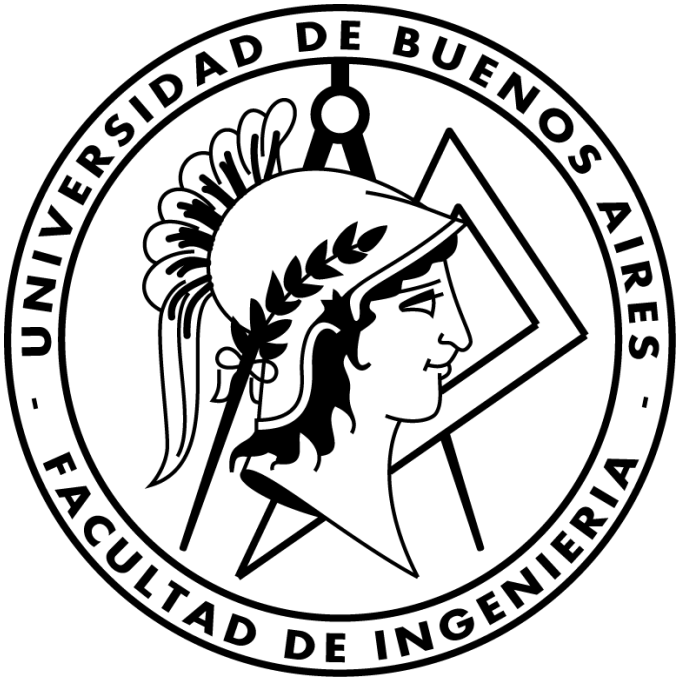
7506 - Organización de Datos: TP 1

Zonaprop - Análisis exploratorio de datos

Grupo 2 - Alumnos:

- Cassi, Julián
- Cozza, Fabrizio
- López Lecube, Lucio
- Rozanec, Matías



Facultad de Ingeniería - Universidad de Buenos Aires
2c 2019

Índice

Introducción

En el presente trabajo práctico se describen los resultados más importantes obtenidos mediante el análisis exploratorio de datos realizado sobre el dataset provisto.

En primer lugar se trabajó sobre el análisis general de las variables, observando con qué features se cuenta, de qué tipo, cuántos valores únicos hay por feature, visualizando las distribuciones de las distintas variables y las posibles relaciones entre ellas.

Una vez que se obtuvo el panorama inicial de los datos, el segundo paso consistió en buscar cosas más allá de lo que se obtiene a primera vista: se intentó utilizar la información espacial para poder graficar de alguna forma los distintos features sobre algún mapa. Este paso se considera de gran importancia, debido a que los nombres de las ciudades en sí mismos no dicen nada, pero proyectando los datos sobre un mapa probablemente se obtengan conclusiones que de otra forma permanecerán invisibles o les sería más difícil encontrar un sentido.

Es importante notar que en el informe se presentan los resultados que tienen algún valor de presentación. Todo lo que se fue intentando en el camino se puede leer en los notebooks en el repositorio.

Objetivos

El objetivo del presente trabajo es obtener un conocimiento lo más detallado posible de los datos dados, teniendo en cuenta que esto debería dejar el camino preparado al segundo trabajo práctico, cuyo objetivo será poder predecir precios desconocidos de propiedades.

Herramientas utilizadas

El análisis exploratorio se realizó en Python 3 con la librería Pandas. Para la visualización se utilizaron las librerías matplotlib, basemap, plotly y seaborn.

Para el control de versiones se utilizó Git; todo el tp se encuentra en GitHub en https://github.com/rozanecm/7506_2c2019_tp1 .

API Google-style Geocoder disponible en http://www.datasciencetoolkit.org/ para obtencion de datos de geolocalizacion.

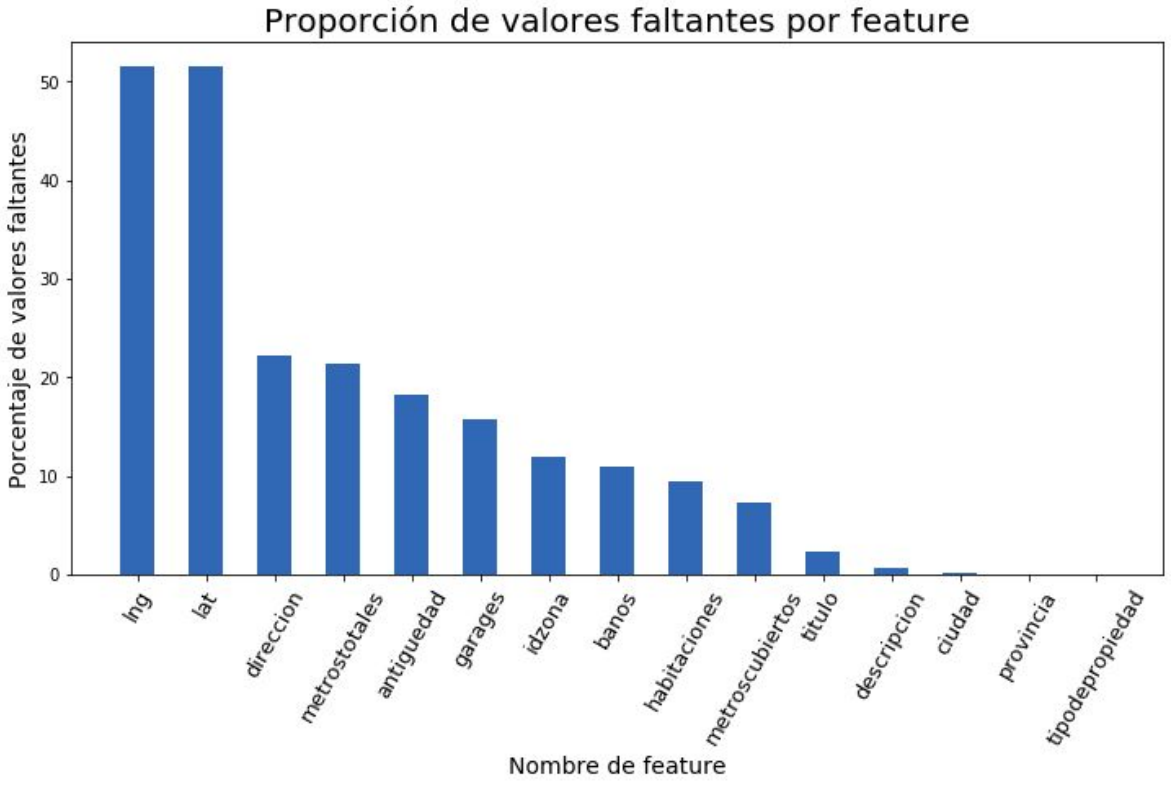
Libreria para trabajar con información en archivos html y xml:
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Herramienta para procesamiento del lenguaje natural: https://spacy.io

Análisis exploratorio

¿Qué datos… no hay?

En primer lugar es de interés conocer con cuántos datos se cuenta. Hay 240000 registros con 23 features cada uno. A continuación se presenta la proporción de valores faltantes por cada uno de los features que tiene al menos algún registro que no cumpla con la totalidad completa.



Latitud - longitud: ¿dirán la verdad?

Veamos ahora si la latitud y longitud son valores reales o si solamente aportan información relativa de ubicación entre las propiedades. En el segundo caso servirían para inferir información de distancia entre las distintas propiedades de forma bastante precisa. Como el dato está presente en solamente la mitad de los registros, habrá que estudiar cómo convendría utilizar este dato.

Se buscaron en Google Maps los primeros registros con los campos de latitud, longitud y ciudad completos, y se verificó que las coordenadas son reales, ya que las mismas pertenecen a la ciudad o quedan fuera por muy poco, por lo tanto la información será considerada confiable.

A partir de esta observación se cree conveniente completar el dato de las coordenadas faltantes con las coordenadas de las ciudad en los casos en que los registros tienen dicha

información. Dicha información contará obviamente con cierto error, pero por otro lado no se puede saber si el valor de las coordenadas no fue alterado levemente para no revelar las ubicaciones exactas, cosa que es totalmente esperable.

Se ha comprobado que los registros que no tienen longitud son los mismos que no tienen latitud, y que de los 123488 registros que no tienen información de latitud y longitud, 28298 no revelan la dirección; 111 tampoco revelan la ciudad, y 83 ni siquiera la provincia.

Para poder completar la información de las coordenadas faltantes se recurrió a los kernels de Kaggle, luego de haber observado que localmente se procesaban 50 requests por minuto y en Kaggle 87. Además Kaggle cuenta con la ventaja de que poder hacer commit de un kernel y dejarlo corriendo sin necesidad de que el usuario mantenga el kernel abierto, por lo que se pudo dividir el dataset de registros faltantes de forma que cada parte tome menos de 9 horas, y se pudo ir dejando que trabaje Kaggle en background mientras el grupo seguía analizando otros aspectos. De cualquier forma fue una tarea que Ílevó bastante tiempo, pero el grupo está convencido de que la misma valió la pena.

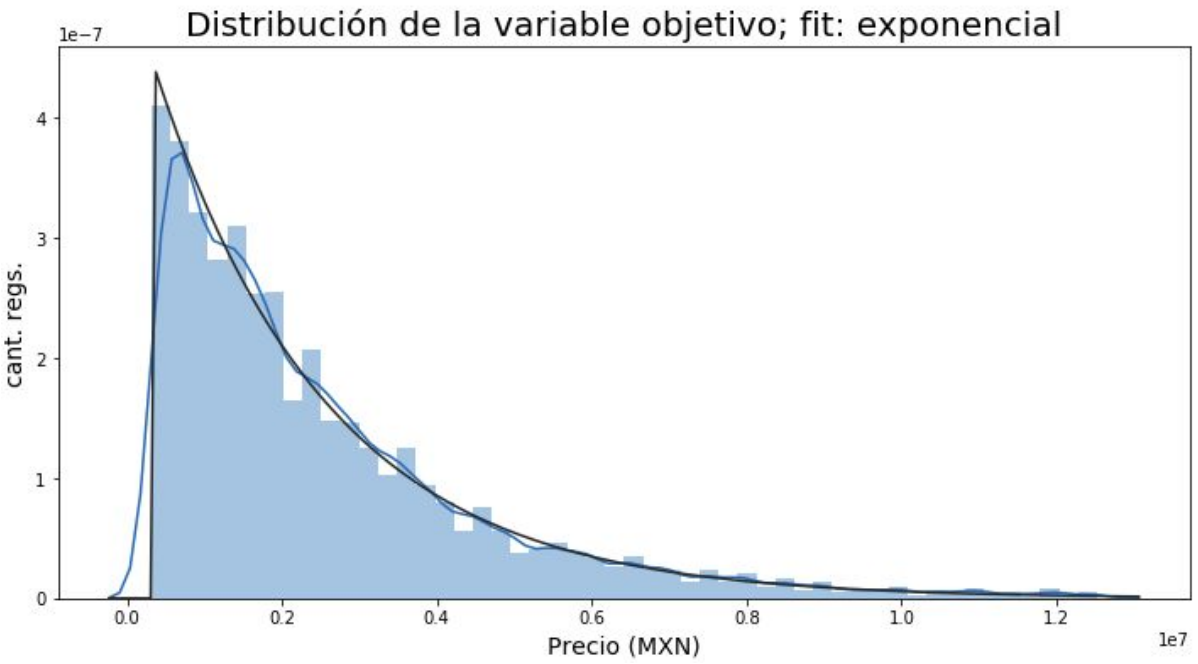
Utilizando la API Google-style Geocoder disponible en http://www.datasciencetoolkit.org/ se procedio a disparar requests por cada registro que contaba con datos en los campos <Dirección> y <Provincia> (un total de 186787) pudiendo lograr 72428 requests exitosos.
De esos requests se han podido recuperar 21886 lat/lng nuevas dejando un número de registros nulos equivalente a 101602.

Se ha observado durante el proceso que entre los registros con coordenadas faltantes algunos tienen direcciones inválidas. Indudablemente sería bueno encontrar una forma de detectar dichos campos y ver si no se ha filtrado información del campo de la descripción.

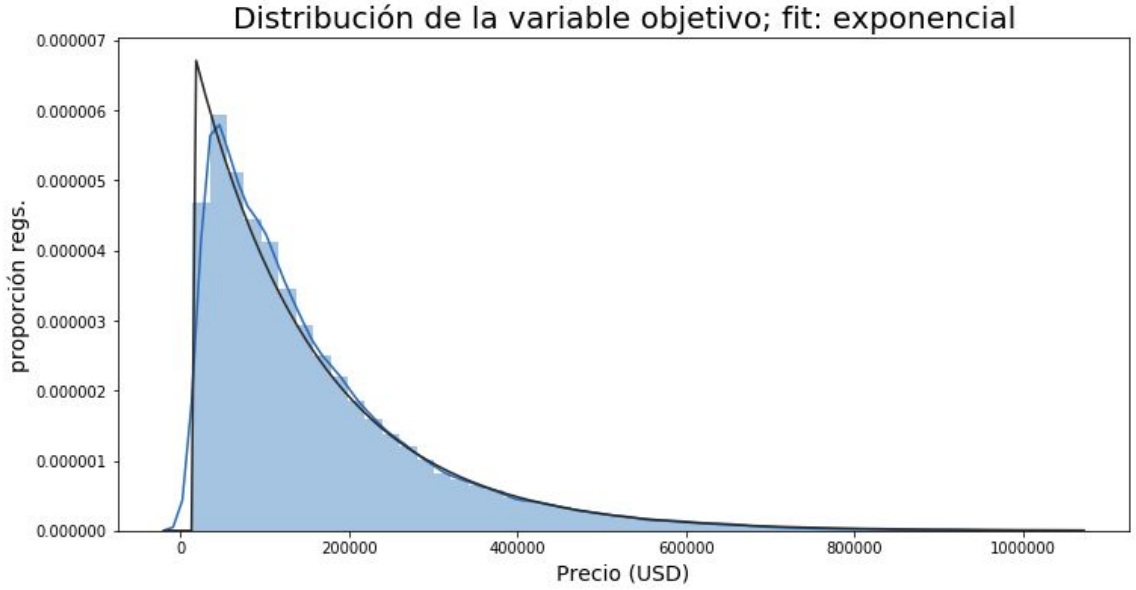
Análisis de la variable objetivo: Precio

A continuación se decidió explorar la variable objetivo, que en este trabajo refleja el precio de las propiedades estudiadas. En primer lugar se obtuvo la imagen del histograma con la distribución. En la misma se logró observar el fenómeno de la cola larga.

Para profundizar el análisis se fitearon varias distribuciones conocidas para ver con cuál coincidía mejor y así poder identificar mejor la distribución que presentan los precios, llegando a la conclusión de que la variable tiene una distribución exponencial. Además se probaron las distribuciones Normal, log-normal, power log-normal, Johnson SU, alpha, beta y burrow, que tienen una forma similar a la exponencial a simple vista, pero que se comprobó que no fittean tan bien a los datos como la exponencial.



Por otro lado se decidió pasar los precios a dólares americanos, ya que se observó que hace cinco años un dólar equivalía a aprox. 13 pesos mex., mientras que hoy día el precio ronda los 20. Como el set de datos parte del año 2012, se considera una parte importante del análisis.



Se puede observar que este gráfico en particular preserva la misma forma, lo cual es lógico, ya que a pesar de haber cierta distorsión en los datos por el valor de la moneda local, la misma no impacta tan fuertemente como para variar la distribución de la variable.

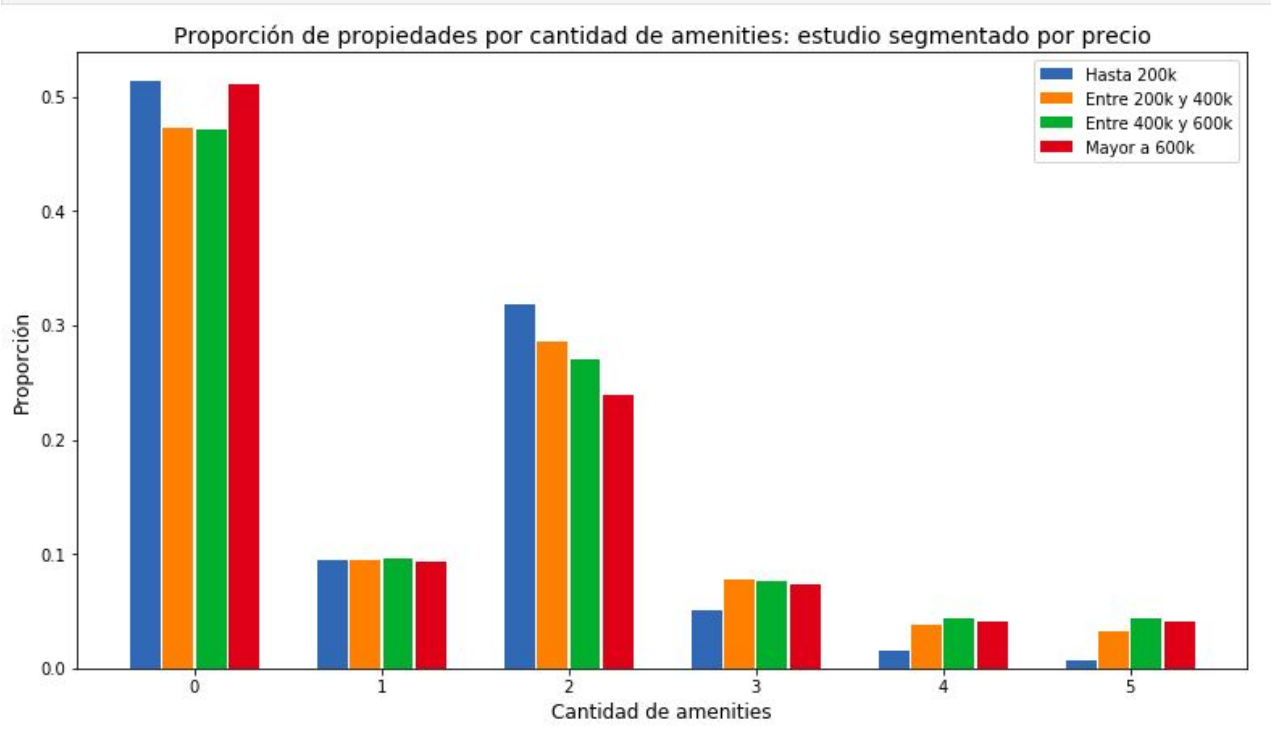
Con esta información se puede adentrar en un estudio de las propiedades segmentado por precios. Se decide fragmentar de la siguiente manera: precio hasta 200k, precio entre 200k y 400k, precio entre 400k y 600k, precio mayor a 600k, todo en dólares.

En primer lugar se observa que la distribución de los tipos de propiedad se mantiene igual (o casi) en todos los casos. Se esperaba en cambio encontrar alguna variación de las clases mayoritarias en los distintos segmentos de precios, pero no se dió.

Precio en relación a las amenities

Por otro lado se decidió observar cómo se distribuyen las proporciones de las propiedades en relación a la cantidad de amenities que tienen en cada segmento.

Llamaremos ‘amenities’ a las propiedades booleanas que indican si la propiedad cuenta con gimnasio, piscina, escuelas o centros comerciales cercanos

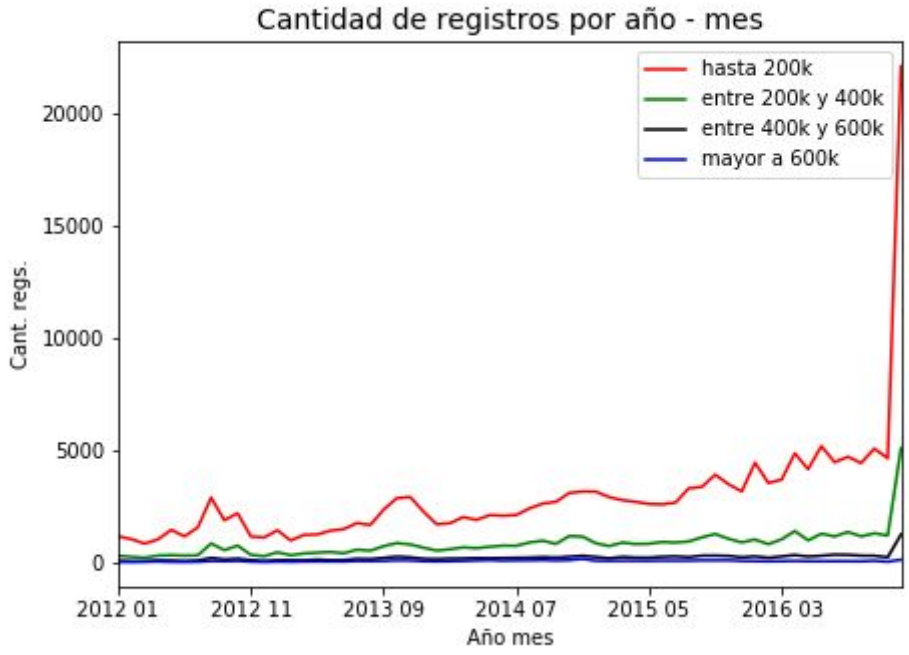


Este plot muestra que en todos los casos las proporciones internas se mantienen. En algunos casos las diferencias son más marcadas que en otros, pero en todos los casos se llega a las mismas conclusiones.

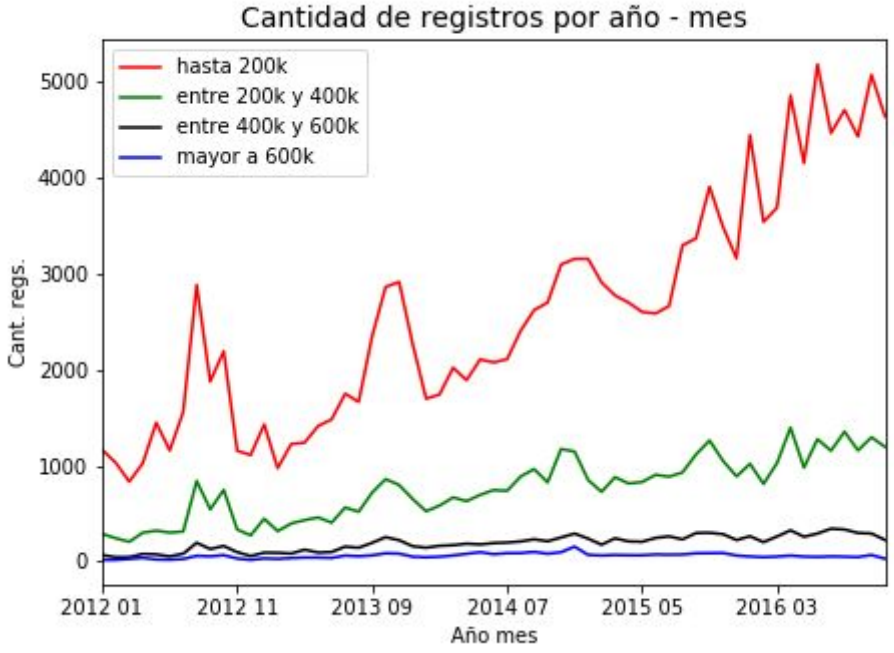
Precio en el tiempo

Al mismo tiempo, es de interés ver cómo evoluciona la cantidad de registros de propiedades que se tiene a través del tiempo, y con qué antigüedad cuentan, para tener una mejor idea como explicación de los precios.

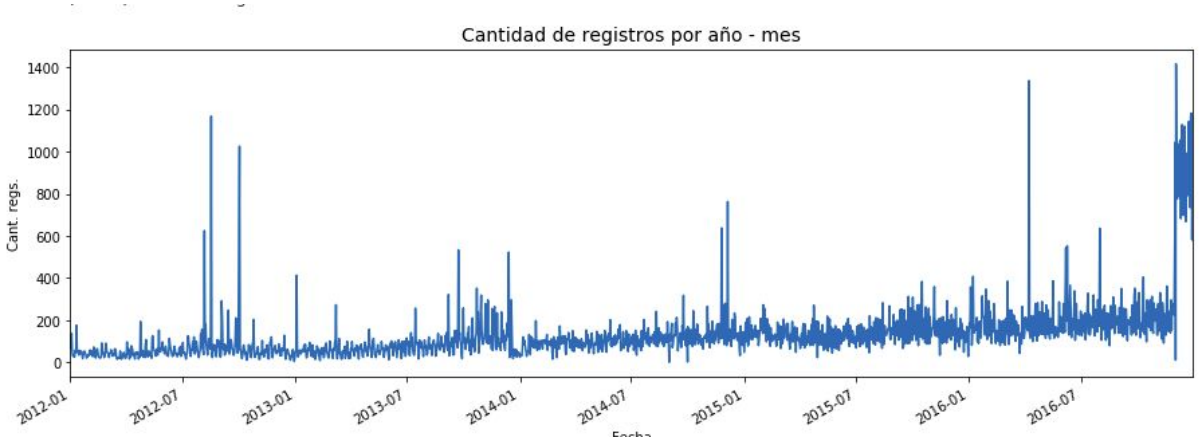
En diciembre de 2016 hay un clarísimo auge de registros, mientras que en los períodos previos la cantidad de registros ha ido teniendo una tendencia a crecer a lo largo del tiempo. Agrupando según la misma fragmentación por precio de antes se obtiene el siguiente gráfico:



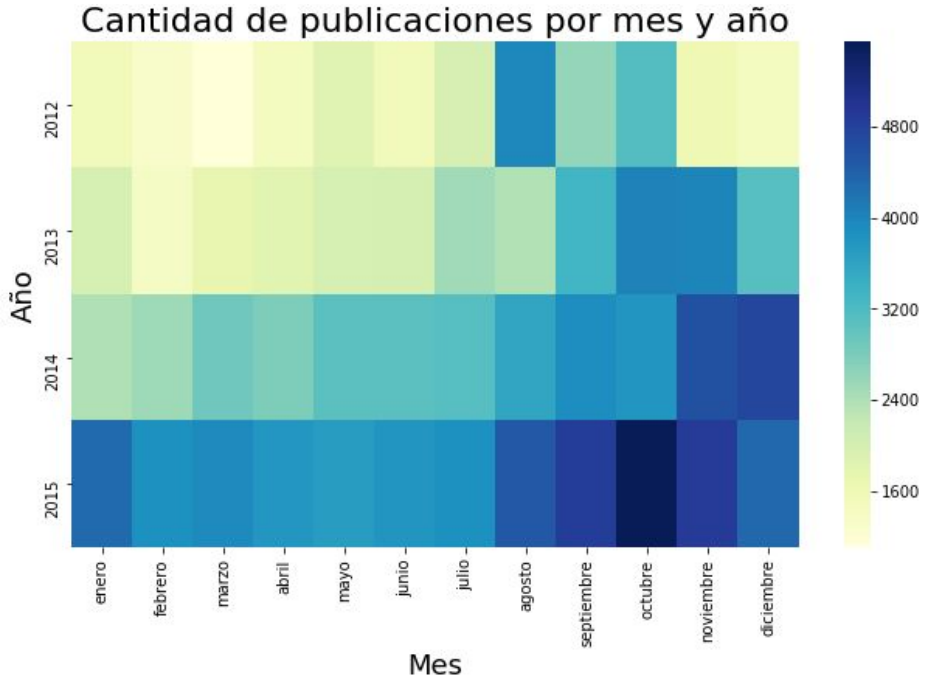
De este gráfico se puede destacar el gran porcentaje de propiedades que se encuentran en los segmentos más baratos y su aumento a lo largo del tiempo.



Eliminando los registros de diciembre 2016 se puede observar un comportamiento cíclico de registro de propiedades a la plataforma. Además, si en lugar de agrupar por año y mes se agrupa por día, se puede ver que hay grandes picos en días puntuales a lo largo de la historia del sitio.



A continuación se muestra la cantidad de registros por combinación de mes y año. Se excluye al 2016 ya que la gran cantidad de registros de diciembre de ese año hacía que el resto de los rectángulos tuviese color casi idéntico. Se observa que en todos los casos la mayor cantidad de publicaciones se produce en la segunda mitad del año.

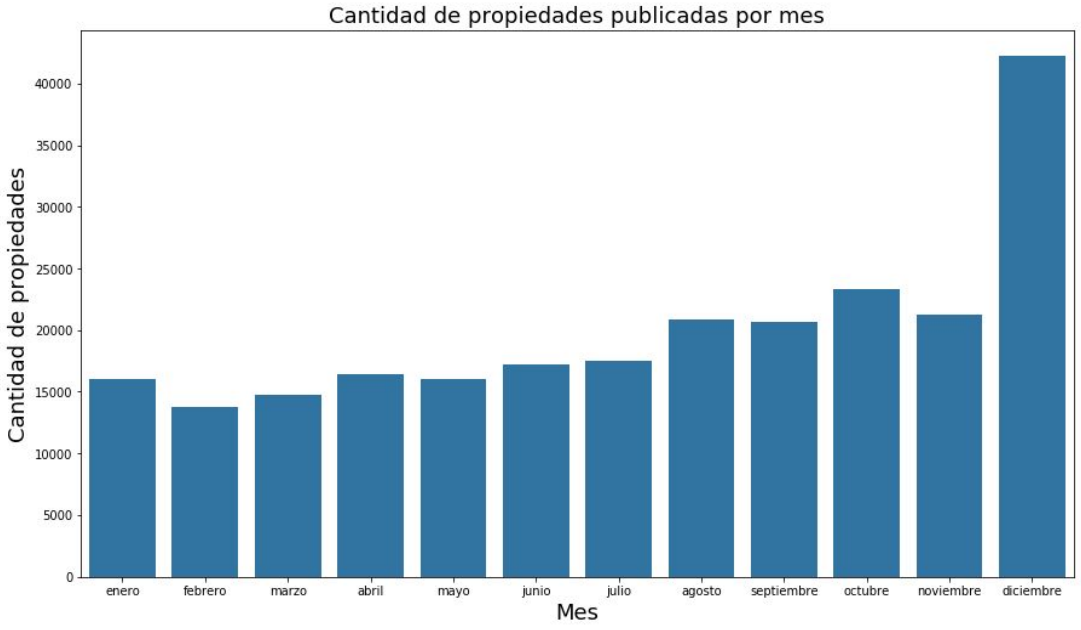


A continuación se muestra el precio promedio de acuerdo a la fecha de registro.

Es muy difícil sacar algún patrón o conclusión a partir de este gráfico. Uno que probablemente tenga mucho más sentido estudiar es la evolución del precio conforme a la antigüedad, donde se espera que a medida que aumenta la antigüedad decrezca el precio.

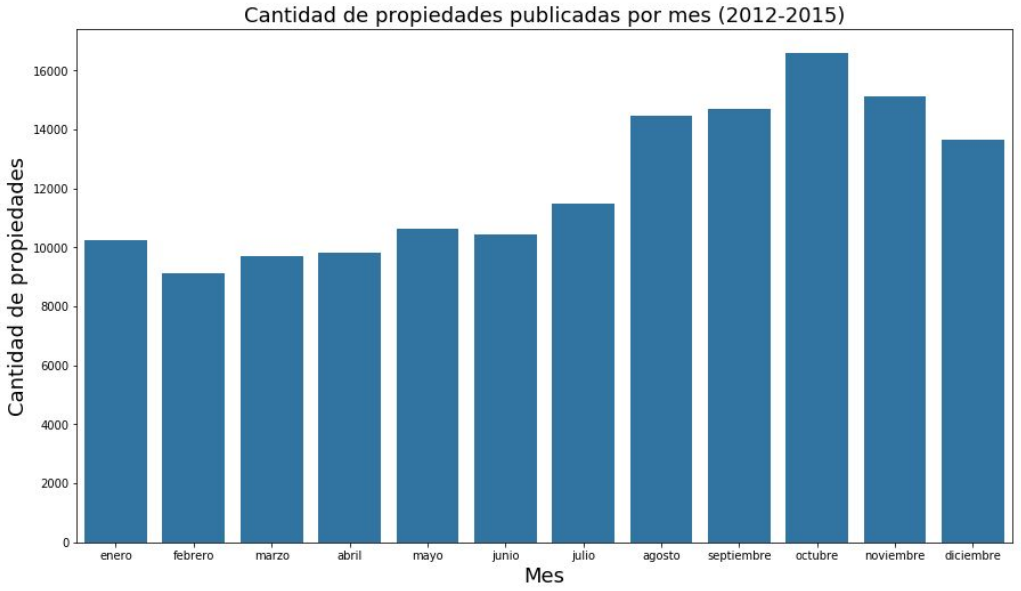
Precio y mes

En el siguiente gráfico se encuentra la cantidad de registros de propiedades por mes, sumando todos los años.



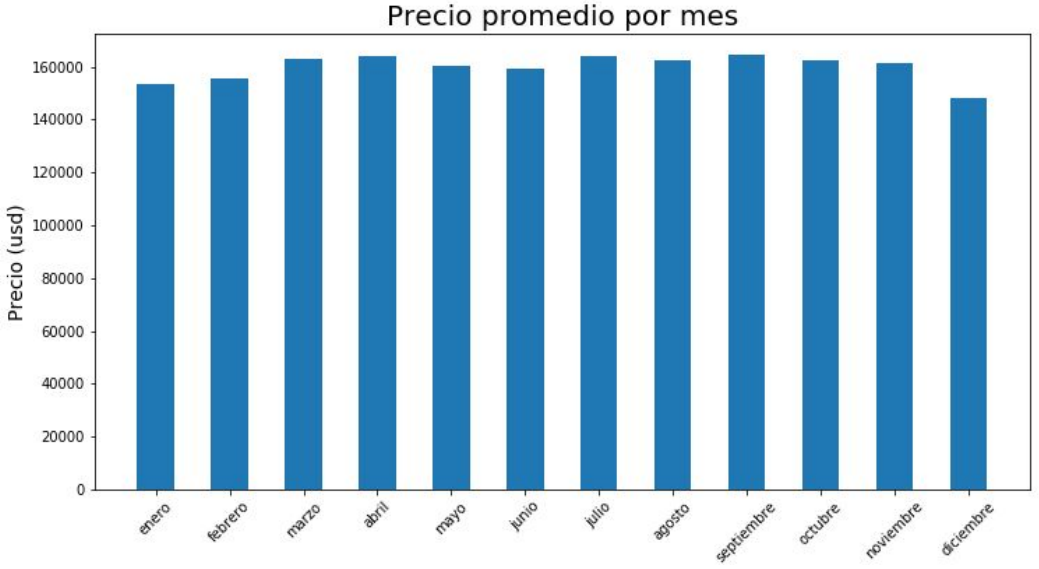
Se observa que diciembre tuvo muchas más publicaciones que los demás meses. Esto es debido a la gran aumento de registros de diciembre de 2016 mencionado anteriormente. Si

tenemos en cuenta sólo los años anteriores, donde no se produjo un aumento tan grande en sólo un mes, se llega al siguiente gráfico:



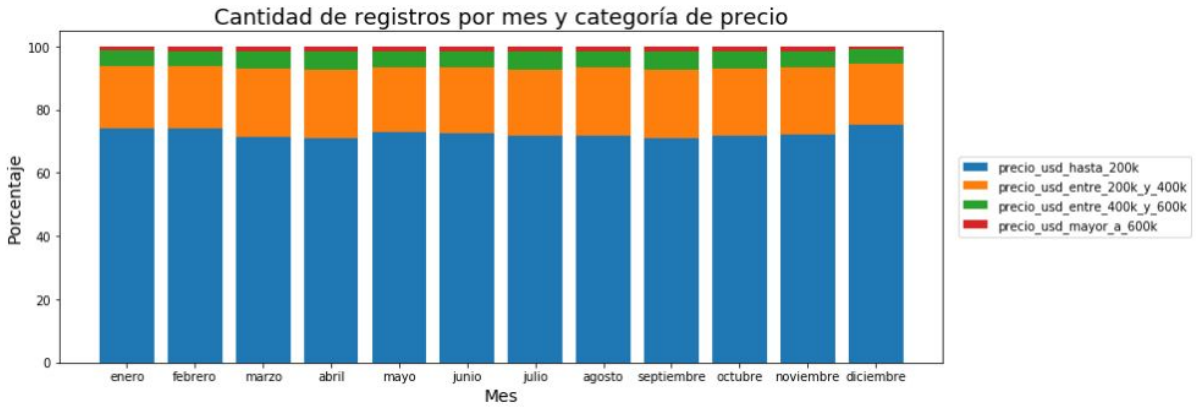
Se aprecia que los últimos cinco meses del año se destacan en cantidad de propiedades, mientras que febrero, marzo y abril son los de menor actividad.

Para determinar si el mes de publicación tiene alguna implicancia en el precio, se calculó el precio promedio por propiedad para cada mes:



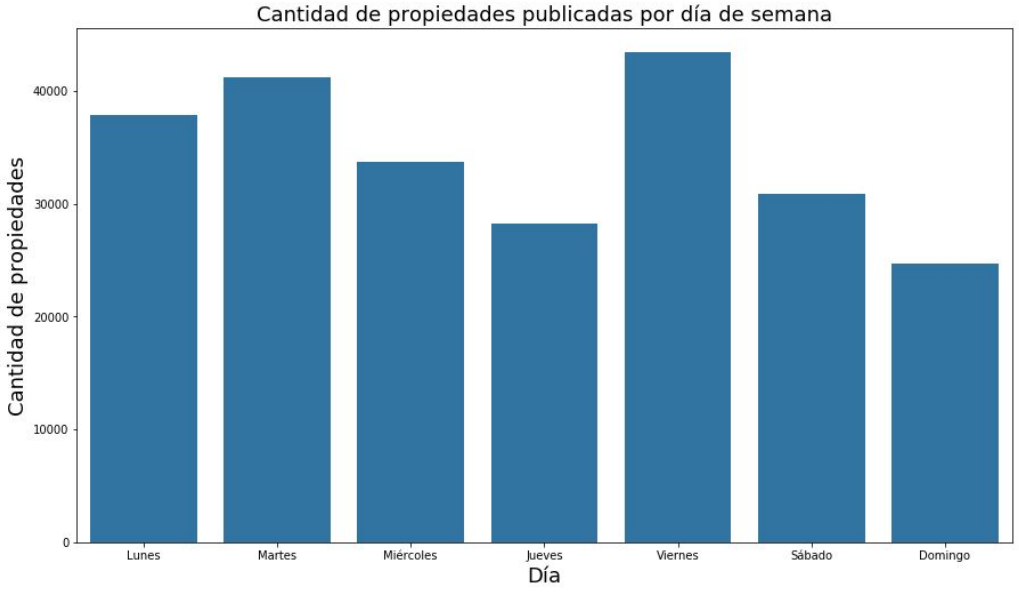
Se observa que la variación de los precios no es muy grande en general, aunque sí se nota un descenso de noviembre a diciembre.

Se obtuvo el porcentaje por mes por categoría de precio. Se nota al amplio dominio de las publicaciones menos costosas. Los meses donde mayor porcentaje representan estas son diciembre, enero y febrero, tal como se había visto en el gráfico anterior.



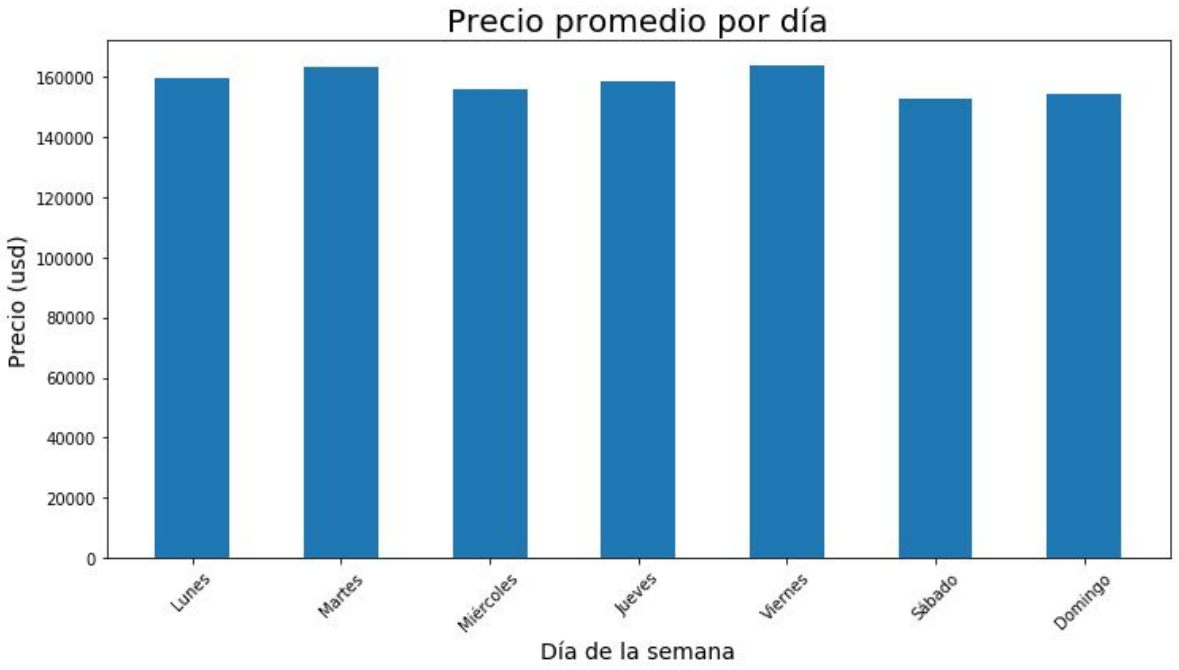
Precio y día

A continuación se encuentra el gráfico de la cantidad de registros por día:



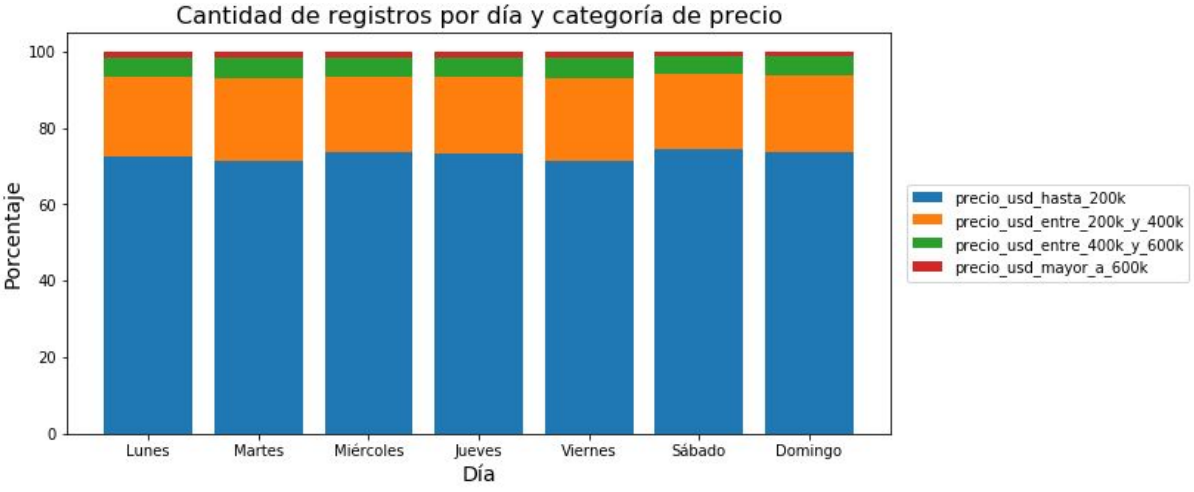
Se ve que la distribución es bastante despareja: los días viernes y martes hay una gran cantidad de registros, seguido de cerca por el lunes. El resto de los días la cantidad es mucho menor, siendo el domingo el punto mínimo.

En el siguiente gráfico se encuentra el precio promedio por día de las propiedades:

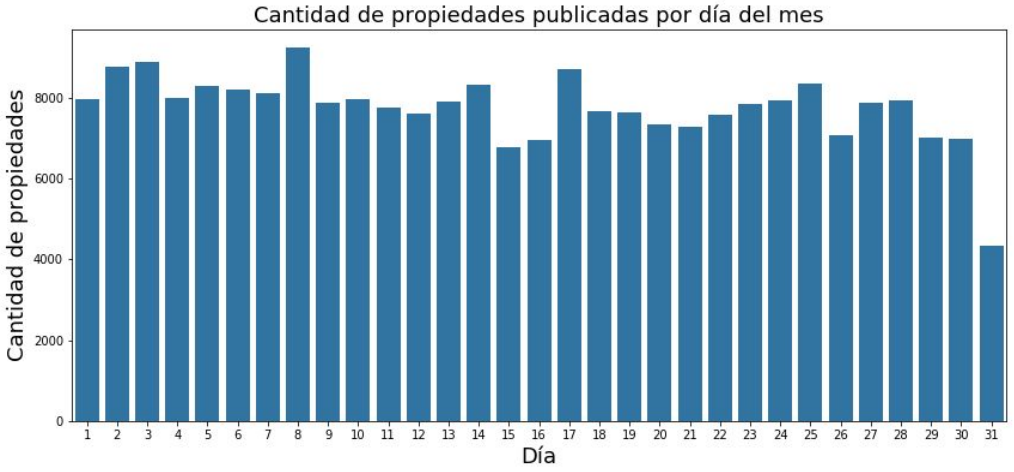


No hay diferencias muy grandes a lo largo de la semana. Los días con precios mayores son martes y viernes (los mismos días con mayor cantidad de registros), mientras que las propiedades menos costosas son publicadas los fines de semana.

Acompañando al gráfico anterior, se obtuvo el porcentaje por día por las categorías de precios antes establecidas. En todos los días hay amplia mayoría de propiedades de las dos categorías inferiores.

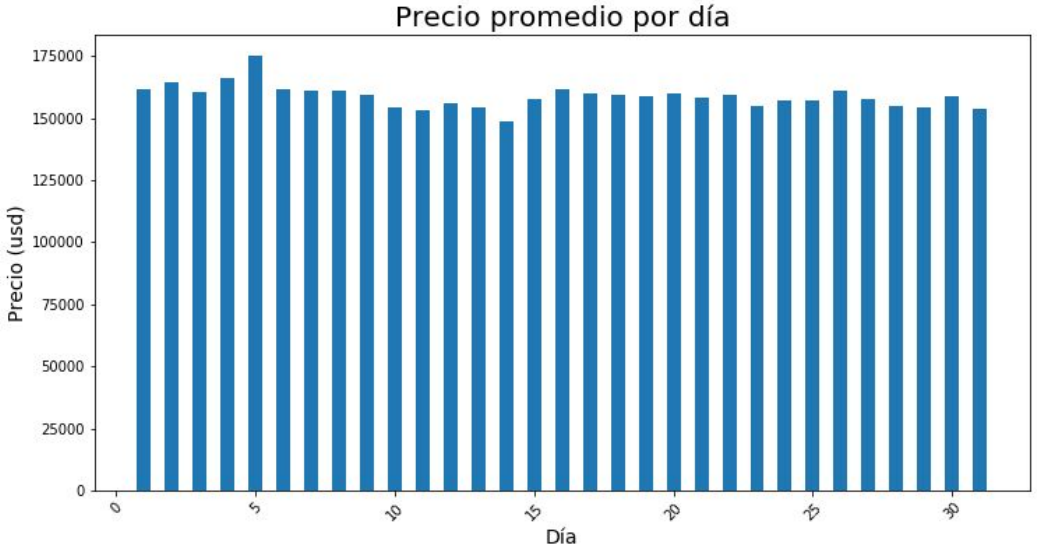


En el siguiente gráfico se encuentra la cantidad de registros total para cada día del mes:

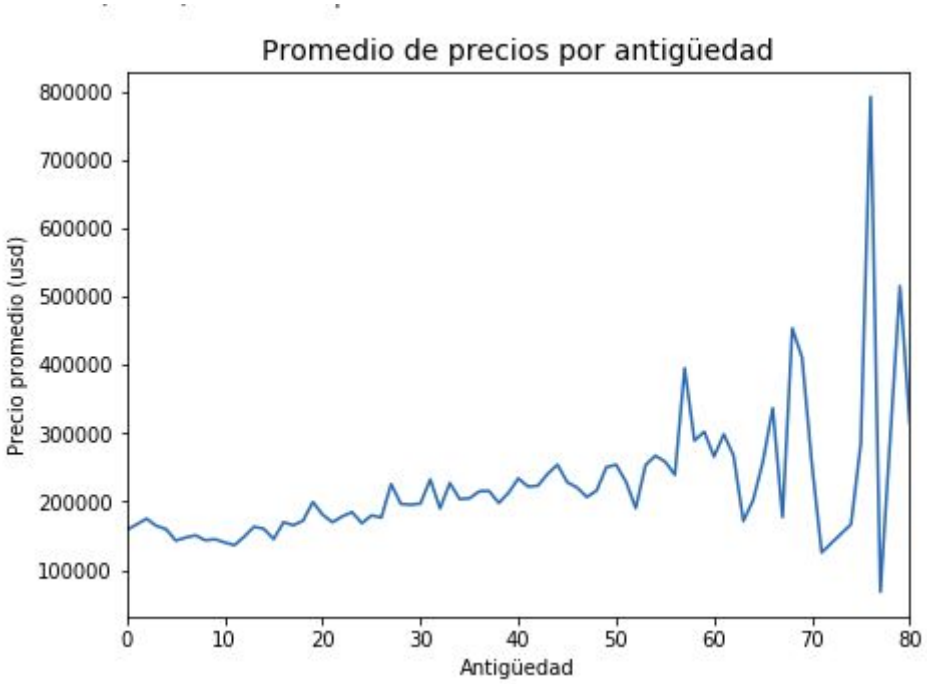


Se observa que en la primera quincena la tendencia es a disminuir en cantidad a lo largo de los días. En cambio, en la segunda quincena no ocurre esto, la distribución es más irregular. La baja cantidad de publicaciones el día 31 se explica porque sólo siete meses llegan a tener tantos días.

Además de la cantidad de propiedades por día, se calculó su precio promedio. No hay una relación directa entre cantidad de publicaciones y precio.



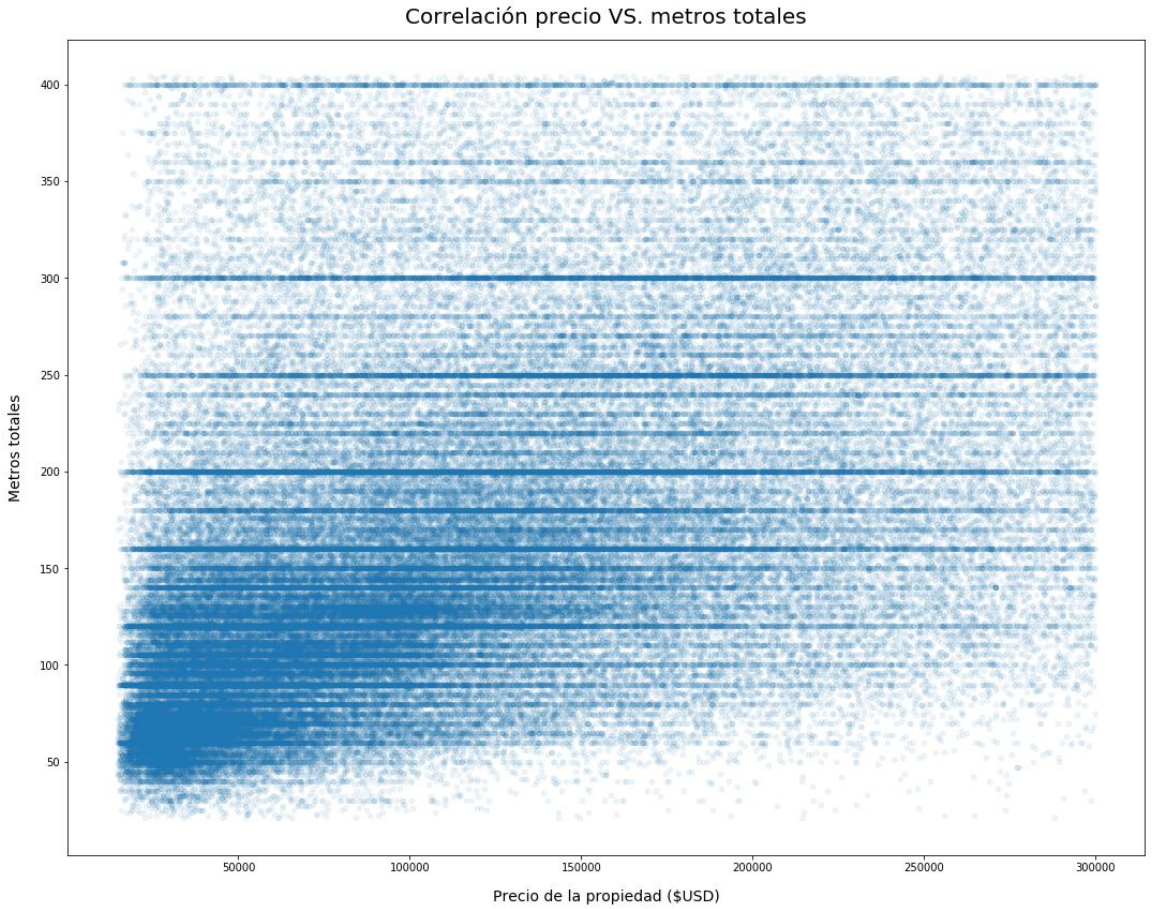
Precio y antigüedad



Sorprendentemente el patrón esperado vale solamente hasta los 10 años. A partir de allí en cierta forma se observa un patrón levemente inverso: a medida que aumenta la antigüedad, los precios suelen elevarse sutilmente. En propiedades muy antiguas se observan precios muy variados, probablemente debido a que depende muchísimo de la calidad inicial de la propiedad y el mantenimiento que ha recibido la propiedad a lo largo del tiempo: construcciones de alta calidad bien mantenidas pueden mantener un valor muy elevado tranquilamente, mientras que otras que no fueron tan bien mantenidas se ven notablemente devaluadas.

Precio y tamaño

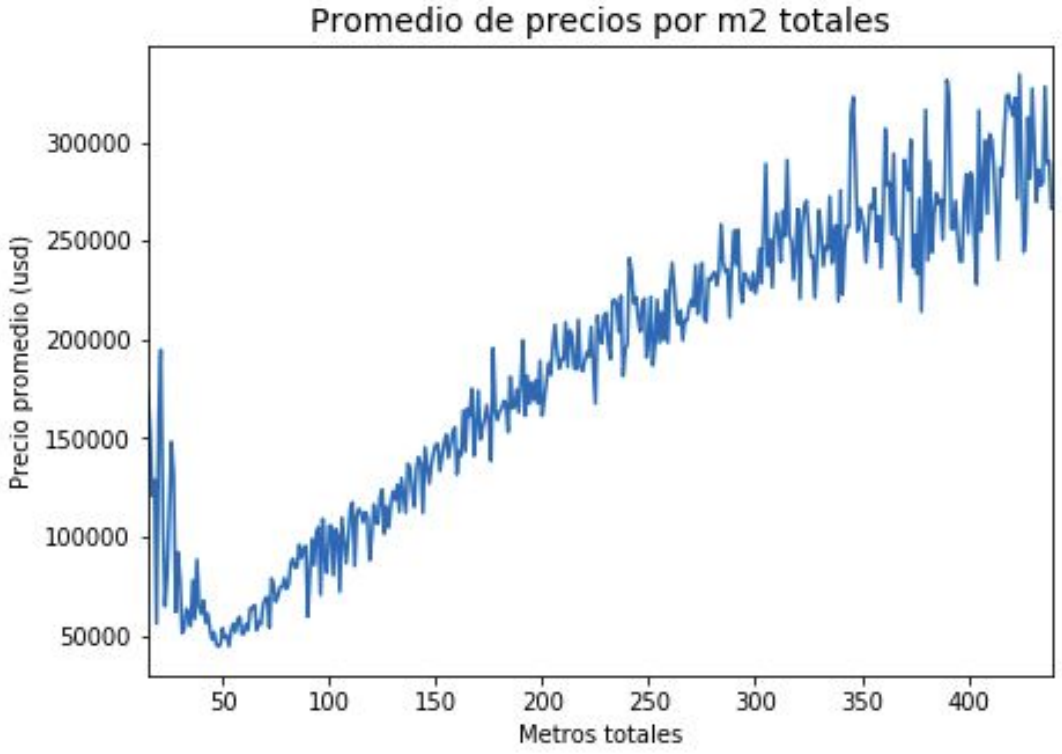
El próximo interrogante surge con respecto al tamaño de las propiedades. En primer lugar se decidió hacer un scatter plot de metros vs. precio, pero el mismo no revela mucha información.



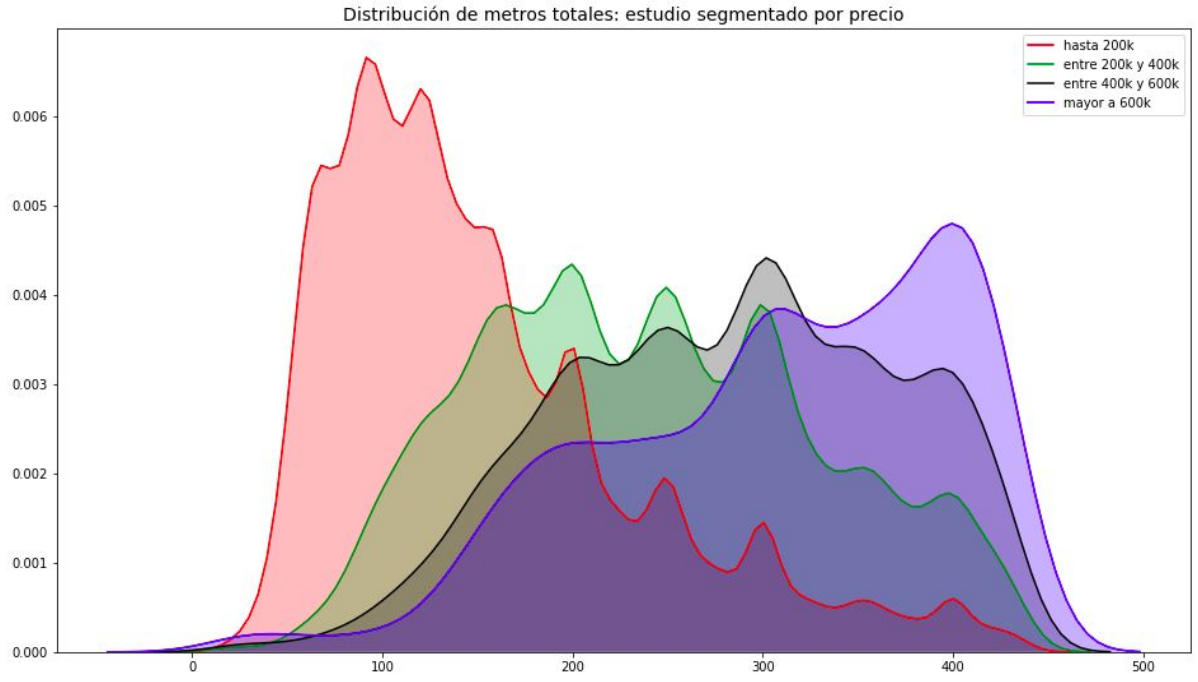
A partir de este gráfico es difícil sacar conclusiones, se ve que hay casos donde los metros se redondean, que es donde se ven las líneas más sólidas, pero es no es fácil llegar a una conclusión porque no se observa ningún patrón más allá de los extremos obvios, por ejemplo, no hay cosas de menos de 30 m2 que salgan arriba de las 100k, aunque por otro lado sí haya cosas baratas de muchos m2.

Es importante notar que tanto para este gráfico como para gráficos siguientes a lo largo del informe se filtró parte de la información para eliminar outliers y hacer los gráficos más entendibles y notables.

Se presenta a continuación la alternativa al mismo, donde se ve mejor la evolución del precio de acuerdo a la cantidad de metros cuadrados de la propiedad. Si bien se realizaron los plots tanto para metros totales y para metros cubiertos, el comportamiento observado del precio es el mismo en ambos casos, por lo que se presenta solamente uno de ambos gráficos.



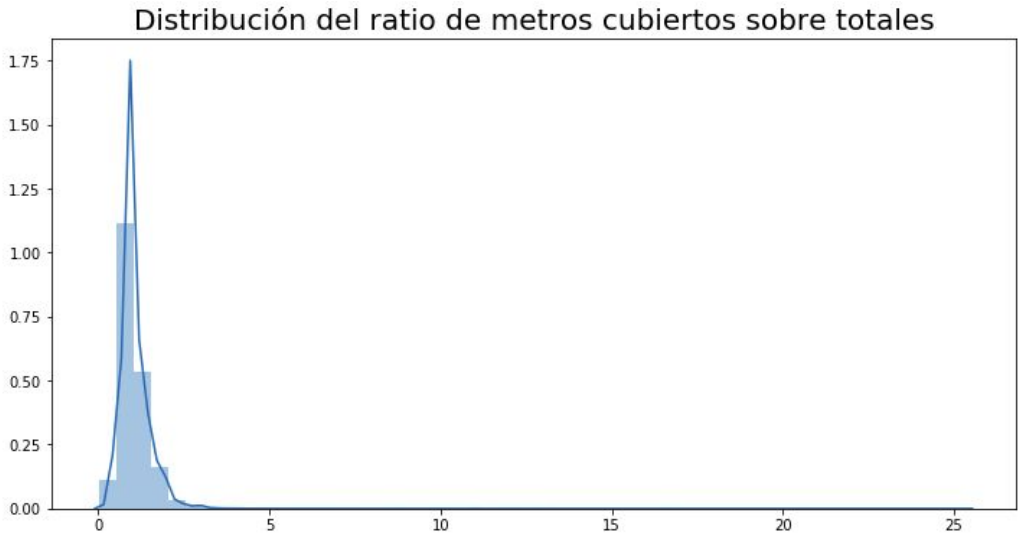
Si se realiza el estudio segmentado, se puede entender mucho mejor cómo es la relación entre precio y cantidad de metros. Se muestra el gráfico para los metros totales, ya que el de metros cubiertos tiene la misma forma, solamente que con picos menos pronunciados.



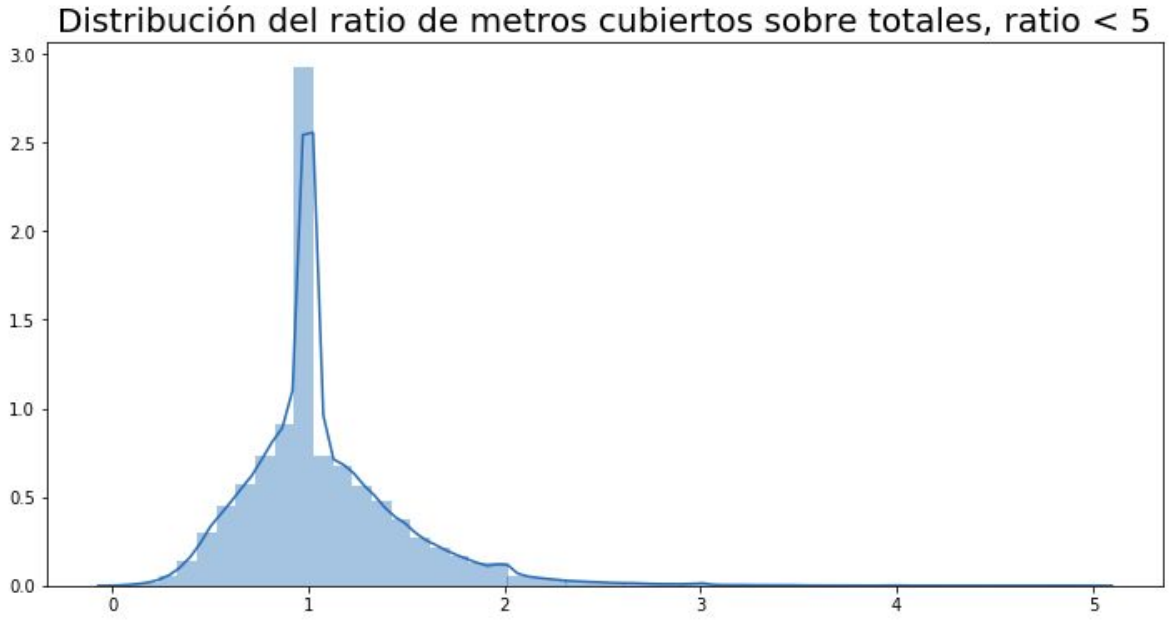
Nótese que la mayor diferencia está entre los grupo de hasta 200k y el que contiene propiedades entre 400k y 600k, quedando el extremo de propiedades más caras más en el centro de la distribución que el grupo precedente. Esto da la pauta que para las propiedades más baratas se puede pensar en que tienen terrenos más chicos, pero a medida que el

precio empieza a crecer, la relación se torna más compleja y empiezan a entrar en juego muchas más cosas.

Veamos ahora la proporción entre los metros totales y los cubiertos, para entender qué tanto difieren ambos números.

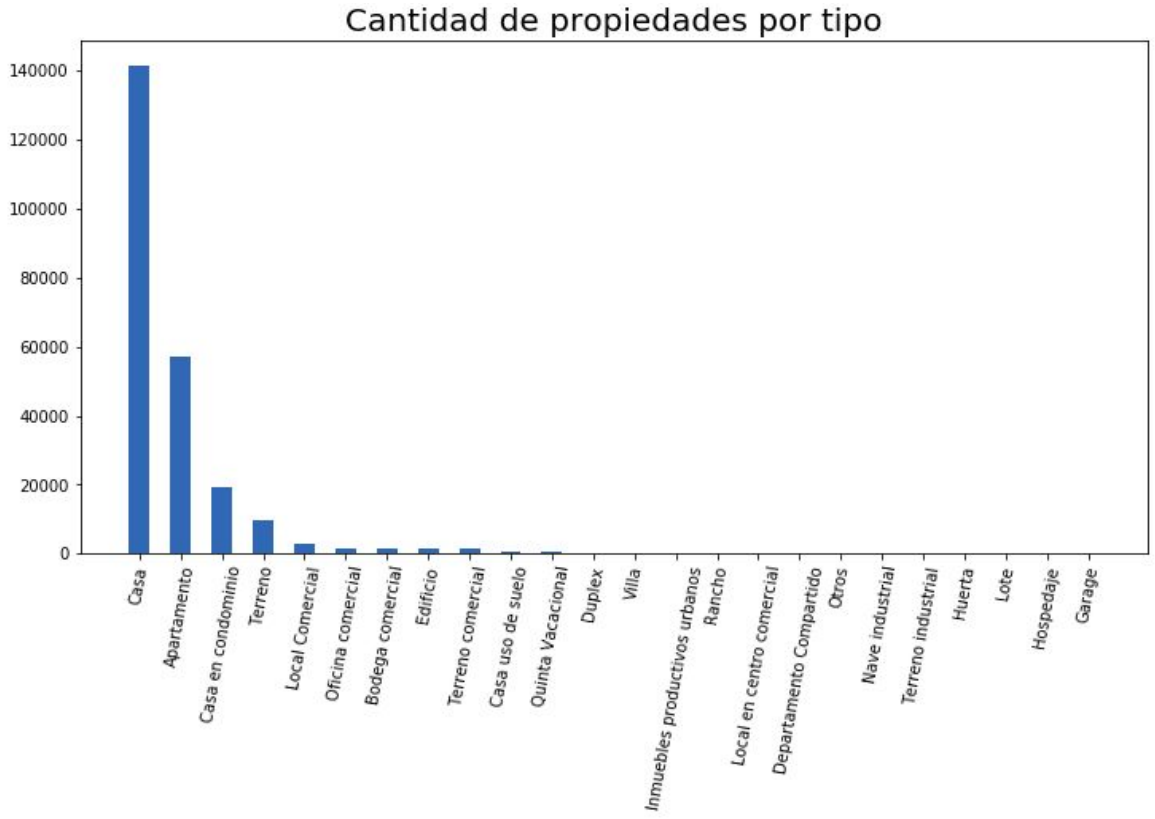


Si bien no se esperaba ver valores tan grandes, esto tiene sentido en el caso de los edificios, donde sobre un lote se erigen muchos pisos de construcción, por lo que el ratio de metros cubiertos sobre los del lote se dispara mucho.

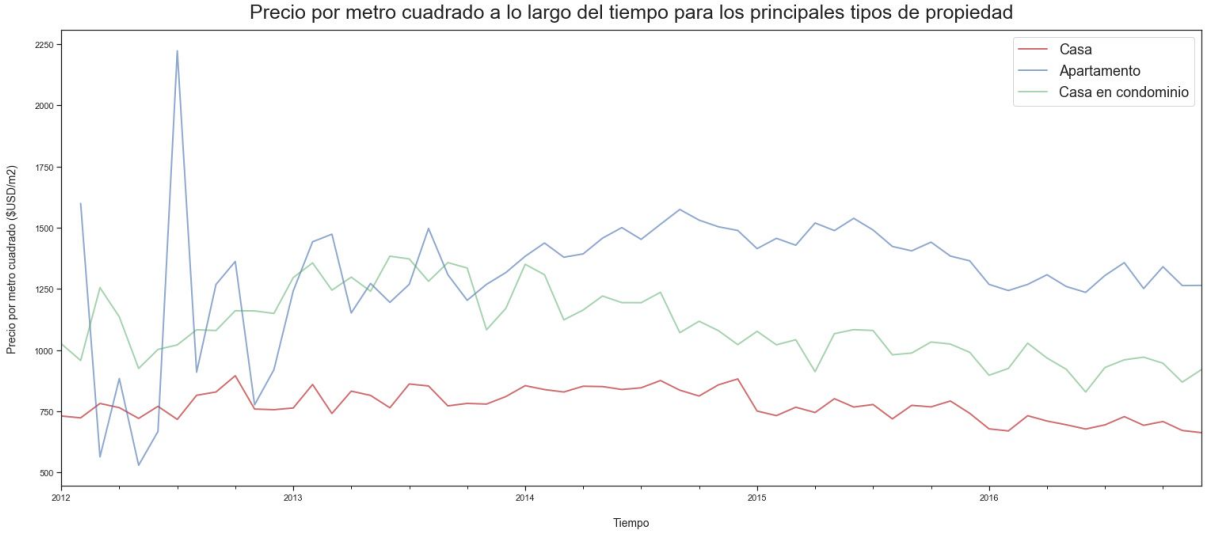


Si filtramos para ver la distribución en los casos donde el ratio es menor a 5, se puede ver que la gran mayoría cumple una relación 1:1, lo cual es esperable: si bien uno no construye sobre todo el lote, una casa de dos pisos que ocupa la mitad del terreno ya cumple con la proporción observada. En la misma línea de razonamiento, es lógico ver que la mayoría de los casos se encuentran entre 0.5 y 2. Esto es además coherente con la cantidad de

registros que hay de casas, mucho mayor a la de cualquier otro tipo de propiedad, como se puede ver en el gráfico abajo.



Teniendo en cuenta la gran cantidad de propiedades atribuidas a Casa, Apartamento y Casa en Condominio como las tres principales, se realiza un estudio en particular de las mismas en el tiempo.



Desde aquí podemos ver que a través de los años las casas han sido las más baratas por m2, lo que lleva a pensar que ya que son más chicas y de menor precio hay mayor cantidad de las mismas.

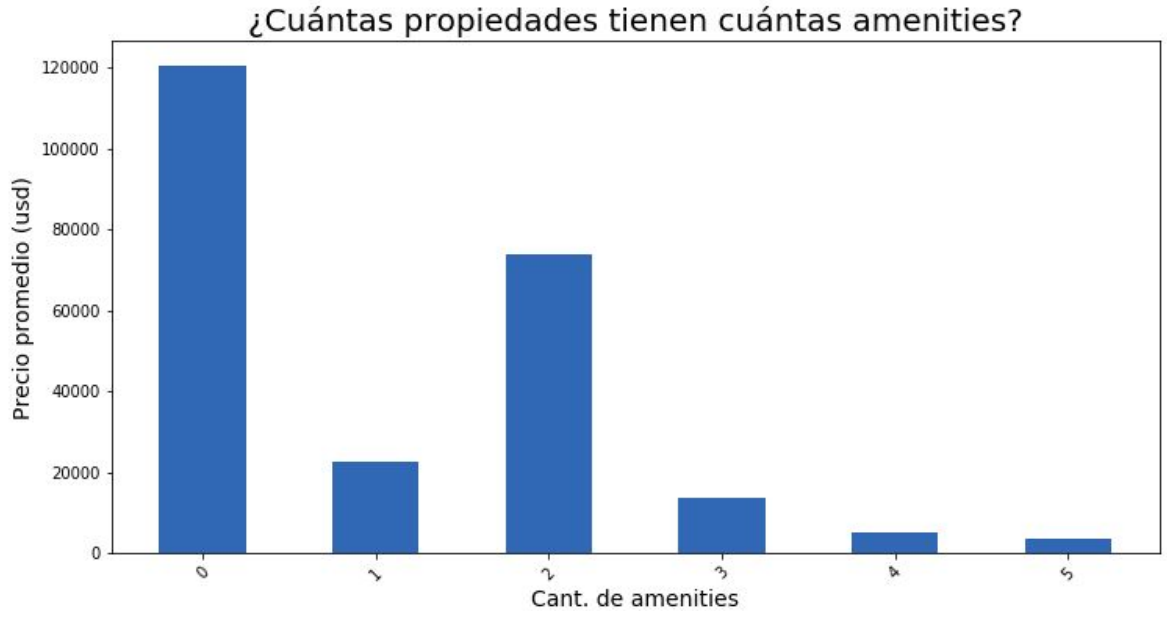
Luego podemos ver que las casas en condominios, en el fondo, son una variación de las casas, ya que son un poco más caras y con más metros.

Por último, los apartamentos. A mediados de 2012 su precio/m2 varió drásticamente, lo que lleva a pensar que la causa de esto fue algún suceso en la economía del país. Más allá de este caso particular, los apartamentos luego se establecieron como aquellos más caros por m2, quizá esto se deba por la ubicación de los mismos y la comodidad que pueda brindar esto mismo. No es lo mismo tener una casa barata con muchos metros en un lugar aislado que un apartamento en una zona céntrica aunque sea más pequeño.

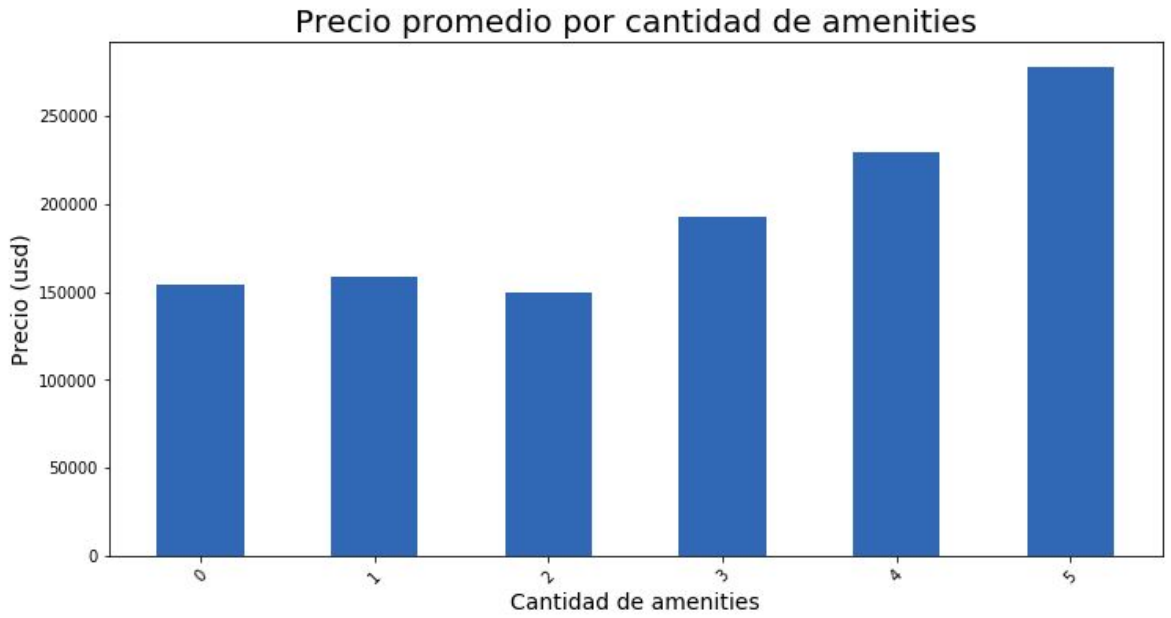
Precio y amenities

A continuación se decide estudiar cómo afectan los distintos atributos de las propiedades a los precios de las mismas.

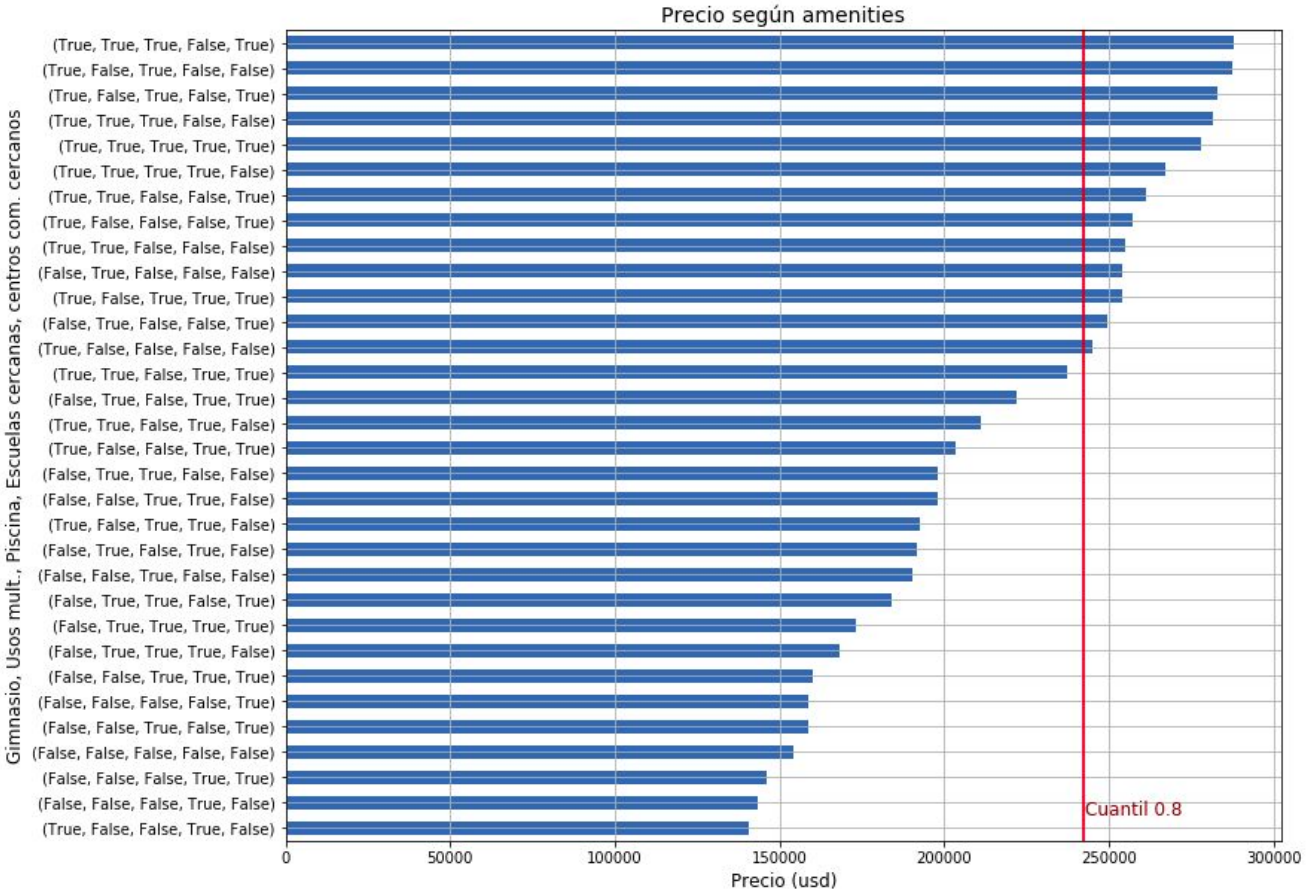
En primer lugar, se ve que la gran mayoría de propiedades o no tienen amenities o tienen dos. Como era de esperar, hay muy pocas propiedades con 4 y 5 amenities.



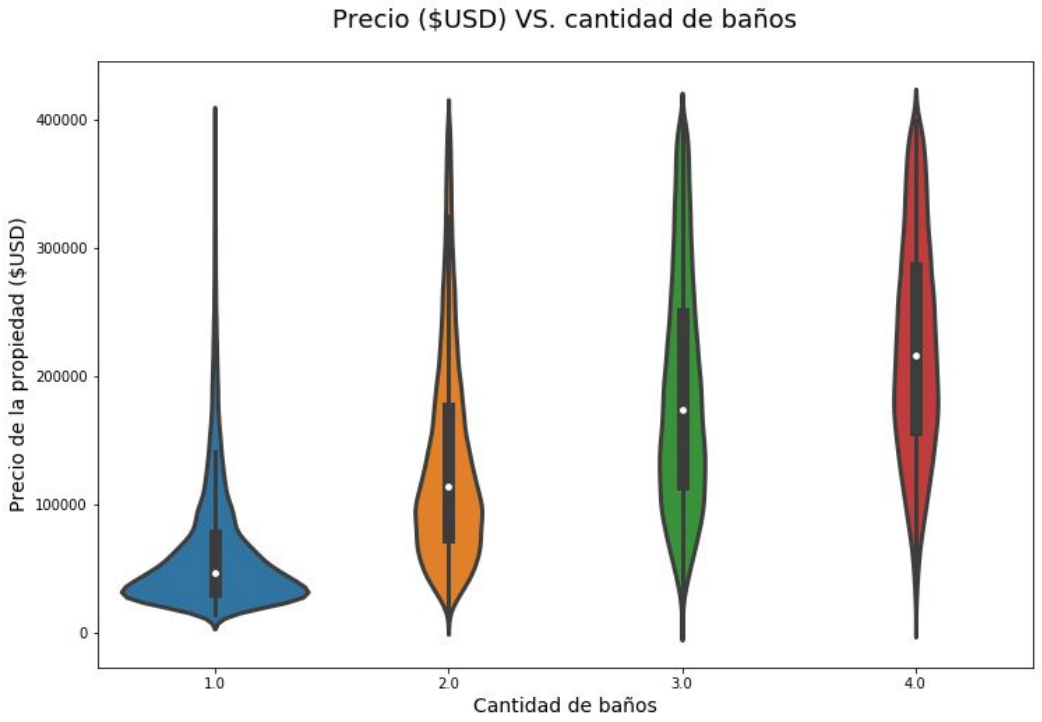
Como era de esperar, el precio de ese sector más exclusivo que cuenta con todas las amenities es marcadamente más elevado. Sin embargo, se esperaba una diferencia más marcada entre el resto de las cantidades.



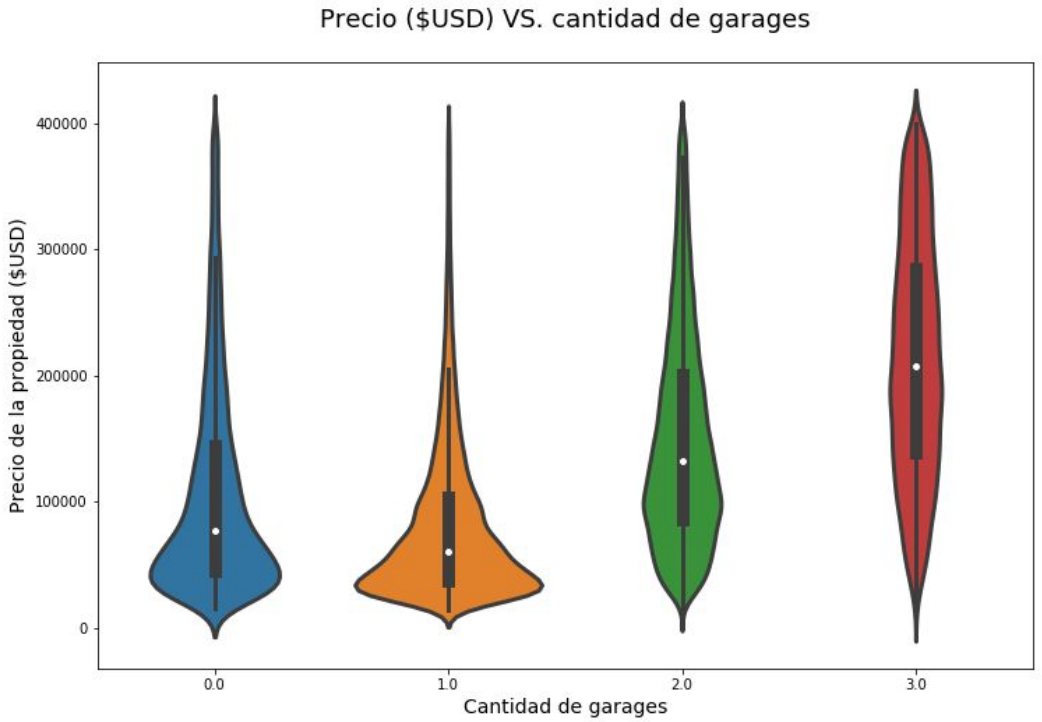
En el gráfico a continuación se detalla el precio promedio por cada combinación de posesión o no de cada amenity. Se marca además el cuantil 0.8 para ver qué tipo de propiedades lo excede. Es interesante observar que en casi todos los casos que lo exceden se trata de propiedades con gimnasio, y las 6 combinaciones más caras incluyen además de gimnasio invariablemente también piscina. Además las últimas cuatro combinaciones no tienen escuelas cercanas. Se podría suponer que estas propiedades están destinadas a personas sin hijos o con hijos que no están en edad escolar, y por lo tanto también ubicadas en zonas donde el caos agregado por una escuela cercana se ve aliviado.



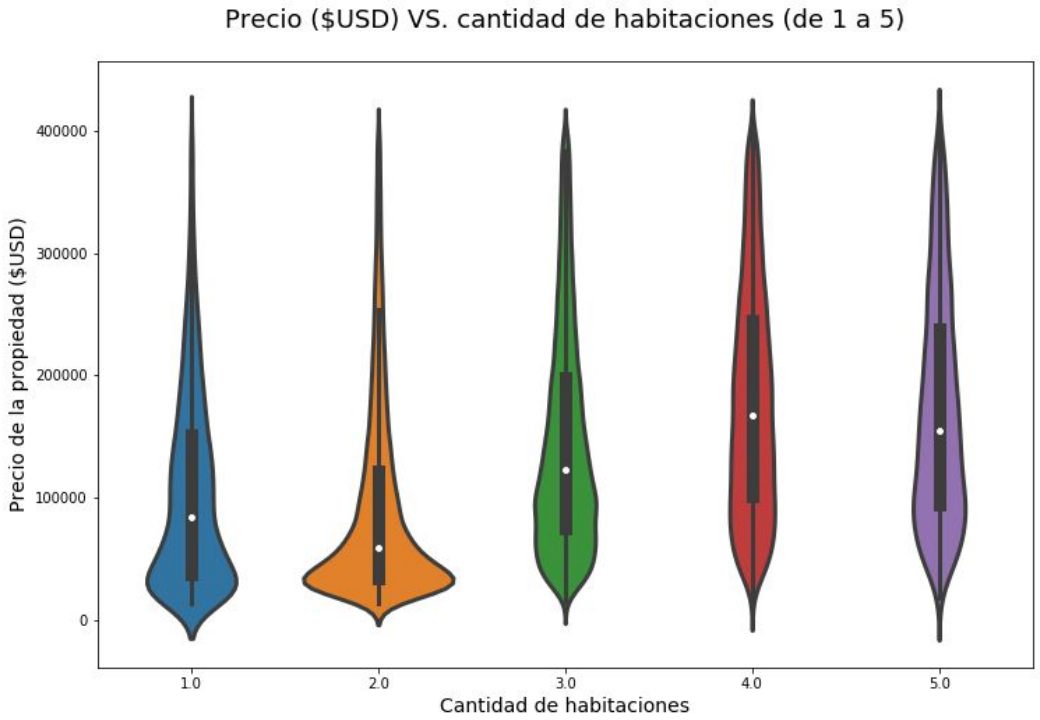
Para tener más perspectiva sobre las características de las amenities y su precio se ven los siguientes gráficos:



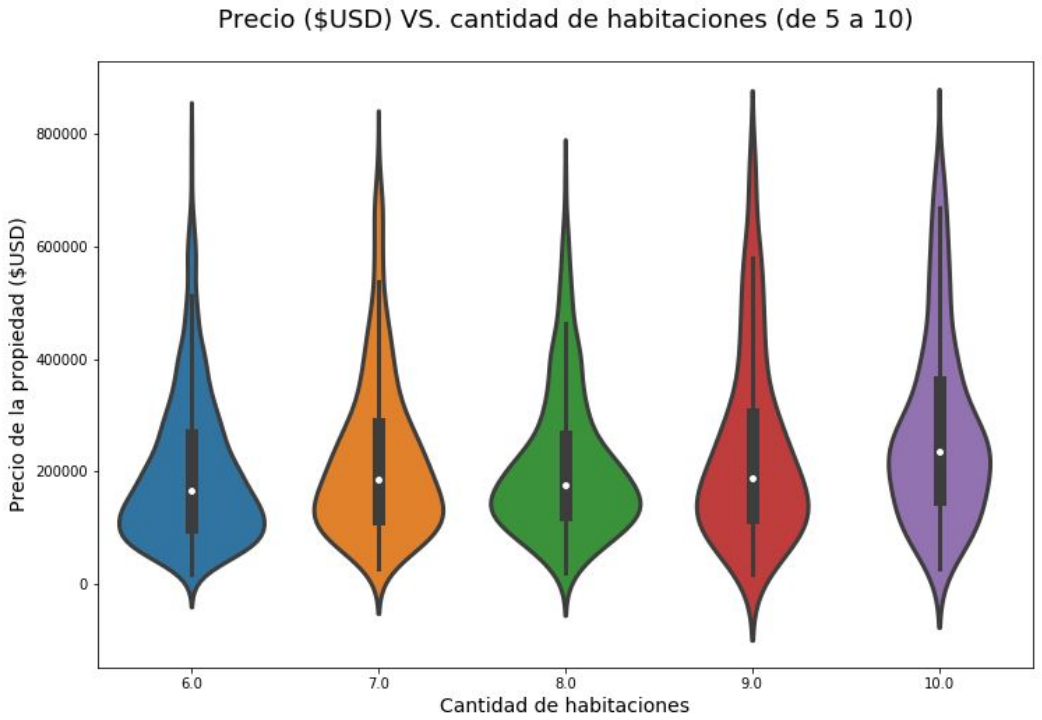
Se puede apreciar que a medida que aumenta la cantidad de baños, el precio también lo hace. Además se puede ver una diferencia mucho más marcada entre un baño y el resto de las cantidades.



Se puede apreciar que a medida que aumenta la cantidad de garages, el precio también lo hace, aunque de alguna manera es conveniente comprar con un garage ya que sale lo mismo o menos que aquellos que no tienen garage. Es lógico pensar que esto dependa del tipo de propiedad, tamaño y su ubicación.

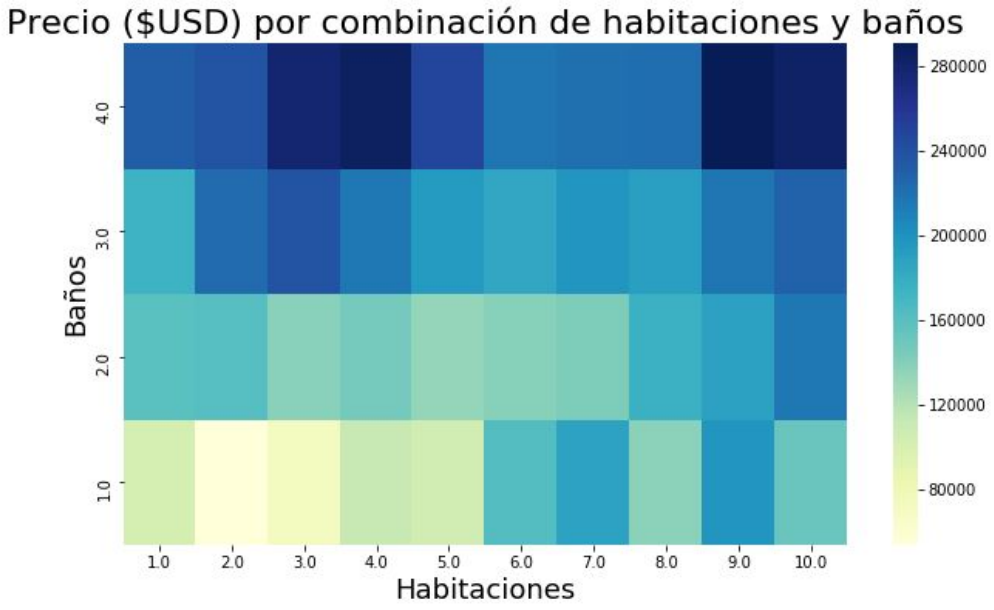


Se puede apreciar que a medida que aumentan la cantidad de habitaciones, el precio también lo hace, aunque si miramos los valores por ejemplo entre 1 y 2 habitaciones, como también entre 4 y 5, en general no hay gran diferencia por lo cual en algunos casos podría ser conveniente comprar por el mismo precio alguna propiedad con alguna habitación más ya que podría ser igual o más barato.



Cuando ya las habitaciones son muchas se puede ver que el aumento del precio no es tan significativo como en los casos previos.

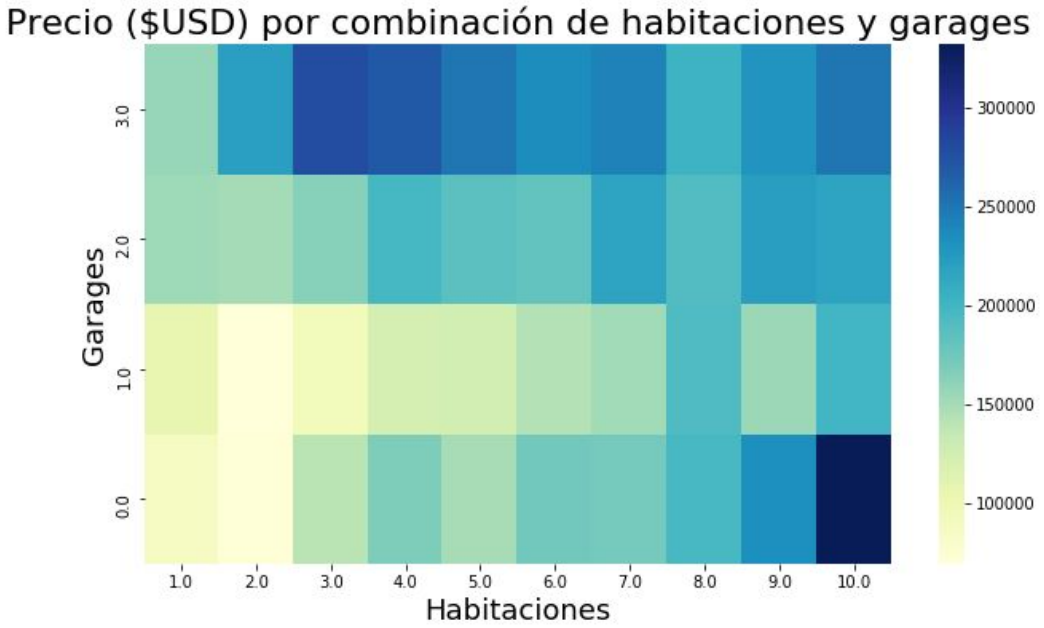
Luego de analizar el efecto de estas variables sobre el precio, se trata de determinar cuál es más determinante sobre éste. Para eso, se realizaron heatmaps comparando de a dos variables. En primer lugar se muestra el heatmap por combinación de habitaciones y baños:



Se observa que el precio aumenta con las habitaciones y los baños. El color es similar dentro de cada fila, pero varía mucho dentro de cada columna, oscureciéndose a medida

que aumenta la cantidad de baños. De aquí se concluye que influye más en el precio la cantidad de baños que de habitaciones.

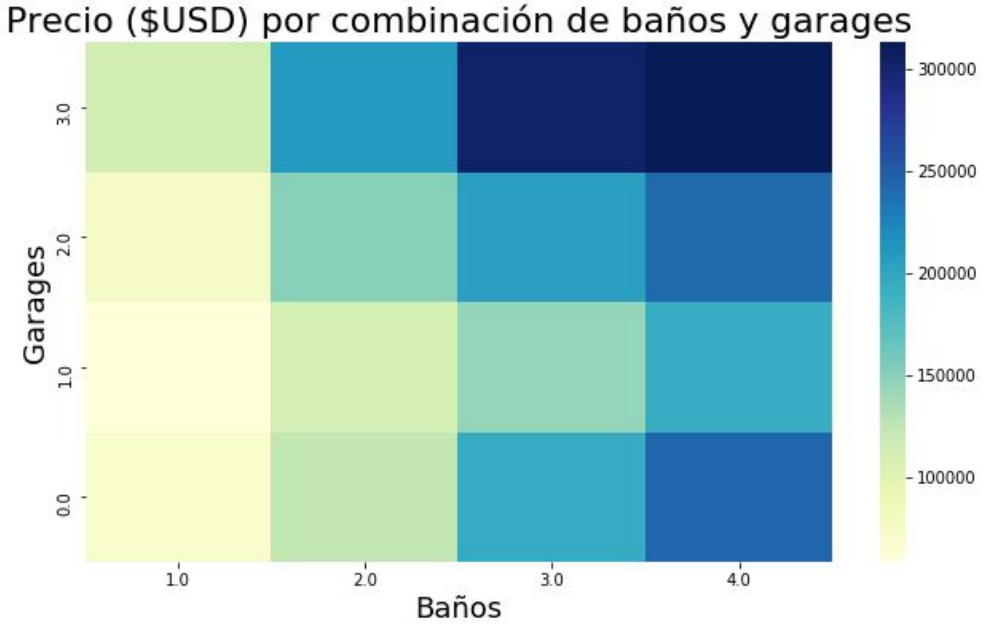
A continuación se examinan el heatmap según habitaciones y garages:



En este caso se ve que la importancia de estos factores es más equilibrado que en el anterior. Hay precios máximos para la máxima cantidad de habitaciones sin haber garages y para pocas habitaciones con varios garages.

Para examinar la relación de baños y garages en cuanto al precio, se realizó otro heatmap comparando el precio según la variación de estos valores:

En



Observando los gráficos se ve que los valores máximos se alcanzan con la mayor cantidad de baños y garages. Se ven también valores altos en los precios para propiedades con 3 o 4 baños independientemente de la cantidad de garages, mientras que las propiedades con 2 o 3 garages con 1 o 2 baños tienen precios bajos. Con este análisis se concluye que la cantidad de baños es más determinante en el precio que la de garages y habitaciones.

Precio y ubicación

Avenidas

Un factor que puede influir en el precio de la propiedad es si la misma se encuentra o no sobre una avenida. Sin embargo, la distribución del precio es prácticamente idéntica en ambos casos, y además tan similar a la distribución no fragmentada mostrada al principio que se omite mostrar el gráfico de las mismas.

Una posible explicación a esta distribución tan similar puede ser a que las propiedades que están cerca de una avenida probablemente sean muchas más que las que están sobre la avenida en sí, y en sentido de ubicación tienen ventajas muy similares, por lo que es de esperar que empujen fuertemente la distribución.

Si bien el promedio del precio de las propiedades ubicadas sobre avenidas es de hecho levemente superior, se puede concluir que la hipótesis probablemente esté bien orientada, pero que se necesitaría profundizar muchísimo más el análisis para llegar a algún resultado significativo.

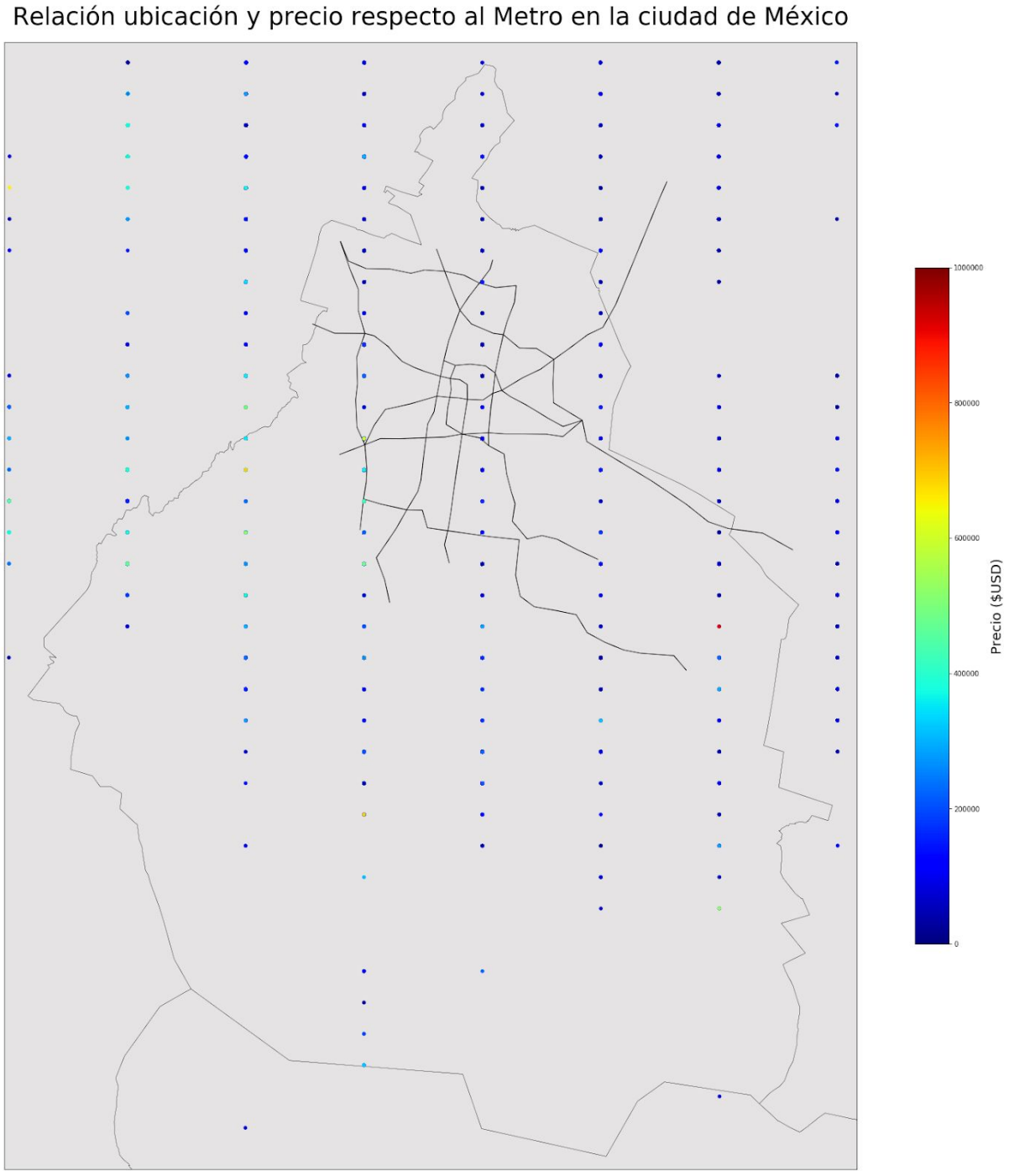
A continuación se muestran dos gráficos de la ubicación de las propiedades en latitud y longitud (con aquellas que se pudieron recuperar inclusive) con sus respectivos precios.

Ambos gráficos tienen el mismo objetivo a mostrar, pero uno fue hecho con la liberia plotly, y el otro con matplotlib y basemap. La diferencia es que con aquel realizado con plotly se puede interactuar, ver los precios para cada punto y hacer zoom en el mapa en el siguiente link:
Observación: no se recomienda intentar interactuar con máquinas de bajos recursos, para dar una idea de esto, para la distancia de la imagen del gráfico dado hay un consumo de 1GB de RAM aproximadamente, y mientras más cercanía, más consumo.
https://plot.ly/~FCozza/112/ubicacion-de-las-propiedades-con-su-respectivo-precio/

La idea de estos gráficos era notar algún tipo de tendencia pero aparentemente, de esta forma no se puede ver mucho más que la zona principal es la del distrito federal y el estado de México.

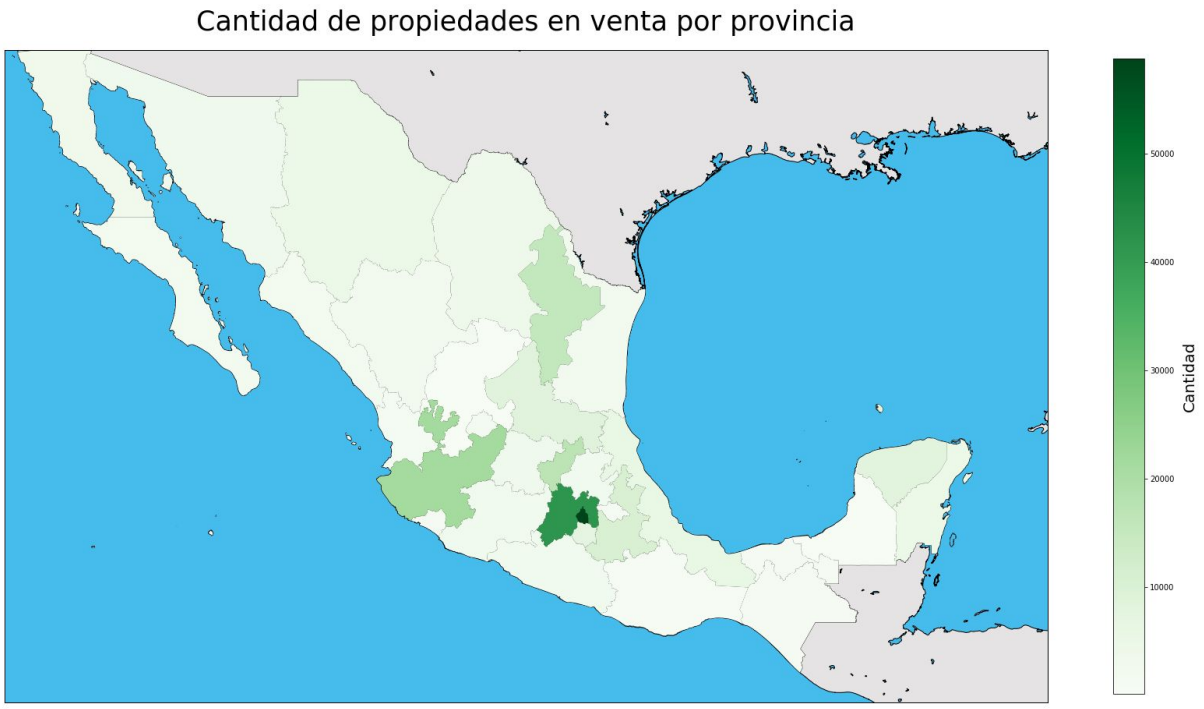
Por lo cual se intentó concentrar y hacer nuevos gráficos en esta zona para intentar relacionarlos con datos externos, pero no hubo éxito ya que las latitudes y longitudes presentadas no son muy amigables.

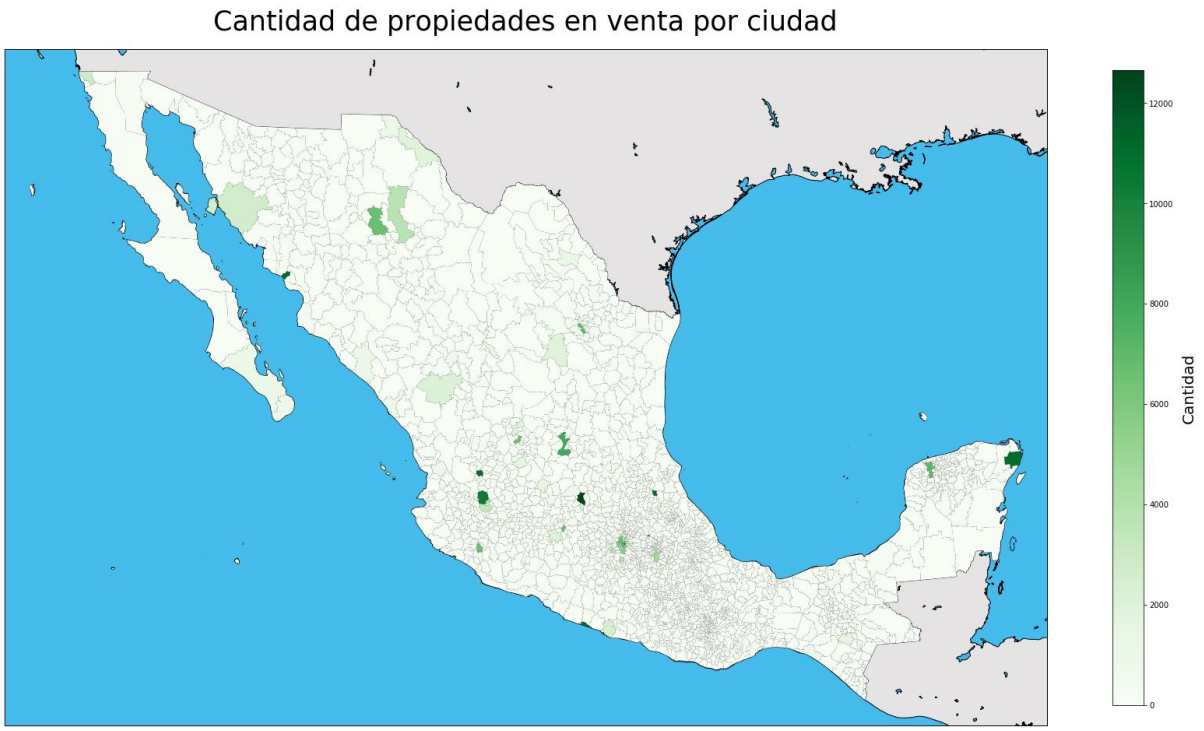
Se muestra a continuación un gráfico en el estado de México con las líneas de Metro



En este gráfico se ve que no hay siquiera dispersión de las latitudes y longitudes y muy pocas variaciones en el rango del precio, por lo cual se decide abandonar este enfoque exhaustivo y relacionado con datasets externos ya que no daria mucha información.

Análisis por provincia

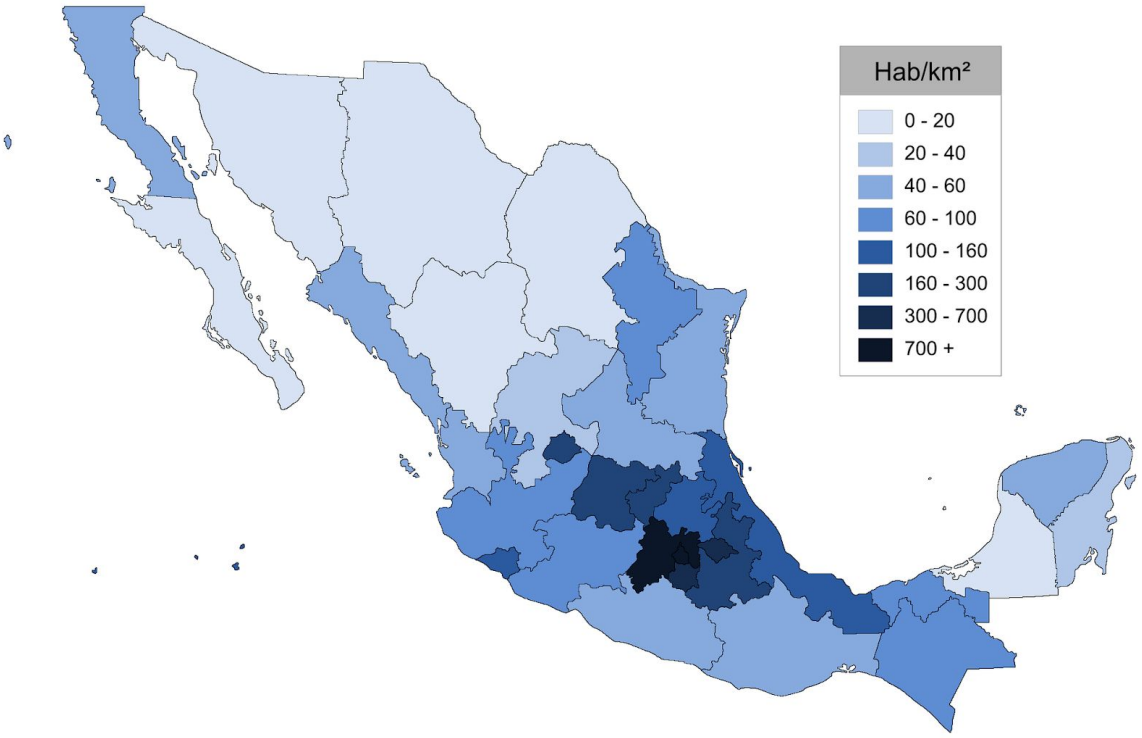




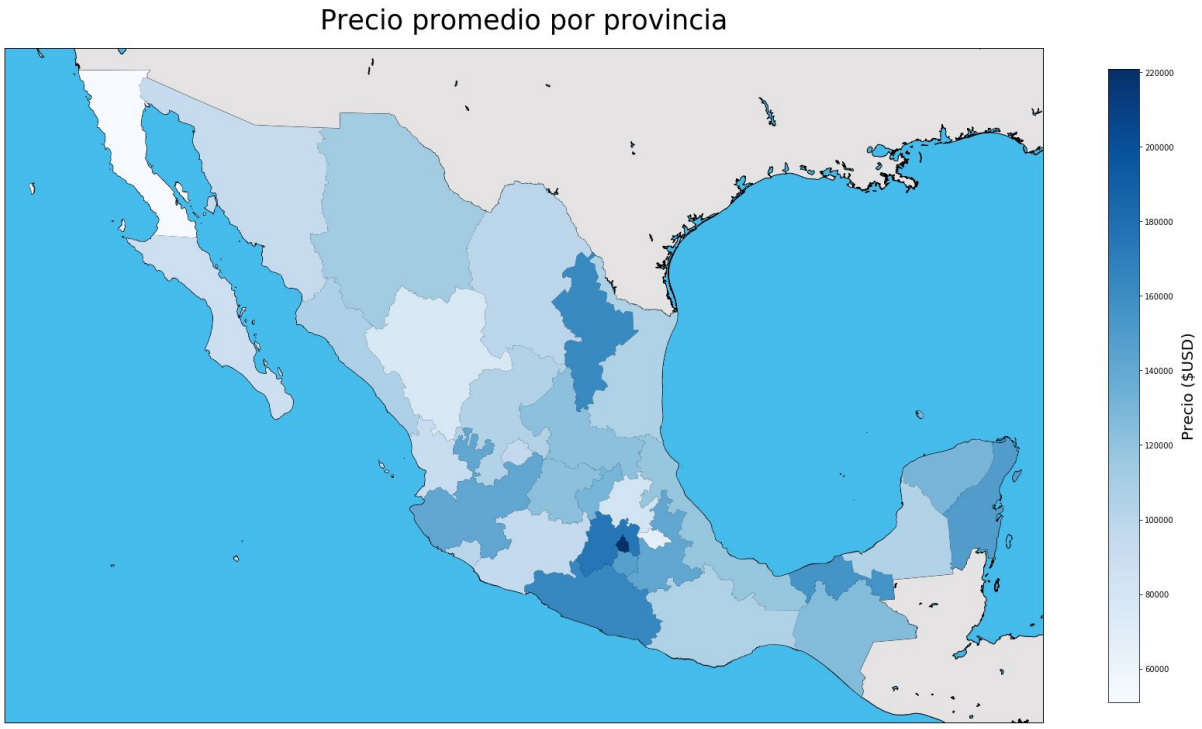
Notar en los gráficos las diferentes escalas con respecto a la cantidad, y si bien en el gráfico de ciudades se pueden hacer algunas menciones honorables, en general este análisis sobre las ciudades no es muy apreciable y para futuros análisis sobre mapas de acuerdo a ciertas variables se optará por usar el mapa sobre las figuras de las provincias.

A partir del choropleth, podemos ver que la mayor cantidad de ventas se concentra en la zona céntrica en el estado de México y su Distrito federal. Esto último coincide muchísimo,

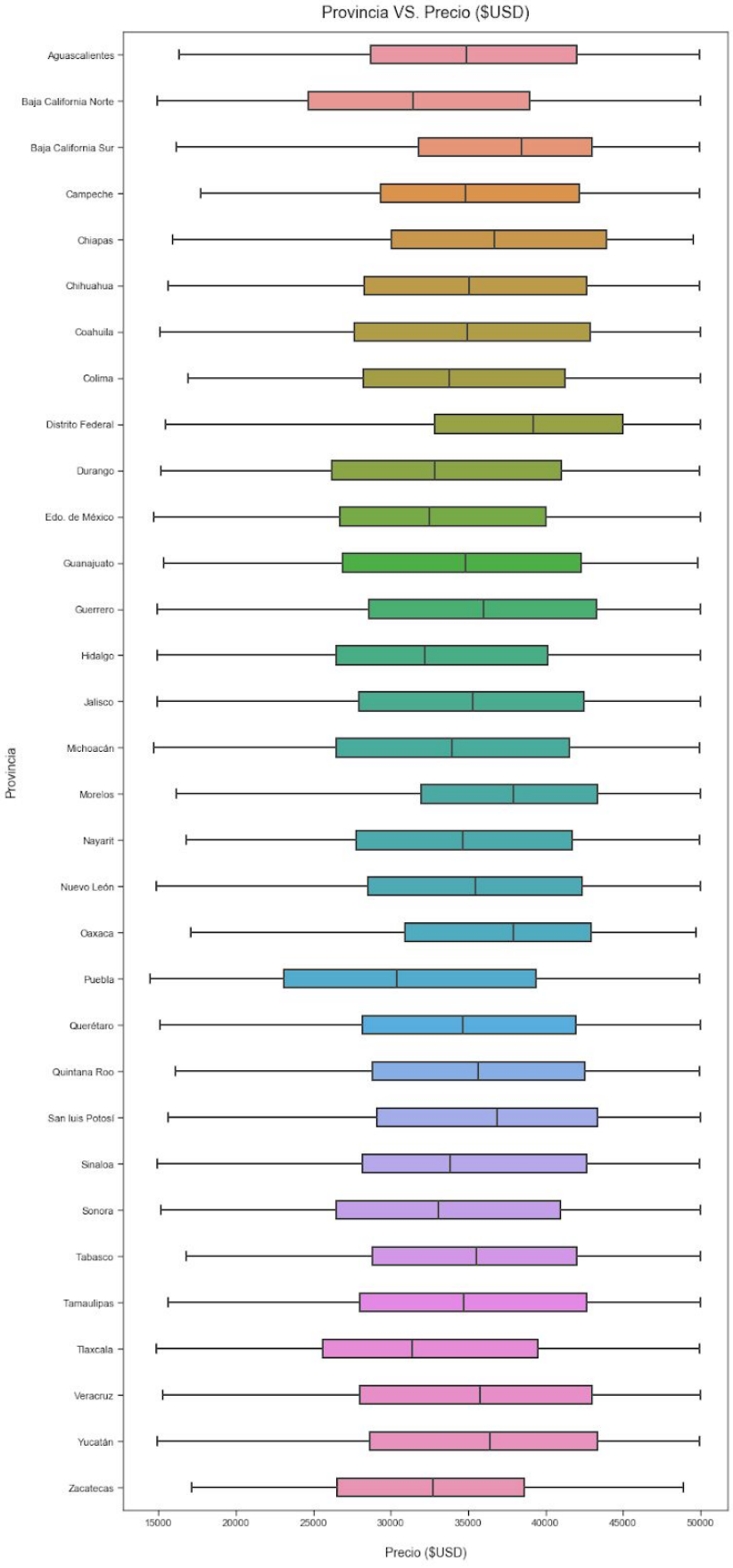
como era de esperar, con la densidad poblacional del país. Se muestra a continuación un gráfico de densidad poblacional para poder apreciar la comparación mencionada.[1]



A continuación se empieza con el análisis de precios por provincia:

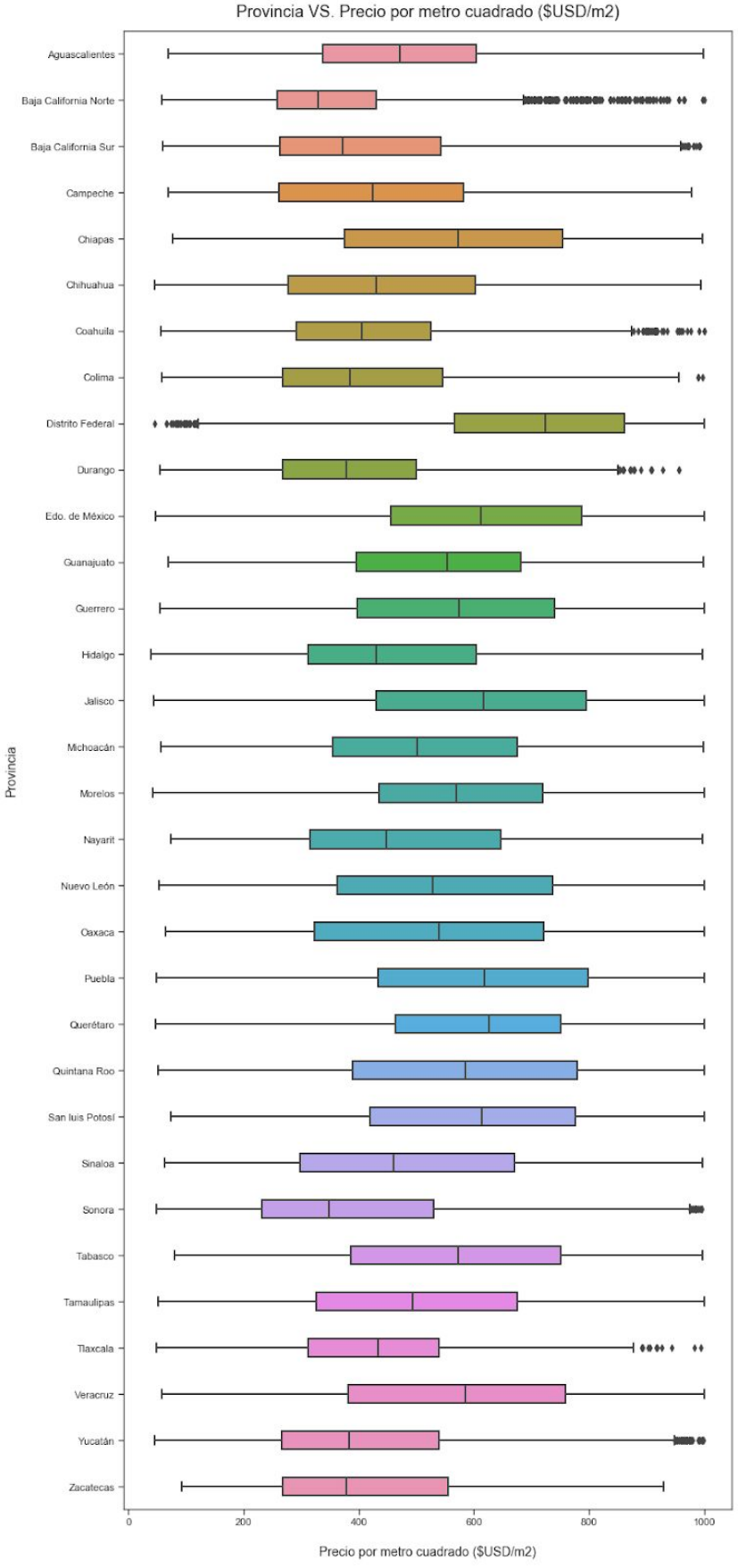


[1] Fuente: https://en.wikipedia.org/wiki/List_of_Mexican_states_by_population_density#/media/File:2012_Density_Map.png



Se destaca que las propiedades más caras (en promedio) se encuentran en el centro de México y su Distrito Federal, y se hace mención honorífica de las provincias de Nuevo León y Guerrero con propiedades relativamente caras.

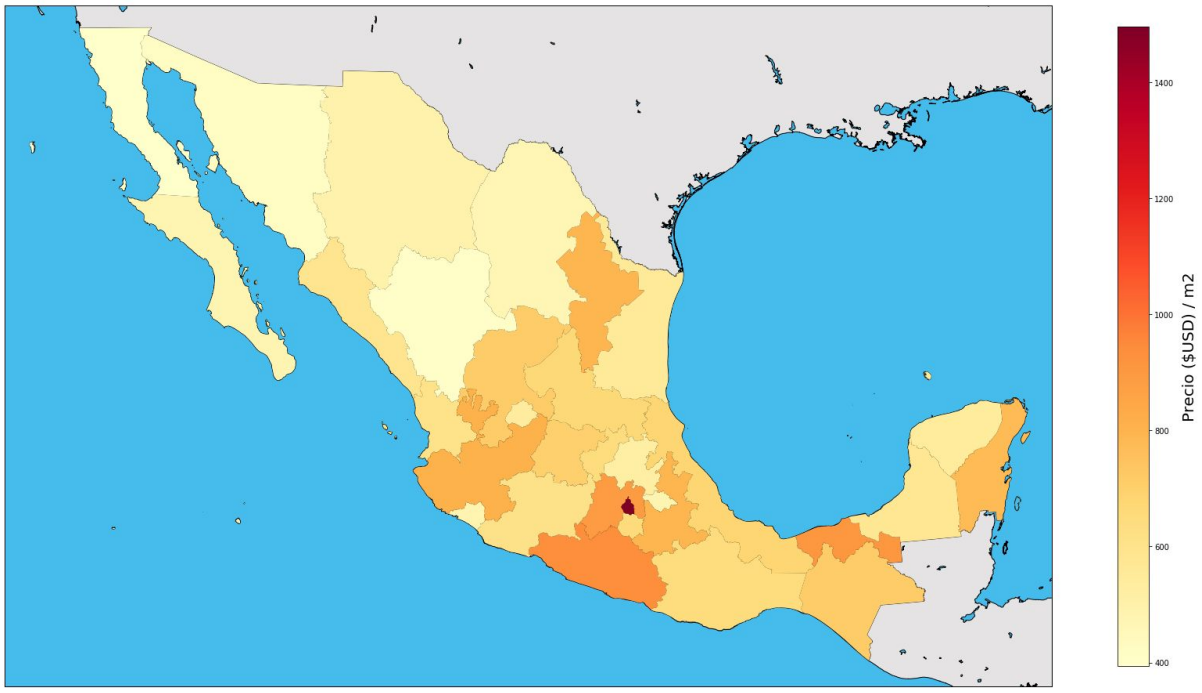
El choropleth a diferencia del boxplot, es clave para detectar estas diferencias.



Teniendo en cuenta el análisis previo del precio promedio, aquí podemos ver que aquellas que se mantienen a un alto precio por metro cuadrado son aquellas del el estado de México y su Distrito Federal y la provincia de Guerrero principalmente, lo cual no es dato menor, ya que en definitiva allí es donde se encuentran las propiedades más caras de todas.

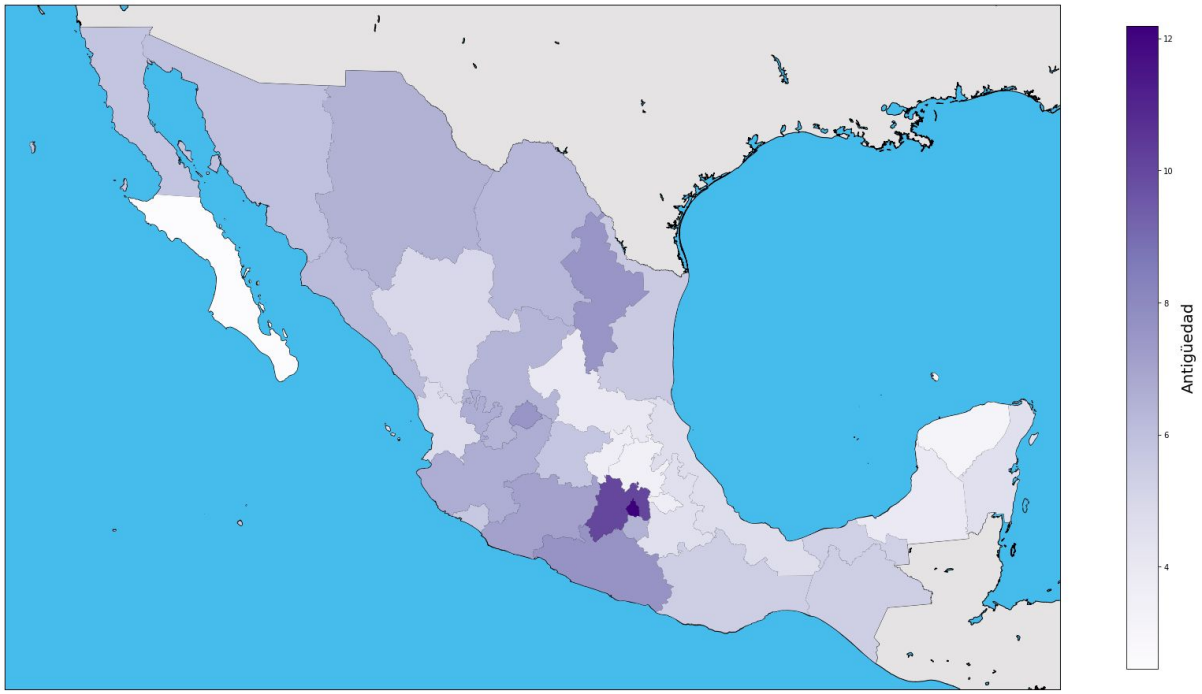
Es lógico pensar que suceda en el centro del país, pero sería interesante ver las caracteristicas provincia de Guerrero para ver que hace subir su valor. (Ver mapa en la siguiente hoja para mayor claridad)

Precio por metro cuadrado promedio por provincia

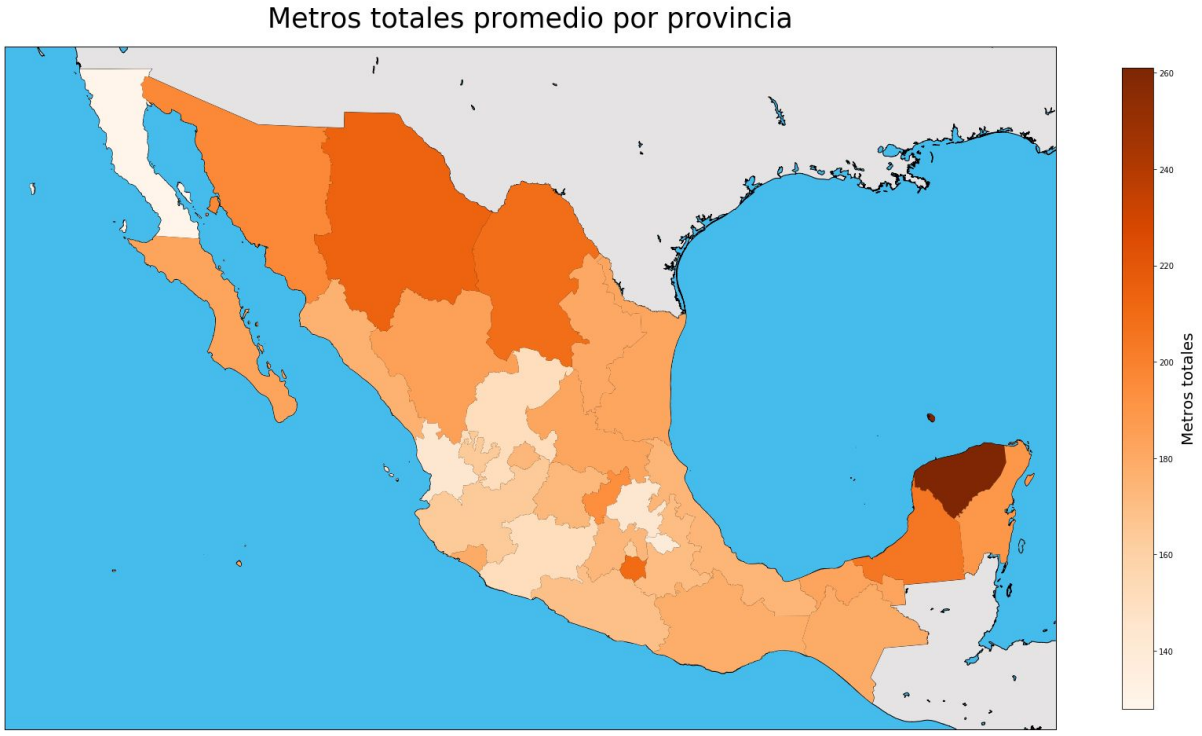


Cambiando el enfoque, ahora podemos ver las demás variables:

Antigüedad promedio por provincia



Se puede ver que los edificios más antiguos se encuentran en el centro de la ciudad de México y su Distrito Federal, lo cual es lógico ya que allí es donde se originó la ciudad.



A diferencia de lo destacado en los previos gráficos del centro de la ciudad, aquí predominan las propiedades con más metros sobre aquellas provincias más grandes, como por ejemplo las provincias de Chihuahua y Coahuila y en particular en la costa en la provincia en la provincia de Yucatán.

Análisis de atributos vs. tipos de propiedad

Para ver esto se conservaron ciertas columnas: *metros totales, metros cubiertos, baños, habitaciones, garages y antigüedad.*

Se separaron estas columnas entre dos grupos debido a naturaleza de los datos y su relación “lógica” entre ellos.

Primer grupo: *habitaciones, garages y baños* (atributos del interior de las propiedades)
Segundo grupo: *metros totales, metros cubiertos y antigüedad* (atributos generales de la propiedad).

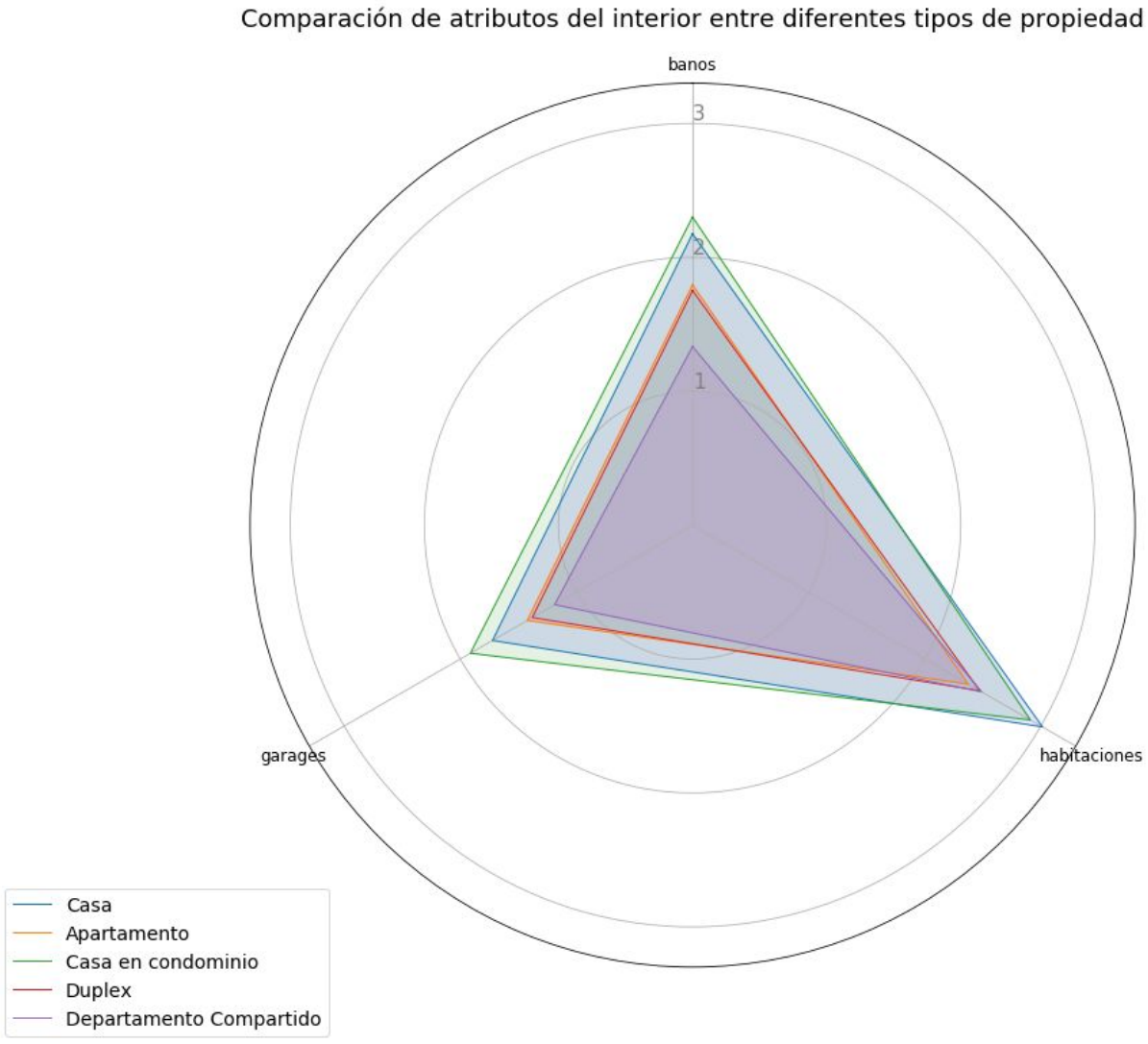
Tipos de propiedad**:** habitaciones, garages y baños

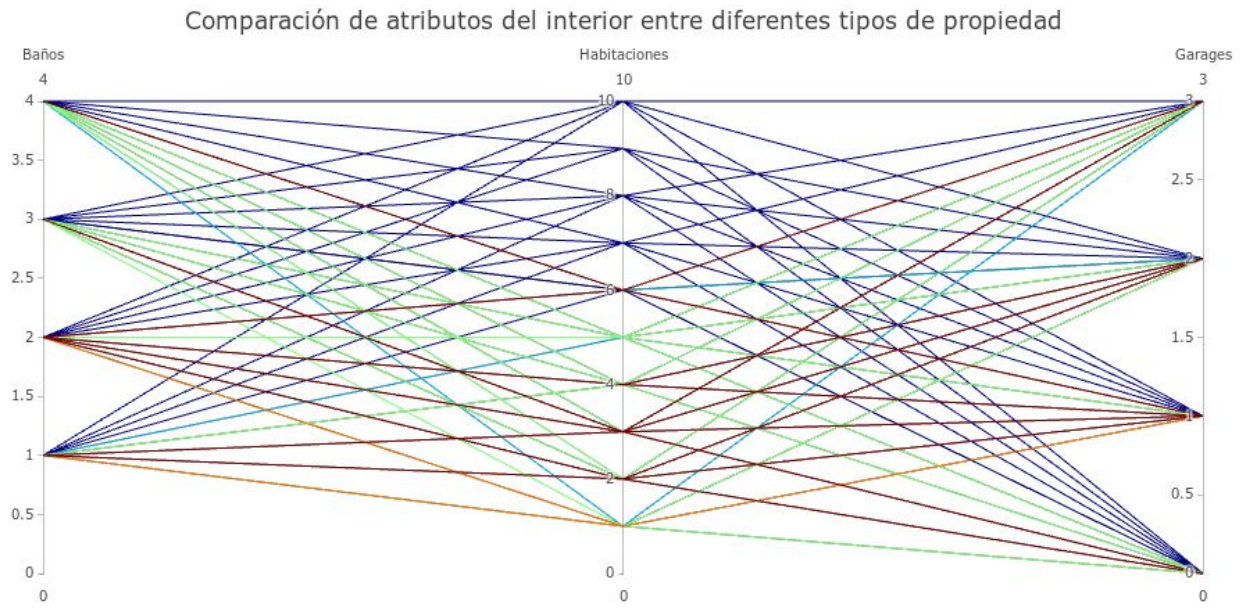
Luego de conservar las mencionadas columnas junto a su tipo de propiedad se eliminaron los valores nulos ya que los gráficos a continuación no aceptaban ese tipo de valores y se optó por no rellenarlos con el valor cero para no modificar la veracidad de los datos.

Se decidió conservar aquellos tipos de propiedad que aún tenían más del 70% de los datos luego de eliminar los NaNs, siendo estos los siguientes:

- Casa
- Apartamento
- Casa en condominio
- Departamento compartido
- Duplex

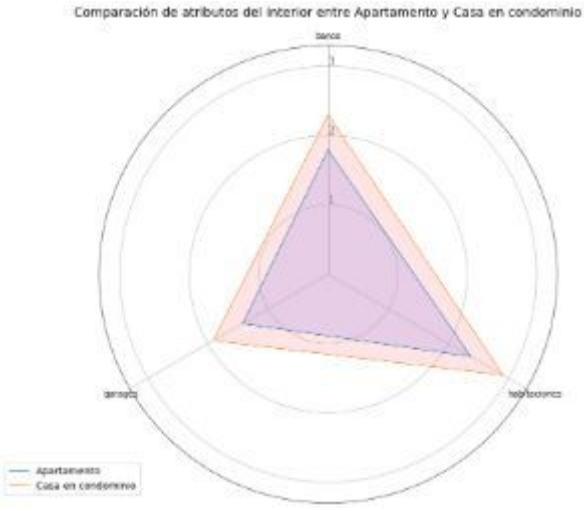
Para los radar charts se usó el valor promedio de todos.

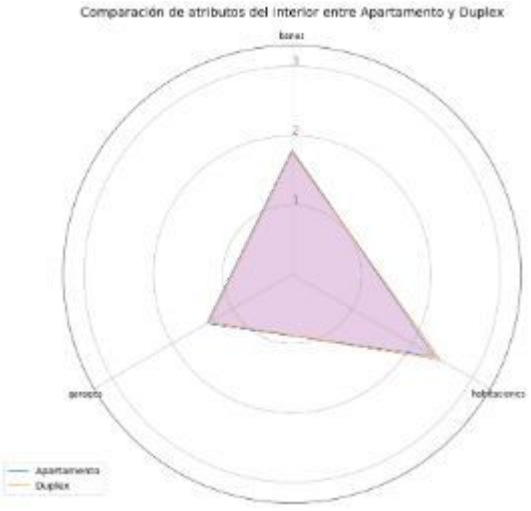




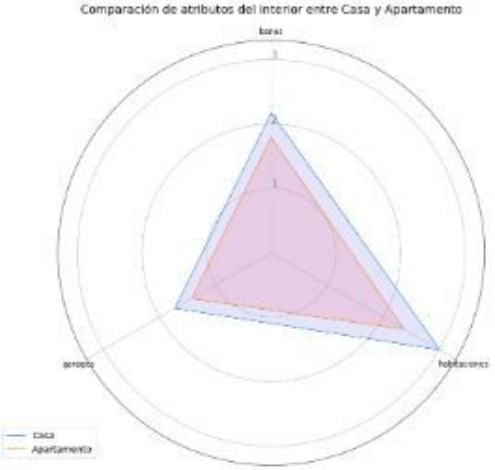
A continuación se optó por separar y hacer los gráficos de las variantes para poder notar más claramente las diferencias y similaridades entre los distintos tipos.

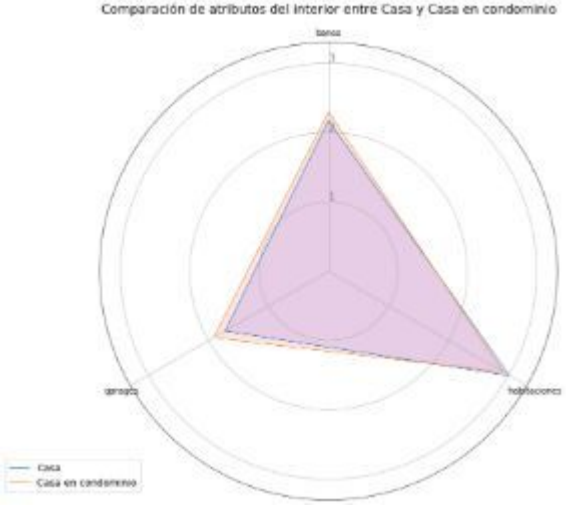
Se utilizaron radar charts ya que son más entendibles e equivalentes a los de coordenadas paralelas para este caso, desde nuestro punto de vista. Es importante notar que se redujeron los tamaños de los gráficos a continuación para reducir el tamaño del informe y se mostró lo importante en los comentarios. Para verlos más en detalle se recomienda acceder al repositorio.

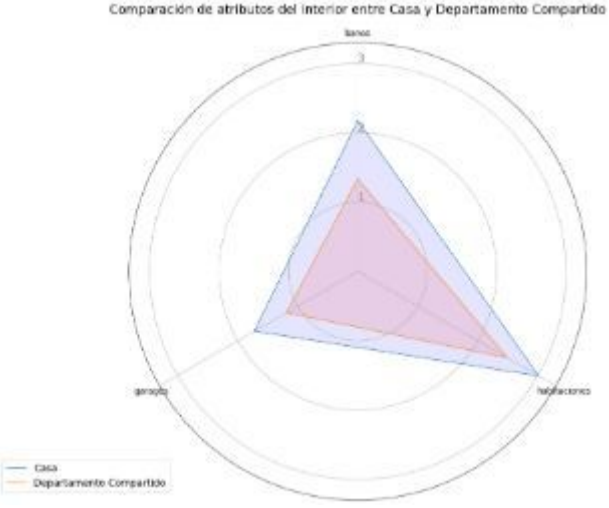

Comparación de atributos del interior entre Apartamento y Casa en condominio
banos
habitaciones
garaje
Apartamento
Casa en condominio

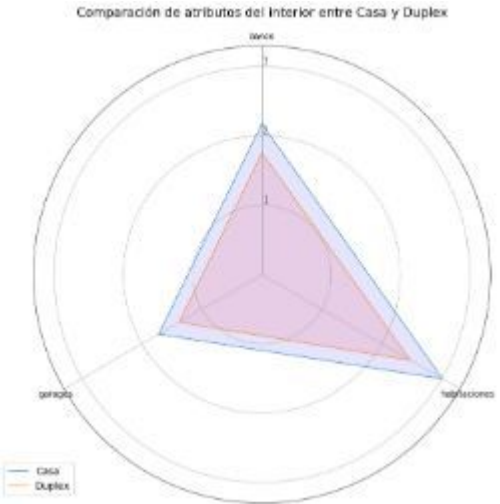

Comparación de atributos del interior entre Apartamento y Duplex
banos
habitaciones
garaje
Apartamento
Duplex

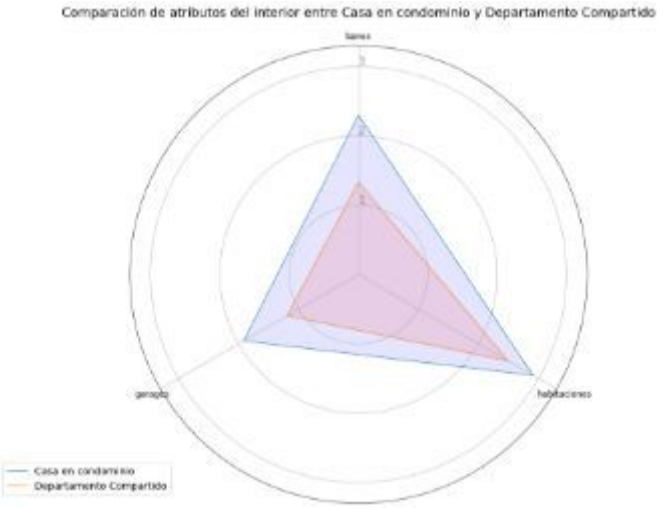

Comparación de atributos del interior entre Apartamento y Departamento Compartido
banos
habitaciones
garaje
Apartamento
Departamento Compartido

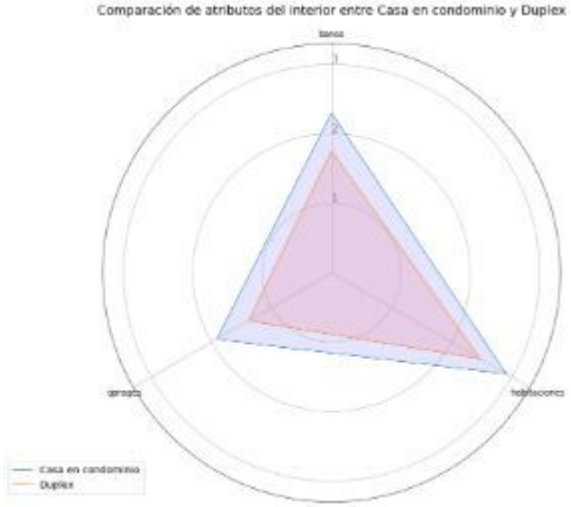

Comparación de atributos del interior entre Casa y Apartamento
banos
habitaciones
garaje
Casa
Apartamento

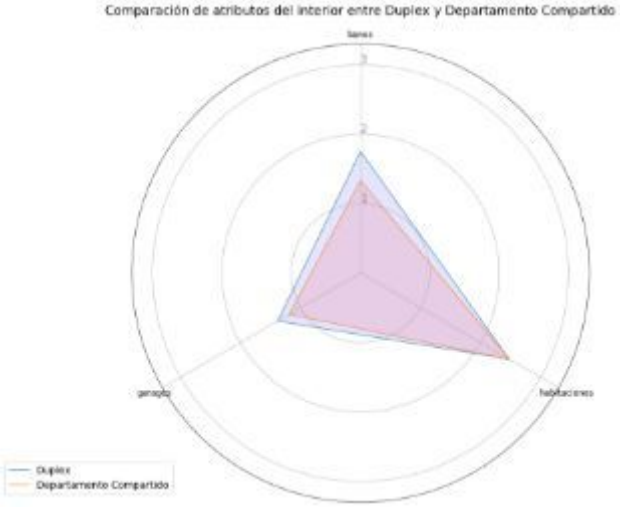

Comparación de atributos del interior entre Casa y Casa en condominio
banos
habitaciones
garaje
Casa
Casa en condominio


Comparación de atributos del interior entre Casa y Departamento Compartido
banos
habitaciones
garaje
Casa
Departamento Compartido









Comentarios sobre estas imágenes:

- *Casa* y *Casa en condominio* son parecidos como es esperado.
- *Apartamento* y *Duplex* son mucho más parecidos que *Apartamento* y *Departamento compartido*
- Aquellas que tienen mayor diferencia entre los tipos son *Casa en condominio* y *Departamento compartido*
- Aquellas que tienen mayor similaridad entre los tipos son *Apartamento* y *Duplex*

Tipos de propiedad: metros totales, metros cubiertos y antigüedad

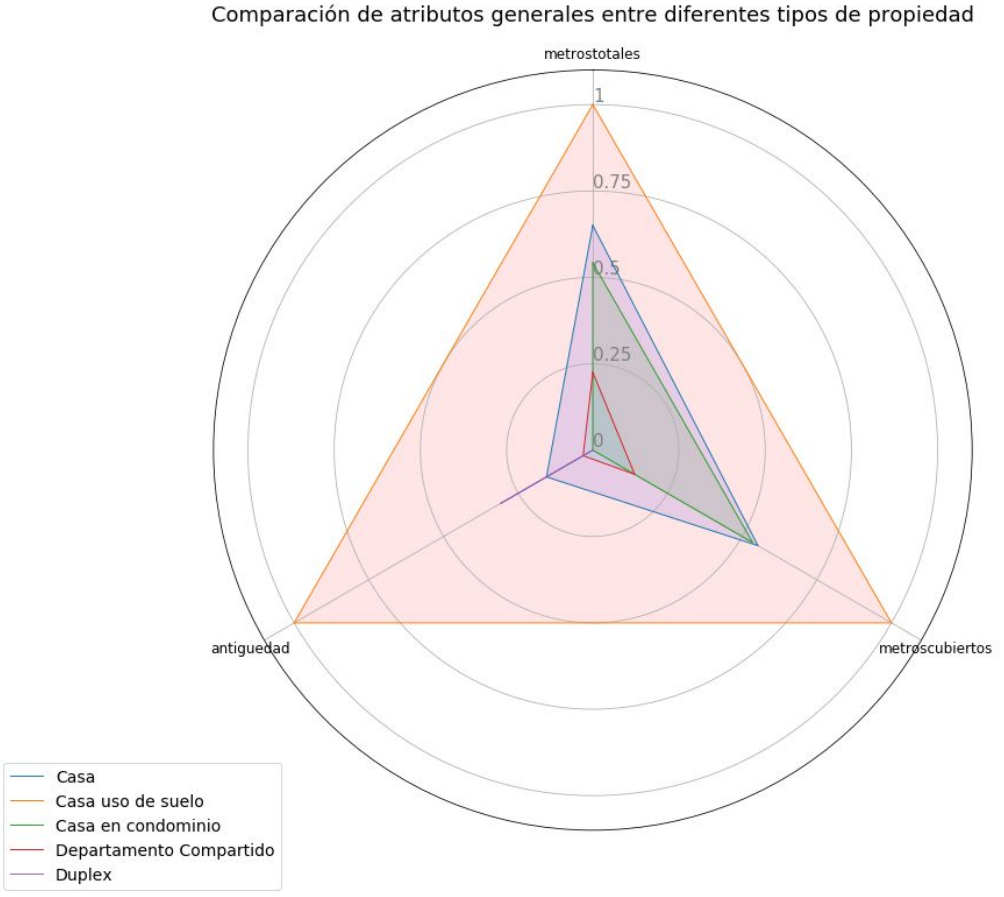
Al igual que en los tipos de propiedad anteriores, luego de conservar las mencionadas columnas junto a su tipo de propiedad se eliminaron los valores nulos por lo ya mencionado.

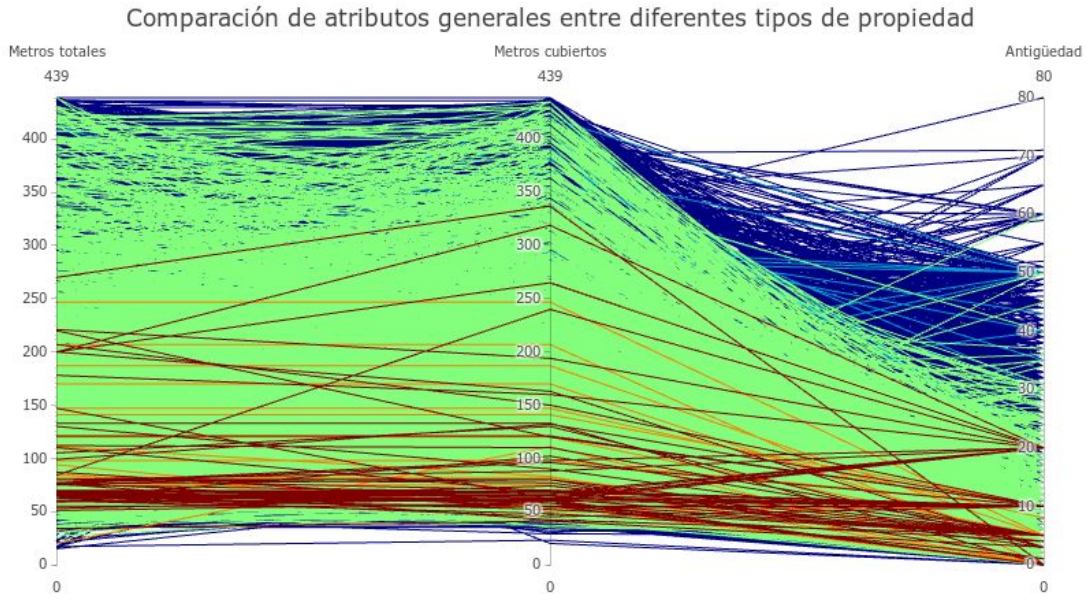
Se decidió conservar aquellos tipos de propiedad que aún tenían más del 60% de los datos luego de eliminar los NaNs, siendo estos los siguientes:

- Casa
- Casa uso de suelo
- Casa en condominio

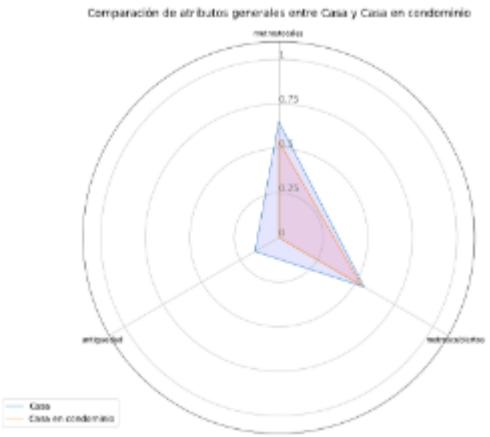
- Departamento compartido
- Duplex

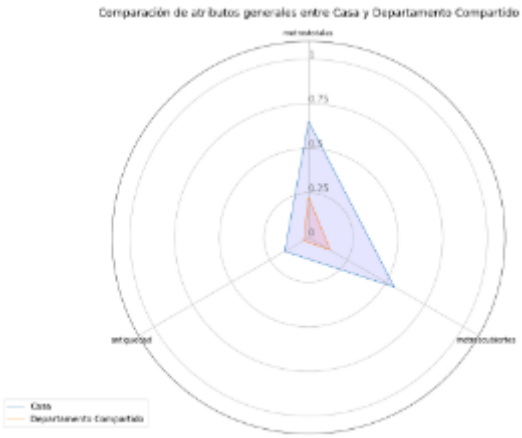
Para los radar charts se usó el valor promedio de todos y se los normalizo por la diferencia de escala que había.

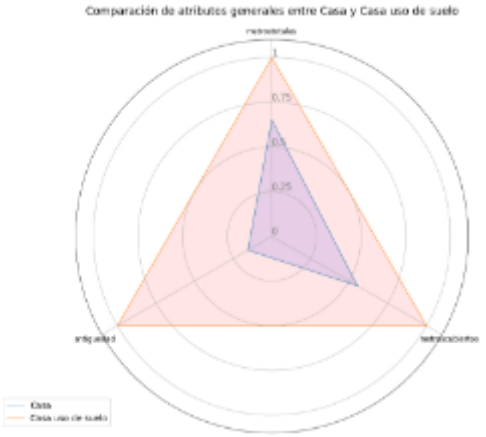


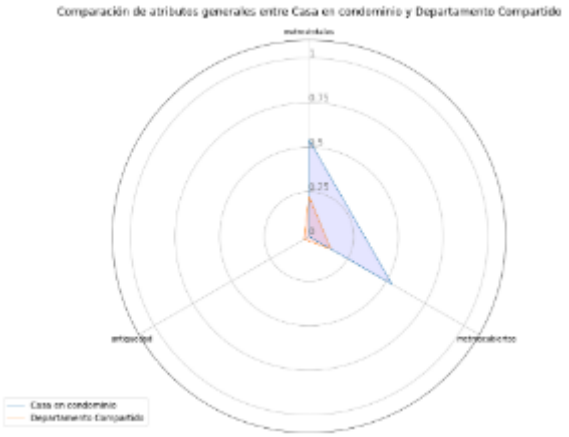


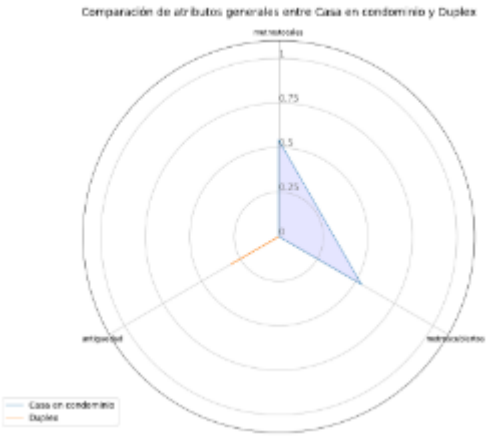
Al igual que con los previos tipos de propiedad, se optó por separar y hacer los gráficos y utilizar radar charts por los motivos ya mencionados. Es importante notar que se redujeron los tamaños de los gráficos a continuación para reducir el tamaño del informe y se mostró lo importante en los comentarios. Para verlos más en detalle se recomienda acceder al repositorio.

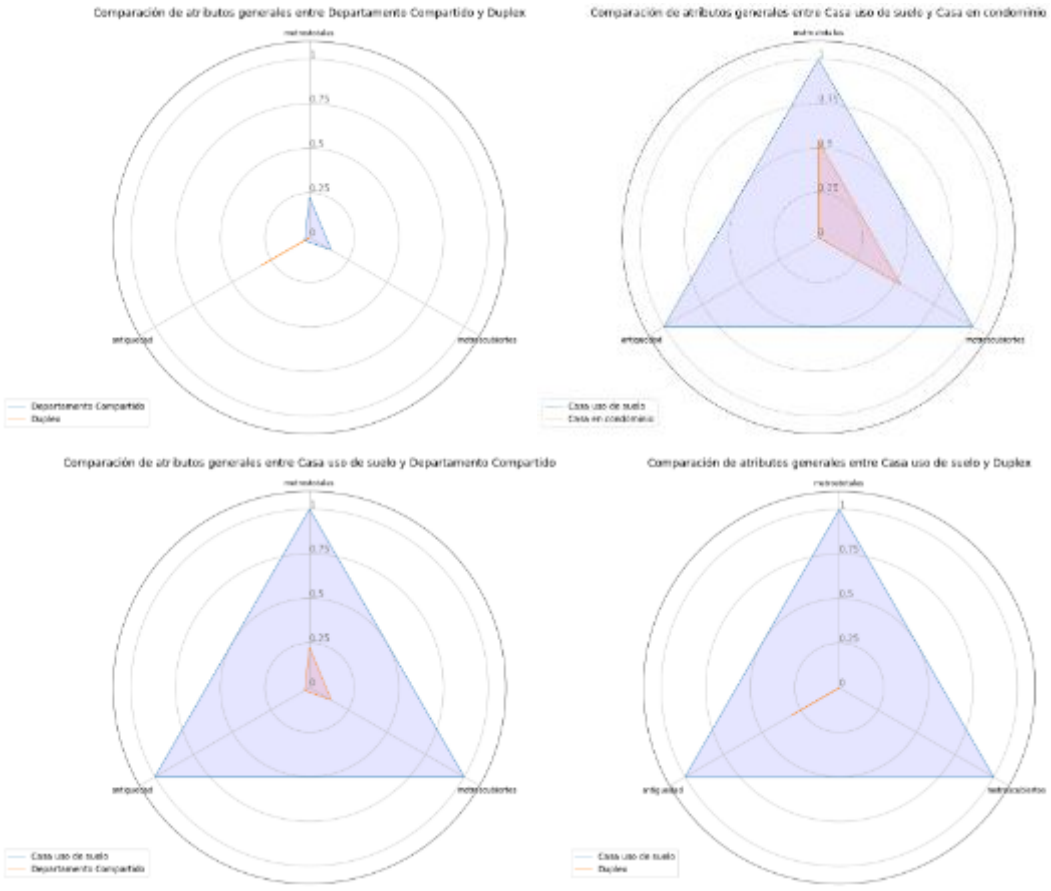












Comentarios sobre estas imágenes:

- *Casa* y *Casa en condominio* son parecidos como es esperado.
- *Casa uso de suelo* y *Casa* no son parecidos como es esperado.
- *Casa uso de suelo* y *Casa en condominio* no son parecidos como es esperado.
- Aquellas que tienen mayor diferencia entre los tipos son *Casa uso de suelo* con *Duplex* o *Departamento compartido*
- Aquellas que tienen mayor similaridad entre los tipos son *Casa* y *Casa en condominio*
- *Casa uso de suelo* son las más antiguas, las que cubren más metros y las que tienen más metros

Análisis de descripciones

Como primera medida decidimos realizar una purificación de las descripciones por cada publicación ya que contenían rastros de lenguaje HTML. Para esa tarea utilizamos la librería Beautiful Soap.

Ya con las descripciones limpias de código procedimos a aplicar NLP(Procesamiento del Lenguaje Natural) eliminando palabras comunes del vocabulario español (conjunciones, preposiciones y verbos comunes) y a su vez descartando palabras con un tamaño menor a cuatro, de esta manera logramos reducir en gran parte el tamaño de texto a analizar para poder generar los gráficos. La herramienta utilizada para esta tarea fue SpaCy.

Tomaremos solamente las 200 palabras más frecuentes en toda la nube de texto.

Para no realizar un solo gráfico en el cual el wordcloud se forme con la completitud de las descripciones se partió el dataset en 3 grandes categorías:

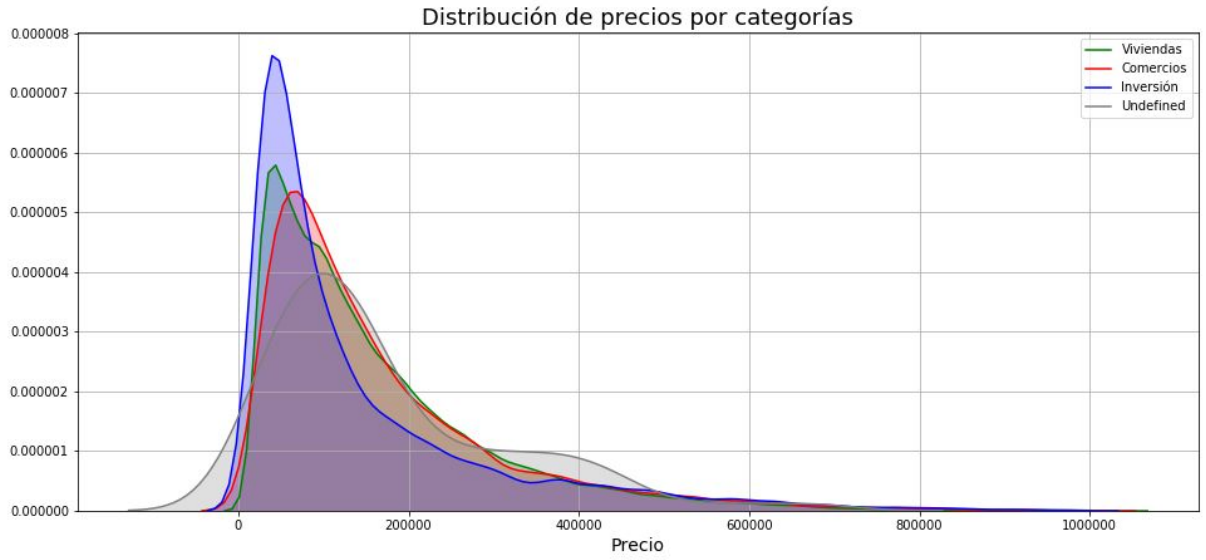
1. **Vivienda:** abarcando las categorías originales "apartamento", "casa en condominio", "casa", "villa", "duplex", "departamento compartido", "hospedaje", "quinta vacacional".

2. **Comercial:** abarcando las categorías "terreno comercial", "local comercial", "oficina comercial", "local en centro comercial", "bodega comercial", "inmuebles productivos urbanos".

3. **Inversión:** abarcando las categorías "casa uso de suelo", "terreno", "edificio", "huerta", "lote", "garage", "otros", "nave industrial", "rancho".

4. **Undefined:** abarcativa de las categorías “otros”, “nan”.

En el siguiente gráfico se encuentra la distribución de precios según estas categorías:



Las tres distribuciones principales muestran modas cercanas, mientras que la de indefinidos está algo desplazada hacia precios más caros. Todas llegan hasta valores altos en precio muy lejanos a la moda.

Conclusiones

En un principio, es importante destacar el posible rol de la información referida a la ubicación de las propiedades. El trabajo de completar coordenadas faltantes se realizó por medio de APIs y acorde a estos recursos se generó una dificultad al hacer muchas consultas para una extensa cantidad de registros, y lamentablemente se llegó al resultado de tan sólo poder recuperar un 10% de los datos.

Hubiese sido satisfactorio poder recuperar aunque sea una mayor proporción de esta información ya que creemos que uno de los supuestos a distinguir es la relación que podría existir entre la ubicación de ciertas propiedades y su precio, pero no hubo la suficiente información para poder justificar esto adecuadamente.

Otra cosa que podría haberse hecho con este tipo de información es relacionar la disposición de las propiedades con datos externos como el sistema de Metro (que se intentó incluir), Metrobús, seguridad, etc… las posibilidades son de alguna manera considerables.

Si bien por suerte se pudo ver tendencias globales en relación a la ubicación con los choropleth provistos, podría ser importante entonces, más allá de este análisis general, intentar en el siguiente trabajo la utilización de una diferente metodología para poder recuperar los datos específicos de la ubicación y así poder darles un mejor uso.

Dejando un poco de lado lo que está fuera de nuestro alcance por falta de datos, podemos hacer hincapié en lo que sí está.
El particular enfoque en intentar ver cómo se distribuye la variable del precio y cómo se relaciona con las posibles amenities y las demás características de las propiedades, nos permitió tener una mejor comprensión para saber qué rumbo tomar en el segundo trabajo.

A continuación algunos comentarios acerca de esto:

1. La importancia del precio por metro cuadrado
2. Relacionar las amenities y no quedarse con la idea de cada una por separado
3. Explorar cuál será la principal propuesta en relación a qué tipo de propiedad se vende.
4. La ubicación es sin duda un factor importante en el precio.
5. Se observó que en todos los casos la mayor cantidad de publicaciones se produce en la segunda mitad del año. Por otro lado, si bien la cantidad de registros que hay por día de semana fluctúa considerablemente, no se encontraron patrones de precios de propiedades por día de la semana de la publicación. Esto mismo se extiende a día del mes.

Dicho esto, creemos haber generado un buen panorama acerca de lo que se debe tener en cuenta para el siguiente TP, por lo que el grupo se siente bien encaminado y con una base sólida acerca de los datos, sus relaciones y el impacto de las distintas variables sobre la variable dependiente, precio.