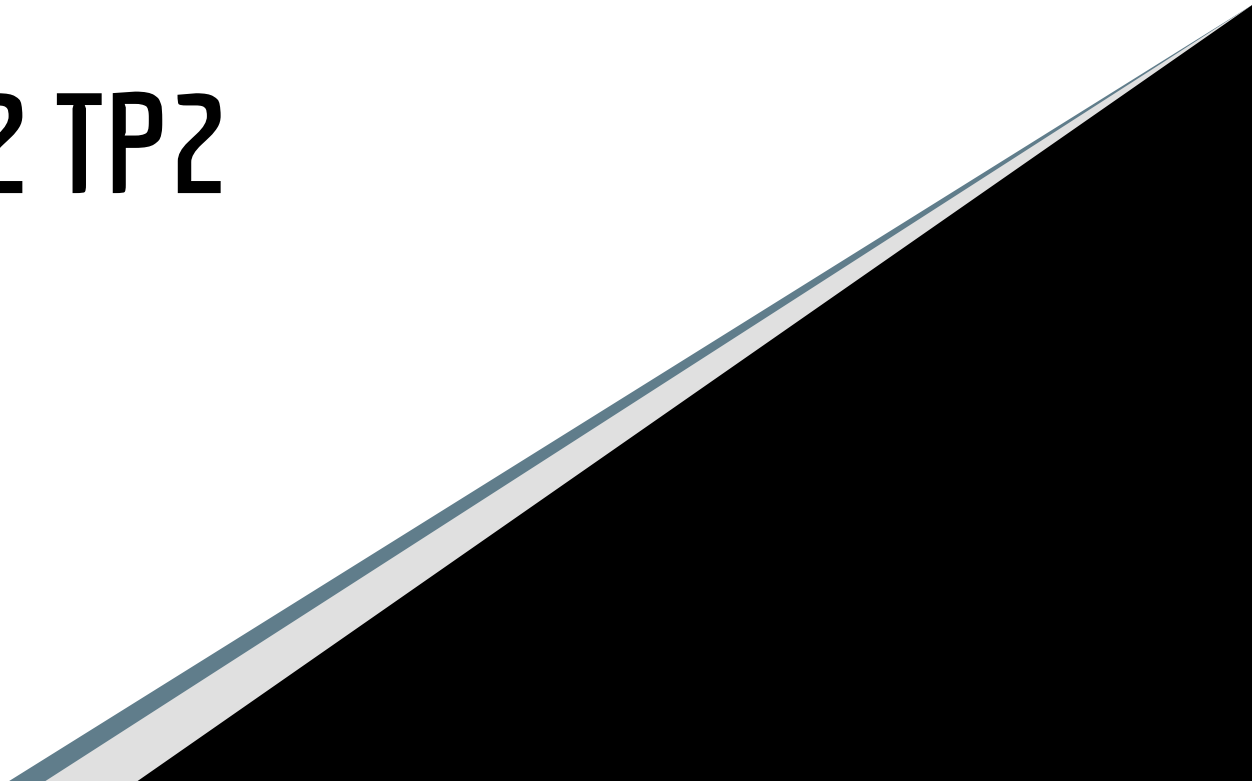
Workshop G2 TP2

Alvarez, Martha
Roa, Fernando
Tasat, Kevin



Objetivo

Entrenar un **modelo** que pueda **predecir el precio de una propiedad** a partir del dataset de Properatti el cual contiene una serie de inmuebles ubicados en diferentes estados de Argentina y se suministran algunas características geográficas y propias de tales inmuebles.

Metodología

1

Entender la data a través de un análisis exploratorio calculando estadísticas y mediante la visualización de variables que a priori consideramos puedan tener un poder de predicción importante.

2

Realizar una limpieza del dataset, imputando registros cuando se considere pertinente y en otros casos omitiendo registros por considerarlos incompletos, además de buscar la consistencia en los datos. También se realizará un análisis de valores extremos.

3

Finalmente, **con la data resultante,** fijar un objetivo en cuanto al tipo de propiedad y/o localizaciones para **entrenar modelos de regresión lineal con y sin regularización.**

Análisis Exploratorio

Análisis exploratorio de la información

```
# Tamaño del dataset

data.shape
```

```
(121220, 26)
```

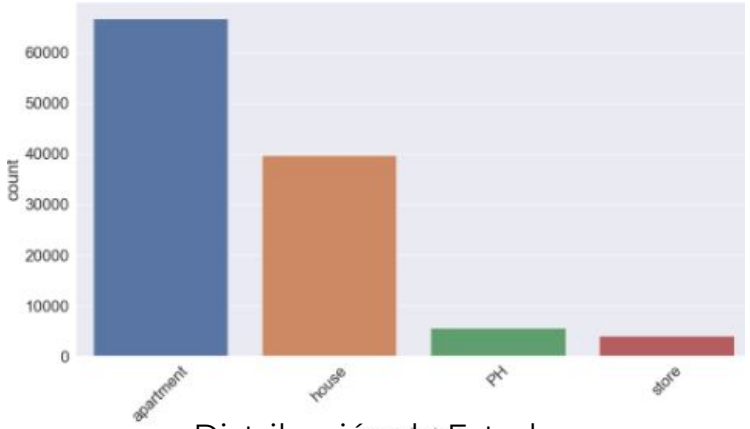
	cant_nulos	porcentaje_nulos
operation	0	0.000000
property_type	0	0.000000
place_name	23	0.000198
place_with_parent_names	0	0.000000
country_name	0	0.000000
state_name	0	0.000000
geonames_id	18180	0.156535
lat-lon	48289	0.415783
lat	48289	0.415783
lon	48289	0.415783
price	17556	0.151162
currency	17557	0.151171
price_aprox_local_currency	17556	0.151162
price_aprox_usd	17556	0.151162
surface_total_in_m2	38378	0.330446
surface_covered_in_m2	18853	0.162330
price_usd_per_m2	49289	0.424393
price_per_m2	30443	0.262123
floor	108481	0.934054
rooms	69941	0.602213
expenses	102055	0.878724
description	2	0.000017
title	0	0.000000

1. Eliminamos 'Unnamed: 0','properati_url','image_thumbnail'
2. Eliminamos duplicados

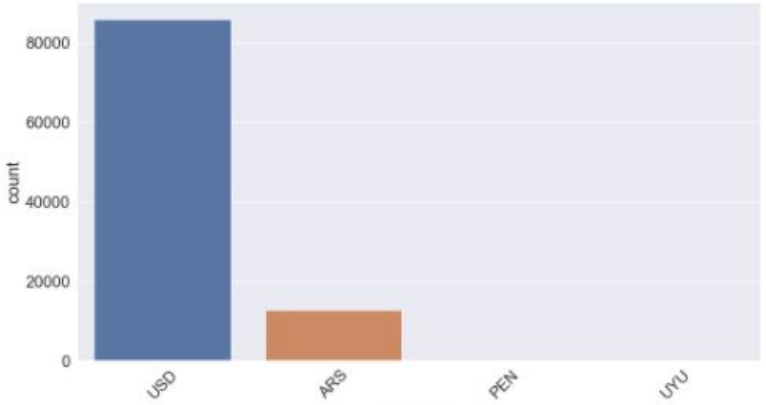
```
data.shape
```

```
(116140, 23)
```

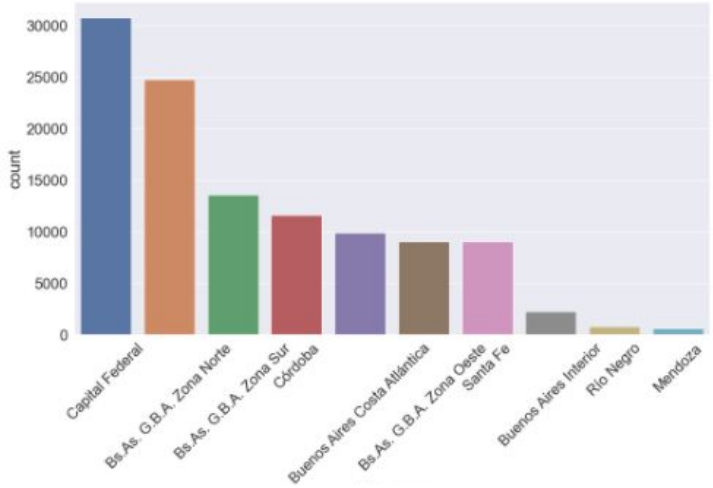
Distribución de Tipo de Propiedad



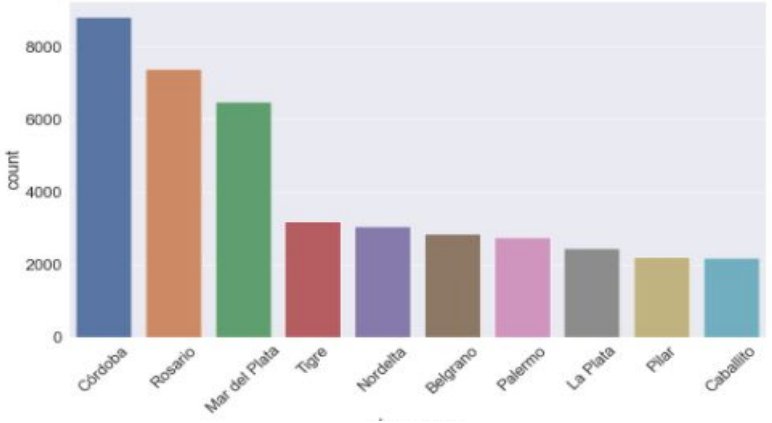
Distribución de Tipo de Moneda

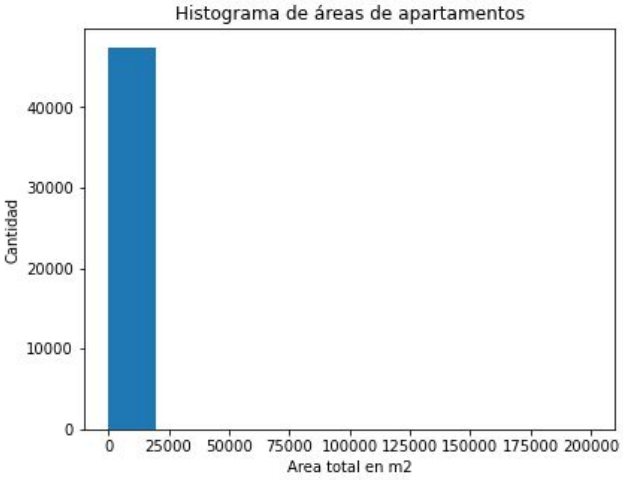


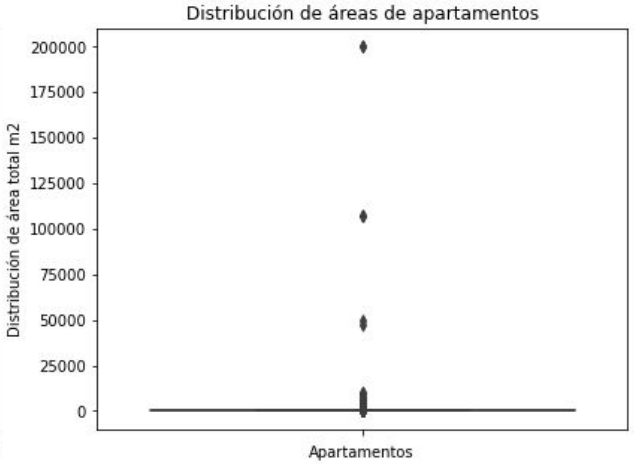
Distribución de Estados

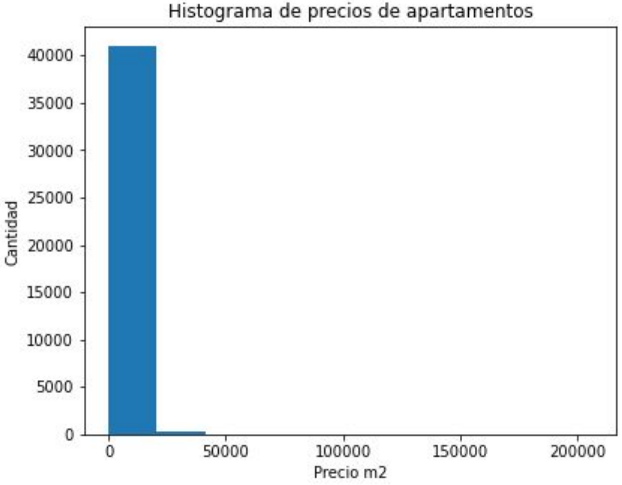


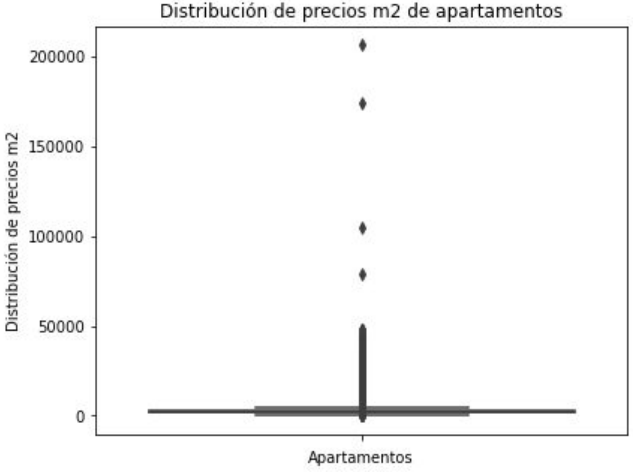
Distribución de Barrios | Localidades




Histograma de áreas de apartamentos
Cantidad
Area total en m2
0
10000
20000
30000
40000
0
25000
50000
75000
100000
125000
150000
175000
200000


Distribución de áreas de apartamentos
Distribución de área total m2
Apartamentos
0
25000
50000
75000
100000
125000
150000
175000
200000


Histograma de precios de apartamentos
Cantidad
Precio m2
0
5000
10000
15000
20000
25000
30000
35000
40000
0
50000
100000
150000
200000


Distribución de precios m2 de apartamentos
Distribución de precios m2
Apartamentos
0
50000
100000
150000
200000

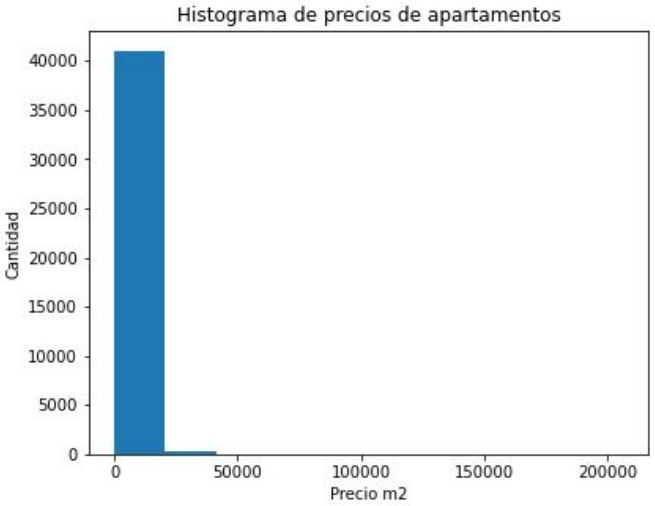
Limpieza

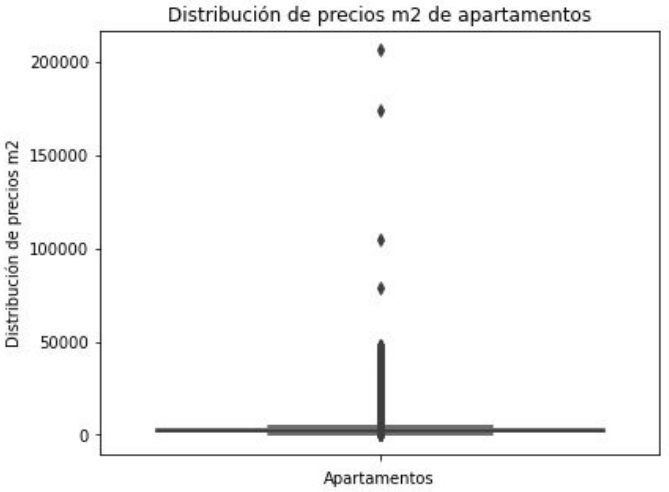
Imputación

	cant_nulos_inicial	porcentaje_nulos_inicial	cant_nulos_final	porcentaje_nulos_final	cant_recuperados
operation	0	0.000000	0	0.000000	0
property_type	0	0.000000	0	0.000000	0
place_name	23	0.000198	0	0.000198	23
place_with_parent_names	0	0.000000	0	0.000000	0
country_name	0	0.000000	0	0.000000	0
state_name	0	0.000000	0	0.000000	0
geonames_id	18180	0.156535	1910	0.156535	16270
lat-lon	48289	0.415783	48289	0.415783	0
lat	48289	0.415783	48289	0.415783	0
lon	48289	0.415783	48289	0.415783	0
price	17556	0.151162	14845	0.151162	2711
currency	17557	0.151171	14847	0.151171	2710
price_aprox_local_currency	17556	0.151162	17185	0.151162	371
price_aprox_usd	17556	0.151162	15217	0.151162	2339
surface_total_in_m2	38378	0.330446	11641	0.330446	26737
surface_covered_in_m2	18853	0.162330	11641	0.162330	7212
price_usd_per_m2	49289	0.424393	49289	0.424393	0
price_per_m2	30443	0.262123	30443	0.262123	0
floor	108481	0.934054	108481	0.934054	0
rooms	69941	0.602213	69941	0.602213	0

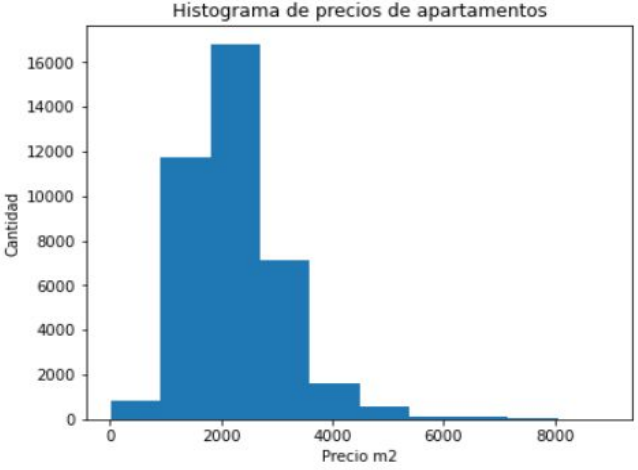
Análisis y Limpieza de Valores Extremos

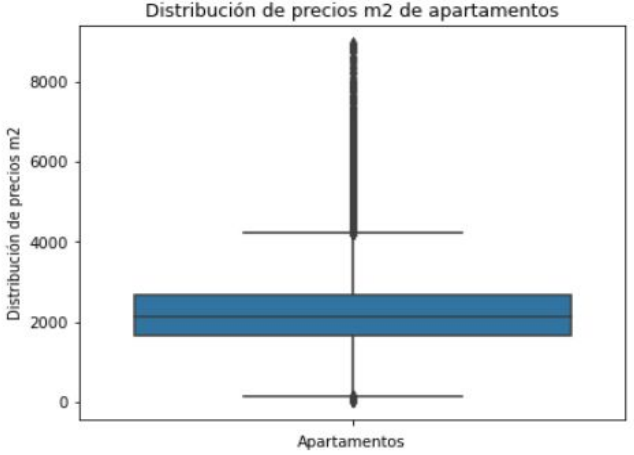
Antes



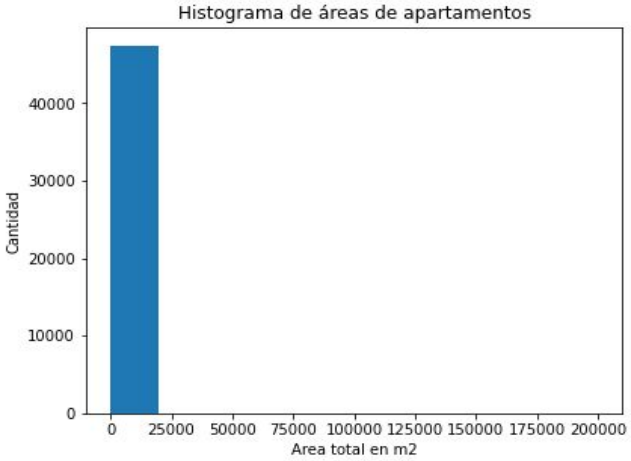


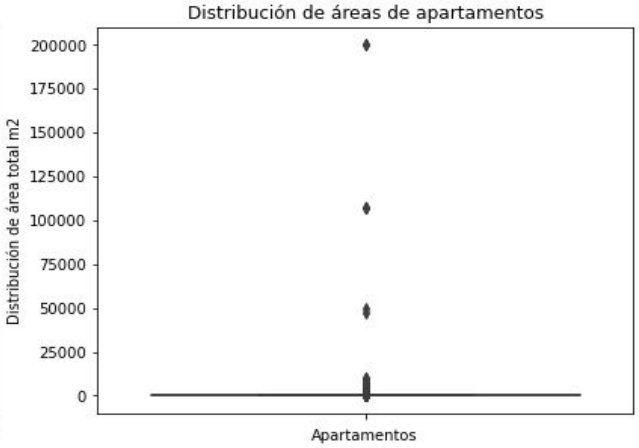
Después



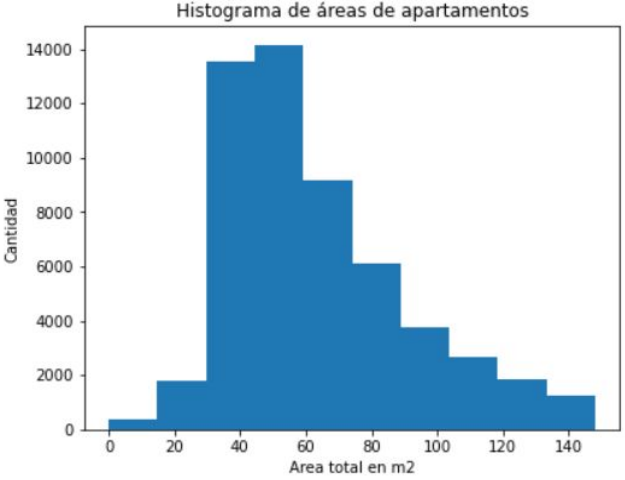


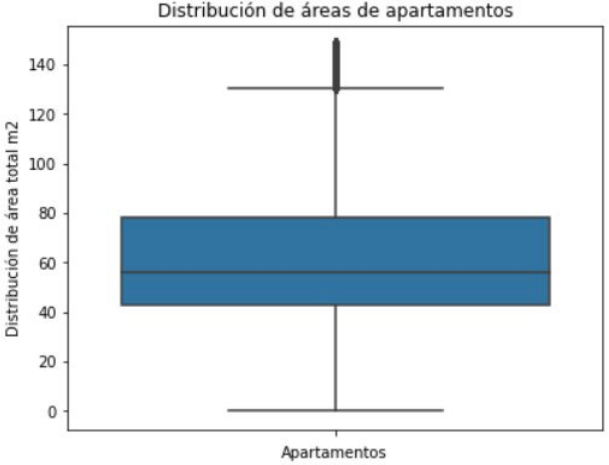
Antes





Después





Modelo:
Regresión Lineal Simple

¿Cómo es el dataset con el que entrenaremos?

	property_type	state_name	place_name	lat	lon	surface_total_in_m2	price_aprox_usd
0	PH	Capital Federal	Mataderos	-34.661824	-58.508839	55.000000	62000.000000
1	apartment	Capital Federal	Mataderos	-34.652262	-58.522982	55.000000	72000.000000
2	apartment	Buenos Aires Costa Atlántica	Centro	-38.002626	-57.549447	35.000000	64000.000000

La **variable target** es **'price_aprox_usd'**.

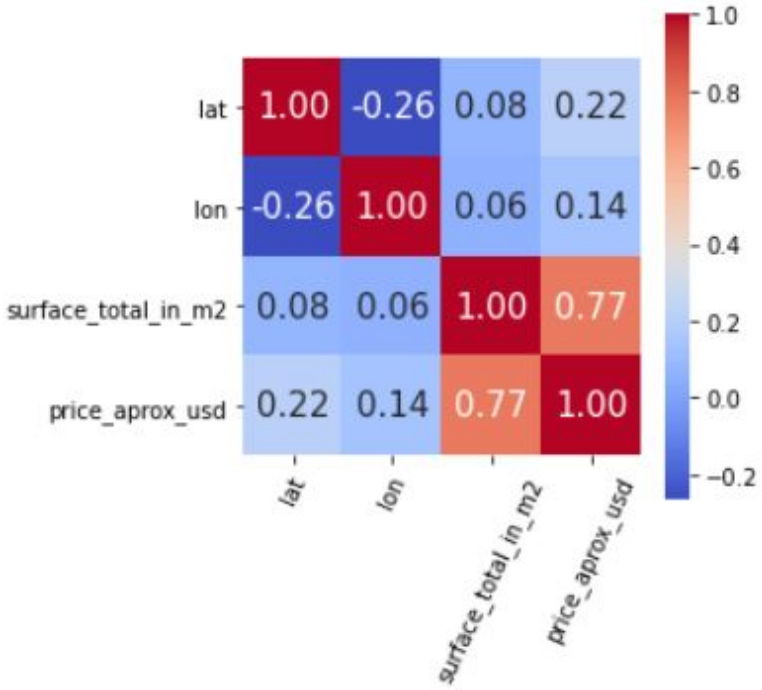
Las **features** que consideramos que tienen un alto poder de predicción del precio total de un inmueble son:

- **property_type**.
- **state_name**.
- **place_name**.
- **lat**.
- **lon**.
- **surface_total_in_m2**.

Nuestro dataset contendrá únicamente el **tipo de propiedad** predominante que es **apartamentos** y el **state_name** predominante en el dataset que es **Capital Federal**.

Shape: (14402, 6)

Análisis de Correlación



Se observa una **buena correlación** de las variables de **ubicación espacial y** de **superficie total con** la **variable target precio total**.
Se construirá un modelo simple con cada variable mencionada

Regresión Precio vs Superficie

OLS Regression Results

Dep. Variable:	price_aprox_usd	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.600
Method:	Least Squares	F-statistic:	2.162e+04
Date:	Wed, 16 Mar 2022	Prob (F-statistic):	0.00
Time:	22:27:57	Log-Likelihood:	-1.8018e+05
No. Observations:	14402	AIC:	3.604e+05
Df Residuals:	14400	BIC:	3.604e+05
Df Model:	1		
Covariance Type:	nonrobust		

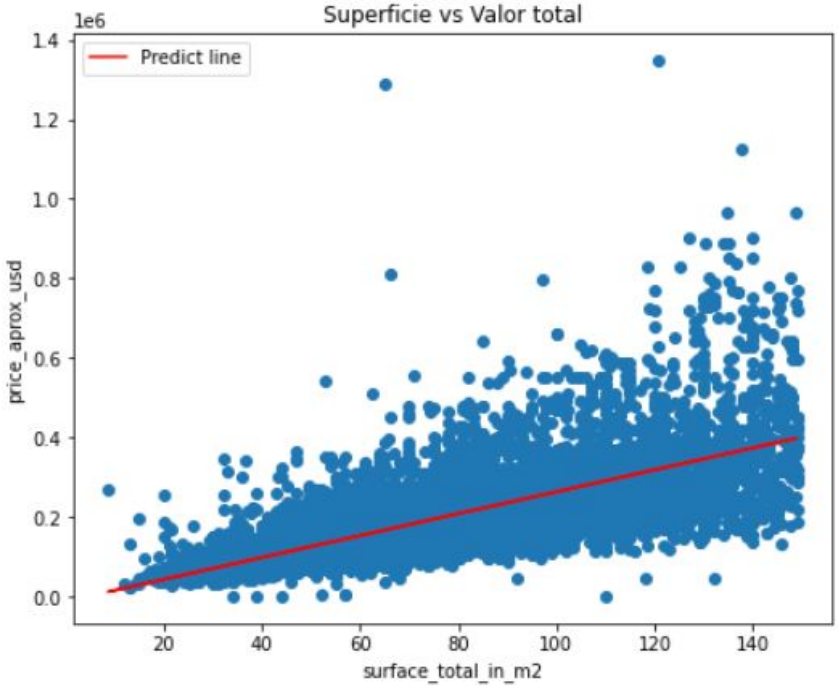
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.078e+04	1311.032	-8.221	0.000	-1.33e+04	-8208.284
x1	2750.2320	18.705	147.035	0.000	2713.569	2786.895

Omnibus:	9448.446	Durbin-Watson:	1.717
Prob(Omnibus):	0.000	Jarque-Bera (JB):	304705.373
Skew:	2.678	Prob(JB):	0.00
Kurtosis:	24.888	Cond. No.	168.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- Obtenemos un R2 de 0.600, el cual nos indica la proporción de la variabilidad que es explicada por la variable surface_total_in_m2, respecto al modelo base que sería el precio promedio de los apartamentos en Capital Federal.
- Con un p-value de 0.000 podemos decir que la relación entre la variable surface_total_in_m2 y price_aprox_usd no es producto del azar.
- Por cada m2 que aumenta el área total, el precio total aumenta en 2750.23 dólares.

Regresión Precio vs Superficie



Se observa una **buena correlación** de las variables de **ubicación espacial y** de **superficie total con** la **variable target precio total**.
Se construirá un modelo simple con cada variable mencionada

Modelo:
Regresión Lineal Múltiple

OLS Regression Results

Dep. Variable:	price_aprox_usd	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	1029.
Date:	Wed, 16 Mar 2022	Prob (F-statistic):	0.00
Time:	22:27:58	Log-Likelihood:	-1.3204e+05
No. Observations:	10801	AIC:	2.642e+05
Df Residuals:	10763	BIC:	2.644e+05
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.635e+05	478.249	341.890	0.000	1.63e+05	1.64e+05
lat	2.085e+04	1715.774	12.150	0.000	1.75e+04	2.42e+04
lon	1.771e+04	1704.270	10.392	0.000	1.44e+04	2.11e+04
surface_total_in_m2	7.135e+04	485.178	147.058	0.000	7.04e+04	7.23e+04
place_name_Balvanera	-5418.9528	611.240	-8.866	0.000	-6617.095	-4220.810
place_name_Barracas	-1305.7192	630.352	-2.071	0.038	-2541.326	-70.113
place_name_Barrio Norte	5127.5139	682.791	7.510	0.000	3789.118	6465.910
place_name_Belgrano	7768.1997	1234.223	6.294	0.000	5348.895	1.02e+04
place_name_Boedo	65.8297	550.514	0.120	0.905	-1013.280	1144.940
place_name_Caballito	9895.7141	854.439	11.582	0.000	8220.857	1.16e+04
place_name_Capital Federal	1300.4718	618.873	2.101	0.036	87.367	2513.577
place_name_Centro / Microcentro	-2412.1924	506.664	-4.761	0.000	-3405.347	-1419.038
place_name_Chacarita	336.0100	567.577	0.592	0.554	-776.545	1448.565
place_name_Coghlan	1023.9599	661.073	1.549	0.121	-271.866	2319.786
place_name_Colegiales	1835.1303	615.893	2.980	0.003	627.867	3042.394

Obtenemos un **R2 ajustado de 0.780**, los cual nos indica que el modelo obtenido explica una proporción significativa de la variabilidad.

Con un nivel de significación del 0.05, podemos decir que los coeficientes de **variables como**: place_name_**Boedo**, place_name_**Chacarita**, place_name_**Coghlan**, place_name_**Parque Patricios**, place_name_**Parque Retiro** y place_name_**Parque Saavedra** tienden a ser nulos, lo que en otras palabras significa que **su aporte al modelo no es significativo**.

Modelo:
Regresión Lineal Múltiple
Interacción entre variables

Realizamos un modelo de regresión lineal simplificado, considerando solamente variables de superficie, latitud y longitud sin interacción entre las mismas.

Luego, lo comparamos contra otro modelo con las mismas variables más su respectiva interacción entre ellas.

	Sin Interacción	**Con Interacción**
Cantidad Variables	3	7*
R2	**0.643**	**0.665**
Valor de Significación	0.05	0.05
Cantidad Variables Significativas	3	7

**Variables interacción: lat_x_lon; surface_cuad; lat_cuad; lon_cuad*

Modelo:
Regresión Lineal Múltiple
Lasso & Ridge

	Lasso	**Ridge**
Cantidad Variables	Todas	Todas
Best Alpha	10	10
Intercepto	163497.95	163503.42
R2 entrenamiento	**0.77955**	**0.77955**
R2 pruebas	**0.752155**	**0.752157**
Overfitting	No	No
Variables Significativas	Beta>1500	Beta>1500

Modelo:
Regresión Lineal Múltiple
Comparación

La **magnitud de los betas** entre los 3 modelos entrenados es **similar**.

Para el modelo **Lasso ningún coeficiente se vuelve 0**. Sin embargo, **algunos Betas son comparativamente bastante pequeños** respecto a las otras variables, lo que nos muestra que el **aporte de la variabilidad que explican es muy bajo**.

Los **betas de los modelos regularizados** en general, **tienden a ser un poco menores respecto a los betas del modelo sin regularización para variables con menor poder predictivo.**

Variable	Beta_sin_regul	Beta_lasso	Beta_ridge
const	163508.432969	163497.953008	163503.422860
lat	20847.232153	20892.030098	20689.382513
lon	17710.659737	17132.410635	17399.957723
surface_total_in_m2	71349.333663	71352.371706	71294.877648
place_name_Balvanera	-5418.952753	-5420.443738	-5449.322368
place_name_Barracas	-1305.719223	-1242.211207	-1319.760624
place_name_Barrio Norte	5127.513927	5103.814610	5108.081641
place_name_Belgrano	7768.199687	7472.069575	7660.336846
place_name_Boedo	65.829708	27.850523	18.928390
place_name_Caballito	9895.714111	9676.487851	9706.183903
place_name_Capital Federal	1300.471841	1223.591872	1240.945051
place_name_Centro / Microcentro	-2412.192371	-2376.985845	-2410.790362
place_name_Chacarita	336.009957	240.513718	285.641519

Conclusión

Tomando un modelo simplificado, con superficie total, latitud, longitud y si el inmueble está en Palermo, Belgrano, Madero, Recoleta o Caballito, y dejando el resto de lugares clasificados como 'otro', seguramente tendríamos un modelo con buen rendimiento y mucho más sencillo.

Modelo:
Prueba datos dummy

Predicciones

Introduzca las características del inmueble para predecir su precio en la siguiente celda:

Nota: Elegir un place_name entre 'Belgrano', 'Palermo', 'Flores', 'Boedo', 'Balvanera', 'Caballito', 'Nuñez', 'Barrio Norte', 'Villa Crespo', 'Puerto Madero', 'Constitución', 'Recoleta', 'Colegiales', 'Villa Urquiza', 'Saavedra', 'Barracas', 'Coghlan', 'Almagro', 'San Telmo', 'Monserrat', 'Villa Devoto', 'San Cristobal', 'Floresta', 'Retiro', 'Capital Federal', 'Chacarita', 'Congreso', 'Villa del Parque', 'Liniers', 'Centro / Microcentro', 'Parque Patricios', 'Once', 'San Nicolás', 'Villa Luro'

```
property_type = 'apartment'  # Única opción
place_name = 'Saavedra'
lat = -34.556875
lon = -58.444444
surface_total_in_m2 = 70
```

```
La predicción de los modelos es:
 Modelo sin regularización: 241155.20361726292
 Modelo lasso: 179244.2095991452
Modelo Ridge: 179482.1856565378
```

MUCHAS GRACIAS