Organización de Datos

Segundo Cuatrimestre 2017

Trabajo Práctico 1

Integrante	Padrón	Correo electrónico
Rodrigo De Rosa	97799	rodrigoderosa@outlook.com
Marcos Schapira	97934	schapiramarcos@gmail.com
Facundo Guerrero	97981	facundoiguerrero@gmail.com

Índice

Parte I
Análisis del precio por m^2

1. Adaptación del DataFrame

Para el análisis particular de cada característica de la información que se posee, se adaptó el DataFrame original para poder analizar dicha información mas fácil y comodamente.

1.1. Filtrado de columnas

Para el análisis de esta cierta característica de las propiedades, consideramos *importantes* sólo a algunas celdas. Estas son:

- `place name` ← `location`
- `price aprox usd` ← `price`
- `surface total in m2` ← `totalSurface`
- `surface covered in m2` ← `coveredSurface`
- `price usd per m2` ← `pricem2`

1.2. Completando el DataFrame

Lo primero que se hizo para realizar este analisis fue completar las columnas faltantes de la mayor cantidad de entradas posibles. Esto es, `location, price, totalSurface, pricem2`. De esta forma, nos permitimos analizar una mayor cantidad de propiedades para realizar un analisis un poco mas correcto.

Para completar el campo de precio por m^2 se necesita que la entrada sobre la que se trabaja cumpla la siguiente condición lógica: $pricem^2 \vee (price \wedge surface)$. Es decir, necesita tener o el precio por metro cuadrado o tanto el precio total como la superficie total.

Si el campo `pricem2` tiene valor, entonces ese será el utilizado. En caso contrario, si tanto el campo `price` como el campo `totalSurface` tienen valor, definimos como nuestro nuevo `pricem2` a la división $\frac{price}{surface}$.

Para esto, necesitamos unificar `coveredSurface` y `totalSurface`, para maximizar nuevamente la cantidad de entradas disponibles. Esto se hace, simplemente, poniendo como `totalSurface` el valor de `coveredSurface` en aquellas entradas donde la primera no tenga valor (consideramos que $total - covered = uncovered$).

Una vez completados todos los `pricem2` posibles, eliminamos todas aquellas entradas que tengan `NaN` como valor (en cualquiera de las celdas que definimos como *importantes*), pues ya no podemos obtener el valor de esa celda de ningun otro lugar.

2. Estudio estadístico de los datos

2.1. Analisis de la distribución de precios

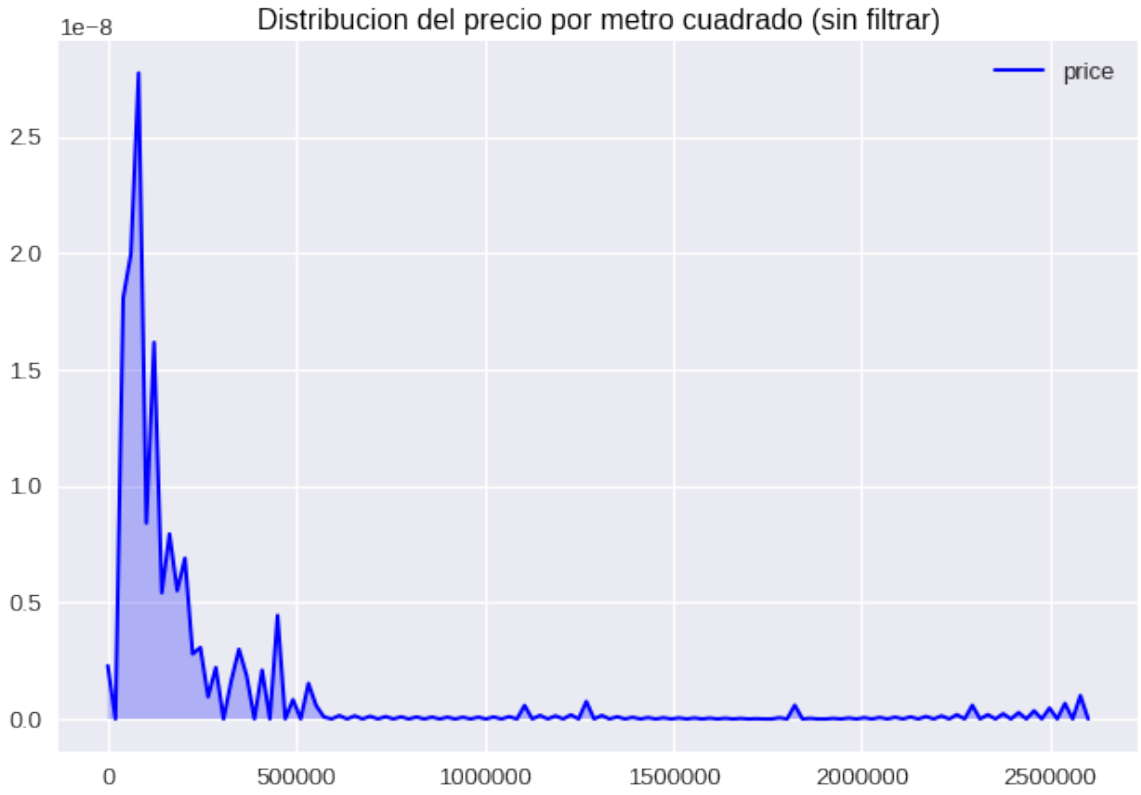
Una vez completado el DataFrame lo mas posible, se realizó un análisis de la distribución de precios. Con esto nos referimos a analizar la variación del precio por metro cuadrado entre todas las propiedades. Es decir, *limpiar los datos que no tienen sentido.*

Para esto le pedimos el `.describe()` a nuesto DataFrame con los percentiles 0,01 y 0,99. Esto nos permite analizar que tan desviados estan los valores máximos y mínimos.

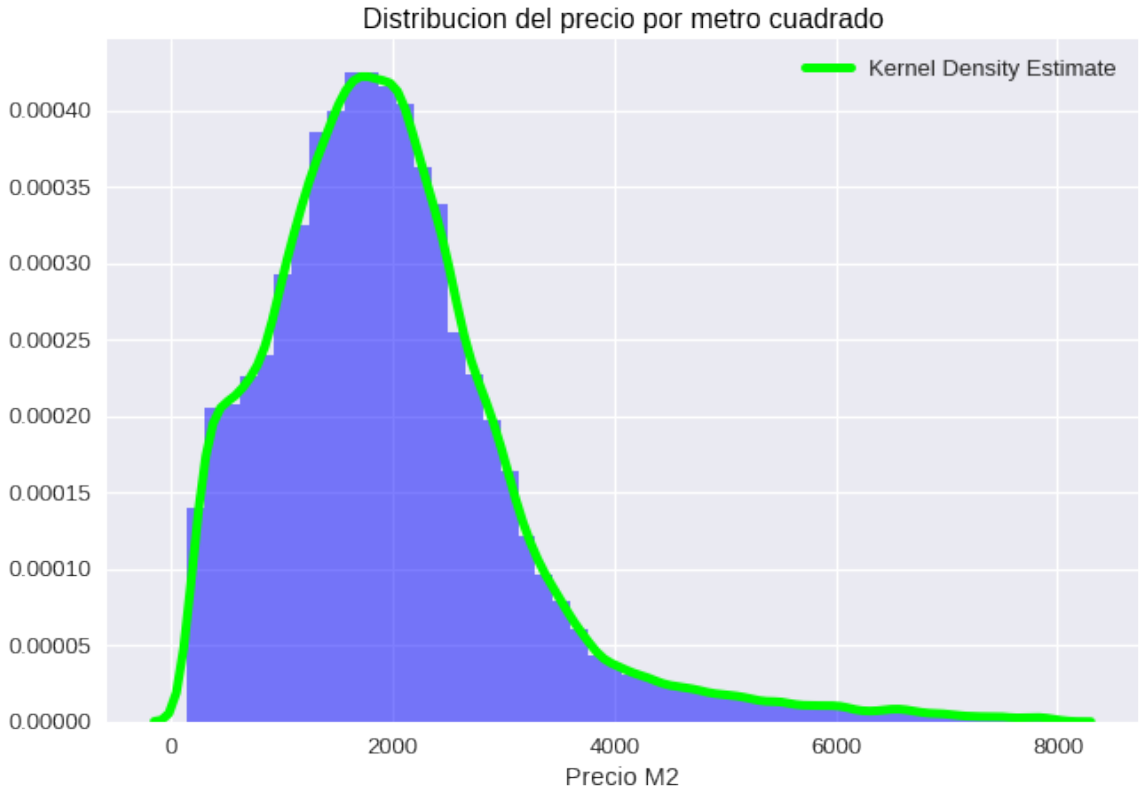
Con los percentiles recién mencionados hacemos un recorte de los datos para lograr una distribución que se asemeje a una Normal lo mas posible. El primer recorte es tanto inferior (> 150USD) como superior (< 18000USD). Como en este nuevo DataFrame la diferencia entre el percentil 0,99 y el máximo es de más del doble, se vuelve a recortar superiormente (< 8000USD).

Luego de esto, la distribución de precios es un poco mejor que antes (las diferencias entre los percentiles $0{,}25, 0{,}5, 0{,}75$ son similares).

A continuación se muestra un gráfico de distribución de precios por metro cuadrado que se obtiene del DataFrame original sin realizar el filtrado recién mencionado. Nos hubiese gustado poder mostrar tanto el KDE como el histograma pero al haber tanta diferencia entre el maximo y los valores principales de la distribución, el histograma era solo una linea. El objetivo de este gráfico es hacer incapié en lo mencionado en el previo párrafo: es necesario filtrar los datos para tener un conjunto de datos con sentido.



En el siguiente gráfico de distribución de precios por metro cuadrado se puede ver que la mayor parte de las propiedades están concentradas en el rango de precios $[150; 4000]$USD y luego hay un drástico decaimiento de cantidad de propiedades para el resto de los precios. Si bien se podría considerar que un recorte sería correcto, a partir de fuentes externas se sabe que ciertos barrios (*i.e.* Puerto Madero) tienen, aproximadamente, un valor medio de 6000USD por metro cuadrado.



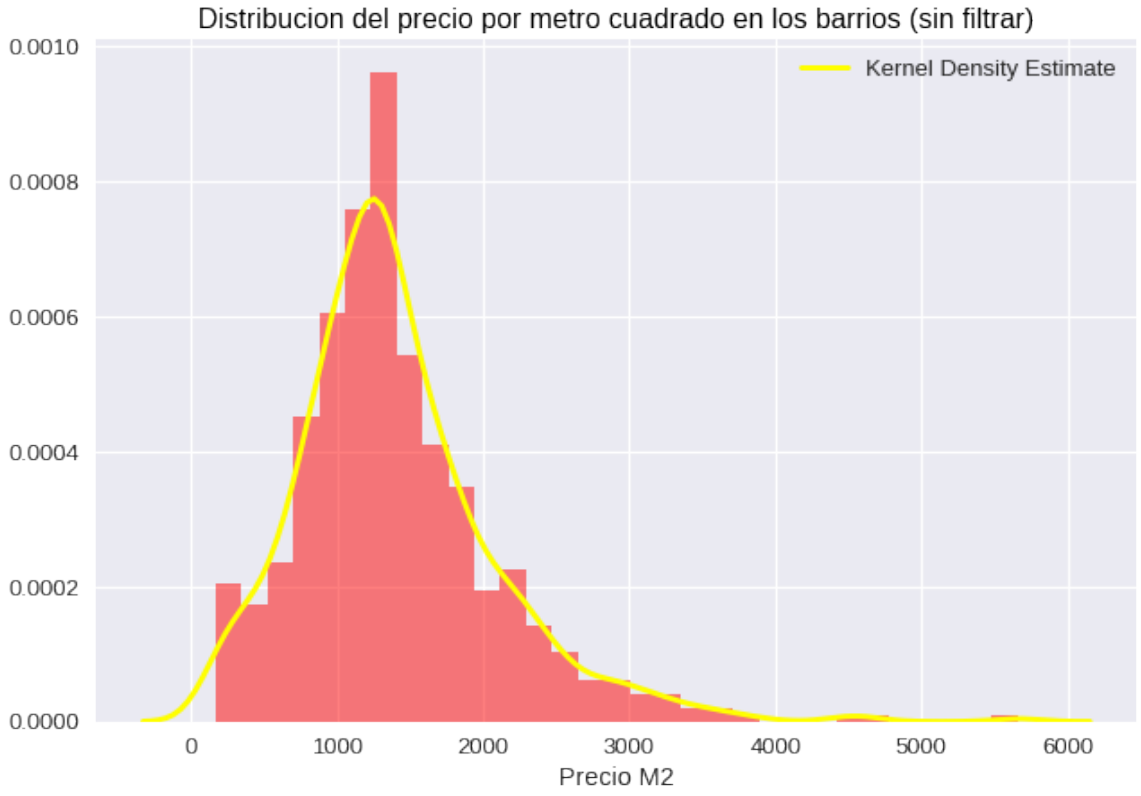
2.2. Agrupando por barrios

Ahora que nuestros datos están tan completos y retocados como querríamos, procedemos a agrupar todas las propiedades de acuerdo al barrio al que pertencen. Una vez que los tenemos agrupados, debemos establecer un

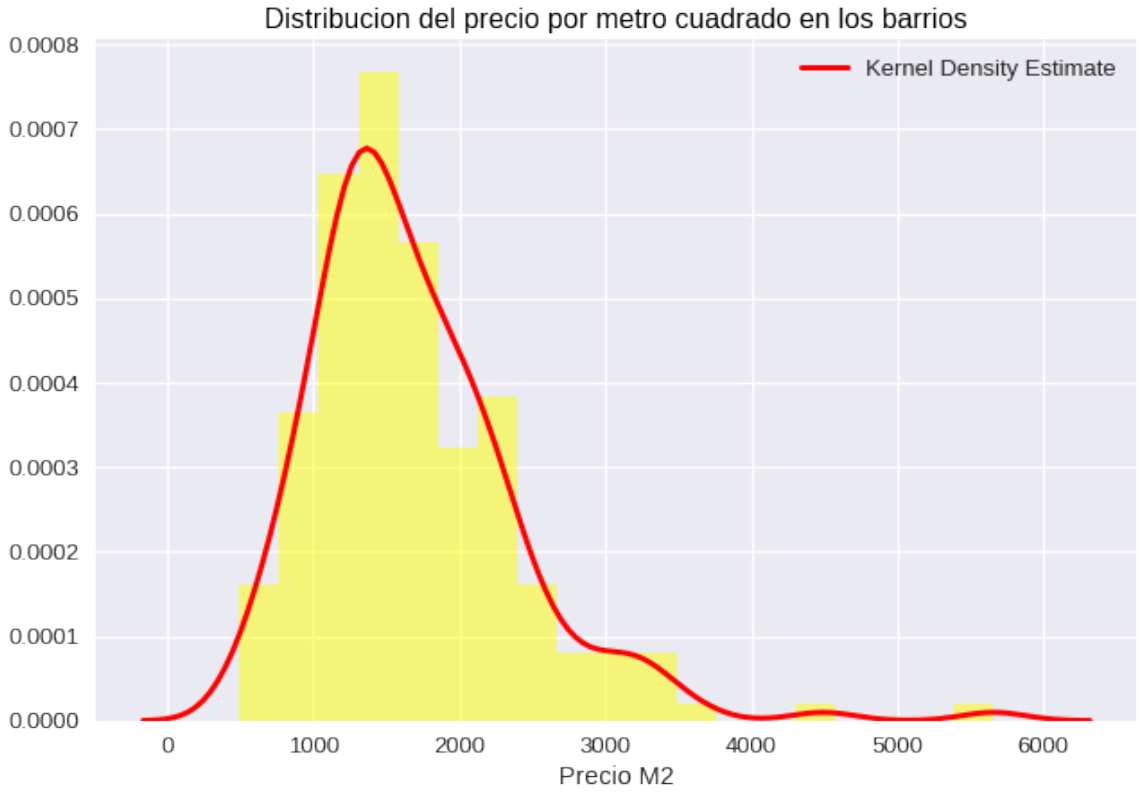
minimo de propiedades por barrio. Pues un barrio que tiene una o dos propiedades podría alterar el estudio de la informacion.

Nuevamente, para esta tarea utilizizamos `.describe()` y resolvemos que utilizaremos como cota inferior 50 propiedades (dos mas que el equivalente a una publicación por mes en los últimos cuatro años).

Aquí, al igual que hicimos antes, mostraremos la distribución antes y después del filtro aplicado. Si bien en escencia no son tan diferentes, podemos observar que desaparecen algunos barrios de la zona de precios altos.



Una vez que eliminamos los barrios problemáticos, si analizamos la distribución de precio por barrio podemos ver que la mayor parte está concentrada en el intervalo $[500; 3500]$USD, mientras que muy pocos (solo tres) superan ese valor.



Podemos ver, además, que la distribución es bastante similar a la anterior (sin agrupar por barrios) aunque, obviamente, con valores menores (pues son promedios).

3. Analizando grupos característicos

En esta sección analizaremos ciertos grupos característicos a partir de la información con la que estamos trabajando.

3.1. Los diez barrios con mayor precio por m^2

Dado que ya estamos felices con la forma en que tenemos dispuestos los datos, comenzaremos por hacer un *Top 10* de los barrios más caros de CABA y GBA.

Para esto, como ya tenemos los datos agrupados, simplemente ordenamos el DataFrame y nos quedamos con los primeros diez. Durante el análisis de esta información, notamos que varios de los barrios que aparecían en este *Top 10* eran subdivisiones del barrio de Palermo. Por esta razón, decidimos incluir dos casos: uno en que consideramos que todos los 'Palermos' son uno solo, y otro en que cada uno es considerado un barrio diferente.

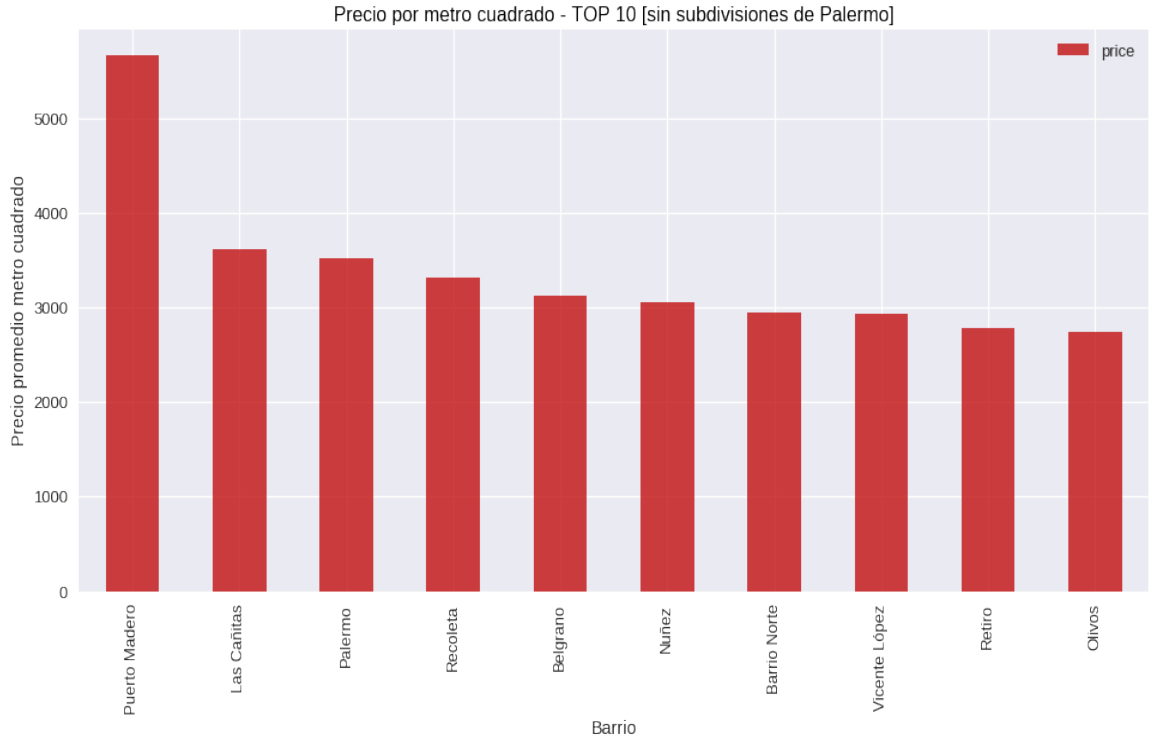
3.1.1. Unificación de Palermo

En este caso, consideramos que todas las subdivisiones de Palermo pertenecen a un sólo barrio.

El resultado obtenido es el siguiente:

Top 10 [Palermo unificado]		
Puesto	Barrio	Precio m^2 [U$D]
1	Puerto Madero	5657
2	Las Cañitas	3612
3	Palermo	3518
4	Recoleta	3316
5	Belgrano	3124
6	Nuñez	3056
7	Barrio Norte	2949
8	Vicente López	2925
9	Retiro	2783
10	Olivos	2737

En la tabla se observa que Puerto Madero tiene un valor mucho mas alto que el resto, de hecho, es mayor al doble del precio del décimo. De todos modos, entre el segundo y el último la variación es más suave. Para aportar a este análisis, se realiza un gráfico de barras:



3.1.2. División de Palermo

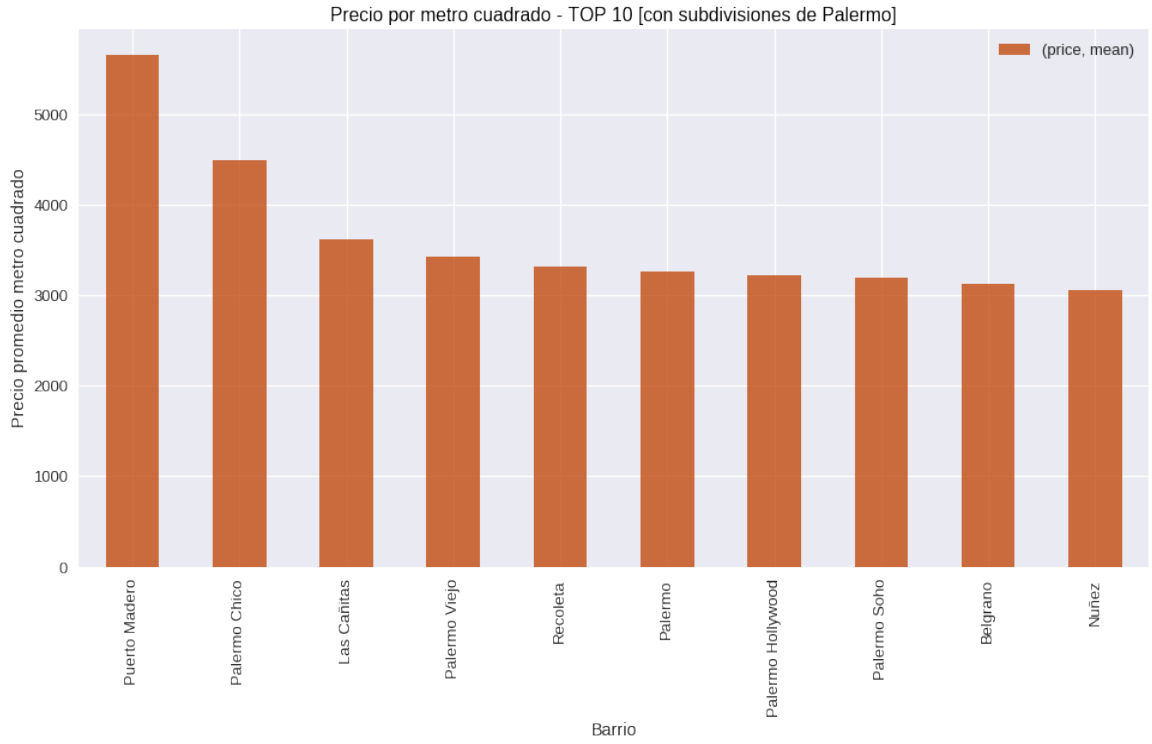
Aquí consideraremos que el barrio al que se nombra Palermo corresponde a todas las secciones de dicho barrio que no son las que ya aparecen en otro grupo.

En este caso, el resultado obtenido es:

Top 10 [Palermo dividido]		
Puesto	Barrio	Precio m^2 [U$D]
1	Puerto Madero	5657
2	Palermo Chico	4489
3	Las Cañitas	3612
4	Palermo Viejo	3419
5	Recoleta	3316
6	Palermo	3260
7	Palermo Hollywood	3224
8	Palermo Soho	3198
9	Belgrano	3124
10	Nuñez	3056

En la tabla podemos ver que, si bien es correcto y es un *Top 10*, esta plagado de subdivisiones de Palermo y no nos permite tener un plano más general.

Aquí el gráfico de barras es muy similar aunque aparece Palermo Chico, que se acerca un poco mas al valor de Puerto Madero. De todos modos, la diferencia entre el primero y el segundo es muy grande como también lo es entre el segundo y el tercero, dejando la relación entre los valores igual de 'no suave'.



De aquí en más, utilizaremos a Palermo como un barrio unificado.

3.1.3. Comentario sobre el Top 10

Este *Top 10* arroja los resultados que se hubieran esperado, pues los únicos dos valores que no pertenecen a CABA corresponden a los primeros dos barrios de GBA en los que se piensa al pensar en los barrios mas caros de Buenos Aires.

Por otro lado, si nos sorprende el hecho de que el m^2 en Barrio Norte sea más barato que Núñez o en Belgrano.

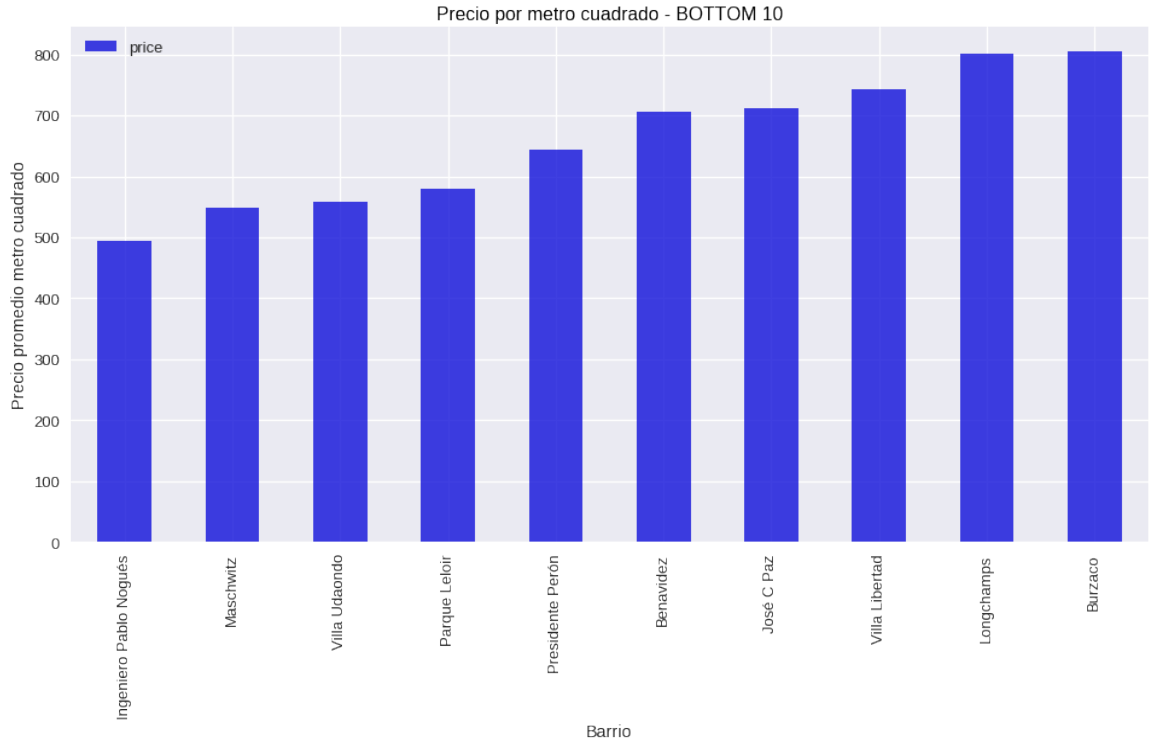
3.2. Los diez barrios con menor precio por m^2

Para esta parte, al igual que antes, ordenamos los datos para analizar cuales son los diez barrios que se encuentran en el *Bottom 10*.

Al realizar este análisis, lo obtenido es:

Bottom 10		
Puesto	Barrio	Precio m^2 [U$D]
1	Ingeniero Pablo Nogués	494
2	Maschwitz	548
3	Villa Udaondo	558
4	Parque Leloir	579
5	Presidente Perón	643
6	Benavidez	705
7	José C Paz	711
8	Villa Libertad	743
9	Longchamps	800
10	Burzaco	804

Si graficamos estos valores al igual que antes podremos ver un ascenso (o descenso) más suave que el del *Top 10*. Si bien el primero es casi la mitad de el último, la variación entre puestos es menor.



Aquí, remitiéndonos a la sección 3.1.3, vemos que los barrios del *Bottom 10* son todos barrios alejados de la ciudad, de los cuales es esperable un bajo valor del m^2.

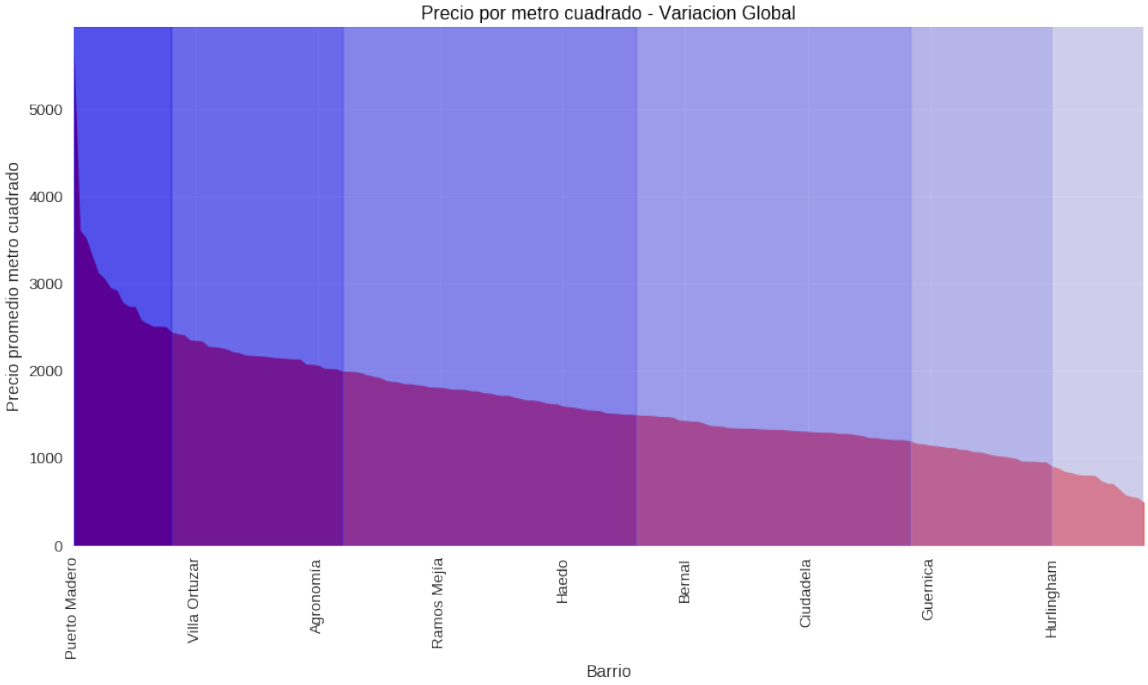
3.3. Dividiendo en secciones

El objetivo de esta parte es determinar diferentes grupos de barrios basados en el precio promedio del m^2 para, de esta manera, analizar que tan suave (o no) es el decrecimiento del valor del m^2 en cada uno de estos grupos.

Los barrios serán divididos en grupos a partir de valores arbitrarios de precio por m^2, que surgen de un análisis de los datos. Los grupos serán:

Divisones		
Numero	min(pricem^2) [USD]	max(pricem^2) [USD]
1	2500	∞
2	2000	2499
3	1500	1999
4	1200	1499
5	950	1199
6	450	949

El siguiente gráfico de área muestra el precio del m^2 y las divisiones (diferentes intensidades de azul) indican los diferentes grupos.



Nos interesa mostrar también que porcentaje de los barrios está incluído en cada uno de estos grupos para saber en cuál de ellos se concentra la mayor parte.

Distribución en grupos		
Grupo	Cantidad de barrios	Porcentaje
1	16	9 %
2	28	16 %
3	48	27 %
4	45	26 %
5	23	13 %
6	16	9 %

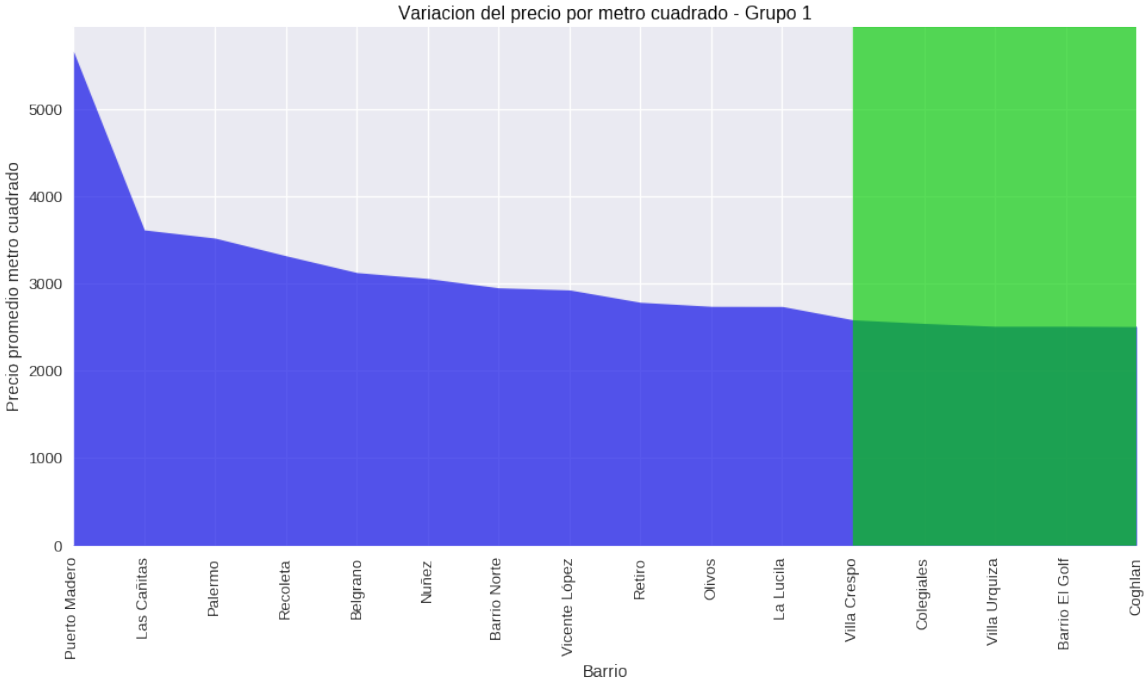
Vemos que, como se puede apreciar en el gráfico, los grupos 3 y 4 son los que concentran a la mayor cantidad de barrios. De hecho, entre ellos dos tienen a más del 50 %.

Ahora procedermos a analizar grupo por grupo.

3.3.1. Grupo 1 - $[2500 : \infty)$U$D

En este grupo se encuentra a los dieciséis barrios más caros de CABA y GBA. Aunque a los primeros diez ya los conocemos de la sección 3.1, lo que nos interesa en esta parte es analizar y comparar cómo varía el precio del metro cuadrado grupo a grupo mas que los nombres propios de cada integrante de cada grupo. En cada uno de los grupos se mostrará un *zoom* del gráfico de área previo, para agregar una visualización al análisis de la variación en cada caso.

En este caso, lo que esperamos encontrar es un inicio con una pendiente muy inclinada y, a medida que nos alejamos del primer barrio (que ya sabemos que es Puerto Madero), un decrecimiento de dicha pendiente, pues la diferencia entre barrio y barrio será mucho menor, aunque seguirá siendo el grupo con mayor variación entre barrios de todos.

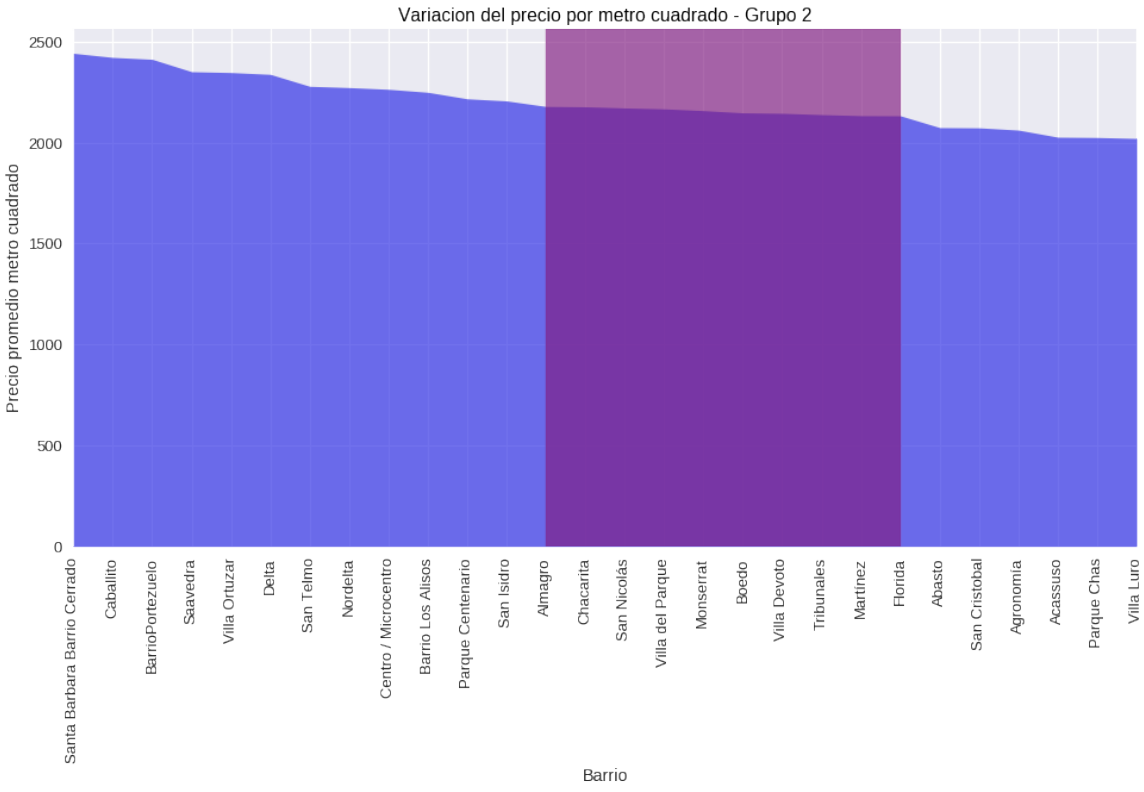


Si bien el gráfico muestra lo que se esperaba, también se ve que a partir del doceavo barrio la curva se vuelve casi constante (en verde), mostrando que hay un cierto subgrupo de barrios con precios muy similares. Esto se puede ver en el gráfico de la sección anterior si se observa el pequeño valle que se forma justo antes de llegar a la división entre el grupo 1 y el 2.

Cabe destacar, además, que si no fuera por el precio extraordinario de Puerto Madero, el grupo tendría una variación mucho menor, pues iría entre 3600USD y 2500USD lo que representaría una variación de 1100USD (30 %) entre el máximo y el mínimo mientras que, actualmente, la variación es de 3100USD (55 %).

3.3.2. Grupo 2 - $[2000 : 2500)$U$D

Observando el gráfico en el que se realizaron las divisiones por grupo esperamos que este segundo grupo tenga un sector medio con pendiente casi nula, donde la diferencia entre el precio en un barrio y el siguiente sea casi nula.



El sector indicado el violeta es aquel que mencionamos previamente con pendiente casi nula. En este subgrupo se encuentran diez barrios y la diferencia entre el primero y el último es de solamente 46USD. Dado los numeros que se manejan, esa diferencia es casi despreciable, teniendo así un subgrupo de diez barrios con el mismo valor para el m^2.

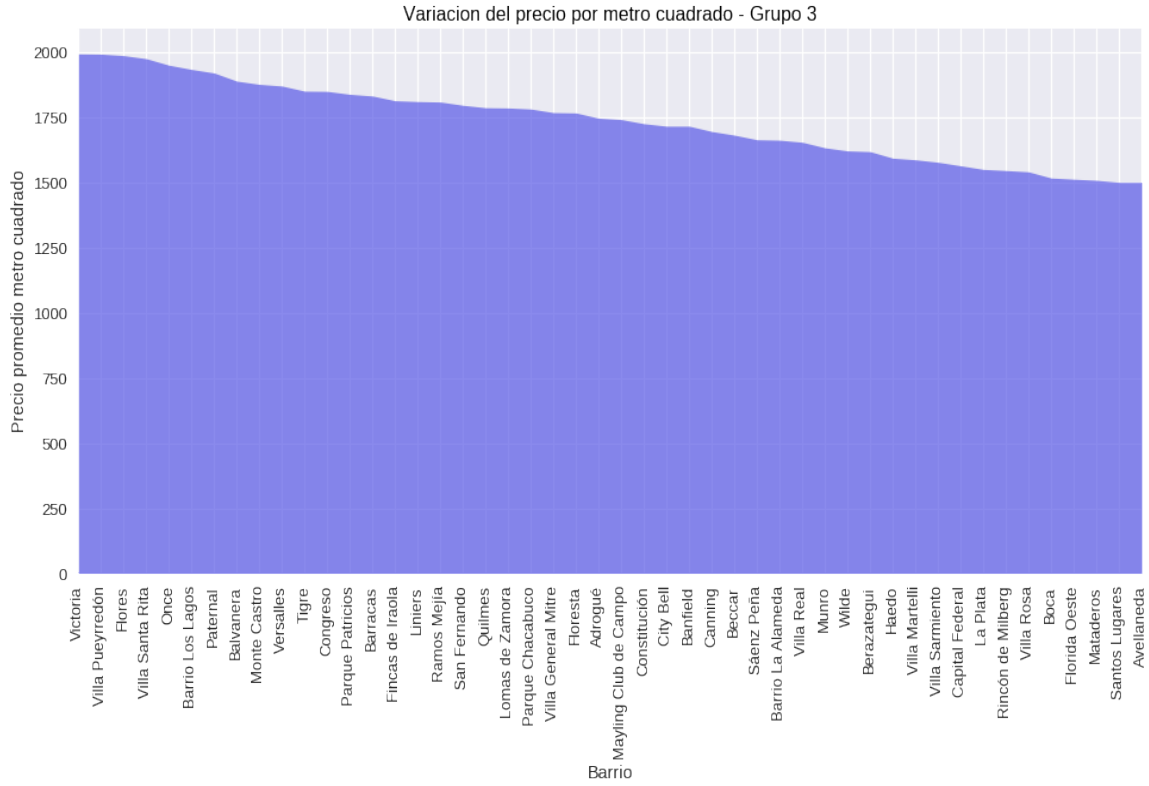
En cuanto al porcentaje de variación, a partir de este segundo grupo lo podemos dividir en dos; uno particular del grupo, en donde el porcentaje se calcula a partir del máximo local, y otro general, en donde el porcentaje se calcula a partir del máximo global (*i.e.* Puerto Madero). La diferencia entre el máximo local y el mínimo es de 422USD, siendo, de esta manera, la variación local del 17 % y la general del 7 %.

Vemos entonces que la diferencia de variación ya decreció en gran medida pasando del grupo 1 al grupo 2.

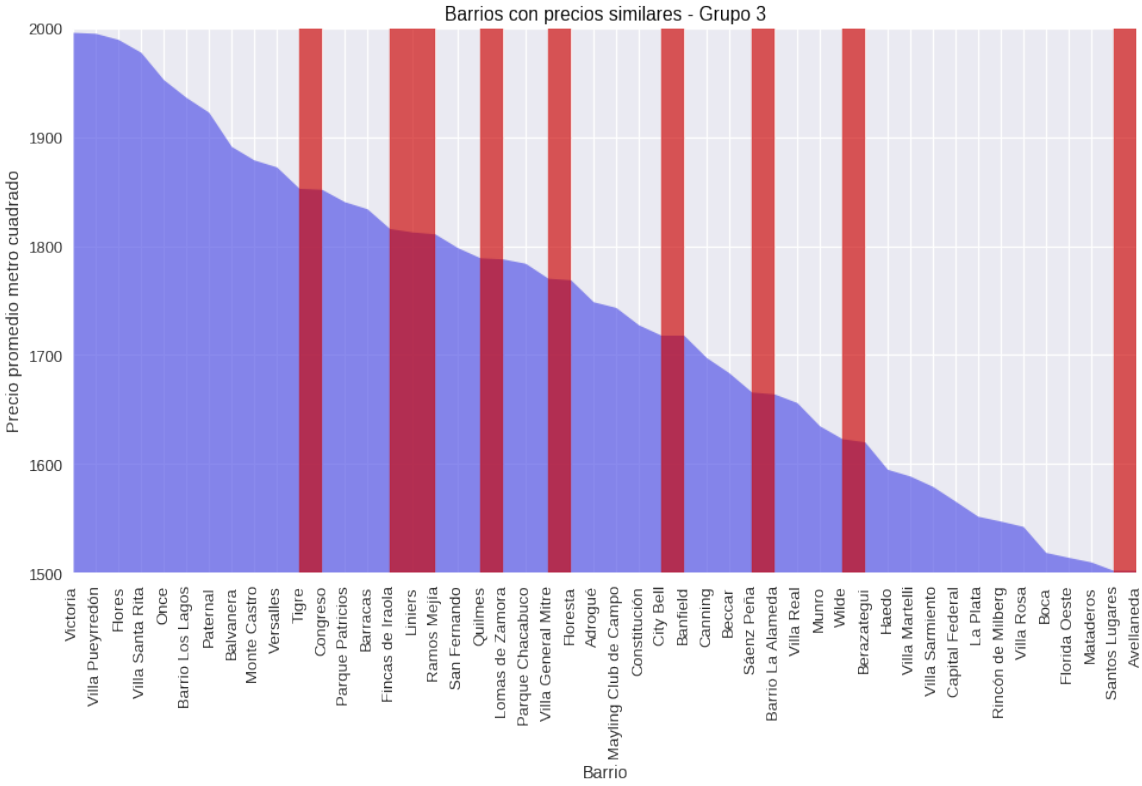
3.3.3. Grupo 3 - [1500 : 2000)U$D

Ahora llegamos al primero de los dos grupos principales. Como vimos antes, este grupo concentra al 27 % de los barrios de CABA y GBA pero (según nos muestra a grandes rasgos el gráfico general) con mayor variación que el próximo grupo, que también concentra a una gran parte de los barrios.

En este caso, esperamos ver una pendiente casi constante en todo el grupo sin regiones con pendiente casi nula pero con una variación pequeña entre un barrio y el siguiente.



En este gráfico se observa lo previamente indicado, el decrecimiento es bastante lineal y no se detectan zonas con precios constantes. De todas formas, si se amplía un poco, se pueden encontrar pares o tríos de grupos que tienen precios similares, como se ve a continuación.



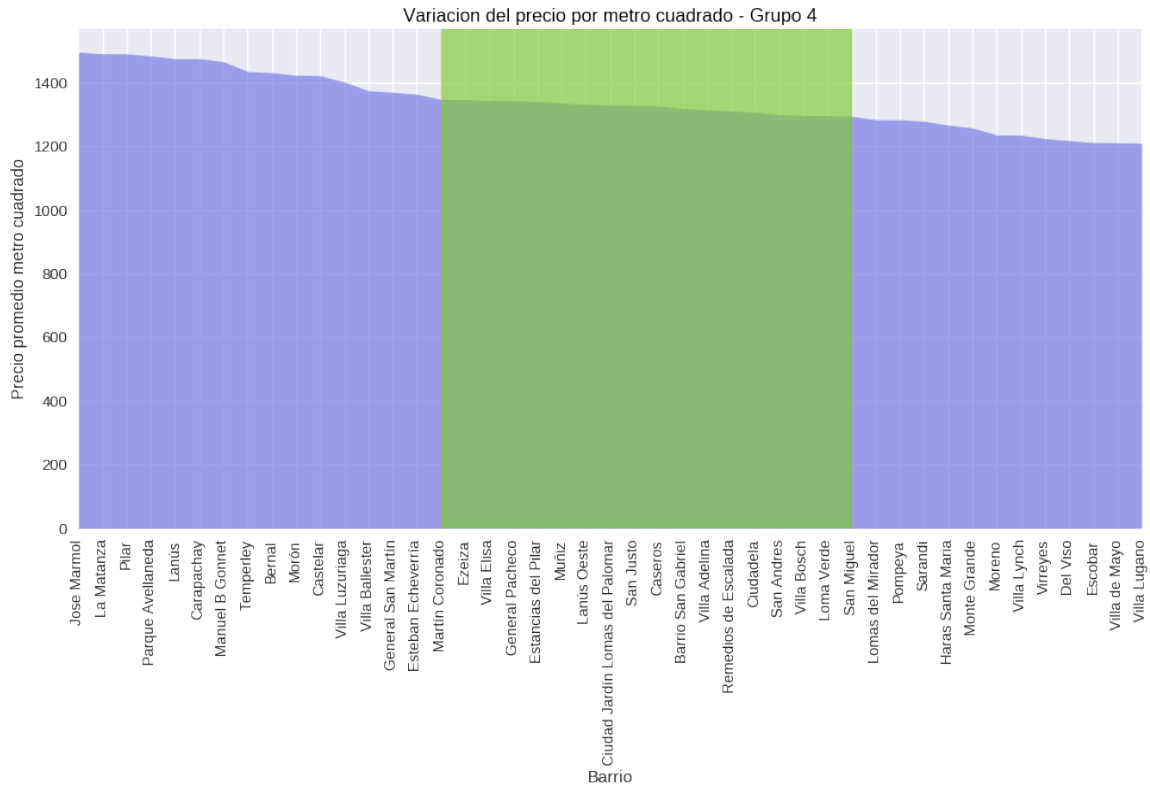
Aquí podemos ver que si bien hay diecisiete barrios que comparten precio con algún otro, no son contiguos como en el grupo 2. De todas maneras, como en este grupo la variación general de precio es menor y la cantidad de barrios es mayor, podríamos encontrar también un grupo de diez barrios con variación de aproximadamente 50USD. Por ejemplo, entre Monte Castro y Ramos Mejía hay una diferencia de 60USD y se encierra un subgrupo de nueve barrios.

Analizando el porcentaje de variación al igual que en el Grupo 2, sabiendo que la diferencia entre el máximo local y el mínimo es de 493USD, el porcentaje de variación local es de un 25 % y el porcentaje de variación global es del 9 %.

Se observa entonces que, lógicamente, ahora que estamos en una zona de precios menores, una diferencia entre mínimo y máximo similar a la del grupo anterior ahora representa una variación mayor. Por otro lado, la variación global es apenas mayor a la del grupo anterior. De todas formas, estos porcentajes están sujetos a la división arbitraria de los grupos hecha previamente (pues el máximo delta es 500USD).

3.3.4. Grupo 4 - $[1200 : 1500)$U$D

Este cuarto grupo es el segundo grupo principal, concentra el 26 % de todos los barrios y es el de menor rango de valores, con 300USD, salvo por el grupo 5, con 250USD. Esto nos dice que la pendiente en este caso será la menor de todas, como se puede ver en el gráfico general. Además, esperamos ver una zona central con pendiente casi nula, en donde se encontrará un grupo muy grande de barrios con una variación en el precio muy pequeña, a diferencia del grupo 3.



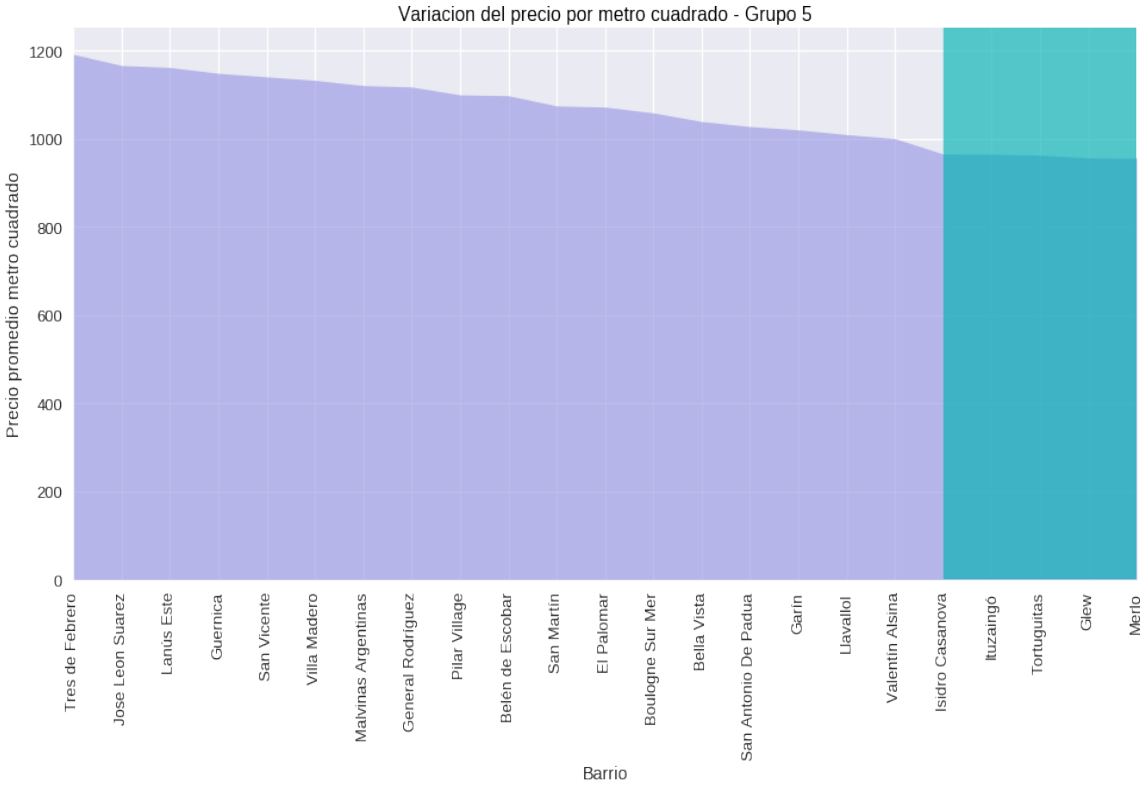
El subgrupo indicado con verde consiste de dieciocho barrios que se diferencian solamente, entre el primero y el último, por 53USD. Como supusimos, en este grupo se encuentra el conjunto de barrios con variación casi nula más grande de todos. Adicionalmente, si se observa el gráfico con atención, en la mayor parte se encuentran subgrupos pequeños con valores casi constantes hasta que hay un pequeño descenso de precio, nuevamente un valor casi constante, etc. dividiendo el grupo en cinco o seis diferentes subgrupos pequeños (o no tanto, como el verde) por lo que no tiene sentido destacarlos como en el grupo 3, pues solo quedarían aquellos pequeños sectores donde la pendiente si toma valor considerable.

En cuanto al porcentaje de variación, sabiendo que la diferencia entre el máximo local y el mínimo es de 286USD, tenemos un porcentaje de variación local del 19 % y uno global del 5 %. Por lo tanto, y como era de esperarse, este grupo tiene los porcentajes mínimos en ambos casos; en el local porque tiene la mayor cantidad de segmentos casi constantes y en el global porque, además de la razón recién mencionada, por tener valores cada vez más bajos. Cabe destacar, además, que el 19 % local está concentrado en ciertos 'saltos' de un barrio a otro; pues, como mencionamos antes, la mayor parte del grupo esta compuesta por subgrupos de precios similares.

3.3.5. Grupo 5 - $[950 : 1200)$U$D

Este quinto y penúltimo grupo comienza una pendiente que incrementa progresivamente hacia el último y sexto grupo. Aquí se concentra el 13 % de los barrios, la mitad que en el grupo anterior, y los precios de los barrios ya se acercan a los minimos que se pueden encontrar.

A partir de la información que brinda el gráfico general, esperamos encontrar una pequeña zona de pendiente casi nula hacia el final del grupo y una variación local relativamente chica.

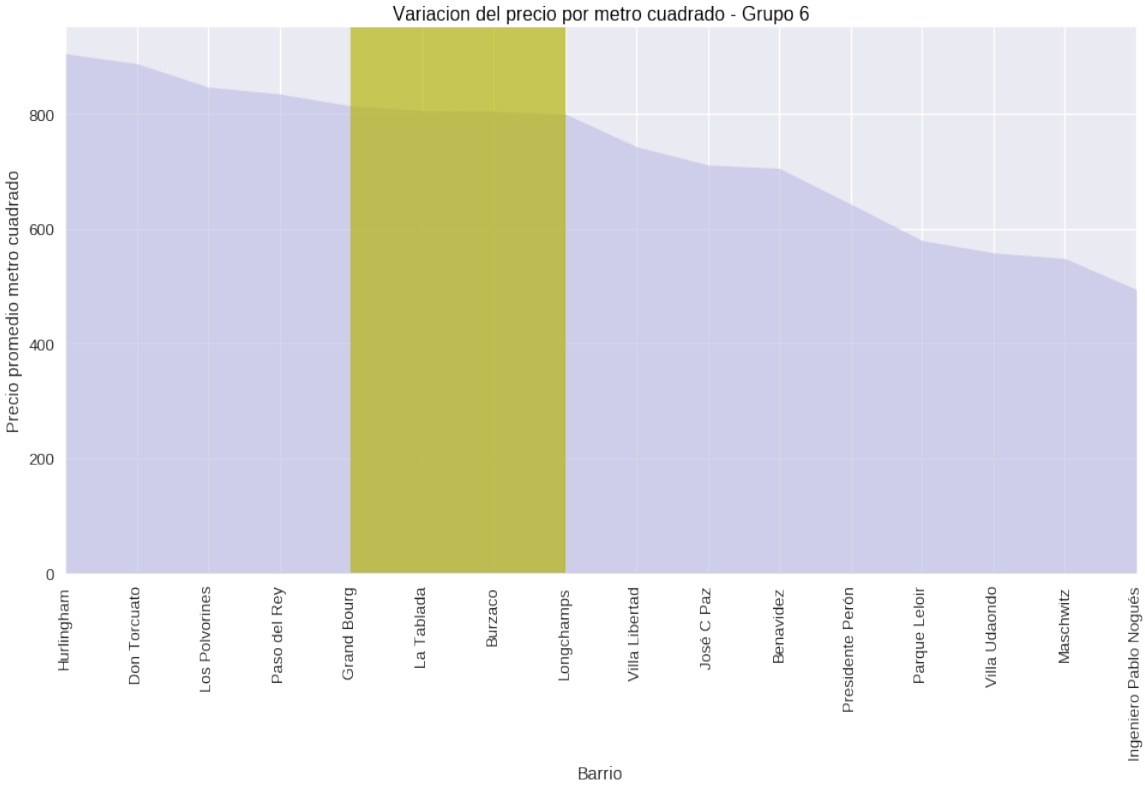


Indicado con color Cyan en el gráfico podemos ver ese pequeño grupo del que hablábamos donde la pendiente es casi nula, mientras que en el resto del grupo hay una pendiente casi constante. El tamaño de este subgrupo es de cinco barrios y la diferencia de precio entre el primero y el último es de 11USD.

Si calculamos el porcentaje de variación, conociendo que la diferencia entre el máximo local y el mínimo es de 235USD, obtenemos un porcentaje de variación local del 20 % y uno global del 4 %. En este caso, el porcentaje local esta distribuido a lo largo de todo el grupo, salvo por el subgrupo final que tiene un valor casi constante.

3.3.6. Grupo 6 - $[450 : 950)$U$D

En este último grupo, que contiene sólo al 9 % de los barrios (al igual que el grupo 1), estaremos analizando a aquellos barrios con menor precio por m^2. A partir del gráfico general sabemos que la pendiente será muy marcada y que la diferencia entre un barrio y el siguiente será importante, salvo por una pequeña parte donde los precios se mantendrán casi constantes.



En amarillo se puede observar la única sección casi constante del grupo, que consta sólo de cuatro barrios y que contienen un rango cuya diferencia entre el mínimo y el máximo es de 14USD. Mientras tanto, el resto del grupo tiene una pendiente considerablemente grande y la diferencia entre el valor máximo y el mínimo es casi del doble.

Para remarcar esto último, calcularemos el porcentaje de variación sabiendo que dicha diferencia es de 410USD. El porcentaje de variación local es del 45 % y el porcentaje de variación global es del 7 %. Aquí la variación local está repartida en todo el grupo pero con mayor participación de la segunda mitad, donde la pendiente es mayor, y con menor participación del sector casi constante.

4. Distribución geográfica

En esta sección se mostrará, con la ayuda de *HeatMaps* cómo están distribuídos los precios por m^2 en CABA y GBA.

4.1. Grupos característicos y su ubicación

Ahora, lo que haremos será repetir la dinámica pero dividiendo en los mismo grupos característicos de antes, para ver que se puede saber de cada grupo según su ubicación.

4.1.1. Grupo 1

4.1.2. Grupo 2

4.1.3. Grupo 3

4.1.4. Grupo 4

4.1.5. Grupo 5

4.1.6. Grupo 6