**Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones**

**Informe Trabajo Práctico I Mentoría**:

“Análisis del mercado inmobiliario de la Ciudad de Buenos Aires 2017”

**Mentor:** Javier Lezama

**Integrantes Grupo:**

-Navarro Agustina

-Calle Manuel

-Tagle Gabriel

**Mayo 2019**

# Introducción

En el siguiente informe se detalla EDA (Exploratory Data Analysis) sobre un dataset con el objetivo de brindar herramientas al lector que le permitan entender los datos, tomar decisiones y entender las relaciones de las variables más importantes que lo componen.

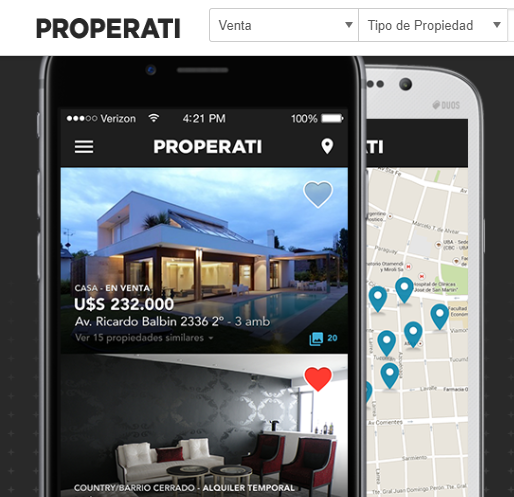
Para la confección del informe se desarrolló una notebook en Jupyter donde se aplicaron los conocimientos de la primera materia de la diplomatura: “Análisis y Visualización de Datos”.

En la notebook usamos las librerías de Numpy, Pandas, Seaborn y Stats que nos permitieron combinar herramientas de manejo de grandes volúmenes de datos, realización de gráficos y análisis estadísticos sobre el dataset.

El mentor de este grupo es Javier Lezama quien nos orientó en lo relacionado al contenido del dataset y a la descripción del caso para poder desarrollar este informe.

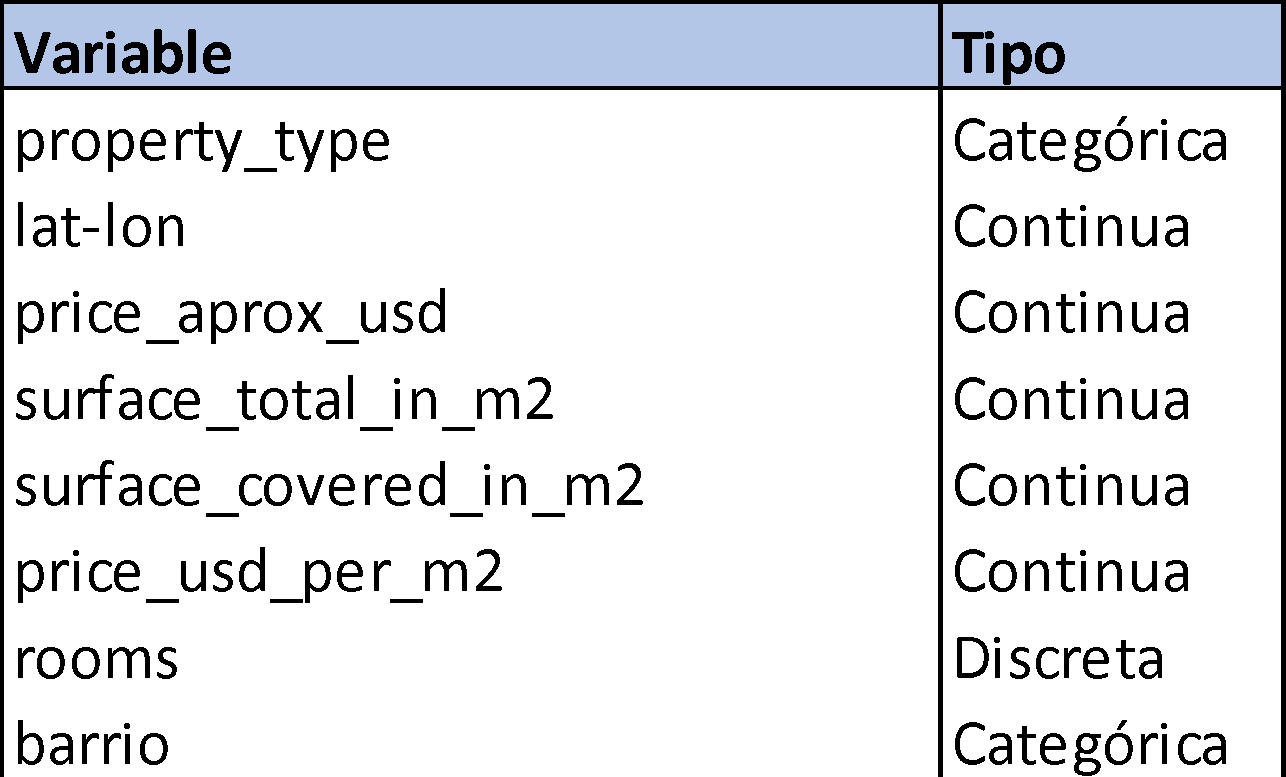
# Descripción del Dataset

El dataset se obtuvo del sitio web [www.properati.com.ar](http://www.properati.com.ar). Está página web almacena datos de propiedades de todo tipo y de varios lugares de Argentina. Este sitio ofrece a los usuarios herramientas para analizar propiedades de manera más eficaz y así poder tomar mejores decisiones.



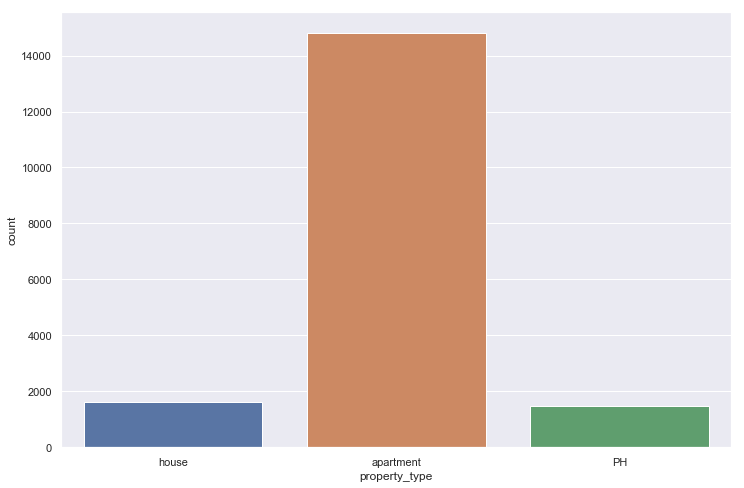
\*Imagen tomada del sitio web www.properati.com.ar

Este dataset contiene datos de propiedades de la ciudad de Buenos Aires del año 2017. En cuanto a la estructura de este, el dataset tiene un tamaño de 18979 filas y 17 columnas. Cada columna presenta información acerca una propiedad. La siguiente tabla resume las variables utilizadas en el análisis y define que tipo de variable corresponde:



# Limpieza de los datos

El dataset tiene dos variables categóricas, una son los barrios y otra el tipo de propiedad. Para comenzar vamos a definir las pautas de nuestro análisis, el cual se basa en analizar propiedades de vivienda. Por este motivo es que filtramos el data set para tener solo apartamentos, casas y PHs.

Una vez realizado el filtro, se observan las cantidades para cada tipo de propiedad: 

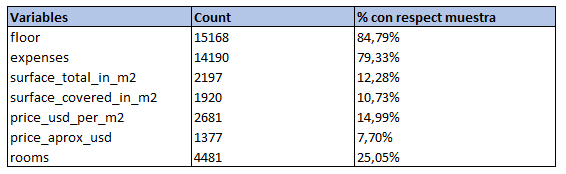
Con este gráfico podemos ver que la mayoría de los datos son departamentos. En estos casos se puede analizar todo completo o dividir la muestra en dos, ya que los departamentos, por tener gran volumen de datos podremos hacer un análisis más profundo y probablemente tenga más sentido evaluar datos de similares.

# Tratamiento de Valores Nulos

Para realizar un análisis productivo de los datos es necesario realizarles un tratamiento previo. Dicho tratamiento consta de la “limpieza de los datos”, verificando los valores que pueden alterar el resultado de nuestro análisis. En este dataset (como en la mayoría de los casos), nos hemos encontrado con valores faltantes. En primer lugar, tratamos de detectar a qué se debía este faltante de información, y cuáles eran las variables más afectadas para ver si influyen o no en el análisis.

En la siguiente tabla vemos la cantidad de valores nulos por variables y las posibles causas de cada uno.

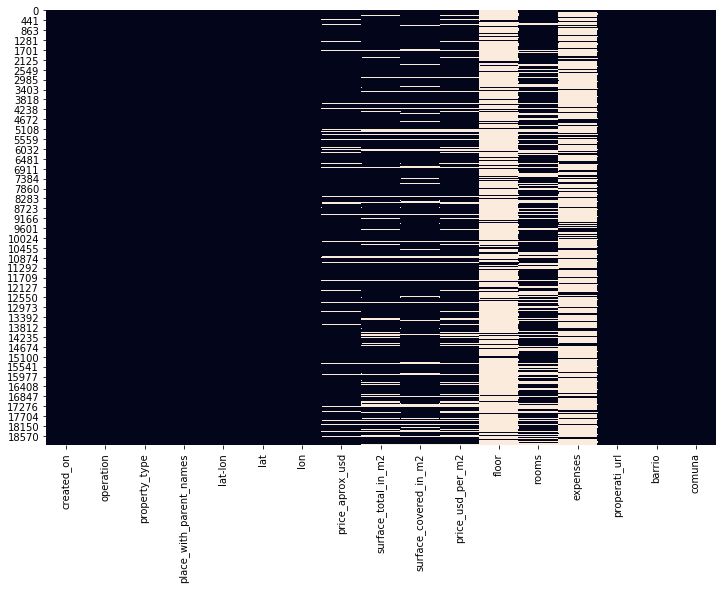
Variables cuantitativas y cantidad de datos.



* Pisos: La mayoría de los datos nulos en esta variable son departamentos, por lo cual podemos interpretar que por pertenecer a esta categoría no tienen más de un piso, por ello el dato es insignificante. No se deben eliminar estas filas ya que no es una variable que nos afecte en el análisis. Se pueden rellenar estos valores con 1, en los casos en que sea un departamento y una casa.
* Expensas: Las expensas no son un dato significante para el análisis del dataset, y al ser un monto que depende del barrio y las condiciones del complejo donde se ubique la propiedad, es muy diverso. Completar con valores como la media no sería una buena opción y eliminarlos tampoco, ya que son el 79% de la muestra
* Precio Total: Esta variable es una de las principales para realizar el análisis del mercado, por lo que vemos que muy pocas propiedades tienen valores faltantes. De todos modos, uno de los criterios que podemos tomar para rellenar estos valores es el de completarlos con el producto del precio por m2 y la superficie. Sin embargo, para aplicar este criterio analizamos si los valores que necesitamos tampoco son nulos, detectamos que en todos los casos es así, por ende, desestimamos este criterio. En segundo lugar, se puede tomar el promedio de precios, pero debe hacerse por zona, ya que cada barrio tiene una cotización diferente según su ubicación.
* Precio por m2: Este valor es un valor calculado, por lo que depende de las variables: precio total y superficie.
* Superficie Total: Esta también es una de las variables principales del análisis, la superficie nos dará la pauta de que tan grande es la propiedad, incluyendo zonas que no estén cubiertas por un techo.
* Superficie Cubierta: Esta variable considera los m2 bajo techo de una propiedad, la cual, debe ser siempre menor o igual a la superficie total. Para el caso de sus valores null, tendremos la opción de copiar el valor de superficie total. En estos casos no es factible obtener un promedio de otros valores ya que puede haber valores que nos desvíen la media y generen datos erróneos.
* Cantidad de ambientes: esta es una de las variables en las que un cuarto de los datos es nulo, y es una variable importante en el análisis ya que suele ser el filtro de búsqueda que más usan los clientes. Esta variable es totalmente discreta y la misma puede deducirse de un cálculo de la cantidad promedio de m2 por habitación, según el tipo de propiedad, ya que el tamaño no será el mismo en casas que en departamentos.

En el estudio del tratamiento de valores nulos vamos a demostrar en gráficos tres situaciones, una es utilizar el dataset con los valores reales, incluyendo nulos, luego el dataset sin los valores nulos y el dataset con los datos faltantes reemplazados por la media de la variable.

A continuación, se presenta el gráfico *heatmap* con los valores nulos de cada variable en la muestra.

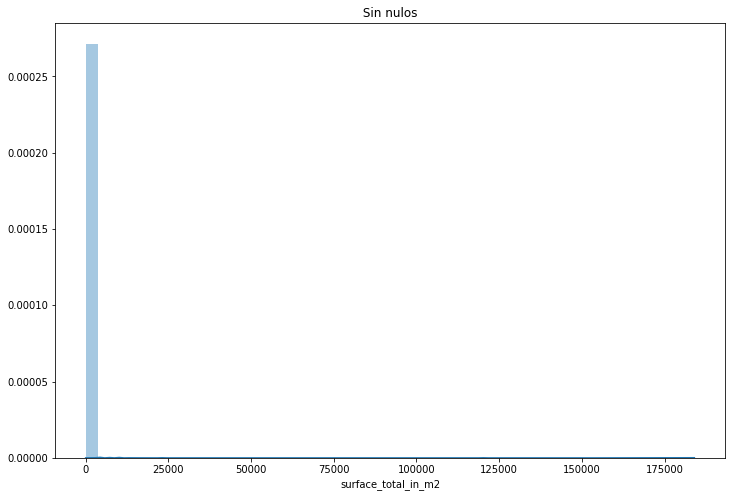
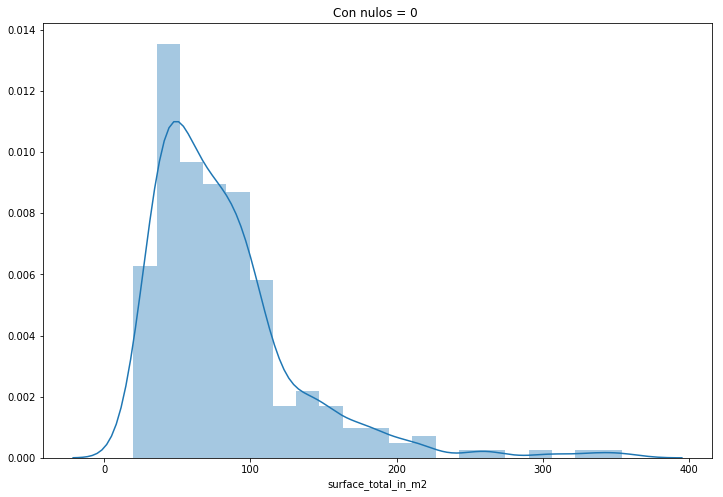


A continuación, se muestran las medias y los cuartiles del set de datos con valores nulos y sin valores nulos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset con valores nulos** | | | | | |
|  | **price\_aprox\_usd** | **surface\_total\_in\_m2** | **surface\_covered\_in\_m2** | **price\_usd\_per\_m2** | **rooms** |
| **count** | 1,65E+10 | 15.691 | 15.968 | 15.207 |  |
| **mean** | 2,88E+11 | 138,05398 | 119,053545 | 2.748,896095 | 3,024689 |
| **std** | 3,89E+11 | 1778,49500 | 1.540,225861 | 1.863,791284 | 1,666531 |
| **min** | 5,04E+09 | 0 | 1 | 958,333 | 1 |
| **25%** | 1,20E+11 | 48 | 43 | 2 | 1 |
| **50%** | 1,80E+11 | 74 | 66 | 2.555,555556 | 3 |
| **75%** | 3,10E+11 | 130 | 111 | 3.160,919540 | 4 |
| **Dataset sin valores nulos** | | | | | |
|  | **price\_aprox\_usd** | **surface\_total\_in\_m2** | **surface\_covered\_in\_m2** | **price\_usd\_per\_m2** | **rooms** |
| **count** | 1,20E+10 | 11.975 | 11.957 | 11.975 | 11.975 |
| **mean** | 3,01E+11 | 141,989729 | 122,899708 | 2.819,156920 | 3.031,566 |
| **std** | 3,87E+11 | 2.032,5793 | 1.776,244655 | 1.951,629155 | 1,.670844 |
| **min** | 5,04E+09 | 10 | 2 | 958,333 | 1 |
| **25%** | 1,25E+11 | 49 | 43,5 | 2.069,605943 | 2 |
| **50%** | 1,89E+11 | 75 | 67 | 2.611,7 | 3 |
| **75%** | 3,30E+11 | 129 | 110 | 3.226,539589 | 4 |

Si analizamos estos valores los que se ven más afectados son los de la superficie total, ya que la media difiere en los tres casos ampliamente, varía de 138 a 141 con valores nulos y sin ellos.

Podemos ver en histogramas como cambian las distribuciones en cada caso y los ejes como se modifican:



Vemos en cada caso que los ejes se modifican al eliminar los valores nulos y se eliminan todos los valores superiores a 400, es por eso que la media varía en cada caso.

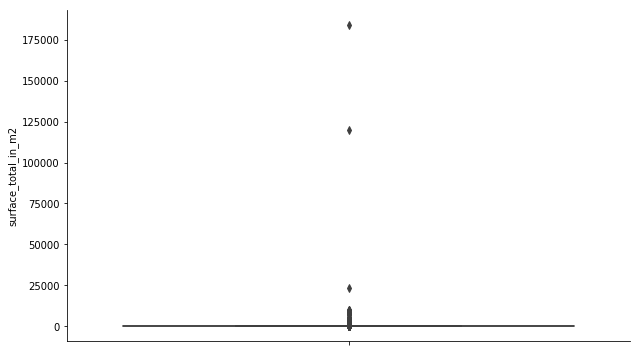
En el primer caso se muestra el histograma de los valores nulos en cero, esto elimina información ya que vemos la falta de datos superiores a 400 y una distribución diferente a los demás casos.

El primer desafío que se plantea es proponer una metodología para completar los mismos ya que la media es una medida no acertada, analizando los siguientes estadísticos vemos por qué:

* Media = 141
* Mediana = 75
* Moda = 50
* Cantidad de datos por sobre la media: 2628 de 17888 datos de la muestra

Lo que nos dice que la media no es representativa y está influenciada por valores extremos, ya que todos los valores se concentran entre 50 y 75 m2. Si realizamos un análisis por barrios de los valores mayores a la muestra, vemos que son propiedades en barrios privados y casas extremadamente grandes; es por ello que el análisis debemos hacerlo por barrios y tipos de propiedades.

Y si analizamos los cuartiles de la superficie en un gráfico de cajas:



Claramente por el análisis de la variable superficie total, vemos que tiene valores outliers, y que la media está afectada por estos valores muy grandes que la desplazan hacia valores mayores. Dichas propiedades son pocas para el análisis de todas las propiedades del dataset y no son representativas al promedio, ya que solo 2628 propiedades están por encima de la media, lo que no representa ni el 25% del total de la muestra. Por lo tanto, deberíamos realizar el análisis por barrios para ver si estos datos que parecen outliers en el dataset completo, lo son por barrios.

Por este motivo los criterios planteados son los siguientes:

**1. Superficie total calculada en base al *precio* total y al precio por m2 de cada propiedad por barrio**

**2. Superficie total calculada en base a la cantidad de *habitaciones* según tipo de propiedad y barrio**

**3. Superficie total calculada como la media de las *superficies* por tipo de propiedad del mismo barrio**

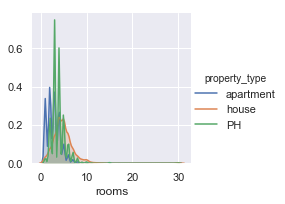
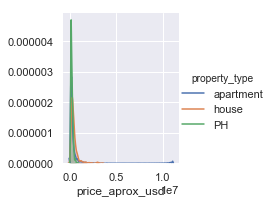
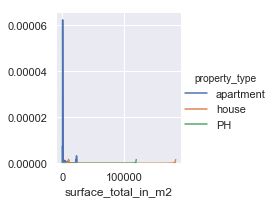
El primer criterio que vamos a tomar es, calcular la superficie total en base al precio total y el precio por m2, en el caso que alguno sea nulo, tomamos el promedio de dicha variable teniendo en cuenta la zona o barrio de la propiedad.

En el caso en que alguno de estos valores sea nulo, tomamos el criterio de la cantidad de habitaciones. Se realiza un promedio de cuál es la superficie por habitación, teniendo en cuenta el tipo de propiedad, ya que un departamento no va a tener los mismos ambientes que una casa y en base a la cantidad de cuartos se calcula los m2 totales.

En el caso en que todos los valores a considerar (precio, precio x m2 y rooms) sean nulos, se realizará un promedio de los m2 por tipo de propiedad y por zona.

El orden expuesto anteriormente es el que se considera al momento de su aplicación y cálculo de la superficie total, ya que si bien el que podría considerarse más aproximado es el número 1, no siempre contamos con los datos ya que muchas veces los valores del precio también son faltantes junto con los de la superficie total.

A continuación, vemos los gráficos de la distribución de cada variable por el tipo de propiedad:



**Análisis del criterio 1: Superficie calculada en base al precio por barrio**

Se realiza el análisis por barrio, del promedio del Precio total de la vivienda sobre el Promedio del precio por m2 y así calculamos la superficie total. Esto lo agregamos al dataset reemplazando los nulos por estos valores. Un ejemplo del cálculo en 5 barrios.



Este valor calculado de la superficie nos da un ***promedio de error de 90,72 m2****,* es decir que dicho valor es diferencia absoluta entre los valores calculados y los valores existentes, la confianza o certidumbre del cálculo.

Si calculamos la media de la superficie total real con valores nulos y la comparamos con la media de la superficie total calculada, se observa que al reemplazar los valores nulos por la media del barrio se disminuye la media real del dataset. Además, si comparamos con la media de la media por barrio vemos que el valor se aleja un poco más de la media total. Por lo tanto, no es una buena medida de media para el dataset.

* Media de surface\_total\_in\_m2\_recalculada = 130.894722
* Media de surface\_total\_in\_m2 (valor original) = 136.451437
* Media de surface\_total\_in\_m2\_mean (por barrio) = 101.27734 6

**Análisis del criterio 2: Superficie calculada en base a la cantidad de habitaciones por tipo de propiedad**

El cálculo consiste en obtener el valor promedio de los m2 que tiene una habitación por tipo de propiedad. Teniendo la cantidad de habitaciones, sobre los m2 totales de la propiedad, obtenemos la superficie de cada pieza.



Una de las complicaciones de aplicar este criterio fue que la variable *rooms* tenía valores nulos, los mismos fueron completados por la media de habitaciones por tipo de propiedad, así como el dato de los m2 por habitación. Con estos valores los llevamos al dataset completo y verificamos el error de esta nueva superficie calculada. En este caso el ***error absoluto es de 98,95 m2****.* Dicho error es mayor al anterior, por lo que no se considera adecuado.

Si calculamos la media de la superficie total real con valores nulos y la comparamos con la media de la superficie total calculada, se observa que al reemplazar los valores nulos por la media de habitaciones por tipo de propiedad aumenta la media real del dataset. Además, si comparamos con la media de la media de habitaciones por tipo propiedad vemos que el valor se aleja un poco más de la media. Por lo tanto, no es una buena medida de media para el dataset

* Media de surface\_total\_in\_m2\_recalculada = 142.876553
* Media de surface\_total\_in\_m2 (valor original) = 136.451437
* Media de surface\_total\_in\_m2\_mean (por barrio) = 143.505702

**Análisis del criterio 3: Superficie calculada en base al promedio de superficies por tipo de propiedad**

En primer lugar, calculamos el promedio por tipo de propiedad:



Luego, llevamos esta media al dataset completo y calculamos el error, ***el mismo es de 98,95***, por lo que comparando con las anteriores no nos conviene asignar esta medición para el reemplazo de valores nulos. Verificando la media de las superficies:

* Media de surface\_total\_in\_m2\_recalculada = 143.505702
* Media de surface\_total\_in\_m2 (valor original) = 136.451437
* Media de surface\_total\_in\_m2\_mean (por barrio) = 143.505702

En conclusión, el criterio que deberíamos aplicar es el primero, ya que obtuvimos un error menor que los demás criterios y no genera pérdidas de información. Por lo que ***la superficie de los datos faltantes puede obtenerse de los promedios de precios totales por barrios sobre el promedio de precio por m2 por barrio***, de modo que el análisis por barrio es un indicador del precio y el tipo de propiedad un indicador importante en el análisis de la superficie.

# Datos erróneos

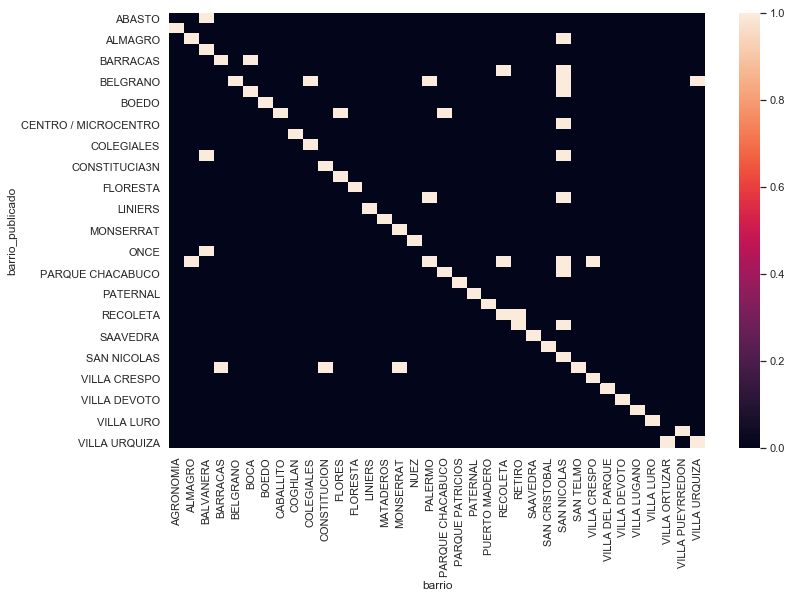
En este caso se analizan los datos que pueden tener errores en el dataset, como son:

* La superficie total debe ser mayor o igual a la superficie cubierta: esto consiste en armar un vector con los datos que no cumplen esta condición e intercambiar los valores.
* Las latitudes y longitudes deben coincidir con los barrios y algunas pueden repetirse para simular algún barrio o ubicación con más popularidad.

Resolviendo este último punto, realizamos un conteo de los barrios que más se repiten y se compara con el barrio publicado, obteniendo la siguiente tabla: (se muestran los que tienen mayor frecuencia y no coinciden en el barrio que se publica y el verdadero)

|  |  |  |  |
| --- | --- | --- | --- |
| **barrio\_publicado** | **barrio** | **conteo** | **barrio\_equals** |
| BARRIO NORTE | RECOLETA | 9224 | False |
| CONGRESO | BALVANERA | 1909 | False |
| NUAEZ | NUEZ | 1710 | False |
| LAS CAAITAS | PALERMO | 1331 | False |
| PALERMO | SAN NICOLAS | 1144 | False |
| CABALLITO | PARQUE CHACABUCO | 679 | False |
| BELGRANO | PALERMO | 629 | False |
| CENTRO / MICROCENTRO | SAN NICOLAS | 620 | False |
| CABALLITO | FLORES | 540 | False |

Acompañando esta tabla presentamos un mapa de calor de los barrios que coinciden y los que no:



En esta tabla podemos ver los barrios que son diferentes entre lo publicado y la ubicación real, por ejemplo, el barrio San Nicolás ha sido reemplazado por la zona Centro y Palermo, porque estas zonas son las que “venden más”. Y en el mapa de calor, encontramos en la diagonal los valores que coinciden y por fuera los valores que difieren entre sí, es decir los barrios que quieren ser otros.

# Quitando valores extremos

Los valores extremos son valores que distorsionan nuestra media y el análisis de los estadísticos de la muestra. A continuación, vamos a detectar que variables contienen valores extremos por medio de los gráficos de cajas:

Gráfico de cajas de cantidad de habitaciones por tipo de propiedad: vemos el valor extremo de 30 habitaciones que puede ser un dato mal cargado o hacer referencia a un hotel, pero este dato no es de interés para el análisis. Notamos que es un dato erróneo por estar en las tres categorías y ser totalmente irracional.

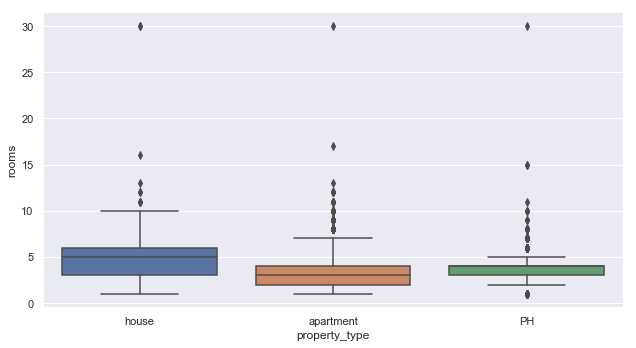


Gráfico de cajas de precio total por tipo de propiedad: vemos en este gráfico todos los valores que están por fuera de la distribución de los datos. Es por esto que se deben eliminar para evitar variaciones en las conclusiones del análisis.

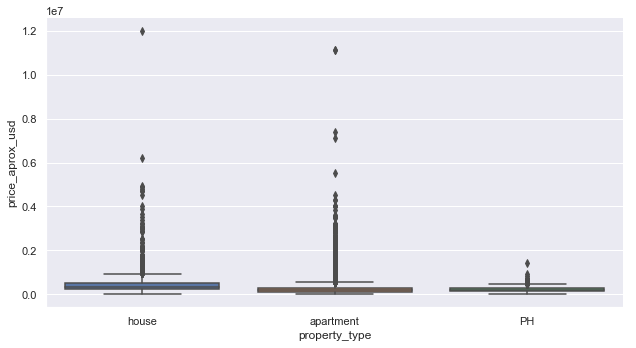
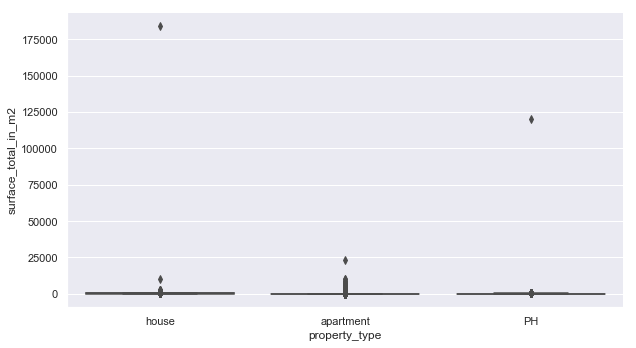
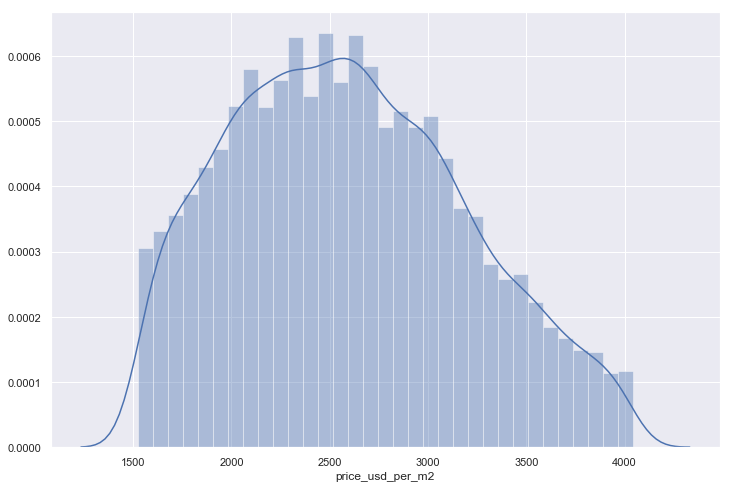
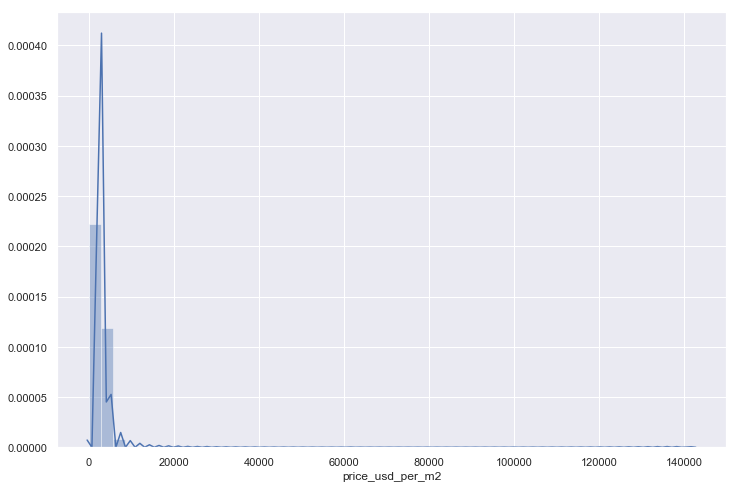


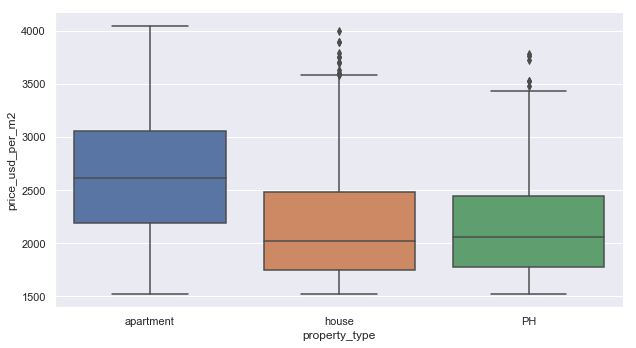
Gráfico de cajas de superficie total por tipo de propiedad: vemos en este gráfico los valores que están por fuera de la distribución de los datos, alguno de ellos puede estar relacionado con el registro de las 30 habitaciones.

En nuestro análisis vamos a demostrar cómo influyen lo valores extremos en la variable de precio. Al eliminar estos valores, se modifican el rango máximo y mínimo de las variables, su media y su desviación, ya que los datos estarán concentrados en los valores correctos. A continuación, presentamos el histograma de la variable con y sin los valores extremos.



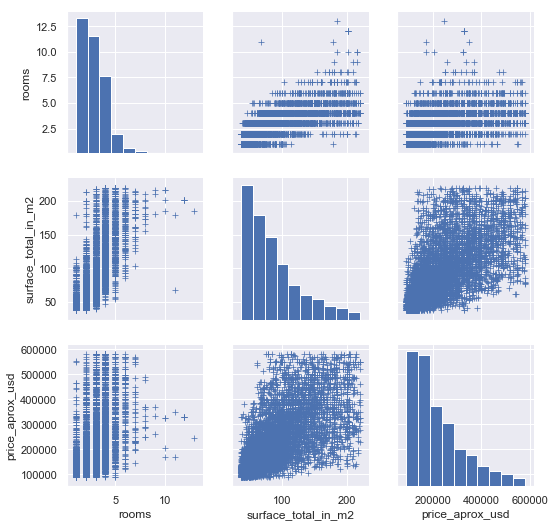
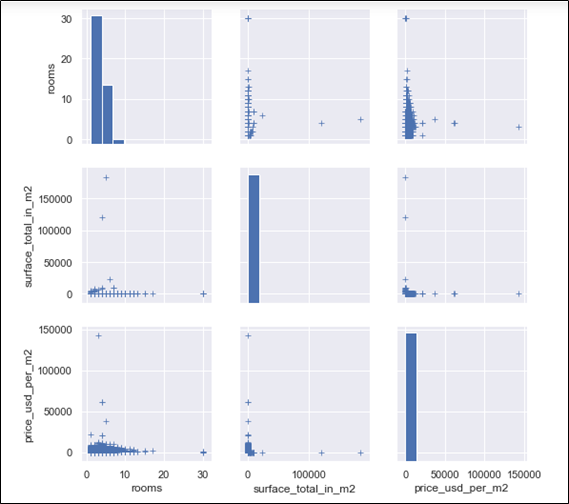
El histograma nuevamente nos muestra que la distribución no tiene grandes picos en los valores y se acerca más a una campana.

En este caso la mediana no cambia ya que la misma indica el valor medio de la distribución y como hemos quitado el 10% de cada lado de la distribución la mediana no cambió. Lo vemos en un gráfico de cajas:



A medida que profundizamos en el análisis podemos hacer un pairplot para ver la relación entre las variables y cómo afecta en ellos los extremos:

Pariplot con valores extremos y sin valores extremos:

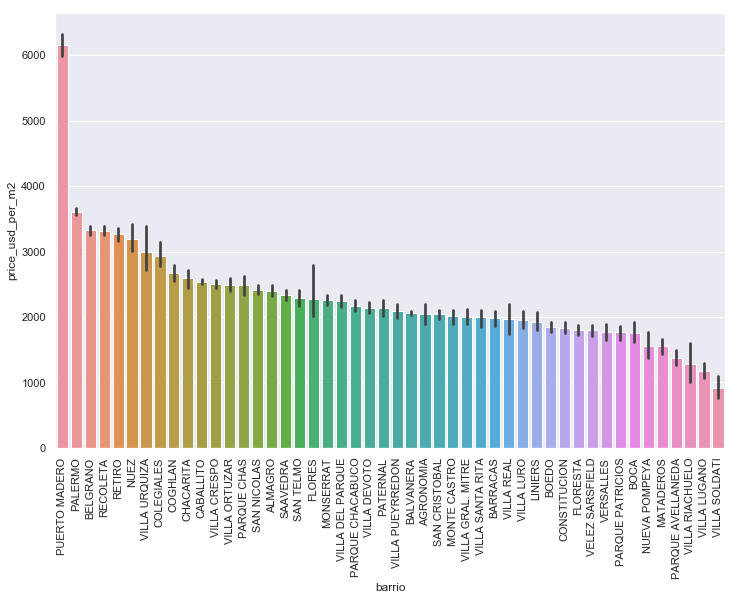


Con estos gráficos podemos ver como es la relación entre las variables quitando los valores que distorsionan las métricas y los análisis. Si nos hubiéramos quedado con los datos sucios las conclusiones serían muy distintas y quizás con algo de sesgo por estos valores distorsionantes.

# Análisis de las propiedades de los barrios

## Barrios más caros y más baratos

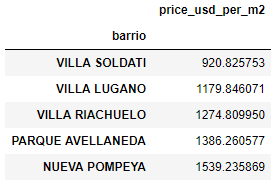
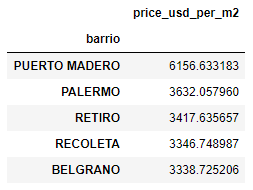
El siguiente gráfico nos provee una información importante a la hora de analizar el precio en dólares por metro cuadrado:



Lo importante al observar este gráfico es que el barrio con precio en dólares por metro cuadrado mayor es PUERTO MADERO, sacando una gran diferencia con el resto de los barrios. Y el de menor precio es VILLA SOLDATI.

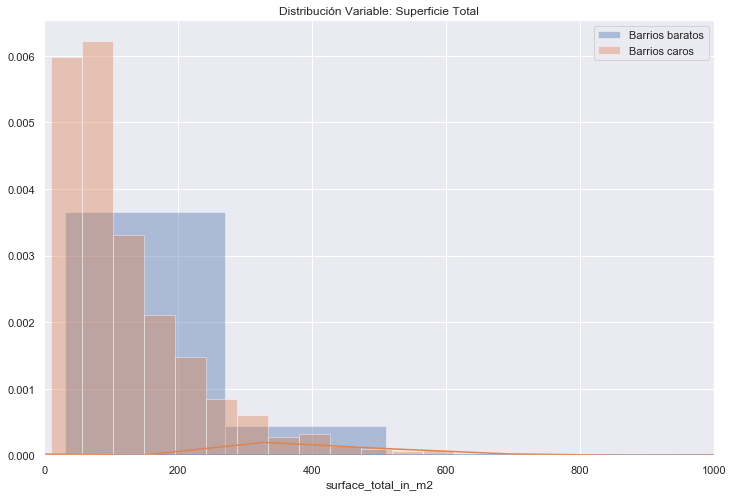
El precio por metro cuadrado es un buen indicador de los precios que se manejan en cada barrio el metro cuadrado sin importar el tipo de propiedad que se vende.

Realizando unas consultas sobre el dataset observamos los 5 barrios más caros (a la izquierda) y los 5 barrios más baratos (a la derecha)

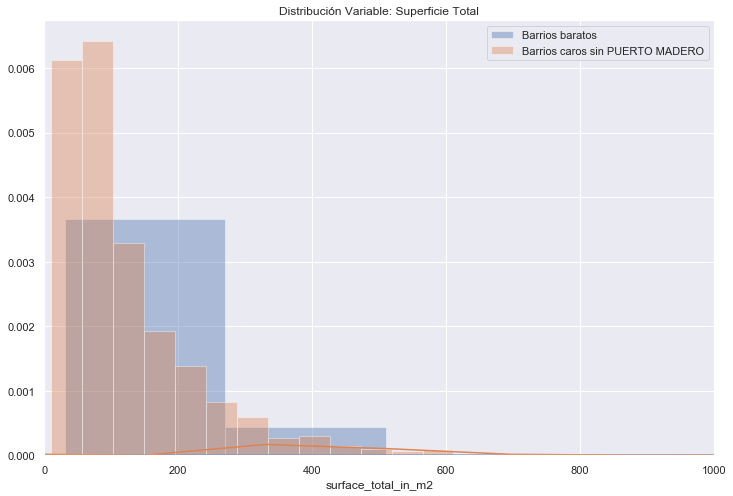


Comparando vemos que PUERTO MADERO tiene un valor de casi 7 veces VILLA SOLDATI por metro cuadrado y es casi el doble que PALERMO que es el segundo barrio más caro. Lo que indica una gran diferencia económica con el resto de los barrios marcando que es una zona de vivienda de gente con mucho dinero.

**Análisis de superficie total de barrios más caros y baratos**



Si analizamos el gráfico (sobre la variable "surface\_total\_in\_m2"), podemos observar muy marcado para "barrios caros", que la distribución parece una exponencial, indicando que las propiedades más caras, tienen menor superficie.



Realizamos el mismo gráfico, pero sacando del dataset al barrio PUERTO MADERO para ver si influye en la distribución y se observa que no influye en el gráfico.

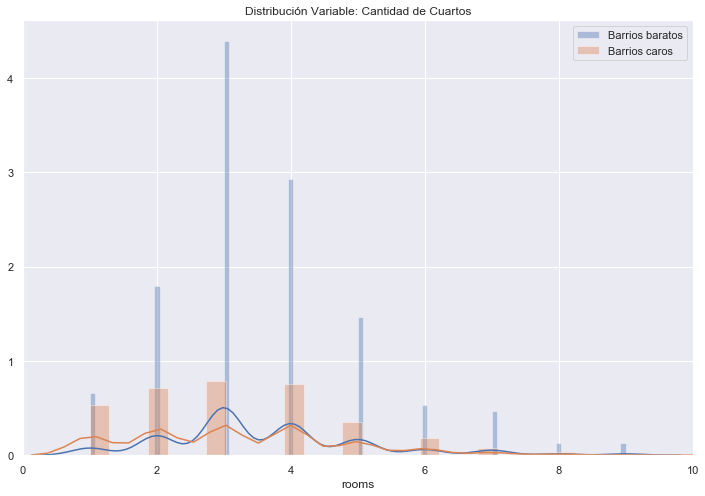
Para el caso de "barrios baratos" también parece una gráfica con distribución exponencial, pero se visualiza escalonada en donde se observa la misma cantidad de valores para un rango de superficies y luego la misma cantidad, pero para otro rango de superficie.

Si realizamos un análisis de asimetría sobre estas variables obtenemos:

* CAF Superficie Total barrios baratos: 13.598080673240528
* CAF Superficie Total barrios caros: 36.04172459279574

Lo que nos indica que ambas gráficas son asimétricas hacia la izquierda por el valor positivo obtenido. Además, como observamos en la gráfica, el mayor número de observaciones de la muestra se encuentran ubicados alrededor de los 100 mts2 para barrios caros y de 200 para barrios baratos, y luego mientras más crece el tamaño de las propiedades hay menos propiedades en la muestra.

**Análisis de habitaciones de barrios más caros y baratos**



Si analizamos el gráfico (sobre la variable "rooms"), podemos observar claramente que la distribución es sobre una variable discreta como es habitaciones y vemos los puntos del 0 al 10 donde se concentran todas las observaciones siento entre el 2, 3 y 4 donde se observa la mayor cantidad de datos.

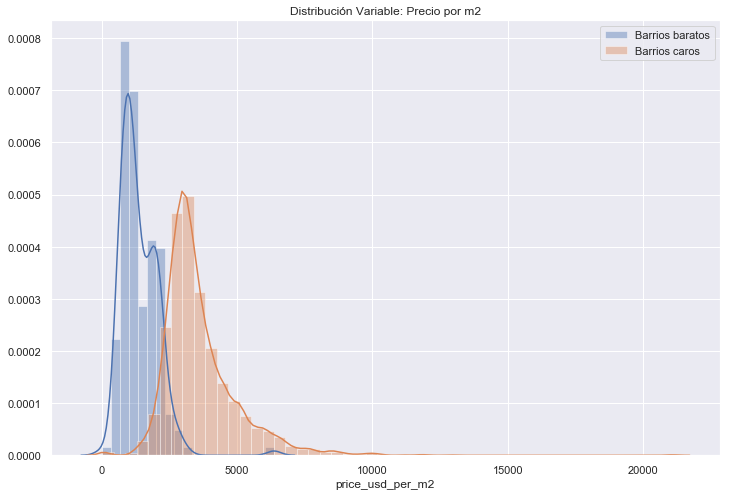
La distribución de "barrios caros" posee un achatamiento en comparación con la de "barrios baratos" ya que la cantidad de barrios del dataset es menor para barrios caros.

Si realizamos un análisis de asimetría sobre estas variables obtenemos:

* CAF Habitaciones barrios baratos: 0.935589064594072
* CAF Habitaciones barrios caros: 2.2639885221102904

Lo que nos indica que ambas gráficas son mínimamente asimétricas hacia la izquierda por el valor positivo obtenido, pero al ser muy cercano a cero entendemos que es mínima la asimetría. Para barrios caros observamos un número mayor y como se ve en la gráfica mayor distribución a la izquierda indicando que la asimetría es mayor, aunque es mínima la diferencia.

**Análisis de precio por m2 de barrios más caros y baratos**



Si analizamos el gráfico (sobre la variable "price\_usd\_per\_m2"), podemos observar para "barrios caros", que la distribución tiene comportamiento de una normal.

Para el caso de "barrios baratos" también parece una gráfica con distribución normal, pero con un “pico” intermedio que varía la distribución normal.

Si realizamos un análisis de asimetría sobre estas variables obtenemos:

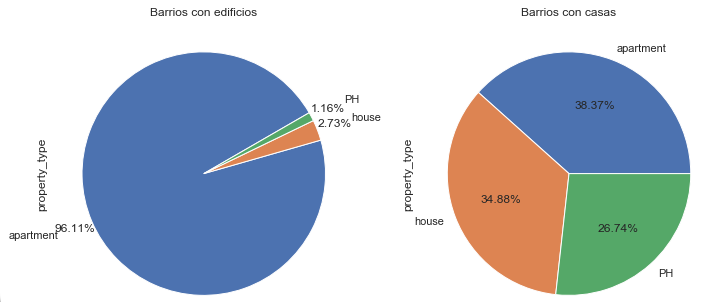
* CAF Precio Por m2 barrios baratos: 2.298934408944905
* CAF Precio Por m2 barrios caros: 4.695993016295685

Lo que nos indica que ambas gráficas son asimétricas hacia la izquierda por el valor positivo obtenido, pero al ser cercano a cero entendemos que es poca la asimetría.

Si observamos barrios caros, la asimetría se observa ya que la gráfica tiende a una normal, pero tiene mayor cantidad de valores a la izquierda del valor medio. Si observamos barrios baratos, la asimetría es menor ya que el rango de valores es menor y se encuentran menos distribuidos observándose una asimetría hacia la izquierda.

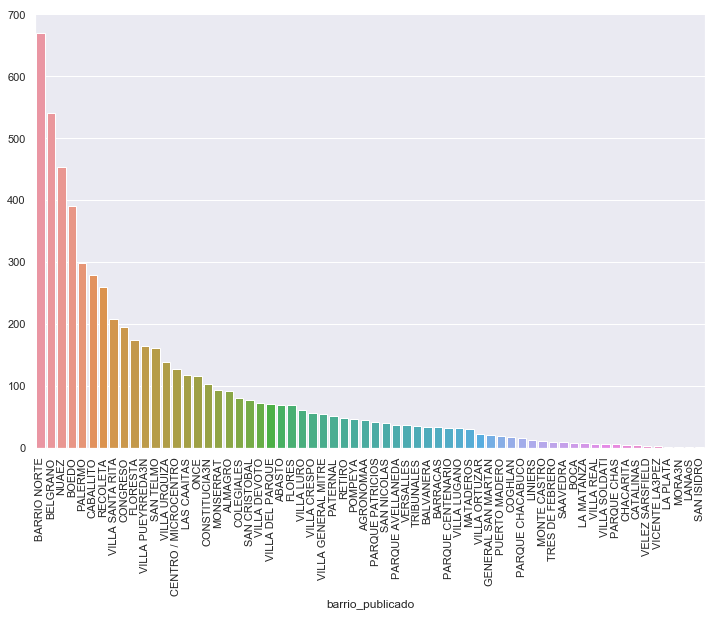
## Barrios con más edificios y barrios con más casas

El mentor planteó un desarrollo en el cual se realizó el análisis de los barrios con más edificios y los barrios con más casas. En este gráfico podemos ver como incluso dentro de los barrios con mayor porcentaje de casas sobre el resto de los barrios, el porcentaje de departamentos es mayor que el porcentaje de casas (Gráfico de la derecha).

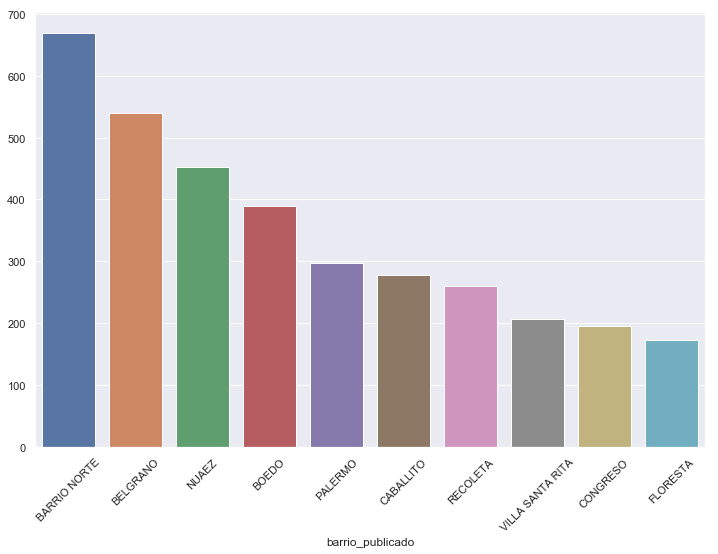


Barrios que quieren ser otros

Para comenzar, realizamos un análisis de todos los barrios que “quieren ser otros”. Nos referimos con esto a aquellas propiedades que pertenecen al barrio “X” pero que están publicados en un barrio “Y”. Se realizó un gráfico en donde se observa la cantidad de barrios publicados que más se repiten siendo que la propiedad es de otro barrio.

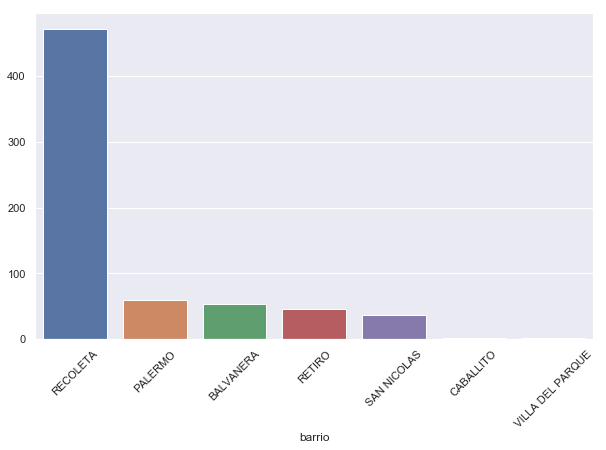


Del total del dataset, 5091 casos son identificados por lo antes descripto. Si tomamos una medida de los 10 barrios más publicados vemos el siguiente gráfico:



Por medio de este gráfico, podemos inferir que estos 10 barrios pueden ser los que más atraen a la clientela para la venta, lo que significa que a la hora de publicar una propiedad si el barrio real de la propiedad no es un barrio “que venda”, se lo categoriza para la publicación en un barrio que puede llamar más la atención a quien quiera comprar una propiedad.

Si elegimos de ejemplo el “BARRIO NORTE” para analizar los barrios que “se quieren parecer” a este observamos lo siguiente:

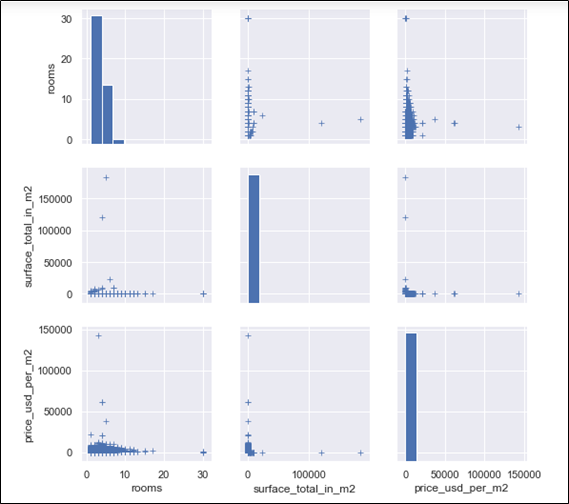


Se observa que en “RECOLETA”, 472 propiedades están publicadas en “BARRIO NORTE”, y siguen barrios como “PALERMO”, “BARVANERA”, “RETIRO” y “SAN NICOLAS” alrededor de 50 propiedades que se “parecen” a “BARRIO NORTE”.

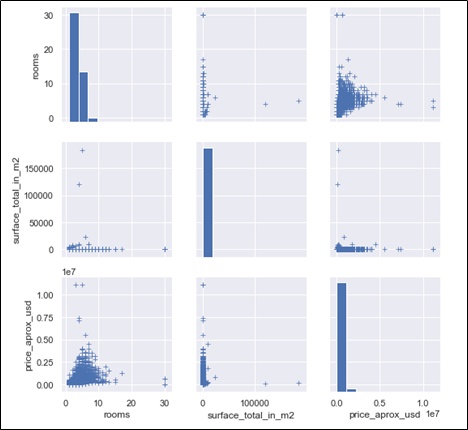
Analizando en el dataset la cantidad de propiedades cuyo barrio es “RECOLETA” es de 1561 y luego si tenemos en cuenta que 472 de esas propiedades están publicadas en “BARRIO NORTE”, más de un 30% de las mismas no quieren pertenecer a “RECOLETA” lo que podría indicar que este barrio no es considerado por la clientela para la contra de propiedades.

# Correlación de variables

Se planteó como desafío el análisis de correlación entre las variables surfacte\_total\_in\_m2, rooms y Price. Por lo que en primer lugar realizamos un pairplot de las tres variables globales, pero considerando el precio por m2:

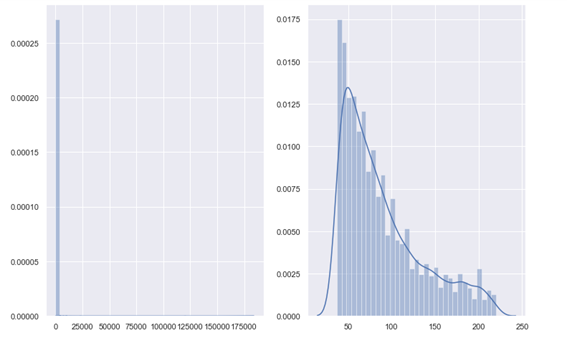


En este gráfico podemos ver que a simple vista no se identifica una correlación marcada entre las variables. Siendo este análisis poco revelador, continuamos el análisis tomando la comparación del precio total de cada propiedad en lugar del precio por m2.

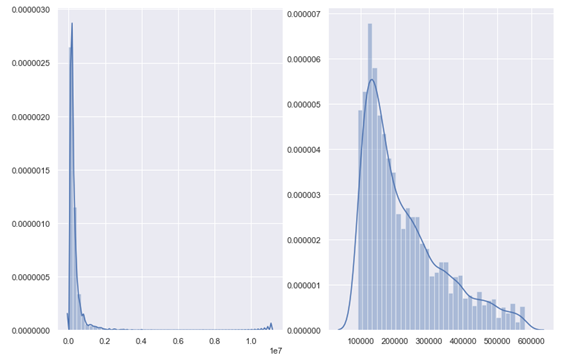


A diferencia del primer pairplot, podemos ver que los datos muestran menor concentración y hay indicios de un grado de correlación. Si bien podemos ver que los gráficos son más indicativos que el primer pairplot realizamos nuevamente el mismo análisis, pero sin considerar los valores extremos en las variables superficie total y precio total.

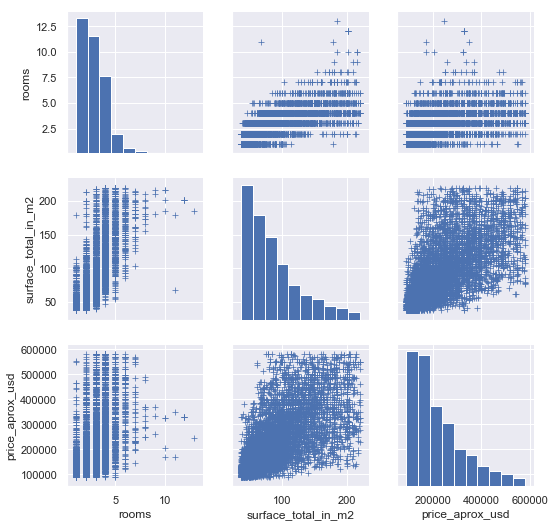
Para quitar los extremos aplicamos la misma metodología mencionada anteriormente donde eliminamos los valores de los 10 percentiles de ambos extremos de la variable. A continuación, podemos ver cómo cambian las distribuciones de cada variable con y sin extremos. (A izquierda el gráfico de distribución con extremos y a derecha sin extremos)



Para el caso de la variable precio total repetimos el proceso:

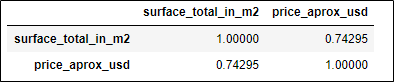


Observando esta mejoría en la distribución de los datos continuamos el análisis con las variables sin los datos extremos. El pairplot siguiente nos permite ver con mayor claridad la correlación entre las variables:



Aquí se puede apreciar que la variable precio total aumenta a medida que aumenta la superficie total y también que la variable precio total aumenta a medida que la propiedad tiene más cuartos. **Sin embargo, vemos en las gráficas que existen propiedades de baja superficie a precios similares de propiedades de gran superficie.**

Calculamos el coeficiente de correlación de spearman para el precio total y la superficie total:

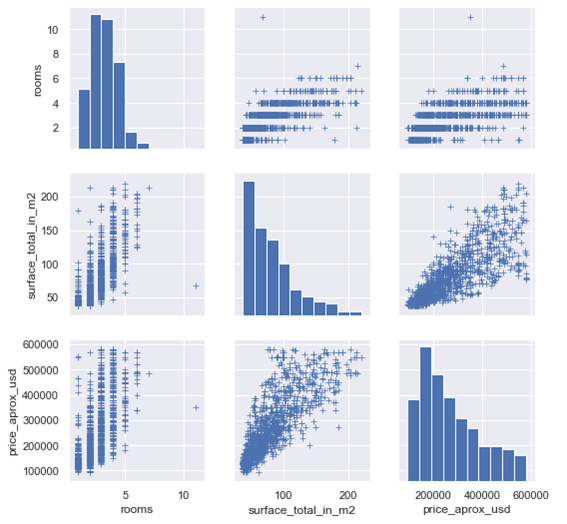


Podemos observar que nos da una correlación positiva y más cercana a 1, lo cual corresponde a lo observado en el gráfico.

Profundizamos aún más el análisis considerando que es importante estudiar la correlación de estas variables haciendo una separación por barrio. Ya que, la variable barrio tiene una gran influencia en el precio. En muchos casos, propiedades de grandes superficies en barrios baratos tienen menores precios que propiedades de superficie chica en barrios caros. Tomamos como ejemplo dos barrios puntuales: Palermo (dentro de los 5 más caros) y Villa Crespo (barrio intermedio). Para acompañar este análisis nos hacemos la siguiente pregunta:

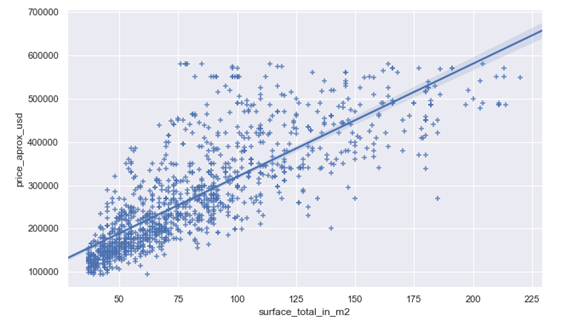
***¿Un departamento pequeño en Palermo es más caro que un gran departamento en Villa Crespo?***

Planteamos el pairplot para el barrio Palermo:

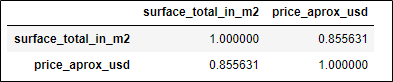


Aquí se puede observar una mayor correlación entre las variables específicas para el barrio Palermo.

Considerando la variable superficie total y precio total realizamos un análisis para ver la recta que nos pueda representar esta relación lineal:

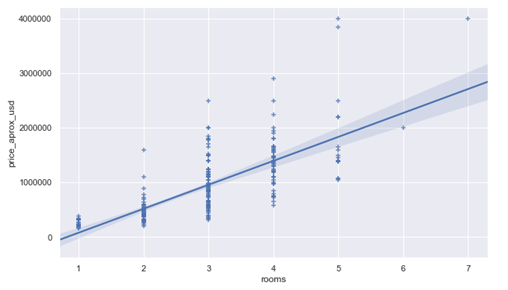


Podemos ver a simple vista que a medida aumenta la superficie aumenta el precio y los puntos se aproximan a una recta. El coeficiente de correlación de spearman obtenido fue el siguiente:

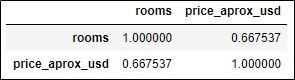


Siendo de 0.85 podemos ver que se confirma lo analizado a simple vista ya que tenemos un intenso grado de correlación positiva.

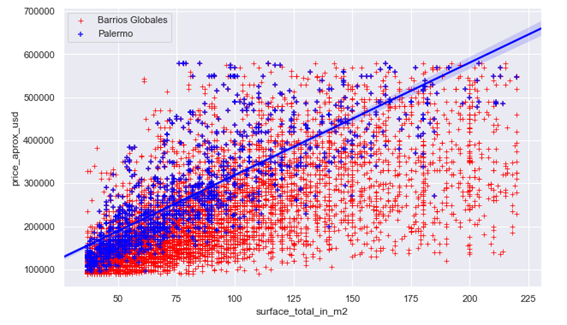
Repetimos el proceso para rooms y precio:



Aparece una correlación entre las variables, pero no tan grande como el caso anterior. Nuevamente aplicamos spearman y obtenemos un coeficiente positivo (aunque menor al anterior):

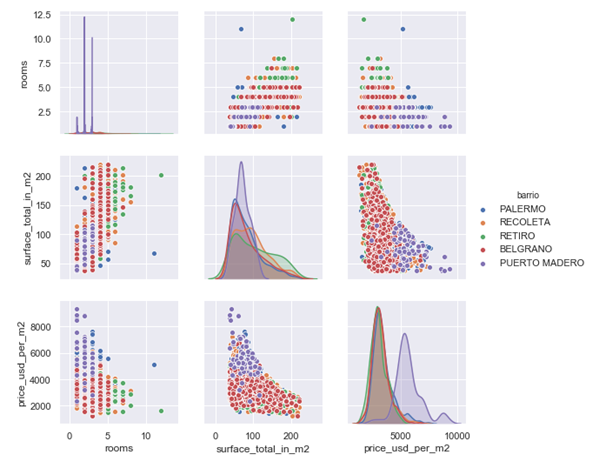


En última instancia, realizamos conjuntamente el scatterplot de las variables superficie total y precio total para todos los barrios conjuntos en comparación a considerar únicamente Palermo.



Se puede observar en un mismo gráfico lo que planteamos anteriormente respecto de la importancia de analizar por barrio, ya que Palermo (en azul) demuestra que sus datos están más cerca de una recta. No así los puntos en rojo que están más dispersos, representando los casos de todos los barrios mezclados.

A modo de cierre, representamos un pairplot diferenciado para los barrios caros:



En el mismo podemos ver como cada uno tiene una tendencia similar y como Puerto Madero (el más caro) se diferencia mucho del resto.

# Conclusiones

El análisis exploratorio realizado sobre este dataset nos sirvió como equipo a comprender un negocio y transformar una gran base de datos en información útil que podría servir para tomar decisiones comerciales o de nuestra vida cotidiana como elegir un buen barrio para vivir.

Aplicando los conocimientos de esta primera materia de la diplomatura pudimos analizar que en las publicaciones de venta o alquiler muchos de los barrios menos solicitados por los usuarios son publicados de manera “engañosa” con ubicaciones en los barrios más caros y lindos. También identificamos que la mayoría de las ofertas de propiedades en Buenos Aires son departamentos y que, a su vez, los barrios más lindos y caros tienen más del 90% de este mismo tipo de propiedad.

En cuanto a la correlación de las variables, pudimos determinar que el precio por m2 no varía mucho dentro de un mismo barrio según la superficie o la cantidad de cuartos de una propiedad. Entonces es un indicador de lo que sale una propiedad en un barrio. En cambio, el precio total de la propiedad tiene una alta correlación positiva con la superficie y la cantidad de cuartos de cada barrio. No así, si se analiza globalmente todos los barrios.

La importancia de separar por barrio a la hora de tomar decisiones con este set de datos es sumamente importante.