Universidad de Buenos Aires

Facultad de Ingeniería

Organización de Datos

Análisis exploratorio - Properati
Repositorio: https://github.com/freedocx/7506-2C-2017
2° Cuatrimestre 2017

Integrantes

Julián Matías Garate	93043
Gastón Montes	89397

Introducción

La empresa PROPERATI brinda un servicio web donde los vendedores pueden publicar propiedades y acercar a los usuarios y/o compradores, las ofertas del mercado inmobiliario. También tiene un blog de noticias donde se realizan, en base a análisis de la base de datos que dispone, información útil del mercado.
En este sentido el presente trabajo se basa en realizar un análisis exploratorio de la base de datos histórica de la empresa para poder encontrar datos que permitan posteriormente realizar un modelo predictivo de precios para las propiedades.
Link: http://www.properati.com.ar/data/

Lenguaje de programación

El lenguaje de programación elegido para realizar este trabajo es R y se utilizaron las siguiente librerías:

Ggplot2: librería de visualizaciones
Ggmap: librería para visualizaciones de mapas
Lubridate: librería para trabajar con dates, para darles formato y poder buscar coincidencias para diferentes unidades de tiempo
Dplyr: librería para trabajar con data frames, hacer joins, merge y agrupaciones.
Matrix: librería para trabajar con formatos matriciales y entre ellos matrices dispersas
Wordcloud: visualización de nubes de palabras en base a la frecuencia de ocurrencia en un texto
FeatureHashing: el feature hashing permite aplicar funciones de hash a un texto y transformar una palabra singular en una columna de una matriz, donde el identificador de la columna es el número resultante de aplicar la función, y el contenido de cada celda en 0 para los textos que no contienen la palabra y 1 para los textos que sí la contienen. El resultado es una matriz dispersa con todas las ocurrencias de las palabras en cada publicación.

Para hacer un análisis rápido de la estructura de datos se utilizó el lenguaje Python Pandas.

Estructura de los datos

Tomando el archivo properati-AR-2017-08-01-properties-sell.csv del set de datos provisto se hace un breve informe sobre los datos que este contiene para poder entender la dimensión de los mismos y las diferentes propiedades con las que podemos trabajar.

Vista rápida

Una vez que se leyeron los datos, se realizó una visualización rápida de los mismos con la función tail() y head() de python pandas.
En esta visualización pudimos ver las columnas que tiene el set de datos y que varias de ellas presentan NaN, por lo que antes de empezar a trabajar con ellos debemos hacer una limpieza de datos corruptos.
También, notamos que varias de las columnas en cuestión no aportan datos significativos a nuestro análisis, como es el caso de la columna “image_thumbnail”.
También notamos que hay varias propiedades que no pertenecen a la zona que estamos analizando (Capital Federal y Gran Buenos Aires) por lo que vamos a tener que filtar las propiedades que no pertenecen a nuestra zona de análisis.
Al ver las fechas tanto del tail como del head, se puede ver que existe cierto orden cronológico de las entradas, aunque eso no lo podamos comprobar.

Info

Al realizar la acción de info() notamos que tenemos un total de 27 columnas con 187481 filas. La mayoría de las columnas son del tipo Object excepto aquellas columnas relacionadas al precio, a las referencias geográficas y a las relacionadas con las cantidades de ambientes y pisos del edificio que son del tipo float.

Columns

Al analizar el output de la función columns, esta nos da los nombres de las diferentes columnas que tiene el dataset:

- Id: Identificador único de la operación.
- Created_on: Fecha de creación de la entrada.
- Operation: Tipo de operación (En este caso todas de venta).
- Property_type: Tipo de propiedad (Casa, Departamento, etc).
- Place_name: Ciudad donde queda la propiedad.
- Place_with_parent_names: Barrio, ciudad, provincia y país donde queda la propiedad separado por |.
- Country_name: País donde queda la propiedad.
- State_name: Provincia donde queda la propiedad.

- Geonames_id: Identificador del lugar de la propiedad en la base de datos de GeoNames (http://www.geonames.org/).
- Lat-lon: Latitud y longitud de la propiedad.
- Lat: Latitud de la propiedad.
- Lon: Longitud de la propiedad.
- Price: Precio de la propiedad.
- Currency: Moneda en la cual está expresada el precio de la propiedad.
- Price_aprox_local_currency: Precio aproximado en moneda local.
- Price_aprox_usd: Precio aproximado en dólares.
- Surface_total_in_m2: Superficie total de la propiedad en metros cuadrados.
- Surface_covered_in_m2: Superficie cubierta de la propiedad en metros cuadrados.
- Price_usd_per_m2: Precio del metro cuadrado en dólares.
- Price_per_m2: Precio del metro cuadrado expresado en la moneda de ´currency´.
- Floor: Piso en donde se encuentra la propiedad.
- Rooms: Cuartos que posee la propiedad.
- Expenses: Expensas que paga por mes la propiedad.
- Properati_url: Link a la publicación de Properati.
- Description: Descripción de la propiedad en la publicación.
- Title: Título de la propiedad en la publicación.
- Image_thumbnail: Imagen principal de la propiedad.

Como notamos, hay varias columnas como lat-lon que no nos suman ningún tipo de información, ya que es dato lo podemos tomar tranquilamente desde lat y lon por separado o el atributo Place_with_parent_names ya que esos datos los tenemos en otros atributos más atómicos.

Describe

Con la función describe() podemos obtener información acerca de los atributos numéricos del dataframe como el promedio, la desviación standar, el mínimo y el máximo de un atributo numérico.
Si bien no podemos tomar mucha información de esta tabla, notamos que el mínimo valor de varias de las columnas es 0. Este es un valor válido pero no significa que tenga sentido,, por ejemplo nadie vendería su propiedad a valor $0, mientras que en nuestro set de datos tenemos que el valor mínimo de una propiedad es 0, ya sea en dólares como en pesos.
Esto nos indica que estamos ante un problema de inconsistencias de datos y que dicho dato con el valor de propiedad 0 debe ser filtrado.
La fila count nos indica la diferencia entre valores NaN encontrados en cada columna.

Acondicionamiento del set de datos

Se procedió a realizar acciones sobre el set de datos con el fin de quedarnos con la información relevante que nos facilite analizar el set de datos y realizar visualizaciones. Por ejemplo, los datos de la columna de imágenes o url no son relevante para nuestro análisis, así como también los de precio en moneda local ya que vamos a trabajar con precios en dólares.
Sin embargo, la utilización de algoritmos de machine learning sobre las imagenes para predecir el precio del inmueble podría ser relevante, pero las imágenes no están normalizadas en resolución y tamaño lo cual requiere un procesamiento extensivo que no vale la pena llevar a cabo.

Filtro de propiedades

Filtro por lugar de ubicación

En el presente análisis, solo se trabaja con las propiedades ubicadas en Capital Federal y Gran Buenos Aires. Por ello, se realiza un filtro de las propiedades que no estén ubicadas dentro de las zonas a analizar y se trabaja con aquellas cuyo atributo “state_name” sea de valor “Bs.As. G.B.A. Zona Norte”, “Bs.As. G.B.A. Zona Sur”, “Bs.As. G.B.A. Zona Oeste” y “Capital Federal”:

```
dataBA = data %>% subset(state_name == "Bs.As. G.B.A. Zona Norte" |
                         state_name == "Bs.As. G.B.A. Zona Sur" |
                         state_name == "Bs.As. G.B.A. Zona Oeste" |
                         state_name == "Capital Federal")
```

Filtro por valores de atributos

Filtro por valor en dólares

Uno de los atributos más importante del set de datos o el más importante es el atributo “Price_aprox_usd” ya que es el atributo que en una segunda instancia vamos a intentar predecir.
Si este atributo encuentra un valor del tipo “NaN” o un valor en 0, filtramos la fila ya que consideramos que el atributo está corrupto y la instancia no nos sirve de nada.
En este caso vamos a crear 2 datasets: uno llamado train y otro llamado test:

```
datatest <- dataBA %>% subset(is.na(price_aprox_usd) | price_aprox_usd
== 0)
datatrain <- dataBA %>% subset(!(is.na(price_aprox_usd) |
price_aprox_usd == 0))
```

```
datatest$price_aprox_usd <- NULL
datatrain$price_aprox_usd <- as.integer(datatrain$price_aprox_usd)
```

Y los datasets resultantes son:

```
dim(datatrain)             [1] 153771     15
dim(datatest)               [1] 20560       14
```

Selección de columnas

No todas las columnas del set de datos nos dan información relevante. Por ello, podemos eliminar ciertas columnas que son irrelevantes para nuestro análisis y así poder trabajar con un set de datos más reducido.
Las columnas a eliminar son:

- Operation: Siempre es sell porque tomamos un dataset especifico de ventas.
- Place_with_parent_names: Como se dijo anteriormente, este campo es una concatenación de otros campos que podemos encontrar en el dataset.
- Country_name: Al realizar los filtros quedan solo datos de Capital Federa y Gran Buenos Aires de Argentina.
- Lat-lon: Como bien se comentó anteriormente, la latitud y longitud también vienen por separado.
- Price: Solo usamos el precio aproximado en dólares.
- Currency: Siempre tomamos en dólares.
- Price_aprox_local_currency: Se usa el precio aproximado en dólares.
- Price_per_m2: Usamos el precio por metro en dólares.
- Properati_url: La URL a la publicación no nos suma ningún valor adicional.
- Image_thumbnail: La imagen de la publicación no nos suma ningún valor adicional.

Como resultado obtenemos:

```
dataBA$operation <- dataBA$place_with_parent_names <- dataBA$country_name <-
dataBA$lat.lon <- NULL
dataBA$price_aprox_local_currency <- dataBA$price <- dataBA$currency <-
dataBA$price_per_m2 <- NULL
dataBA$properati_url <- dataBA$image_thumbnail <- NULL
dataBA$created_on <- dataBA$geonames_id<- NULL
```

Acondicionamiento de textos

Se toma el título y la descripción de las propiedades, se realiza una limpieza, encoding y una concatenación de ellos al fin de tener solo 1 atributo con los textos resultantes
Para esto se realizan los siguientes pasos:

1. Los textos están cargados en UTF-8, se los convierte a ASCII.
2. Se une la descripción con el título en una variable.

3. Se pasan todos los textos a minúscula.
4. Se remueven las “stopwords” del diccionario español, los números y los dígitos especiales.
5. Se quitan los espacios en blanco extras.

Código que ejecuta estas acciones:

```
dataText <- datatrain[c("description","title","price_aprox_usd")]
dataText$texto = paste(dataText$description,dataText$title, sep = " ")
dataText$texto <- tolower(gsub("[^[:alnum:] ]", " ",dataText$texto)) %>%
       removeWords(.,stopwords("spanish")) %>%
      gsub("[[:digit:]]+", " ",.) %>%
      gsub(" *\\b[[:alpha:]]{1,2}\\b *", " ",.) %>%
       gsub("\\s+", " ",.) %>% as.character(.)
```

Analisis

Análisis de textos

Se toma el campo texto del dataset y se hace un split para quedarse con las palabras sin contar la existencia extra de la misma en cada texto.
Luego, se aplica una función de hash con un tamaño de 2^24 y se conserva el mapping para no perder referencia:

```
f <- ~ split(texto, delim = " ", type = "existence")
d1 <- hashed.model.matrix(  f ,
                            data = dataText, hash.size = 2^24,
                            create.mapping = T)
```

La matriz resultante es una matriz dispersa, por ello, se aplican ciertas transformaciones con el fin de filtrar aquellas que tengan pocas apariciones.
Para crear un wordcloud se segmentan los datasets según el precio:

```
dataText1 <- dataText %>% subset(price_aprox_usd < 50000)
dataText2 <- dataText %>% subset(price_aprox_usd >= 50000 &
price_aprox_usd <80000)
dataText3 <- dataText %>% subset(price_aprox_usd >= 80000 &
price_aprox_usd <115000)
dataText4 <- dataText %>% subset(price_aprox_usd >= 115000 &
price_aprox_usd <150000)
dataText5 <- dataText %>% subset(price_aprox_usd >= 150000 &
price_aprox_usd <200000)
dataText6 <- dataText %>% subset(price_aprox_usd > 200000)
```

Frecuencia de palabras con precio menor a U$S50000:

Frecuencia de palabras con precio mayor a U$S50000 y menor a U$S80000

Frecuencia de palabras con precio mayor a U$S80000 y menor a U$S115000

Frecuencia de palabras con precio mayor a U$S115000 y menor a U$S150000

Frecuencia de palabras con precio mayor a U$S150000 y menor a U$S200000

Frecuencia de palabras con precio mayor a U$S200000

Para los wordclouds anteriores se filtraron las palabras que eran predominantes en todos los segmentos:

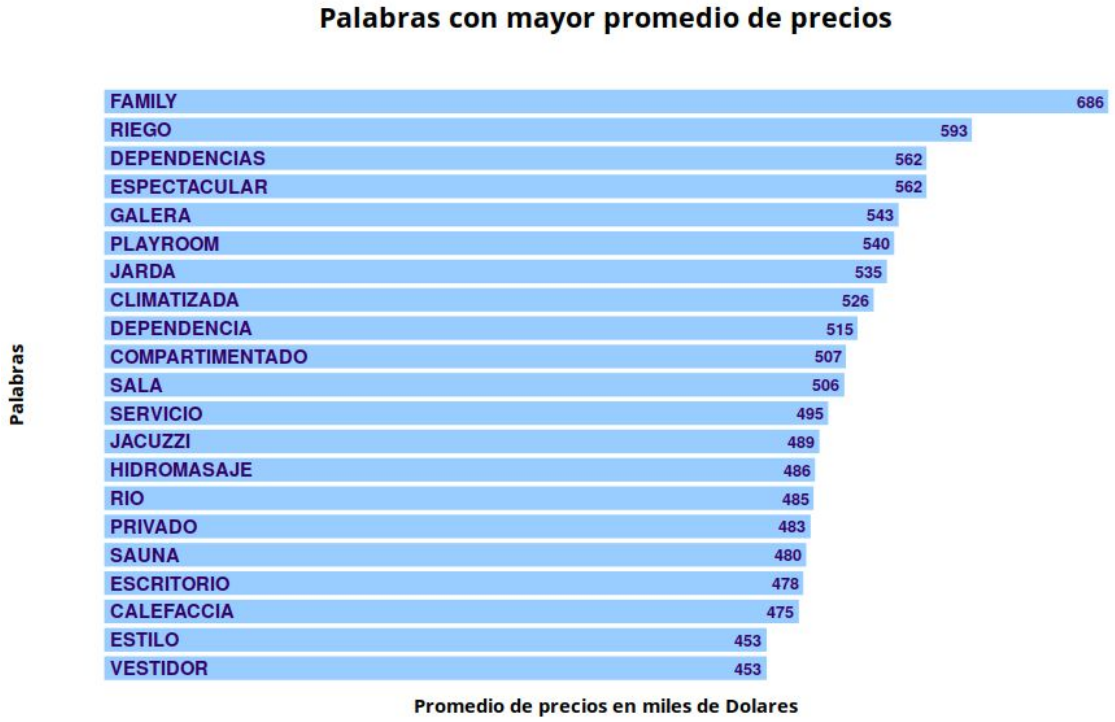
```
aver<-aver[!(aver$name=="departamento" |
               aver$name=="venta" |
               aver$name=="cocina" |
               aver$name=="bano" |
               aver$name=="null" |
               aver$name=="ambientes" |
               aver$name=="ambiente"  |
               aver$name=="dormitorios" |
               aver$name=="dormitorio" |
               aver$name=="comedor" |
               aver$name=="living" |
               aver$name=="completo"),]
```

Promedio de precio por palabras

Se puede apreciar en las nubes, que ciertas palabras son más grandes, por consecuencia son más comunes en publicaciones que tienen determinada franja de precios como por ejemplo “monoambiente”, “suite”, “balcón”, “parrilla”, etc.
Posteriormente se hace un promedio de precios por palabras, filtrando las palabras con pocas ocurrencias (mayores a 5000 ocurrencias).

Palabras con menor promedio de precios

Palabras con mayor promedio de precios

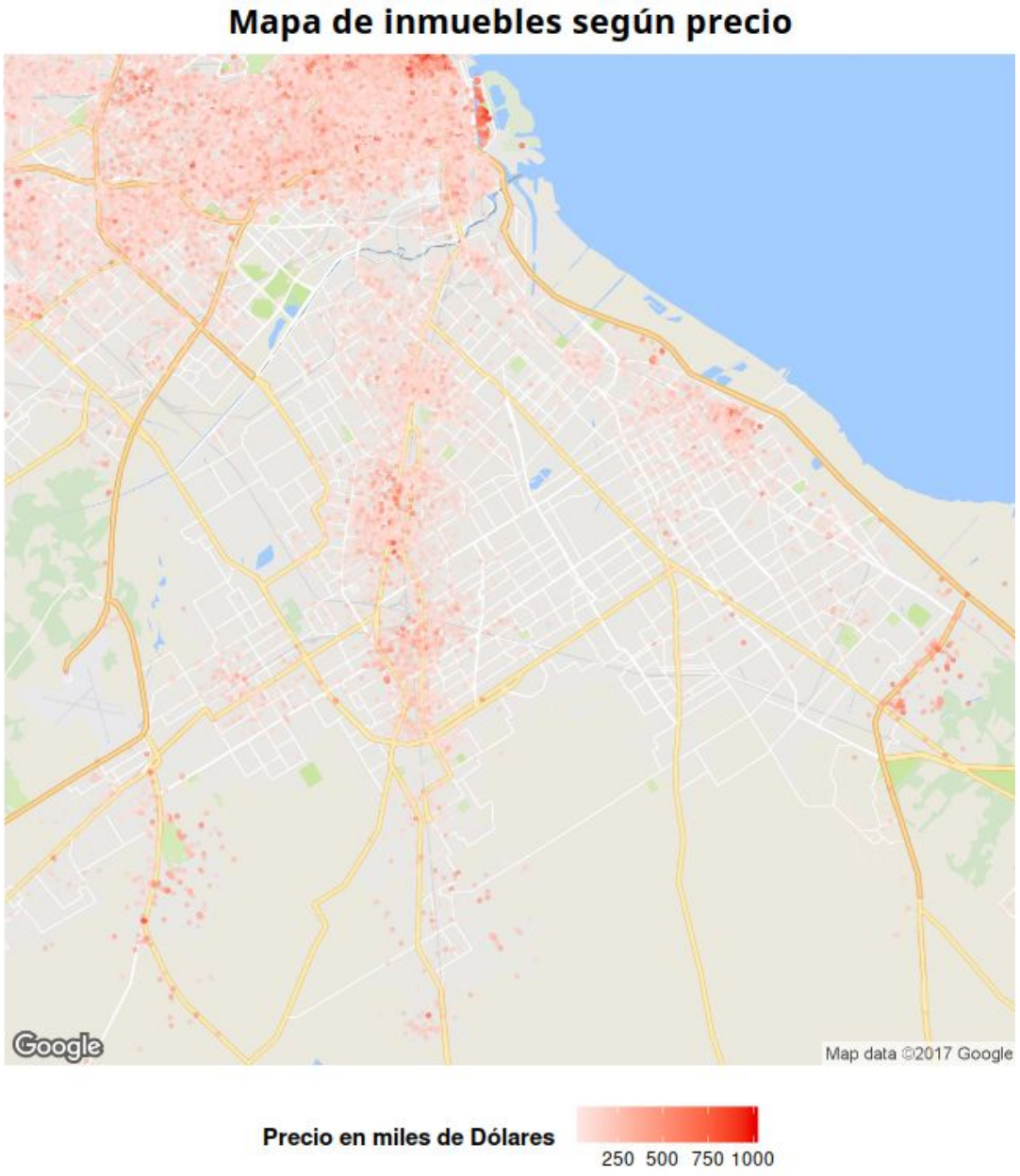


Se puede apreciar en estos últimos gráficos de barras que las palabras “jacuzzi”, “hidromasaje”, “sauna”, “riego”, “playroom”, etc; son palabras características de precios altos y “monoambiente”, “créditos”, “lavarropas”, “colectivos” de precios bajos.

Análisis de localización

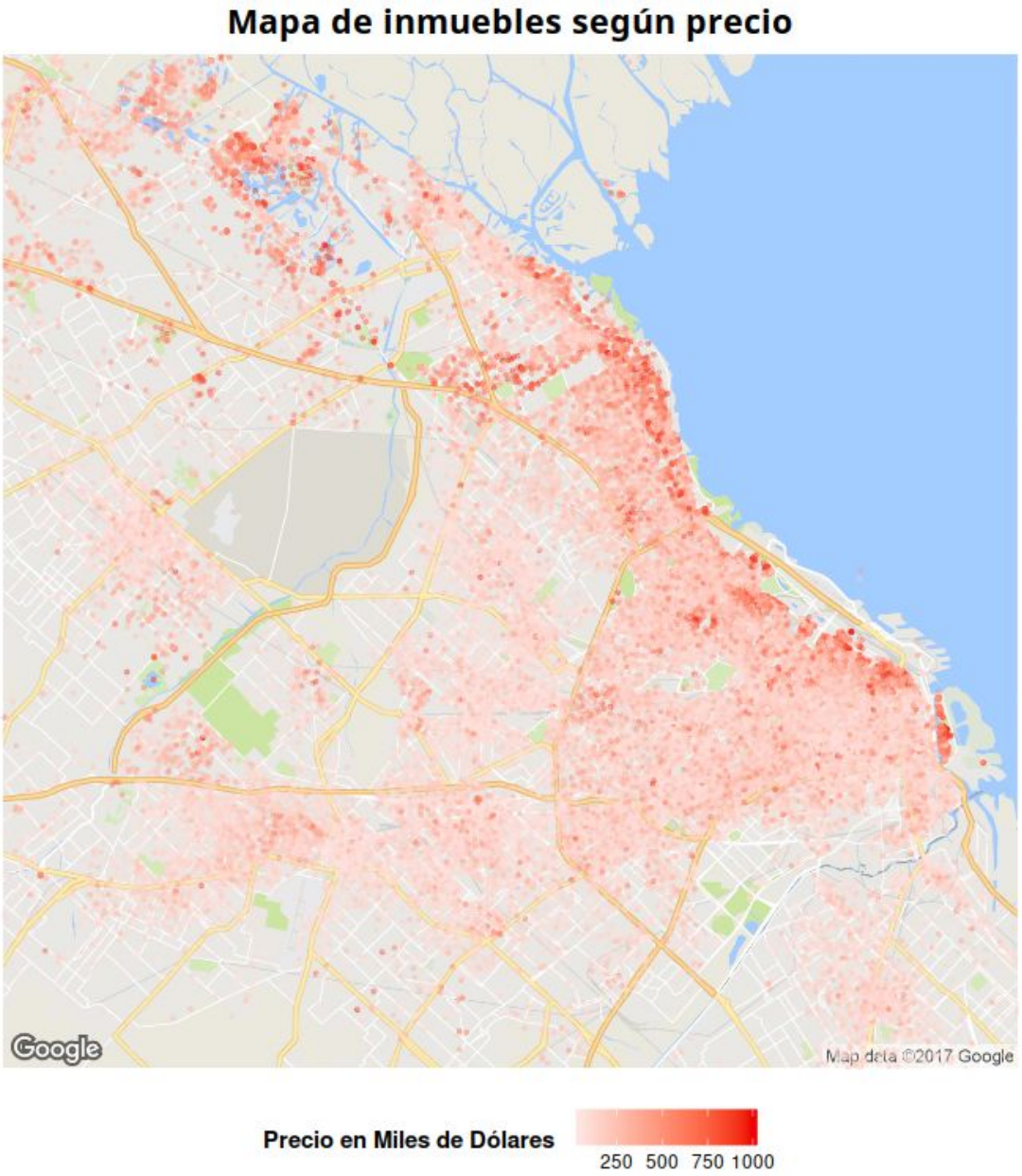
La localización del inmueble parece ser determinante para el precio del mismo, hay zonas que son evidentemente más caras que otras como se muestra a continuación, principalmente el área del corredor norte de la Capital Federal, Zona Norte, Nordelta y Puerto Madero.
Se separaron los datos entre los inmuebles de valor menor a U$S1000000 y mayor a ese precio como así también por zonas aledañas a Capital Federal (También incluída).

Zona Sur inmuebles de valor menor a U$S1000000



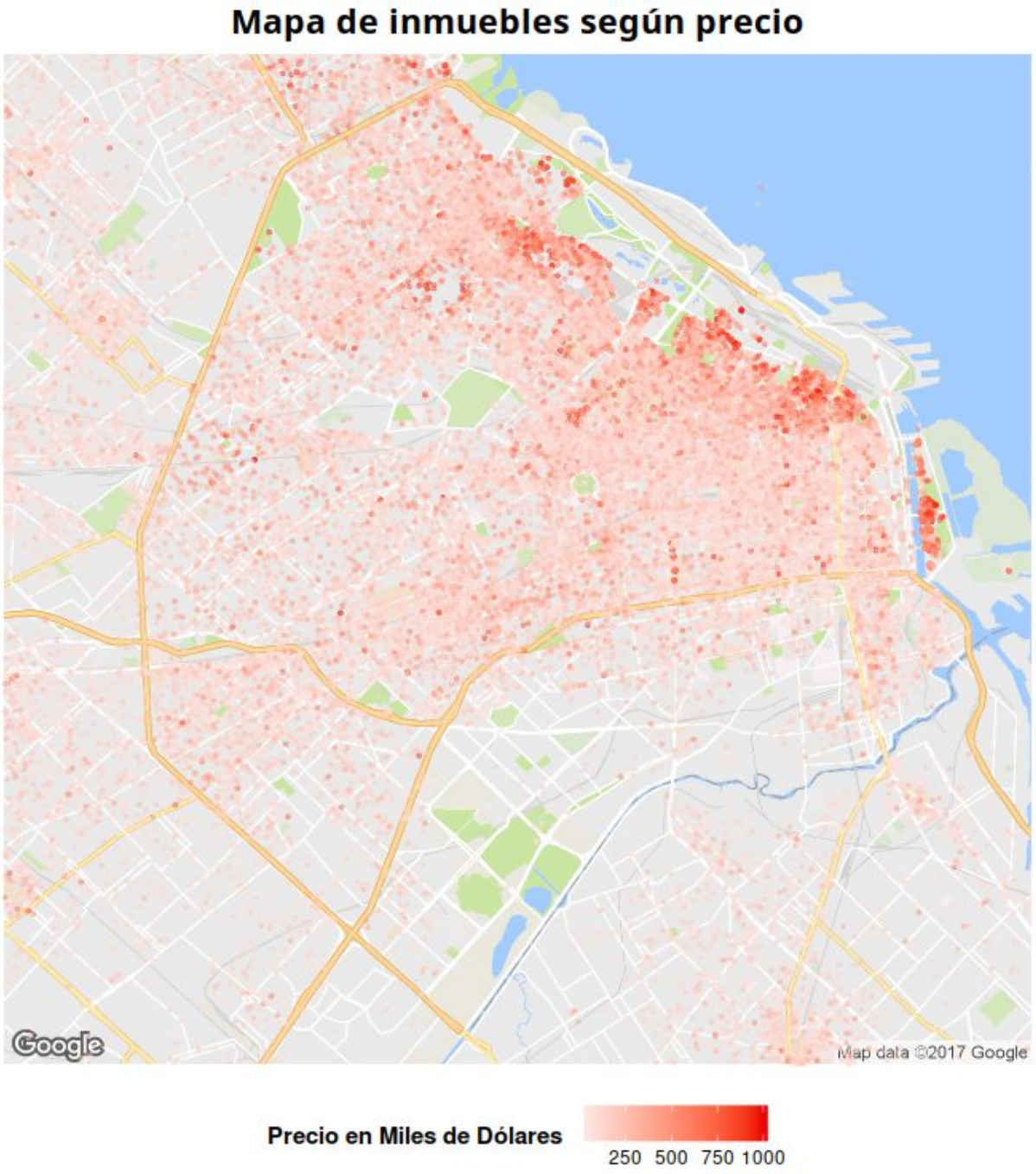
En el primer mapa se ve la Zona Sur de inmuebles menores a un millón de dólares, cuando el rojo se ve más intenso el precio del inmueble es más elevado.
Se aprecian tres zonas marcadas con rojo, Lomas De Zamora, Quilmes, Monte Grande.

Zona norte y Capital inmuebles de valor menor a U$S1000000



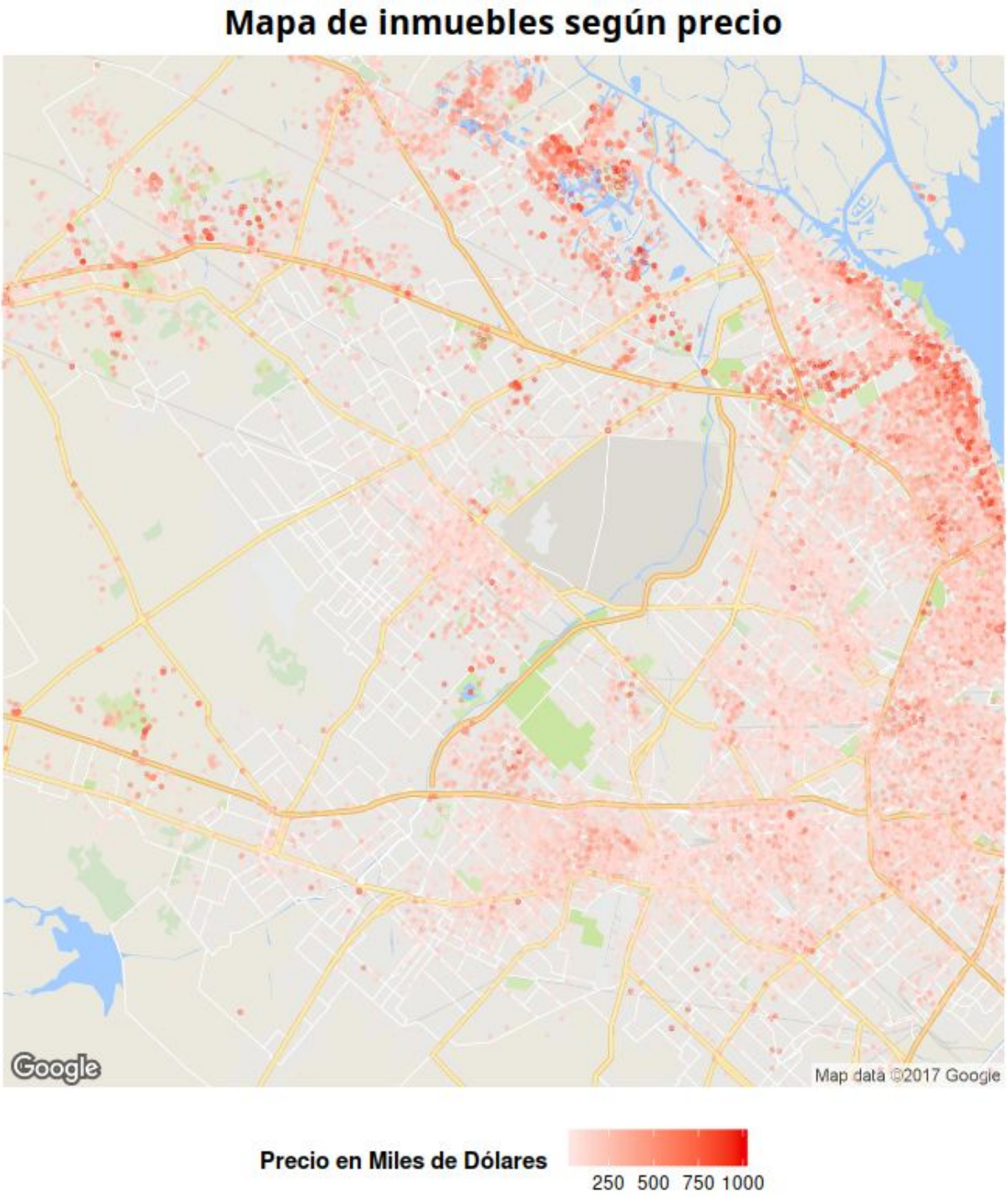
En el mapa de Capital Federal y Zona Norte del Gran Buenos Aires de inmuebles menores a un millón de dólares, queda bien marcado el corredor que empieza en Retiro y termina en Nordelta, también se puede ver la marcado Puerto Madero.

Capital Federal inmuebles de valor menor a U$S1000000



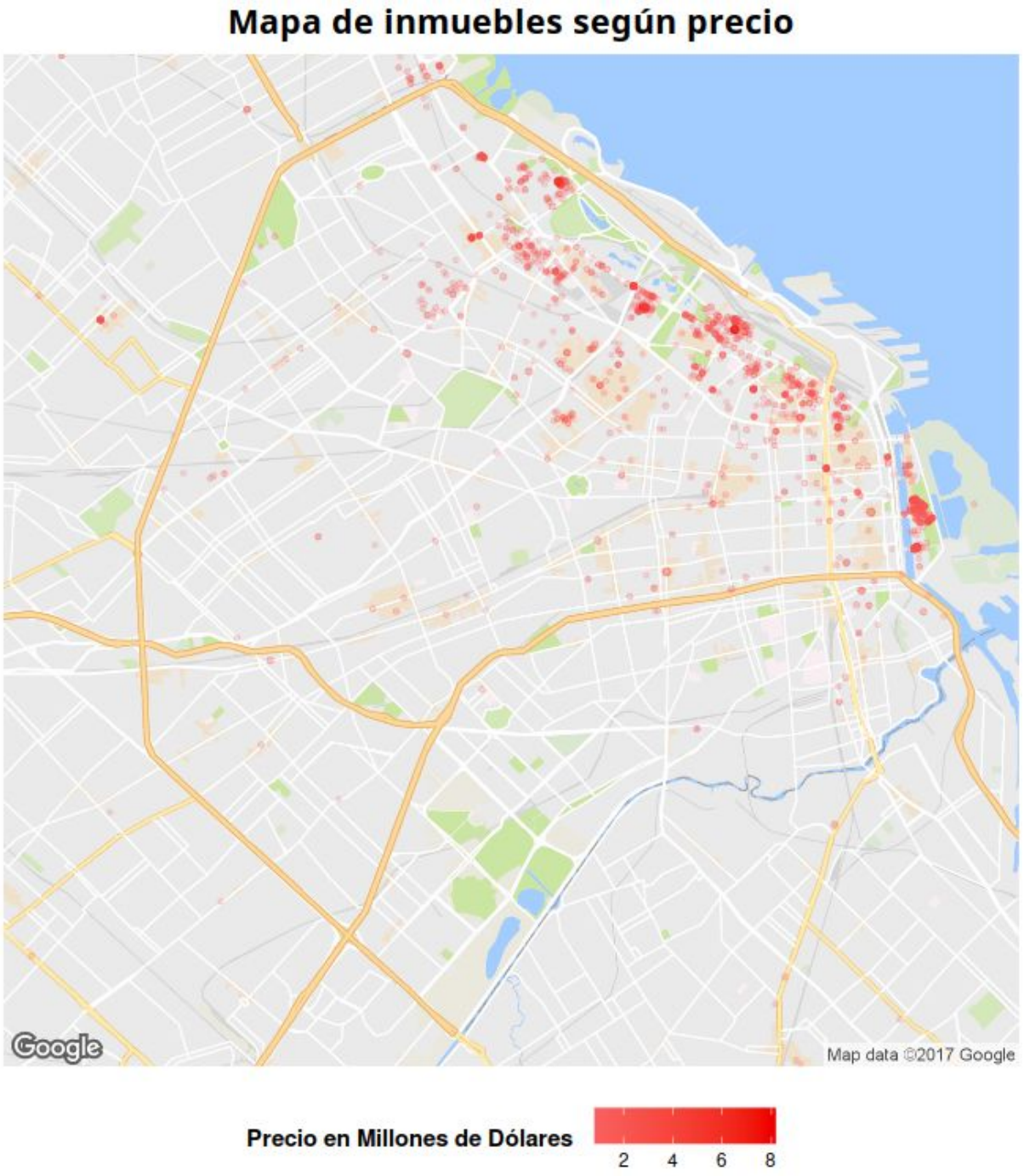
Como bien ya se vió en el mapa anterior, en Capital Federal existe un corredor bien marcado alrededor de lo que es la avenida del Libertador empezando en Retiro y terminando en su límite con la provincia de Buenos Aires.
Este corredor también se extiende por lo que es Puerto Madero.

Zona oeste inmuebles de valor menor a U$S1000000



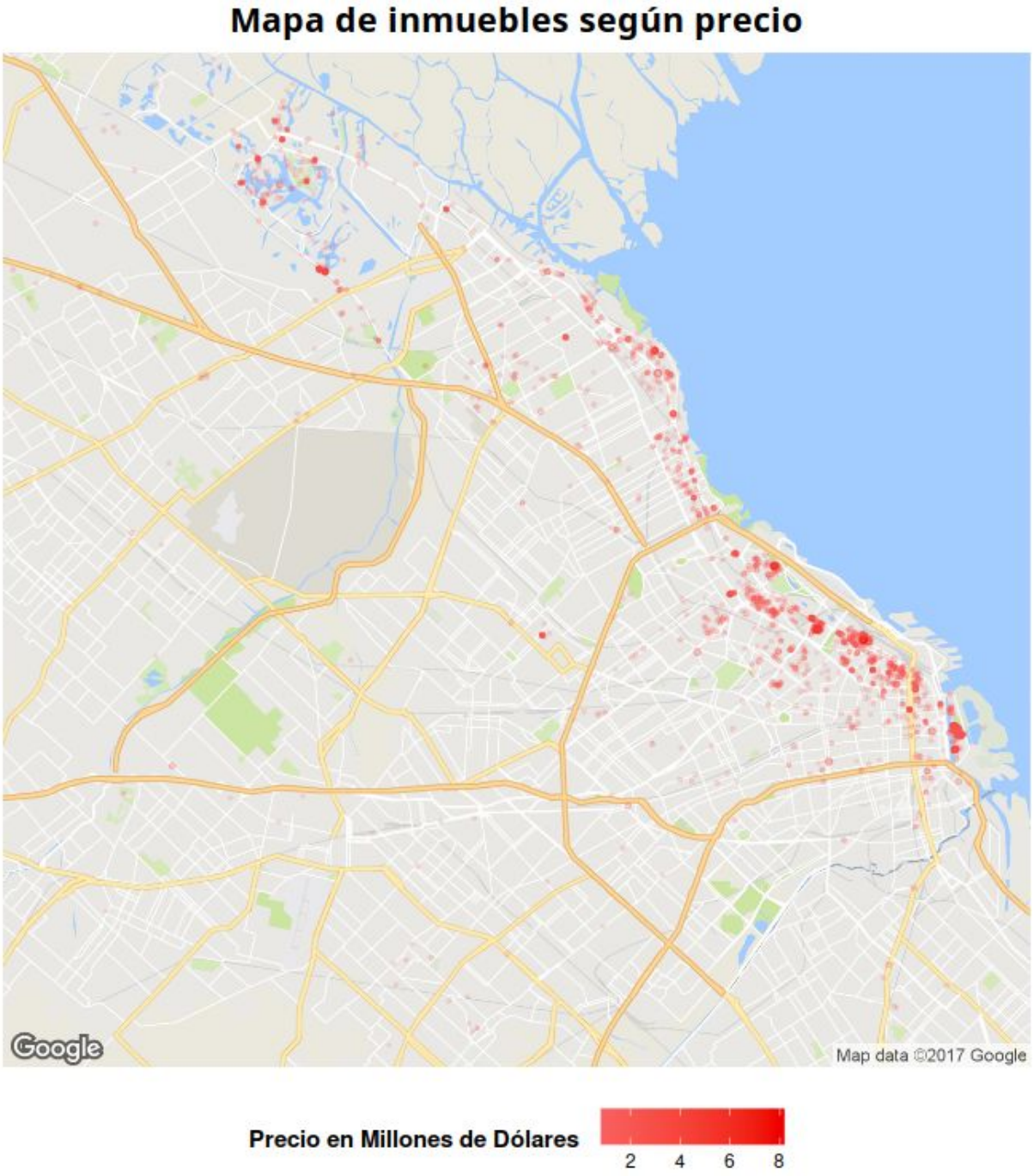
En este mapa podemos divisar tanto la zona oeste como la zona norte del Gran Buenos Aires.
Como bien se aprecia, la zona oeste tiene casos aislados de inmuebles que superen los U$S750000 dólares.

Capital Federal inmuebles de valor mayor a U$S1000000



En este mapa queda aún más marcado el corredor que empieza en retiro y se extiende hacia la zona norte del Gran Buenos Aires con su apéndice en Puerto Madero.
Se pueden focalizar también, zonas muy específicas que podrían marcar puntos concretos hacia los cuales se pueden calcular distancias.

Zona norte y Capital Federal inmuebles de valor mayor a U$S1000000



Este mapa es una extensión del anterior y es muy visible el corredor que empieza en Retiro (o Puerto Madero) y termina en Nordelta.

Análisis de localidades

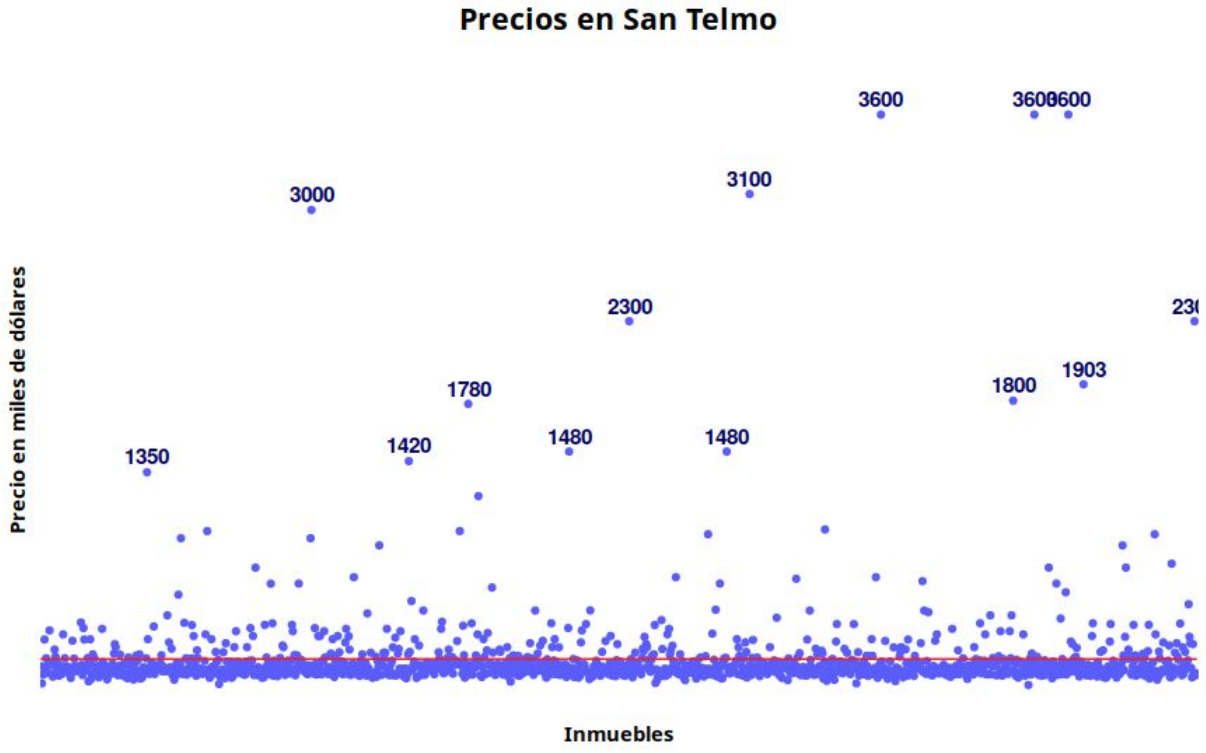
Para esta sección se van a analizar los precios de las publicaciones según la localidad en la que se encuentran.

Datos corruptos

Antes de analizar los precios por localidades se realizó un análisis de los datos presentes en ciertos barrios que no concordaban con la realidad ya que existen inmuebles de precios muy superiores a la media, por esto, se filtraron estos inmuebles ya que se consideraron como datos corruptos.

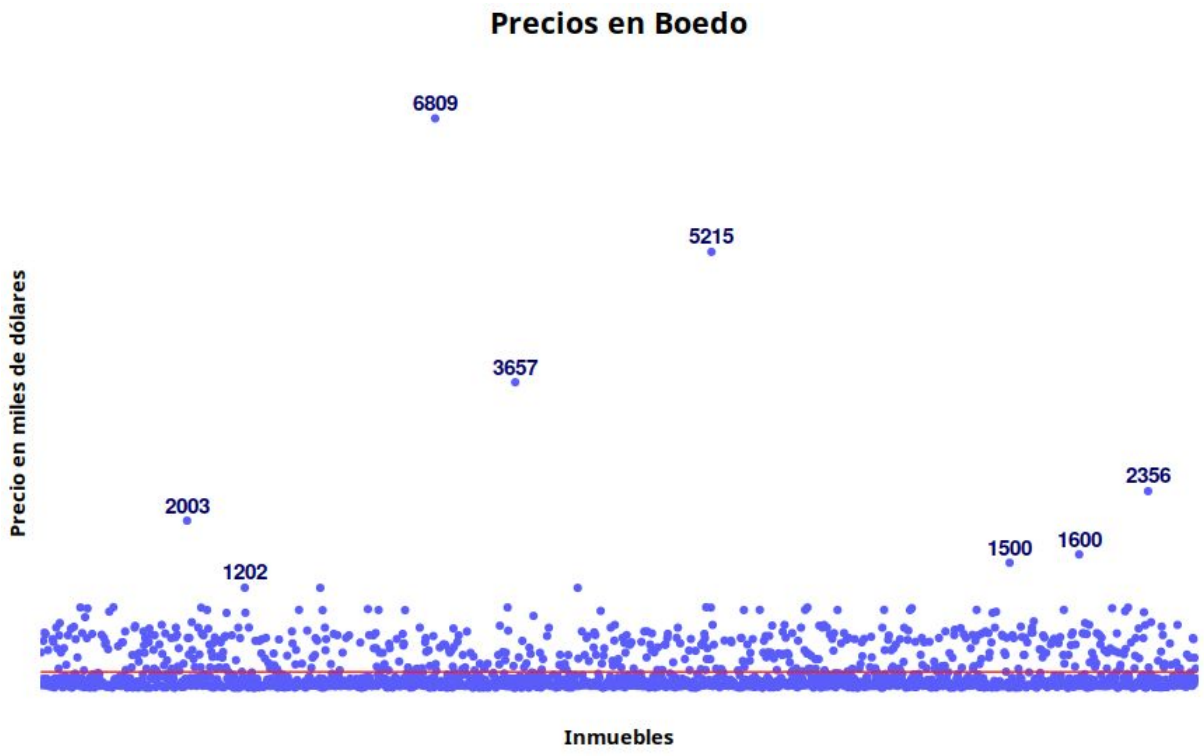
San Telmo

San Telmo tiene un inmueble que cotiza a 46 millones de dólares levantando el promedio casi en 30 mil dólares. Al filtrar este caso el gráfico de puntos se normaliza y se puede apreciar mejor la disposición. El promedio resultante es de 173 mil dólares se aprecia en la línea roja.



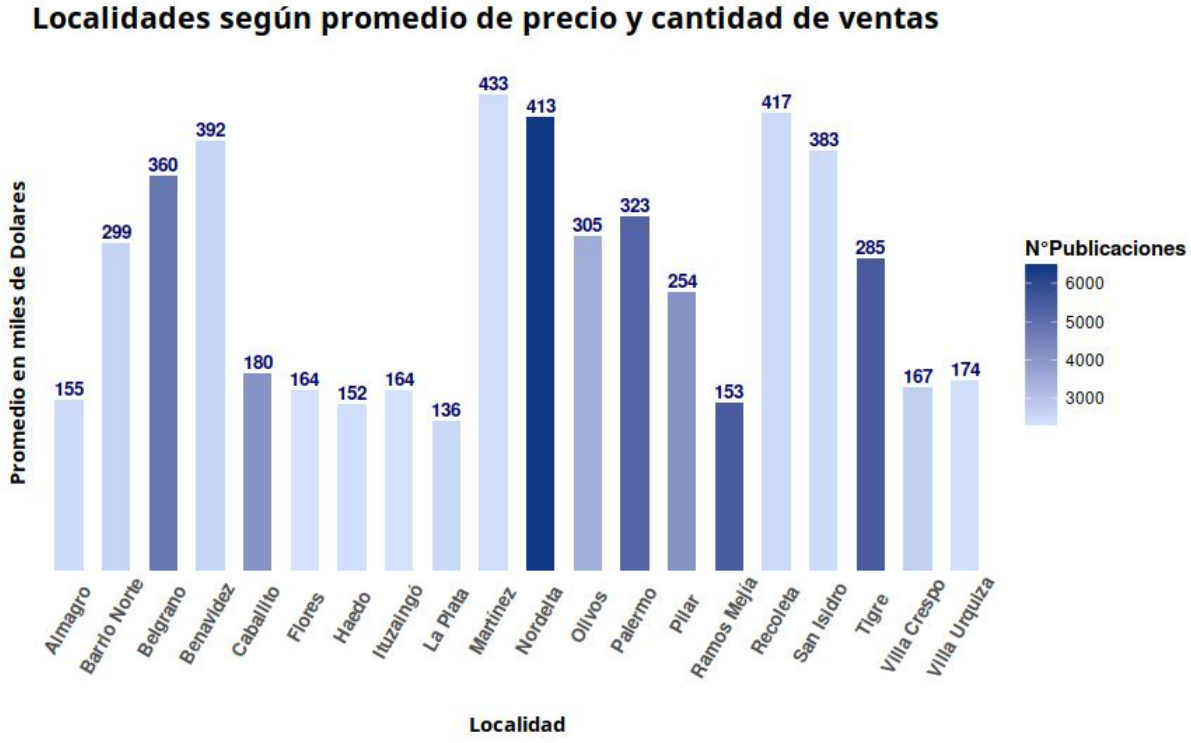
Boedo

También se analizó boedo con un par de publicaciones que rompen el promedio, pero por la cantidad el promedio no se mueve mucho.



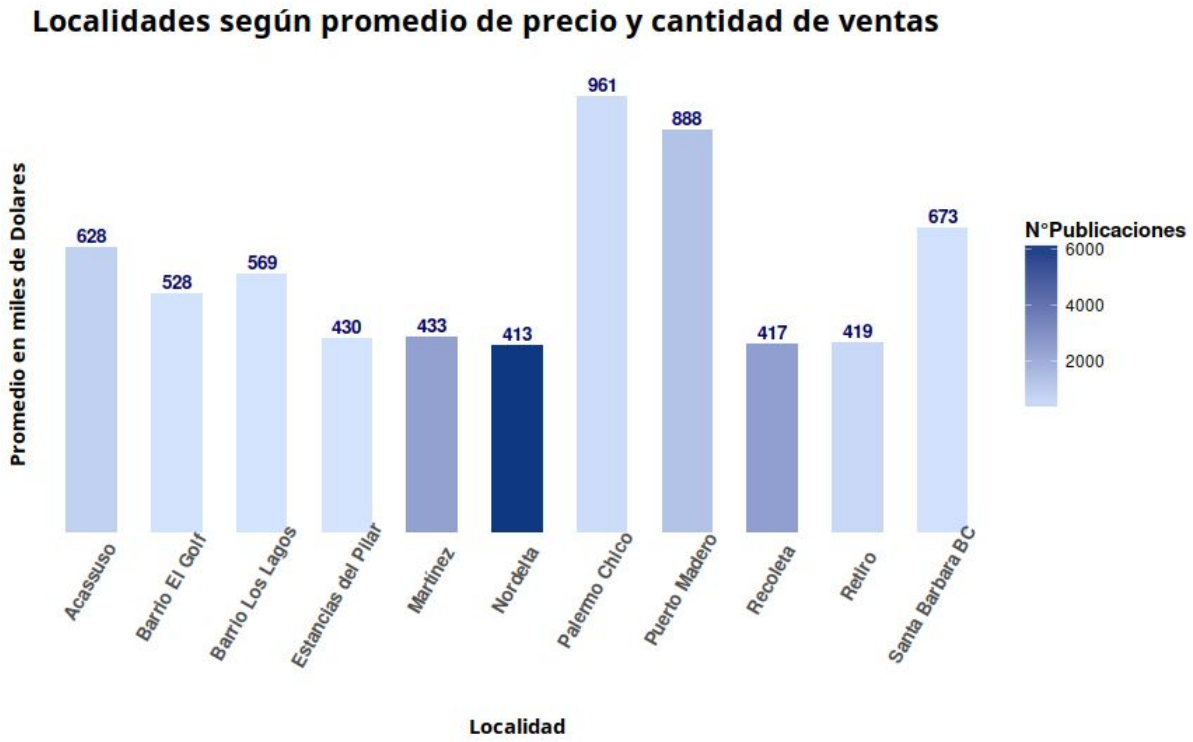
Localidades por precio y cantidad de publicaciones

Para realizar este análisis, se filtraron las localidades con menos de 2000 publicaciones.



Se puede ver como Nordelta, Palermo, Tigre, Belgrano tienen promedio de precios altos y tienen gran cantidad de publicaciones.

En el siguiente gráfico, el filtro de cantidad de publicaciones se llevó a 1000:

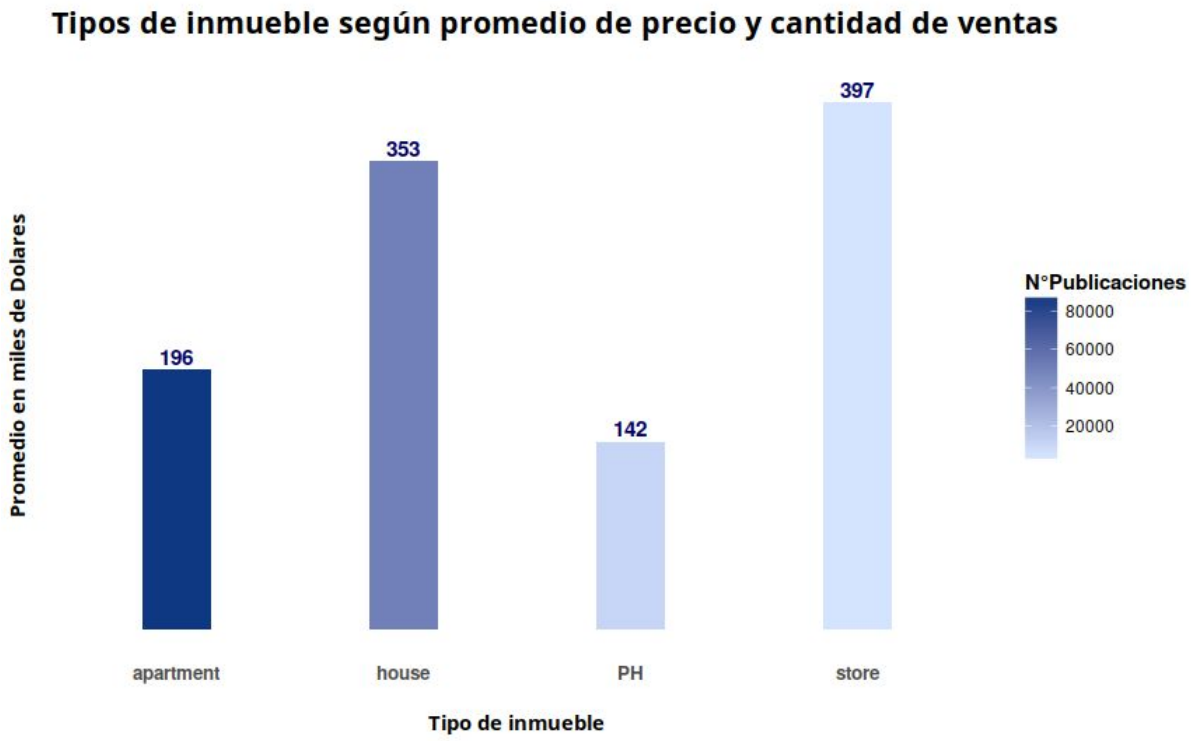


Se nota como barrios como Puerto Madero, Palermo Chico y Santa Bárbara tienen un promedio alto de precio pero la cantidad de publicaciones que existen en esos barrios son pocas (Entre 1000 y 2000 publicaciones).

Análisis por tipo de inmueble

El tipo de inmueble es un atributo fundamental a la hora de analizar el precio de los mismos.

Inmuebles por tipo y por cantidad de publicaciones

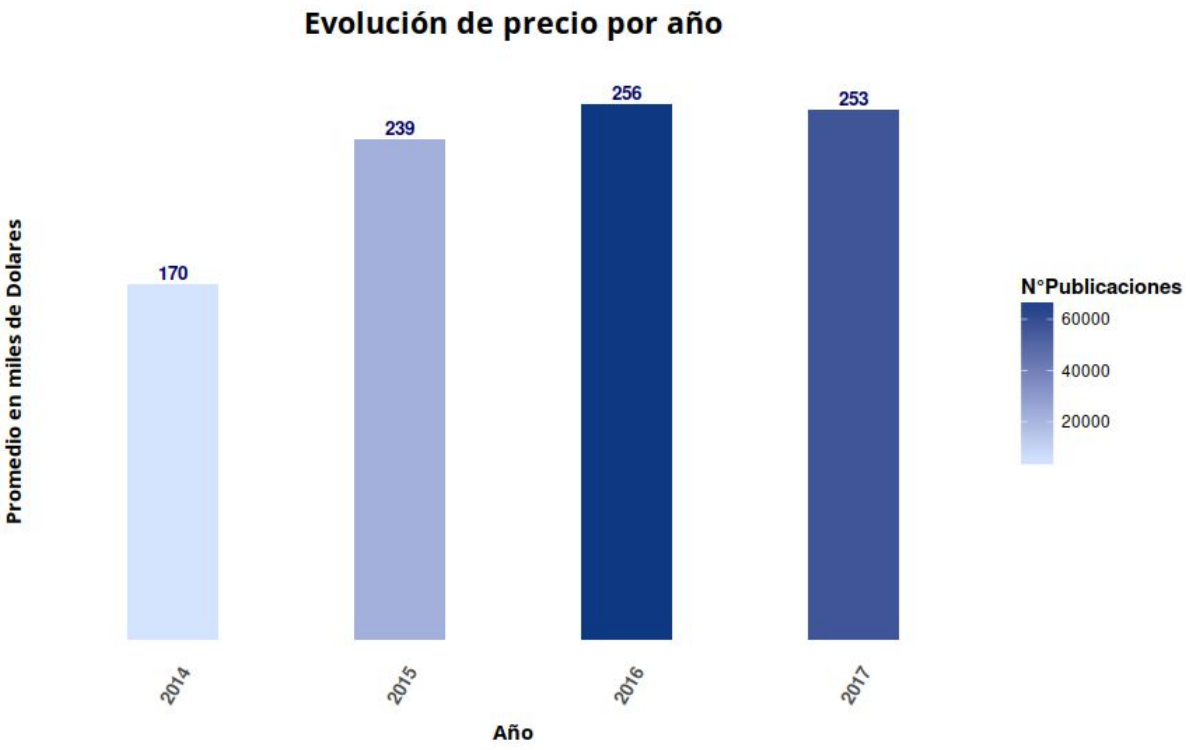


Se nota que las publicaciones de departamentos son mucho mayora que las de los demás tipos de inmuebles.
A pesar de esto, el promedio del valor de las casas y los negocios son ampliamente superiores al valor promedio de los departamentos mientras que el valor promedio de los PH son los menores.

Análisis por fecha

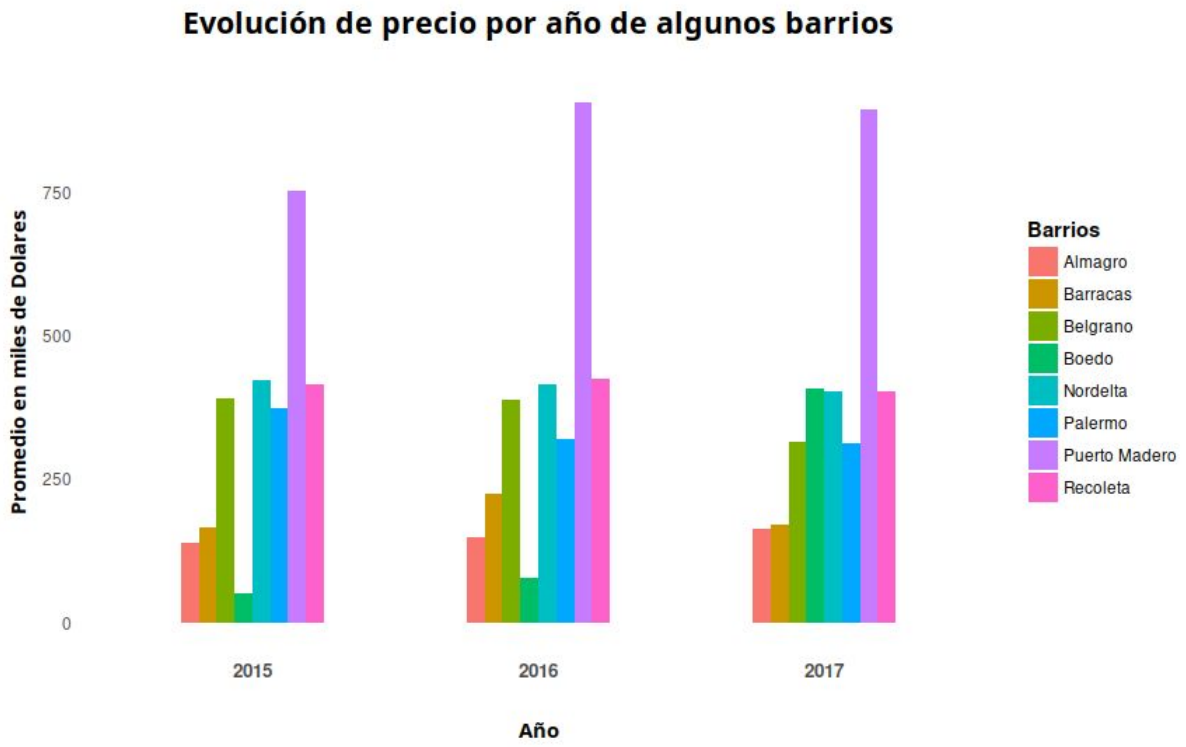
Las fechas de las publicaciones generalmente suelen determinar variaciones en los precios de venta de los inmuebles.
A continuación se analiza la variación de los precios en función de las fechas de venta.

Evolución del precio y cantidad de publicaciones por año



Se puede ver que el número de publicaciones está creciendo en el tiempo, teniendo en cuenta que el año 2017 está todavía en curso.
El año 2012 y el 2013 tienen muy pocas publicaciones y se filtraron para hacer la comparación, igualmente el año 2014 tienen muchas menos publicaciones que el 2016.

Evolución precio por año en barrios



También se puede ver como los el promedio de precio de cada barrio va variando levemente según el año de la publicación. En el gráfico se eligieron algunos barrios como ejemplos.
El precio promedio de Boedo viene creciendo a lo largo del tiempo y tuvo un crecimiento muy marcado entre 2016 y 2017.

Conclusiones

Las descripciones y los títulos de las publicaciones contienen información útil que es característica de distintas franjas de precios.
La localización del inmueble es bastante relevante dado que hay zonas donde se concentran las propiedades más caras, una posibilidad para la predicción es la utilización de distancias a estas zonas. De la misma manera hay barrios más caros que otros.
La utilización de algoritmos de machine learning sobre las imagenes para predecir el precio del inmueble podría ser relevante, pero las imágenes no están normalizadas en resolución y tamaño.