UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

Proyecto Final – Módulo I
Análisis del Mercado Inmobiliario en México de 2013 a 2016

PRESENTA

Edgar David Cardoso Olvera
314551885

PROFESOR

Carla Paola Malerva Resendiz

DIPLOMADO

Ciencia de Datos

28 de noviembre de 2020

Índice general

Capítulo 1

Introducción

Una de las necesidad básicas y primordiales del ser humano es la vivienda, es por eso que el sector inmobiliario es tan importante para los países. Este mercado puede afectar factores tanto políticos como económicos, como se vio en la crisis inmobiliaria de 2008 que daño a toda la economía mundial.

En México, siendo uno de los países con mayor población mundial, el mercado inmobiliario es uno de los sectores que mantiene a flote el Producto Interno Bruto. En el primer trimestre de 2020 el mercado de bienes raíces represento un 10 % del PIB en México, contribuyendo con 2.4 billones de pesos según datos de Inegi.

1.1. Objetivo

El objetivo principal de este proyecto es analizar y conocer el comportamiento del mercado inmobiliario mexicano a partir del año 2013 a 2016, esto con ayuda de la base de datos más grande del mundo, Internet. Se analizarán componentes primordiales en anuncios de venta y renta de inmuebles en internet, como lo son: fechas, precios, tipo de inmueble, ubicación, superficie, descripción, entre otras. Todo lo anterior con la finalidad de comprender uno de los mercados más grandes e importantes en México.

1.2. Metodología del Análisis

Al recabar la información del mercado inmobiliario en México se realizará un análisis con respecto a la calidad de los datos, esto con el fin de mejorar la toma de decisiones y facilitar el procesamiento de la información. Lo anterior tienen como objetivo conocer la falta de información en la base de datos principal y comprender que características son importantes para el proyecto. De igual manera se le dará un formato homogéneo a la información para así poder trabajar de una manera más eficiente.

Después se realiza un análisis estadístico para conocer las distribuciones de los datos y su comportamiento, esto es necesario para identificar valores fuera de los rangos comunes dentro de la base. Este mismo análisis ayudará a la creación de visualizaciones, que facilitaran la compresión de la información del mercado inmobiliario.

1.3. Fuente de Datos

El conjunto de datos se obtuvo de la página web https://data.world/. La fuente original de los datos contiene dos archivos de inmuebles publicados en México del año 2013 al 2016 divididos por el tipo de operación: Venta y Renta. Los archivos anteriores fueron unidos para crear una base de datos más completa y funcional con un total de 188,525 registros. Un agradecimiento al usuario @properati por haber proporcionado esta fuente de datos esencial para el análisis inmobiliario en México.

Capítulo 2

Análisis Exploratorio de Datos

El análisis exploratorio de la información tiene como objetivo reconocer patrones significativos en nuestros datos, encontrar irregularidades en la información y comprende de manera rápida y eficaz el conocimiento que estos transmiten. También brindará información relevante para crear visualizaciones para un mejor entendimiento.

2.1. Diccionario de Datos

En la siguiente tabla se da a conocer las características que se tienen de cada publicación dentro de la tabla de datos, esto es fundamental para conocer con que herramientas contamos para el análisis de la información.

Nombre de la Variable	Tipo de Dato	Descripción
created_on	String	Fecha de Publicación
operation	String	Tipo de Operacion: Venta o Renta
property_type	String	Tipo de Propiedad: Casa, Departamento, Tienda, PH
place_name	String	Nombre de Municipio o Colonia
place_with_parent_names	String	Ubicacion con País, Estado, Municipio, Colonia
geonames_id	String	ID Geofráfico
lat-lon	String	Latitud y Longitud Unidas
lat	Float64	Latitud con Decimales
lon	Float64	Longitud con Decimales
price	Float64	Precio de Inmueble
currency	String	Tipo de Moneda en la Publicación
price_aprox_local_currency	Float64	Precio Aproximado en Pesos Mexicanos
price_aprox_usd	Float64	Precio Aproximado en Dólares Americanos
surface_total_in_m2	Float64	Superficie total en m2
surface_covered_in_m2	Float64	Superficie construida en m2
price_usd_per_m2	Float64	Precio en Dólares Americanos por m2
price_per_m2	Float64	Precio por m2
floor	Float64	Piso de Departamento
rooms	Float64	Número de Cuartos
expenses	Float64	Gastos Adicionales
properati_url	String	Dirección Web
description	String	Descripción del Inmueble
title	String	Título del Anuncio
image_thumbnail	String	Dirección Imagen

2.2. Tipos de Variables

- Variables Continuas:
  lat, lon, price, price_aprox_local_currency, price_aprox_usd, surface_total_in_m2, surface_covered_in_m2, price_per_m2, price_usd_per_m2, floor, rooms, expense
- Variables Discretas:
  operation, property_type, currency, estado, municipio, colonia, geonames_id
- Variables de Fechas:
  created_on
- Variables de Texto:
  properati_url, description, title, image_thumbnail

2.3. Completitud

Nombre de la Variable	% De Nulos
created_on	0.00 %
operation	0.00 %
property_type	0.00 %
place_name	0.00 %
place_with_parent_names	0.00 %
geonames_id	100.00 %
lat-lon	16.94 %
lat	16.94 %
lon	16.94 %
price	1.39 %
currency	1.39 %
price_aprox_local_currency	1.39 %
price_aprox_usd	1.39 %
surface_total_in_m2	48.21 %
surface_covered_in_m2	3.95 %
price_usd_per_m2	58.79 %
price_per_m2	11.23 %
floor	84.77 %
rooms	96.78 %
expenses	99.89 %
properati_url	0.00 %
description	0.00 %
title	0.00 %
image_thumbnail	2.93 %

Algunas variables como geonames_id que contiene un sólo dato o la variable expense con 207 datos serán eliminados en los siguientes pasos ya que contienen mas del 90 % de información nula, lo cuál perjudica el análisis de la información.

2.4. Estadística Descriptiva

	c_lat	c_lon	c_price	c_price_aprox_local_currency
count	156134	156134	185454	185454
mean	21.034041540242	-99.0133530053932	2559721.73365363	3183243.06557337
std	2.97352809912203	5.17021181128755	6038333.16636718	7963734.31420877
min	14.843818	-117.228632	50	940.42
10 %	18.917542	-103.438171	13500	14864.16
20 %	19.16150848	-101.00246024	155000	252693.13
30 %	19.367960277	-100.356804	580000	672019.707
40 %	19.454332	-99.6030794	950000	1090307.49
50 %	20.1242815	-99.2277873	1400000	1565706.81
60 %	20.6965424	-99.168088	1880000	2105776.52
70 %	21.039522	-98.979619523	2570554	2873765.76
80 %	22.272658848	-97.837899572	3567479.99999999	4063873.78
90 %	25.643093	-89.636932	5900000	6936675.98
max	41.577487	99.206936	945000000	936451275.4

	c_price_aprox_usd	c_surface_total_in_m2	c_surface_covered_in_m2	c_price_per_m2
count	185454	97630	180635	166908
mean	169244.919603299	448.098811840623	1993.31448501121	16247.7332532109
std	423411.453007493	3017.1541337506	563709.794917386	531832.067114607
min	50	-396	-324	0.008
10 %	790.29	0	46	99.7603322270001
20 %	13435.05	35	70	1813.6720144
30 %	35729.575	90	90	6428.571429
40 %	57968.87	120	120	8333.333333
50 %	83244.64	160	150	9941.0024365
60 %	111958.77	200	190	11410.25641
70 %	152790.8	282	235	13279.209701
80 %	216065.81	400	300	16190.47619
90 %	368805.38	674	431	23500
max	49788727.19	200000	230303030	199500000

Al observar las variables de coordenadas vemos que los valores máximos y mínimos de latitud y longitud difieren en las coordenadas máximas y mínimas del territorio mexicano, al crear los distintos gráficos se analizaran estos casos particulares.

En el caso del precio de propiedades se analizará el valor máximo de esta variable ya que existen propiedades con un precio superior a los 900 millones de pesos mexicanos.

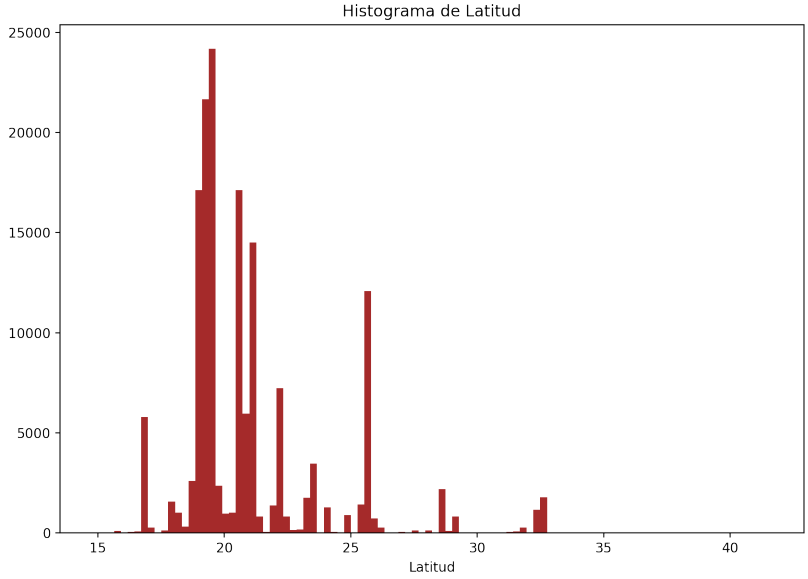
La variable superficie total será analizada detalladamente, ya que en el primer percentil la longitud total es de 0 metros cuadrados.

2.5. Visualización de Datos

Esta sección tiene como objetivo comprender de forma intuitiva y visual la información del mercado inmobiliario en México, brindará la capacidad de reconocer patrones en los datos, encontrar información errónea y comprender cuales son las características principales para este análisis.

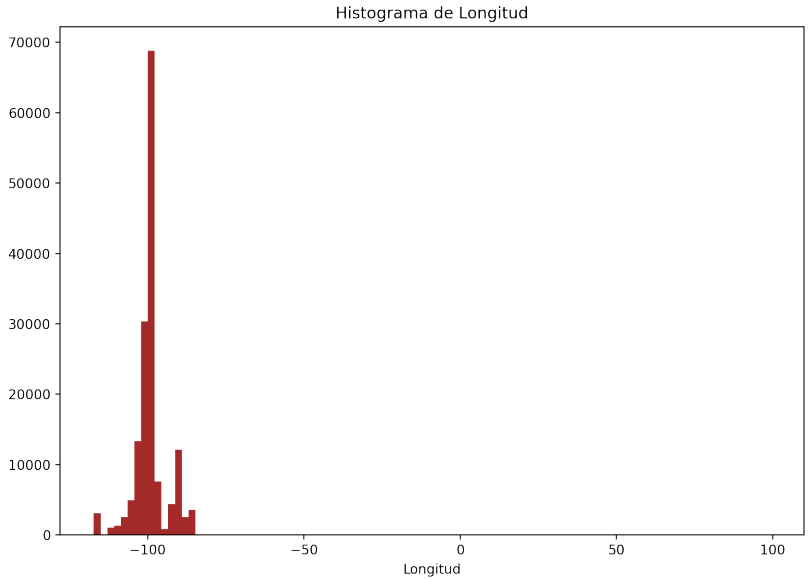
2.5.1. Geográfica

Latitud



En el gráfico anterior se observa la distribución de las latitudes geográficas, se puede notar, gracias al rango de latitudes, que existen coordenadas mayores a 35, estos son valores inusuales ya que no existe una cantidad significativa de valores para dicha latitud.

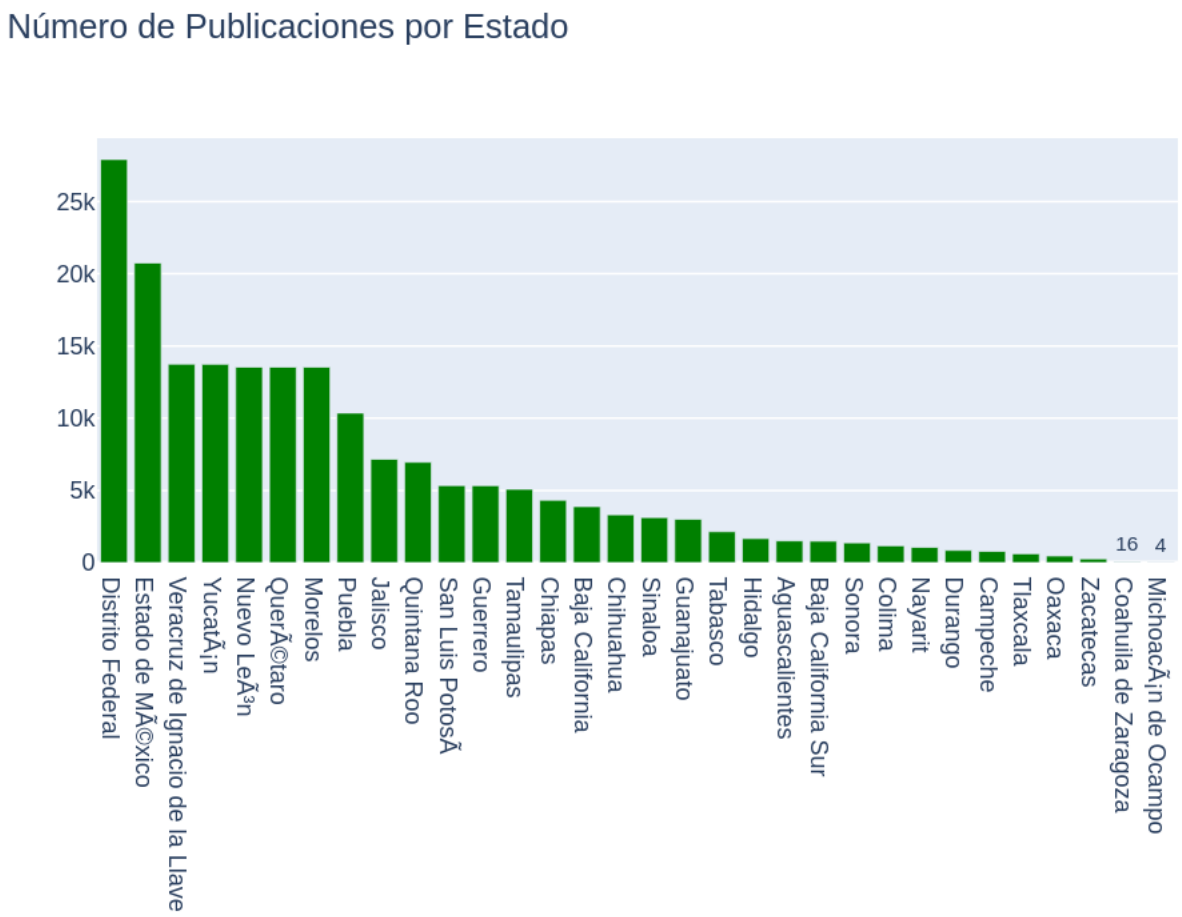
Longitud



De igual manera notamos que el rango de longitudes es muy amplio y hay valores que estan en una longitud mayor a 90, lo que indica la existencia de valores extremos en esta variable.

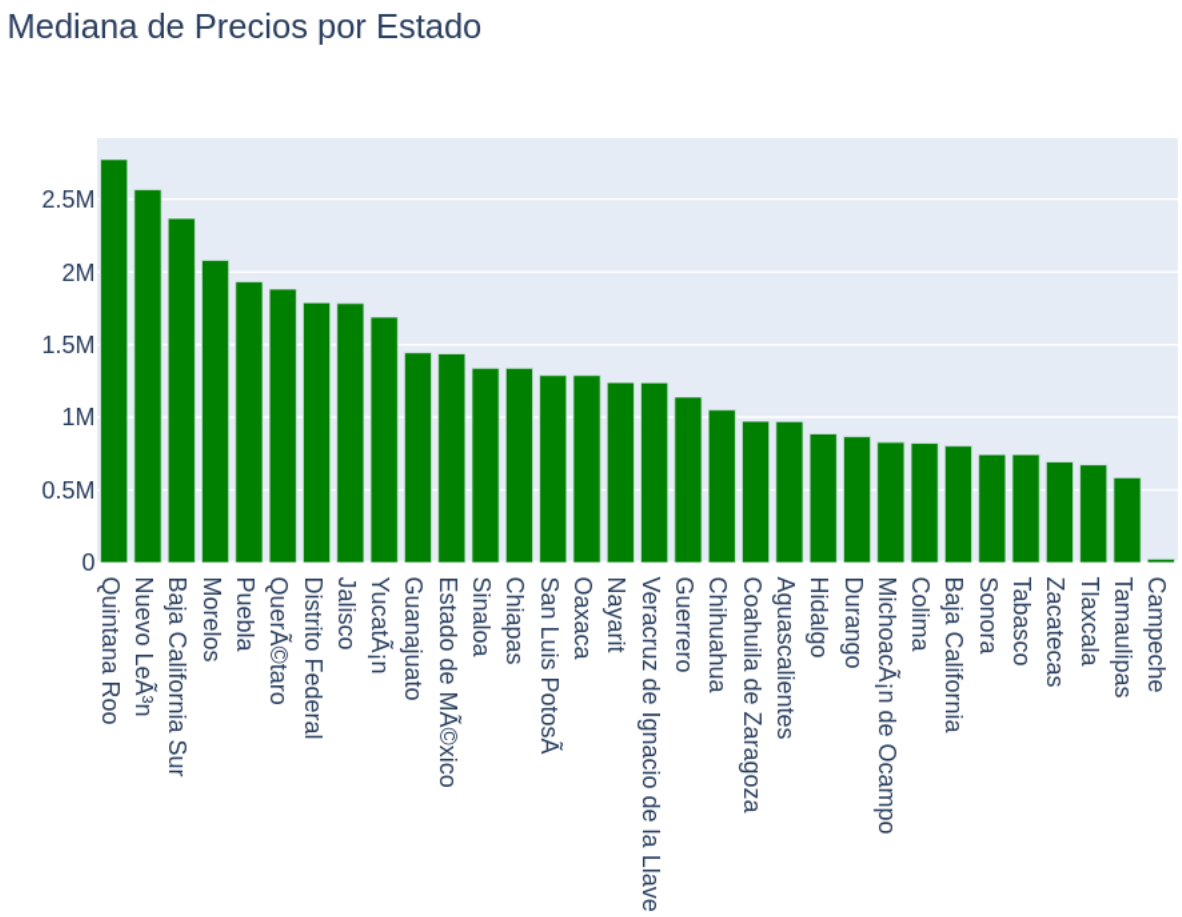
2.5.2. Estados

Publicaciones



Se puede notar que los estados con mayor población en México tambien cuenta con el mayor número de publicaciones de inmuebles, esto puede ser ocasionado por la alta demanada que existen en estos sitios.

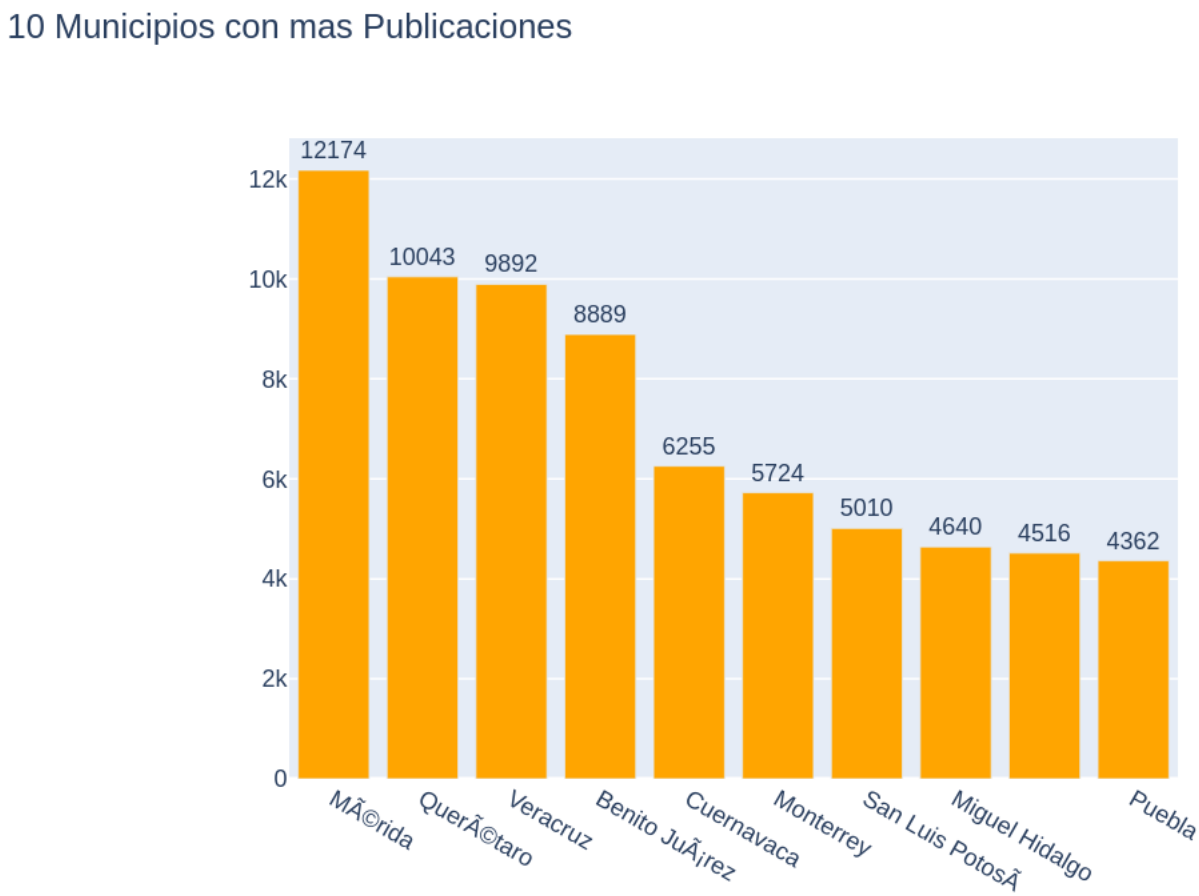
Precios



Observamos que los lugares con una mayor mediana de precios son estados donde su mayor ingreso económico es el turismo. El incremento en precios puede ser debido a una cercanía a sitios vacacionales con un gran flujo de visitantes. De igual manera observamos que estos estados son de clima caliente, lo que puede indicar que las personas buscan climas más cálidos.

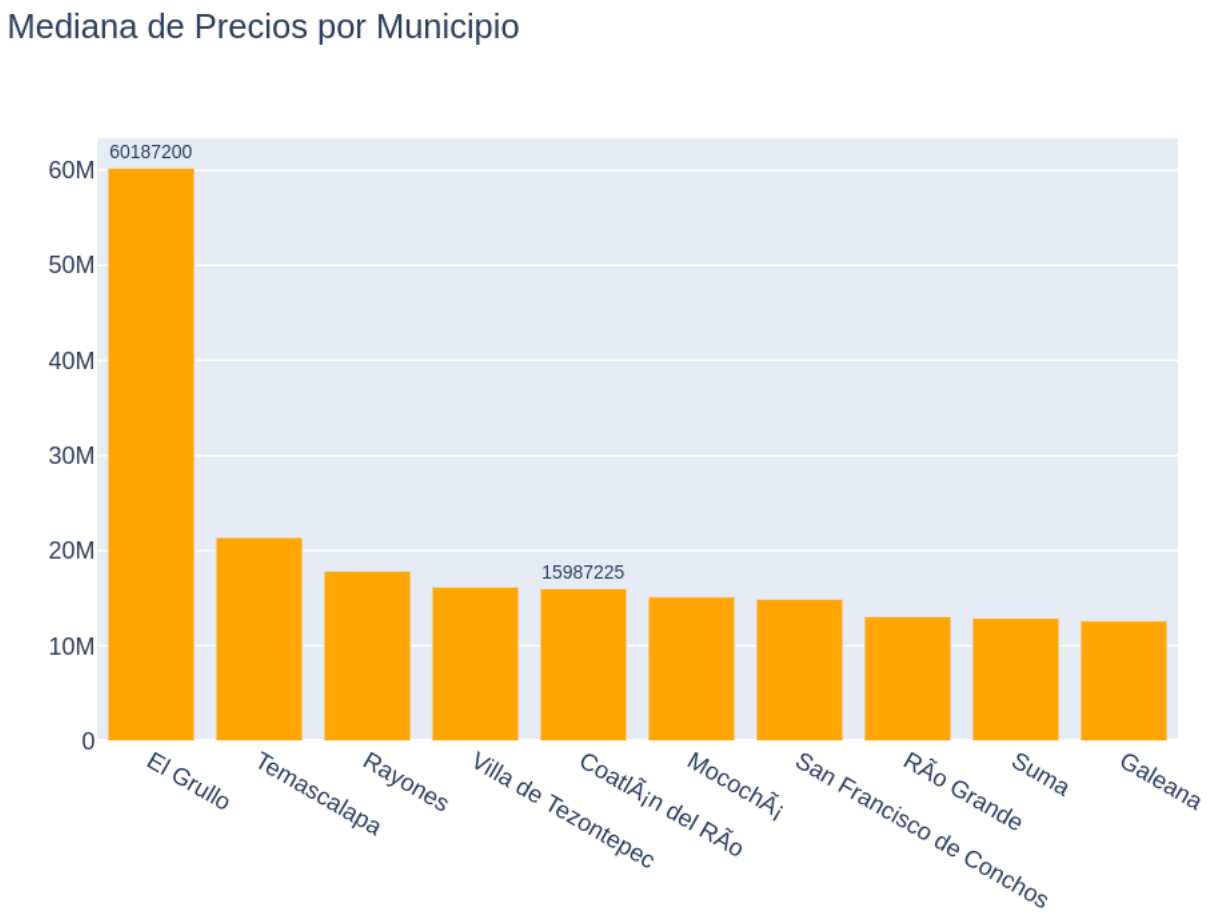
2.5.3. Municipios

Publicaciones



La mayoría de publicaciones de municipios se encuentran en las capitales de los estados, aunque existe una gran cantidad de valores nulos dentro de dicha variable.

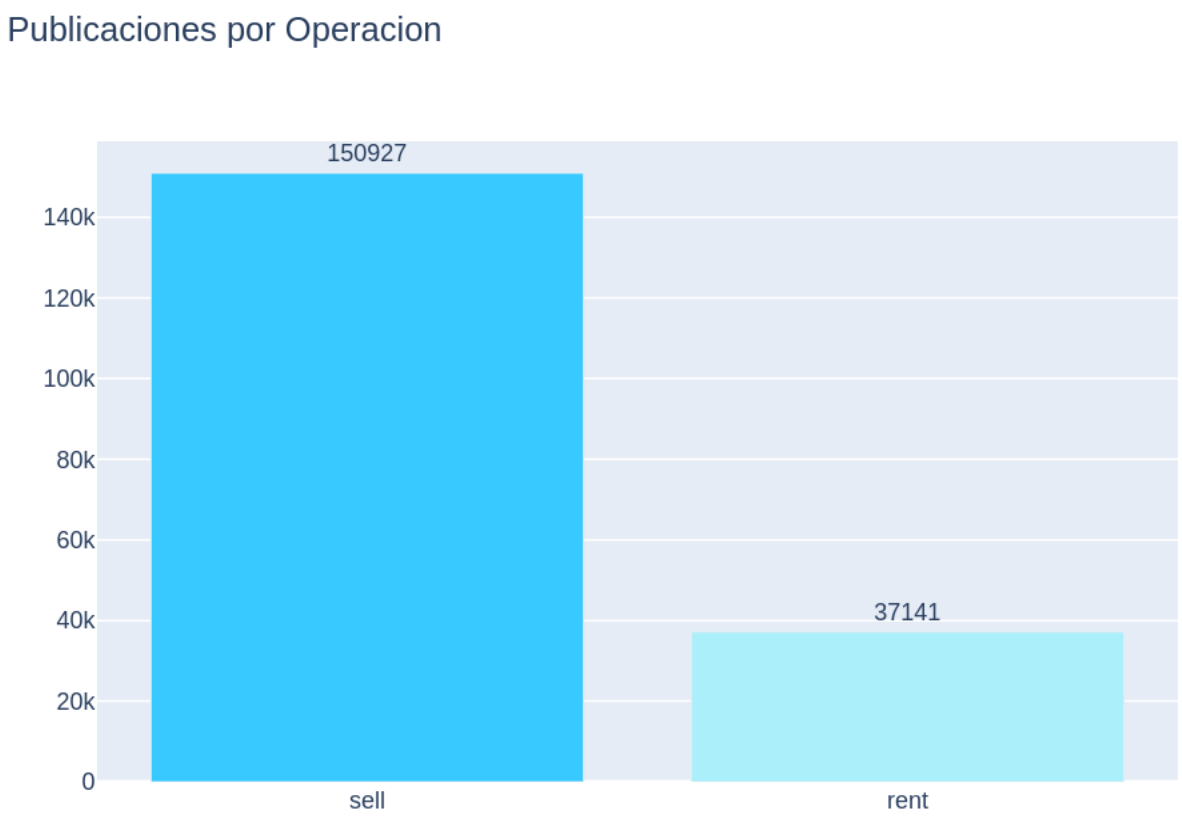
Precio



Observamos que los municipio que en mediana tienen un precio mayor son municipios no tan conocidos, esto puede ser debido a un error en el precio o dichos municipios contienen pocas publicaciones de inmuebles con valores muy altos.

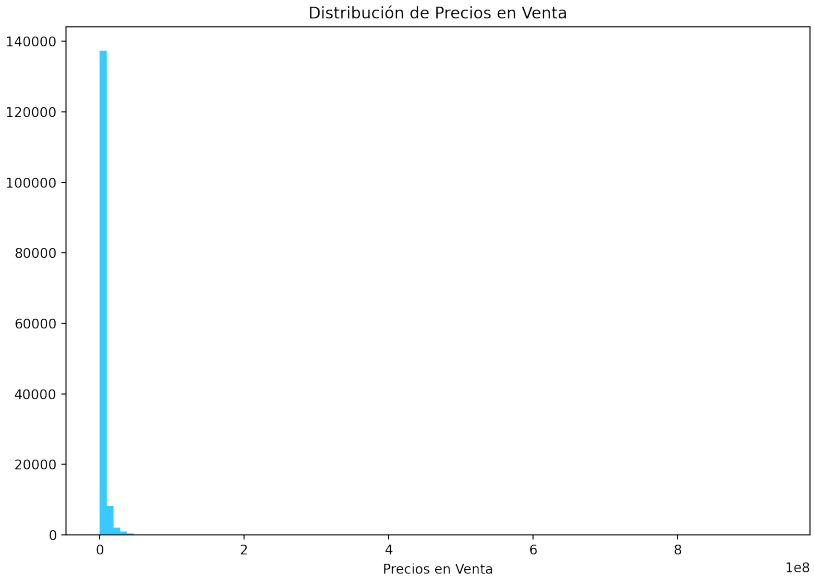
2.5.4. Venta y Renta

Publicaciones



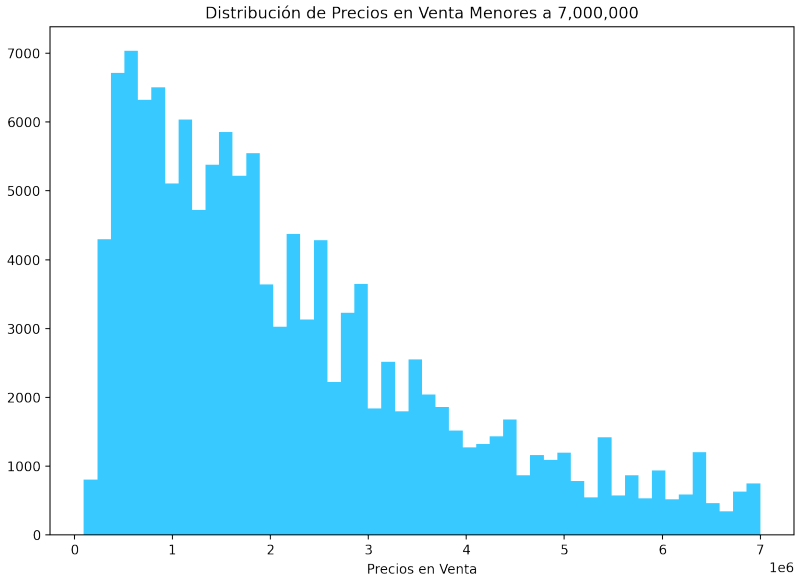
Existe una mayor cantidad de inmuebles en venta dentro de todo el conjunto de datos.

Precio de Venta



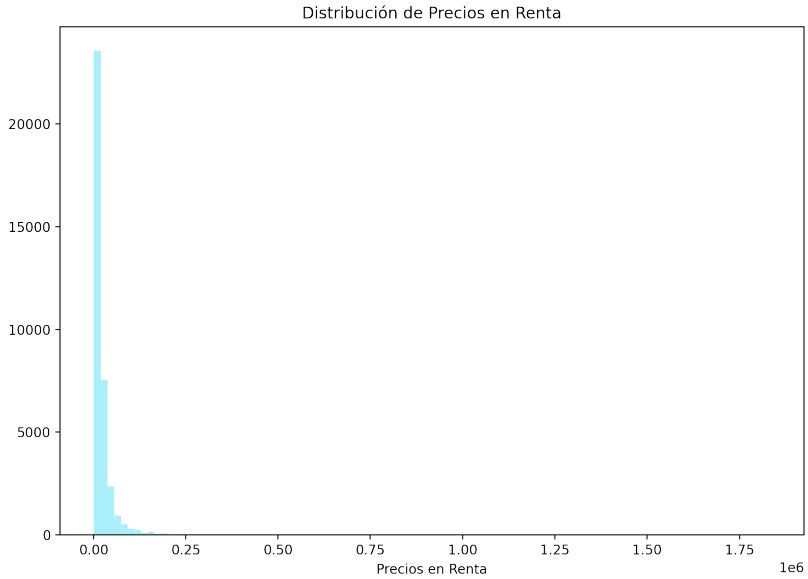
Observamos que la distribución de precios de venta se concentra a la izquierda de la gráfica, lo que nos da a entender que existen valores extremos que deben ser analizados, ya que la gráfica indica que existen inmuebles con valores mayores a los 900 millones de pesos.

Precio de Venta Menor a 7 Millones de Pesos



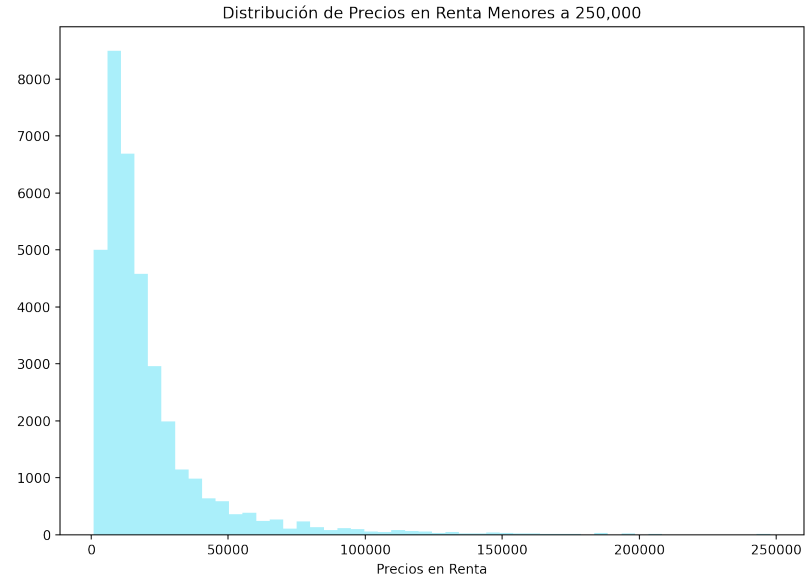
Al limitar las publicaciones por un precio menor a los 7 millones de pesos, que representa el 90 % de los precios según el análisis descriptivo hecho anteriormente, observamos una distribución más realista del mercado inmobiliario mexicano.

Precio de Renta



De igual manera que con la venta, la distribución de precios de los inmuebles en renta contiene valores extremos que afectan el análisis de este mercado.

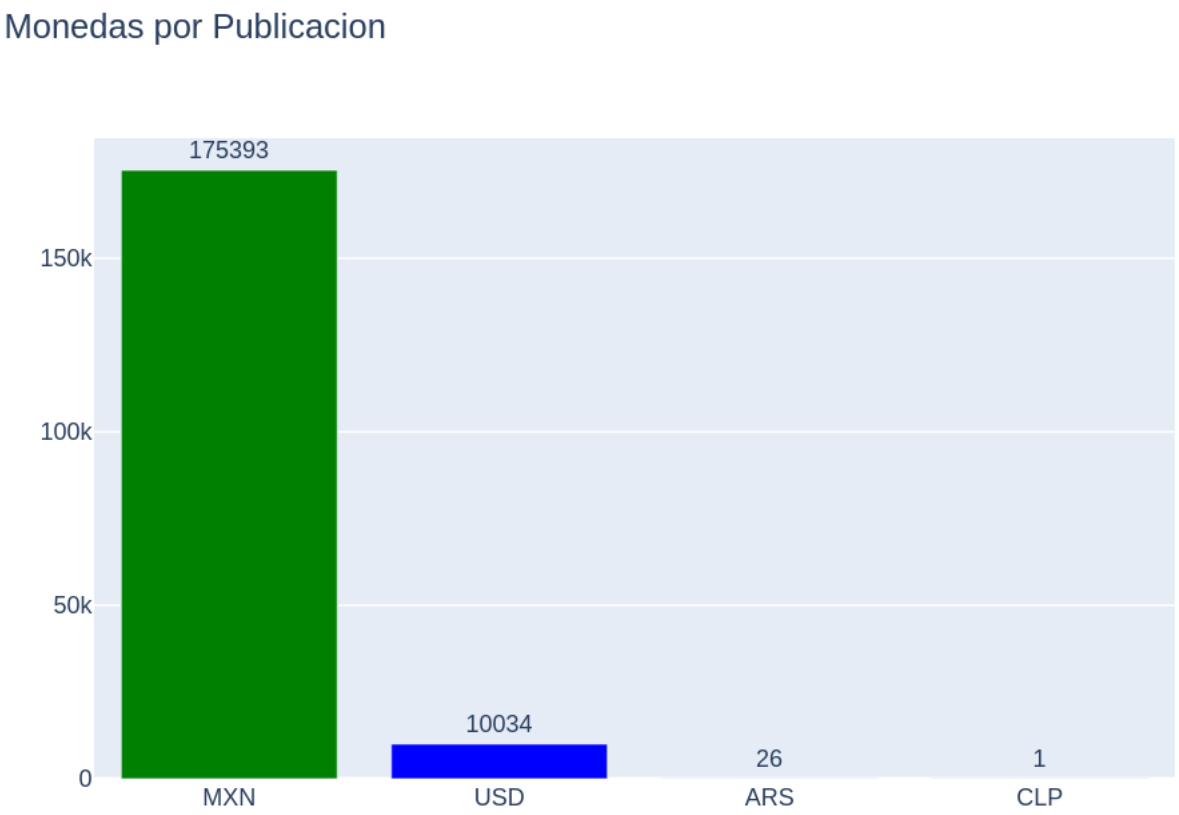
Precio de Renta Menor a 250 Mil Pesos



Limitando las publicaciones menores de 250 mil pesos de renta observamos nuevamente una distribución más realista del mercado mexicano.

2.5.5. Moneda

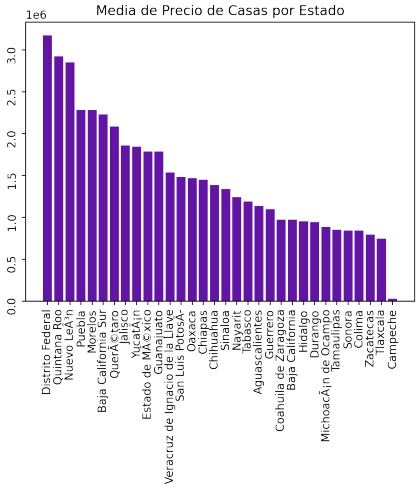
Publicaciones por Moneda

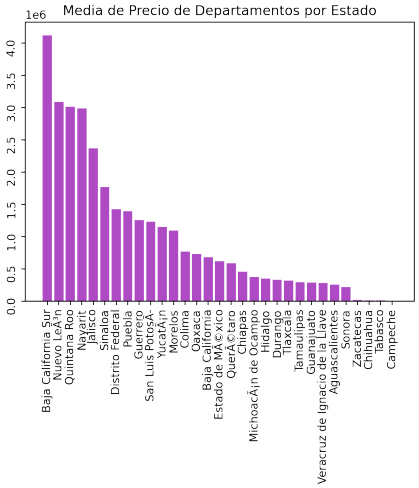


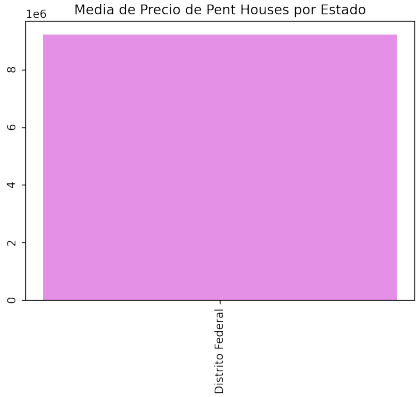
Se puede observar que no todas las monedas en las publicaciones son Pesos mexicano, lo que puede generar un sesgo en la información, estos registros deben ser eliminados para una comprensión correcta del mercado.

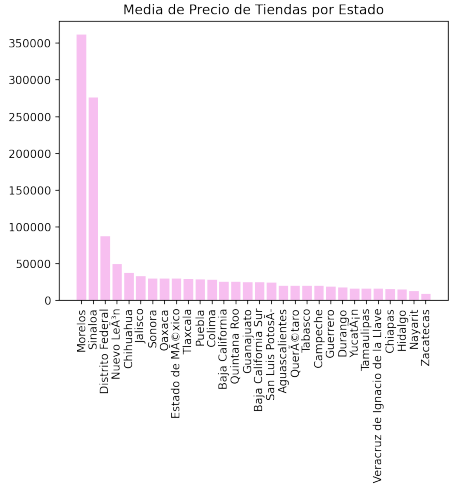
2.5.6. Tipo de Inmueble

Distribución de Precios por Tipo de Inmueble





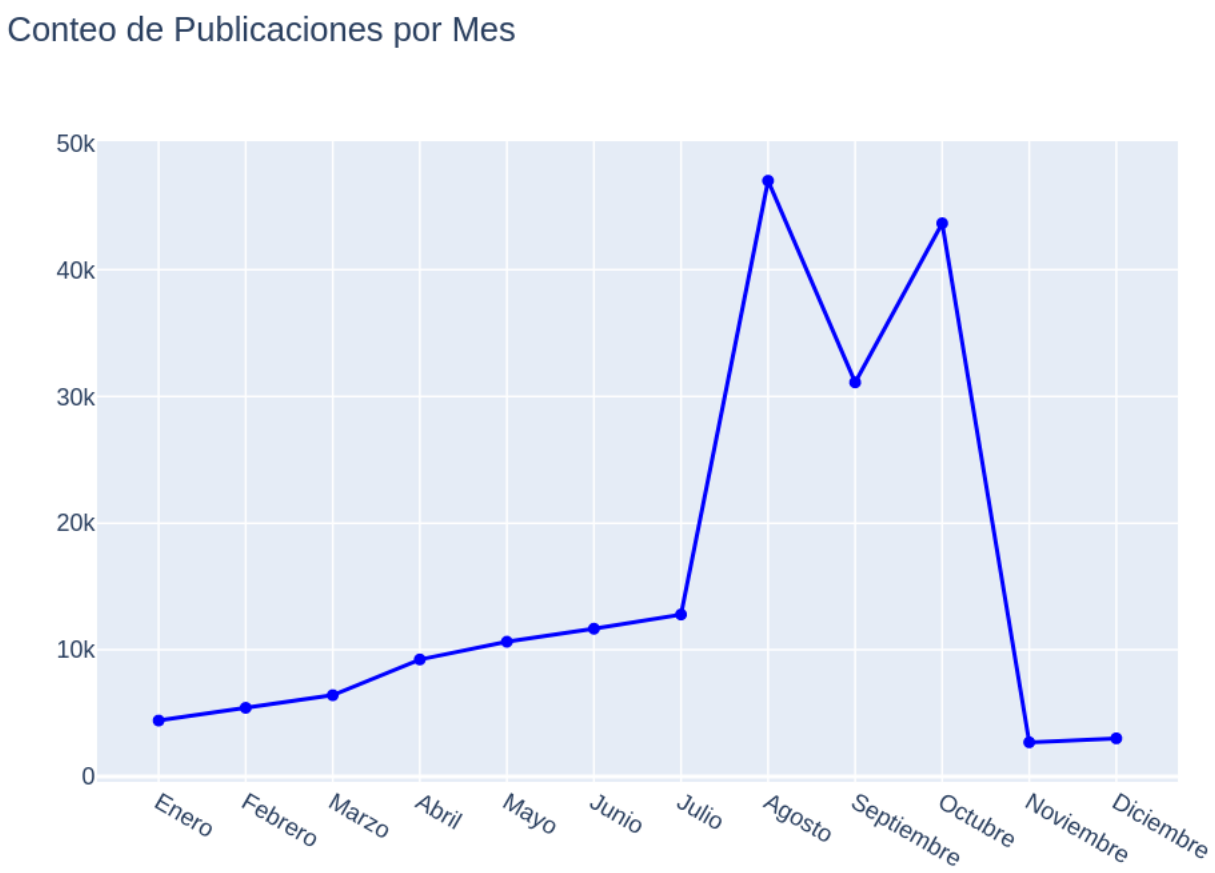




Observamos que las casas y departamentos tienen una distribución parecida en los precios, con una reducción de precios en los departamentos. No existe una distribución de precios en el mercado de Pent House, ya que solo existen publicaciones en la Ciudad de México. Por últimos los precios en tiendas son muy bajo a excepción de los estados de Morelos y Sinaloa, esto puede ser debido a la renta de bodegas en dichos estados.

2.5.7. Fechas

Publicaciones por Mes

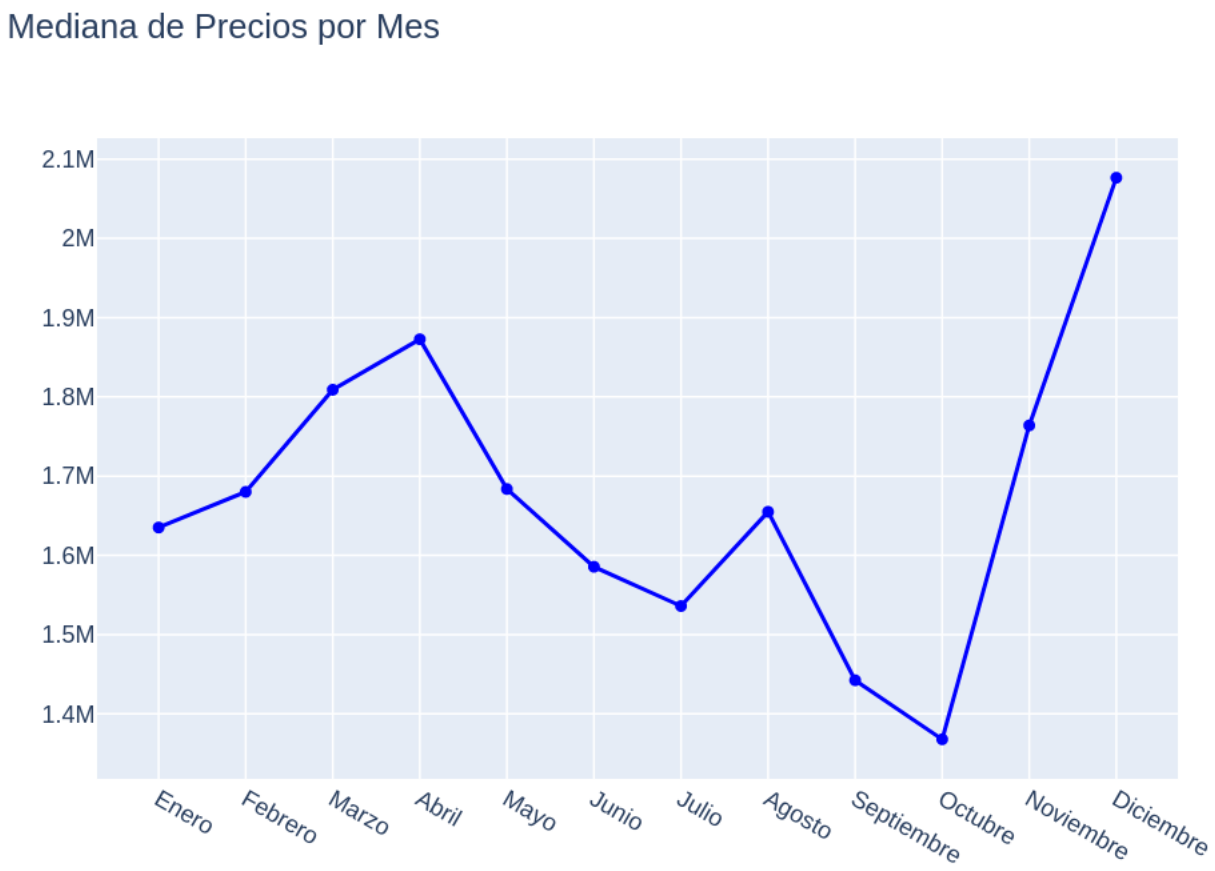


Se puede destacar el aumento de publicaciones en los meses de Agosto, Septiembre y Octubre. Esto puede ser debido a las vacaciones de verano, donde las personas tienen más tiempo libre en los meses de Junio y Julio para tomar fotos y crear descripciones de sus inmuebles para así publicarlos en los meses subsecuentes.

Publicaciones por Día

Los días con mayor número de publicaciones son los sábados, esto puede ser ocasionado al calendario laboral en México, ya que la mayoría de los trabajadores descansan sábados y domingos, aunque en domingo no existen tantas publicaciones, probablemente originado por la cultura mexicana de convivencia familiar en este día.

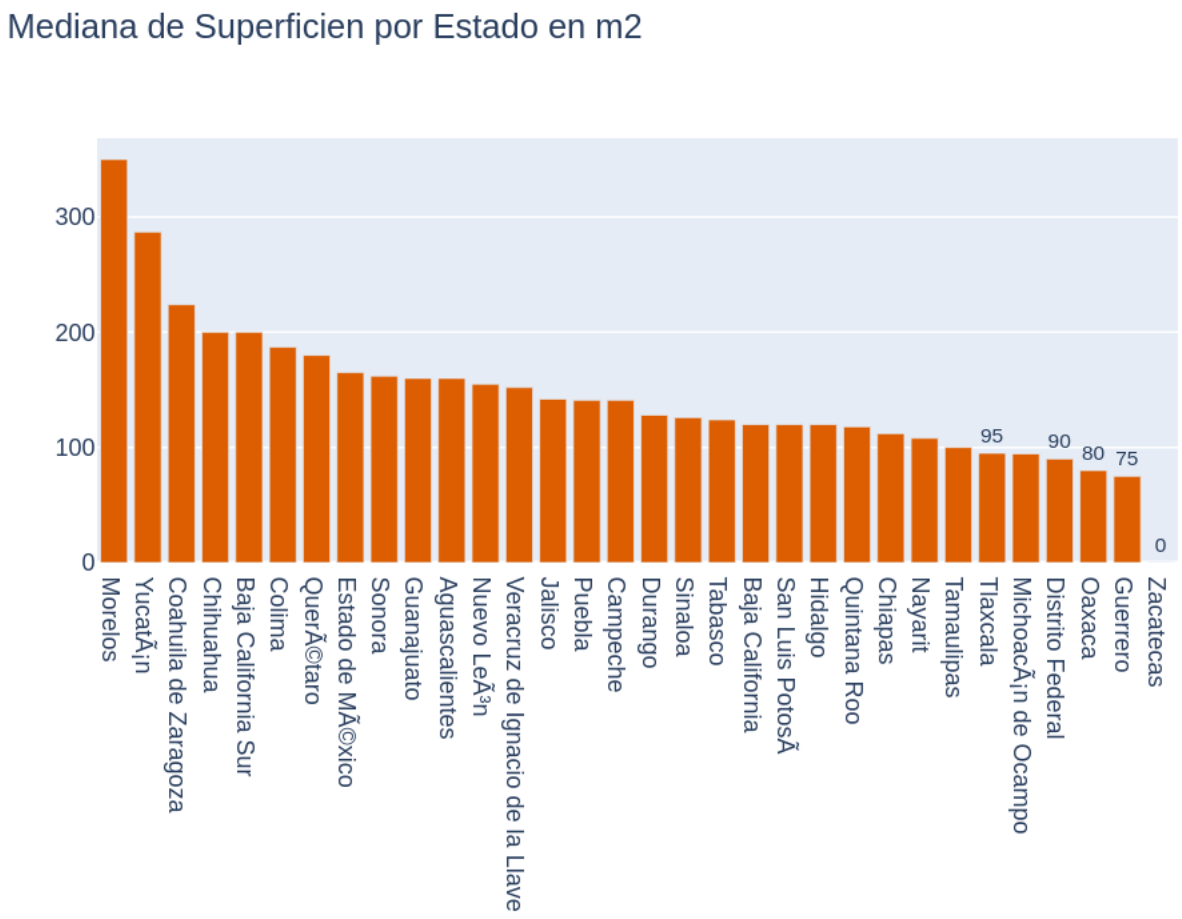
Precios por Mes



A comparación del aumento de publicaciones en los meses de Agosto a Octubre, los precios bajan en dichos meses, con un valor mínimo en el mes de Octubre. Este dato debe ser comparado con las tasas de créditos hipotecarios o la cantidad de préstamos que se realizan en dicho mes, para comprender porque se origina este fenómeno.

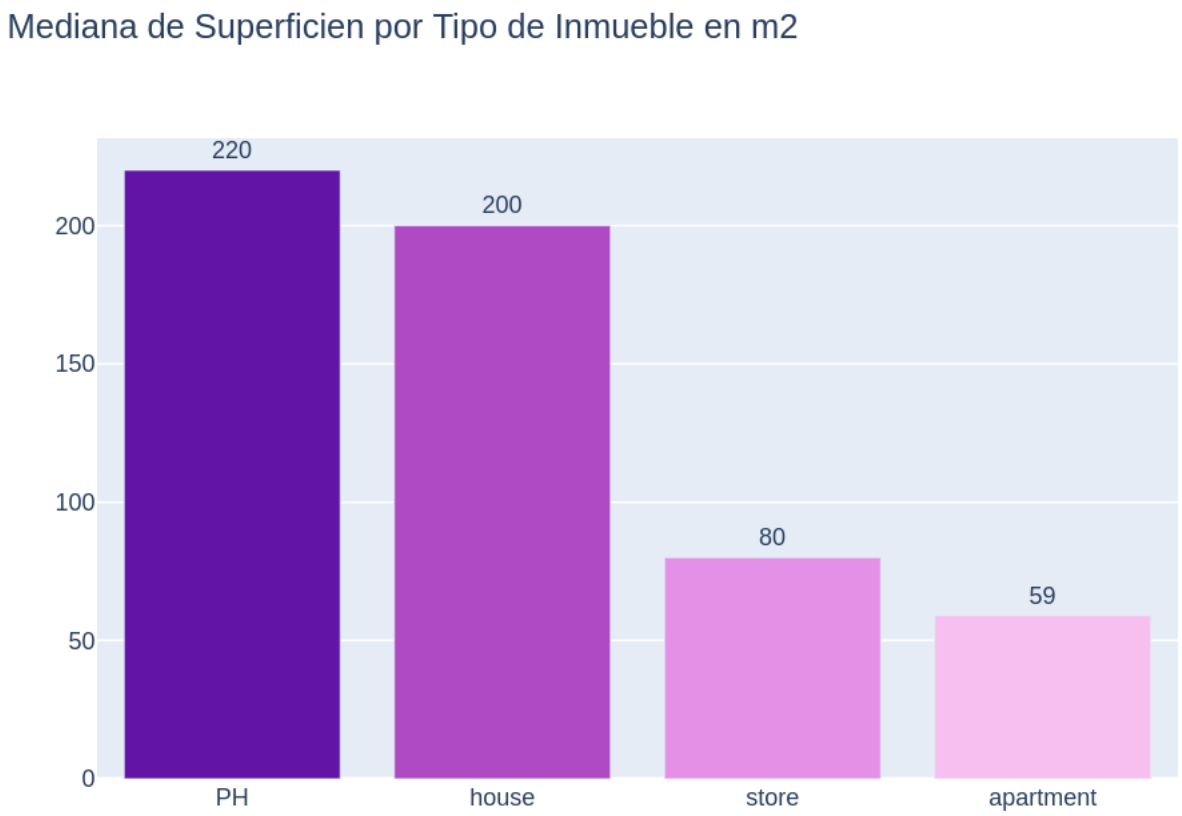
2.5.8. Superficie

Estado



Se puede observar que la mediana en superficie de un inmueble está relacionada con el tamaño del estado en el que se encuentre, entre más grande el estado, mayor superficie tendrán sus inmuebles.

Tipo de Inmueble



Se puede notar que los Pent House tienen una mayor superficie en mediana que los demás inmuebles, esto es producido a que existe una cantidad muy baja de ellos y sólo existe este tipo de inmueble en la Ciudad de México.

2.5.9. Análisis de Texto

Títulos de Publicaciones

Es notable que la mayoría de los títulos de publicaciones contienen el tipo de inmueble y la operación que se desea realizar, ya sea venta o renta, de igual manera muchas publicaciones contienen la ubicación de la casa o algún adjetivo positivo para dicha publicación.

Títulos de Publicaciones

Dentro de la descripción de las publicaciones observamos que las palabras más utilizadas son cuartos que contienen los inmuebles, como sala, cocina, recámara, comedor, entre otros. Por lo que es común publicar las descripciones de los distintos cuartos.

Capítulo 3

Tratamiento y Limpieza de Variables

El tratamiento y limpieza de las variables, ya sean discretas o continuas, es importante para la creación correcta de modelos predictivos de machine learning. Al no tener un conjunto de datos normalizado, con valores extremos y con valores ausentes, el modelo aprenderá a interpretar estos datos, lo que puede llevar a un sesgo y producirá una predicción incorrecta al utilizarlo. Es por eso que esta sección es una de las más importantes y conlleva la mayor parte del tiempo al realizar un análisis de los datos, ya que será un factor decisivo en los resultados finales.

3.1. Valores Extremos — Outliers

Los registros con valores extremos son considerados conjuntos de datos que se encuentran fuera del rango normal de nuestro conjunto total de datos, estas pueden estar en cualquier variable, desde discretas hasta continuas.

Estos valores son los que afectan de sobremanera la creación correcta de un modelo predictivo, ya que se deben ajustar a valores que normalmente no se observarían y crearían un sesgo en el resultado final.

Estos registros deben ser identificados con métodos estadísticos ya que, al tener una gran cantidad de datos, no se podrán diferenciar de una manera sencilla. Al ser identificados, estos valores, serán eliminados de la tabla final.

3.1.1. Outliers Variables Discretas

Al analizar los valores que se tienen dentro de las variables discretas observamos que la variable relacionada al tipo de moneda dentro de la publicación contiene distintas divisas, por lo cual los inmuebles que sean registrados con otro tipo de moneda que no sean pesos mexicanos serán eliminados.

La siguiente tabla muestra la cantidad de registros que se cuentan por tipo de monedas en la tabla de datos original.

Tipo de Moneda	Cantidad de Registros
MXN	175,393
USD	10,034
ARS	26
CLP	1

Existen 10,061 publicaciones con monedas distintas al peso mexicano y representan un 5.43 % del total de registros, estos serán eliminados.

3.1.2. Outliers Variables Continuas

Al tener dos tipos de operación (venta y renta) dentro de nuestro conjunto de datos se realizó un análisis de valores extremos en ambas categorías, esto debido a la diferencia significativa de precios y tipos de inmuebles que se tienen en cada una de las categorías.

Venta

features	n outliers IQR	n outliers Percentil	n outliers Z-Score	n outliers IQR %
c_lat	16847	64171	0	11.94
c_lon	24440	64171	0	17.32
c_price	11792	77569	1157	8.36
c_price_aprox_local_currency	11809	77593	1162	8.37
c_price_aprox_usd	11809	77593	1162	8.37
c_surface_total_in_m2	6161	37186	0	4.37
c_surface_covered_in_m2	7654	74328	0	5.42
c_price_per_m2	12057	71367	0	8.54

features	n outliers Percentil %	n outliers Z-Score %	total outliers	% outliers
c_lat	45.47	0.00	5924	4.20
c_lon	45.47	0.00	15225	10.79
c_price	54.96	0.82	7454	5.28
c_price_aprox_local_currency	54.98	0.82	7478	5.30
c_price_aprox_usd	54.98	0.82	7478	5.30
c_surface_total_in_m2	26.35	0.00	3370	2.39
c_surface_covered_in_m2	52.67	0.00	6673	4.73
c_price_per_m2	50.57	0.00	6486	4.60

Tomando en cuenta el método de IQR podrían eliminarse en el mejor de los casos 24,440 registros, lo que representa un 17.31 % de la información total en la operación de venta. Se usaron distintos métodos para eliminar outliers para no tener errores por algún sesgo en el método.

En la siguiente tabla se observan el número de registros al eliminar los datos extremos por variable, esta tabla solo cuenta con información de la operación venta.

Variable	Total de Elementos al Eliminar Valores
c_lat	135,209
c_lon	124,638
c_price	117,458
c_price_aprox_local_currency	117,437
c_price_aprox_usd	117,437
c_surface_total_in_m2	115,679
c_surface_covered_in_m2	113,169
c_price_per_m2	108,653

Renta

features	n outliers IQR	n outliers Percentil	n outliers Z-Score	n outliers IQR %
c_lat	1040	16009	0	3.04
c_lon	6520	16023	0	19.03
c_price	2476	18841	243	7.23
c_price_aprox_local_currency	2475	18837	235	7.22
c_price_aprox_usd	2475	18837	235	7.22
c_surface_total_in_m2	1964	10988	0	5.73
c_surface_covered_in_m2	2153	17306	0	6.28
c_price_per_m2	1880	15522	0	5.49

features	n outliers Percentil %	n outliers Z-Score %	total outliers	% outliers
c_lat	46.73	0.00	1040	3.04
c_lon	46.77	0.00	3105	9.06
c_price	54.99	0.71	1733	5.06
c_price_aprox_local_currency	54.98	0.69	1731	5.05
c_price_aprox_usd	54.98	0.69	1731	5.05
c_surface_total_in_m2	32.07	0.00	850	2.48
c_surface_covered_in_m2	50.51	0.00	1636	4.78
c_price_per_m2	45.31	0.00	1404	4.10

Observamos un cambio en la cantidad de outliers, ya que casi en todas las variables la cantidad es menor. Pero en el caso de la longitud hay una mayor cantidad de outliers, estos deberán ser eliminados correctamente ya que podrían afectar el análisis.

En la siguiente tabla se observan el número de registros al eliminar los datos extremos por variable, esta tabla solo cuenta con información de la operación venta.

Variable	Total de Elementos al Eliminar Valores
c_lat	33,220
c_lon	31,046
c_price	29,407
c_price_aprox_local_currency	29,405
c_price_aprox_usd	29,405
c_surface_total_in_m2	28,870
c_surface_covered_in_m2	28,133
c_price_per_m2	27,177

De igual manera que en el análisis anterior observamos una mayor cantidad de valores extremos en la variable longitud. La variable precio contiene el mismo número de outliers en sus tres variables correlacionadas y las variables de superficie contienen una menor cantidad de valores extremos a comparación de las otras variables.

De los 34,260 registros que existían en la tabla original se mantuvieron 27,177, una reducción del 20.67 % con respecto al total de registros de renta.

3.2. Valores Ausentes

De la misma manera que con los valores extremos usaremos las tablas de venta y renta de manera separada para imputar los valores ausentes en cada tabla. En las siguientes tablas se observará la completitud de cada variable por cada tipo de operación.

Venta

Variable	Total de Elementos al Eliminar Valores
v_operation	0.000000
v_property_type	0.000000
c_lat	20.047307
c_lon	20.047307
c_price	0.000000
v_currency	0.000000
c_price_aprox_local_currency	0.000000
c_price_aprox_usd	0.000000
c_surface_total_in_m2	55.184855
c_surface_covered_in_m2	3.860915
c_price_per_m2	8.030151
t_properati_url	0.000000
t_description	0.000000
t_title	0.000000
t_image_thumbnail	3.021546
v_estado	0.000000
v_municipio	0.000000
v_colonia	77.001095

Renta

Variable	Total de Elementos al Eliminar Valores
v_operation	0.000000
v_property_type	0.000000
c_lat	16.679545
c_lon	16.679545
c_price	0.000000
v_currency	0.000000
c_price_aprox_local_currency	0.000000
c_price_aprox_usd	0.000000
c_surface_total_in_m2	41.089892
c_surface_covered_in_m2	4.161607
c_price_per_m2	19.111749
t_properati_url	0.000000
t_description	0.000000
t_title	0.000000
t_image_thumbnail	3.039335
v_estado	0.000000
v_municipio	0.000000
v_colonia	81.705118

Podemos observar que las variables de longitud y latitud están incompletas pero la variable de estado y municipio no lo esta por lo que con la combinación de ambas variables se completaran las latitudes y longitudes faltantes. Las variables que no cuenten con esta información geográfica serán eliminadas.

Notamos que la variables c_surface_total_in_m2 contiene más del 40 % de valores nulos y existe la variable c_surface_covered_in_m2 que tiene una alta correlación con la anterior. Por lo tanto, se eliminará la variable de superficie total. Para completar los valores ausentes que existen en la variable c_surface_covered_in_m2 se utilizará la mediana de los tipos de inmuebles que se encuentren en el mismo estado y municipio que el registro ausente.

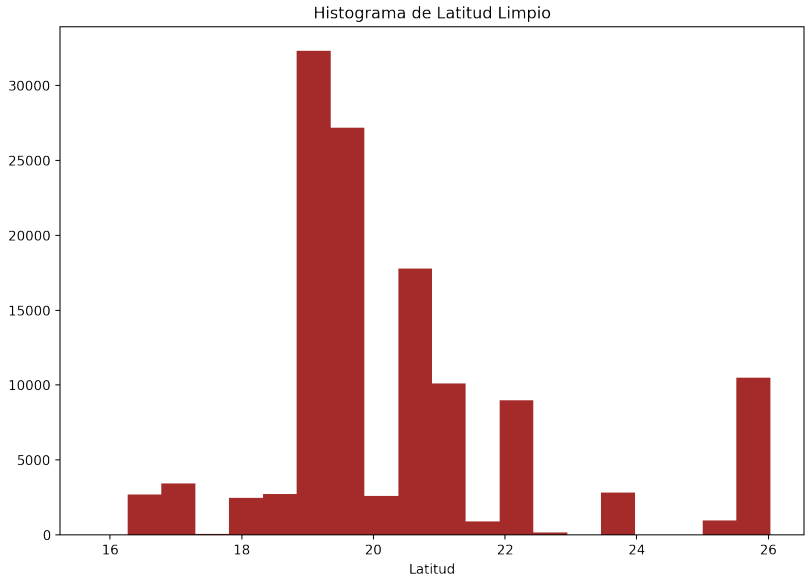
Al ser t_image_thumbnail una variable que es difícil de usar en el análisis estadístico será eliminada de ambas tablas. De igual manera la variable v_colonia será eliminada de ambas tablas por la alta cantidad de valores nulos que presentan en las dos tablas de datos.

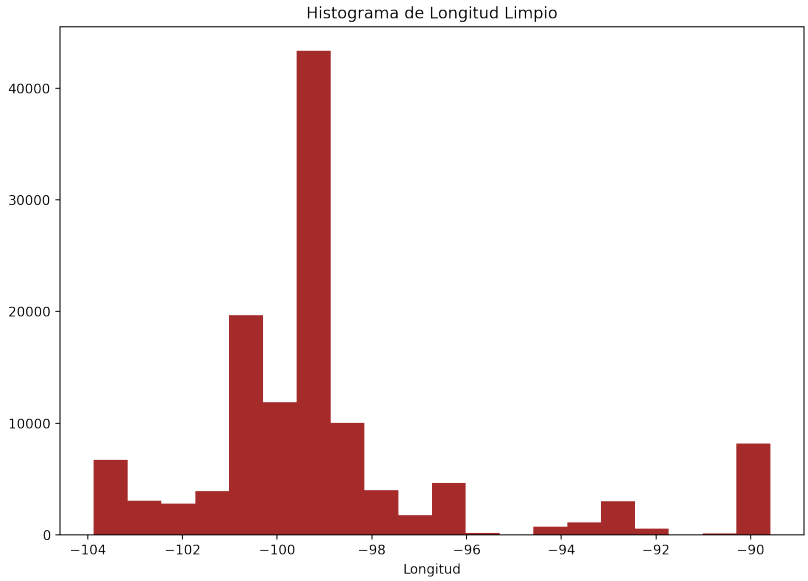
Con los tratamientos a los valores nulos pasamos de una tabla de 175,393 registros con 19 columnas, a una tabla con 125,684 registros y 16 columnas, es decir, observamos una reducción de registros del 28.34 % de la tabla de datos, al igual que una reducción del 15.78 % de las variables. Esta nueva tabla ayudará a facilitar el análisis de los datos ya que no contiene valores extremos, ni ausentes.

3.3. Visualización de Datos Limpios

3.3.1. Geográfica

Latitud y Longitud





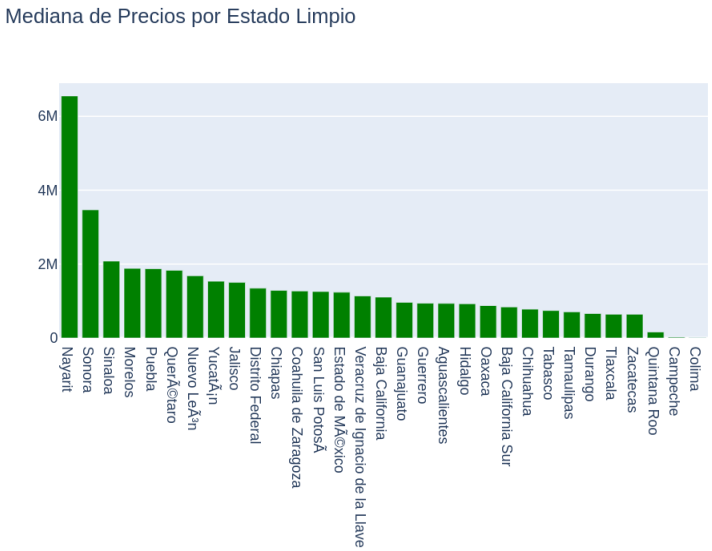
Observamos un gran cambio en el rango de latitudes, esto nos indica que todos los registros con latitud mayor a 26 eran erróneos o estaban muy alejados de los datos normales. También vemos una concentración de los datos en un rango de -104 a -90 en las longitudes. En los datos originales el rango llegaba a 100 lo que indica un gran error en los datos.

3.3.2. Estados

Cantidad de Publicaciones

Vemos una reducción notable en las publicaciones de los estados, existen cambios de posiciones en los estados y vemos la desaparición de los registros de Michoacán, ya que únicamente contaba con cuatro publicaciones.

Mediana de Precios



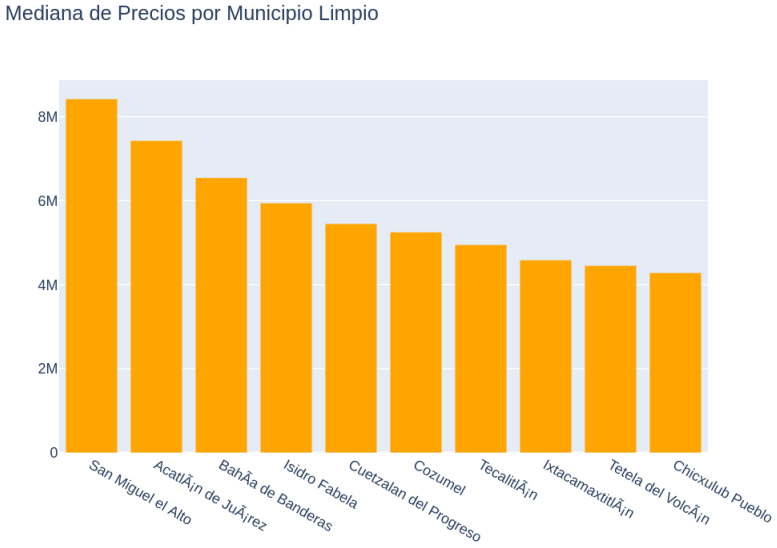
Es notable un cambio radical en los precios de Nayarit, esto puede ser debido a la baja cantidad de datos que existen en este estado, de igual manera los precios en los estados con las ciudades más importantes en el país redujeron el precio. Se observa que el estado de Quintana Roo tuvo la mayor reducción, esto puede ser ocasionado por el alto precio que tienen los inmuebles en dicho estado por la cercanía a zonas turísticas.

3.3.3. Municipios

Cantidad de Publicaciones

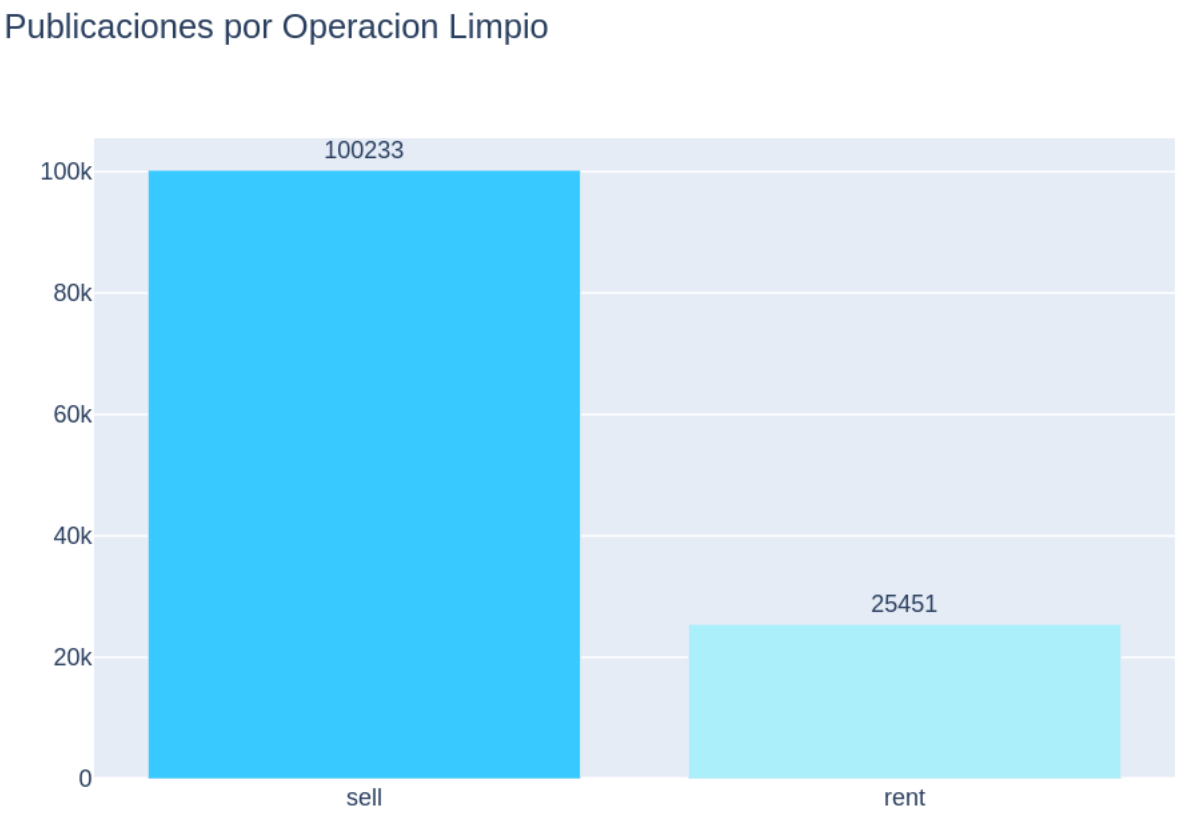
La mayoría de la ciudades con mayor número de publicaciones se conservan, aunque el orden cambia con respecto a los datos originales. También podemos ver que los datos de Mérida se encuentran dentro de la tabla limpia aunque este sea un municipio que se encuentra en los extremos del país, es decir no se eliminaron los estados que se encuentran en las regiones fronterizas del país.

Mediana de Precios



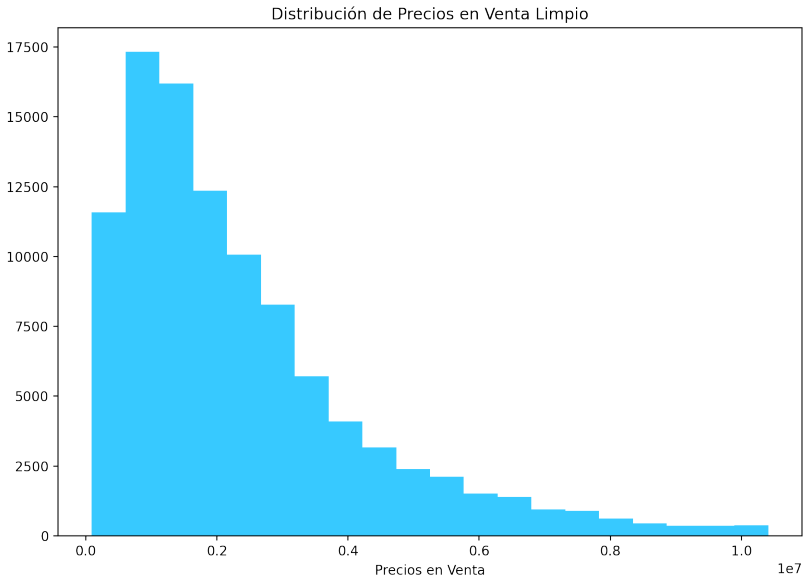
Vemos un cambio radical en los municipios con mediana más alta, observamos un cambio en el rango de precios acorde al país, ya que pasamos de un rango superior de 60 millones de pesos a 8 millones de pesos.

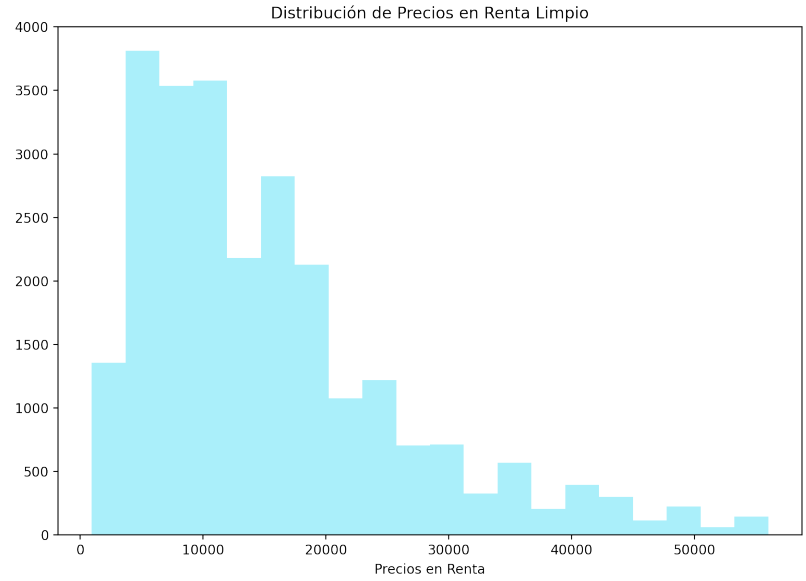
3.3.4. Venta v.s. Renta



Notamos una mayor reducción en las publicaciones de venta que de renta al limpiar la tabla de datos. Pasaron de 150,000 publicaciones a 100,000.

Distribución de Precios Venta y Renta

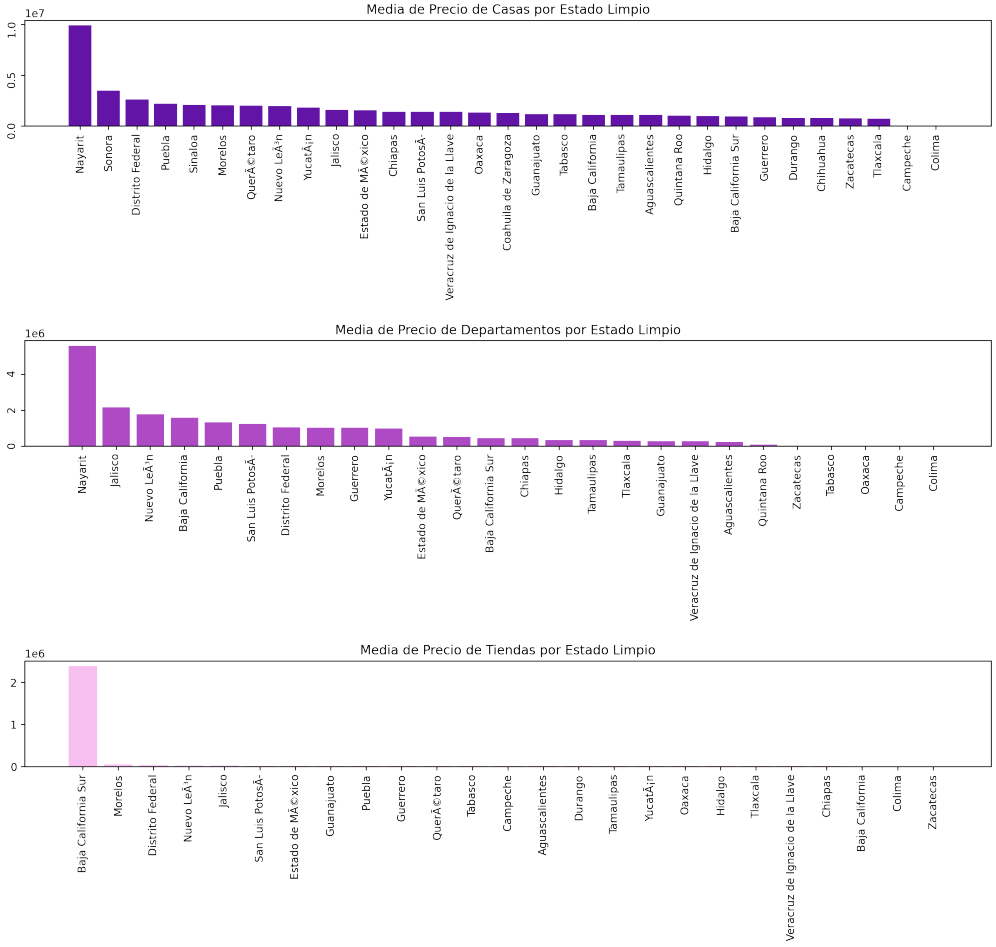




Observamos un cambio favorable en el rango de precios ya que van desde menos de un millón de pesos hasta los 10 millones. En los datos originales el rango de precios va desde 0 hasta los 900 millones de pesos.

De igual manera vemos un gran cambio en el rango de precios de rentas, ya que el original va desde los 0 pesos hasta los 1.75 millones de pesos. El nuevo rango tiene una mayor concentración de precios de renta entre los 4 mil a los 10 mil pesos, valores más realistas para el mercado mexicano.

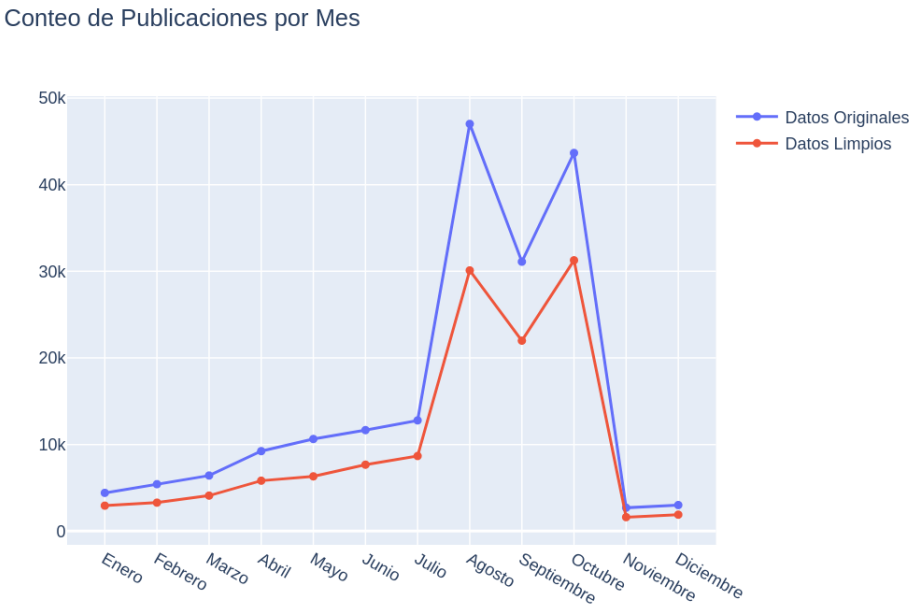
3.3.5. Tipo de Inmueble



Vemos un cambio en el rango de precios en los tres tipos de inmueble, de igual manera observamos que la mediana de precios en Nayarit es muy alta, esto debido a la poca cantidad de datos con los que se cuenta. Vemos que los Pent House se eliminaron ya que se consideran datos extremos al sólo existir registros en la Ciudad de México.

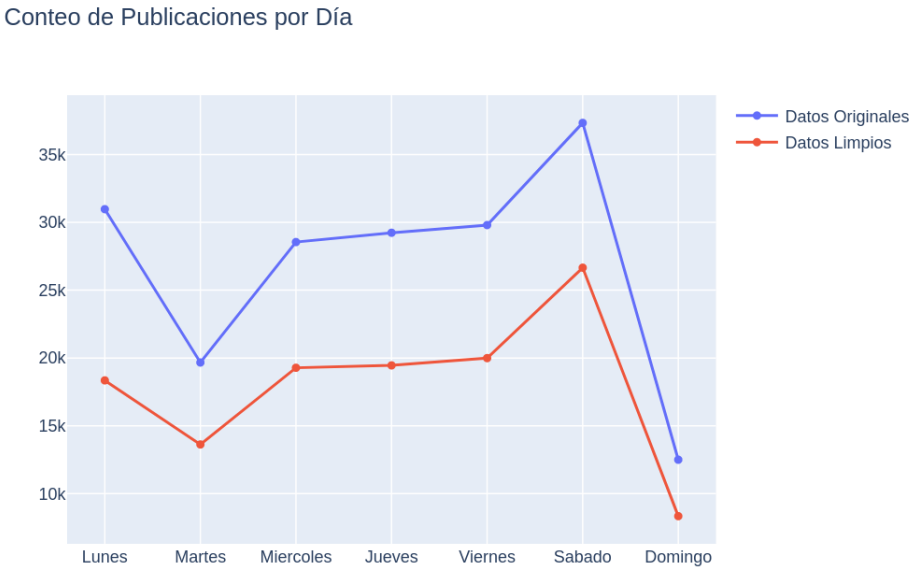
3.3.6. Fechas de Publicación

Publicaciones por Mes



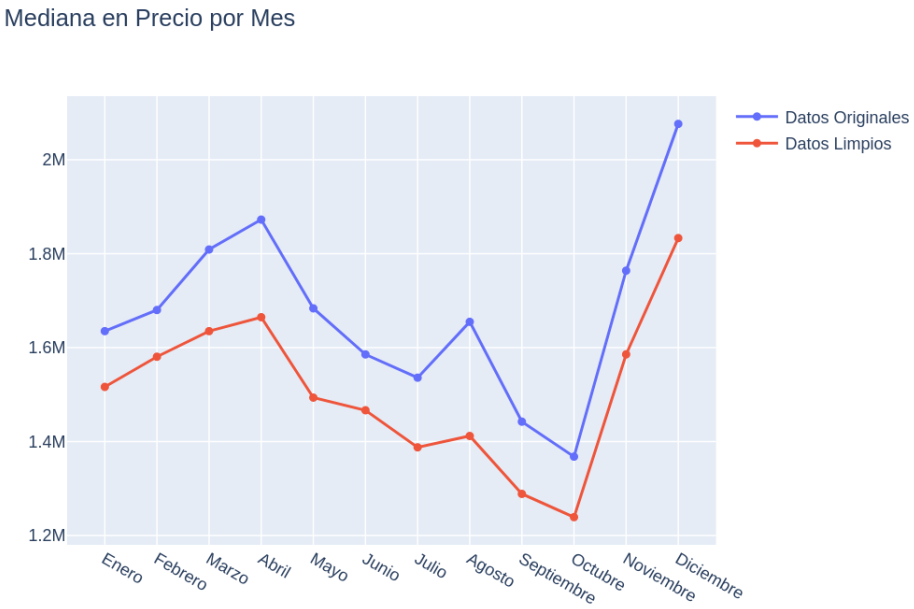
Existe un comportamiento igual a la serie original, pero desplazado algunas unidades hacia abajo, esto debido a la eliminación de los precios más altos en las tablas de datos.

Publicaciones por Día



De igual manera observamos un comportamiento parecido entre las dos gráficas pero desplazado. Observando que las publicaciones entre lunes y martes no tienen un cambio tan brusco como se ve en los datos originales e igualmente el día domingo no contiene una gran diferencia entre las publicaciones originales y limpias.

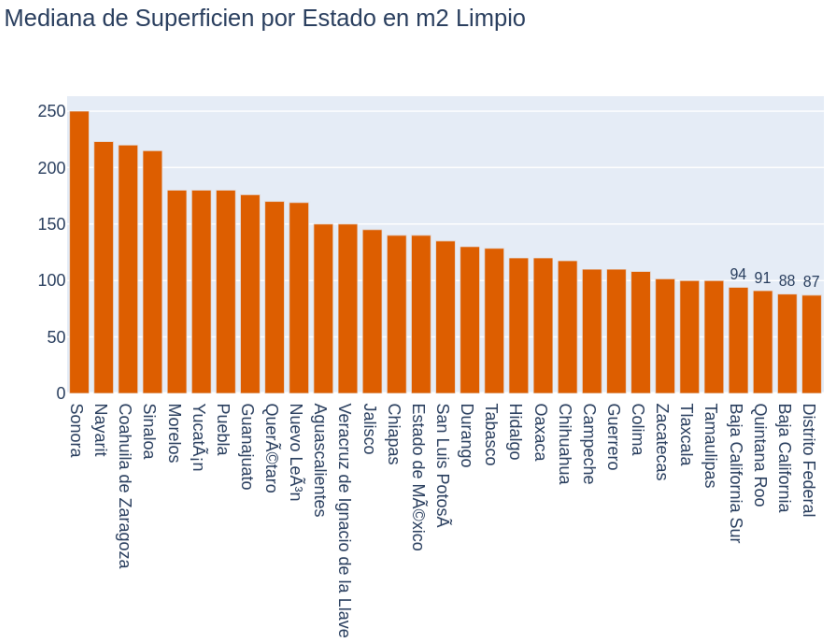
Precio por Mes



Por último, notamos que en los datos originales no se puede apreciar de buena manera la reducción de precios entre los meses de abril y octubre, pero con los datos limpios notamos que los precios de los inmuebles bajan en estos meses teniendo un mínimo en el mes de octubre.

3.3.7. Superficie

Superficie por Estado



Con los datos limpios podemos notar de mejor manera como los inmuebles en los estados más grandes tienen mayor superficie que los estados más chicos y más conglomerados. Sonora el segundo estado más grande de todo México y la Ciudad de México el estado más chico de la república.

Superficie por Tipo de Inmueble

Vemos que se elimina los Pent House e incrementan la superficie en cada un de los tipos de inmueble, de igual manera podemos notar que los apartamentos tienen una mayor superficie a comparación de las tiendas, esto es más común en el mercado inmobiliario.

Capítulo 4

Ingeniería de Variables — Feature Engineering

El objetivo de la ingeniería de variables es crear nuevas características a partir de la información que proporcionan las distintas variables dentro de la tabla de datos. Esto puede servir para obtener nueva información relevante, segmentar variables discretas que tengan una gran cantidad de datos o resumir alguna característica que posea mucha información. Estas nuevas variables estarán altamente correlacionadas con las variables originales, pero en la siguiente sección se reducirán las características con menos importancias para el modelo.

Las variables relacionadas con el precio no serán modificadas ya que estas se utilizaran como variable objetivo y esto podría sesgar el entrenamiento correcto del modelo.

4.0.1. Variables Fecha

v_anio_2015	v_anio_2016	v_mes_2	v_mes_3	...
0	0	0	0	...
0	0	0	0	...
0	0	0	0	...
0	0	0	0	...
0	0	0	0	...

Se crean variables indicadoras para fecha, con división de años, meses y día de la semana. Al crear estas variables indicadoras se elimina la primera variable categórica para no tener multicolinealidad con las demás variables creadas.

4.0.2. Variables de Tipo de Operación

v_operation_sell
1
1
1
1
1

Se crea una variable indicadora para la operación al no ser una variable ordinal, de igual manera se elimina la columna de v_operation_rent para eliminar la relación entre las variables.

4.0.3. Variables de Tipo de Propiedad

v_property_type_house	v_property_type_store
1	0
1	0
0	0
1	0
1	0

Eliminamos la variable v_property_type y obtenemos dos nuevas variables las cuales indican si el tipo de inmueble es casa o tienda, si ambas indican 0 el inmueble es un departamento.

4.0.4. Variables Latitud y Longitud

...	v_conteo_geo
...	10.0
...	1.0
...	10.0
...	10.0
...	10.0

Se crea una nueva variable con el conteo de publicaciones por latitud y longitud.

4.0.5. Variable Título

...	title_casa	title_centro	title_departamento	title_fracc
...	1	0	0	0
...	1	0	0	0
...	0	0	1	0
...	1	0	0	0
...	1	0	0	0

Se agregaron las 10 palabras más repetidas en los títulos y su número de apariciones en el título y se eliminó la variable t_title. Se realizará lo mismo para la descripción.

4.0.6. Variable Descripción

...	descr_baao	descr_br	descr_casa	descr_cocina
...	2	0	0	1
...	0	0	1	0
...	1	0	0	1
...	1	0	0	1
...	1	0	0	1

4.0.7. Variable Estado

...	**v_region_norte**	**v_region_sur**
...	0	0
...	0	0
...	0	0
...	0	0
...	0	0

Se crean variables categóricas (norte, centro y sur) para poder dividir los estados de una forma más reducida.

4.0.8. Variable Estado

...	v_tamanio_grande	v_tamanio_mediano
...	0	1
...	1	0
...	0	1
...	0	1
...	0	1

Se crean dos variables categóricas con valores del tamaño de la localidad, si la localidad tiene un mayor número de publicaciones que la mediana de publicaciones será grande. De igual manera para un tamaño mediano y chico.

La tabla final, al hacer la limpieza de variables y crear nuevas características a partir de las variables originales, cuenta con 125,684 registros y 57 variables. Esta gran cantidad de características se reducirá en la siguiente sección.

Capítulo 5

Reducción de Dimensiones

Al tener una gran cantidad de variables, no se puede analizar de forma correcta cada una de ellas. También muchas de estas variables están relacionadas entre si, esto puede ocasionar problemas al momento de la modelación, ya que unas características ya explican a otras de manera adecuada.

La reducción de dimensiones se dividirá en dos secciones, selección de características y un análisis de componentes principales.

5.1. Selección de Características

Dentro de esta sección se eliminarán las variables que no tengan una importancia para el modelo. Esta selección se realizará mediante métodos estadísticos debido a la gran cantidad de datos. Los métodos esencialmente eliminarán las variables que estén altamente relacionadas entre si, contengan una variación muy baja entre los datos o posean una relación tan baja con el precio que no influya si estas características llegaran a cambiar.

Para comenzar, las variables v_currency, t_properati_url al contener valores únicos o tener información en texto que no servirá para la parte de modelacíon se eliminarán. Igualmente las variables c_price_aprox_local_currency, c_price_aprox_usd, c_price_per_m2 serán eliminadas ya que estás fueron calculadas con la información de precio, la variable objetivo.

Variables altamente relacionadas con otras
v_region_sur
v_tamanio_grande
v_tamanio_mediano
title_renta
title_venta
title_casa
title_departamento
title_local
descr_comedor
descr_sala

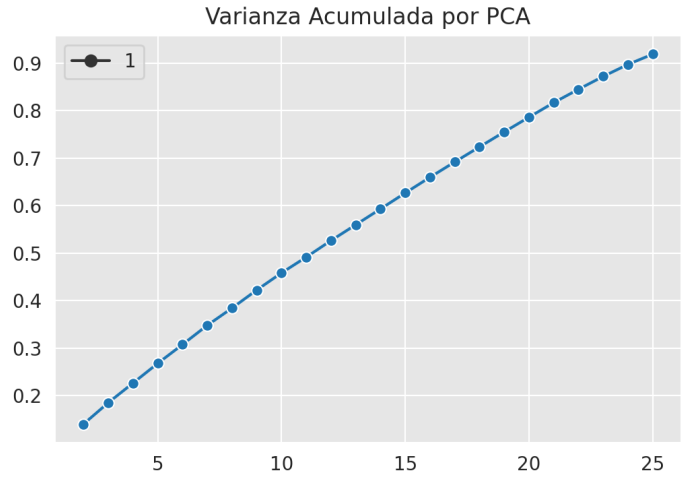
Variables de poca relevancia
c_lat
c_lon
v_anio_2015
v_anio_2016
v_mes_8
v_mes_10
descr_id
descr_nocnok
descr_sistema

Observamos que las variables de región sur, o el tamaño del municipio no son relevantes, esto puede ser a que están muy relacionadas con la zona geográfica. De igual manera las variables de título como renta, venta, casa , departamento entre otras están muy relacionadas con las variables indicadoras que indican el tipo de propiedad o tipo de operación.

La variables de año y algunos meses tampoco son relevantes para la predicción del precio, esto nos quiere decir que los años no importarán para el análisis del mercado.

5.2. Análisis de Componentes Principales — PCA

Con el análisis anterior se obtuvo una conjunto de datos reducido a 33 características, estas se tomarán para hacer un análisis de componentes principales. El análisis consiste en crear una combinación de todas las características, dándole peso a las variables con mayor varianza explicativa, esto es para poder dividir de mejor manera el conjunto de datos.



Observamos un crecimiento muy lento en la varianza acumulada, ya que con 24 variables podemos obtener un 90 % de la varianza explicada. Es decir que reduciendo 9 variables podemos explicar el 90 % de varianza.

Capítulo 6

Conjunto de Entrenamiento y Prueba

Con el conjunto utilizado en la selección de características se crearon dos nuevos conjuntos, una tabla de entrenamiento de modelos que cuenta con 87,978 registros (aproximadamente un 70 % de los datos) y una tabla de prueba que cuenta con 37,706 registros. La tabla de entrenamiento servirá como entrada en un modelo de estadístico para ser entrenado y poder predecir los precios de las casa según sus características. El conjunto de prueba ayudará a validar el funcionamiento correcto de dicho modelo.

Estructura Final de la Tabla

d_created_on	c_surface_covered_in_m2	v_mes_2	v_mes_3	...
2016-09-22	150.0	0	0	**...**
2016-07-18	450.0	0	0	**...**

...	v_mes_6	v_mes_7	v_mes_9	v_mes_11
...	0	0	0	0
...	0	0	0	0

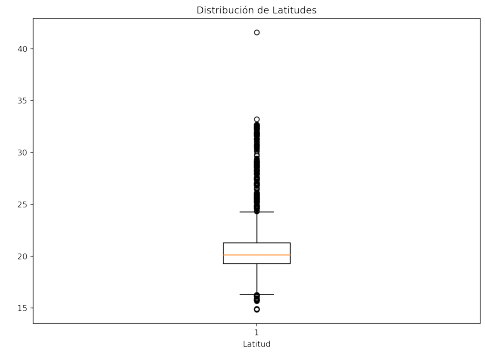
...	title_fracc	title_lomas	title_residencial	title_san
...	0	0	0	0
...	0	0	0	1

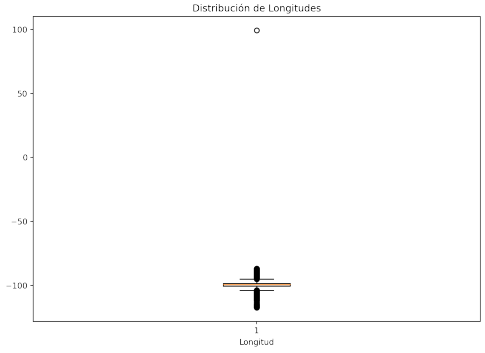
...	descr_casa	descr_cocina	descr_recamaras	v_region_norte
...	1	1	1	0
...	0	0	0	0

Capítulo 7

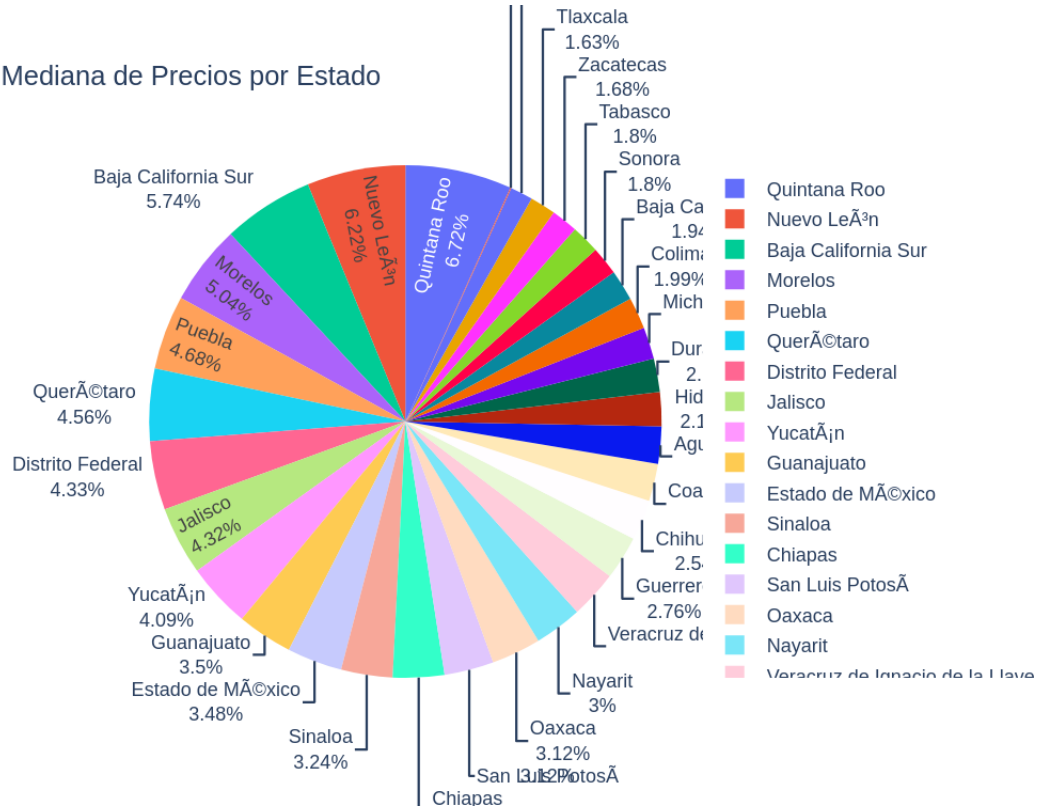
Apéndice

7.1. Gráficos Adicionales

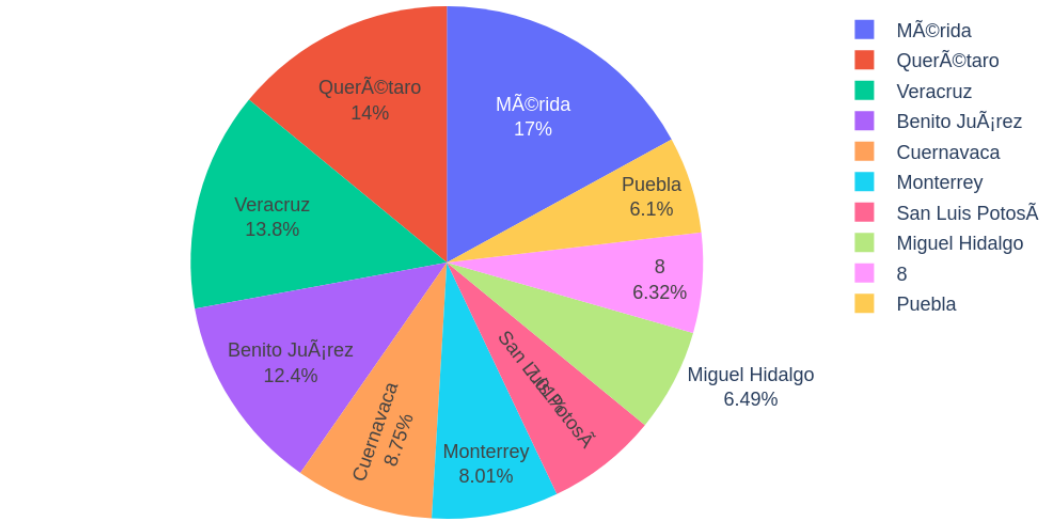




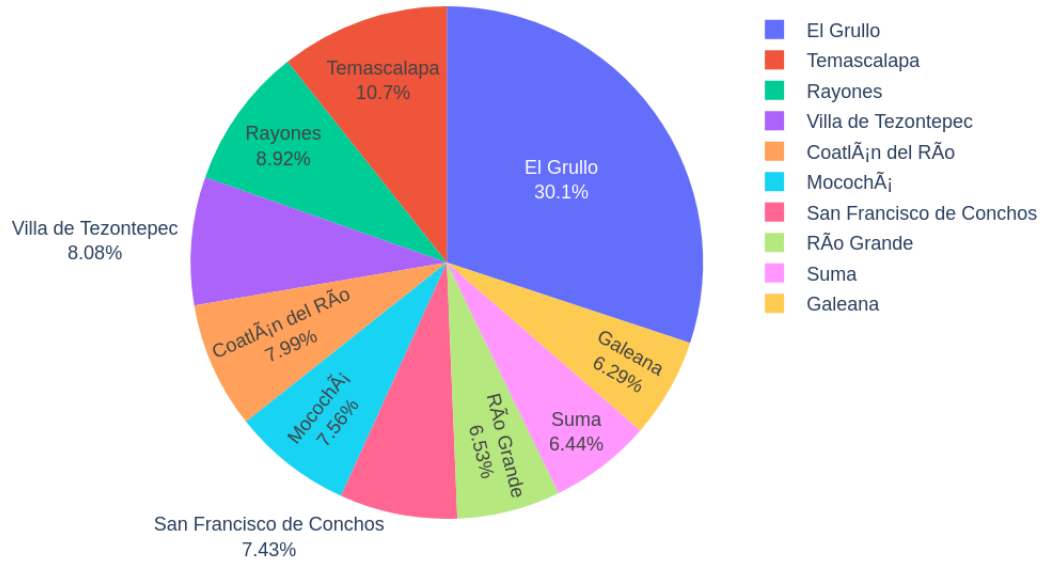

Número de Publicaciones por Estado
Distrito Federal 14.8%
Estado de MÃ©xico 11%
Veracruz de Ignacio de la Llave 7.31%
YucatÃ¡n 7.3%
Nuevo LeÃ³n 7.2%
QuerÃ©taro 7.2%
Morelos 7.19%
Puebla 5.51%
Jalisco 3.81%
Quintana Roo 3.7%
Nayarit 0.565%
Colima 0.624%
Sonora 0.731%
Distrito Federal
Estado de MÃ©xico
Veracruz de Ignacio de la Llave
YucatÃ¡n
Nuevo LeÃ³n
QuerÃ©taro
Morelos
Puebla
Jalisco
Quintana Roo
Guerrero
Tamaulipas
Chiapas
Baja California
Chihuahua


Mediana de Precios por Estado
Quintana Roo 6.72%
Nuevo LeÃ³n 6.22%
Baja California Sur 5.74%
Morelos 5.04%
Puebla 4.68%
QuerÃ©taro 4.56%
Distrito Federal 4.33%
Jalisco 4.32%
YucatÃ¡n 4.09%
Guanajuato 3.5%
Estado de MÃ©xico 3.48%
Sinaloa 3.24%
Chiapas
Oaxaca 3.12%
Nayarit 3%
Veracruz de
2.76%
Tlaxcala 1.63%
Zacatecas 1.68%
Tabasco 1.8%
Sonora 1.8%
Quintana Roo
Nuevo LeÃ³n
Baja California Sur
Morelos
Puebla
QuerÃ©taro
Distrito Federal
Jalisco
YucatÃ¡n
Guanajuato
Estado de MÃ©xico
Sinaloa
Chiapas
Oaxaca
Nayarit

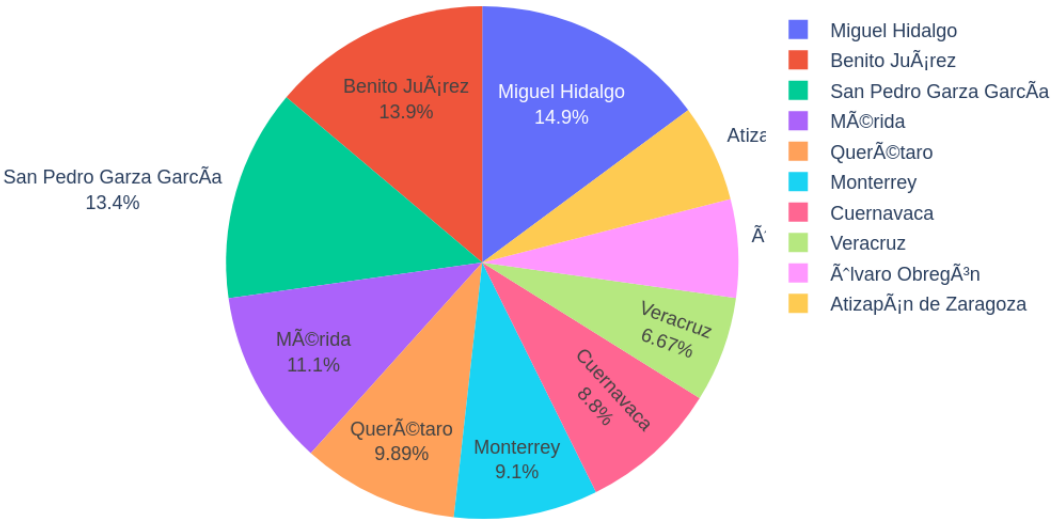
10 Municipios con mas Publicaciones



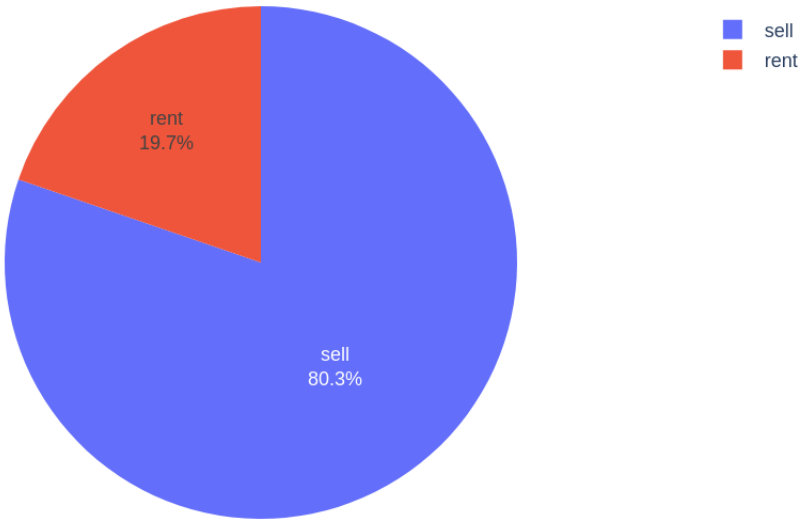
Mediana de Precios por Municipio

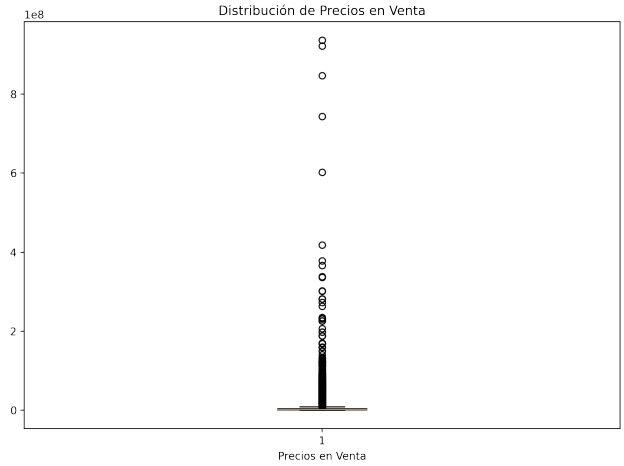


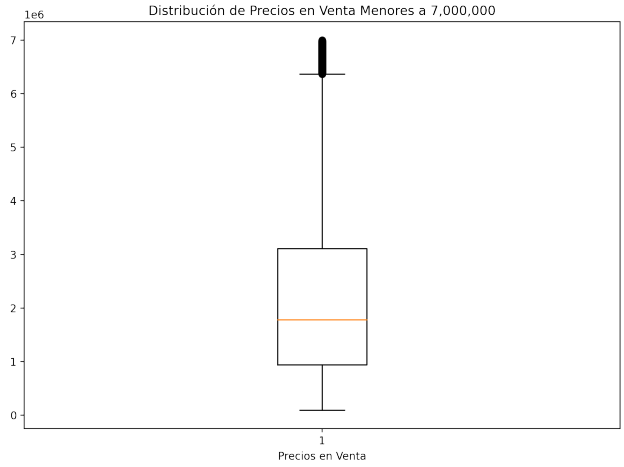
Suma de Precios por Municipio

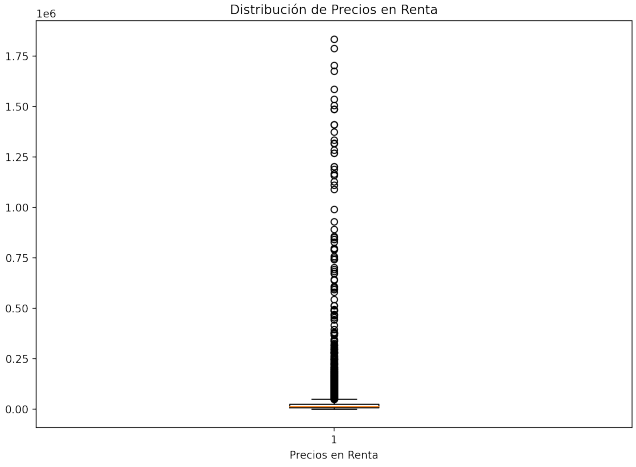


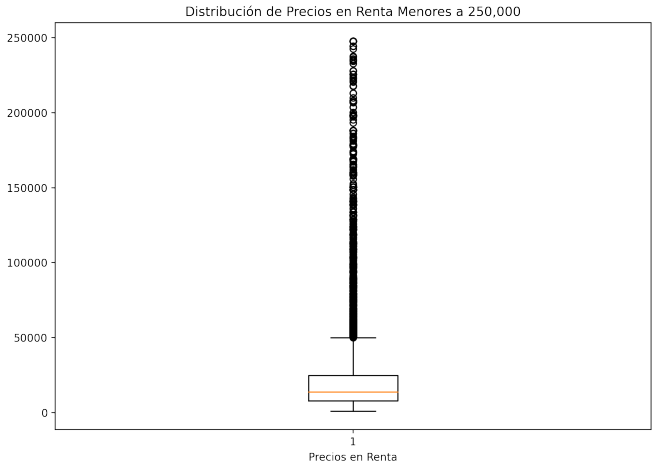
Publicaciones por Operacion

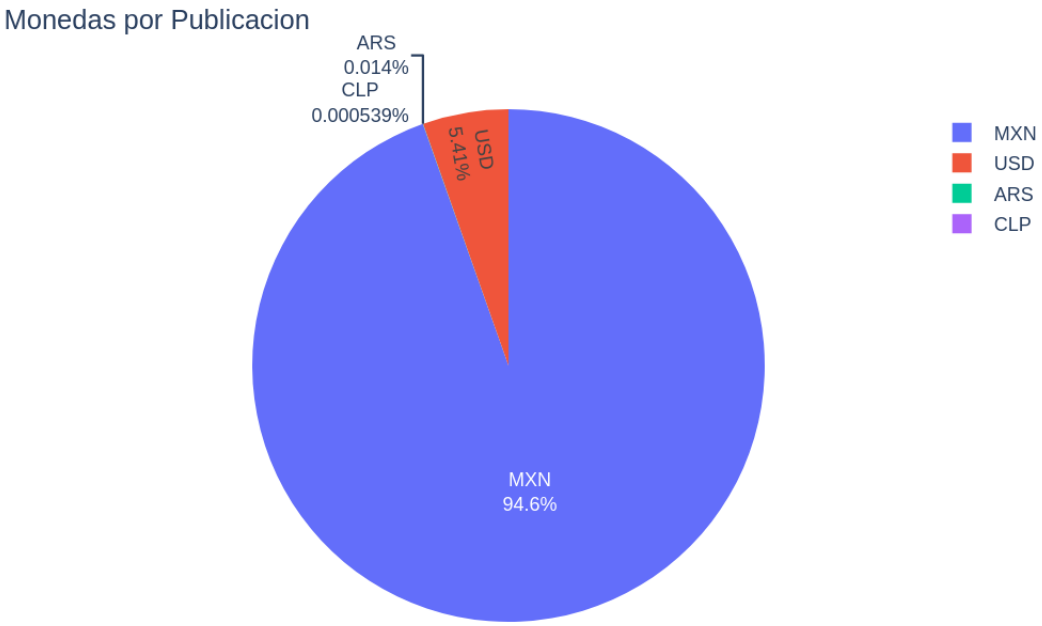



Distribución de Precios en Venta
1e8
8
6
4
2
0
1
Precios en Venta


Distribución de Precios en Venta Menores a 7,000,000
1e6
7
6
5
4
3
2
1
0
1
Precios en Venta


Distribución de Precios en Renta
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
Precios en Renta


Distribución de Precios en Renta Menores a 250,000
250000
200000
150000
100000
50000
0
1
Precios en Renta


Monedas por Publicacion
ARS
0.014%
CLP
0.000539%
USD
5.41%
MXN
94.6%
MXN
USD
ARS
CLP