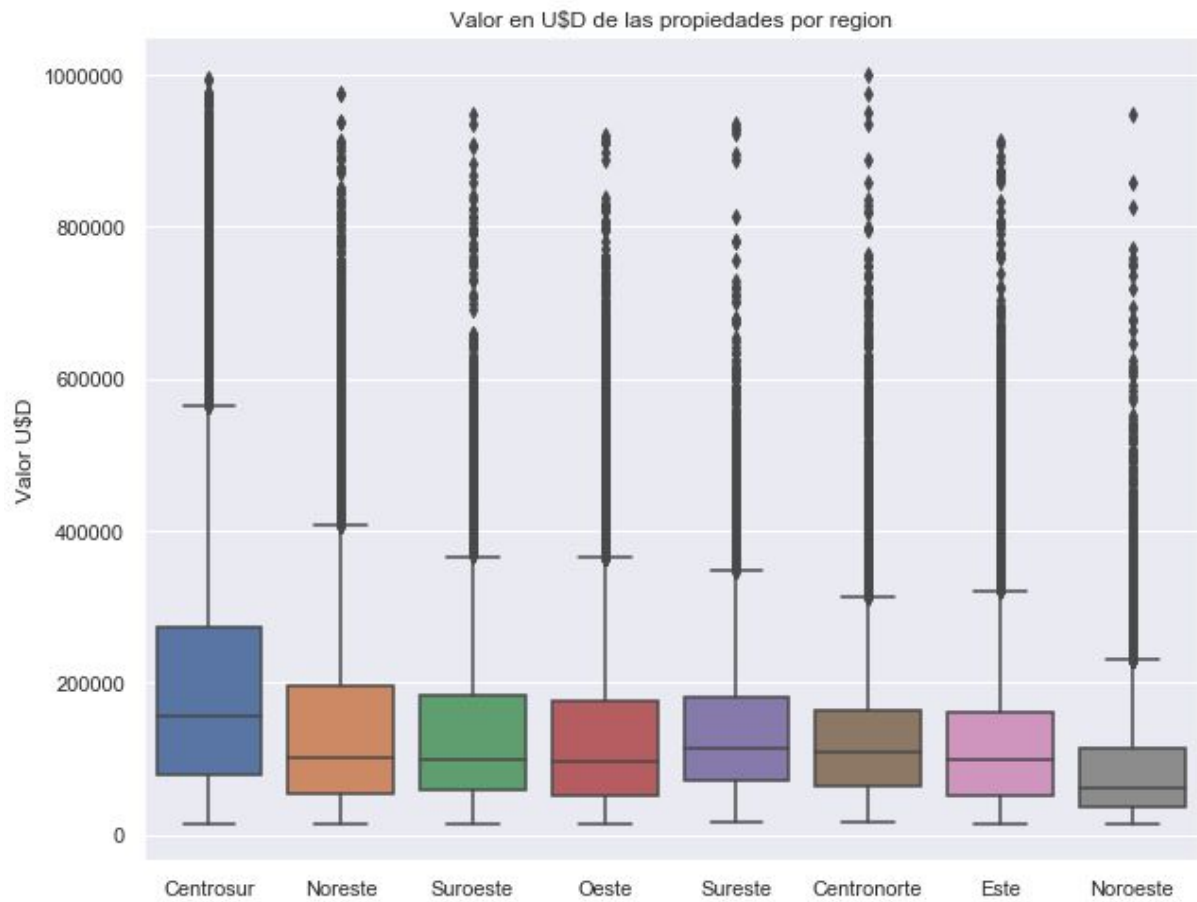
2.13.2. Cantidad de propiedades por región



La región del Centro Sur a pesar de ser de las regiones con menos cantidad de estados que la componen y la menor en cuanto a territorio, es la que tiene la mayor cantidad de publicaciones, siendo que el total de publicaciones del resto de México es alrededor de 132 mil y la región del Centro-Sur es de 107 mil.

Aparte de ser la región con más publicaciones, se destaca por tener la mayoría de las propiedades con los valores más altos de México, esto posiblemente se debe a que esta pequeña región de solo 3 estados abarca a la capital de México.

2.13.3. Distribución de los precios (USD) para cada región.



Se puede observar que la región con más concentración de propiedades de valor alto corresponden a la región Centro Sur. En las demás regiones se puede observar que la distribución de precios es muy similar.

3. Conclusión

Uniendo todo lo analizado en el transcurso de éste trabajo podemos extraer ciertos puntos centrales que nos resultaron interesantes y nos permiten llegar a algunas conclusiones.

3.1. ZonaProp y los precios

Como se observó la cantidad de publicaciones aumentó de gran manera en Diciembre de 2016 y previo a ésto el crecimiento fue continuo, por lo que si se mantiene el trabajo realizado en éste tiempo sería esperable observar aún más usuarios en el sitio en los años siguientes.

En cuanto a los precios dentro de México sucedieron importantes cambios en la moneda que influyen directamente en los precios de las propiedades. Manteniendo el peso mexicano dentro de las publicaciones marcaría un gran crecimiento y diferencia de precio entre publicaciones realizadas en un lapso de tiempo cada vez menor. De ésta manera una posible solución a este problema de comparación en las publicaciones sería agregar o reemplazar el precio a dólares, lo que generaría una mayor estabilidad en publicaciones subidas en meses diferentes y sería más sencillo para los usuarios que estén buscando una propiedad, en caso de que publicaciones más viejas no tengan su precio actualizado.

3.2. Tipos principales

El sitio se encuentra claramente dominado por tipo de propiedades destinadas a ser habitadas, como son las casas y apartamentos, el resto no tiene gran impacto en comparación dentro de ZonaProp. Los esfuerzos en desarrollo de nuevas herramientas dentro del sitio así como campañas publicitarias, etc. deberían tener el foco en éstos tipos habitables entonces. De ésta manera se podría esperar un impacto mucho mayor dentro de la actividad del sitio.

3.3. Determinación del precio de una propiedad

Un gran aporte al sitio sería ayudar a los usuarios que estén publicando su propiedad a determinar el precio de la misma. Gran mayoría de las secciones de éste trabajo tuvieron el enfoque en encontrar el efecto de cada característica de una propiedad

en el precio de la misma, por lo que uniendo todos estos datos se podría brindar herramientas a los usuarios de ZonaProp que, en base a ubicación, metros cuadrados, habitaciones, servicios, cercanías, etc., pueda orientar al usuario a encontrar el precio adecuado para su propiedad.

3.4. Ayudar a encontrar una propiedad

Con toda la información obtenida se podría orientar a usuario que estén buscando comprar una propiedad, en base a lo que están esperando encontrar. Completando los puntos sobre los cuales se tiene información en una publicación se podría brindar mucha información a los clientes: dónde será más fácil encontrar una publicación de esas características, cuántas publicaciones puede encontrar con características iguales o muy similares, que precio esperar en las mismas etc.

3.5. Ayudar a crear una buena publicación

Por último, nos parece interesante que la información obtenida pueda ser de ayuda a aquellos usuarios que estén creando una publicación en ZonaProp. Se mencionó previamente que se podría ayudar a calcular el precio aproximado de una propiedad, pero también se podría a recomendar a los usuarios como crear sus títulos y descripciones, recordando todos aquellos términos que se repiten en la mayoría de las publicaciones y ayudarán a que una nueva publicación no esté en desventaja (en cuanto a información y atracción a posibles compradores) frente a otras.