FACULTAD DE INGENIERIA
Universidad de Buenos Aires


TRABAJO PRÁCTICO 1

# ANÁLISIS EXPLORATORIO DE DATOS

## [75.06 / 95.58] ORGANIZACIÓN DE DATOS

SEGUNDO CUATRIMESTRE DE 2019


INTEGRANTES

| Apellido | Nombre | Padrón | Mail |
|---|---|---|---|
| Aguerre | Nicolás | 102145 | naguerre@fi.uba.ar |
| Parafati | Mauro | 102749 | mparafati@fi.uba.ar |
| Secchi | Ana María | 99131 | asecchi@fi.uba.ar |

# ÍNDICE



## LINKS PRÁCTICOS

.

- Repositorio (GitHub): github.com/nicomatex/datos_tp1_2c2019.
- Kaggle: www.kaggle.com/nicolasaguerre/tp1-2c2019

# 1 INTRODUCCIÓN

En este informe se busca documentar el análisis exploratorio realizado sobre un dataset que contiene información sobre las publicaciones realizadas en **ZonaProp México**, una página web que brinda la posiblidad de disponibilizar una propiedad para su venta.

A lo largo del desarrollo de este informe, buscaremos plantear distintas interrogantes que permitan entender mejor la naturaleza de estos datos. Exploraremos las relaciones que existen entre las distintas variables, con el objetivo de encontrar dependencias que no resulten obvias a simple vista.

Finalmente desarrollaremos una serie de conclusiones o *insights* que resulten de interés para el lector.

# 2 ANÁLISIS INTRODUCTORIO AL DATASET

## 2.1 Estructura de los datos

Como primer paso en el análisis exploratorio, y antes de poder empezar a plantearnos cualquier interrogante, resulta necesario entender cómo están formados los datos.

El dataset proporcionado consta de un listado de propiedades de México que fueron publicadas en **ZonaProp** para su venta. Cada publicación cuenta con la siguiente información:

- **id**: Identificador numérico de cada publicación.
- **titulo**: Título con el que se publico la propiedad.
- **descripcion**: Descripción asignada a la propiedad.
- **tipodepropiedad**: Indica si se trata de una casa, un departamento, etc.
- **direccion**: Calle y/o altura de la propiedad.
- **ciudad**: Ciudad en la que se encuentra ubicada.
- **provincia**: Provincia en la que se encuentra ubicada.
- **antiguedad**: Cantidad de años desde que se construyó.
- **habitaciones**: Cantidad de habitaciones en la propiedad.
- **garages**: Cantidad de garajes que posee.
- **banos**: Cantidad de baños que posee.
- **metroscubiertos**: Metros cubiertos de la propiedad.
- **metrostotales**: Metros totales de la propiedad.
- **idzona**: Identificador numérico para la zona en la que se encuentra.
- **lat**: Latitud geográfica.
- **lng**: Longitud geográfica.
- **fecha**: Fecha en la que fue realizada la publicación.
- **gimnasio**: Se indica si posee o no gimnasio.

- **ususmultiples**: Se indica si posee o no SUM.
- **piscina**: Se indica si posee o no piscina.
- **escuelascercanas**: Se indica si la propiedad se encuentra en la cercanía de escuelas.
- **centroscomercialescercanos**: Se indica si la propiedad se encuentra en la cercanía de centros comerciales.
- **precio**: Precio con el que se publicó la propiedad.

Como vemos, tenemos suficiente información para poder encontrar relaciones interesantes entre las distintas variables que conforman una publicación.

### 2.2 Preparación de los datos

Realizado el análisis introductorio a los distintos datos que fueron proporcionados, consideramos de vital importancia realizar una *preparación* de los mismos, a fin de poder facilitar el trabajo posterior. Con *preparación* nos referimos más específicamente a:

- **Manejo de campos faltantes**: Como en todo set de datos, es esperable encontrarnos con publicaciones que no se encuentren completas y que tengan datos ausentes, valores a los que nos referiremos como **valores nulos** o **nulls**. Frente a estas situaciones, es necesario tomar un criterio para poder trabajar. El criterio elegido en esta ocasión fue mantener todas las publicaciones que contengan valores nulos para no perder datos de manera innecesaria. Por ejemplo, si buscamos hallar una relación entre el precio y la cantidad de ambientes, no nos interesaría que el valor de latitud sea nulo para dicha publicación.
- **Tipos de los datos**: Con el principal objetivo de ahorrar memoria, es importante realizar un proceso de conversión a los distintos tipos de datos que sean provistos para llevarlos a su *tipo ideal*, es decir, a un tipo de dato que sea capaz de mantener la misma información utilizando menos recursos. En otros casos, también nos interesará convertir los datos a otro tipo para poder facilitar su análisis como puede ser el caso de la **fecha**.

## 3 INTERROGANTES

Conocida ya la estructura básica del set de datos y transformado el mismo para lograr aprovechar de manera óptima los recursos y facilitar su manipulación, nos proponemos realizar un **brainstorming** (*lluvia de ideas*) con el objetivo de identificar a las variables que resulten de interés analizar.

### 3.1 Distribución de variables

Para comenzar, sería interesante tener a disposición una serie de gráficos que nos introduzcan rápidamente en el dataset. Para lograr esto, abarcaremos las siguientes interrogantes:

- *¿Cómo se distribuyen las propiedades según tipo de propiedad?*
- *¿Cómo se distribuyen según provincia? ¿Y según ciudad?*
- *¿Qué proporción de propiedades publicadas tienen pileta? ¿Y qué proporción gimnasio? ¿Y garaje? ¿Y SUM? Etc.*

- *¿Cuántas propiedades cuentan con escuelas cercanas? ¿Cuántas con centros comerciales?*
- *¿Cuál es la proporción entre metros cubiertos y metros totales en general? ¿Y por tipos de propiedad?*
- *¿Qué tan antiguas son las propiedades en general?*
- *¿Qué tantas habitaciones tienen las propiedades?*

### 3.2 Precio

Normalmente, el primer parámetro que uno fija a la hora de buscar una propiedad. Es por esto que será nuestro eje central en el análisis, con los siguientes objetivos en mente:

- *Encontrar su distribución fijando distintos parámetros como sean las ciudades, provincias, etc.*
- *Encontrar factores que influyan de manera directa o indirecta en el mismo.*
- *¿Cómo se distribuye el precio del metro cuadrado en las distintas ciudades?*

### 3.3 Localización

El atributo más importante de cualquier propiedad es su **localización geográfica**. Con esta dirección buscaremos profundizar en los siguientes aspectos:

- *Encontrar la distribución de las publicaciones a lo largo del territorio Mexicano: ¿Qué tanta disponibilidad hay en cada ciudad? ¿Y en cada provincia? ¿Dónde se encuentran las propiedades más nuevas y donde las más antiguas? Etc.*
- *Encontrar tendencias dependientes de la misma.*
- *¿Qué localizaciones son más propensas a tener piscinas en sus propiedades? ¿Y gimnasios? Etc.*
- *¿Cómo afectan las cercanías a escuelas, gimnasios, etc. al precio? ¿Varía esto según la ciudad?*

### 3.4 Evolución de ZonaProp en México

Teniendo información sobre las fechas de publicación, resulta interesante plantear las siguientes interrogantes:

- *¿En qué fechas tuvo ZonaProp más actividad?*
- *¿Cómo fue la evolución de publicaciones realizadas a lo largo del tiempo?*
- *¿Cuáles son las provincias más activas? ¿Y cuáles las ciudades?*

### 3.5 Aspectos secundarios

Aprovechando la cantidad de información que poseemos de cada publicación, nos proponemos utilizarla para obtener alguna conclusión en los siguientes apartados:

- *Aproximación a la devaluación de la moneda Mexicana frente al dólar (utilizando para esto la evolución del precio del metro cuadrado).*

# 4 DISTRIBUCIÓN DE LOS DATOS

Un primer aspecto a tener en cuenta del análisis es: ¿Cómo se distribuyen los datos en sí? En esta sección se analizará la distribución y proporciones de los datos, con el motivo de tener en cuenta su estructura básica y su composición a la hora de analizar otras variables.

## 4.1 Geográfica

La interrogante que se desea responder con este análisis es: ¿De que partes de México provienen las publicaciones pertenecientes a este dataset? Para responder a esta interrogante, se presenta el siguiente gráfico, representando la **densidad de publicaciones según ubicación geográfica**:


**Figura 1:** Distribución de precio por provincia

Se puede observar que el grueso de los datos caen en el centro de México: **Distrito Federal**. También podemos ver que algunos puntos caen fuera del territorio, clara muestra de que nuestro dataset está 'sucio', es decir, que posee datos que fueron cargados de manera errónea.

## 4.2 Según tipo de propiedad

Tenemos muchos datos, pero, ¿qué proporción de los mismos pertenece a **casas**? ¿Y qué proporción a apartamentos? Utilizaremos un **doughnut plot** para ilustrar esto:

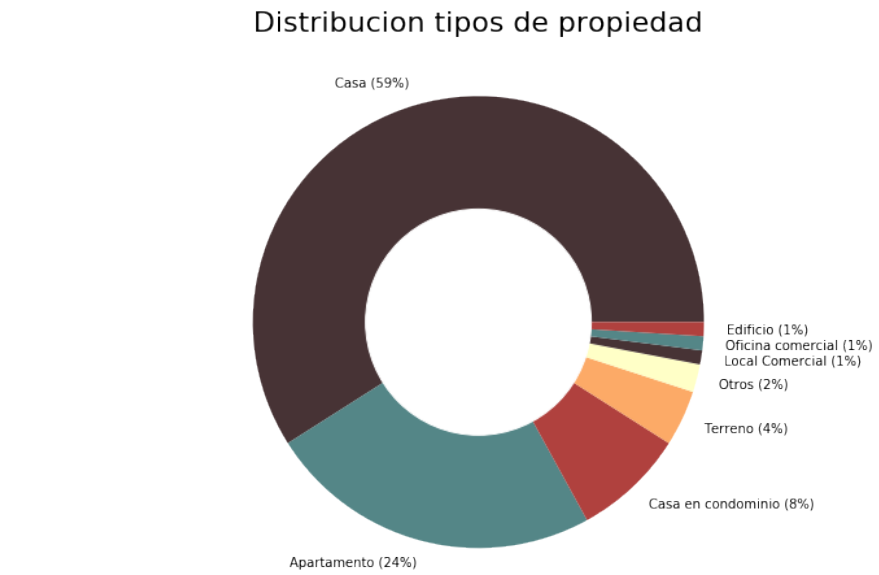


**Figura 2:** Distribución según tipo de propiedad

Vemos entonces que nuestro *dataset* esta formado principalmente por **Casas** y **Apartamentos**. Basaremos nuestros futuros análisis en esta importante conclusión.

### 4.3 Según atributos extra (piscinas, garajes, gimnasios, etc.)

Nos interesa saber, antes de analizar en profundidad, cuál es la proporción de propiedades que cuentan con atributos extra como por ejemplo piscina, garaje, etc. Para ilustrar esto, elegimos un **stacked bar plot**:

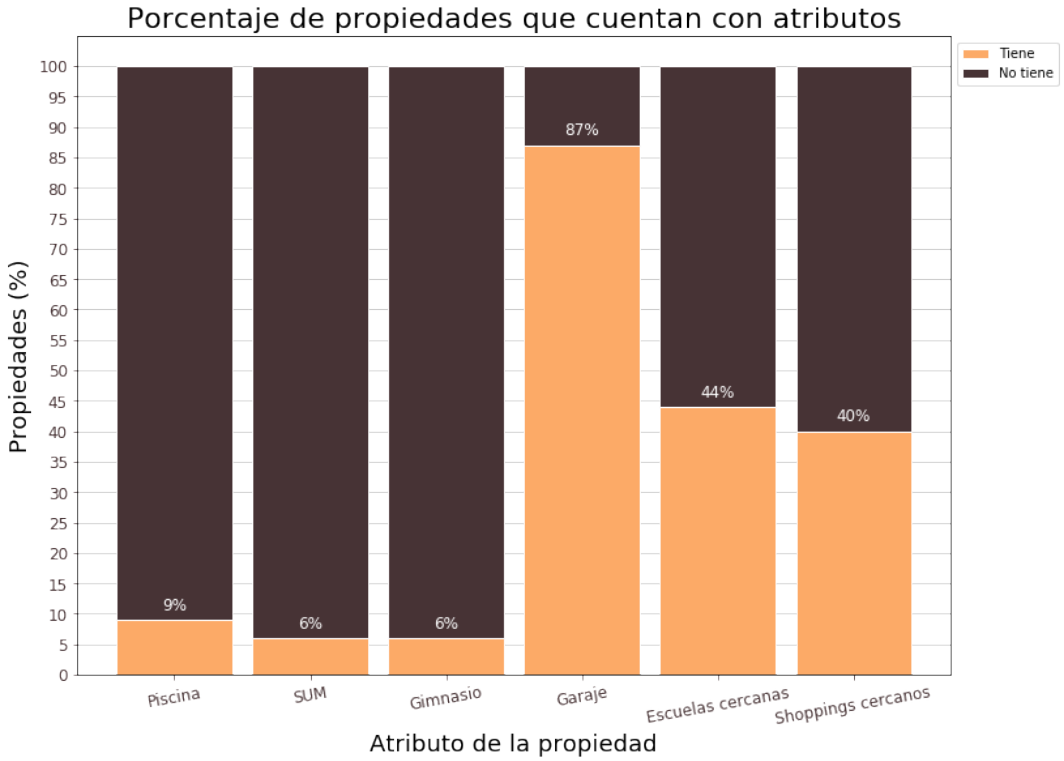


**Figura 3:** Porcentajes de propiedades que cuentan con un determinado atributo

Resulta muy llamativo ver que casi un 90 % de las propiedades **tienen garaje**, por lo que luego analizaremos si hay alguna relación entre la antigüedad de la propiedad con los garajes que posee, para ver si encontramos alguna tendencia que explique este fenómeno.

En cuanto a las cercanías a escuelas, también nos llevamos una sorpresa: el 44 % de las propiedades cuentan con una cercana. Si confiamos en la información que nos brinda ZonaProp (en realidad no sabemos cuál es el requisito que imponen para poder decir que la escuela está cerca), podemos afirmar que **casi la mitad de las propiedades brindan facilidades a la hora de pensar en el estudio de los niños**.

A su vez, observamos que 4 de cada 10 propiedades tienen un shopping cercano, lo que probablemente influya de forma directa sobre el precio. Indagaremos en esto más tarde.

## 4.4 Relación entre atributos extra

¿Existe alguna relación entre los atributos extra de una propiedad? Por ejemplo, de las propiedades que tienen piscina, ¿cuales tienen también gimnasio? Se ilustra en el siguiente gráfico dicha relación:

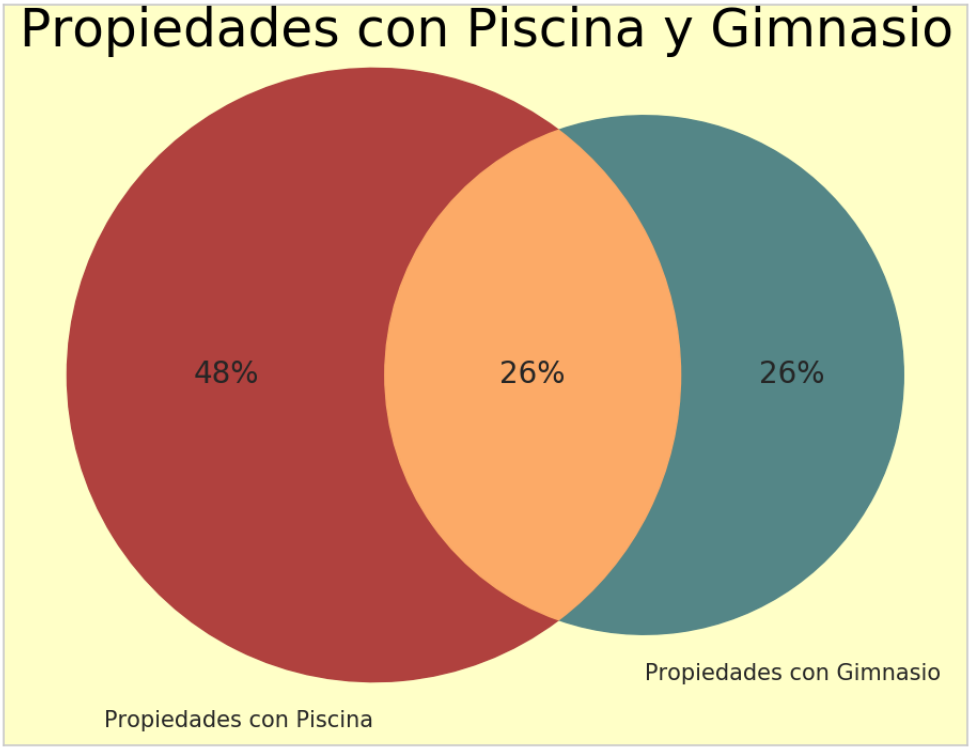


**Figura 4**: Propiedades con piscina y/o con gimnasio

## 4.5 Proporción de metros cubiertos por metros totales

Puede resultar interesante en el futuro saber que proporción de los metros totales de una propiedad están cubiertos. Para representar dicha interrogante, realizaremos un **stacked bar plot**:

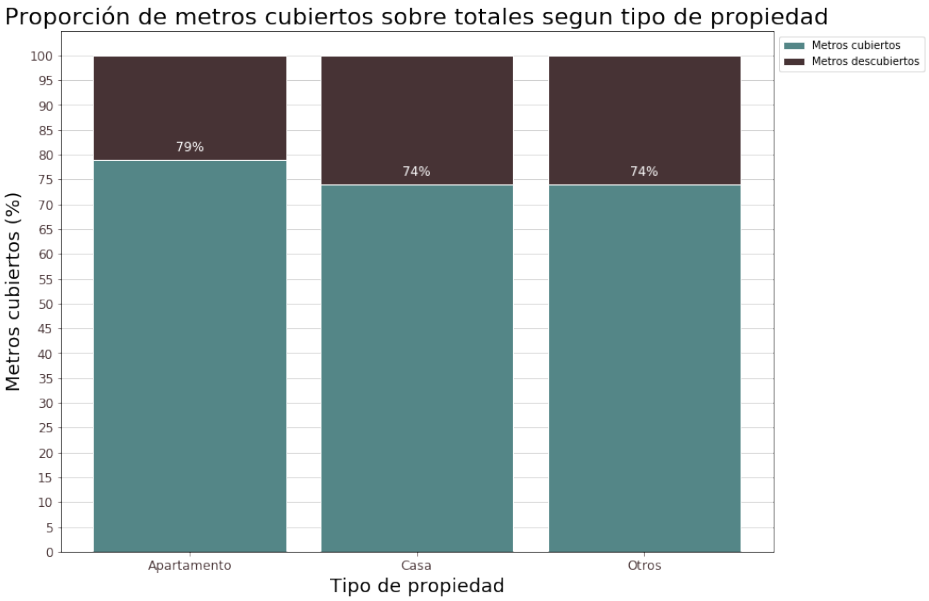


**Figura 5:** Proporción de metros cubiertos sobre totales según tipo de propiedad

Como resultado se observa que en general, aproximadamente el **75 % de los metros de las propiedades son cubiertos**. En particular, podemos ver que para los *apartamentos*, la cifra asciende hasta **casi un 80 %**: tiene mucha lógica, debido a que los apartamentos en general no poseen metros descubiertos. De hecho, nos llama la atención que tengan un 20 % de metros descubiertos.

## 4.6 Según antigüedad

Ahora intentaremos encontrar la estructura de los datos en cuánto a la antigüedad de las propiedades. ¿Dónde se encuentra el grueso de las publicaciones? Utilizaremos un **bar plot** para ilustrar los resultados:

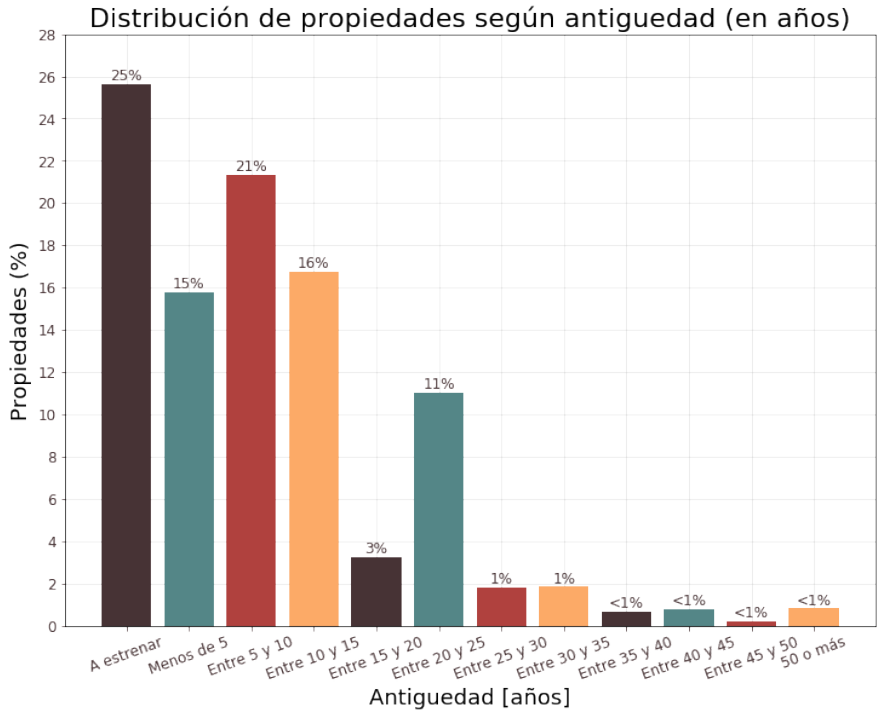


**Figura 6:** Distribución de propiedades según antigüedad

Observado el gráfico, podemos extraer las siguientes primeras impresiones:

- Una de cada cuatro casas está **a estrenar**.
- Un 60 % de las casas tienen **menos de 10 años** de antigüedad.
- Sólo un 10 % de las casas tienen **más de 25 años** de antigüedad.

### 4.7 Según cantidad de habitaciones

¿Qué tendencia marca a las propiedades en cuanto a la cantidad de habitaciones que estas poseen? Intentaremos responder esto utilizando un **pie plot**:

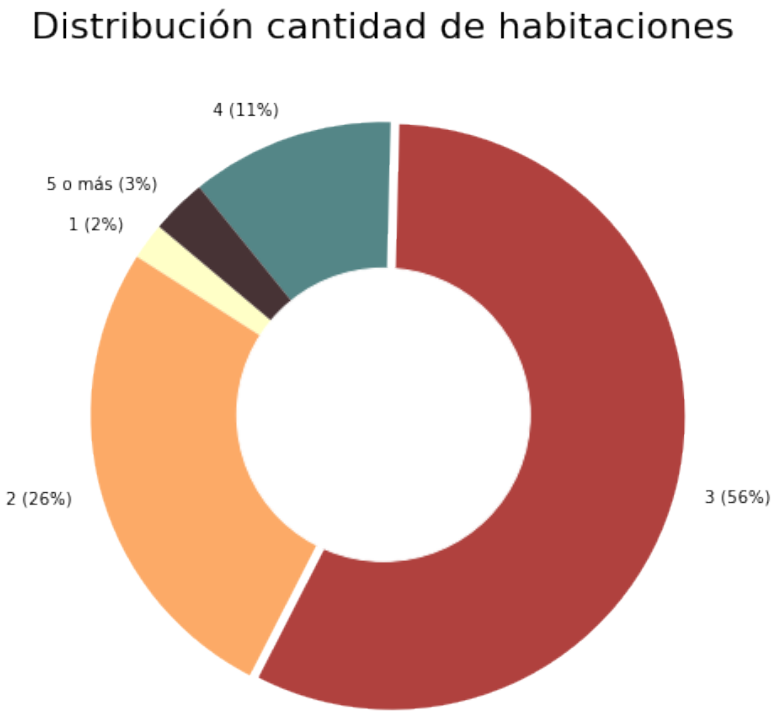


**Figura 7:** Distribución de propiedades según cantidad de habitaciones

Vemos que un 56 % de las propiedades tienen **tres habitaciones**. Probablemente esto se deba a que una familia tipo está compuesta por una pareja con dos hijos, familia que precisaría de tres habitaciones para vivir cómodos. También se considera que tres habitaciones es el estándar a seguir en los inmuebles modernos.

Algo que nos llama la atención de este gráfico es que sólo un 2 % de las propiedades son **mono-ambientes**. De todas formas, podemos rápidamente darnos cuenta del porqué: como vimos anteriormente, un 60 % de las publicaciones son **casas**, y prácticamente ninguna casa tiene sólamente una habitación.

# 5 ANÁLISIS DEL PRECIO

La primera pregunta que surge de forma casi inmediata al mirar el set de datos planteado es la siguiente: ¿Cómo afectan las diferentes variables al precio de las propiedades? Para responder esta pregunta, se analizará si existe **correlación** entre cada una de las mismas y el precio final de la propiedad.

## 5.1 Distribución según ubicación

### 5.1.1 *Por provincia*

Empezamos a continuación el análisis de la dependencia del precio para con la zona geográfica. ¿Cómo se distribuyen los precios de las propiedades según la provincia en la que ésta se encuentre? Para este análisis, se muestra a continuación un *BoxPlot* que muestra los rangos de precio para cada una de las provincias.

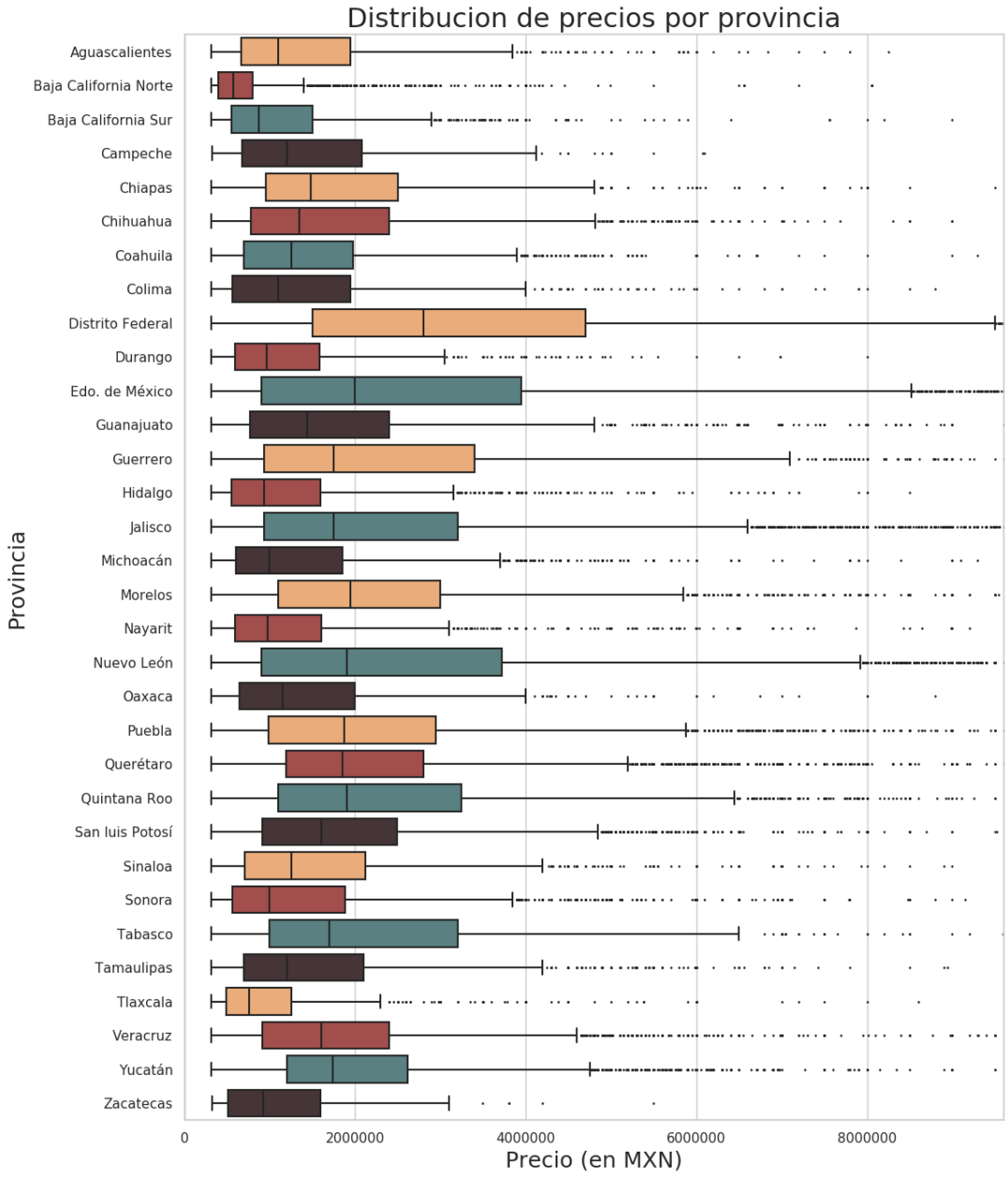


**Figura 8:** Distribución de precio por provincia

Como era de esperarse, el rango de precios mas amplio (y cuyo limite superior es mas elevado) corresponde a Distrito Federal, seguido por Nuevo Leon y Estado de México. Cabe destacar que en este gráfico, algunos de los *outliers* quedan por fuera de la escala horizontal.

### 5.1.2 *Por Ciudad*

Ahora analizamos como varía el precio promedio según **ciudad**, teniendo en cuenta las ciudades con más de **2000 publicaciones**. Consideramos que para aproximar mejor el promedio, son de mayor interés, en este caso, las ciudades que con más publicaciones cuentan.

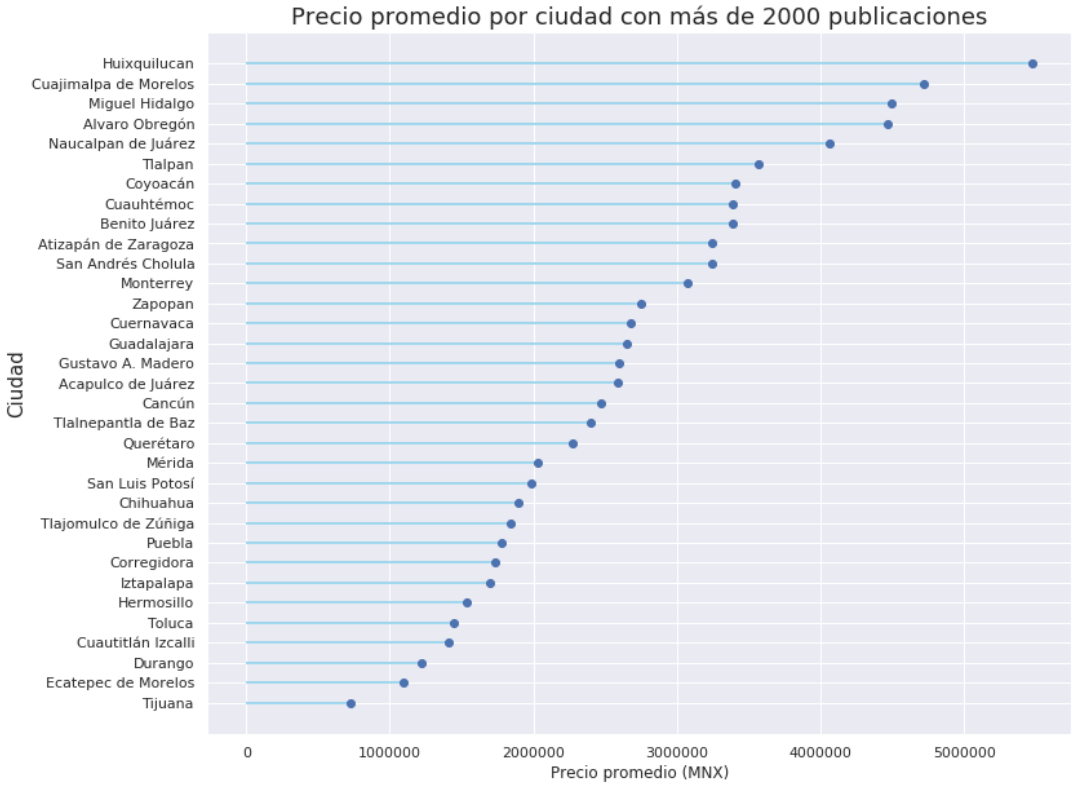


**Figura 9:** Precio promedio por ciudad

En lo alto se encuenta la ciudad de **Huixquilucan**, con un precio promedio de 5.475.000 MXN, lo cual no llama la atención por ser par de Estado de Mexico. Por otro lado, en lo bajo se encuentra la ciudad de **Tijuana**, con un precio promedio de 728.000 MXN. Ciudades **turísticas** como **Cancún**, **Acapulco de Juárez** y **Guadalajara** rondan aproximadamente por el mismo precio promedio, 2.500.000 MXN.

### 5.1.3 *Ciudades mas caras y mas baratas*

Ahora analizaremos cuáles son las diez **ciudades más caras** y cuáles las diez **más baratas**.

Como puede ser interesante para el lector, distinguir entre los principales tipos de propiedad, hemos decidido realizar el análisis para casas y para apartamentos por separado (como se observó anteriormente en *Distribución de los tipos de propiedad*, el 90 % de los datos está formado por *Casas* y *Apartamentos*). También es importante aclarar que se tomó como criterio trabajar con las ciudades que tengan al menos 500 publicaciones para evitar posibles *outliers*.

Primero, analizaremos los resultados para las **casas**:

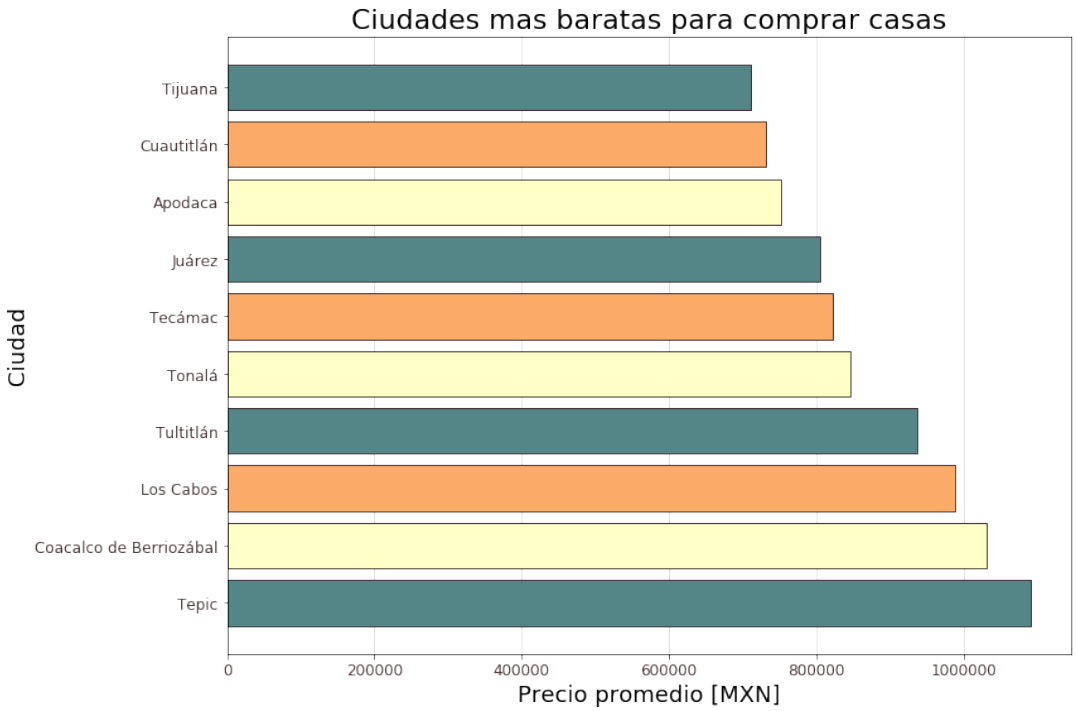


**Figura 10:** Diez ciudades más baratas (casas)

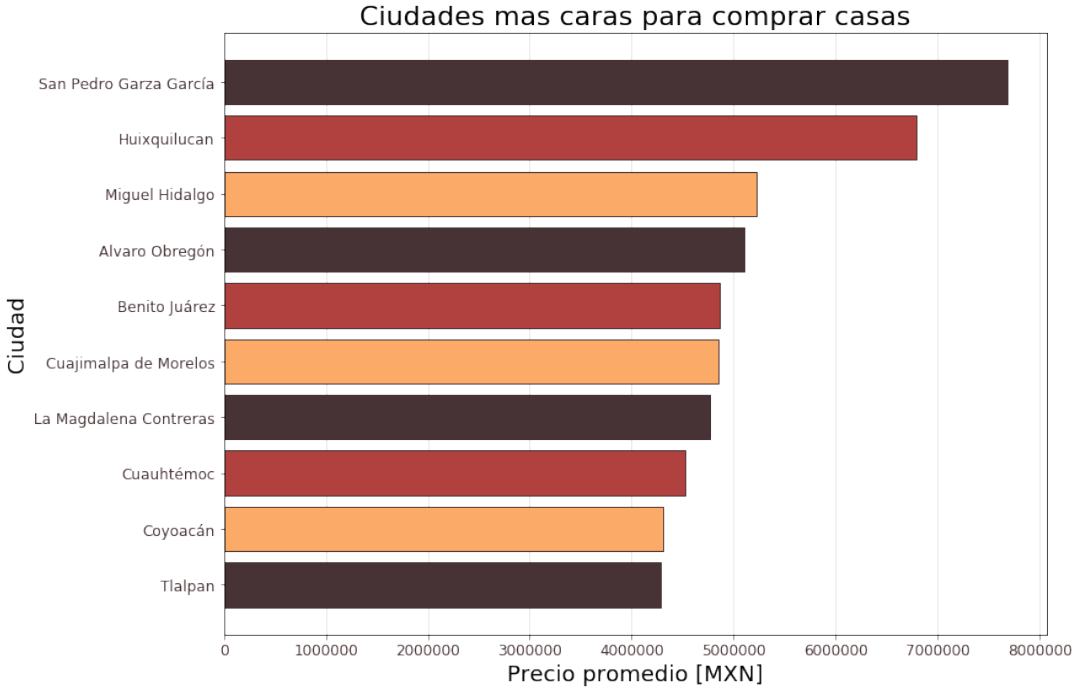


**Figura 11:** Diez ciudades más caras (casas)

Podemos ver que la ciudad de **San Pedro Garza García** (ciudad que forma parte del área metropolitana de **Monterrey**) se posiciona como la ciudad más cara a la hora de buscar casas con un precio rondando los 7.500.000 MXN. Es seguida por **Huixquilucan**, ciudad que pertenece a **Distrito Federal** (la provincia más cara de todo México según vimos anteriormente), cuyo precio se acerca a los 7.000.000 MXN.

En el otro extremo, vemos que **Tijuana**, seguida por **Cuautitlán**, son las ciudades en las que es más barato comprar una casa, con un precio al rededor de los 700.000 MXN. Es interesante observar que Tijuana es la ciudad que se encuentra más al **sur** de México, en la frontera con **Estados Unidos**. Frente a lo que uno creía, resulta que es más barato comprar una casa allí.

Ahora, representamos los resultados para los **apartamentos**:

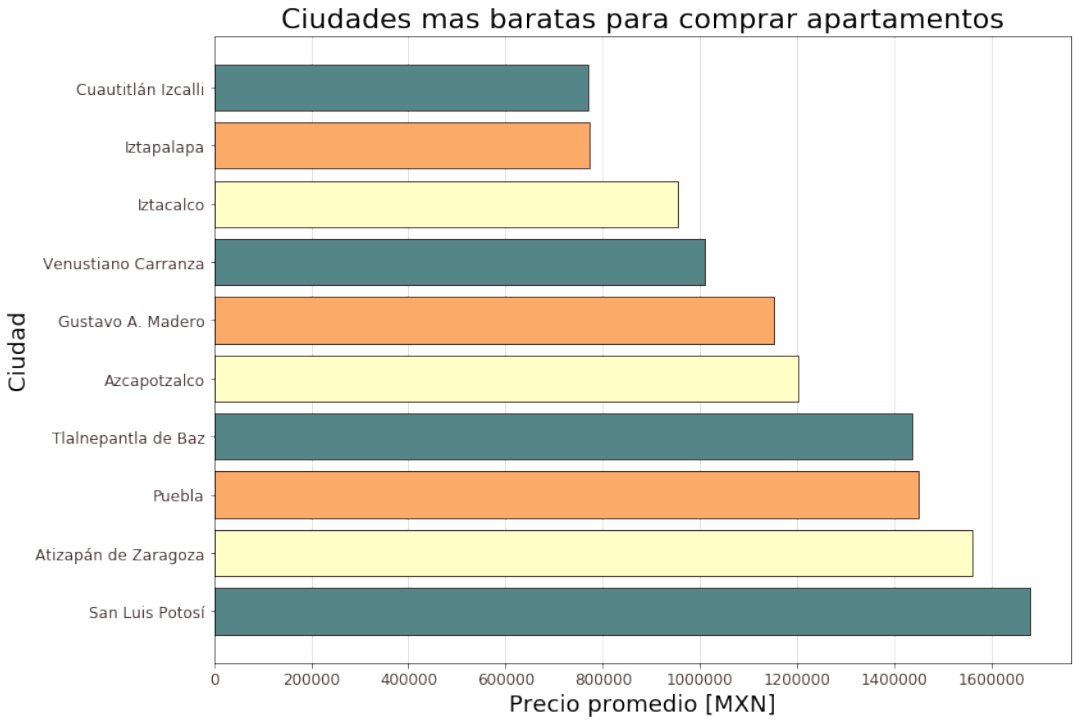


**Figura 12:** Diez ciudades más baratas (apartamentos)

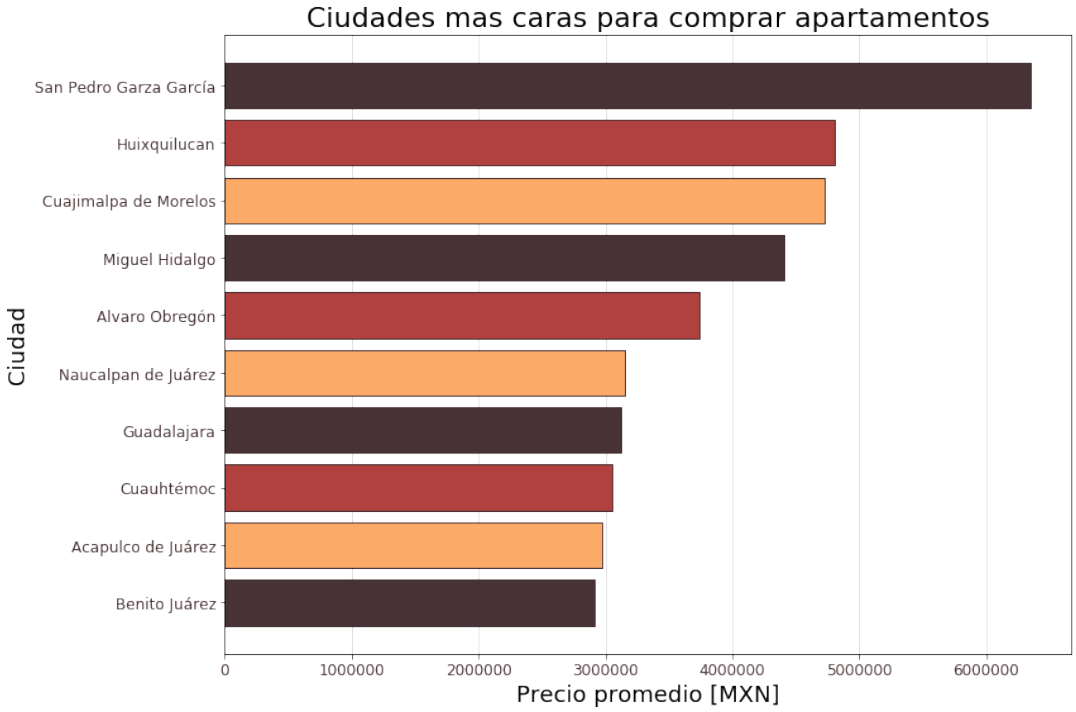


**Figura 13:** Diez ciudades más caras (apartamentos)

En cuanto a los apartamentos, vemos que **San Pedro Garza García** es nuevamente la ciudad más cara, pero esta vez, por mucha diferencia contra su seguidora, que siguie siendo **Huixquilucan**. En este caso, comprar un apartamento en San Pedro Garza García costará aproximadamente 6.500.000 MXN, mientras que uno en Huixquilicuan no llega a costar 5.000.000 MXN.

Por el otro lado, se ve que **Cautitlán Izcalli**, que resulta ser una ciudad vecina a **Cautitlán**, es la más barata para comprar apartamentos, al igual que su seguidora, **Iztapalapa**. Ambas tienen un precio que ronda los 800.000 MXN.

### 5.1.4 *Según latitud*

Otra pregunta interesante es ¿existe una relación entre el precio de una propiedad y que tan al norte dicha propiedad se encuentre? Quizás por una cuestión de cercanía con Estados Unidos, o por alguna otra razón. Para analizar lo recién mencionado, se muestra a continuación gráficos correspondientes al precio promedio según la latitud, tanto para todas las propiedades, como para casas y apartamentos por separado.

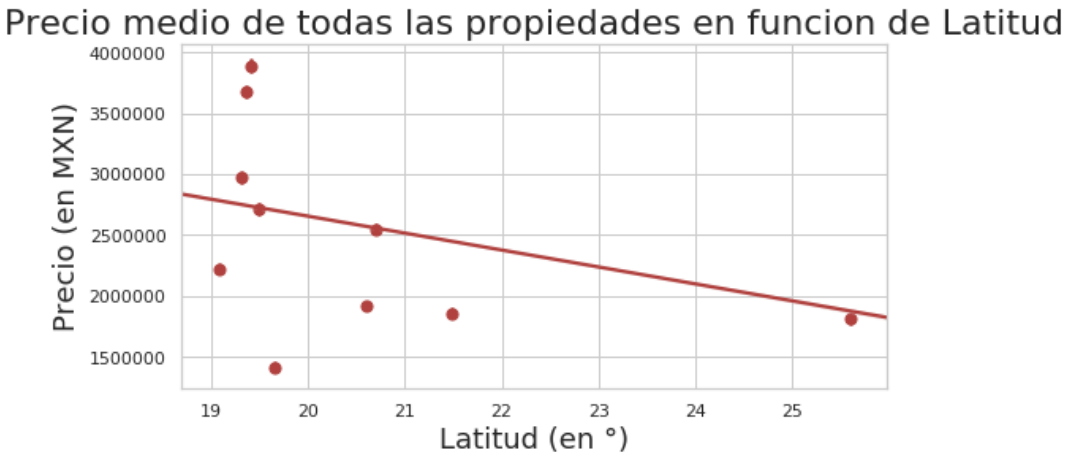


**Figura 14:** Dependencia del precio con la latitud

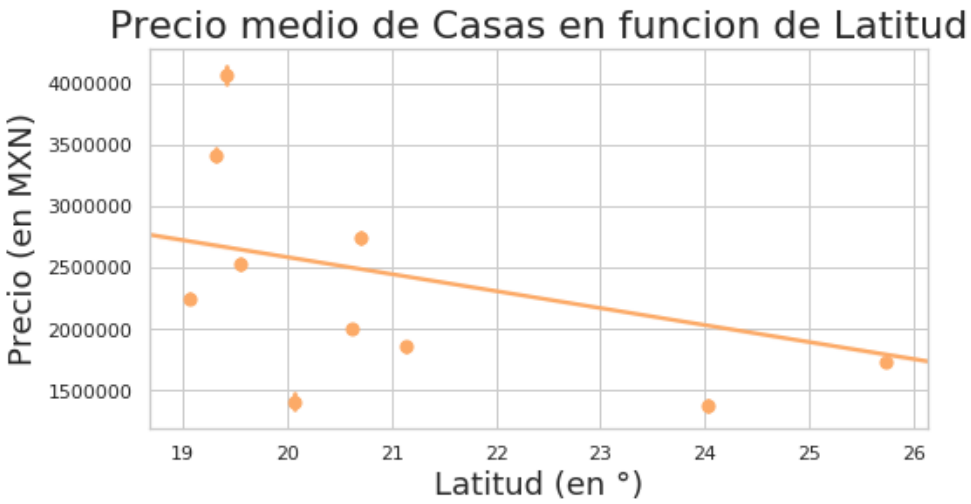


**Figura 15:** Dependencia del precio con la latitud (casas)

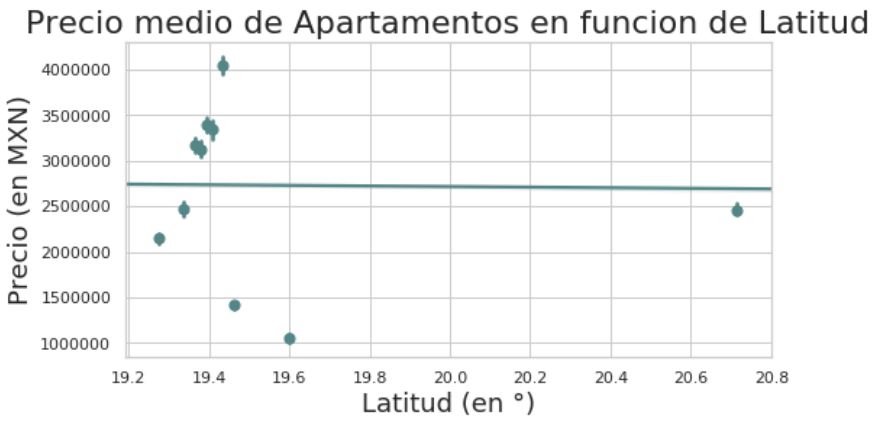


**Figura 16:** Dependencia del precio con la latitud (apartamentos)

Parece existir una correlación negativa entre la latitud de las propiedades y el precio de las mismas, a pesar de que esta correlación no se ve reflejada si solo se analizan los apartamentos. Esta tendencia se ve marcada cuando se analiza el total de las propiedades dado que la mayoría de estas son casas (como se vio anteriormente en los gráficos de distribución).

### 5.1.5 *Según ubicación geográfica (latitud y longitud)*

Siguiendo en la misma linea del apartado anterior, ahora intentaremos ver de forma gráfica cuales son las ubicaciones más caras en México: Esperamos ver algún pico fuerte en Distrito Federal, ya que como se ha observado anteriormente, resulta ser la provincia más cara de todo México.

Para este gráfico hemos decidido utilizar un **HeatMap**:

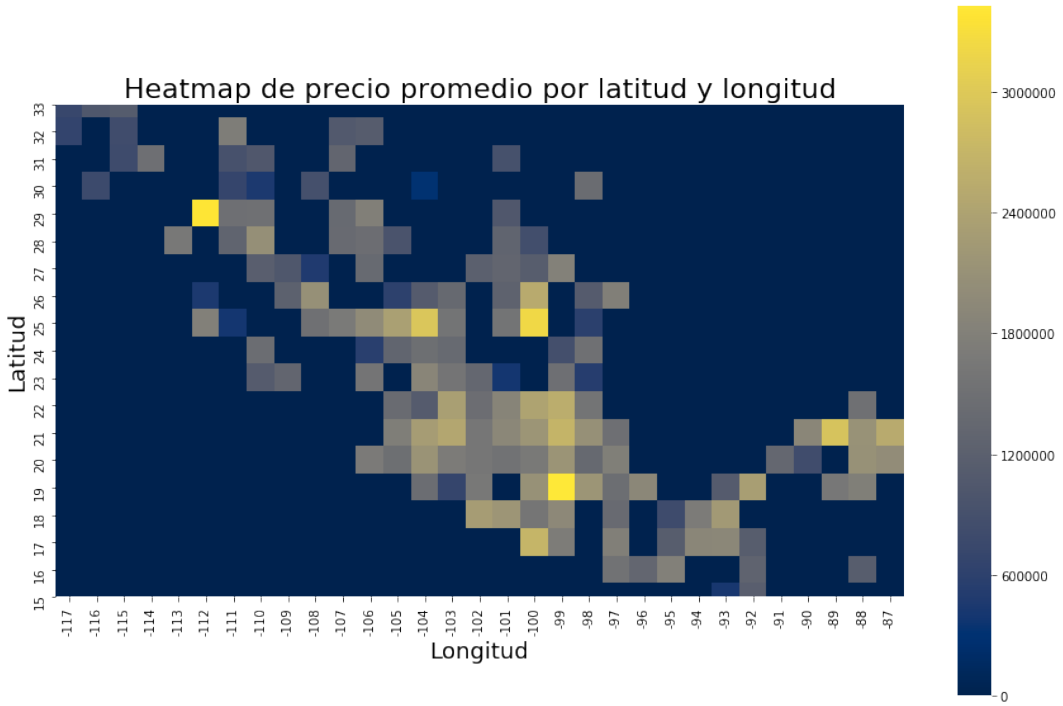


**Figura 17:** Distribución de precios por latitud y longitud

Como vemos, se ve el pico en la zona de **Distrito Federal** (Latitud: 19, Longitud: -99) y en los alrededores. Tiene sentido, dado que se trata de la provincia con mayor movimiento de todo México.

## 5.2 Distribución según escuelas y centros comerciales cercanos

### 5.2.1 *Por escuelas cercanas*

Habíamos visto que el 44 % de las propiedades cuentan con una escuela cercana. Ahora veamos si existe alguna relación entre el precio y la cercanía a un colegio:

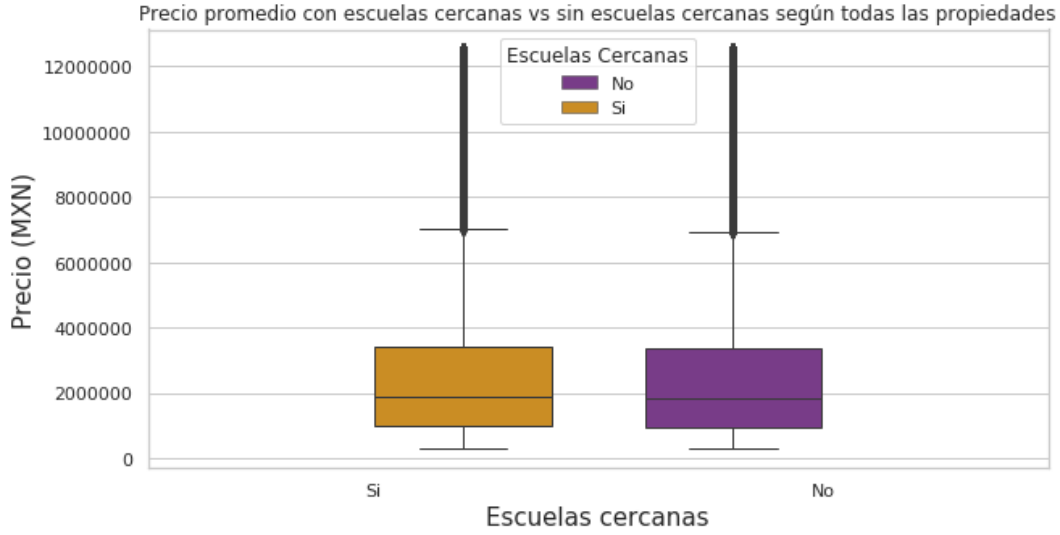


**Figura 18:** Precio promedio de propiedades según cercanía a escuelas

El precio promedio no parece variar demasiado, de hecho casi no hay diferencia alguna. Veamos entonces si se resalta alguna diferencia interesante, filtrando el tipo de propiedad por casa, apartamento, casa en condominio y duplex, lugares donde pueden llegar a residir niños:

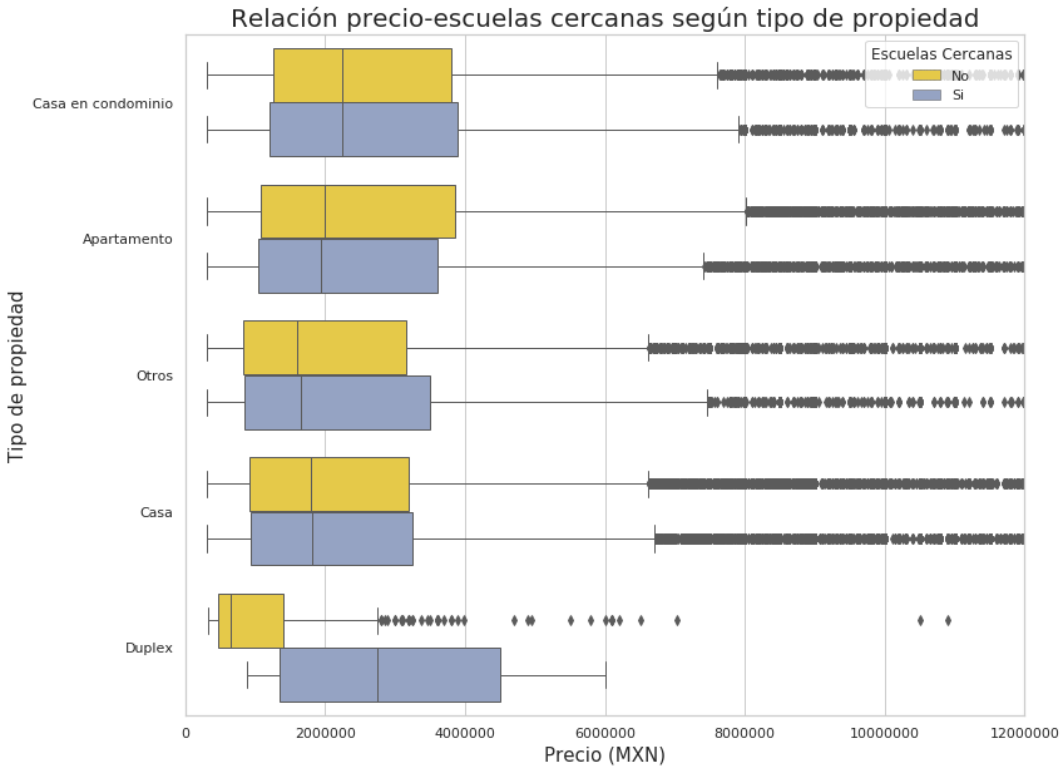


**Figura 19:** Distribución de precio por escuelas cercanas según lugares típicos de residencia

Observando el boxplot, la distribución se da bastante similar en las casas y apartamentos. Sin embargo, el tipo de propiedad 'Duplex'llama bastante la atención: el precio se eleva bastante para aquellos duplex que se encuentran en cercanía con algún colegio. Esto puede deberse a que los duplex cuentan con mayor espacio cubierto que un apartamento, por lo tanto son más propensos a ser habitados por familias numerosas que llevan a los niños al colegio.

### 5.2.2 *Por centros comerciales cercanos*

En esta sección analizaremos la relación del precio con la cercanía a algún centro comercial.

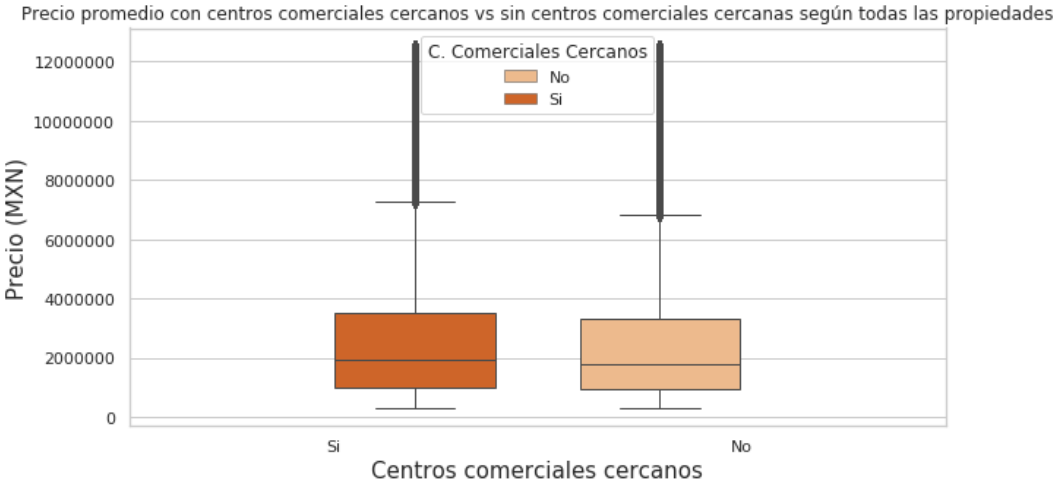


**Figura 20:** Precio promedio de propiedades según cercanía a centros comerciales

Previamente se había observado 4 de cada 10 propiedades tienen un shopping cercano, y esto podía significar una influencia directa sobre el precio. Para sorpresa, no parece haber diferencia alguna de precio cuando comparamos contra todos los tipos de propiedades. Sin embargo, podemos aproximarnos un poco más: ¿Qué pasará si tomamos en cuenta ciertos tipos de propiedad específicos?

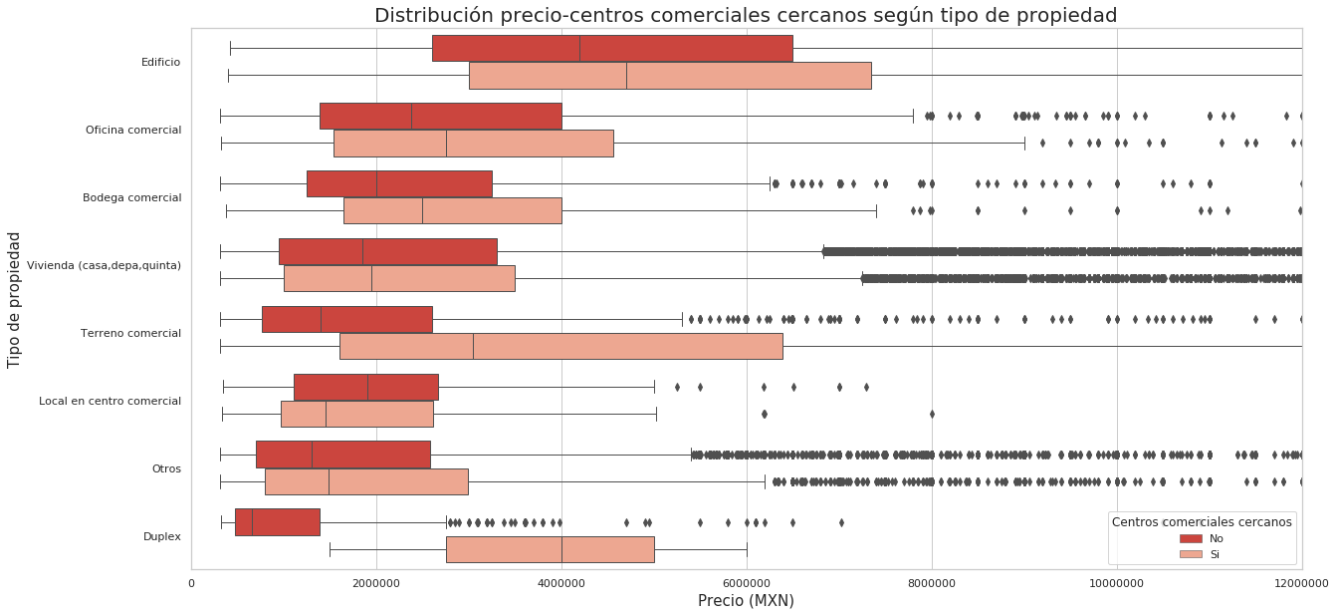


**Figura 21:** Distribución de precio por cercanía a centro comercial según tipo de propiedad

Aquí pueden observarse ciertos resultados interesantes: para viviendas clásicas (ya sea casa, apartamento, quinta) no hay una diferencia manifestada importante. En cambio para propiedades de tipo 'comercial', sí que se puede ver un resalte mayor entre la diferencia de precio. Si observamos los valores para **Terreno comercial** tiene mucho sentido que un sea más caro estando en cercanía con un centro comercial que si no lo estuviera y es ahí en donde impacta directamente sobre el precio.

*Nótese además que hay propiedades de tipo* ***Local en centro comercial*** *que dan negativo para cercanía; los consideramos absurdos, ya que se esperaría que todos ellos den positivo*

## 5.3 Distribución según antiguedad

Se podría llegar a pensar que cuánto más nuevo un inmueble, más caro es. Pero, es realmente así? El siguiente **Heatmap** distribuye el precio promedio de cada tipo de propiedad teniendo en cuenta su antigüedad:

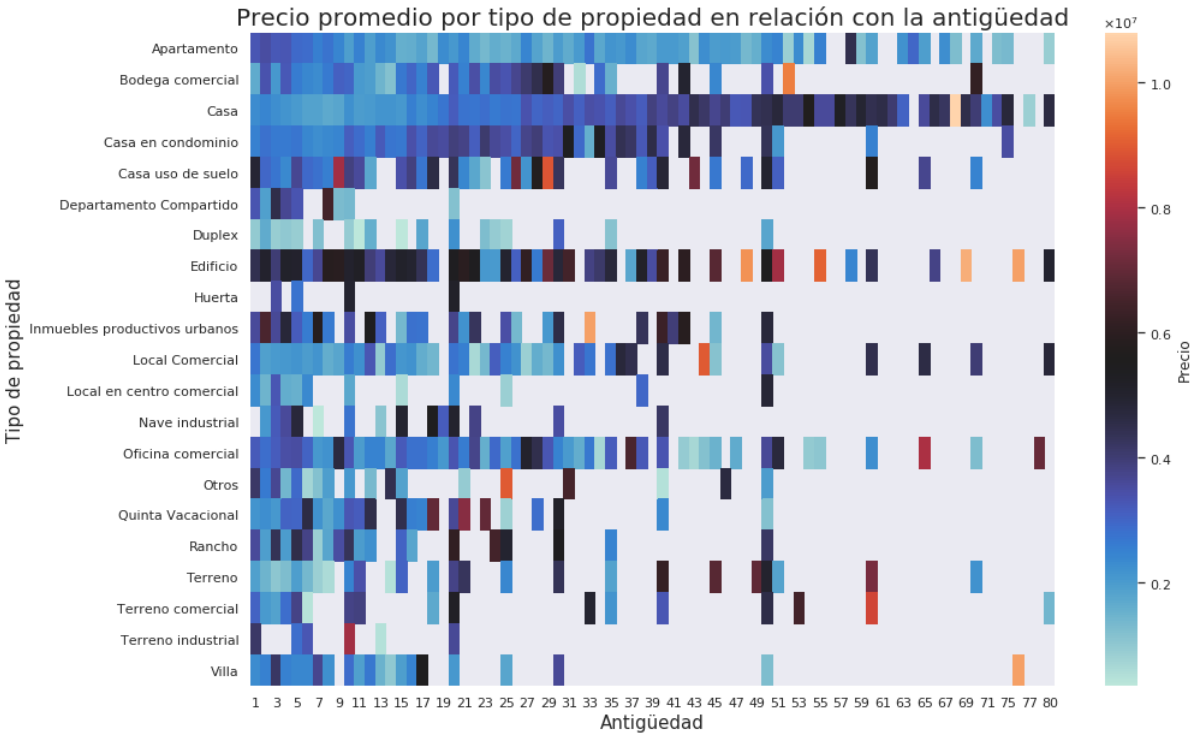


**Figura 22:** Precio promedio por tipo de propiedad según antigüedad. Fondo gris representa valores nulos

Se puede observar la falta de valores para ciertas celdas; sin embargo hay una cantidad significativa de datos para analizar por casas y apartamentos. A continuación se analizarán dichos casos.

### 5.3.1 *Por Casas*

.

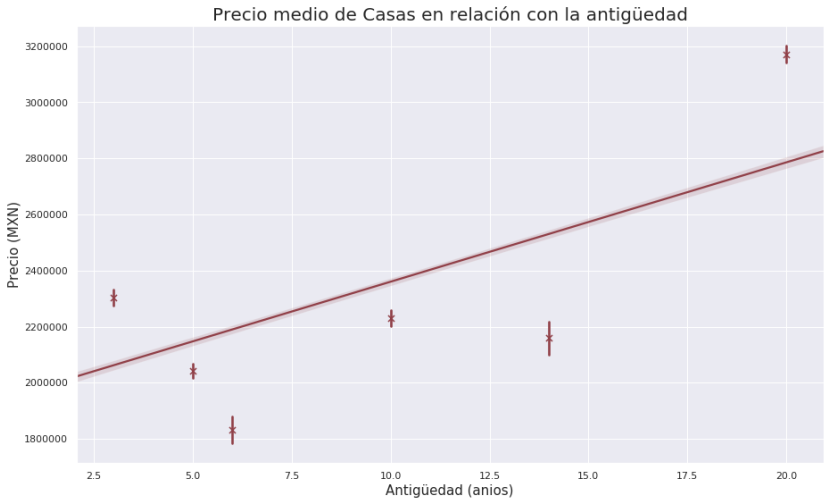


**Figura 23:** Precio promedio de casas en relación con la antigüedad

Como se observa en este gráfico y en el anterior Heatmap, el precio promedio de las casas tiende a ascender al pasar de los años. Podemos interpretar que las casas solían construirse en terrenos más grandes y con otros tipos de materiales, generalmente más costosos. Además, terrenos más grandes significan casas más grandes, en comparación con las casas modernas, que suelen ser más pequeñas y con una cantidad más escasa de ambientes. Si bien esto no llega a justificar con precisión estos resultados, sería algo a tener en cuenta.

Observemos el siguiente **Scatterplot** :

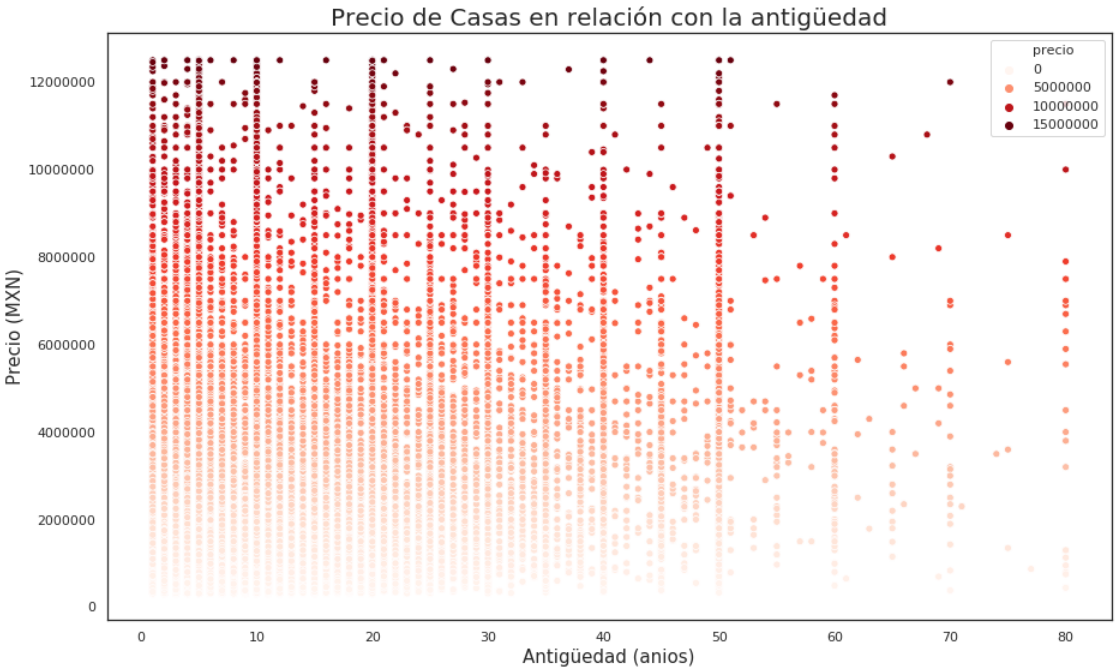


**Figura 24:** Distribución de precio de las casas según antiguedad

Aquí se resalta el precio directo de cada casa, según su antigüedad. A medida que aumenta la antigüedad, van bajando la cantidad de casas menos costosas.

### 5.3.2 *Por Apartamento*

.

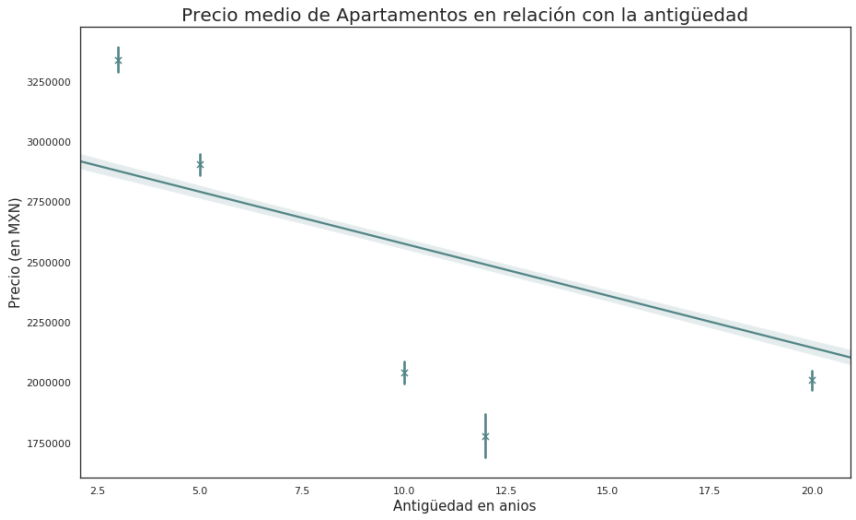


**Figura 25:** Precio promedio de apartamentos en relación con la antigüedad

Analizando el **Regplot**, se observa que a **mayor modernidad, más costoso** será el apartamento. Esto llega a tener sentido, ya que los apartamentos modernos tienden a ser más costosos por la disponibilidad que ofrecen de distintos amenities. Chequearemos esto más adelante.

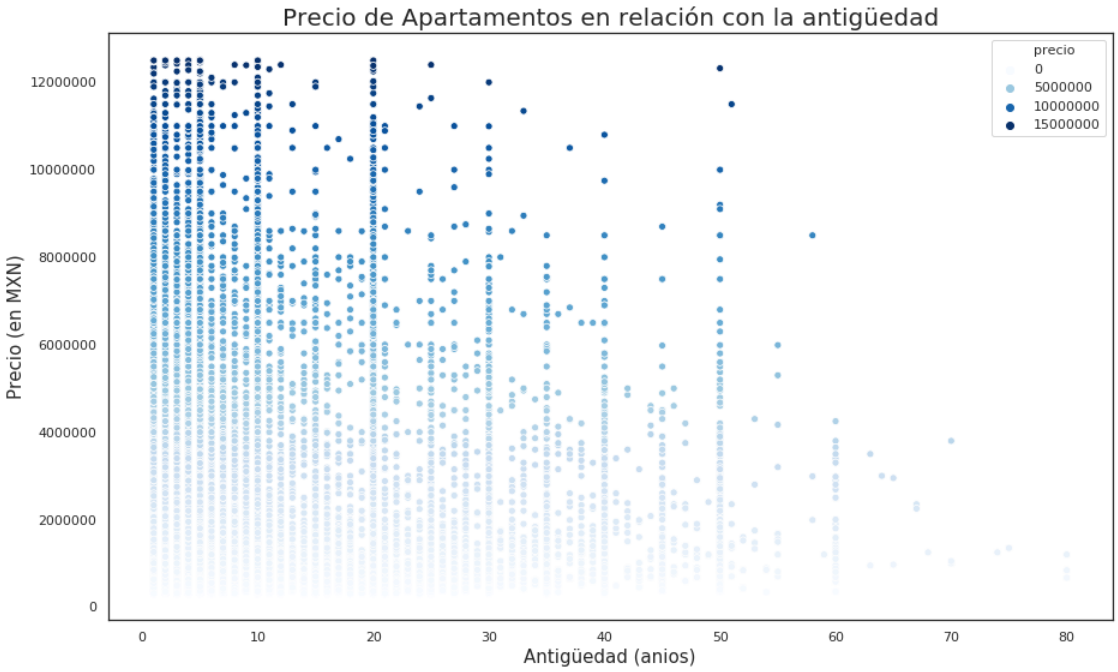


**Figura 26:** Distribución de precio de los apartamentos según antiguedad

Aquí observamos claramente la tendencia de que a mayor antigüedad del apartamento, más valor pierde la propiedad; aquel salto se puede percibir especialmente entre los valores 0 a 10 años y valores de 10 a 20 años.

## 5.4 Distribución según otros parámetros

### 5.4.1 *Por cantidad de habitaciones*

Claramente, esperamos que aquellas propiedades que tengan mayor cantidad de **habitaciones** sean más caras: esperamos ver una relación claramente lineal. De todas formas, es válido verificarlo.

Como se trata de una variable discreta, y también nos interesa mostrar los valores separando Casas de Apartamentos, para este apartado hemos decidido utilizar un gráfico de barras (**BarPlot**):

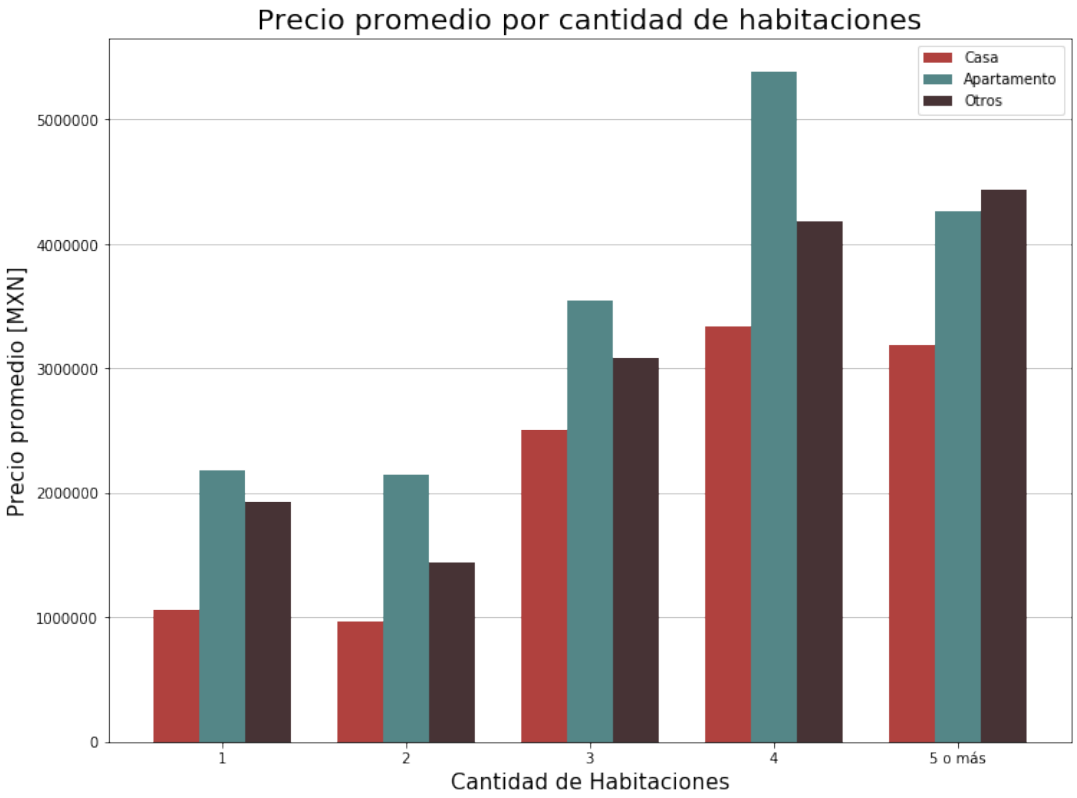


**Figura 27:** Distribución de precios por cantidad de habitaciones

Se puede ver, como se esperaba, una relación lineal. De todas formas, observamos algunos factores que pueden ser interesantes:

- El precio promedio es prácticamente el mismo cuando se tiene una o dos habitaciones.
- Hay un gran salto entre tener dos habitaciones a tener tres. Esto puede deberse a que los apartamentos de tres habitaciones suelen ser los más demandados por las familias tipo.
- El pico se alcanza para propiedades con cuatro habitaciones. Luego, se ve que el precio promedio se estabiliza.
- Si analizamos los *Apartamentos* por separado, vemos que su evolución parece ser exponencial entre aquellas propiedades que tienen dos habitaciones y las que tienen cuatro: vemos que un apartamento con dos habitaciones tiene un precio que ronda los 2.000.000 MXN, mientras que uno que tiene 4 habitaciones tiene un precio que ronda los 5.500.000 MXN, es decir que prácticamente se triplíca el valor. Esto seguramente se deba a que los apartamentos de más de 3 habitaciones son escasos y más modernos.

Como este gráfico es muy general, nos resulta de interés ver como evoluciona el precio según la cantidad de habitaciones en zonas de alta demanda y precio, como por ejemplo, **Distrito Federal**. Uno esperaría que al ser una zona muy densa y cara, el salto entre las propiedades que ofrecen pocas habitaciones a aquellas que ofrezcan más, sea mucho más marcado.

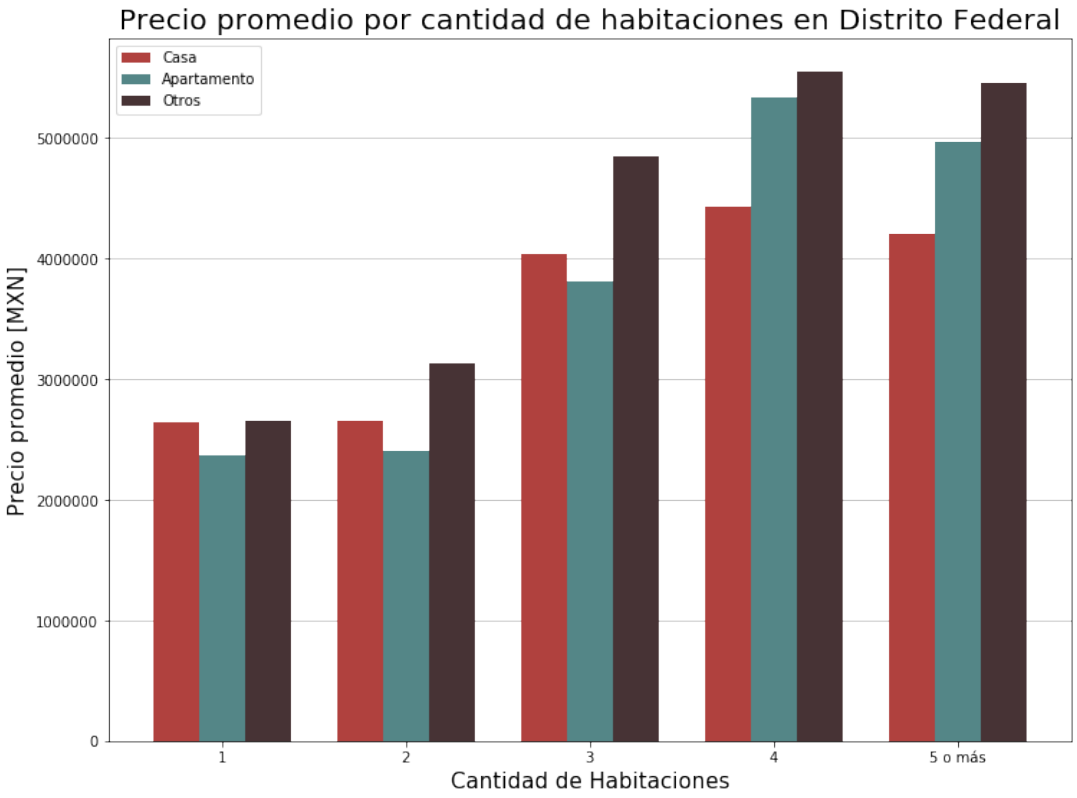


**Figura 28:** Distribución de precios por cantidad de habitaciones (Distrito Federal)

Contrario a lo esperado, cuando se analiza **Distrito Federal** en particular, los resultados parecen tender a estabilizarse entre los distintos tipos de propiedad. También se ve que, nuevamente, el precio de los *monoambientes* es prácticamente el mismo que el de las propiedades que cuentan con dos habitaciones.

### 5.4.2 *Por cantidad de metros cuadrados*

En cuanto a la **distribución del precio por metros cuadrados totales**, es normal esperar que la relación sea **totalmente lineal**: mientras más metros cuadrados se tienen, más cara es una propiedad. De todas formas vamos a verificar que esto se cumpla y analizar la magnitud de la pendiente de la recta que mejor ajuste.

Para representar esta relación, se eligió utilizar dos ajustes lineales en base a puntos (**RegPlot**): uno para los metros totales, y otro para los metros cubiertos. Analizaremos luego los resultados de ambos gráficos obtenidos.

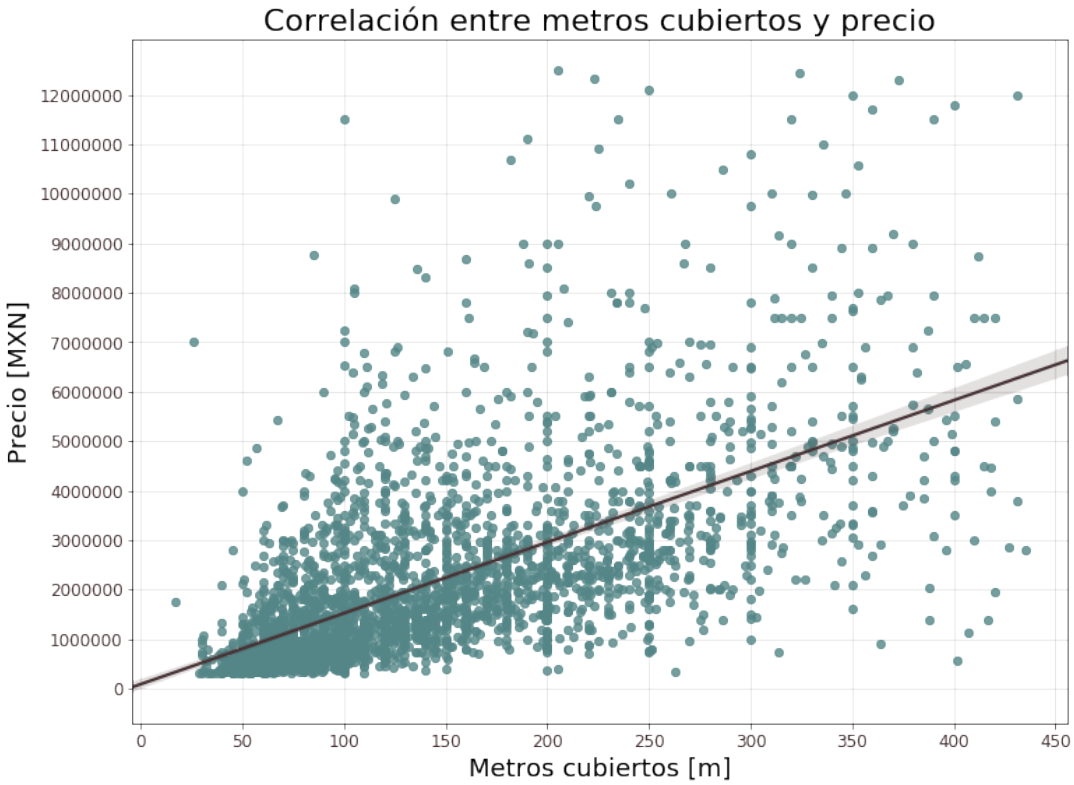


**Figura 29:** Dependencia del precio con los metros cuadrados cubiertos

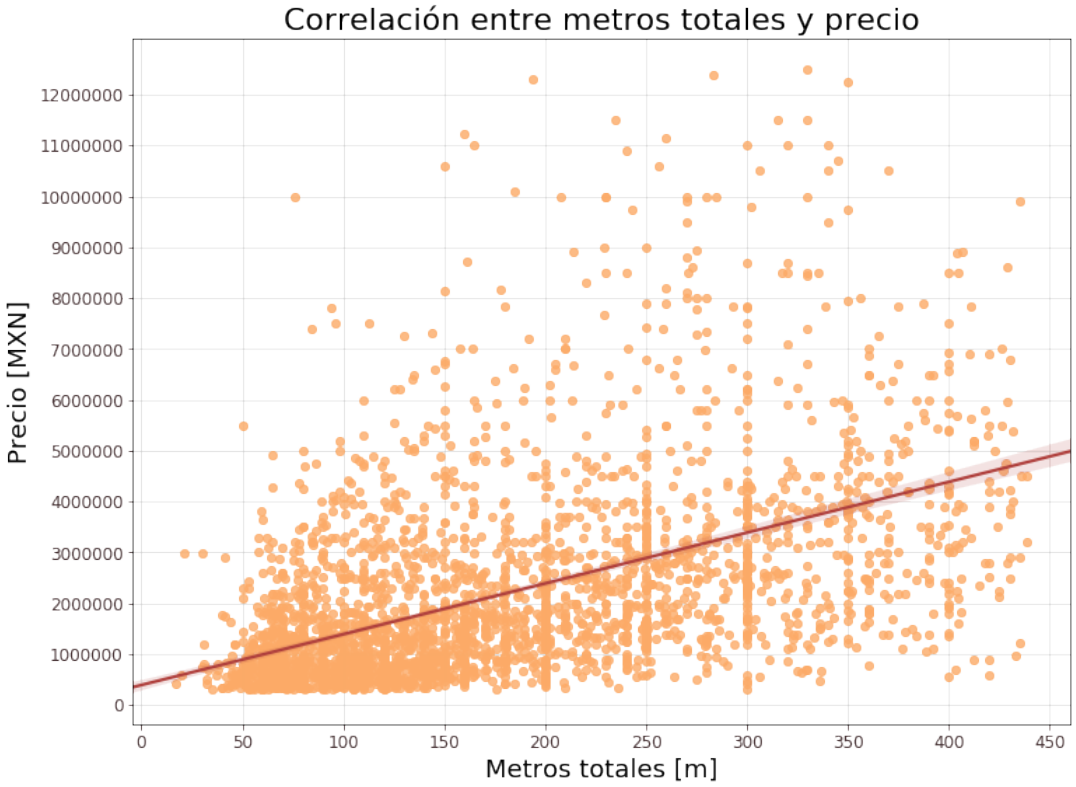


**Figura 30:** Dependencia del precio con los metros cuadrados totales

Claramente vemos que se cumple la relación lineal que esperábamos. Sin embargo, resulta de interés analizar la pendiente de la recta que mejor ajusta en ambos casos:

- En el caso de los **metros totales**, vemos una recta que crece con una pendiente aproximada de 10.000 MXN/m (podemos verificar este resultado luego analizando el *precio promedio* del metro cuadrado).
- En el caso de los **metros cubiertos**, se ve una pendiente aproximada de 15.000 MXN/m, es decir, un 150 % de la pendiente que encontramos para los metros totales.

Se puede concluir entonces, lo que desde un principio era esperado: la relación es lineal, y a medida que aumentan los metros cubiertos, el precio aumenta más rápidamente en comparación con el aumento producido por la suba de los metros totales.

## 5.5 Precio promedio del metro cuadrado

Habiendo realizado ya un profundo análisis de los factores que influyen de manera directa sobre el precio, tales como la ubicación geográfica, la cantidad de metros, la cantidad de habitaciones, etc., se propone ahora realizar un análisis **intensivo** específico sobre el **precio promedio del metro cuadrado**.

El objetivo es ver de forma clara como varía el mismo, calculado de manera general y promediada, según la provincia. A partir de los gráficos buscaremos hallar conclusiones interesantes

Se eligió para representar esta interrogante, un **lollipop plot**, una especie de gráfico híbrido entre un **scatter plot** y un **bar plot**:

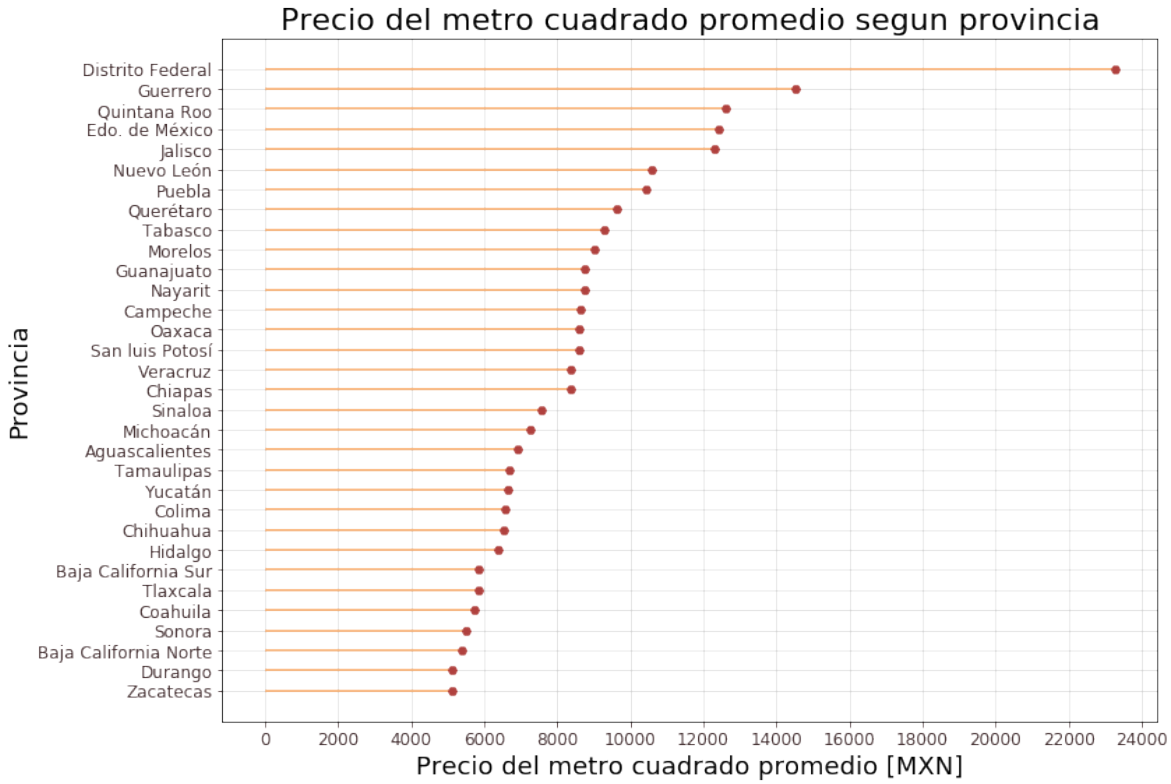


**Figura 31:** Precio promedio del metro cuadrado por provincia

Se ve claramente que **Distrito Federal** es, por mucho, la provincia más cara en todo México: vemos que el precio promedio de un metro cuadrado cuesta aproximádamente 23.000 MXN.

### 5.5.1 *Diferencias respecto del precio promedio del metro cuadrado*

Resulta entonces muy interesante plantear la siguiente interrogante: ¿Qué tanto más cara es cada provincia respecto del **precio promedio del metro cuadrado** teniendo en cuenta todo México?

Utilizaremos un **bar plot** para mostrar los resultados:

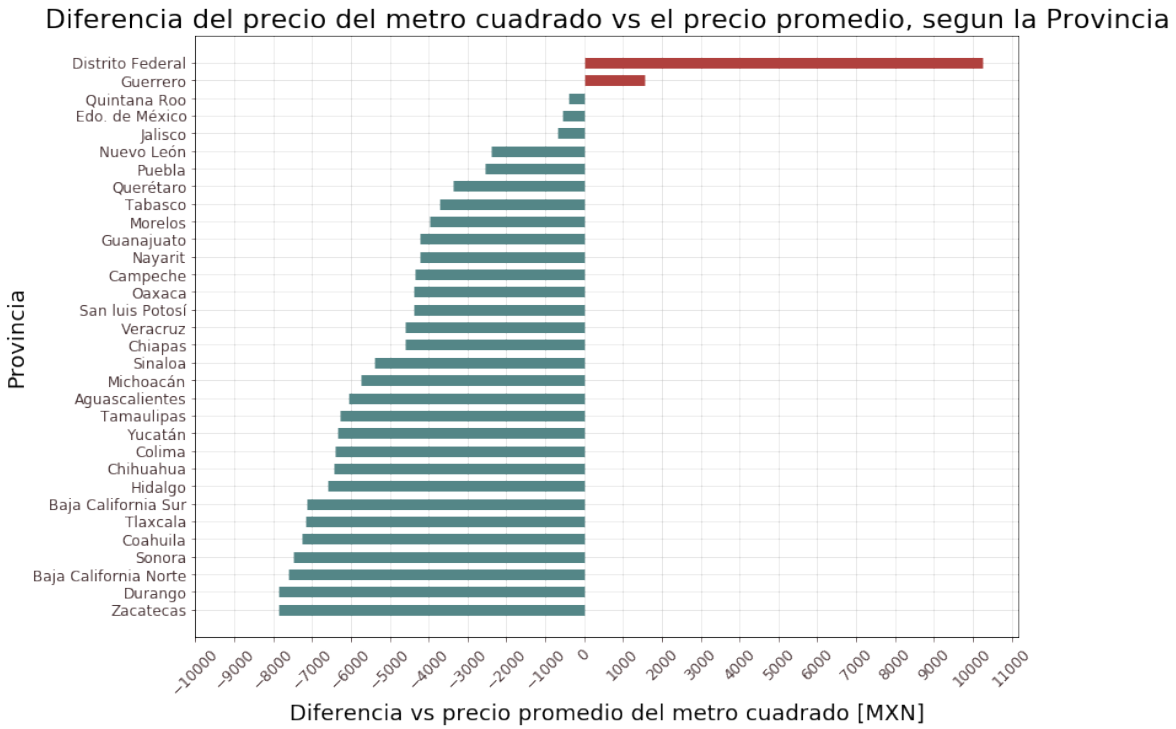


**Figura 32:** Diferencias respecto del precio promedio del metro cuadrado para cada provincia

Podemos observar que la diferencia entre **Distrito Federal** y el resto de las provincias es abismal. De hecho, **Ditrito Federal** y, en una medida considerablemente menor, **Guerrero**, son las únicas provincias que tienen un precio más alto al promedio en cuanto a metros cuadrados. Todas las demás, caen por debajo. Resulta impactante ver estos resultados gráficamente: hay una diferencia de 18.000 MXN por metro cuadrado entre la provincia más barata y la más cara.

## 5.6 Precio promedio según amenities

Es un hecho que las amenities como gimnasio, piscina o salón de usos múltiples agregan al valor de la propiedad. En el siguiente gráfico se ilustran los precios promedios para las propiedades que contienen dichas amenities, y, de todas las propiedades que contienen alguna de estas cosas, que porcentaje de ellas tiene cada amenity.

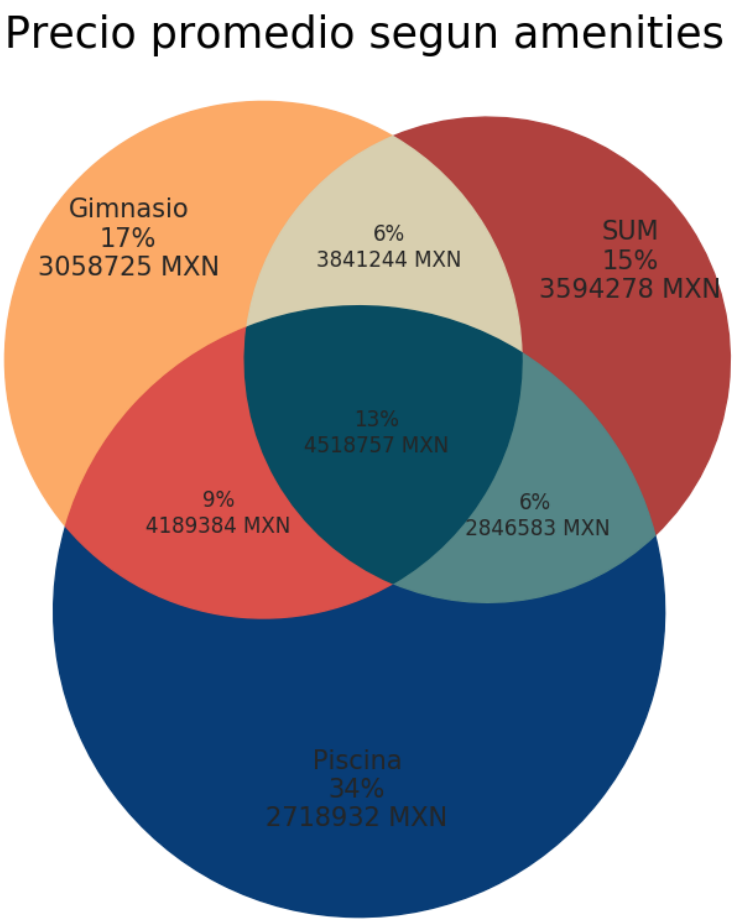


**Figura 33:** Precio promedio según amenities

Podemos apreciar que de todas las amenities, la piscina es la mas barata, y que existe una marcada tendencia a incrementar el precio mientras mayor cantidad de amenities se tengan: Las propiedades que tienen tanto Salón de Usos Múltiples como gimnasio y piscina son aquellas que presentan el mayor precio promedio respecto de las demás categorías hechas en este gráfico.

# 6 ANÁLISIS SEGÚN ANTIGÜEDAD

Otro aspecto que puede resultar interesante es verificar como las características de las propiedades han ido variando con el pasar de los años. En el análisis que sigue se analiza la evolución de varios parámetros de interés como función de la antigüedad. Comencemos por observar la distribución de la antigüedad por cada provincia:

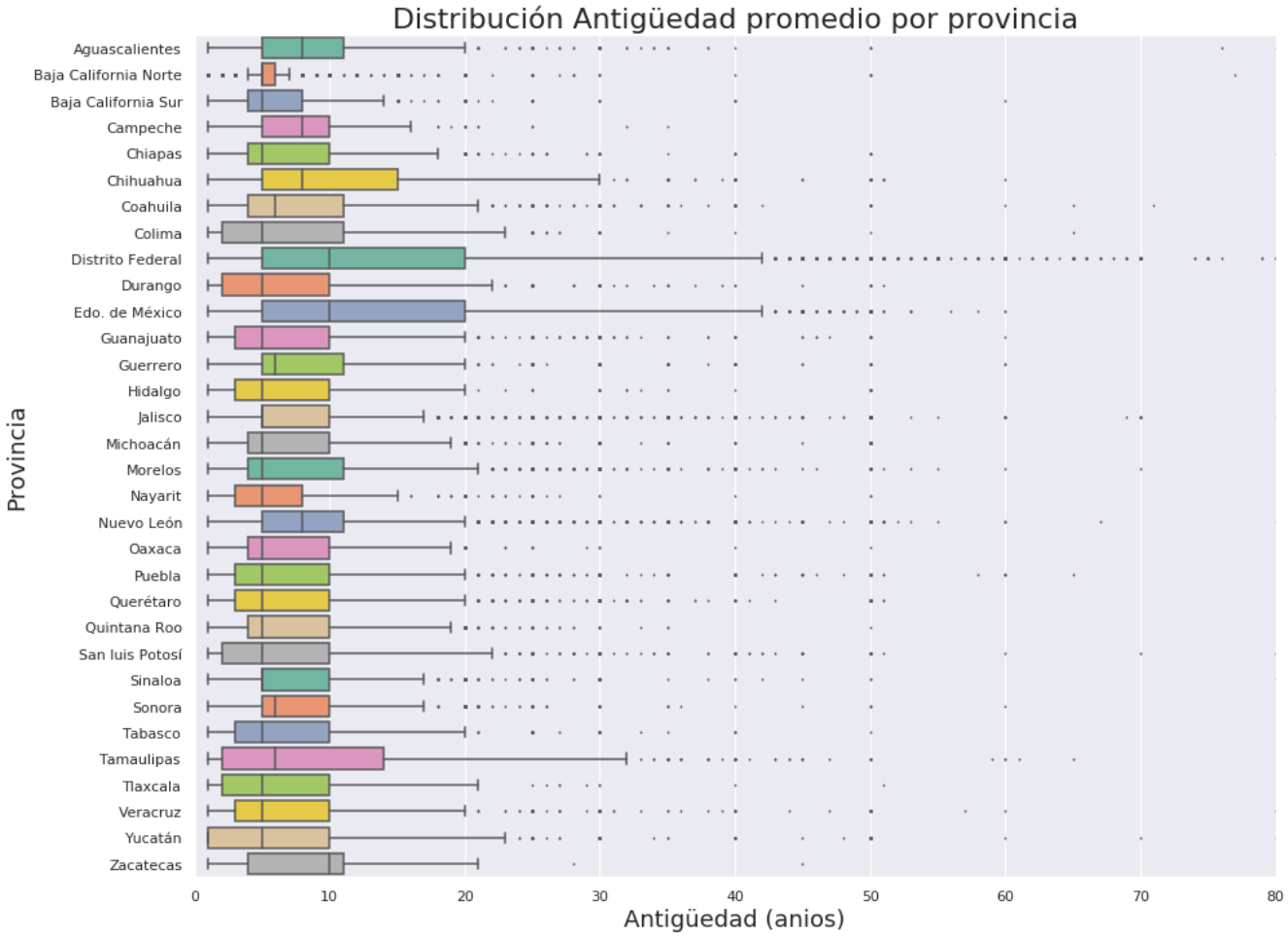


**Figura 34:** Distribución de la antigüedad promedio por Provincia

Como es de esperar, las reconocidas ciudades antiguas como el Distrito Federal y Estado de Mexico son las que presentan un mayor promedio.

## 6.1 Superficie según la antigüedad

Un primer aspecto posiblemente interesante es analizar como varió el tamaño de las propiedades con el correr de los años. Se incluyen a continuación gráficos correspondientes a la correlación entre la superficie cubierta, la superficie total y la antigüedad, siendo el análisis realizado con la totalidad de las publicaciones cuya antigüedad y superficie son valores válidos.

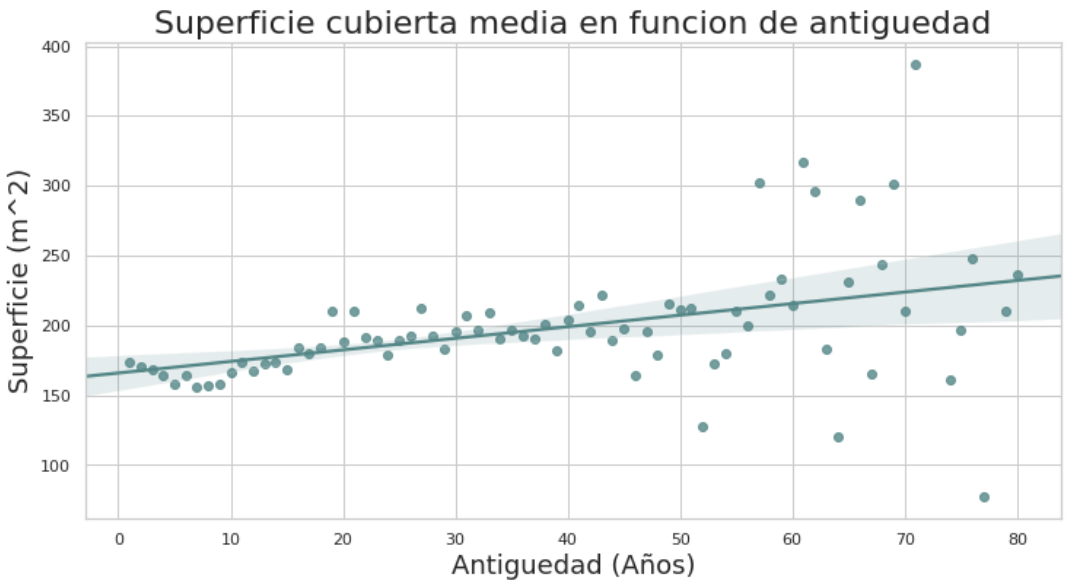


**Figura 35:** Superficie cubierta media según la antigüedad

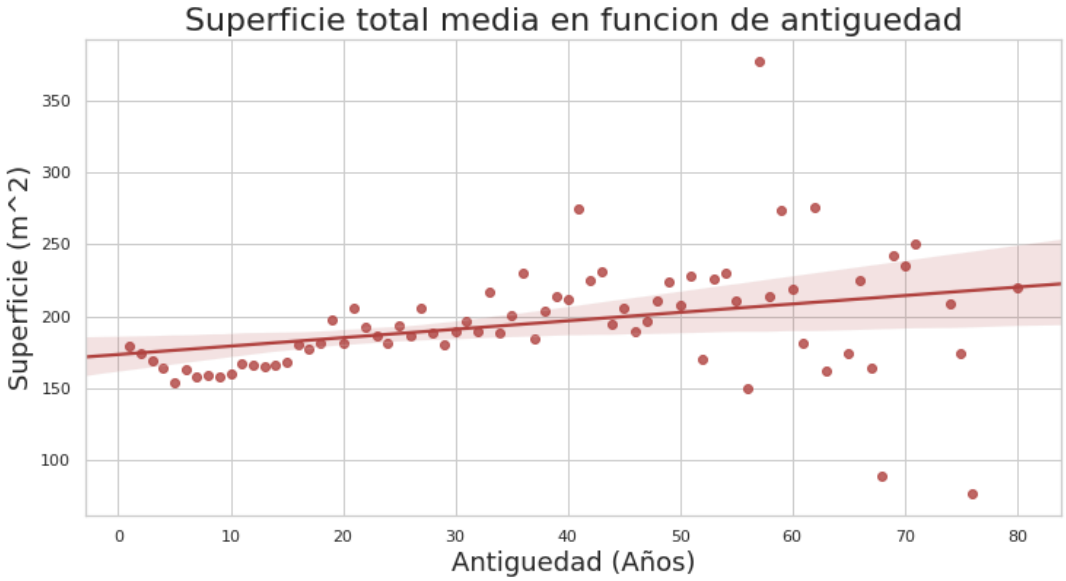


**Figura 36:** Superficie total media según la antigüedad

Como era de esperarse, la superficie tiene una tendencia creciente con la antigüedad, esto es, las propiedades mas antiguas tienen una superficie mayor que las propiedades mas nuevas, aunque los datos de propiedades mas antiguas presentan una mayor dispersión en cuanto a la superficie.

## 6.2 Cantidad de baños según la antigüedad

Quizás sea notorio que las casas mas antiguas tienen una cantidad inusual de baños. El propósito del siguiente análisis es verificar si este set de datos manifiesta dicha correlación. Para este análisis, se discretizó la variable antigüedad en intervalos.

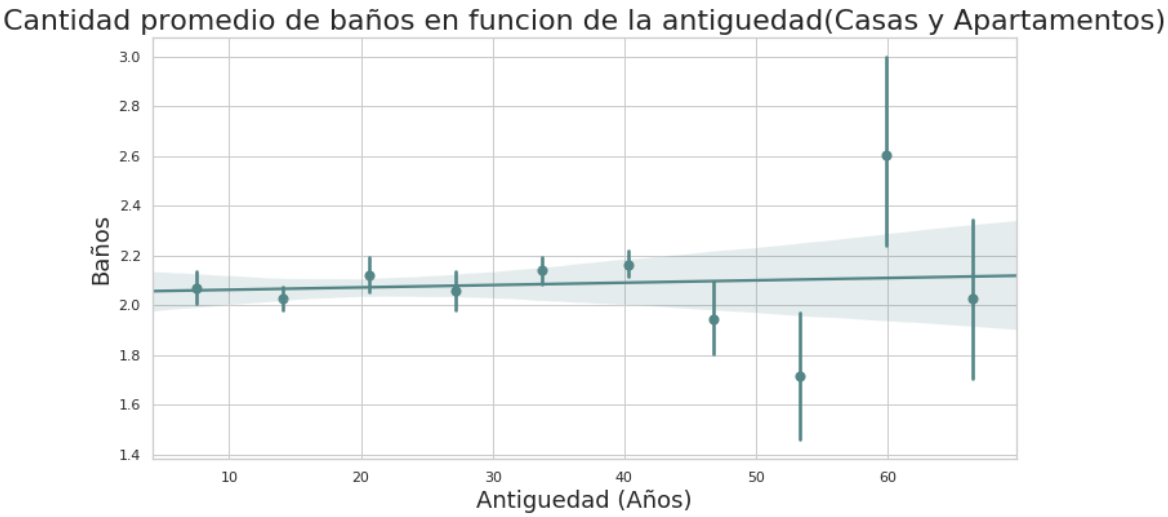


**Figura 37:** Cantidad de baños según la antigüedad

Si bien no existe una tendencia marcada a aumentar la cantidad de baños con la antigüedad, podemos observar que el desvío estándar de los datos según la antigüedad aumenta, se vuelve cada vez más grande. Esto puede deberse a la diversidad esperada.

## 6.3 Cantidad de garajes según la antigüedad

De forma análoga al análisis anterior, se busca una correlación entre la cantidad de garajes y la antigüedad de la propiedad. En este análisis, no se discretizó la variable antigüedad en intervalos, con el objetivo de ver si los resultados hallados se manifestaban de forma diferente.

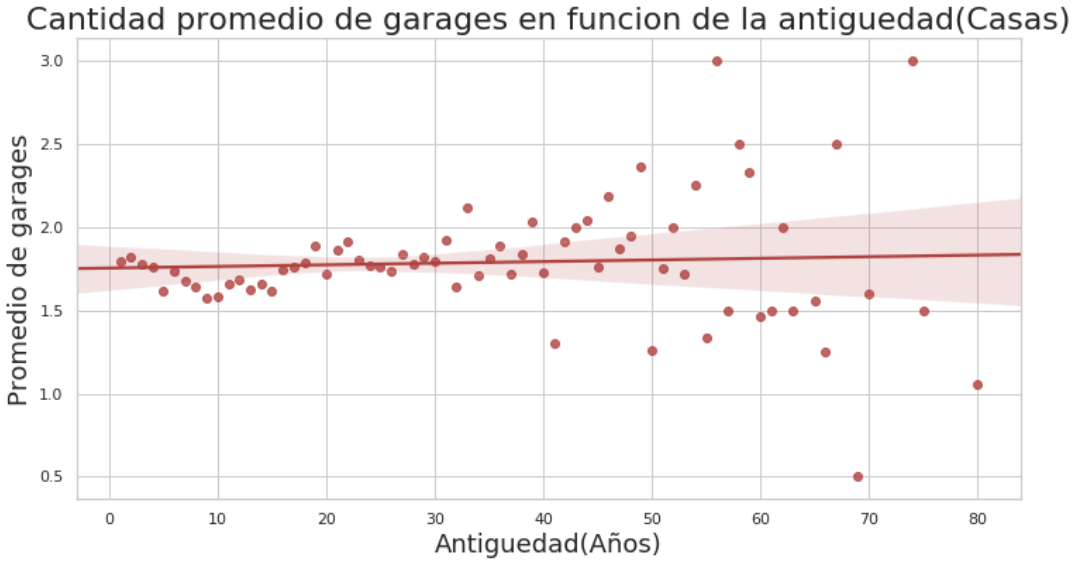


**Figura 38:** Cantidad de garajes según la antigüedad

Al igual que en el caso de los baños, los datos para las propiedades mas antiguas están mas dispersos, pero no hay una tendencia marcada respecto de la antigüedad.

# 7 EVOLUCIÓN DE ZONAPROP

## 7.1 Evolución temporal

Se aprecia en el siguiente gráfico la evolución de cantidad de publicaciones en ZonaProp, discriminadas por tipo.

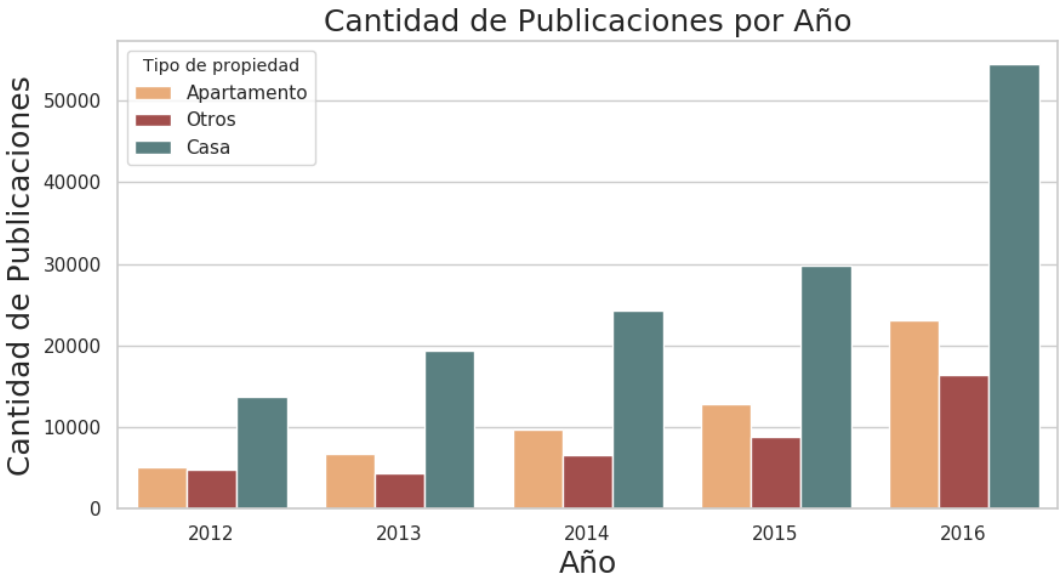


**Figura 39:** Publicaciones de cada tipo según el año

Como es de esperarse, la mayor cantidad de publicaciones para todos los años corresponde a Casas, y la cantidad de publicaciones total ha sido creciente.

### 7.1.1 *Fechas de mayor actividad*

Realizando un análisis mas fino, pretendemos ver en que épocas del año se concentra la mayor cantidad de publicaciones. Con estos fines, se ilustra a continuación un heatmap de cantidad de publicaciones según el año y el mes.

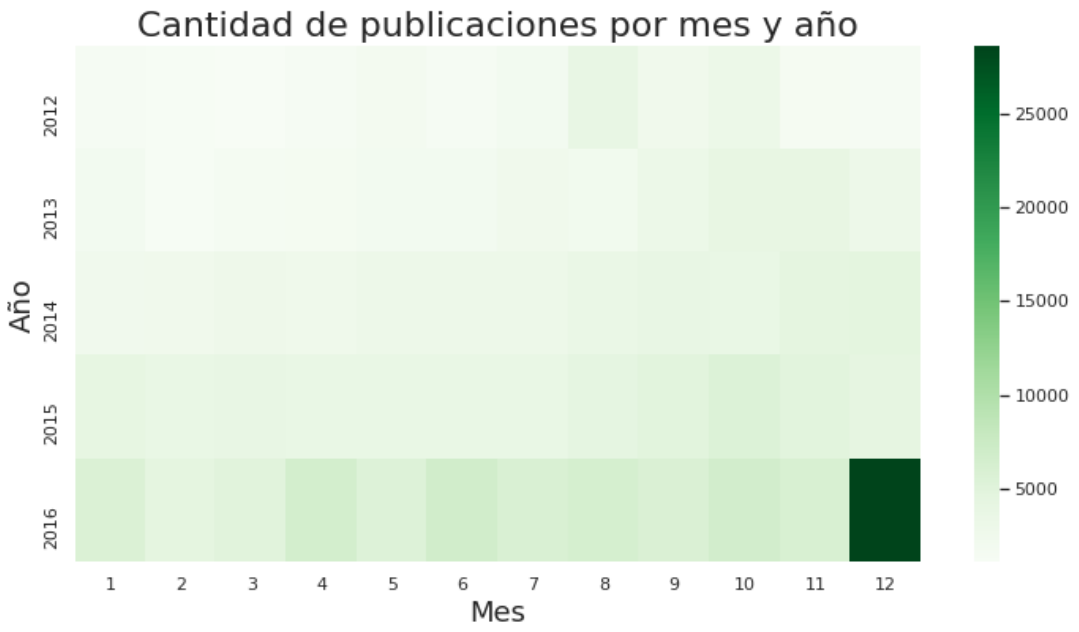


**Figura 40:** Cantidad de publicaciones según año y mes

Por alguna razón (quizás debido a factores mas propios del dataset que de la distribución real de publicaciones), parece ser que la mayor cantidad de publicaciones se concentran en Diciembre de 2016. A fin de analizar la distribución del resto de los años, se ilustra a continuación un heatmap considerando todos los años hasta 2015.

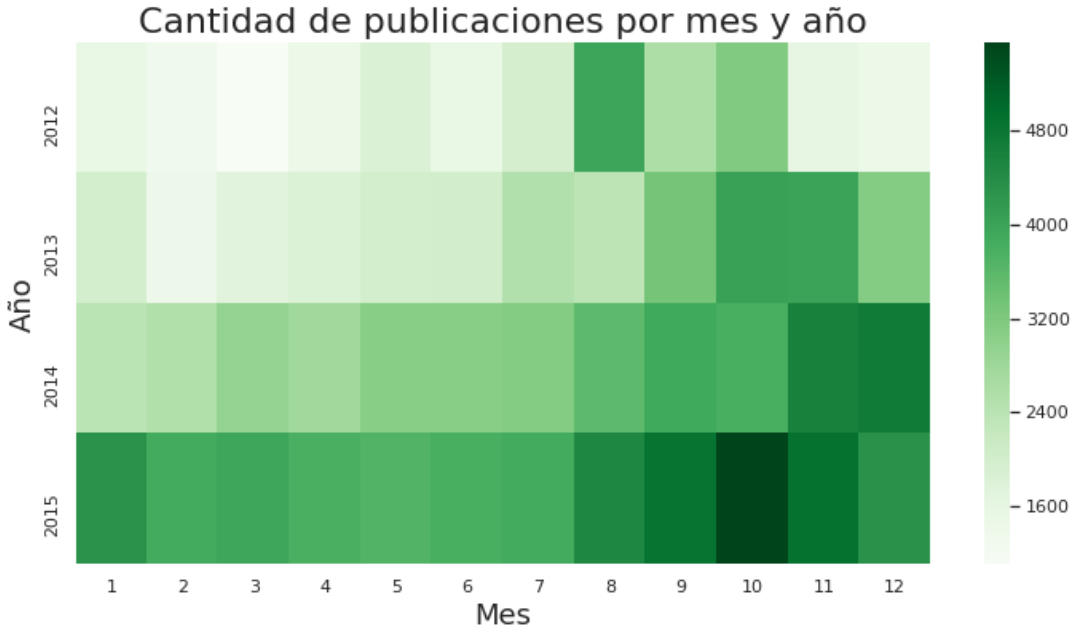


**Figura 41:** Cantidad de publicaciones según año y mes, sin contar el 2016

Se puede apreciar en este heatmap que existe una tendencia a aumentar en la cantidad de publicaciones hacia el último tercio de cada año, y también el ya mencionado incremento de cantidad de publicaciones en los años mas recientes.

# 8 ASPECTOS SECUNDARIOS DE INTERÉS

## 8.1 Devaluación de la moneda

Sería interesante poder utilizar los datos que tenemos sobre las propiedades puestas a la venta entre 2012 y 2016 para extraer alguna especie de información sobre la **economía mexicana**.

Para esto, proponemos analizar la evolución del precio del metro cuadrado promedio a lo largo de estos años: si la curva resultado fuese constante, esto indicaría que los precios se mantuvieron a lo largo del tiempo y que la moneda mexicana no se devaluó frente al *dólar estadounidense**. Por el contrario, si la curva resulta creciente, esto nos dará una idea de los índices de inflación mexicana y de la devaluación de la moneda.

Para este gráfico, discretizamos las variables según Mes y Año, y luego utilizamos un simple **line plot** para mostrar la evolución:

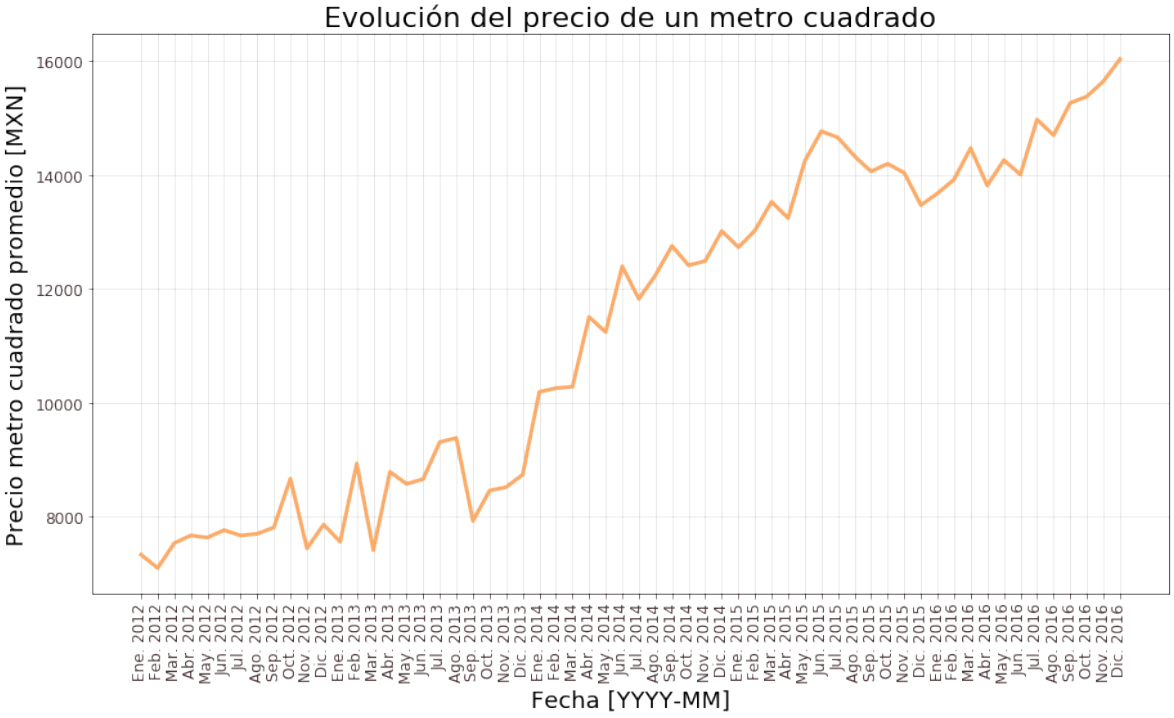


**Figura 42:** Evolución del precio promedio del metro cuadrado a lo largo del tiempo

Vemos que, más allá de los picos de subidas y bajadas, se ve una clara tendencia: **el precio promedio del metro cuadrado duplicó su valor en cuatro años**. Es importante aclarar, nuevamente, que todos estos resultados están atados a los datos que nos proporciona **ZonaProp**, y que debido a negligencia en la carga de datos puede que los mismos no sean fieles a la realidad. Sin embargo consideramos a los mismos como una buena aproximación.

* *Analizamos los resultados en función del dólar estadounidense ya que la misma se suele tomar como moneda de referencia debido a la ínfima inflación que presenta Estados Unidos. Además, en el caso particular de México, toma mayor importancia debido a la cercanía geográfica que presentan.*

## 8.2 Nivel de Ingresos según Entidad Federativa (Provincia)

Consideramos importante tener alguna visión aproximada del **nivel de poder adquisitivo** de las personas que residen en México. Es por esto que para esta sección, decidimos trabajar paralelamente con un dataset* proporcionado por el **Gobierno de México** a través de su amplia página web.

El dataset cuenta con **133.855 filas y 6 columnas**. Contiene datos desde el año 2005 al 2019, sin embargo solo fueron tomados los datos desde el año 2012 al 2016 (Que corresponden al período de publicaciones del dataset proporcionado por Navent). El dataset fue debidamente depurado y filtrado para descartar datos no deseados.

Cada fila cuenta entonces con la siguiente información:

- **Entidad_Federativa**: Provincia
- **Nivel_Ingresos**: El nivel de ingresos medido en salarios. Pueden ser:
  - *No especificado*
  - *No recibe ingresos*
  - *Menos de 1 salario minimo*
  - *1 salario minimo*
  - *Mas de 1 hasta 2 salarios minimos*
  - *Mas de 2 hasta 3 salarios minimos*
  - *Mas de 3 hasta 5 salarios minimos*
  - *Mas de 5 hasta 10 salarios minimos*
  - *Mas de 10 salarios minimo*
- **Numero_personas**: Cantidad de personas que cobran poseen dicho nivel de ingreso.

Teniendo en cuenta la **distribución de precios por provincia** de la figura 8, queremos ver si existe algún tipo de relación de precios de las propiedades entre el precio promedio de la provincia y el nivel de ingresos de la población en dicho lugar geográfico.

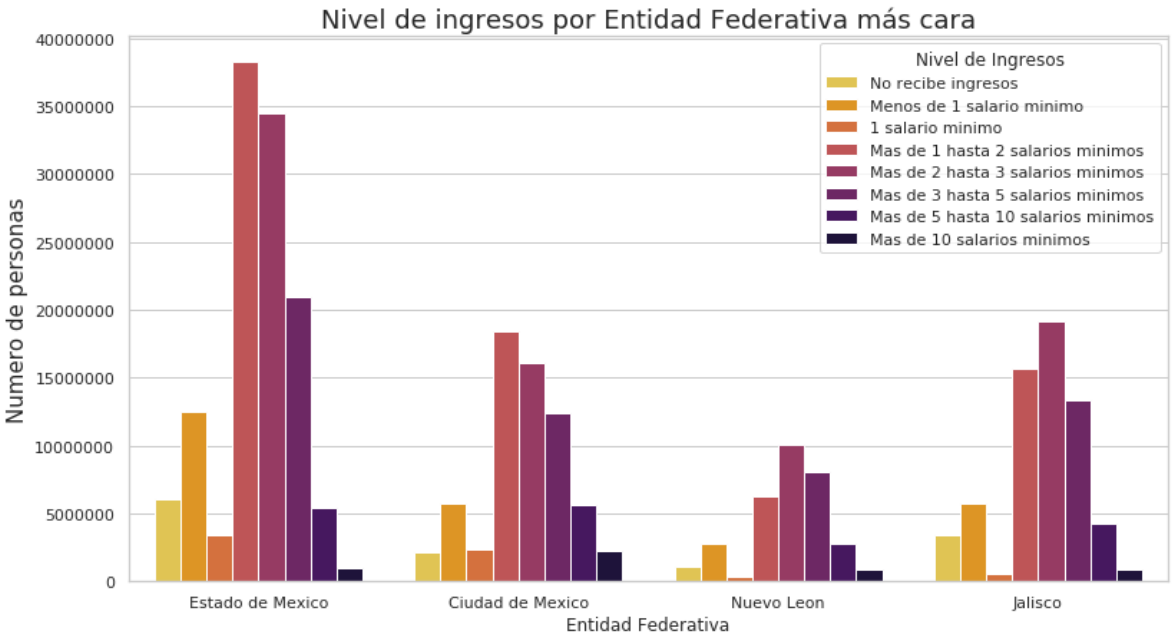


**Figura 43:** Nivel de ingresos de la población con respecto a entidades federativas (provincias) costosas para vivir (2012 - 1016)

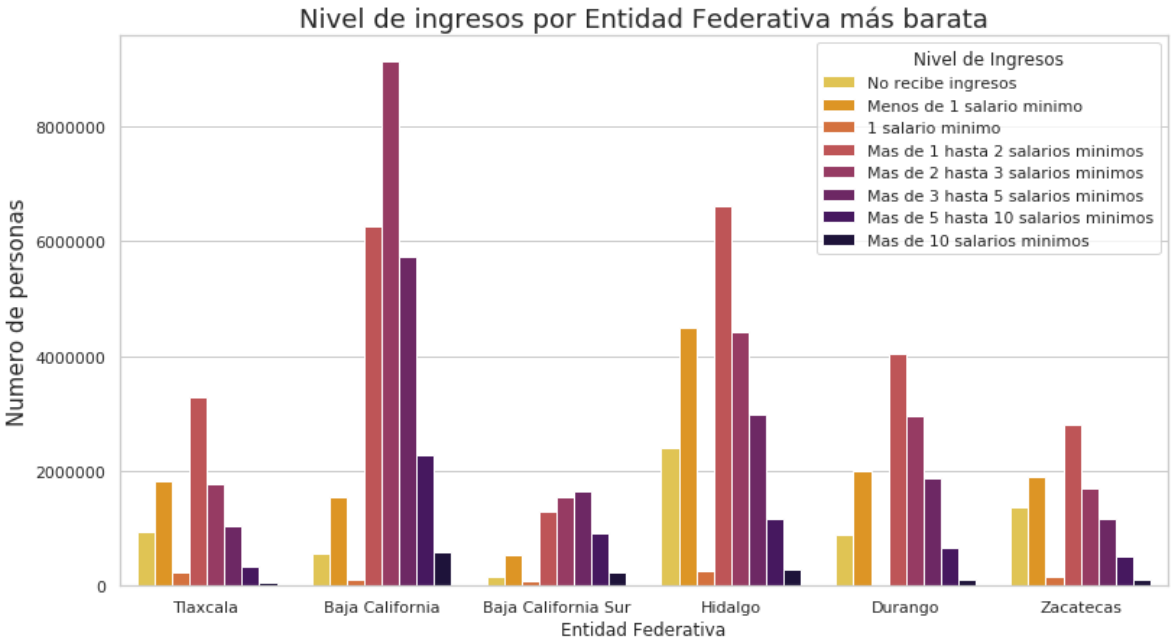


**Figura 44:** Nivel de ingresos de la población con respecto a entidades federativas (provincias) más económicas para vivir (2012 - 1016)

- En las entidades federativas más costosas se obtiene un valor esperable: Son muchos más los individuos que cobran por encima de 2 hasta 5 salarios mínimos que menos de 1 salario mínimo.
- Baja California Norte es una de las provincias más económicas de México, sin embargo hay una diferencia abismal entre personas que ganan mas de 3 hasta 5 salarios mínimos y personas que cobran 1 salario mínimo ó menos.
- En el Estado de Mexico, a pesar de ser uno de los más desarrollados, hay una importante cantidad de personas que cobran menos de un salario minimo.
- El nivel de ingresos por entidad federativa más económica, la cantidad de personas que ganan menos de 1 salario mínimo es prácticamente constante, con un pico elevado en Hidalgo y un pico más bajo en Baja California Sur.
- En comparación con ambos gráficos, hay mucha una cantidad significativa de personas que cobran más de 5 salarios mínimos en las provincias más caras que en las provincias más baratas.
- Hay más cantidad de personas en la figura 43 que en la figura 44, lo cual se puede interpretar que en las provincias más caras, hay un mayor índice de población.
- El salto de ganar un salario mínimo y más de uno, es mucho más violento en las provincias más caras, mientras que, en las provincias más baratas es más suave, a excepción quizás de Baja California.
- En las provincias más caras, el pico se alcanza para un nivel de ingresos de más de 1 hasta 3 salarios mínimos, ya que ambas barras van prácticamente iguales. Luego, se ve que el nivel promedio se estabiliza. Por otro lado, en las provincias más económicas, se observa un pico pronunciado de más de uno hasta dos salarios mínimos, con la excepción de Baja California, que tiene su pico en más de dos hasta tres salarios mínimos.
- La cantidad de personas que ganan un salario mínimo permanece básicamente constante, sin picos importantes.

# 9 CONCLUSIONES

Para comenzar es importante destacar que, si bien consideramos que el análisis realizado fue bastante profundo, se podrían haber hallado resultados más interesantes si se hubiera tenido información respecto de si las **publicaciones resultaron o no en una venta**, y de ser así, **cuanto tiempo transcurrió entre la realización de la publicación y la venta efectiva**. A partir de estos datos hubiera sido posible plantearse nuevas interrogantes que resulten de interés, como por ejemplo: ¿Qué tipo de propiedad se vendió más cada año? ¿Cuál es la influencia de los atributos tales como gimnasio, pileta, etc. sobre las compras (es decir, que tanto la gente busca dichos atributos en las propiedades que efectivamente compran)? ¿Cuáles son las descripciones de propiedad que atran más potenciales compradores? Y así podríamos seguir, y sin duda obtendríamos algunas conclusiones mucho más interesantes.

También es importante aclarar nuevamente, como ya se hizo anteriormente en el desarrollo del informe, que estas conclusiones se basan en el análisis de 300.000 publicaciones de ZonaProp que contaban con muchos outliers, y muchos datos poco coherentes (probablemente como consecuencia de la carga negligente de los mismos, o de la falta de validación por ZonaProp a la hora de aceptar una publicación). Se detectaron, por ejemplo:

- Sólo el 30 % de los mismos presentaban información coherente sobre los **metros cubiertos** y sobre los **metros totales**.
- Sólo el 50 % posee información sobre su **latitud y longitud**. Descubrimos también que, de los datos que poseen valores de **latitud y longitud**, muchos de estos no son coherentes con la información que proporcionan en sus campos de **Ciudad** y/o **Provincia**, así como otros de estos geográficamente se encuentran en el océano.
- Si bien desde el principio suponemos que nos fueron proporcionados los datos de ZonaProp entre 2012 y 2016 (existen 60 meses entre los cuales se distribuyen las publicaciones, por lo que de ser uniforme la distribución, deberíamos tener aproximádamente 5000 datos en cada mes), se descubrió que el 10 % de los datos, es decir, 30.000 publicaciones, pertenecen al mes de **Diciembre del 2016**, es decir, el último mes del cuál se tiene información. Esto nos da para pensar si realmente todas esas publicaciones pertenecen efectivamente a dicho mes o si por el contrario tenemos datos para los cuáles no sabemos su fecha de publicación.
- El dataset nos proporciona información sobre si la propiedad se encuentra o no en cercanía a escuelas o centros comerciales, pero no tenemos idea según qué criterio. Hubiese resultado interesante conocer **el radio** para el cuál ZonaProp considera que una propiedad es cercana a un determinado punto, para entonces poder analizar por ejemplo, por medio de intersecciones, la ubicación geográfica de algunas de las escuelas y/o centros comerciales.

## 9.1 Insights

A continuación se presenta un listado de los insights que pueden llegar a resultar interesantes que fueron descubiertos a lo largo del análisis:

- La **popularidad** de la plataforma **ZonaProp** ha ido creciendo casi de manera uniforme a lo largo de de los últimos años, y se observa que en el 2016 se publicaron **el doble de Casas y Apartamentos** que en 2015 (Atención: es posible que estos datos no sean fieles a la realidad por la situación expresada en el apartado anterior).

- Se descubrió una tendencia bien marcada a publicar las propiedades **en el último cuatrimestre del año**.
- Analizando la evolución del precio promedio del metro cuadrado según la fecha, observamos que el mismo ha ido subiendo a un ritmo muy elevado, al punto de llegar a **duplicar su valor en tan solo cinco años**. Suponemos que este aumento está fuertemente relacionado con la devaluación de la moneda mexicana.
- Resulta muy interesante observar la siguiente tendencia inmobiliaria: **la mitad** de las publicaciones poseen **tres habitaciones**. De hecho, si observamos en el precio en función de la cantidad de habitaciones, vemos que hay un salto muy grande entre las propiedades que poseen dos habitaciones y las que poseen tres, así como el salto que existe, en el caso de los *apartamentos*, entre los que poseen tres y los que poseen cuatro.
- **Casi un 90 %** de las propiedades incluyen uno o más **garajes**. Esto nos resultó muy inesperado, aunque analizando esta observación con detenimiento, debemos recordar que la mayoría de los datos son propiedades de tipo **Casa** (más propensas a tener garajes) y de la provincia **Distrito Federal**, zona que presenta mayor densidad poblacional. De todas formas, resulta interesante observar este dato.
- Respecto de la **ubicación geográfica**, se observa una importante tendencia: **el precio disminuye mientras más al norte nos encontremos**. De hecho, se observa que la mitad inferior de México posee un precio promedio mucho más elevado a las zonas de mayor cercanía a los Estados Unidos, relación que en principio creímos que podría ser al revés.
- Un 25 % de las propiedades puestas en venta se encuentran **a estrenar**, y más de la mitad (un 60 %) tienen **menos de 10 años de antigüedad**. Esto nos puede marcar dos posibles causas): el mercado inmobiliario en México es **muy moderno**, y/o aquellas personas poseedoras de propiedades antiguas no tienen intereses en venderlas.
- Es probable que la distribución del dataset no refleje la distribución real de las propiedades en venta en México (por ejemplo, que la mayoría de las propiedades en venta son casas y no apartamentos). Esto puede ser indicador de que **los sitios web para compra-venta de propiedades suelen ser utilizados más tanto por compradores como por poseedores de casas y apartamentos**, en contraste con otros tipos de propiedades, como por ejemplo locales comerciales, oficinas, etc., y por ende la información que se le hace disponible a quienes navegan por dicho sitio resulta ser de interés para un grande pero acotado segmento de la población. Nuestra **propuesta** aquí es lanzar un **programa especial para promocionar la venta de propiedades especiales**, como por ejemplo, la de oficinas, locales comerciales, terrenos, etc. Si se agranda el enfoque, se conseguiría **mayor cantidad de transacciones** y por lo tanto **mayor cantidad de ganancia**.
- Se observa en el análisis que la mayor cantidad de los datos provienen de **Distrito Federal**, lo cual es razonable, considerando que es allí donde es mayor la densidad poblacional de México. En vista de esto, nuestra propuesta para ZonaProp es la siguiente: debería **incrementarse la publicidad del sitio en localidades que presenten una cantidad de publicaciones comparativamente baja** en relación a su cantidad de habitantes (como podría serlo Ciudad Juárez), con programas especiales, promociones, y beneficios para quienes se incorporen a la plataforma. Por ejemplo, dado que hemos visto que el coste promedio de las propiedades es muy distinto en Distrito Federal respecto del resto de las provincias, una opción podría ser personalizar las comisiones que

se cobran dependiendo de la ubicación de la propiedad, para animar a los sectores más aislados a incorporarse por un costo relativamente bajo.

* *Fuente del dataset: https://datos.gob.mx/busca/dataset/indicadores-estrategicos-poblacion-ocupada-por-nivel-de-ingresos*