



Facultad de Estudios Superiores

Acatlán

Proyecto Final – Módulo I

Análisis del Mercado Inmobiliario en México de 2013 a
2016

Cardoso Olvera Edgar David

Diplomado en Ciencia de Datos

Noviembre 2020

Índice general

Índice general	2
1. Introducción	4
1.1. Objetivo	4
1.2. Metodología del Análisis	4
1.3. Fuente de Datos	4
2. Análisis Exploratorio de Datos	5
2.1. Diccionario de Datos	5
2.2. Tipos de Variables	6
2.3. Completitud	6
2.4. Estadística Descriptiva	7
2.5. Visualización de Datos	7
2.5.1. Geográfica	8
2.5.2. Estados	9
2.5.3. Municipios	11
2.5.4. Venta y Renta	13
2.5.5. Moneda	16
2.5.6. Tipo de Inmueble	17
2.5.7. Fechas	18
2.5.8. Superficie	21
2.5.9. Análisis de Texto	23
3. Tratamiento y Limpieza de Variables	24
3.1. Valores Extremos Outliers	24
3.1.1. Outliers Variables Discretas	24
3.1.2. Outliers Variables Continuas	25
3.2. Valores Ausentes	26
3.3. Visualización de Datos Limpios	29
3.3.1. Geográfica	29
3.3.2. Estados	30
3.3.3. Municipios	31
3.3.4. Venta v.s. Renta	32
3.3.5. Tipo de Inmueble	34
3.3.6. Fechas de Publicación	35
3.3.7. Superficie	37
4. Ingeniería de Variables Feature Engineering	38
4.0.1. Variables Fecha	38
4.0.2. Variables de Tipo de Operación	38
4.0.3. Variables de Tipo de Propiedad	39
4.0.4. Variables Latitud y Longitud	39
4.0.5. Variable Título	39

4.0.6. Variable Descripción	39
4.0.7. Variable Estado	39
4.0.8. Variable Estado	40
5. Reducción de Dimensiones	41
5.1. Selección de Características	41
5.2. Análisis de Componentes Principales PCA	42
6. Conjunto de Entrenamiento y Prueba	43
7. Modelación Supervisada	44
7.1. Objetivo	44
7.2. Procesamiento Antes de Modelación	44
7.3. Selección de Variables de Modelación Venta	45
7.4. Resultados de Modelación en Venta, Modelos de Regresión	45
7.5. Selección de Variables de Modelación Renta	46
7.6. Resultados de Modelación en Renta, Modelos de Regresión	46
7.7. Resultados de Modelación Venta y Renta, Modelos de Regresión	47
7.8. Prueba Real	48
8. Modelación No Supervisada	49
8.1. Objetivo	49
8.2. Procesamiento Antes de Modelación	49
8.3. Visualización con Reducción de Dimensiones	49
8.3.1. Análisis de Venta	50
8.3.2. Análisis de Renta	51
8.4. Perfilamiento de Inmuebles	51
8.4.1. Análisis de Venta	51
8.4.2. Análisis de Renta	53
9. Modelo Redes Neuronales	56
9.1. Procesamiento de Variables Venta	56
9.2. Resultados de Modelación en Venta, Redes Neuronales	57
9.3. Procesamiento de Variables Renta	58
9.4. Resultados de Modelación en Renta, Redes Neuronales	59
9.5. Resultados de Modelación Venta y Renta, Redes Neuronales	60
10. Conclusiones	62
11. Apéndice	63
11.1. Gráficos Adicionales	63

Capítulo 1

Introducción

Una de las necesidades 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 dañó 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 representó 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 qué 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 facilitarán la comprensión 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 qué 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 Operación: 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	Ubicación con País, Estado, Municipio, Colonia
geonames_id	String	ID Geográfico
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.

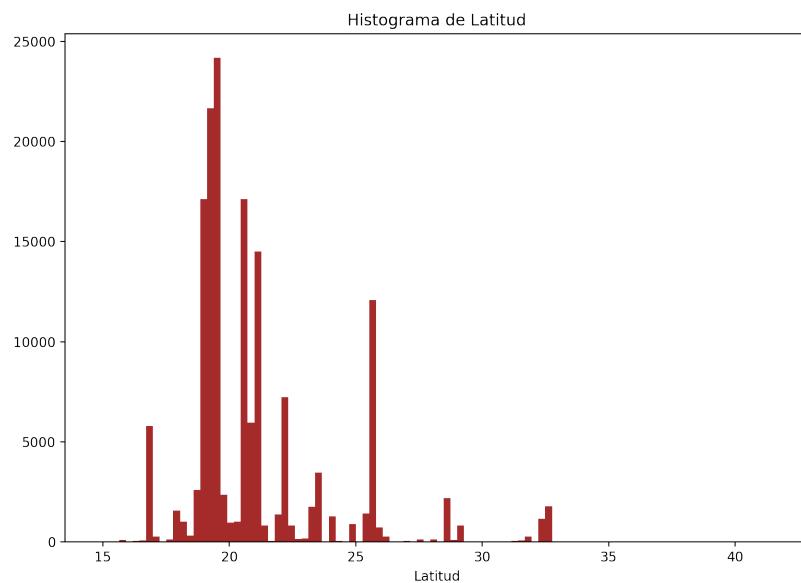
La variable superficie total será analizada detalladamente, ya que en el primer percentil la longitud total es de 0 metros cuadrados.

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.

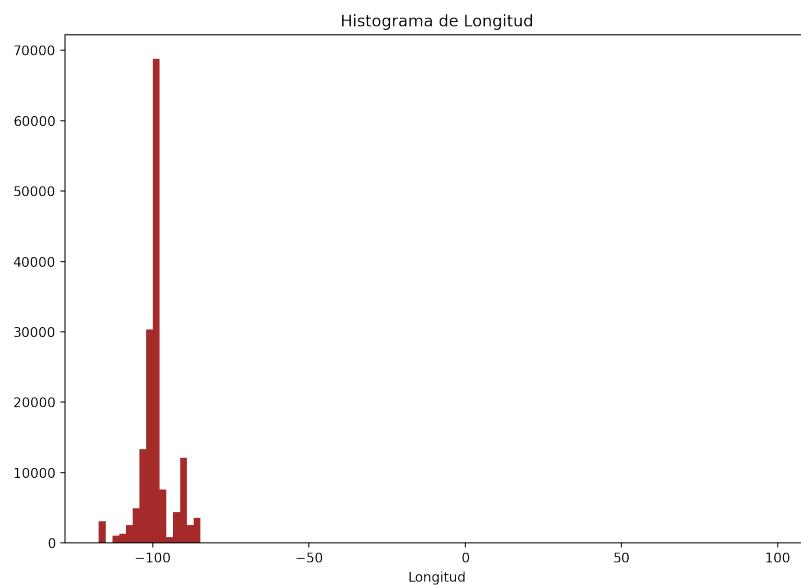
2.5.1. Geográfica

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

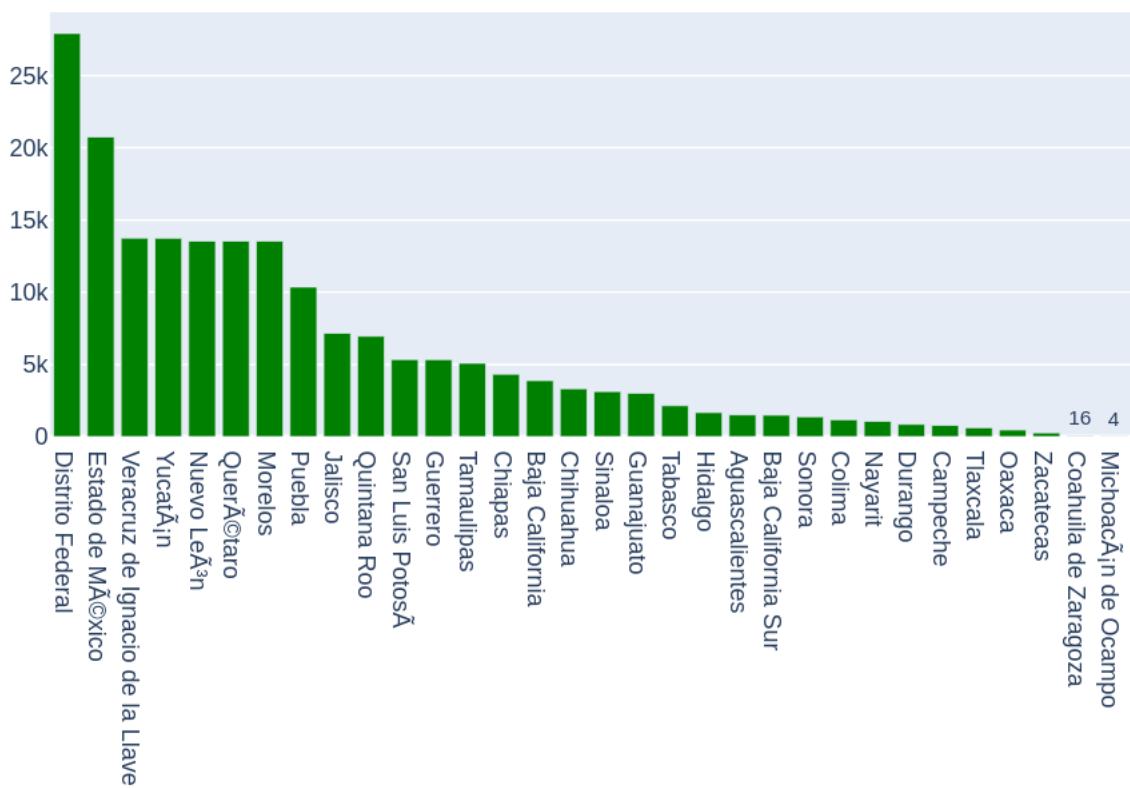


De igual manera notamos que el rango de longitudes es muy amplio y hay valores que están en una longitud mayor a 90, lo que indica la existencia de valores extremos en esta variable.

2.5.2. Estados

Publicaciones

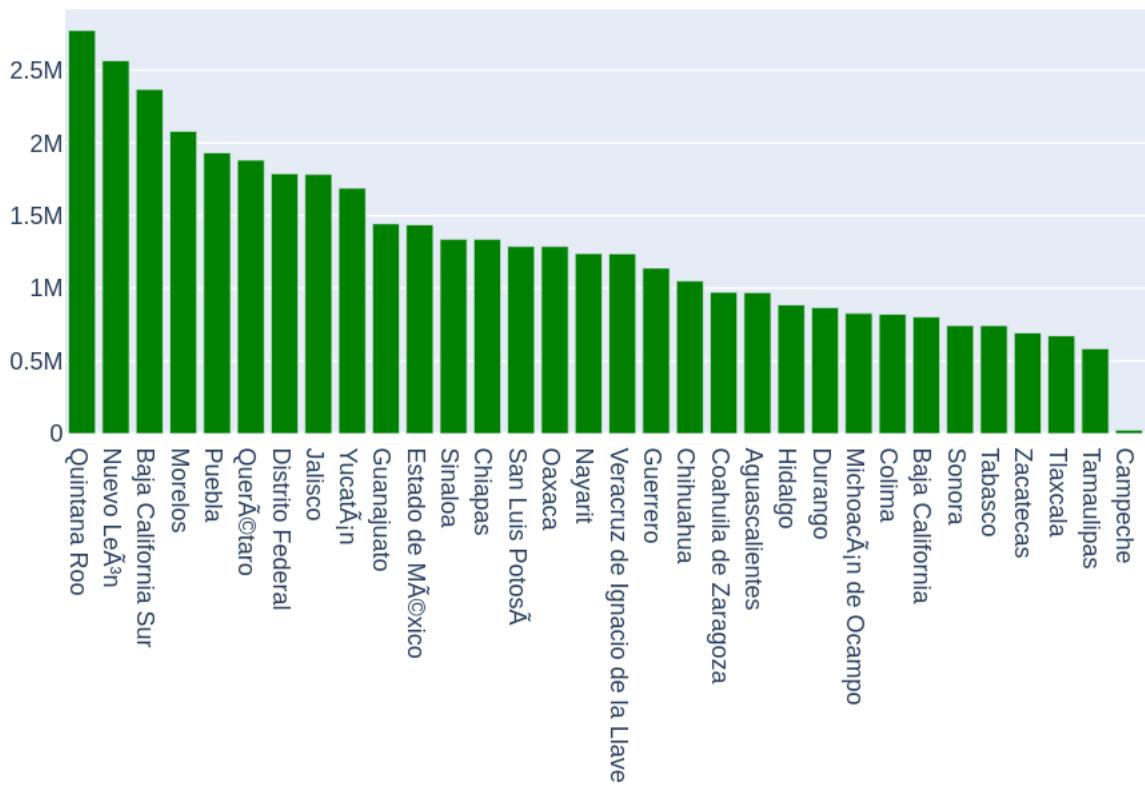
Número de Publicaciones por Estado



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

Precios

Mediana de Precios por Estado

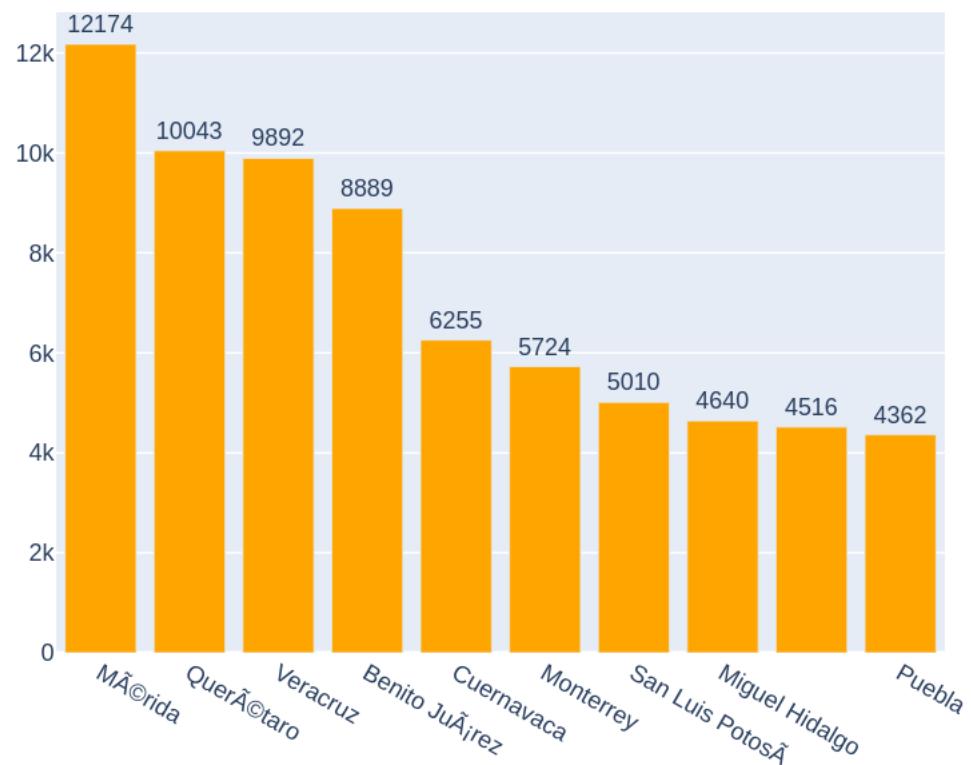


Observamos que los lugares con una mayor mediana de precios son estados donde su mayor ingreso económico es el turismo. El incremento en precios puede ser debido a una cercanía a sitios vacacionales con un gran flujo de visitantes. De igual manera observamos que estos estados son de clima caliente, lo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

2.5.3. Municipios

Publicaciones

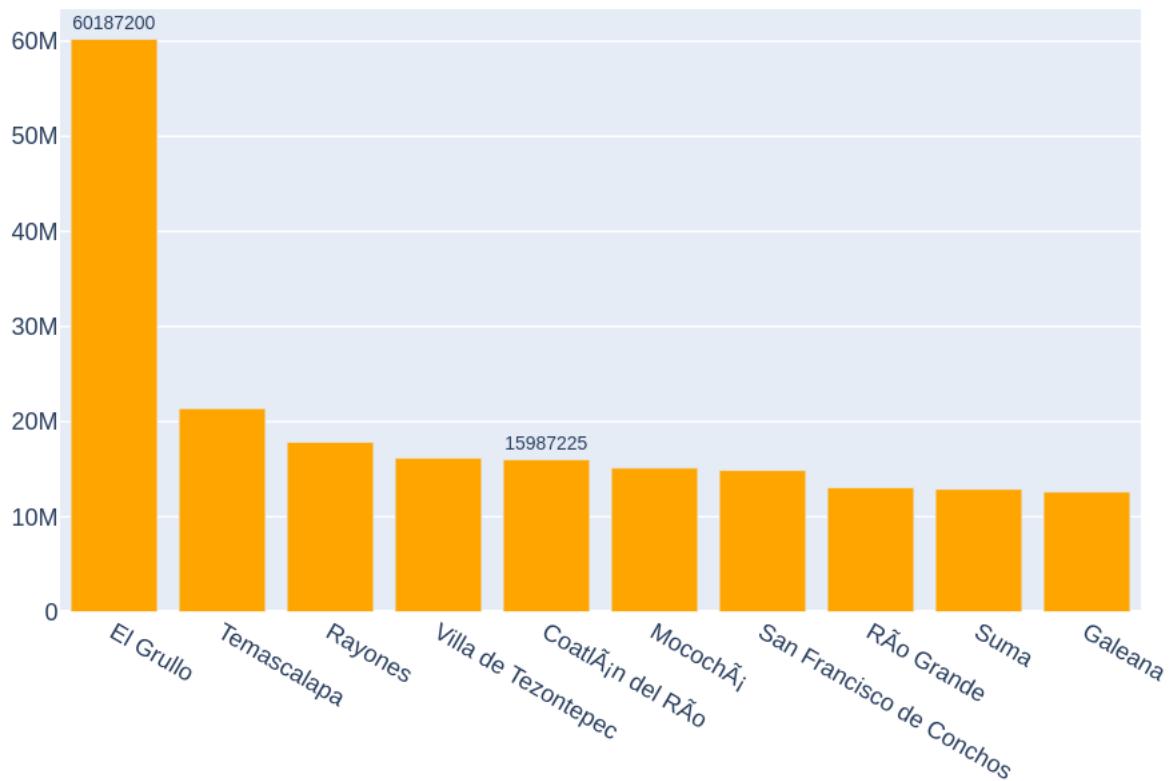
10 Municipios con mas Publicaciones



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

Mediana de Precios por Municipio

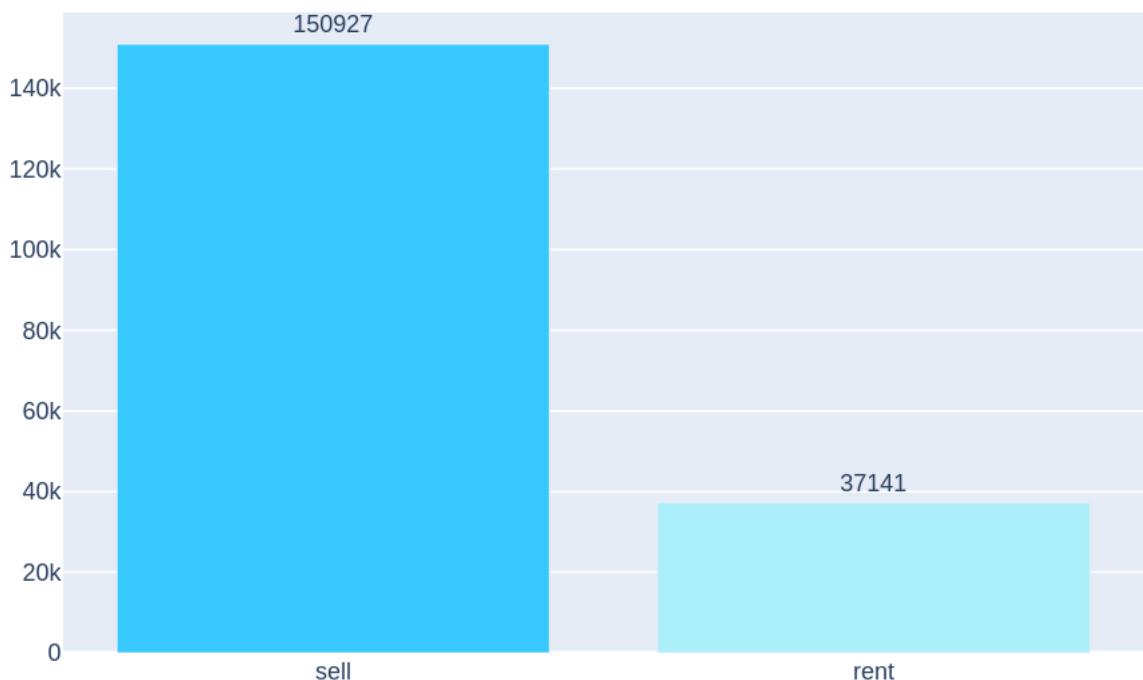


Observamos que los municipio que en mediana tienen un precio mayor son municipios no tan conocidos, esto puede ser debido a un error en el precio o dichos municipios contienen pocas publicaciones de inmuebles con valores muy altos.

2.5.4. Venta y Renta

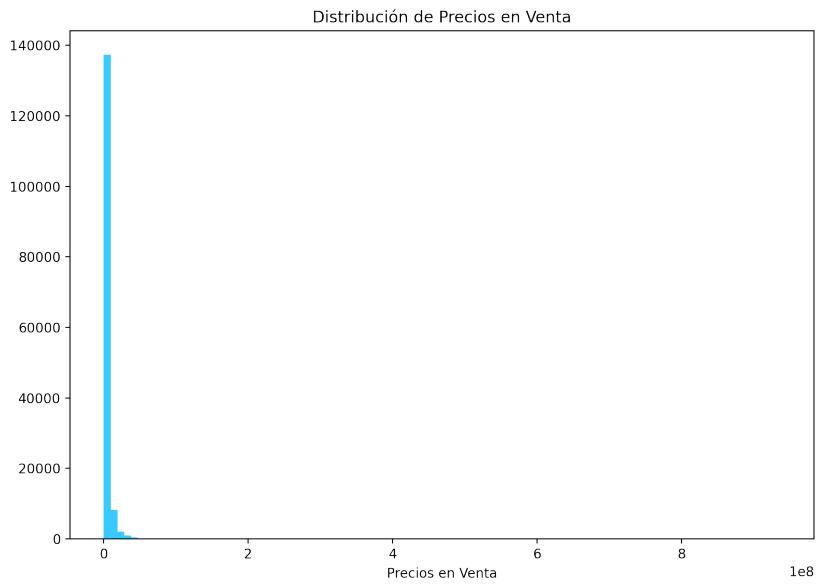
Publicaciones

Publicaciones por Operacion



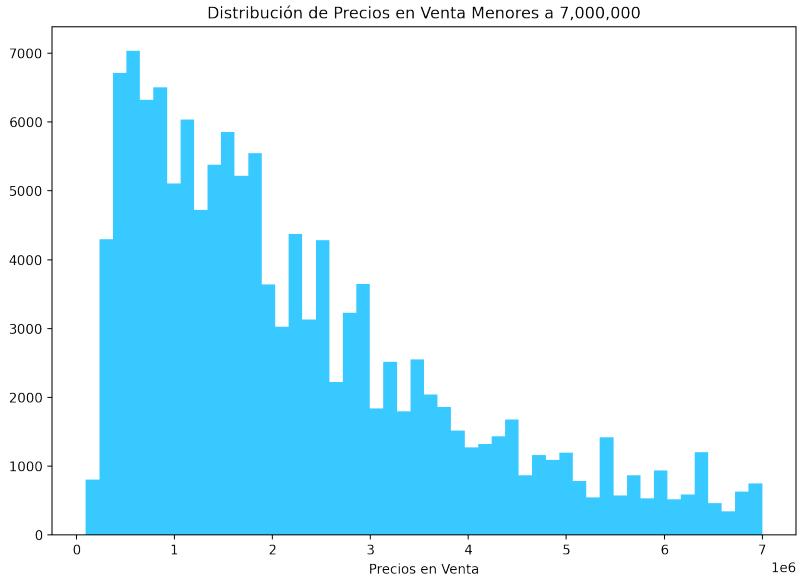
Existe una mayor cantidad de inmuebles en venta dentro de todo el conjunto de datos.

Precio de Venta



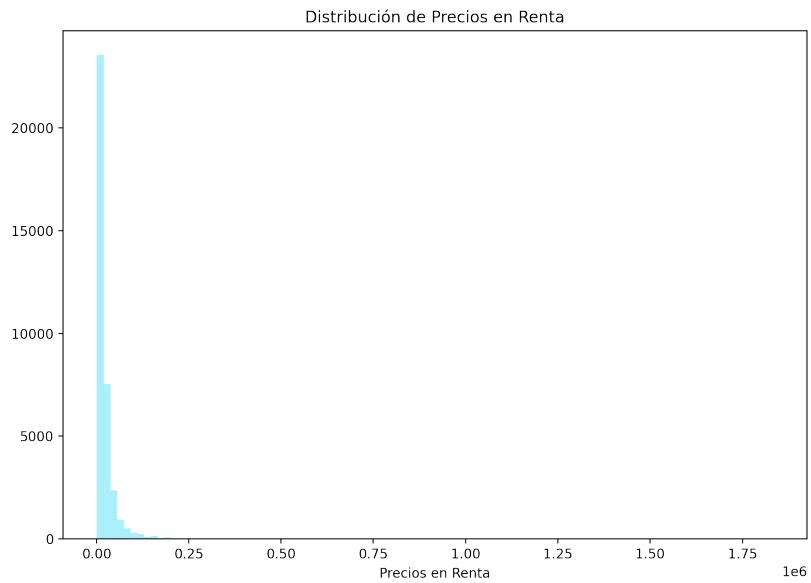
Observamos que la distribución de precios de venta se concentra a la izquierda de la gráfica, lo que nos da a entender que existen valores extremos que deben ser analizados, ya que la gráfica indica que existen inmuebles con valores mayores a los 900 millones de pesos.

Precio de Venta Menor a 7 Millones de Pesos



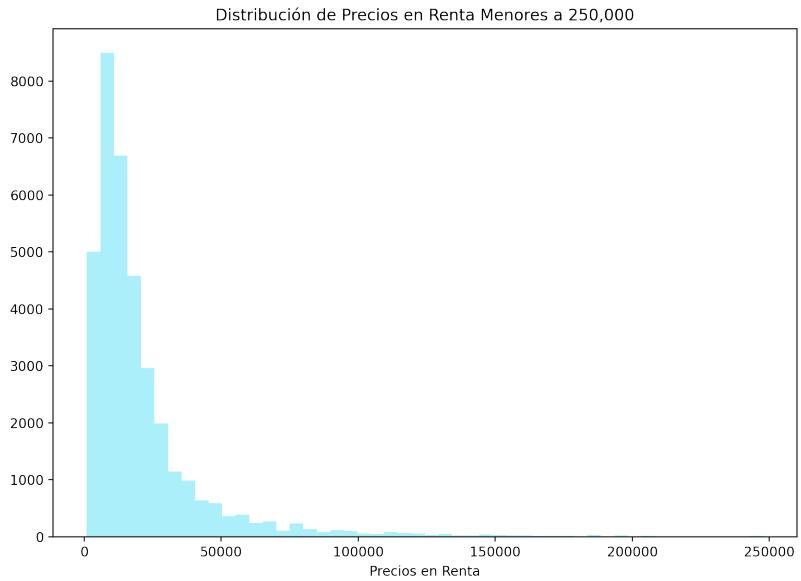
Al limitar las publicaciones por un precio menor a los 7 millones de pesos, que representa el 90 % de los precios según el análisis descriptivo hecho anteriormente, observamos una distribución más realista del mercado inmobiliario mexicano.

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

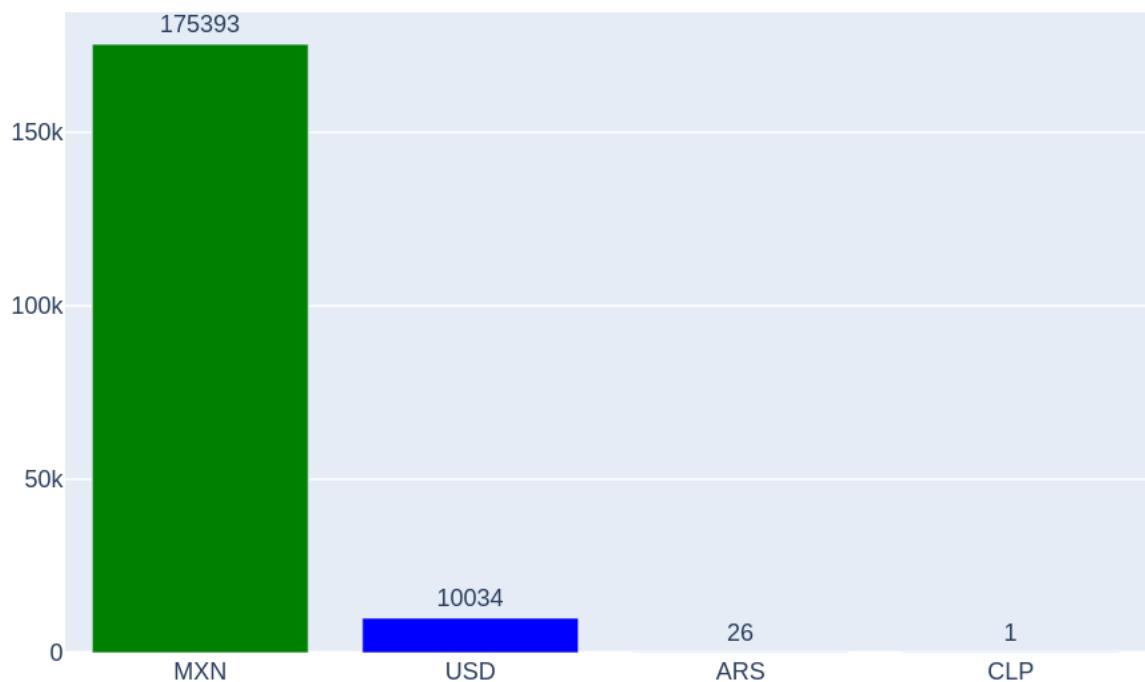


Limitando las publicaciones menores de 250 mil pesos de renta observamos nuevamente una distribución más realista del mercado mexicano.

2.5.5. Moneda

Publicaciones por Moneda

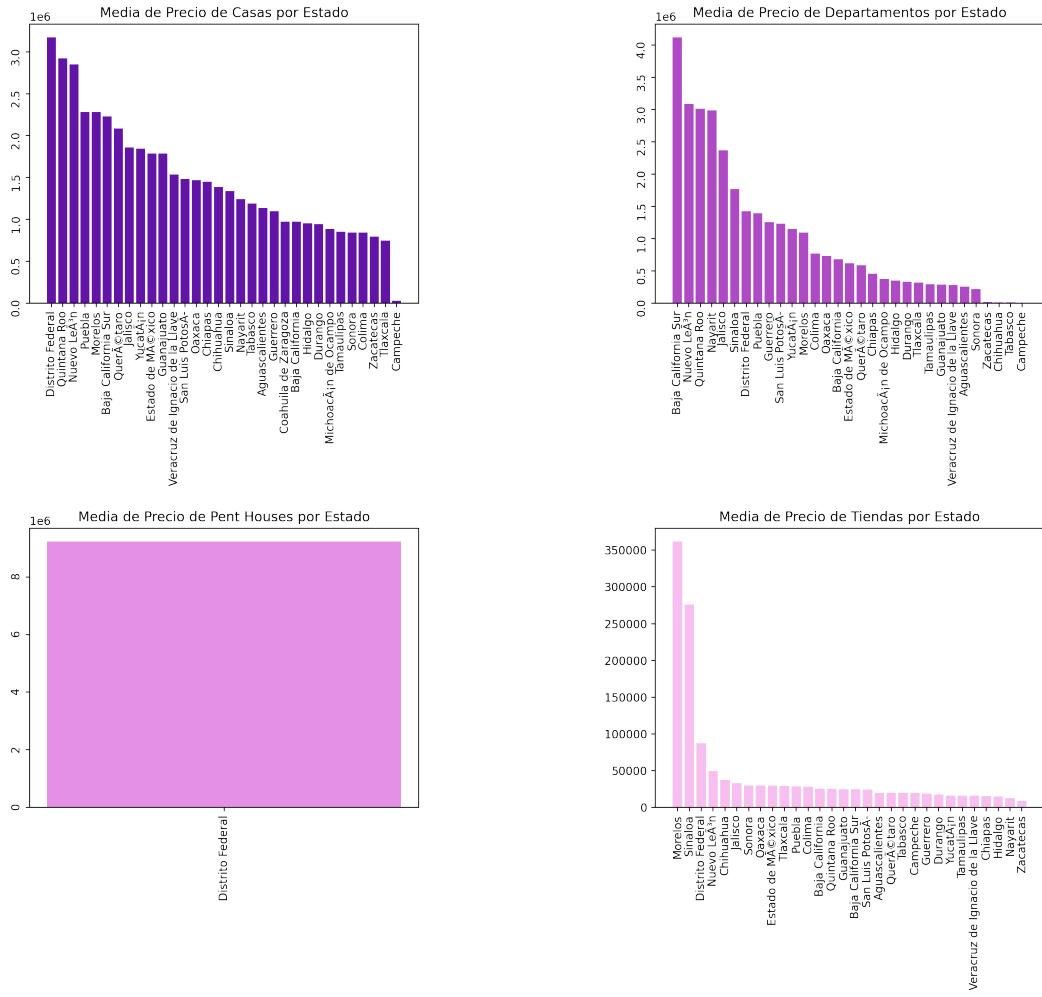
Monedas por Publicacion



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 Inmuble

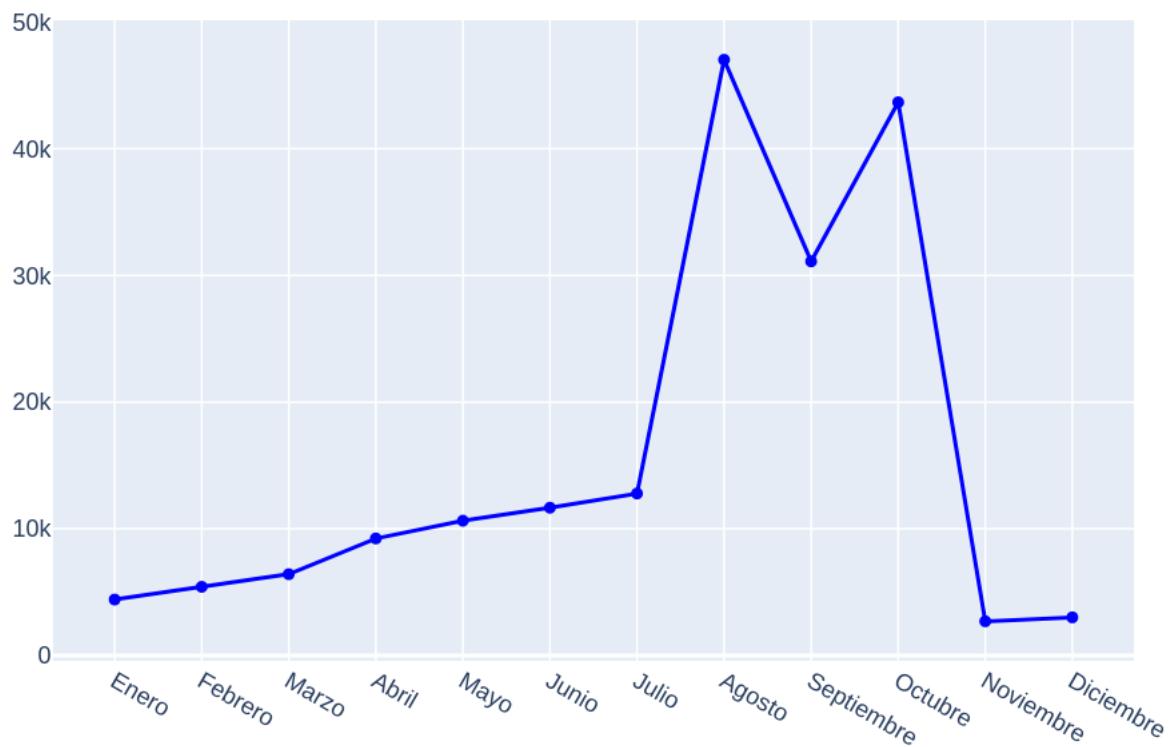


Observamos que las casas y departamentos tienen una distribución parecida en los precios, con una reducción de precios en los departamentos. No existe una distribución de precios en el mercado de Pent House, ya que solo existen publicaciones en la Ciudad de México. Por último los precios en tiendas son 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

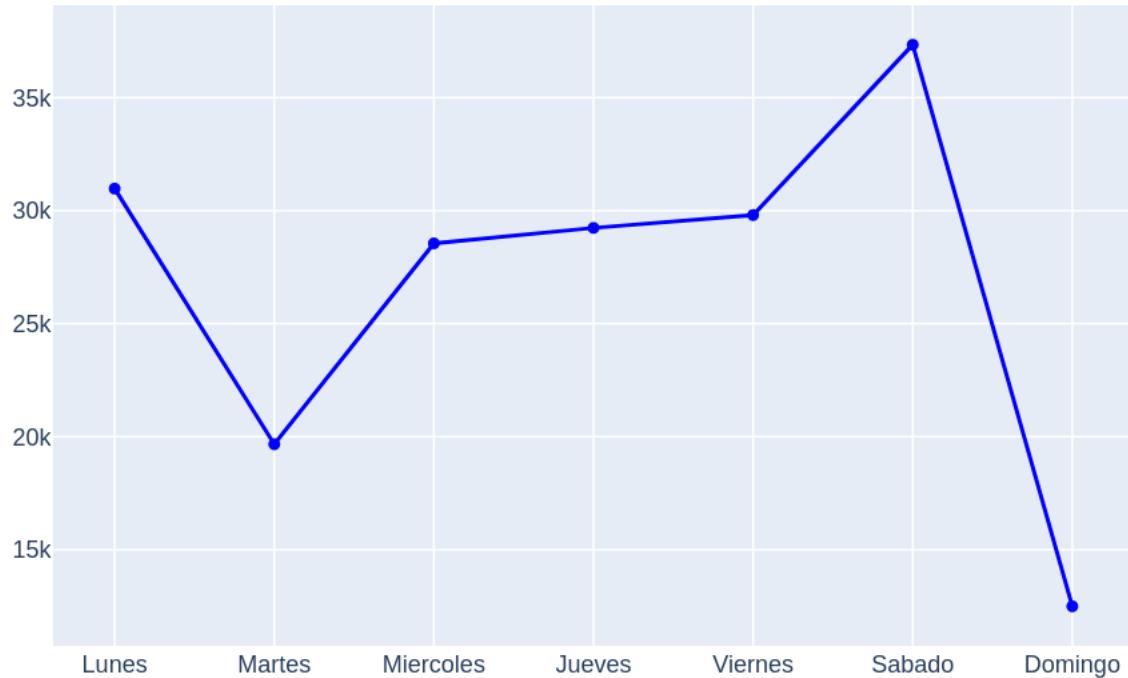
Conteo de Publicaciones por Mes



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

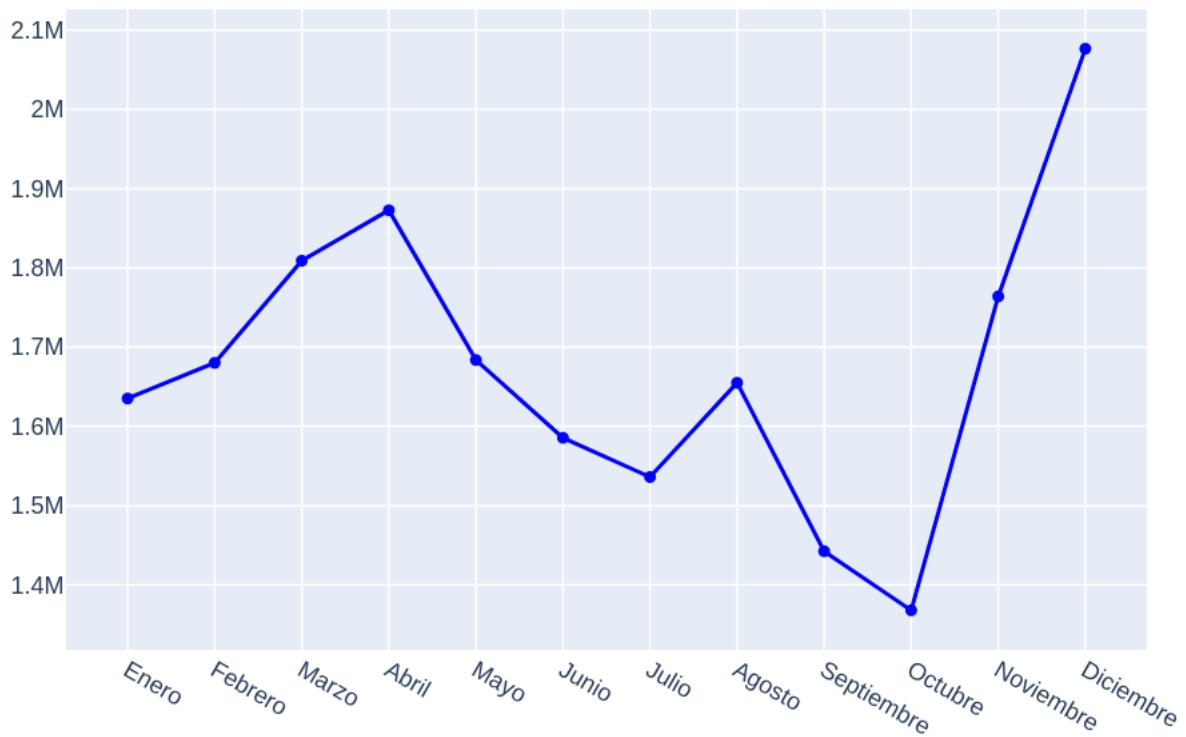
Conteo de Publicaciones por Día



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

Mediana de Precios por Mes

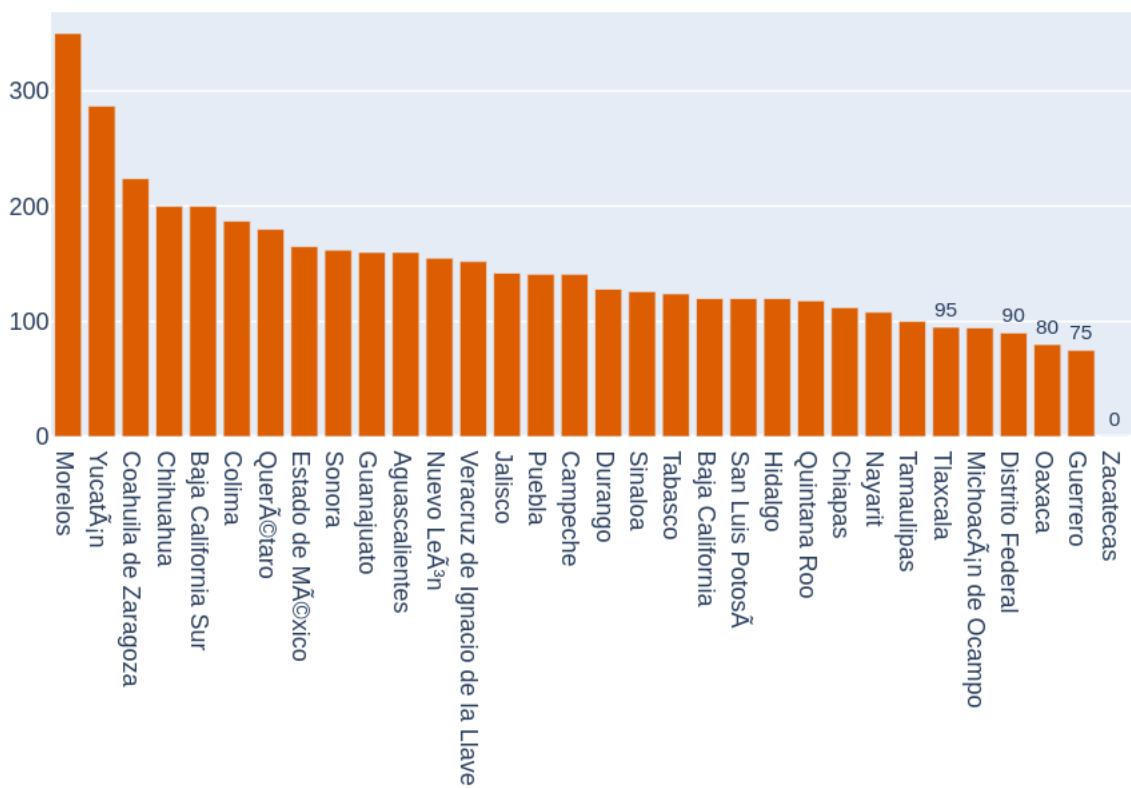


A comparación del aumento de publicaciones en los meses de Agosto a Octubre, los precios bajan en dichos meses, con un valor mínimo en el mes de Octubre. Este dato debe ser comparado con las tasas de créditos hipotecarios o la cantidad de préstamos que se realizan en dicho mes, para comprender porque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

2.5.8. Superficie

Estado

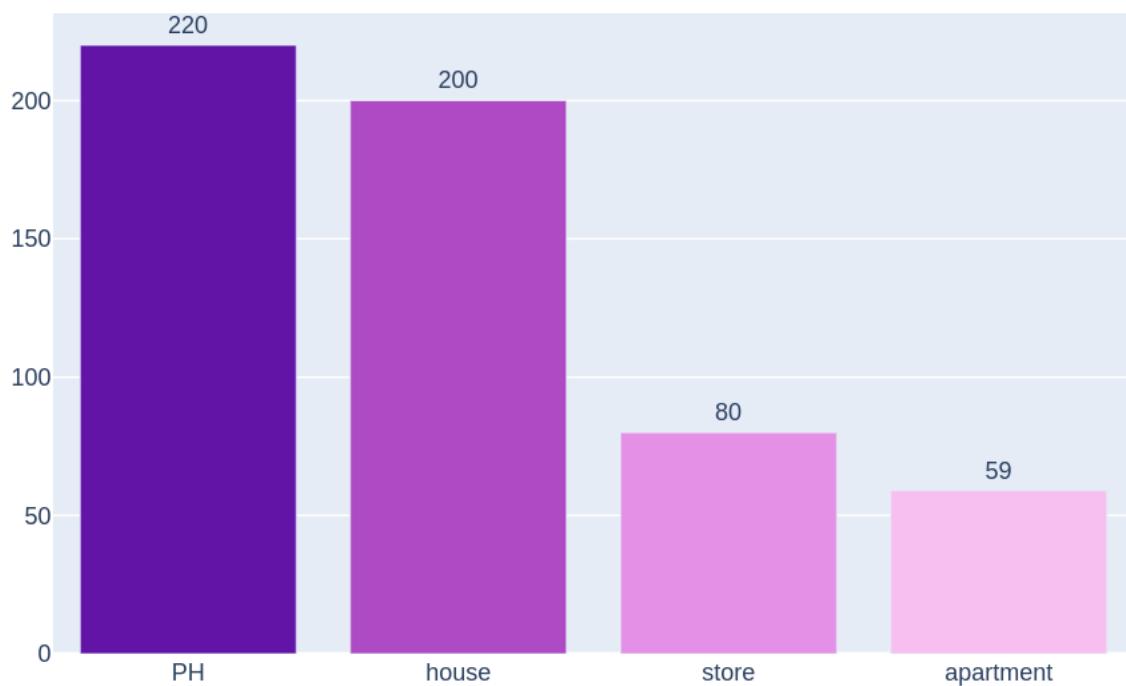
Mediana de Superficien por Estado en m²



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

Mediana de Superficie por Tipo de Inmueble en m²



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 renta.

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_operati_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_operati_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.

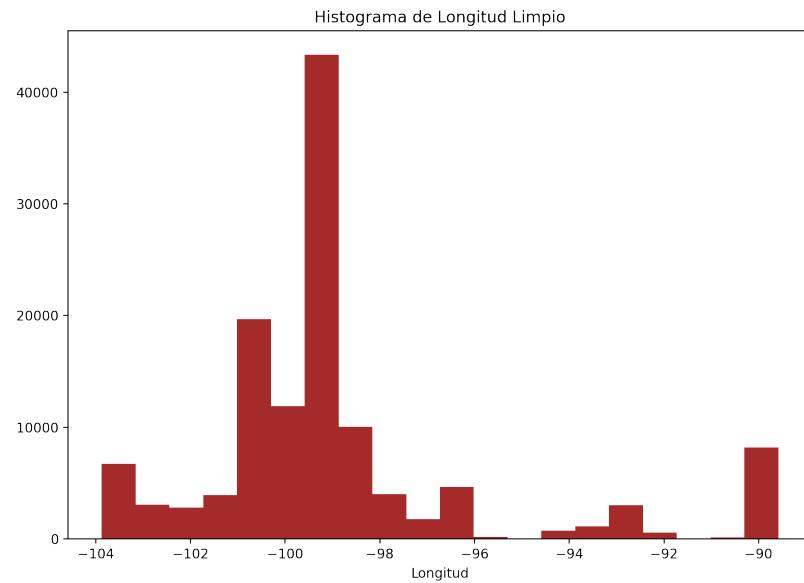
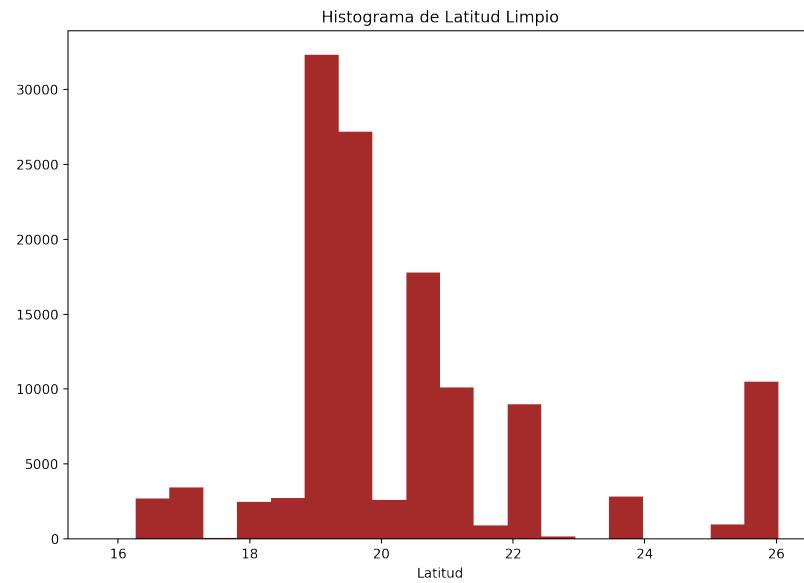
Al ser t_image_thumbnail una variable que es difícil de usar en el análisis estadístico será eliminada de ambas tablas. De igual manera la variable v_colonia será eliminada de ambas tablas por la alta cantidad de valores nulos que presentan en las dos tablas de datos.

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

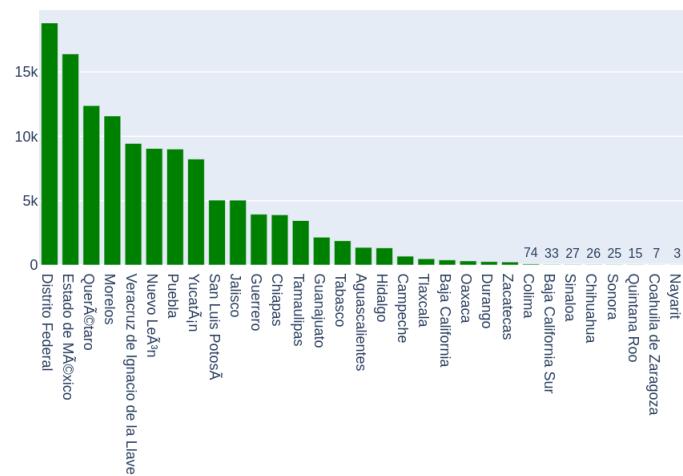


Observamos un gran cambio en el rango de latitudes, esto nos indica que todos los registros con latitud mayor a 26 eran erróneos o estaban muy alejados de los datos normales. También vemos una concentración de los datos en un rango de -104 a -90 en las longitudes. En los datos originales el rango llegaba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

3.3.2. Estados

Cantidad de Publicaciones

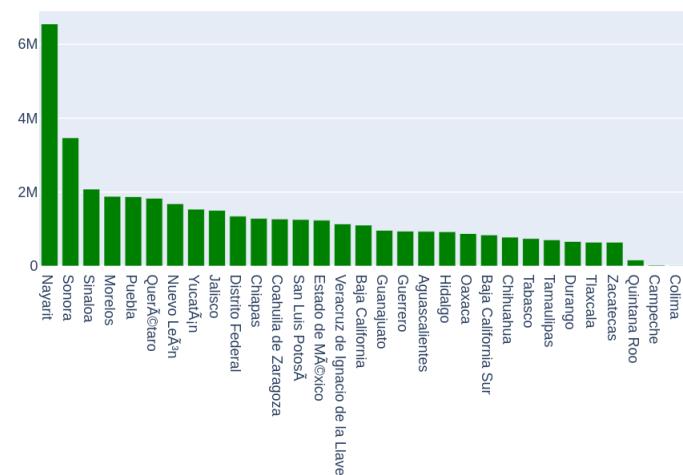
Número de Publicaciones por Estado Limpios



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

Mediana de Precios por Estado Limpio



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 a cercanía a zonas turísticas.

3.3.3. Municipios

Cantidad de Publicaciones

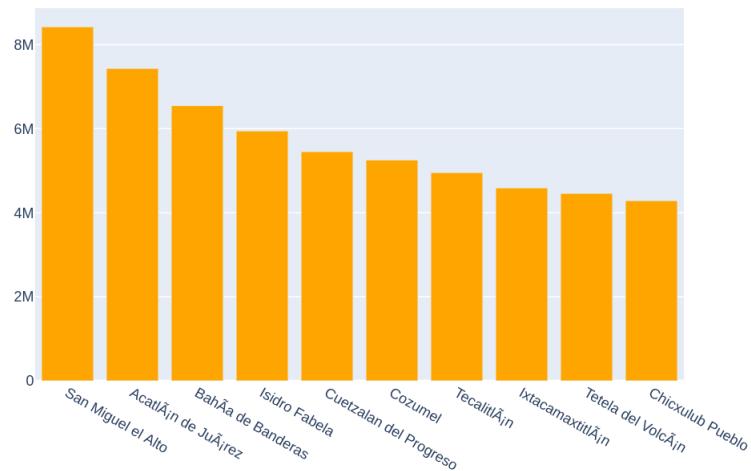
10 Municipios con mas Publicaciones Limpio



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

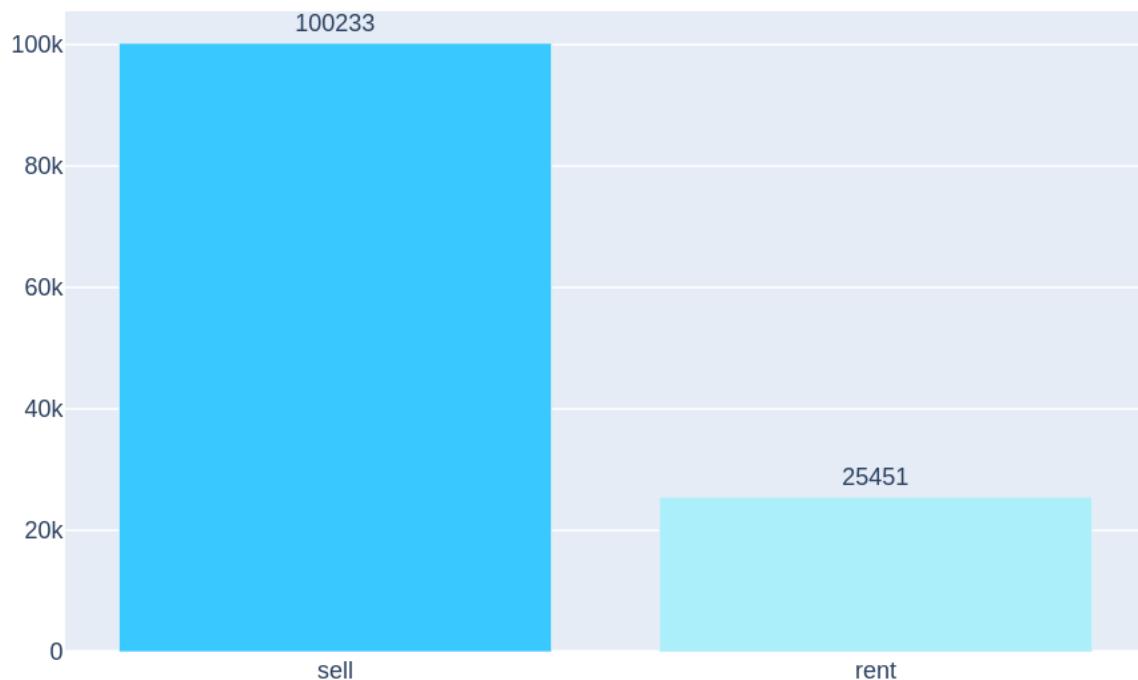
Mediana de Precios por Municipio Limpio



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.

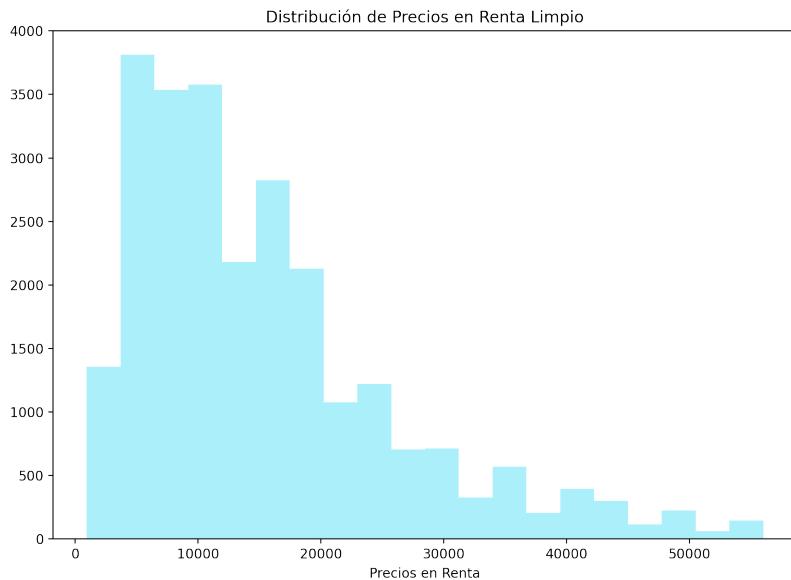
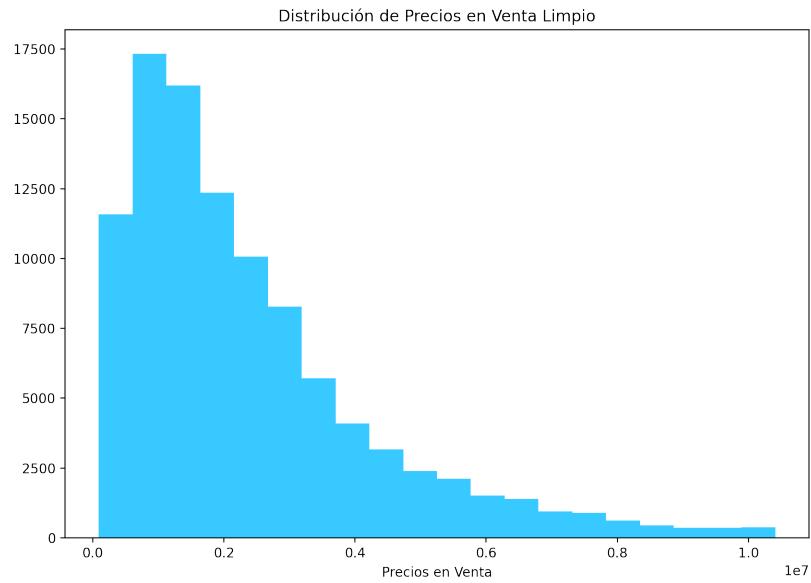
3.3.4. Venta v.s. Renta

Publicaciones por Operacion Limpio



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.

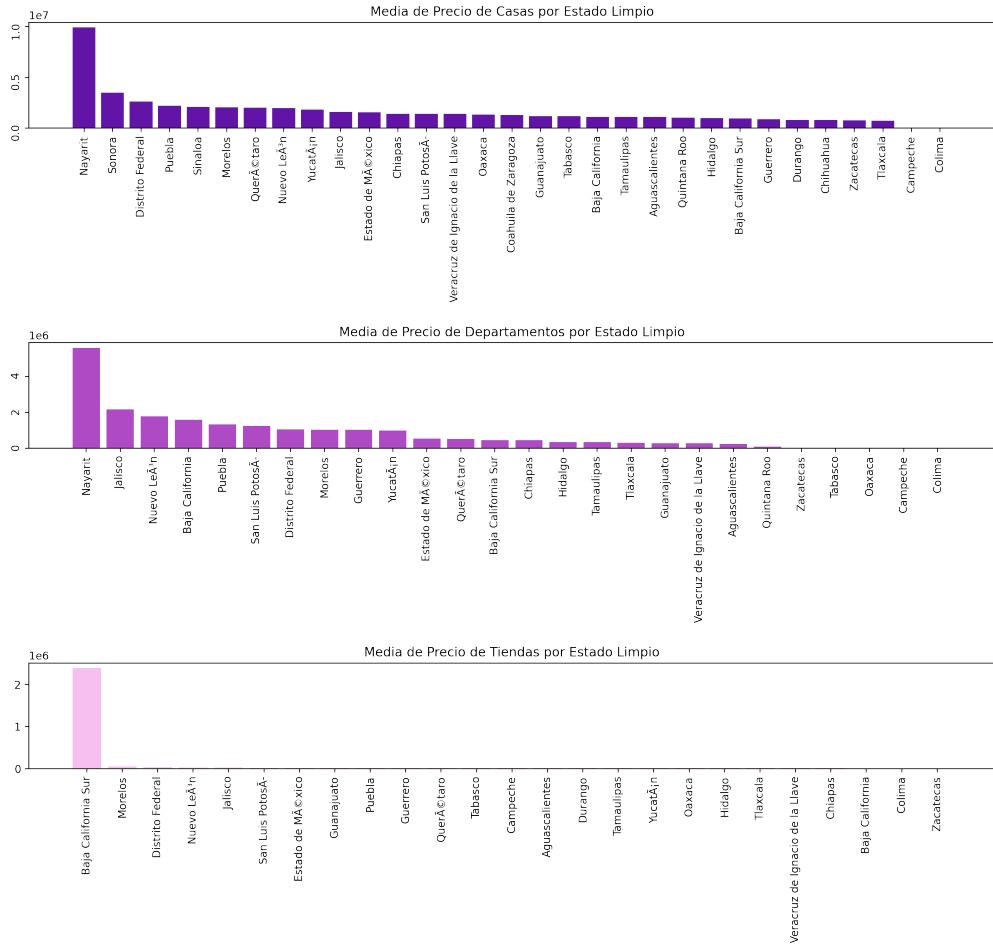
Distribución de Precios Venta y Renta



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

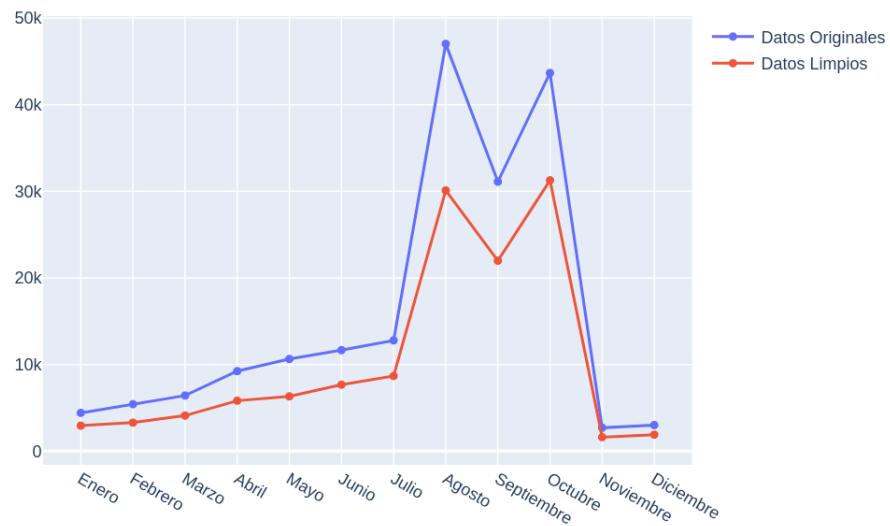


Vemos un cambio en el rango de precios en los tres tipos de inmueble, de igual manera observamos que la mediana de precios en Nayarit es muy alta, esto debido a la poca cantidad de datos con los que se cuenta. Vemos que los Pent House se eliminaron ya que se consideran datos extremos al sólo existir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

3.3.6. Fechas de Publicación

Publicaciones por Mes

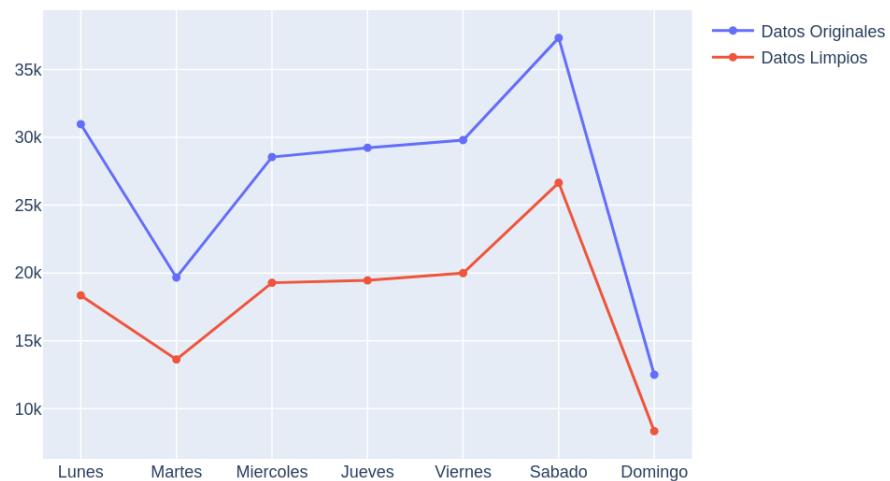
Conteo de Publicaciones por Mes



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

Conteo de 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

Mediana en Precio por Mes

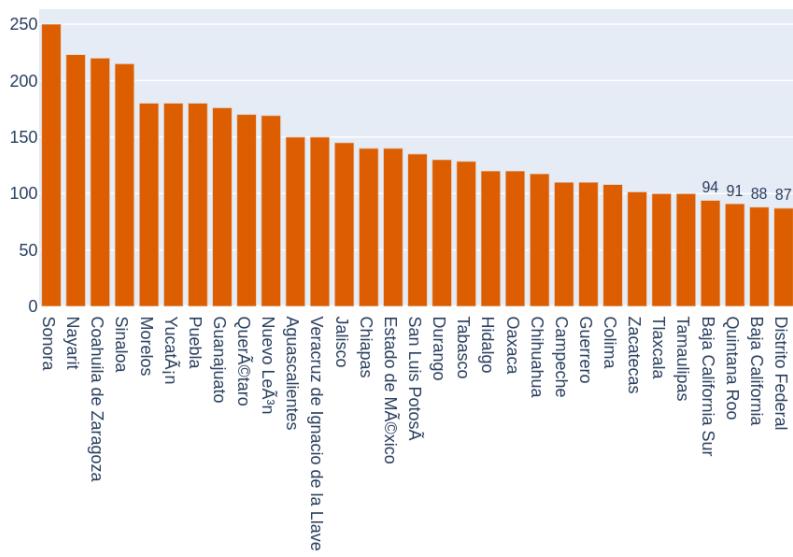


Por último, notamos que en los datos originales no se puede apreciar de buena manera la reducción de precios entre los meses de abril y octubre, pero con los datos limpios notamos que los precios de los inmuebles bajan en estos meses teniendo un mínimo en el mes de octubre.

3.3.7. Superficie

Superficie por Estado

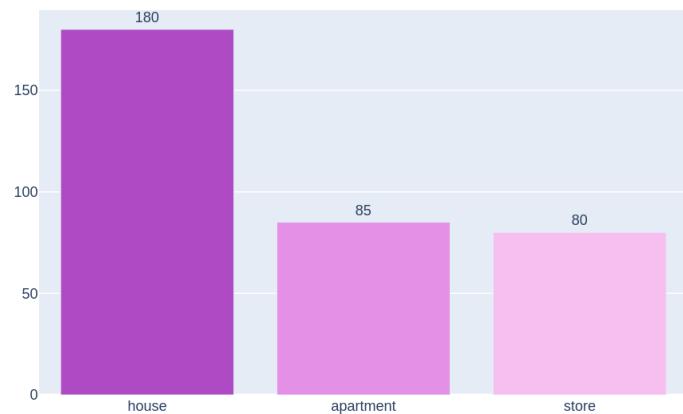
Mediana de Superficien por Estado en m² Limpio



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

Mediana de Superficie por Tipo de Inmueble en m² Limpio



Vemos que se elimina los Pent House e incrementan la superficie en cada uno 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_baa0	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_tamano_grande	v_tamano_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 sí, 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 sí, 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 modelación 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.

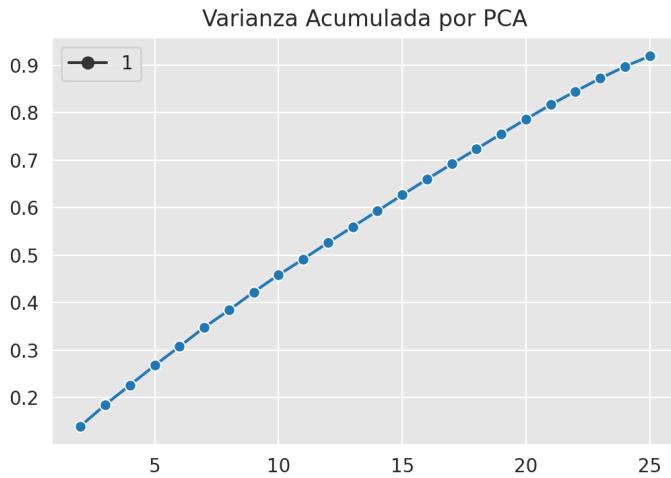
Variables altamente relacionadas con otras	Variables de poca relevancia
v_region_sur	c_lat
v_tamanio_grande	c_lon
v_tamanio_mediano	v_anio_2015
title_renta	v_anio_2016
title_venta	v_mes_8
title_casa	v_mes_10
title_departamento	descr_id
title_local	descr_nocnok
descr_comedor	descr_sistema
descr_sala	

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

Modelación Supervisada

7.1. Objetivo

El objetivo principal de esta sección es encontrar un algoritmo o modelo de Machine Learning que puede predecir, con el mínimo error y máxima generalización, el precio de inmuebles en venta y renta dentro de la República Mexicana entre los años 2013 y 2016. Estos modelos se entrenarán con la información procesada en la sección anterior, para así lograr la máxima generalización en nuestro modelo.

7.2. Procesamiento Antes de Modelación

Con la intención de mejorar los resultados en la modelación, se agregaron variables de enriquecimiento que podrán contribuir a un mejor entendimiento del problema.

Las primeras características que se agregaron fueron las variables indicadoras por estados, si el inmueble se encuentra en un estados se agregará un 1 a dicha variable, esto con el motivo de ayudar al modelo a conocer la ubicación de la propiedad.

Las segundas características fueron los datos demográficos y económicos de cada estado. Lo anterior fue hecho para agregar información sobre cada estado de manera separada. Se agregaron las variables de superficie, población por estado, además de la tasa hipotecaria promedio en el año de venta. A partir de las variables anteriores se realizó ingeniería de variables para calcular la densidad de población por metro cuadrado. Así terminando con una tabla analítica de 84 variables.

En la siguiente tabla se pueden observar las nuevas variables y algunos ejemplos en ellas.

c_lat	c_lon	c_poblacion	...	c_densidad	c_tasa	v_conteo_geo
20.712125	-101.353925	5853677	...	191.246635	0.102192	1
18.899422	-99.182983	4584471	...	78.234629	0.102192	42
20.593513	-100.352842	6168883	...	179.819361	0.103133	1

Otro punto importante a considerar son los tipos de operación que existen en los inmuebles (venta y renta), esto afecta a la generalización del modelo debido a la diferencia significativa de precios en inmuebles con características similares, se decidió dividir la tabla analítica en dos conjuntos, venta y renta, al finalizar la modelación y selección se juntaran en una función final para poder predecir de forma correcta y sin ningún sesgo por operación. El conjunto de venta cuenta con 100,233 registros y renta cuenta con 25,451 registros.

7.3. Selección de Variables de Modelación Venta

Al tener una gran cantidad de variables se tomaron solo las 53 variables más significativas para la predicción de precios de inmuebles en venta, en el siguiente cuadro se observan las variables seleccionadas.

'c_lat'	'descr_comedor'
'c_lon'	'descr_id'
'c_surface_covered_in_m2'	'descr_nochnok'
'c_poblacion'	'descr_recamaras'
'c_superficie'	'descr_sala'
'c_densidad'	'descr_sistema'
'v conteo_geo'	'v_region_norte'
'v_anio_2015'	'v_tamano_grande'
'v_anio_2016'	'v_tamano_mediano'
'v_mes_8'	'v_estado_Baja California'
'v_mes_9'	'v_estado_Baja California Sur'
'v_mes_10'	'v_estado_Chihuahua'
'v_dia_semana_Lunes'	'v_estado_Coahuila de Zaragoza'
'v_dia_semana_Sábado'	'v_estado_Colima'
'v_property_type_house'	'v_estado_Distrito Federal'
'v_property_type_store'	'v_estado_Durango'
'title_casa'	'v_estado_Estado de México'
'title_departamento'	'v_estado_Hidalgo'
'title_local'	'v_estado_Jalisco'
'title_lomas'	'v_estado_Nayarit'
'title_residencial'	'v_estado_Nuevo León'
'title_san'	'v_estado_Puebla'
'title_venta'	'v_estado_Querétaro'
'descr_baa'	'v_estado_San Luis Potosí'
'descr_br'	'v_estado_Sinaloa'
'descr_casa'	'v_estado_Yucatán'
'descr_cocina'	

Observamos que la ubicación espacial, las nuevas variables demográficas, algunos meses y días de la semana especiales (en estos meses se tiene un aumento de precio y existe una mayor cantidad de publicaciones), algunas características en la descripción y los estados económicamente más activos, son las características más significativas para predecir el precio de venta de un inmuebles.

7.4. Resultados de Modelación en Venta, Modelos de Regresión

Para la modelación de este problema se eligió utilizar 3 algoritmos de Machine Learning y 4 modelos ensamble. En la siguiente tabla se observan las métricas y resultados en el conjunto de entrenamiento por cada modelo entrenados.

Modelo de Regresión	Métrica R2	MAE
Regresión Lineal	0.562	854,091.5
Ridge	0.562	854,083.3
Decision Tree	0.547	774,320.3
Red Neuronal	0.659	710,153.1
Random Forest	0.75	599,425.25
XGBoost	0.77	578,222.16
Voting Regressor (RandomForest, XGBoost)	0.769	572,575.66

Al entrenar 7 modelos y compararlos con métricas como R2 y MAE, es notable que los modelos con mejor rendimiento son los "modelos ensamblados" los modelos de redes neuronales. Para mejorar dicho performance y eliminar el overfitting se decidió crear un modelo ensamble Vote Regressor, este está conformado de un bosque aleatorio y un xgboost. El modelo de voto puede no tener las mejores métricas pero tiene la mejor generalización, es decir bajo overfitting, lo que ayudará a predecir de mejor manera el precio de inmuebles externos al conjunto de datos.

7.5. Selección de Variables de Modelación Renta

De igual manera que en la sección de venta, al tener una gran cantidad de variables se tomaron solo las 53 variables más significativas para la predicción de precios de inmuebles en renta, en el siguiente cuadro se observan las variables seleccionadas.

'c_lat'	'descr_br'
'c_lon'	'descr_casa'
'c_surface_covered_in_m2'	'descr_cocina'
'c_poblacion'	'descr_comedor'
'c_superficie'	'descr_id'
'c_densidad'	'descr_nocnok'
'c_tasa'	'descr_recamaras'
'v_anio_2015'	'descr_sala'
'v_anio_2016'	'descr_sistema'
'v_mes_2'	'v_region_norte'
'v_mes_4'	'v_tamanio_grande'
'v_mes_5'	'v_tamanio_mediano'
'v_mes_7'	'v_estado_Baja California'
'v_mes_10'	'v_estado_Colima'
'v_mes_11'	'v_estado_Distrito Federal'
'v_dia_semana_Lunes'	'v_estado_Durango'
'v_dia_semana_Miercoles'	'v_estado_Estado de México'
'v_property_type_house'	'v_estado_Guerrero'
'v_property_type_store'	'v_estado_Hidalgo'
'title_casa'	'v_estado_Jalisco'
'title_centro'	'v_estado_Nuevo León'
'title_departamento'	'v_estado_Puebla'
'title_local'	'v_estado_Querétaro'
'title_lomas'	'v_estado_San Luis Potosí'
'title_renta'	'v_estado_Veracruz de Ignacio de la Llave'
'title_san'	'v_estado_Yucatán'
'descr_baa'	

Observamos, de nuevo, que la ubicación espacial es un punto importante dentro de nuestras variables, de igual manera las nuevas características añadidas. Se observa una mayor cantidad de meses en la fecha de publicación y un cambio en los días de la semana. De la misma forma que en el análisis de venta observamos que los estados de la república con mayor participación económica son los más significativos para nuestro modelo.

7.6. Resultados de Modelación en Renta, Modelos de Regresión

Para la modelación de este problema se eligió utilizar 3 algoritmos de Machine Learning y 4 modelos ensamble. En la siguiente tabla se observan las métricas y resultados en el conjunto de entrenamiento por

cada modelo entrenados para renta.

Modelo de Regresión	Métrica R2	MAE
Regresión Lineal	0.419	5,930.13
Ridge	0.419	5,932.83
Decision Tree	0.436	5,447.92
Red Neuronal	0.556	5,013.16
Random Forest	0.594	4,834.75
XGBoost	0.634	4,504.10
Voting Regressor (RandomForest, XGBoost)	0.626	4,580.12

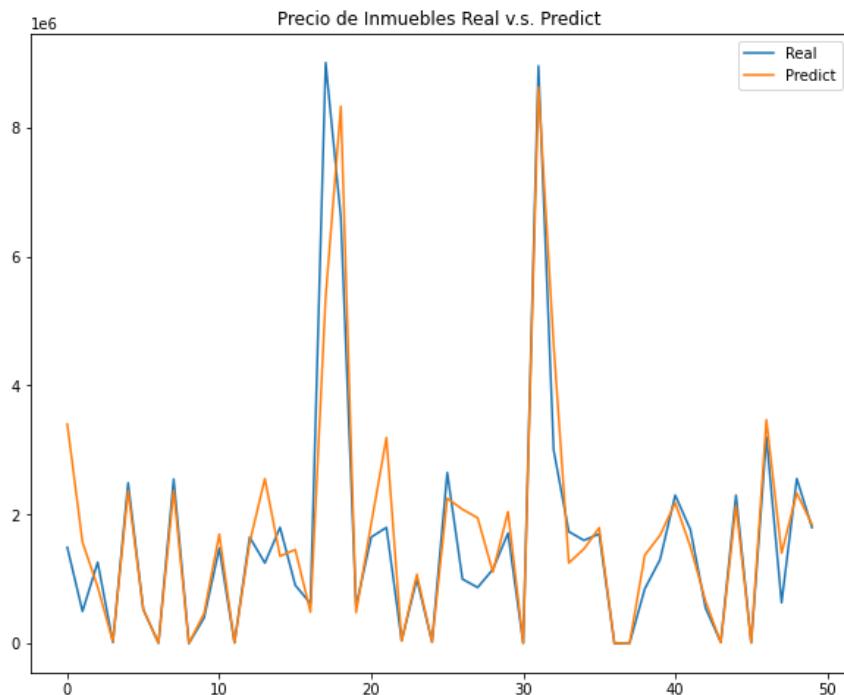
Al igual que en los modelos de venta, se entrenaron 7 modelos y al compararlos con métricas como R2 y MAE, es notable que los modelos con mejor rendimiento son los "modelos ensamblaz" los modelos de redes neuronales. Para mejorar dicho performance y eliminar el overfitting se decidió crear un modelo ensamble Vote Regressor, este está conformado de un bosque aleatorio y un xgboost.

7.7. Resultados de Modelación Venta y Renta, Modelos de Regresión

Al tomar los modelos seleccionados en cada una de las secciones anteriores y juntarlos se obtuvieron métricas más altas en el conjunto de pruebas. Este modelo será utilizado para predecir el precio de inmuebles ya sea tanto en venta como en renta. En el siguiente cuadro se pueden ver las métricas del modelo en conjunto.

Modelo de Regresión Final	Métrica R2	MAE
Modelo de Regresión en Conjunto	0.824	457,544.02

En la siguiente gráfica se puede observar la comparación entre valores reales y predichos por el modelo de regresión.



7.8. Prueba Real

Con los siguientes datos en la tabla extraídos de una página de venta y renta de inmuebles, se hará la prueba final del modelo de predicción. Lamentablemente no se pudo conseguir registros de 2016, por lo cuál se utilizaron datos de 2021. Lo anterior generará un error en la predicción, pero se hará la prueba para conocer los resultados del modelo entrenado.

d_created_on	c_lat	c_lon	c_price	c_surface_covered_in_m2	...
2016-10-21	19.316667	-99.25	780000	210	...
2016-01-05	25.67507	-100.31847	16000000	590	...
2016-02-17	20.66682	-103.39182	12000	130	...
2016-02-23	20.97537	-89.616996	17000	160	...

La comparación entre el precio real y el predicho por el modelo se encuentra en la siguiente tabla.

Precio Real	Predicción
7,800,000	3,990,672.32
16,000,000	7,283,819.36
12,000	19,925.72
17,000	9,789.32

El resultado de la métrica R2 es de 0.5360.

Gracias a los resultados anteriores se puede tener noción de los siguientes pasos para este modelo de regresión. Se debe complementar la tabla de datos con información actual del mercado inmobiliario, esto para tener una predicción con mayor precisión en un futuro. De igual manera se deben de explorar nuevas técnicas de modelación y selección de variables para tener el mejor modelo posible para esta problemática.

Capítulo 8

Modelación No Supervisada

8.1. Objetivo

La finalidad de este capítulo es encontrar patrones no visibles dentro de nuestro conjunto de datos enriquecido con ayuda de nuevas técnicas de aprendizaje no supervisado e igualmente con el apoyo del conocimiento previo adquirido a lo largo de este proyecto. Estos patrones no son fáciles de identificar por lo que se utilizarán técnicas de Machine Learning aplicadas a la búsqueda de nuevos grupos dentro de nuestros datos y la visualización de dichas divisiones en dimensiones reducidas.

8.2. Procesamiento Antes de Modelación

Al notar que las variables de enriquecimiento añadidas en el capítulo pasado fueron de gran ayuda para los modelos supervisados, se utilizarán nuevamente para poder encontrar patrones de una manera más fácil. Recordando que se agregaron las variables bandera que indican el estado donde se encuentra el inmueble. De igual manera se agregaron datos demográficos y económicos, como los son la superficie del estado, población y tasa hipotecaria promedio, con las cuales se crearon nuevas variables como la densidad poblacional, obteniendo así una tabla con 84 variables con características de cada inmueble.

Finalmente, antes de entrar a la visualización con dimensiones reducidas, es necesario recordar la división de inmuebles en venta y renta que existe. Se seguirá trabajando de manera separada para no encontrar divisiones notorias al momento de utilizar las técnicas de modelación no supervisada, por lo que se tendrán análisis para los inmuebles que se encuentren en venta y otro para aquellos inmueble por arrendar.

8.3. Visualización con Reducción de Dimensiones

Al tener una gran cantidad de datos dentro de nuestra tabla analítica, es difícil visualizar los patrones dentro de una gráfica debido a las limitaciones humanas. Es por esto que por medio de los métodos de PCA, MDS y t-SNE se podrán reducir las dimensiones al combinar la información de cada una de las características que se tienen por inmueble, y así poder encontrar agrupamiento o patrones dentro de todas las variables.

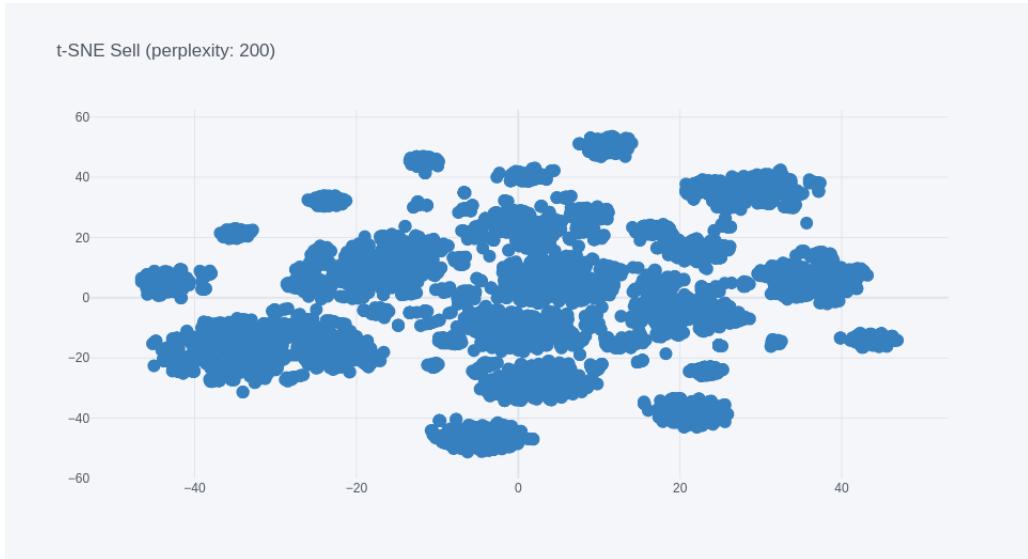
Para el uso de las técnicas anteriores, es necesario tener a todas las variables en la misma escala, ya que por las distintas dimensiones de las características, los algoritmos podrían dar más pesos a las variables con escalas más grandes. A continuación se mostrarán las tablas con la media y la varianza de cada variable.

8.3.1. Análisis de Venta

	c_lat	c_lon	c_surface_covered_in_m2	...	v_estado_Zacatecas
mean	20.3415	-98.8365	174.7653	...	0.0015
variance	4.3000	9.4280	13611.0774	...	0.0015

Visualización Venta

Al comparar las distintas gráficas que dan los múltiples algoritmos se decidió utilizar la gráfica generada por el método t-SNE.



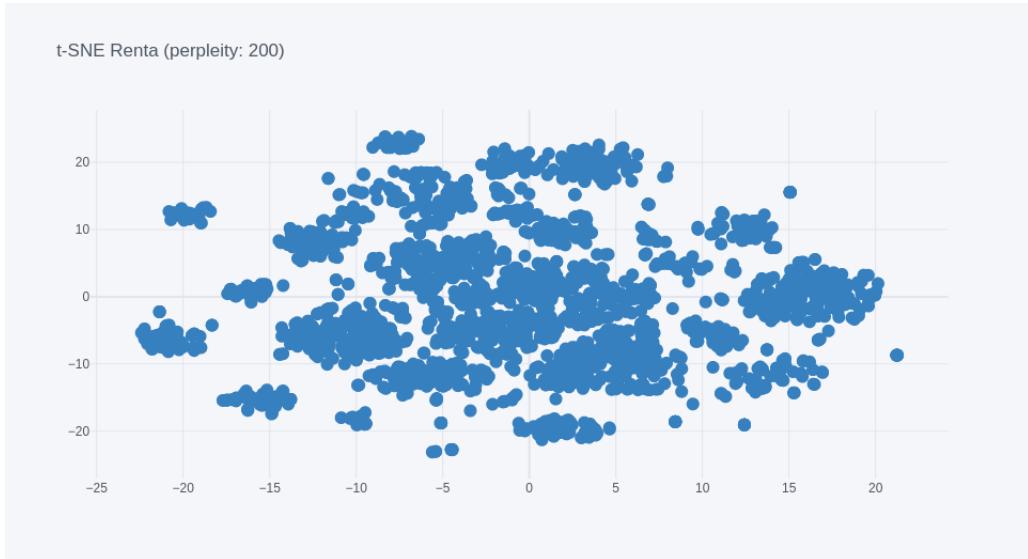
Al observar las gráficas de reducción de dimensiones, notamos que el método que tiene divisiones más específicas es t-SNE. Con este método observamos por lo menos 10 grupos, que nos pueden ayudar a la elección de cantidad de clusters a generar.

8.3.2. Análisis de Renta

	c_lat	c_lon	c_surface_covered_in_m2	...	v_estado_Zacatecas
mean	20.5170	-98.4521	143.3521	...	0.0035
variance	5.0580	11.6563	12769.1826	...	0.0035

Visualización Venta

Igualmente, al comparar las distintas gráficas que dan los múltiples algoritmos se decidió utilizar la gráfica generada por el método t-SNE con un parámetro de 200 en complexity.



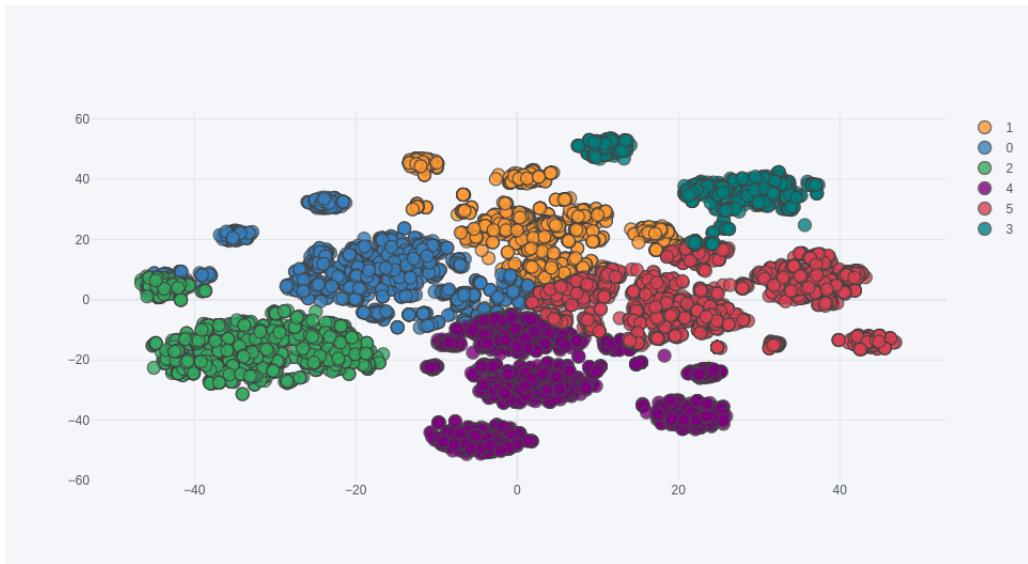
Observando los resultados de los tres métodos de reducción de dimensiones podemos notar que el método t-SNE es el mejor en cuanto a la separación tomando en cuenta la visualización. Al observar la gráfica de t-SNE es notable una división de al menos 10 grupos dentro de los registros de renta. Esta división puede funcionar al elegir el número de clusters a generar en los próximos pasos.

8.4. Perfilamiento de Inmuebles

Al analizar las gráficas del capítulo anterior se observa una división entre los datos de cada inmueble. El objetivo de esta sección es encontrar las características de los inmuebles dentro de la tabla analítica para poder determinar los distintos grupos de casas y departamentos que se tienen. Igualmente estos grupos darán a conocer las oportunidades de compra y venta que se tienen en los bienes raíces de la República Mexicana, encontrando los grupos con mayor balance entre calidad, ubicación y precio.

8.4.1. Análisis de Venta

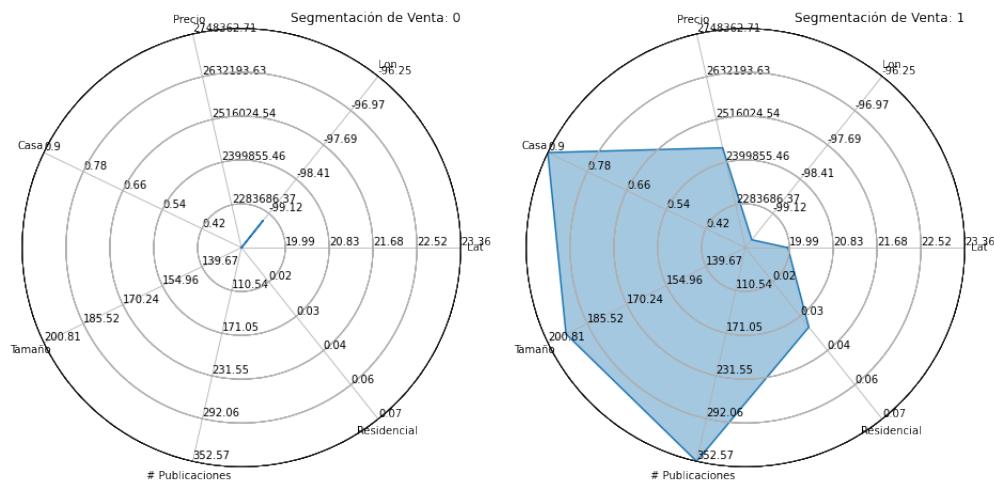
Con la reducción de dimensiones propuesta en la sección anterior se eligió un número total de seis segmentos dentro de nuestro conjunto de inmuebles en venta. Cada uno cuenta con características distintas a los otros, como se podrá observar a continuación.



Segmentación de Venta

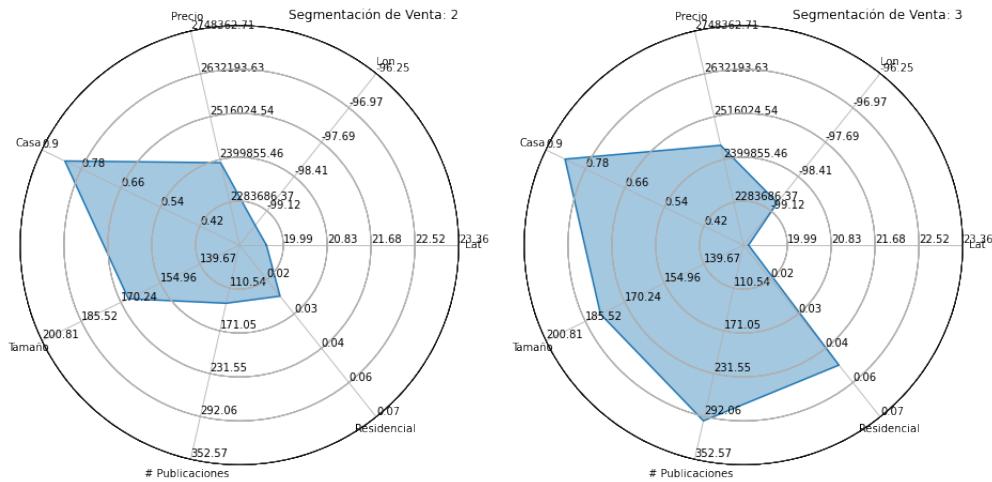
Segmento 0: En el primer segmento se encuentran los departamentos en el centro del país, estos cuentan con una superficie menor y un menor precio de venta. En general, hay muy pocas publicaciones de este tipo de inmueble por lo que esto puede ser un factor para bajar el precio del bien al no tener un punto de comparación.

Segmento 1: En este grupo se encuentran las publicaciones de casas más comunes en el país, tienen un precio promedio y una superficie mayor que la mayoría de los inmuebles, ya que en su mayoría son casas, estas no se encuentran en zonas residenciales como la gran parte de las publicaciones.



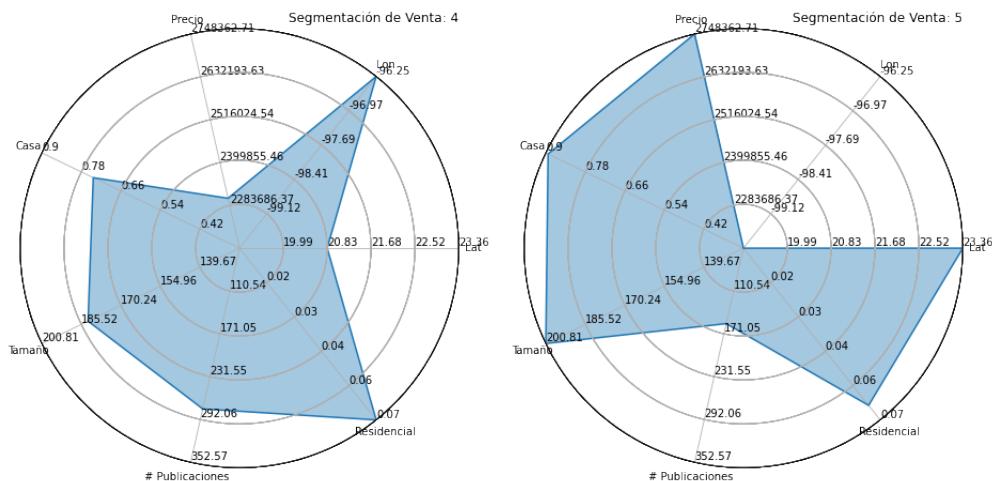
Segmento 2: En este segmento se encuentran las casas cerca de zonas industriales, estas cuentan con una superficie y precio promedio, por lo que están al alcance de todos los trabajadores de estas zonas en el Bajío del país.

Segmento 3: El grupo cuenta con las bienes en venta en los estados más grandes de la república, por lo que su tamaño es mayor de lo normal. Igualmente, existe muchas publicaciones con estas características, por lo que los precios no son tan elevados ya que existe una gran competencia en la venta de estos inmuebles. Estos bienes cuentan con la mejor relación superficie/precio a comparación de los demás segmentos.



Segmento 4: En general este grupo de inmuebles cuenta con casas cerca del Golfo de México y el Caribe, es decir son casa muy cercanas al mar y a la playa. Además, son casas con una mayor superficie que el promedio pero con un precio acorde al promedio nacional. Finalmente, estas casas se encuentran en zonas residenciales, por lo que son una gran oportunidad para los amantes del mar.

Segmento 5: Los inmuebles en este grupo cuenta con la mayor superficie de todas y el mayor precio de todos, se encuentran en zonas cercanas a la frontera, próximos a los estados de Nuevo León y Tamaulipas, por lo que se puede decir que son las grandes casas que se encuentran en la zona de Monterrey y San Pedro Garza García, una de las colonias más caras de todo México.



8.4.2. Análisis de Renta

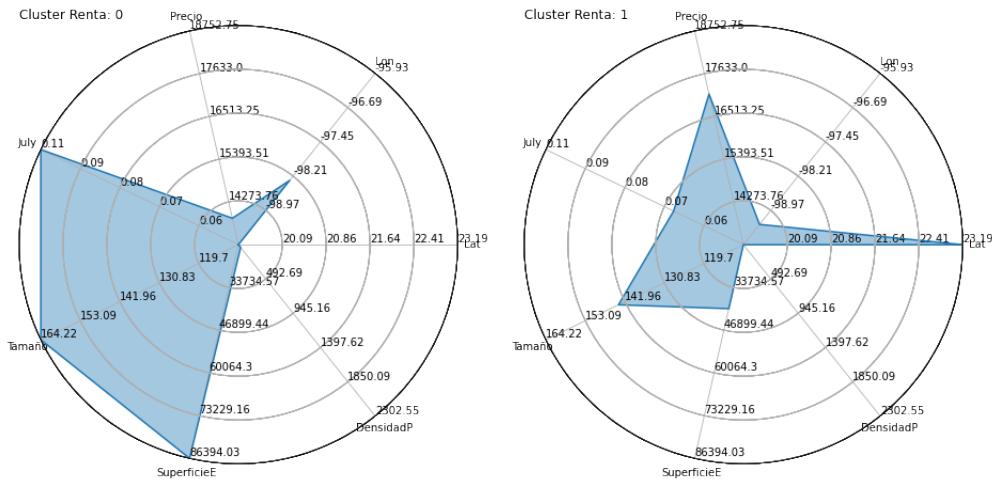
Igualmente, se analizó la información de rentas de la tabla analítica de datos. En ella se encontraron distintos tipos de inmuebles, como aquellas casas vacacionales cerca de la ciudad, o inmuebles cerca de zonas industriales para los trabajadores, o aquellos lujosos departamentos en las playas del Caribe.



Segmentación de Renta

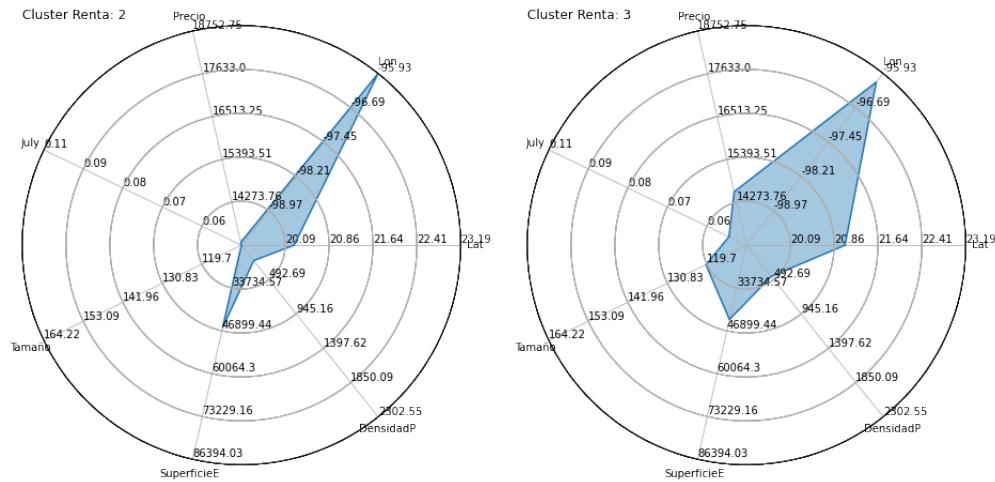
Segmento 0: El segmento se puede identificar como las casas de verano cercanas a la Ciudad de México, estas cuentan con un precio intermedio y una superficie más grande que el promedio de los otros inmuebles, se encuentran en estados donde la densidad de población es baja, esto es ideal para descansar.

Segmento 1: En este grupo se encuentran las grandes casas ubicadas en los estados del norte del cerca de Golfo de México, estas casas se encuentran en estados muy grandes con poca densidad de población y cuentan con un precio de renta más alto del normal. Lo anterior puede indicar que son bienes cerca de la zona fronteriza y en ciudades importantes como Monterrey.



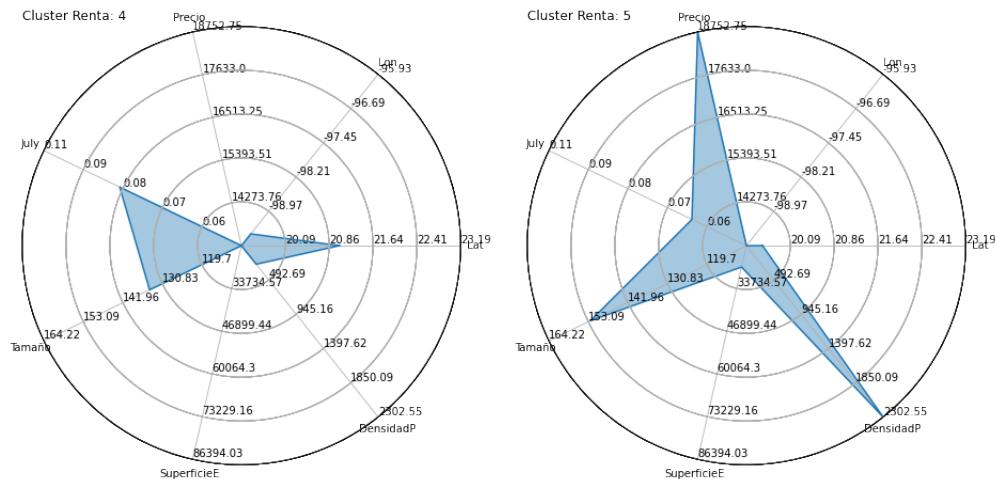
Segmento 2: Este segmento cuenta con las casas cercanas a los puertos del Golfo de México, cuentan con una superficie y precio menor al promedio, esta es una buena opción para vacacionar cerca del mar si el presupuesto es limitado.

Segmento 3: Al igual que el segmento anterior, esta agrupación de inmuebles se encuentra cerca del Golfo de México muy cercano a la parte del Caribe, cuenta con bienes más grandes que el segmento anterior y un precio más elevado, por lo que podemos decir que estas son casa para vacacionar de un nivel más alto. En este grupo están ubicadas las departamentos y casa en Yucatán y Quintana Roo, lo que vuelve las rentas más caras.



Segmento 4: Este grupo de inmuebles se encuentran las casas en zonas industriales en el Bajío de México, cuentan con menor superficie y hay una mayor concentración de personas. Son casas para los trabajadores de todas las industrias.

Segmento 5: En este segmento se ubican todos los inmuebles en renta de la Ciudad de México y el Estado de México, ya que se cuenta con un mayor precio de renta y menor superficie a comparación de otros. Igualmente, estos inmuebles están ubicados en zonas con mayor concentración de población.



Finalmente, con estos resultados es posible observar que la mayoría de las publicaciones, tanto de venta como de renta, se encuentran concentradas en el occidente, centro y sur del país. Con esto segmentos es posible encontrar las características de cada zona y es notable el comportamiento de precios y superficie en cada estado, ya que los segmentos cerca de las grandes ciudades tienden a tener un precio mayor, aunque en la zona norte se cuenta con mayor superficie de inmueble a comparación de la Ciudad de México. También, es posible observar que cerca de las grandes ciudades siempre existen zonas vacacionales donde la mayoría de los ciudadanos renta o compra una casa para descansar de la gran densidad de población que viven día a día.

Capítulo 9

Modelo Redes Neuronales

Actualmente los modelos de redes neuronales han ayudado a mejorar de sobremanera la inteligencia artificial, esto imitando la naturaleza y el funcionamiento de las neuronas reales. Las redes neuronales artificiales están compuestas de un gran número de componentes interconectados, con un fin común y específico.

En este capítulo se creará un modelo de redes neuronales multicapa con el objetivo de conseguir un modelo con mejores métricas y predicciones que los obtenidos en el capítulo 7. Si los resultados no mejoran se utilizarán los modelos anteriormente entrenados. Al igual que en secciones anteriores se dividirá la información de venta y renta, ya que las variables de entrenamiento y la target tienen distinta naturaleza entre los tipos de operación.

9.1. Procesamiento de Variables Venta

Se utilizarán las 53 variables más significativas que se ocuparon para entrenar los modelos principales, esto con el objetivo de tener una comparación certera y sin modificaciones. La siguiente tabla contiene el nombre de las variables con las que se entrenará.

'c_lat'	'descr_comedor'
'c_lon'	'descr_id'
'c_surface_covered_in_m2'	'descr_nocnok'
'c_poblacion'	'descr_recamaras'
'c_superficie'	'descr_sala'
'c_densidad'	'descr_sistema'
'v_conteo_geo'	'v_region_norte'
'v_anio_2015'	'v_tamanio_grande'
'v_anio_2016'	'v_tamanio_mediano'
'v_mes_8'	'v_estado_Baja California'
'v_mes_9'	'v_estado_Baja California Sur'
'v_mes_10'	'v_estado_Chihuahua'
'v_dia_semana_Lunes'	'v_estado_Coahuila de Zaragoza'
'v_dia_semana_Sábado'	'v_estado_Colima'
'v_property_type_house'	'v_estado_Distrito Federal'
'v_property_type_store'	'v_estado_Durango'
'title_casa'	'v_estado_Estado de México'
'title_departamento'	'v_estado_Hidalgo'
'title_local'	'v_estado_Jalisco'
'title_lomas'	'v_estado_Nayarit'
'title_residencial'	'v_estado_Nuevo León'
'title_san'	'v_estado_Puebla'
'title_venta'	'v_estado_Querétaro'
'descr_baaoo'	'v_estado_San Luis Potosí'
'descr_br'	'v_estado_Sinaloa'
'descr_casa'	'v_estado_Yucatán'
'descr_cocina'	

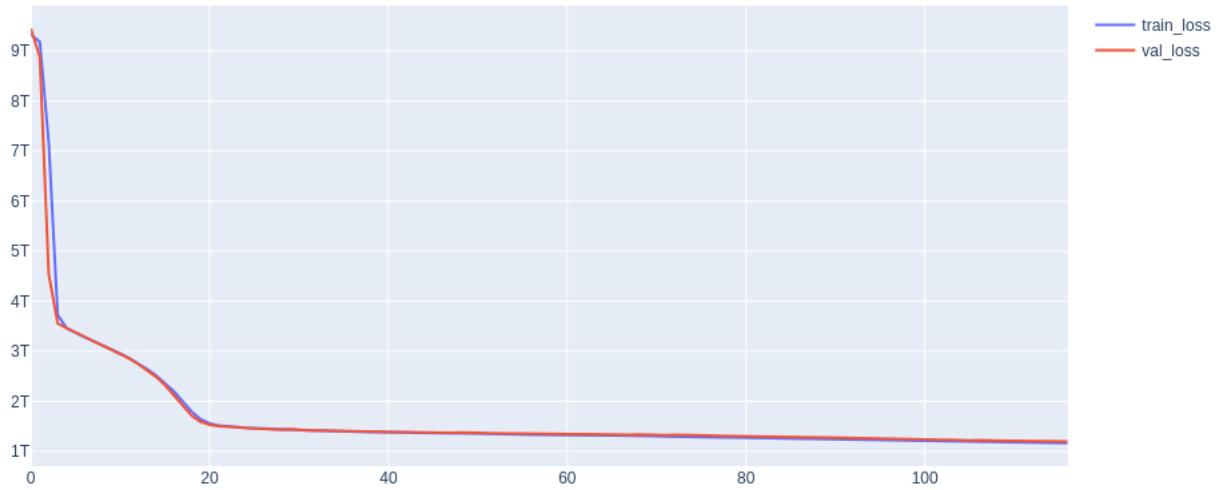
9.2. Resultados de Modelación en Venta, Redes Neuronales

Para el modelo multicapa se utilizaron 4 capas ocultas Densas, las cuales utilizaron capas Dropout con el fin de evitar el overfitting. Otro de los métodos para reducir el sobre ajuste fue utilizar un Early Stopping en el modelo para hacer entrenar de manera rápida y eficaz.

En la siguiente tabla se encuentran las métricas de R2 y Error Medio Absoluto de entrenamiento y validación, al igual que la gráfica de a función de pérdida mostrando el nulo sobre ajuste del modelo.

	R2 Score	MAE
Train	0.6872	681,324.00
Test	0.6744	693,711.02

Loss



Se puede observar que el nuevo modelo con redes multicapa es mejor que el generado en el capítulo 7, ganando tanto en métricas de R2 y Error Medio Absoluto, pero al compararlo con modelos de ensamble o Boosting no alcanza los mismos niveles que estos, ya que cuenta con un déficit de 8 puntos porcentuales y cerca de 100 mil unidades de Error Medio Absoluto.

En la siguiente sección se modelarán los inmuebles en renta dentro de la muestra inmobiliaria.

9.3. Procesamiento de Variables Renta

Igualmente, se utilizarán las 53 variables más significativas que se ocuparon para entrenar los modelos principales, esto con el objetivo de tener una comparación certera y sin modificaciones. La siguiente tabla contiene el nombre de las variables con las que se entrenará.

'c_lat'	'descr_br'
'c_lon'	'descr_casa'
'c_surface_covered_in_m2'	'descr_cocina'
'c_poblacion'	'descr_comedor'
'c_superficie'	'descr_id'
'c_densidad'	'descr_nocnok'
'c_tasa'	'descr_recamaras'
'v_anio_2015'	'descr_sala'
'v_anio_2016'	'descr_sistema'
'v_mes_2'	'v_region_norte'
'v_mes_4'	'v_tamanio_grande'
'v_mes_5'	'v_tamanio_mediano'
'v_mes_7'	'v_estado_Baja California'
'v_mes_10'	'v_estado_Colima'
'v_mes_11'	'v_estado_Distrito Federal'
'v_dia_semana_Lunes'	'v_estado_Durango'
'v_dia_semana_Miercoles'	'v_estado_Estado de México'
'v_property_type_house'	'v_estado_Guerrero'
'v_property_type_store'	'v_estado_Hidalgo'
'title_casa'	'v_estado_Jalisco'
'title_centro'	'v_estado_Nuevo León'
'title_departamento'	'v_estado_Puebla'
'title_local'	'v_estado_Querétaro'
'title_lomas'	'v_estado_San Luis Potosí'
'title_renta'	'v_estado_Veracruz de Ignacio de la Llave'
'title_san'	'v_estado_Yucatán'
'descr_baa'	

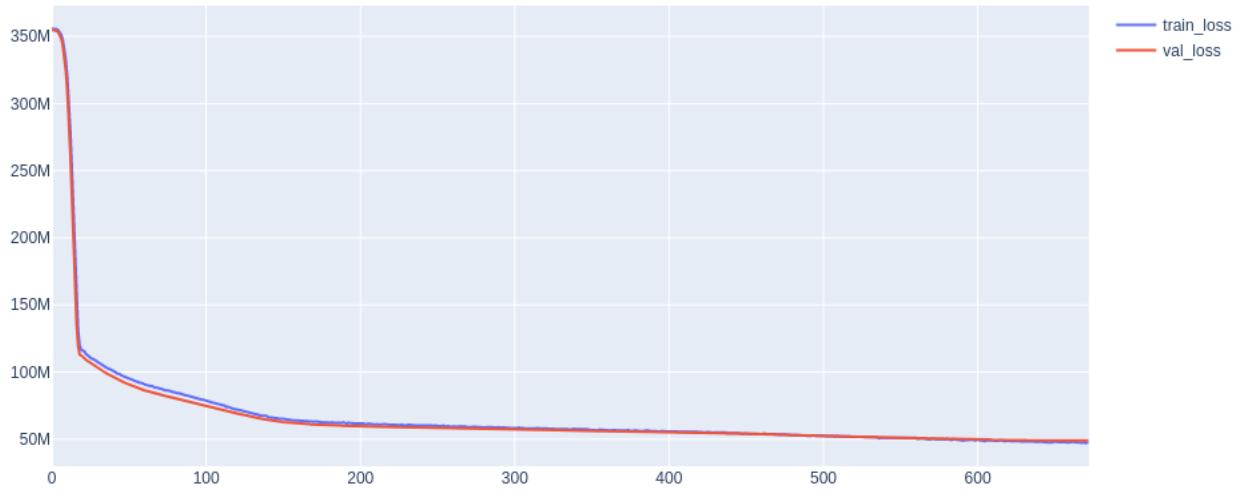
9.4. Resultados de Modelación en Renta, Redes Neuronales

Para el modelo multicapa se utilizaron 3 capas ocultas Densas, las cuales utilizaron capas Dropout con el fin de evitar el overfitting. Otro de los métodos para reducir el sobre ajuste fue utilizar un Early Stopping en el modelo para hacer entrenar de manera rápida y eficaz.

En la siguiente tabla se encuentran las métricas de R2 y Error Medio Absoluto de entrenamiento y validación, al igual que la gráfica de a función de pérdida de renta mostrando el nulo sobre ajuste del modelo.

	R2 Score	MAE
Train	0.6222	4,656.31
Test	0.5704	4,919.67

Loss



Al igual que en el modelo de ventas, se ha mejorado el modelo de redes neuronales que se tenía en el capítulo 7, pero este nuevo modelo no supra a los modelos de ensamble o de Boosting.

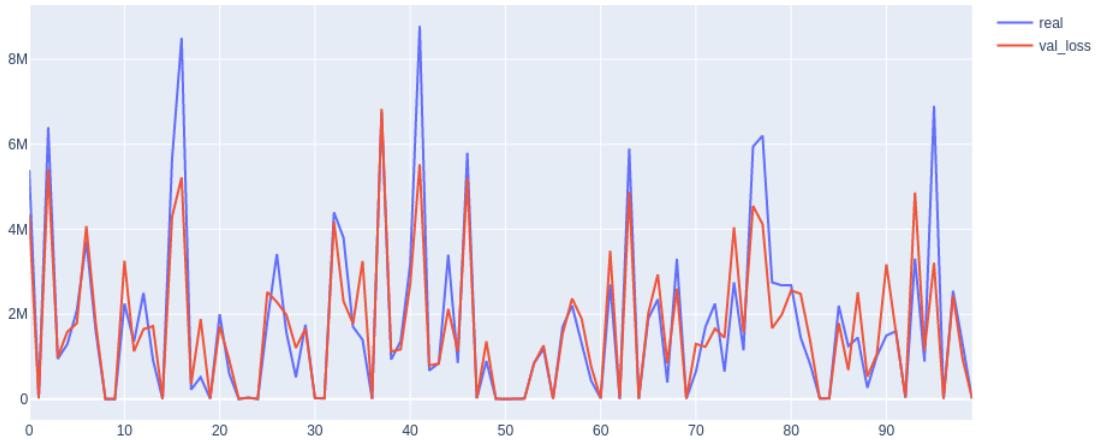
9.5. Resultados de Modelación Venta y Renta, Redes Neuronales

Al momento de juntar los modelos de Venta y Renta se obtienen métricas más altas que los modelo individuales, las métricas se muestran en el cuadro siguiente.

	R2 Score	MAE
Train	0.7239	592,160.65

La siguiente gráfica muestra el desempeño de las predicciones.

Loss



En conclusión, es notable el resultado más bajo que aquel que se tenía con el modelo ensamblado del capítulo 7, por lo que para la predicción de los precios, tanto para venta y renta, se utilizará el modelo ensamble con el Bosque Aleatorio y XGBoost, el cual cuenta con las siguientes métricas.

Modelo de Regresión Final	Métrica R2	MAE
Modelo de Regresión en Conjunto	0.824	457,544.02

Capítulo 10

Conclusiones

A partir del análisis exploratorio es notable la diferencia de inmuebles dentro de cada estado y su similitud con el mismo. Las variables como superficie, número de publicaciones y las variables socio económicas afectan de manera notable al precio de un inmueble, así que al encontrar datos más granulares para cada inmueble se podrá mejorar la predicción del modelo.

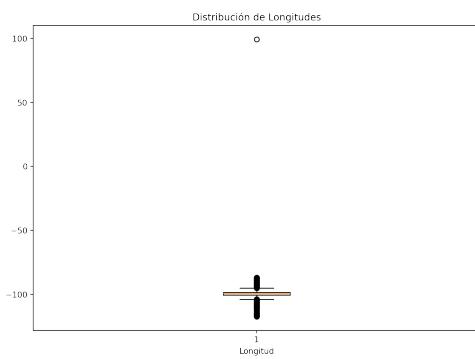
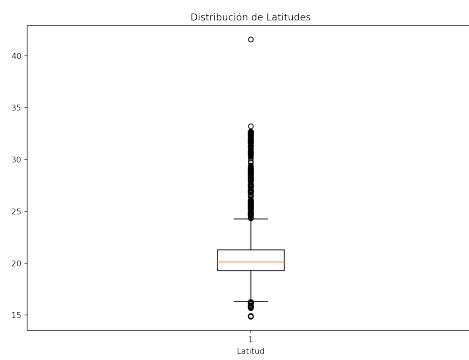
Al analizar las gráficas y métricas, es notable el buen ajuste del modelo, con este se podrán hacer predicciones de valor para conocer el precio adecuado de un inmueble a partir de información histórica, esto ayudará a encontrar oportunidades inmobiliarias dentro del mercado. Al no ser un problema trivial el determinar el precio de una casa, se necesitan modelos complejos para poder ajustar la predicción del precio, es conveniente utilizar modelos de ensamble para su predicción, esto mejora de sobre manera las métricas de ajuste y generará el modelo final de forma eficiente. Es necesario conseguir más información histórica y de distintos años para poder generalizar el modelo y poder detectar los cambios de precio en el tiempo.

Finalmente, es de destacar las agrupaciones encontradas dentro de los inmuebles en México. La división entre los distintos sectores se encuentran muy marcados, ya que se cuentan con casas para trabajadores, inmuebles en la costa, grandes residencias en la frontera norte del país, entre otras. Para mejorar los ajustes del modelo predictivo se recomienda la creación y entrenamiento de modelos particulares para cada uno de las agrupaciones detectadas.

Capítulo 11

Apéndice

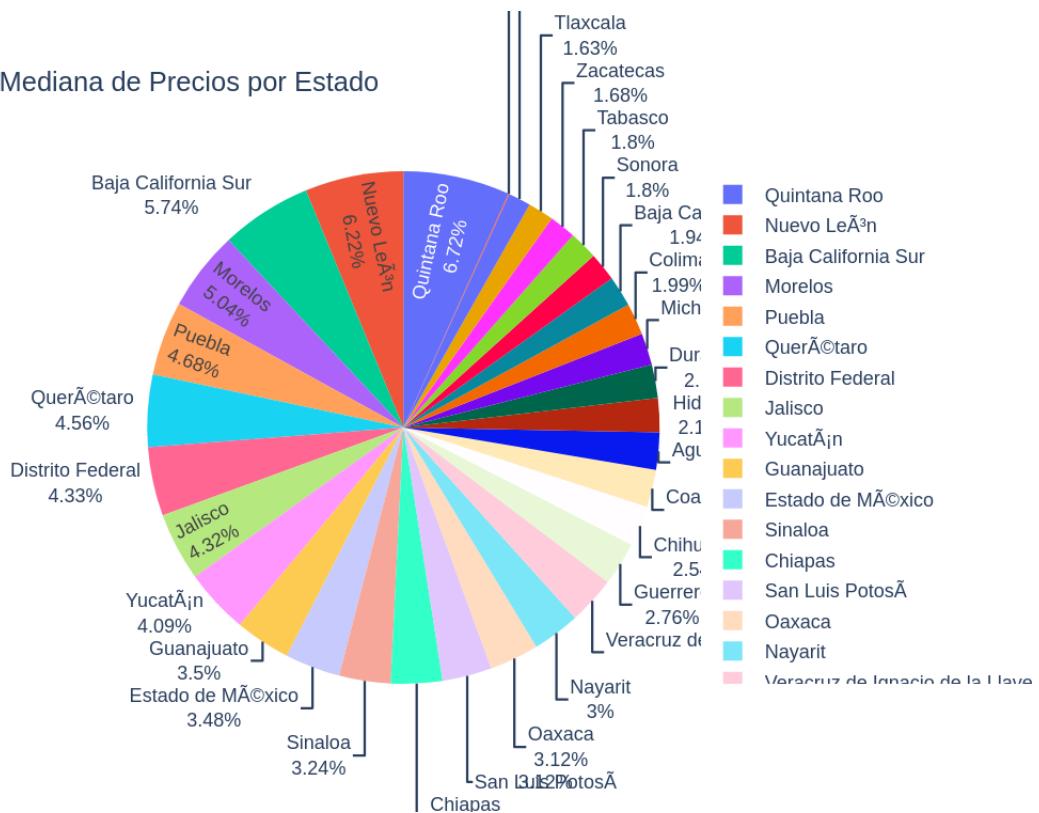
11.1. Gráficos Adicionales



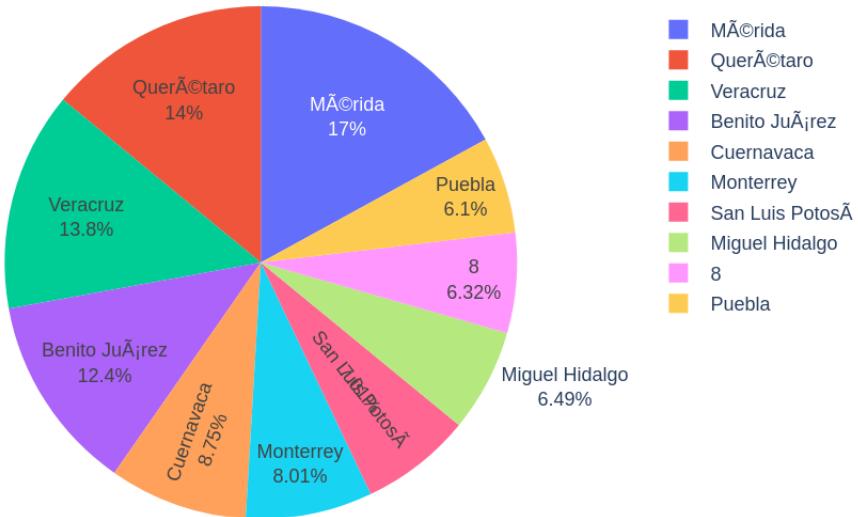
Número de Publicaciones por Estado



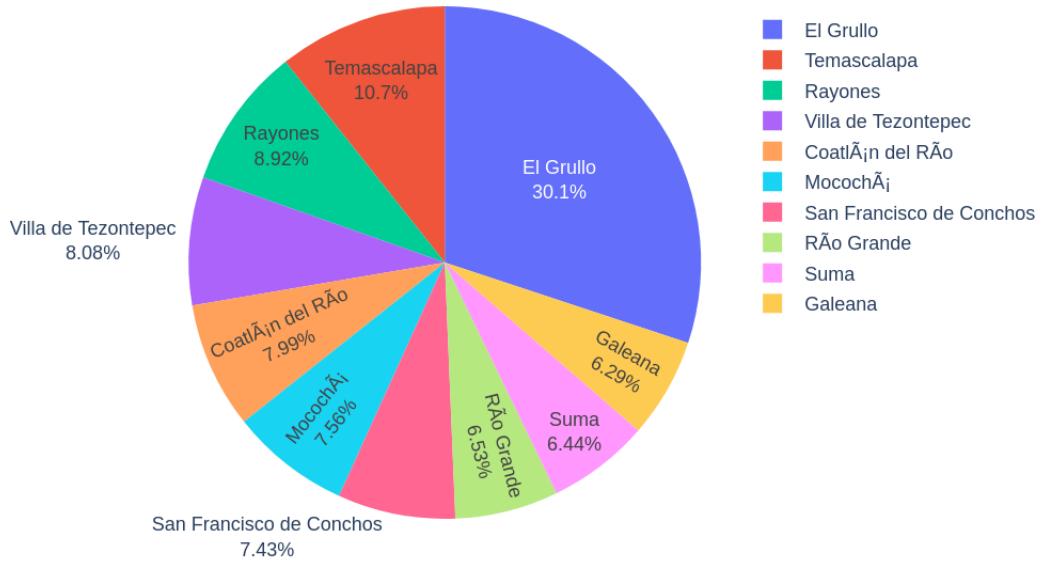
Mediana de Precios por Estado



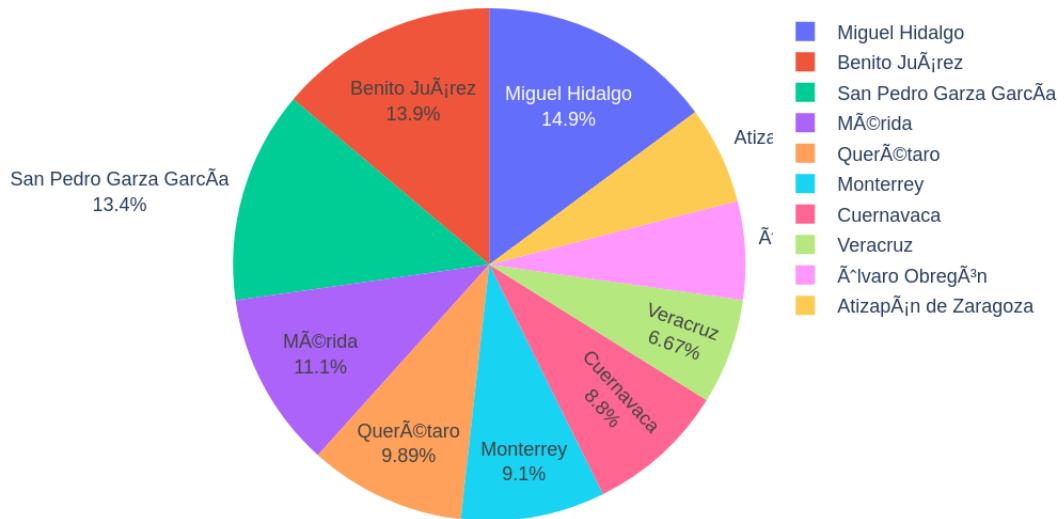
10 Municipios con mas Publicaciones



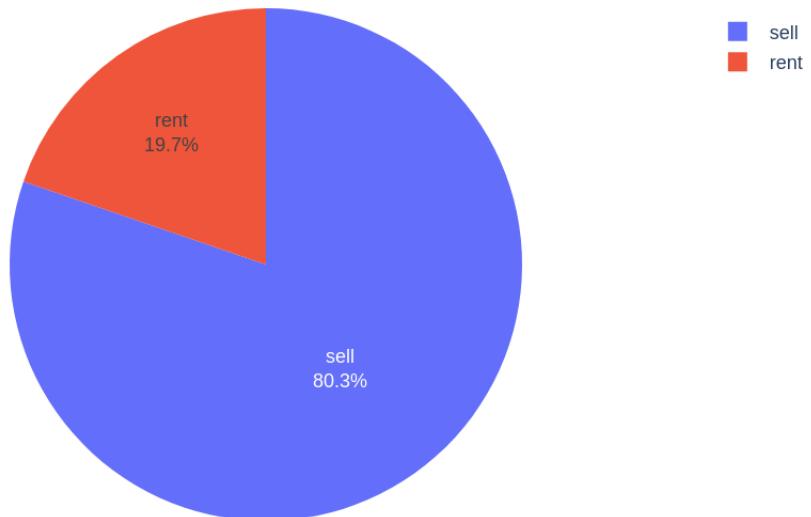
Mediana de Precios por Municipio

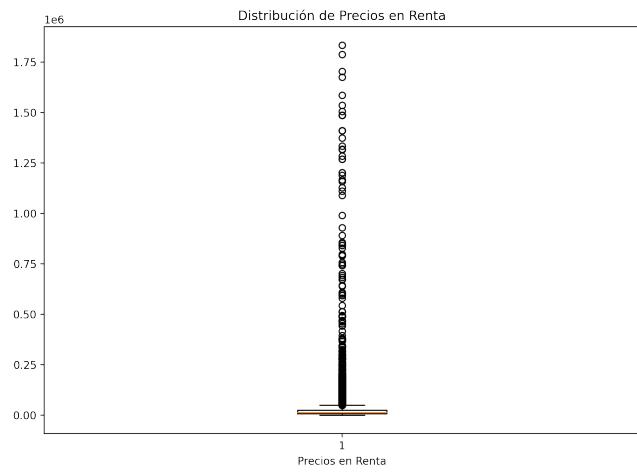
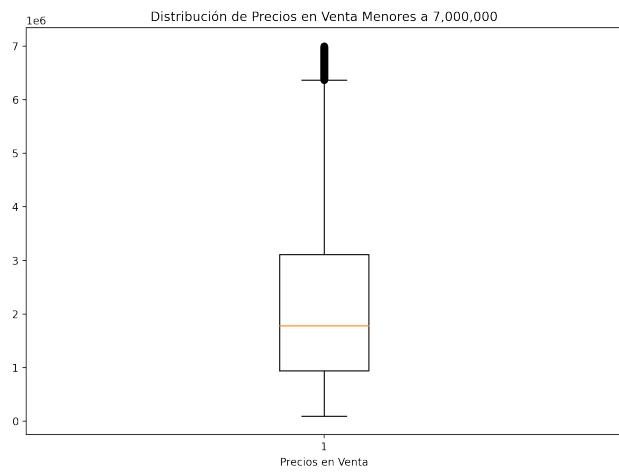
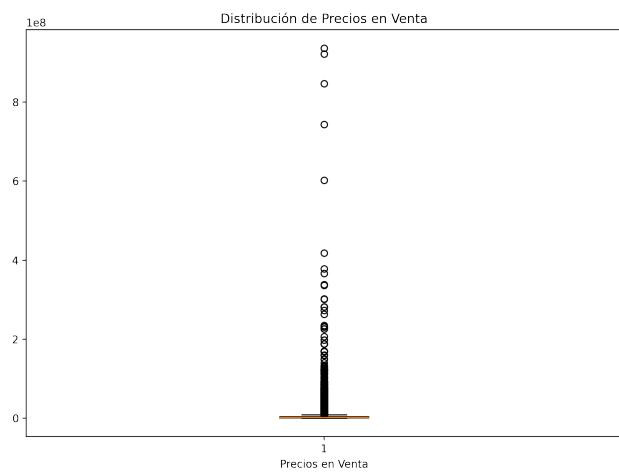


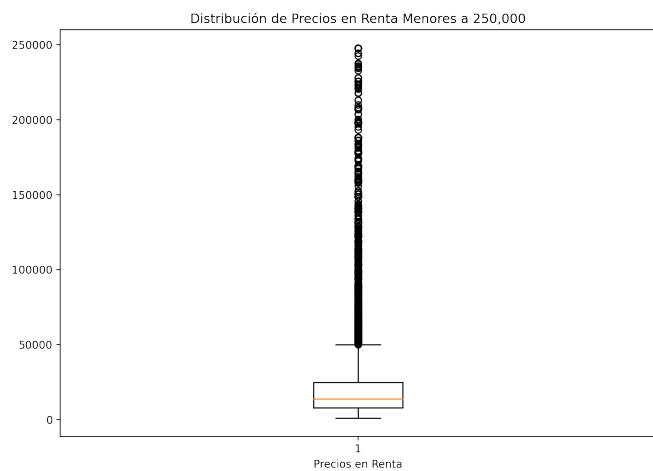
Suma de Precios por Municipio



Publicaciones por Operacion







Monedas por Publicacion

