Proyecto Limpieza de Datos

PROPERATI

- Sankowicz, Javier
- Izquierdo, Stephanie
- Alcoleas, Noelia P.

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías
- Resolución de problemas de formato

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías
- Resolución de problemas de formato
- Asignación de formatos adecuados

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías
- Resolución de problemas de formato
- Asignación de formatos adecuados
- Corrección de valores erróneos

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías
- Resolución de problemas de formato
- Asignación de formatos adecuados
- Corrección de valores erróneos
- Completar datos faltantes

Alcance del Proyecto

A partir de la base de datos recibida, se iniciaron tareas de limpieza y organización de datos. Se pretende lograr una base de datos depurada, que sea eficiente para un posterior análisis y predicción

- Conocimiento de Base
- Estandarización de categorías
- Resolución de problemas de formato
- Asignación de formatos adecuados
- Corrección de valores erróneos
- Completar datos faltantes
- Organización correcta del dataset

El Dataset recibido, tiene 25 variables organizadas en columnas y consta de 121.220 registros en filas.

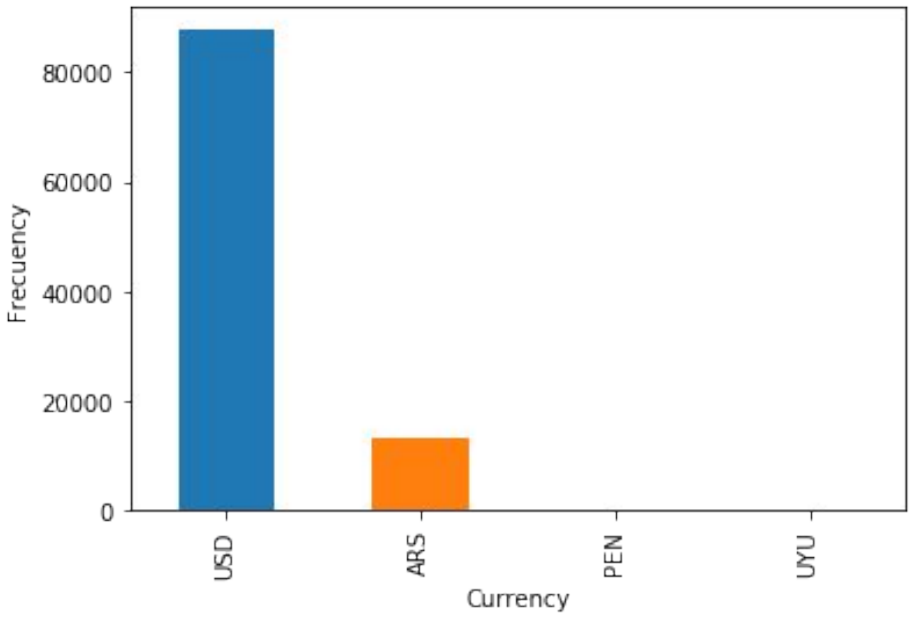
Variables Originales Presentes en el DataSet:

```
In [3]: data = pd.read_csv('properatti.csv', index_col=0)
        data.shape

Out[3]: (121220, 25)
```

Operación del aviso : sell, rent
Tipo de la propiedad : house, apartment, ph, store
Nombre del lugar
Nombre del lugar + nombre de sus 'padres'
País
Estado:
ID de geonames del lugar (si está disponible) Latitud,Longitud
Precio original del aviso
Moneda original del aviso : ARS, USD
Precio del aviso en moneda local : ARS
Precio aproximado en USD:
Superficie en m² : referido a la superficie total
Superficie cubierta en m² :
Precio en USD/m² : Precio aproximado en dólares / Superficie Total en m2
Precio por m²: Precio / superficie cubierta en m2
N° de piso, si corresponde:
Ambientes:
Expensas:
URL en Properati
Descripción
Título
URL de un thumbnail de la primer foto

En la Columna currency, encontramos avisos cotizados en pesos uruguayos y en moneda peruana. Se decidió reclasificar ambos valores a dólares.

```
In [14]: data['currency'].value_counts()

Out[14]: USD    87587
         ARS    13219
         PEN        2
         UYU        1
         Name: currency, dtype: int64
```



De la revisión de formatos, vimos:

- variables numéricas: datos tipo float
- variables categóricas: datos tipo object

```
In [13]: data.dtypes

Out[13]: operation                    object
         property_type                object
         place_name                   object
         place_with_parent_names      object
         country_name                 object
         state_name                   object
         geonames_id                 float64
         lat-lon                      object
         lat                         float64
         lon                         float64
         price                       float64
         currency                     object
         price_aprox_local_currency  float64
         price_aprox_usd             float64
         surface_total_in_m2         float64
         surface_covered_in_m2       float64
         price_usd_per_m2            float64
         price_per_m2                float64
         floor                       float64
         rooms                       float64
         expenses                    float64
         properati_url                object
         description                  object
         title                        object
         image_thumbnail              object
         dtype: object
```

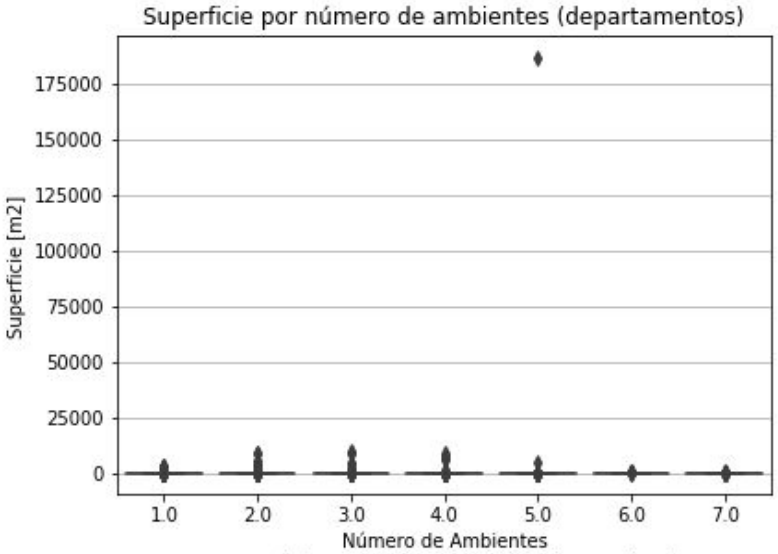
Corrección previa de los datos:

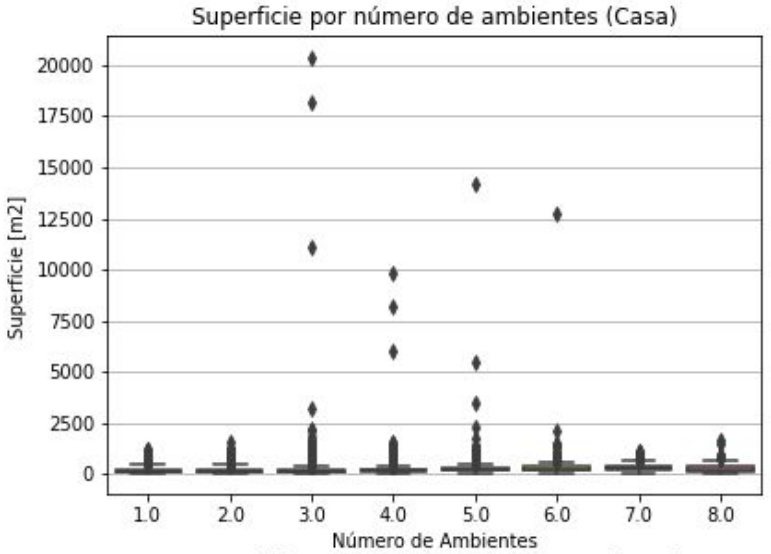
- Surface total in m2: Se determinó una superficie mínima válida.
- Place name: Datos faltantes y datos mal cargados, todos correspondientes a Tigre.
- Currency: Se convirtieron las monedas.
- Validación por Outliers

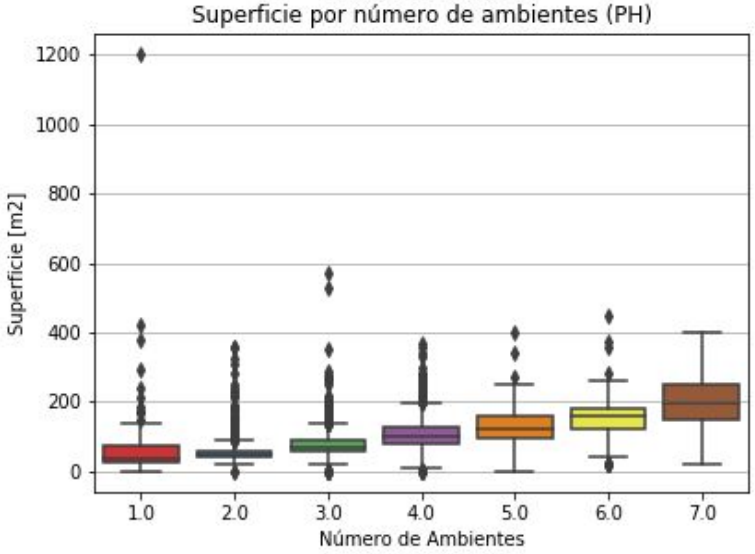
```
data.isnull().sum()
```

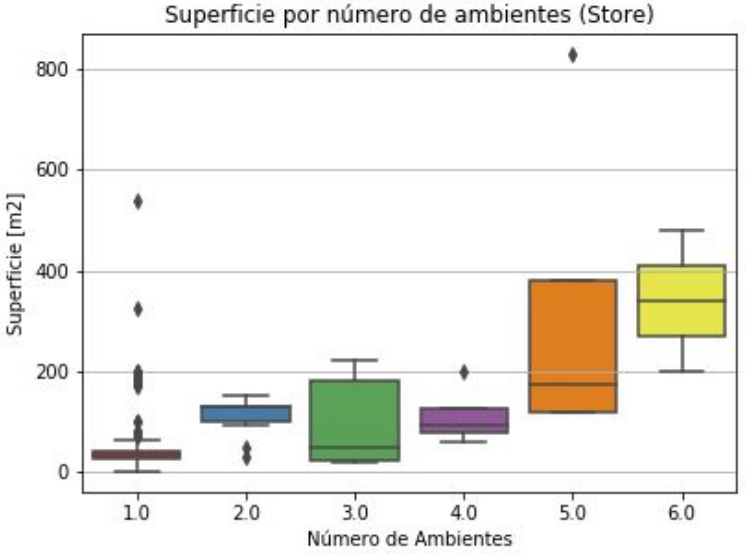
```
operation                     0
property_type                 0
place_name                   23
place_with_parent_names       0
country_name                  0
state_name                    0
geonames_id               18717
lat-lon                   51550
lat                       51550
lon                       51550
price                     20410
currency                  20411
price_aprox_local_currency    20410
price_aprox_usd           20410
surface_total_in_m2       39328
surface_covered_in_m2     19907
price_usd_per_m2          52603
price_per_m2              33562
floor                    113321
rooms                     73830
expenses                 106958
properati_url                 0
description                   2
title                         0
image_thumbnail            3112
dtype: int64
```

Validación de datos por Outliers Multivariados





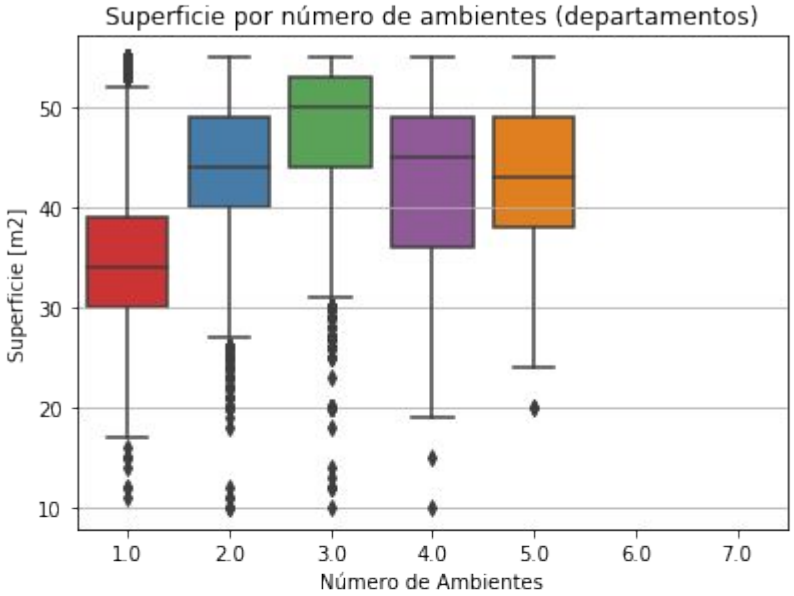


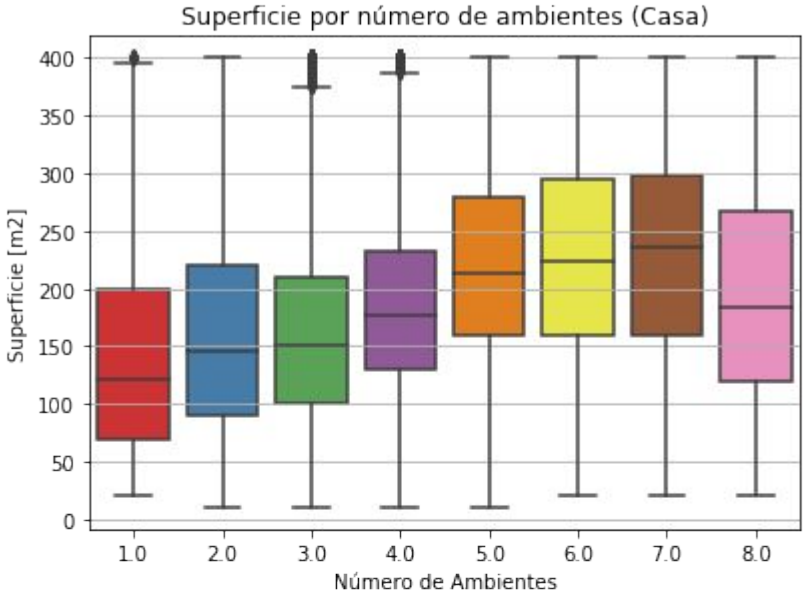


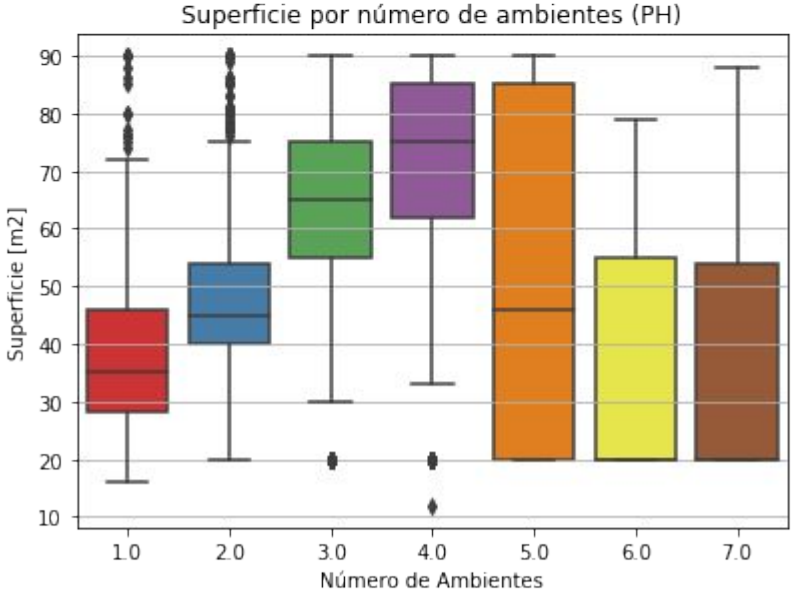
Corrección

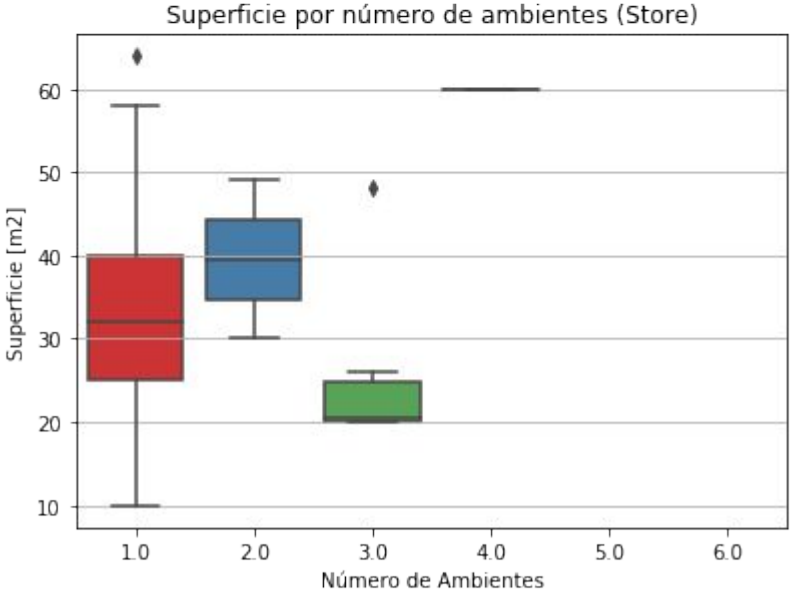
Valores Erróneos

PROPERATI







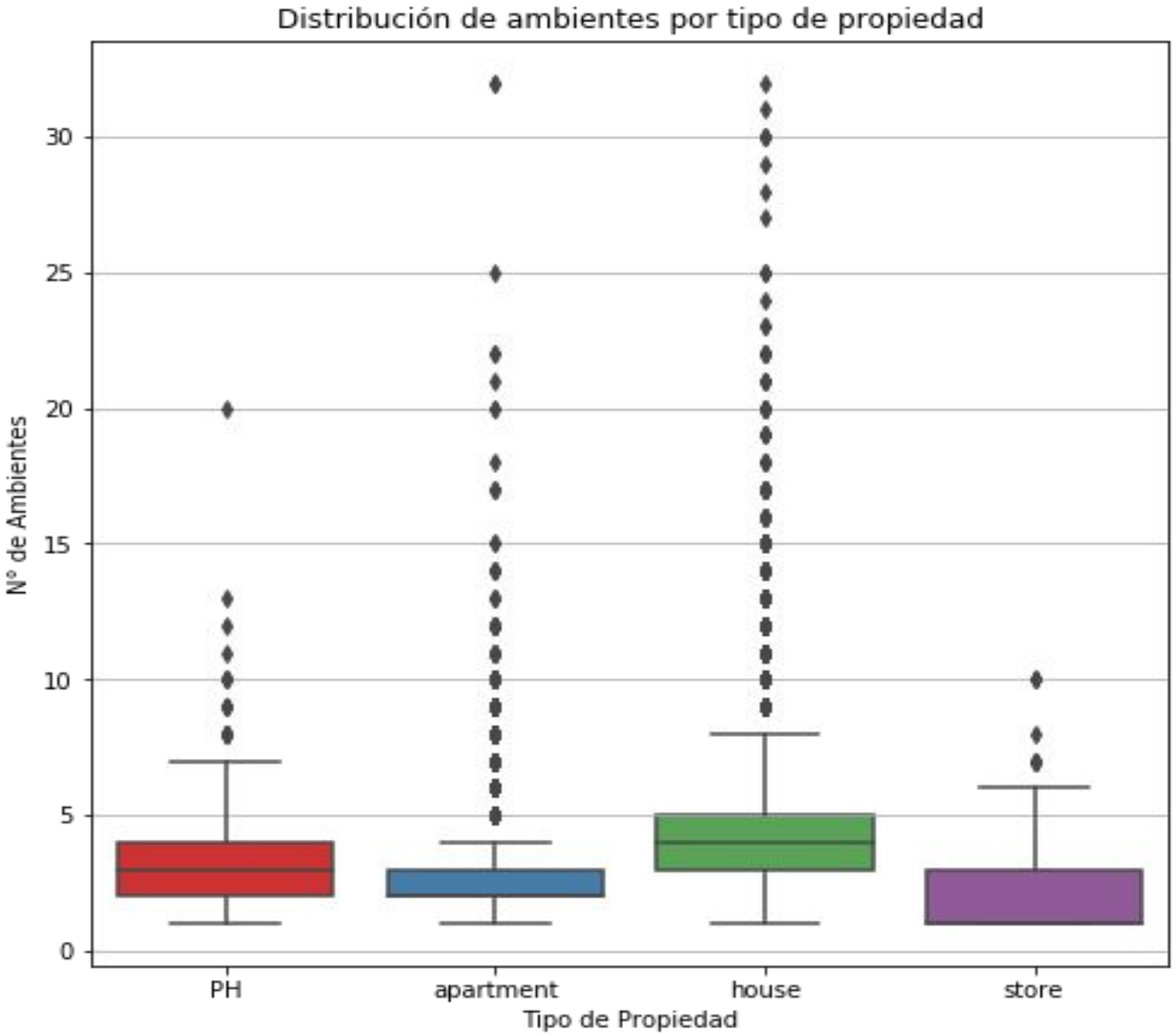


Corrección

Valores Erróneos

PROPERATI

ANTES

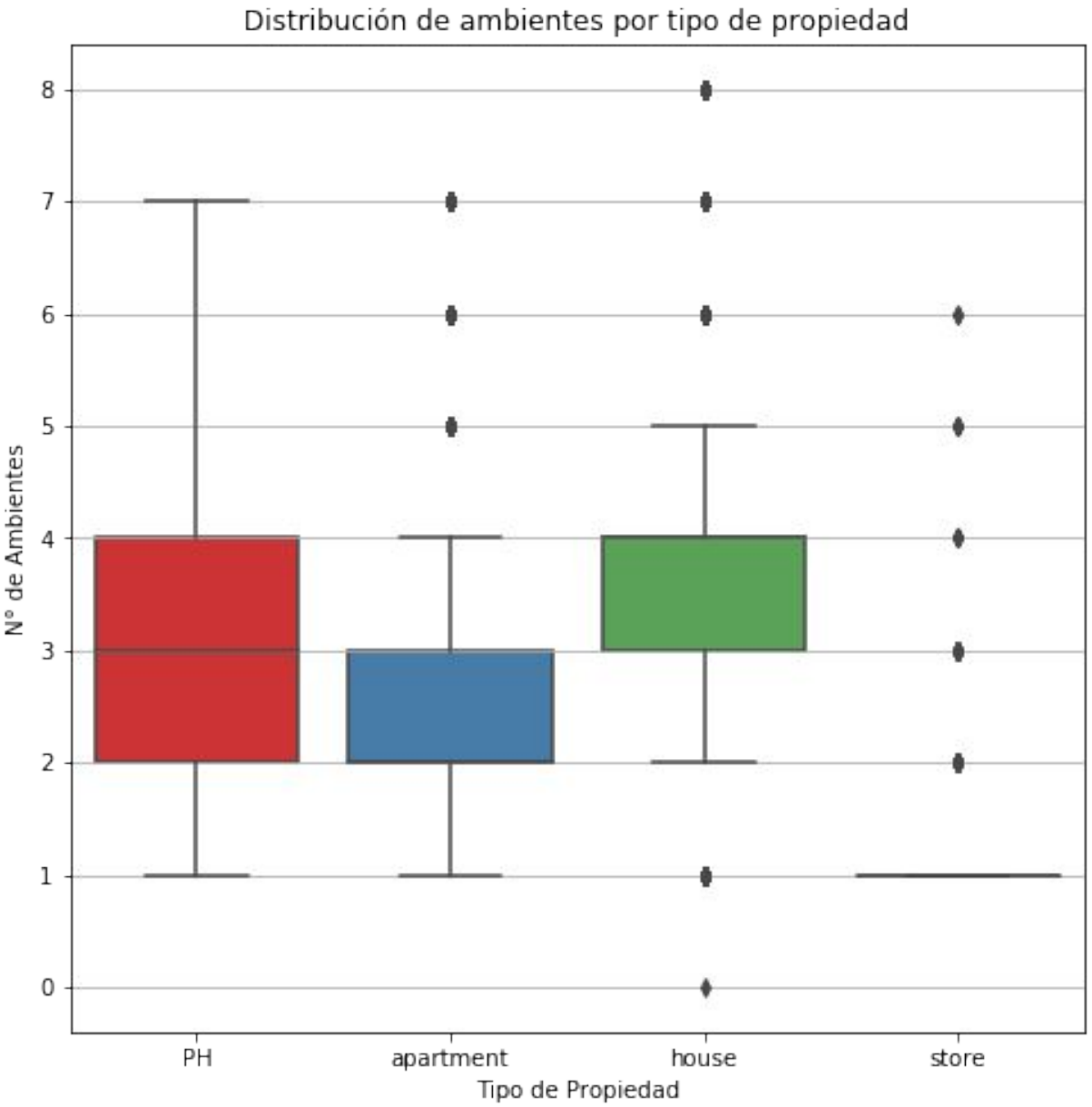


Valores Erróneos

Corrección

PROPERATI

DESPUÉS



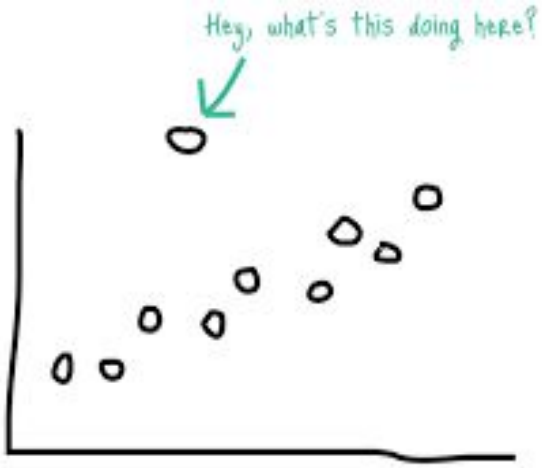
PROPERATI

Valores Erróneos

Corrección

Estrategias:

- Búsqueda con Regex
- Álgebra sobre Variables Dependientes
- Índices
- Imputación por media/mediana/moda.

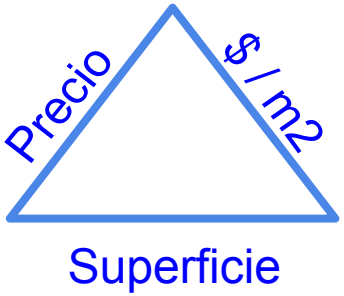


ANCIENT EGYPTIAN REGEXP

- **Expresiones Regulares:** Búsqueda sobre ‘title’ y ‘description’.

Departamento de 3 Ambientes al Contrafrente. Son 50 M2 totales . Piso 10° . El Edificio cuenta con Grupo Electrógeno. Antigüedad 50 años. Cocina 1.80 X 4.50. Lavadero 1.80 X 1.80, Living-Comedor 3 X 4.50, Dormitorio 2.80 X 3 con placard, Dormitorio 2.50 X 2.60 con placard.Baño Completo. Baulera Independiente en sótano.

- **Álgebra sobre Variables Dependientes**: Aplicadas sobre las distintas variables de precios y superficies presentes en el dataset.



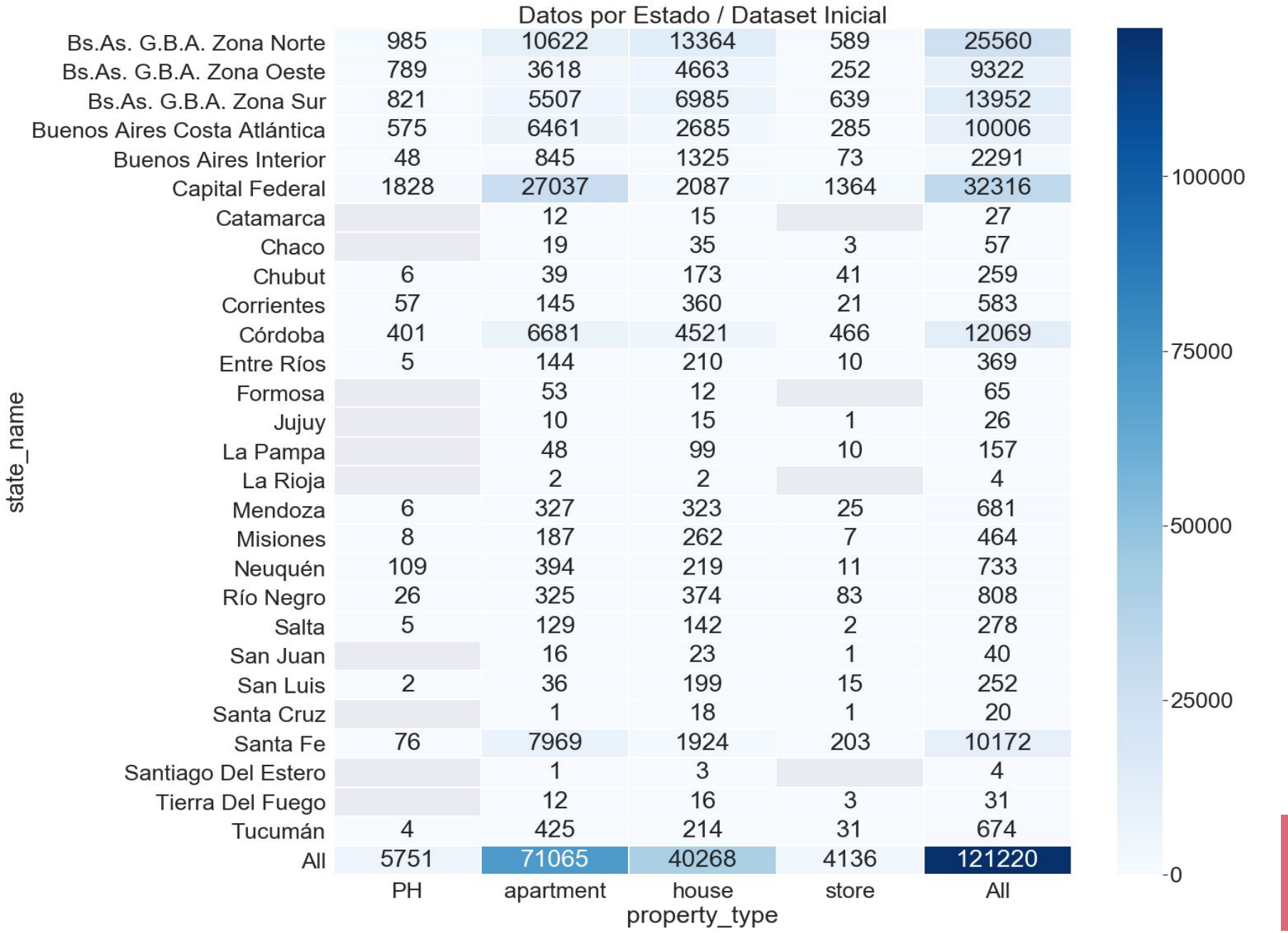
- **Construcción de Índices**

  - Índice de construcción por barrio: relaciona las superficies totales y superficies cubiertas por barrio

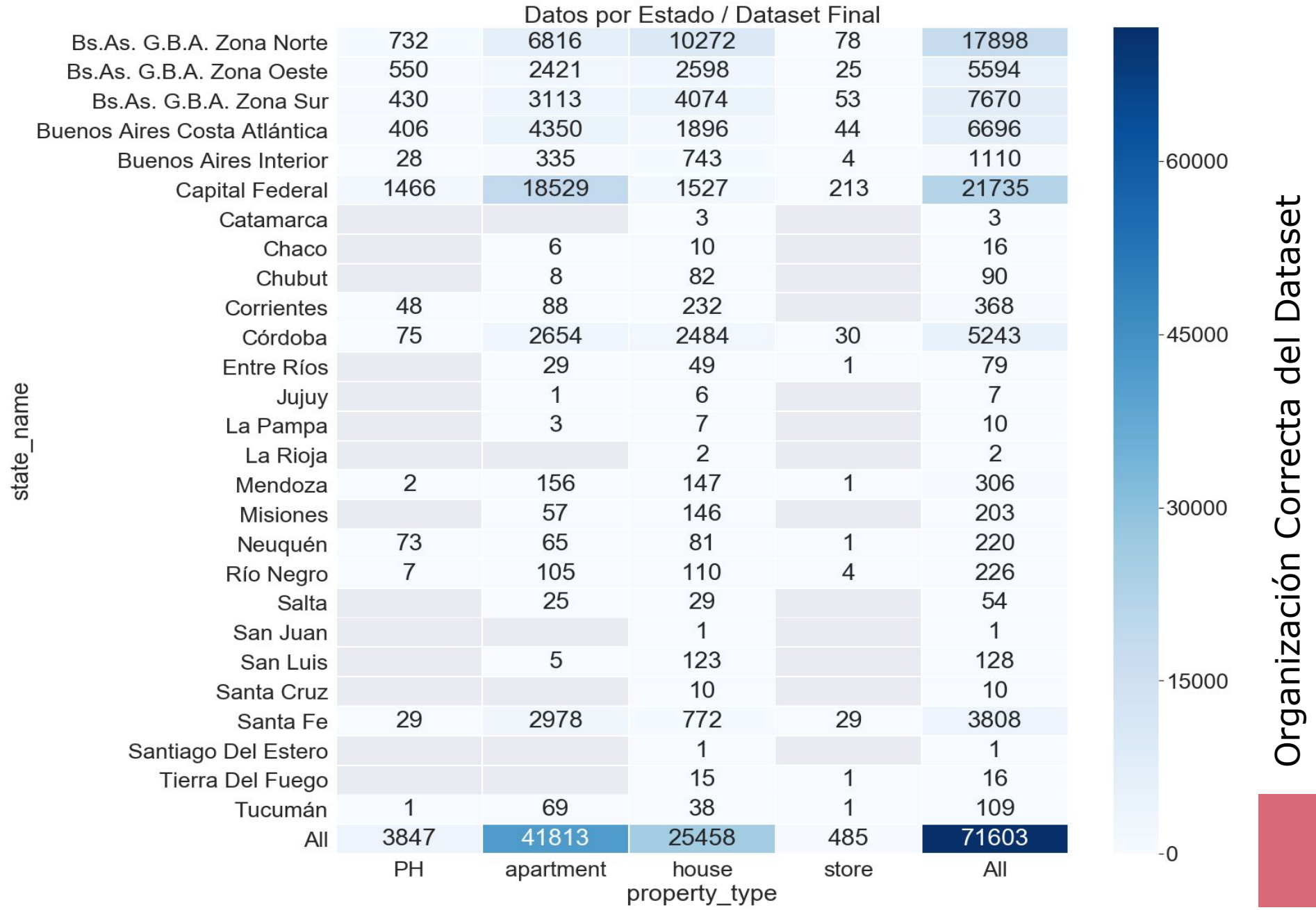
$$Indice\ de\ Construcción = \frac{SuperficieCubierta}{SuperficieTotal}$$

Dataset Final:

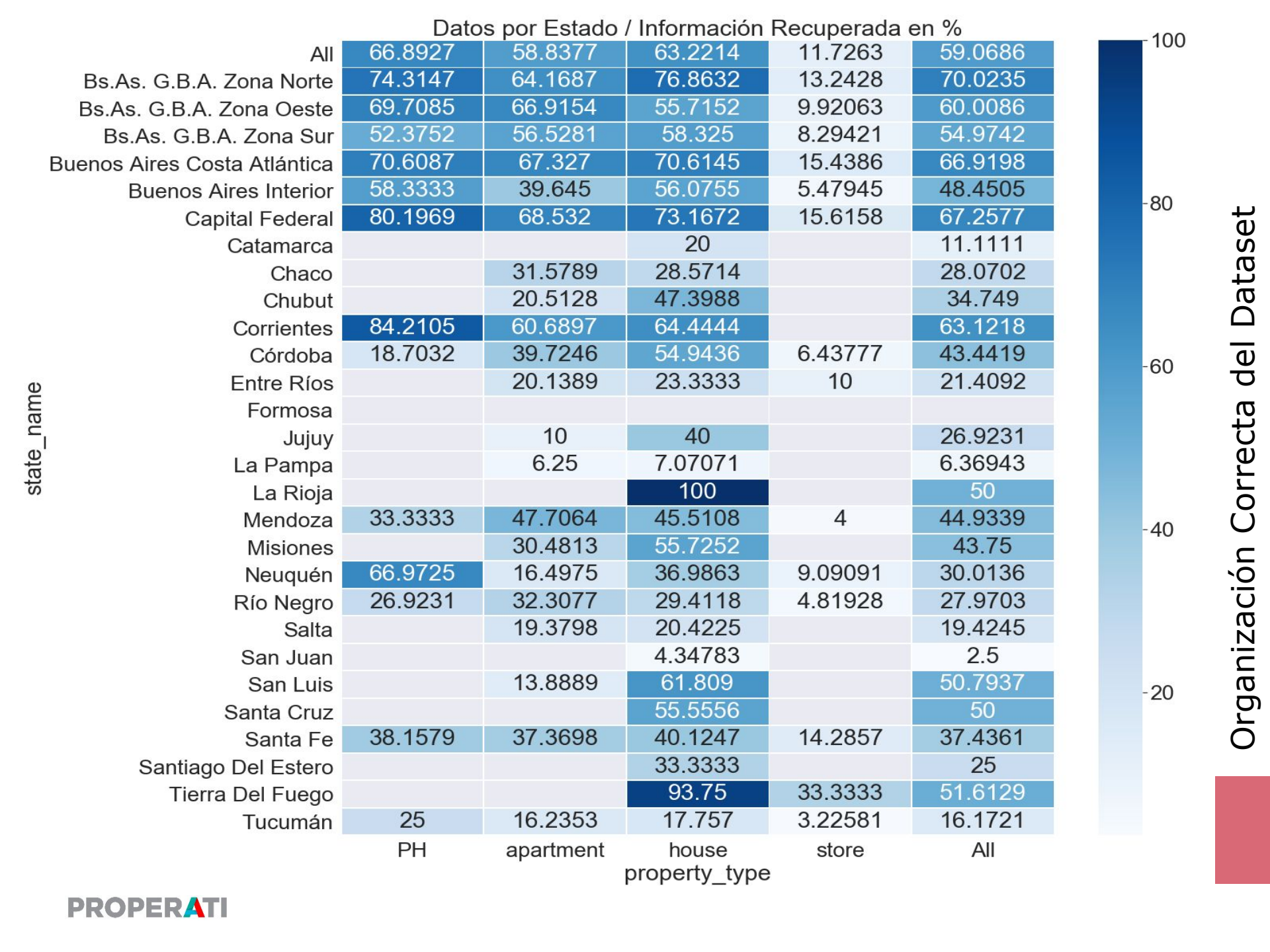
- Eliminación de columnas innecesarias.
- Tomamos en cuenta características que aportaran información relevante para la regresión del desafío 2.
- Drop de datos faltantes.
- Variables dummies: Cochera, Pileta, SUM, Patio, Balcón, etc.


Datos por Estado / Dataset Inicial
Bs.As. G.B.A. Zona Norte 985 10622 13364 589 25560
Bs.As. G.B.A. Zona Oeste 789 3618 4663 252 9322
Bs.As. G.B.A. Zona Sur 821 5507 6985 639 13952
Buenos Aires Costa Atlántica 575 6461 2685 285 10006
Buenos Aires Interior 48 845 1325 73 2291
Capital Federal 1828 27037 2087 1364 32316
Catamarca 12 15 27
Chaco 19 35 3 57
Chubut 6 39 173 41 259
Corrientes 57 145 360 21 583
Córdoba 401 6681 4521 466 12069
Entre Ríos 5 144 210 10 369
Formosa 53 12 65
Jujuy 10 15 1 26
La Pampa 48 99 10 157
La Rioja 2 2 4
Mendoza 6 327 323 25 681
Misiones 8 187 262 7 464
Neuquén 109 394 219 11 733
Río Negro 26 325 374 83 808
Salta 5 129 142 2 278
San Juan 16 23 1 40
San Luis 2 36 199 15 252
Santa Cruz 1 18 1 20
Santa Fe 76 7969 1924 203 10172
Santiago Del Estero 1 3 4
Tierra Del Fuego 12 16 3 31
Tucumán 4 425 214 31 674
All 5751 71065 40268 4136 121220
PH apartment house store All
property_type
state_name
100000
75000
50000
25000
0
Organización Correcta del Dataset

PROPERATI


Datos por Estado / Dataset Final
state_name
Bs.As. G.B.A. Zona Norte 732 6816 10272 78 17898
Bs.As. G.B.A. Zona Oeste 550 2421 2598 25 5594
Bs.As. G.B.A. Zona Sur 430 3113 4074 53 7670
Buenos Aires Costa Atlántica 406 4350 1896 44 6696
Buenos Aires Interior 28 335 743 4 1110
Capital Federal 1466 18529 1527 213 21735
Catamarca 3 3
Chaco 6 10 16
Chubut 8 82 90
Corrientes 48 88 232 368
Córdoba 75 2654 2484 30 5243
Entre Ríos 29 49 1 79
Jujuy 1 6 7
La Pampa 3 7 10
La Rioja 2 2
Mendoza 2 156 147 1 306
Misiones 57 146 203
Neuquén 73 65 81 1 220
Río Negro 7 105 110 4 226
Salta 25 29 54
San Juan 1 1
San Luis 5 123 128
Santa Cruz 10 10
Santa Fe 29 2978 772 29 3808
Santiago Del Estero 1 1
Tierra Del Fuego 15 1 16
Tucumán 1 69 38 1 109
All 3847 41813 25458 485 71603
PH
apartment
house
store
All
property_type
60000
45000
30000
15000
0
Organización Correcta del Dataset

PROPERATI


Datos por Estado / Información Recuperada en %
state_name
All 66.8927 58.8377 63.2214 11.7263 59.0686
Bs.As. G.B.A. Zona Norte 74.3147 64.1687 76.8632 13.2428 70.0235
Bs.As. G.B.A. Zona Oeste 69.7085 66.9154 55.7152 9.92063 60.0086
Bs.As. G.B.A. Zona Sur 52.3752 56.5281 58.325 8.29421 54.9742
Buenos Aires Costa Atlántica 70.6087 67.327 70.6145 15.4386 66.9198
Buenos Aires Interior 58.3333 39.645 56.0755 5.47945 48.4505
Capital Federal 80.1969 68.532 73.1672 15.6158 67.2577
Catamarca 20 11.1111
Chaco 31.5789 28.5714 28.0702
Chubut 20.5128 47.3988 34.749
Corrientes 84.2105 60.6897 64.4444 63.1218
Córdoba 18.7032 39.7246 54.9436 6.43777 43.4419
Entre Ríos 20.1389 23.3333 10 21.4092
Formosa
Jujuy 10 40 26.9231
La Pampa 6.25 7.07071 6.36943
La Rioja 100 50
Mendoza 33.3333 47.7064 45.5108 4 44.9339
Misiones 30.4813 55.7252 43.75
Neuquén 66.9725 16.4975 36.9863 9.09091 30.0136
Río Negro 26.9231 32.3077 29.4118 4.81928 27.9703
Salta 19.3798 20.4225 19.4245
San Juan 4.34783 2.5
San Luis 13.8889 61.809 50.7937
Santa Cruz 55.5556 50
Santa Fe 38.1579 37.3698 40.1247 14.2857 37.4361
Santiago Del Estero 33.3333 25
Tierra Del Fuego 93.75 33.3333 51.6129
Tucumán 25 16.2353 17.757 3.22581 16.1721
PH apartment house store All
property_type
100 80 60 40 20
Organización Correcta del Dataset

PROPERATI

¡Gracias!

PROPERATI