

# Problema

Presentar un modelo que sirva para predecir o calcular el precio de propiedades por metro cuadrado.

# Pregunta

Qué características de una propiedad dentro de la Comuna 13 influyen en el precio por metro cuadrado de casas, PHs y departamentos.

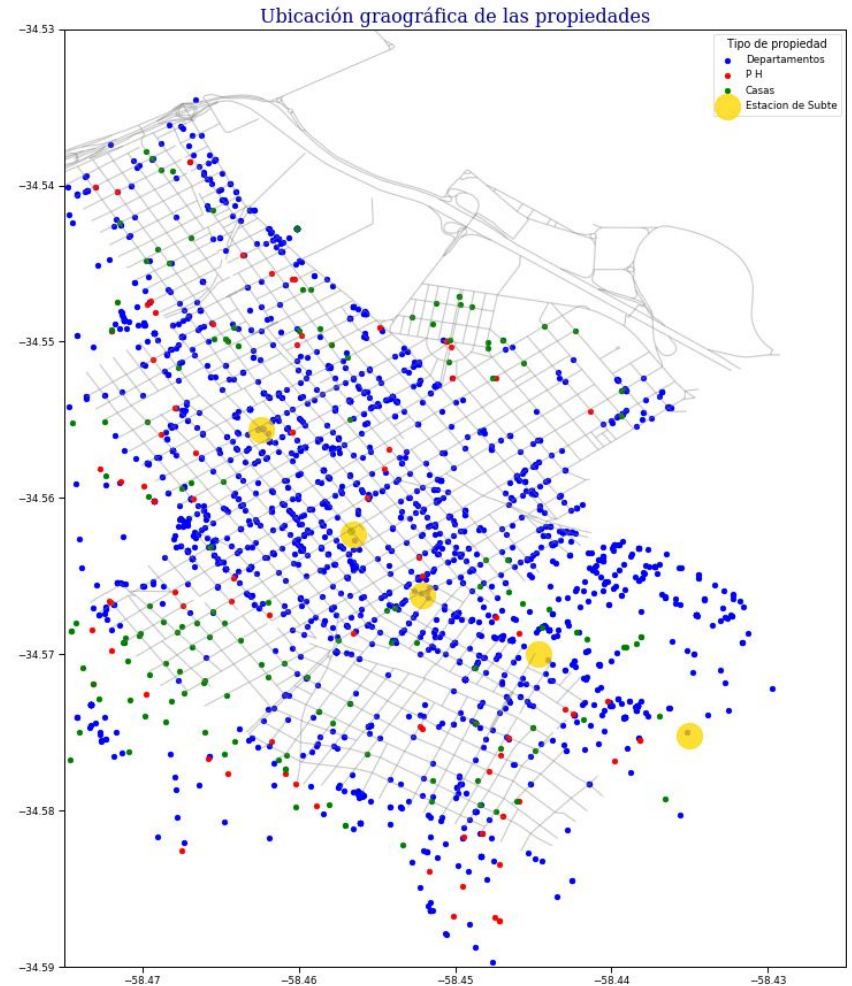
# Dataset / Información disponible

Dataset de Properati del primer semestre de 2017

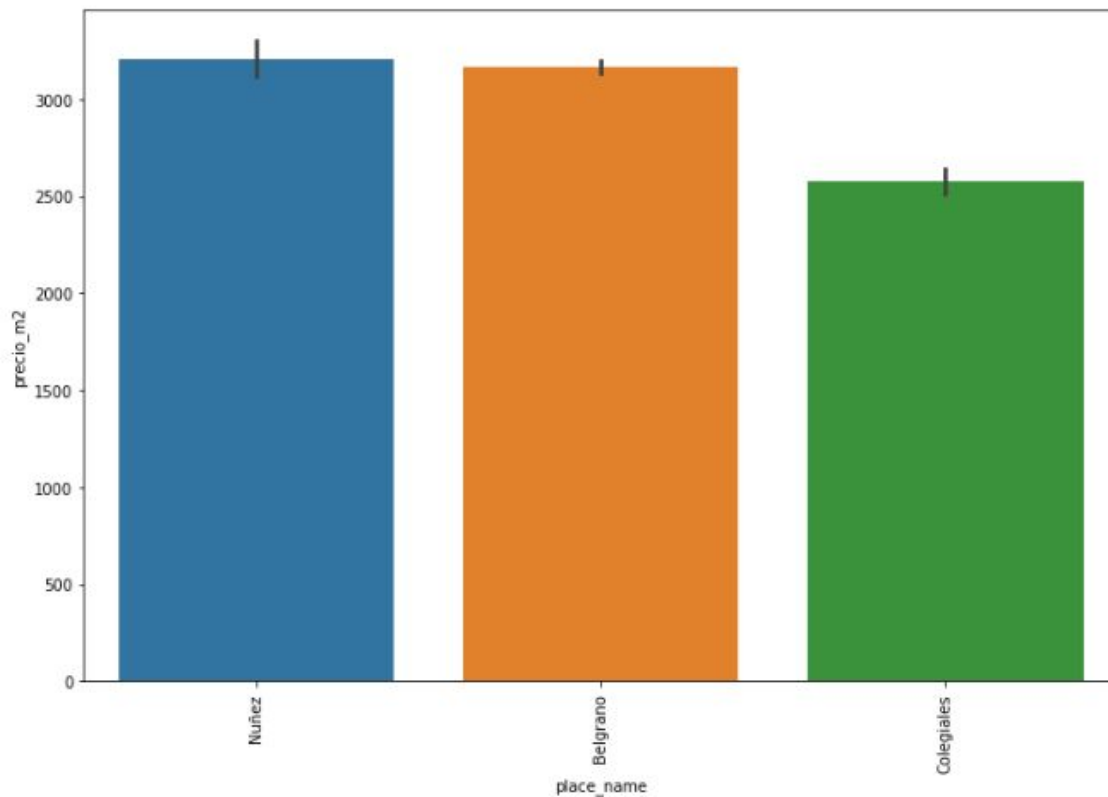
# Propiedades que tomamos para el análisis

Comuna 13  
(Nuñez, Colegiales,  
Belgrano)

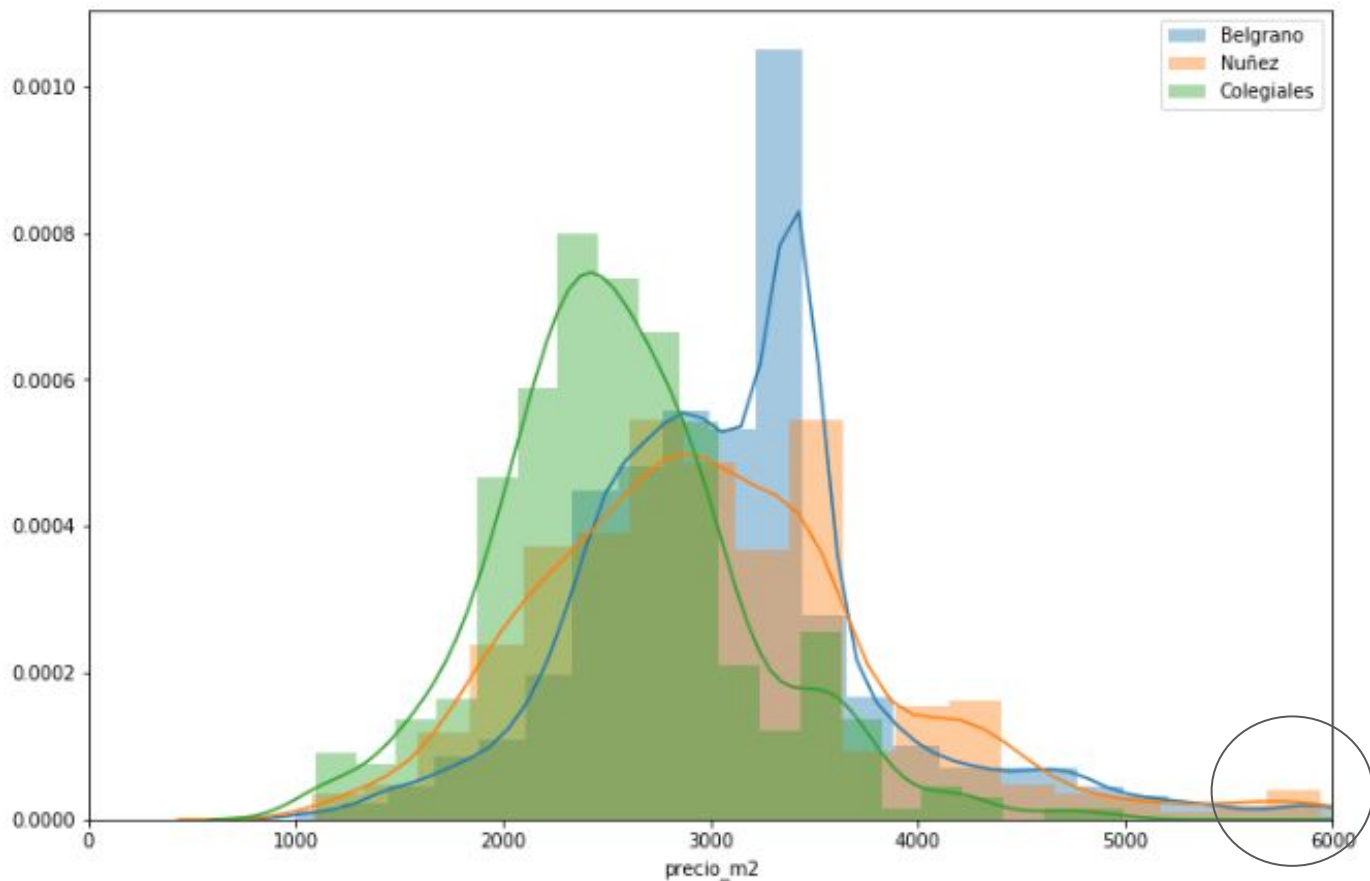
Departamentos 3727  
Casas 192  
PHs 102



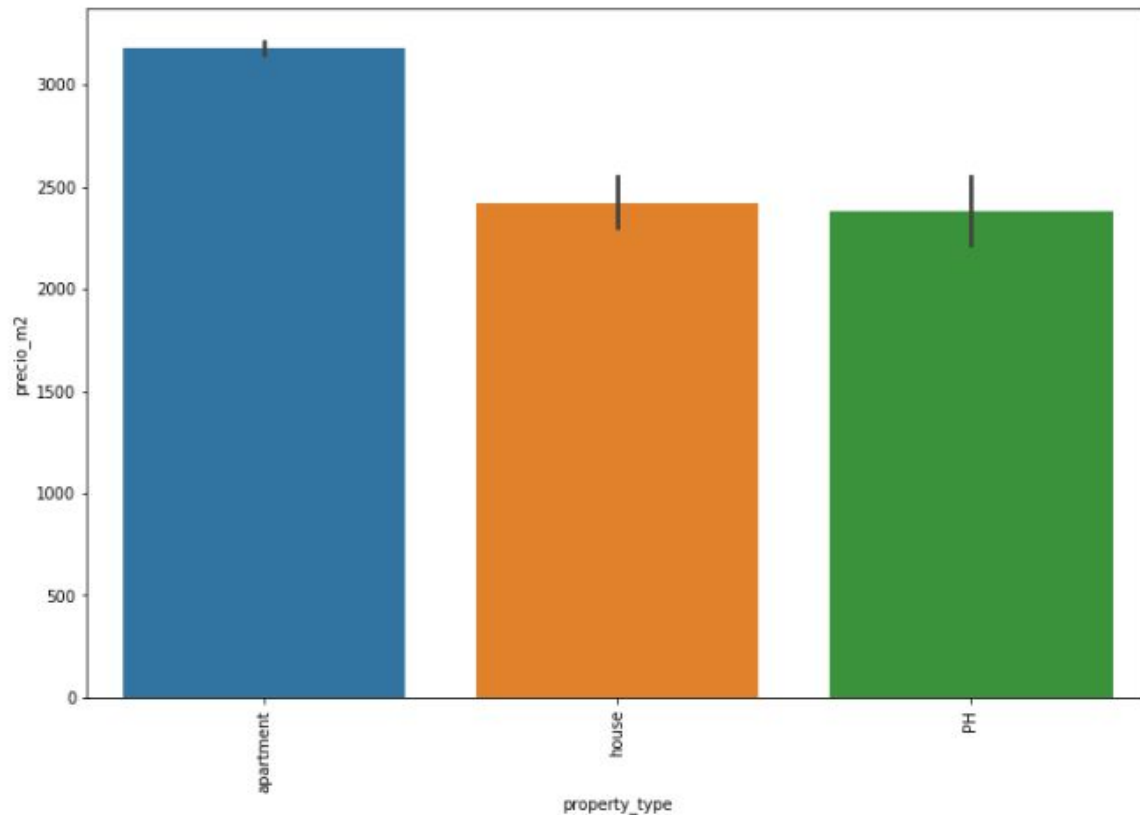
Medimos qué barrio tiene el precio por m2 más caro



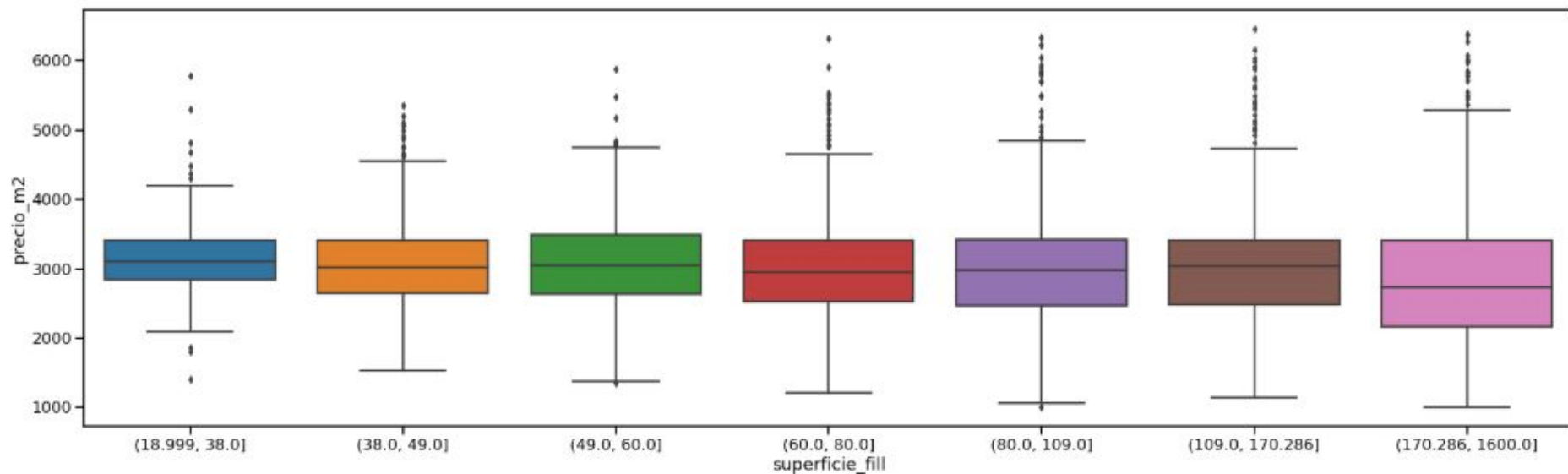
## Otra manera de verlo con la distribución



# Qué tipo de propiedad es más cara por m2



## Descripción de precios por m2 en dólares por categorías de superficie en boxplot

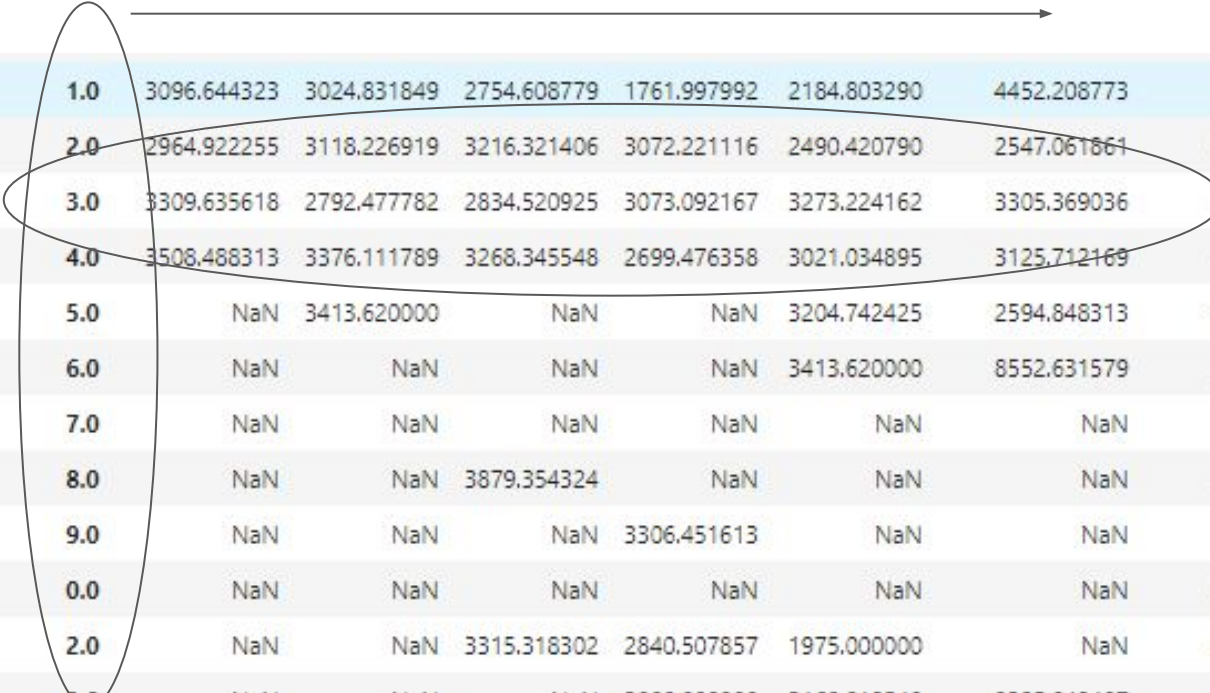


## Descripción de la media del precio por m2 por superficie, tipo de propiedad y cantidad de ambientes

apartment	1.0	3096.644323	3024.831849	2754.608779	1761.997992	2184.803290	4452.208773	NaN
	2.0	2964.922255	3118.226919	3216.321406	3072.221116	2490.420790	2547.061861	2191.247116
	3.0	3309.635618	2792.477782	2834.520925	3073.092167	3273.224162	3305.369036	3298.075667
	4.0	3508.488313	3376.111789	3268.345548	2699.476358	3021.034895	3125.712169	3836.972864
	5.0	NaN	3413.620000	NaN	NaN	3204.742425	2594.848313	4001.723505
	6.0	NaN	NaN	NaN	NaN	3413.620000	8552.631579	3035.226001
	7.0	NaN	NaN	NaN	NaN	NaN	NaN	1761.506141
	8.0	NaN	NaN	3879.354324	NaN	NaN	NaN	3089.298479
	9.0	NaN	NaN	NaN	3306.451613	NaN	NaN	NaN
house	0.0	NaN	NaN	NaN	NaN	NaN	NaN	2000.000000
	2.0	NaN	NaN	3315.318302	2840.507857	1975.000000	NaN	2312.552031
	3.0	NaN	NaN	NaN	3000.000000	3160.919540	2325.042407	2423.111862

Encontramos que más superficie y más ambientes no necesariamente significa mayor precio por m2 siempre

superficie por m2 de forma creciente

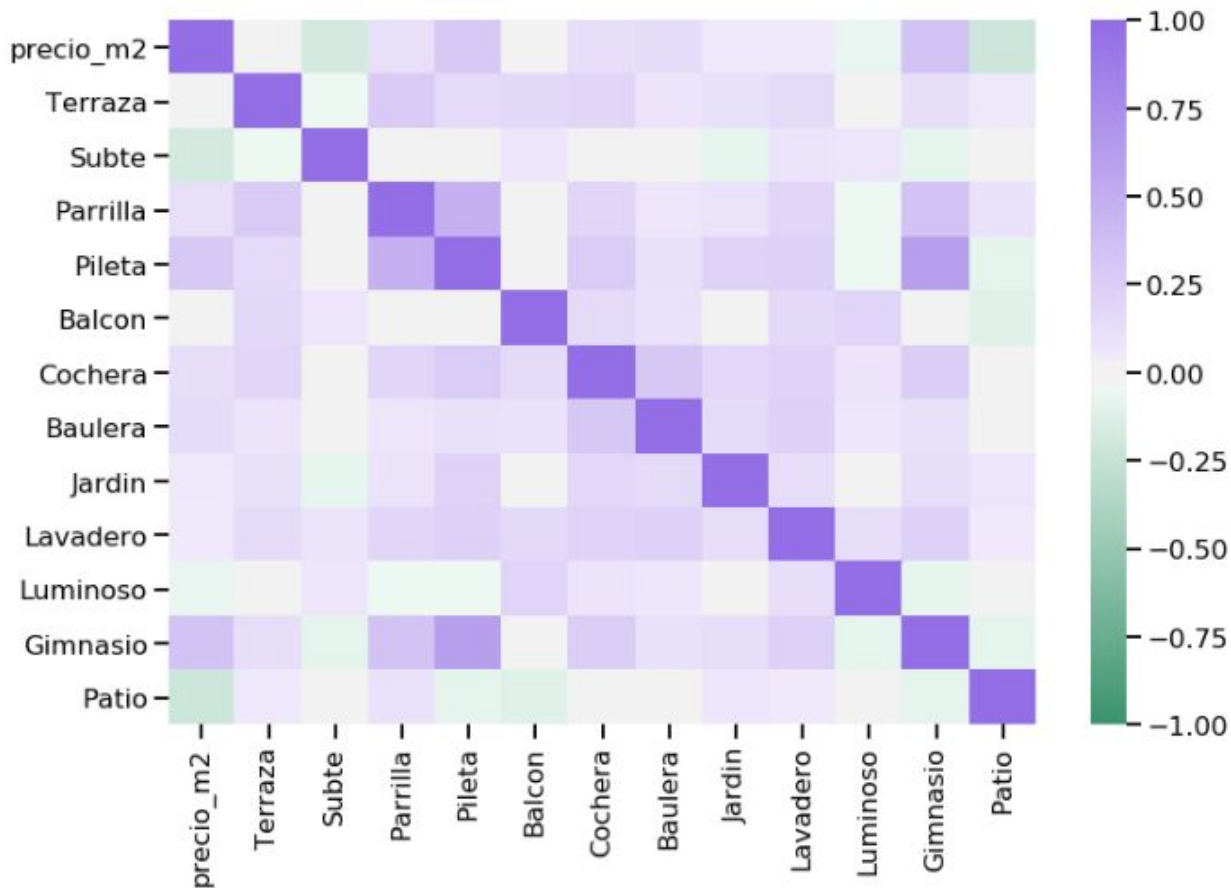


apartment	1.0	3096.644323	3024.831849	2754.608779	1761.997992	2184.803290	4452.208773	NaN
	2.0	2964.922255	3118.226919	3216.321406	3072.221116	2490.420790	2547.061861	2191.247116
	3.0	3309.635618	2792.477782	2834.520925	3073.092167	3273.224162	3305.369036	3298.075667
	4.0	3508.488313	3376.111789	3268.345548	2699.476358	3021.034895	3125.712169	3836.972864
	5.0	NaN	3413.620000	NaN	NaN	3204.742425	2594.848313	4001.723505
	6.0	NaN	NaN	NaN	NaN	3413.620000	8552.631579	3035.226001
	7.0	NaN	NaN	NaN	NaN	NaN	NaN	1761.506141
	8.0	NaN	NaN	3879.354324	NaN	NaN	NaN	3089.298479
	9.0	NaN	NaN	NaN	3306.451613	NaN	NaN	NaN
house	0.0	NaN	NaN	NaN	NaN	NaN	NaN	2000.000000
	2.0	NaN	NaN	3315.318302	2840.507857	1975.000000	NaN	2312.552031
	3.0	NaN	NaN	NaN	3000.000000	3160.919540	2325.042407	2423.111862

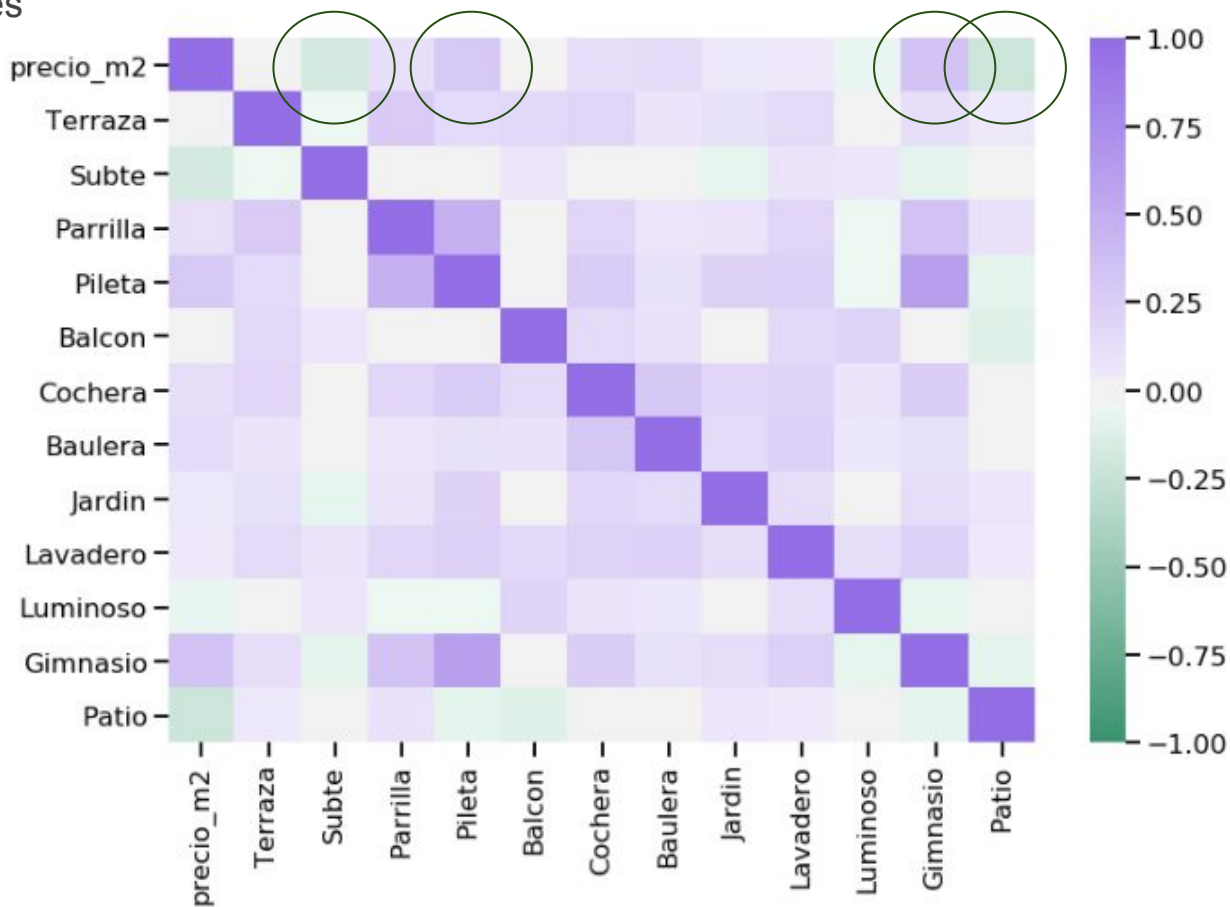
Ambientes



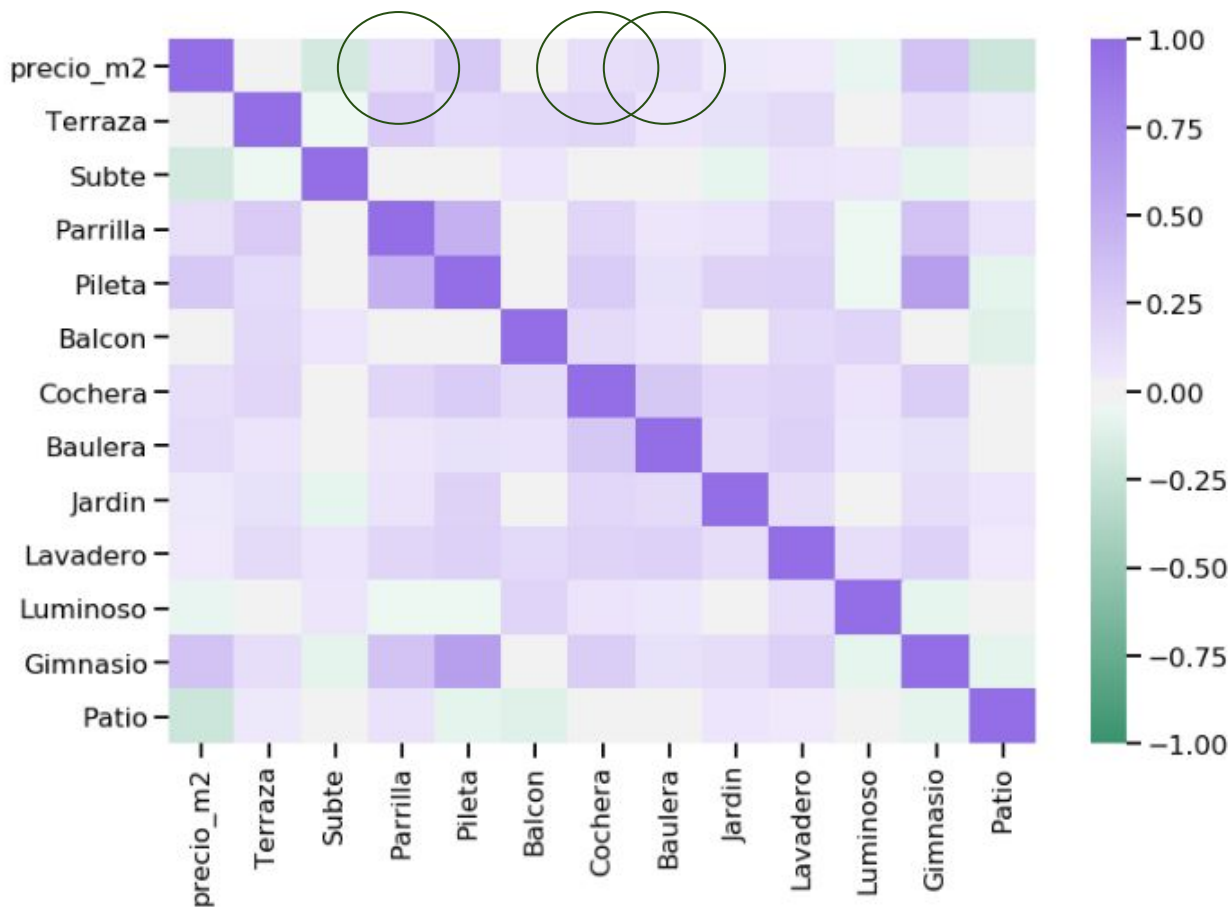
## Correlación precio m2 por palabras clave



Correlación más fuerte entre precio por m2 y palabras clave subte, pileta, gimnasio y patio. Tener en cuenta que esto solo se refiere a la palabra clave SUBTE, no a la distancia real con respecto a las propiedades



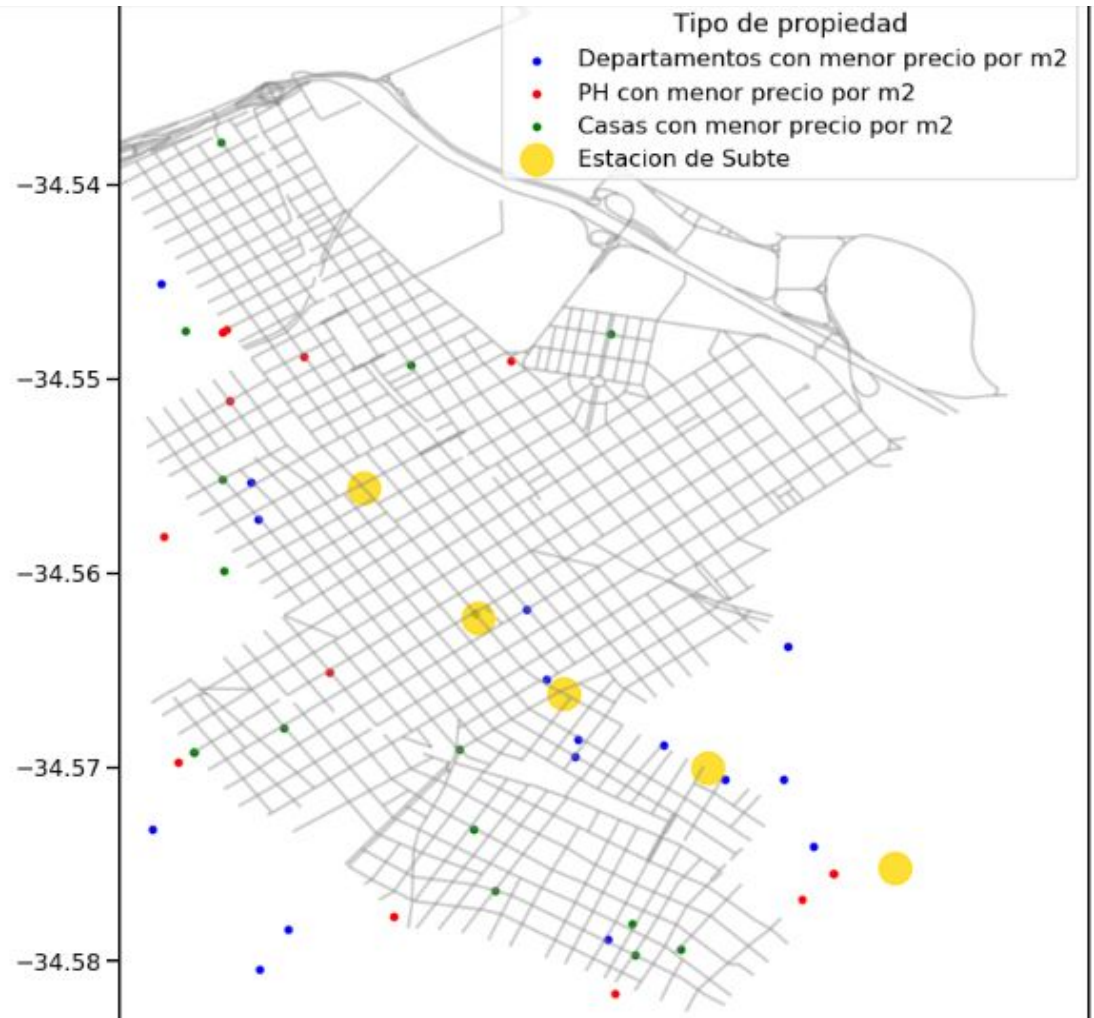
En segundo lugar las otras palabras clave que están conectadas al precio por m2 son Parrilla, Cochera y Baulera



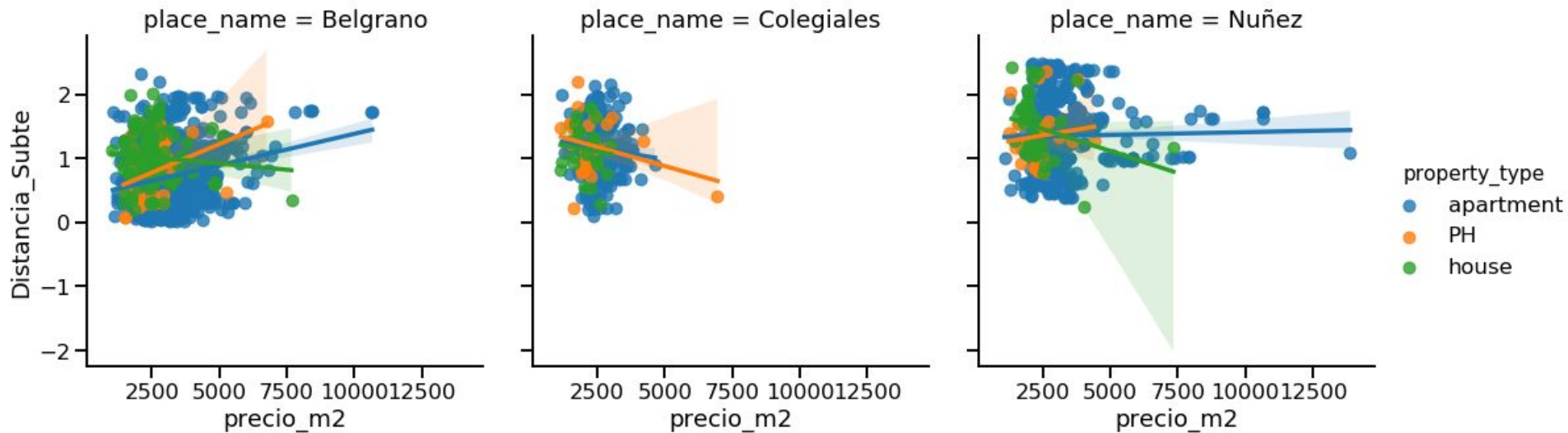
Qué tanto varía el precio por m2 con y sin palabra clave **subte** para cada tipo de propiedad en cada barrio de la Comuna 13

		Subte	False	True
place_name	property_type			
Belgrano	PH		2557.0	2491.0
	apartment		3414.0	3005.0
	house		2486.0	2561.0
Colegiales	PH		2232.0	2227.0
	apartment		2645.0	2648.0
	house		2145.0	2211.0
Nuñez	PH		2417.0	2074.0
	apartment		3418.0	2979.0
	house		2285.0	2478.0

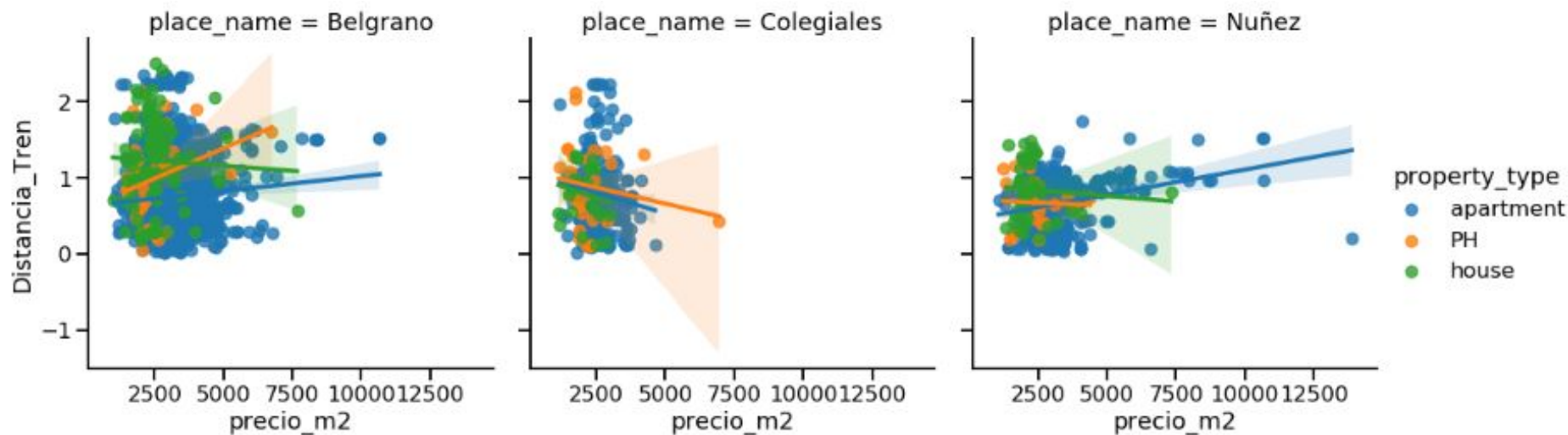
Observamos que sobre la franja de las estaciones de subte de la línea D se encuentran varios departamentos que tienen menor precio por m2



La relación entre distancia de una estación de subte y el precio del m2 no parece tener una influencia significativa cuando se lo analiza por Barrio.

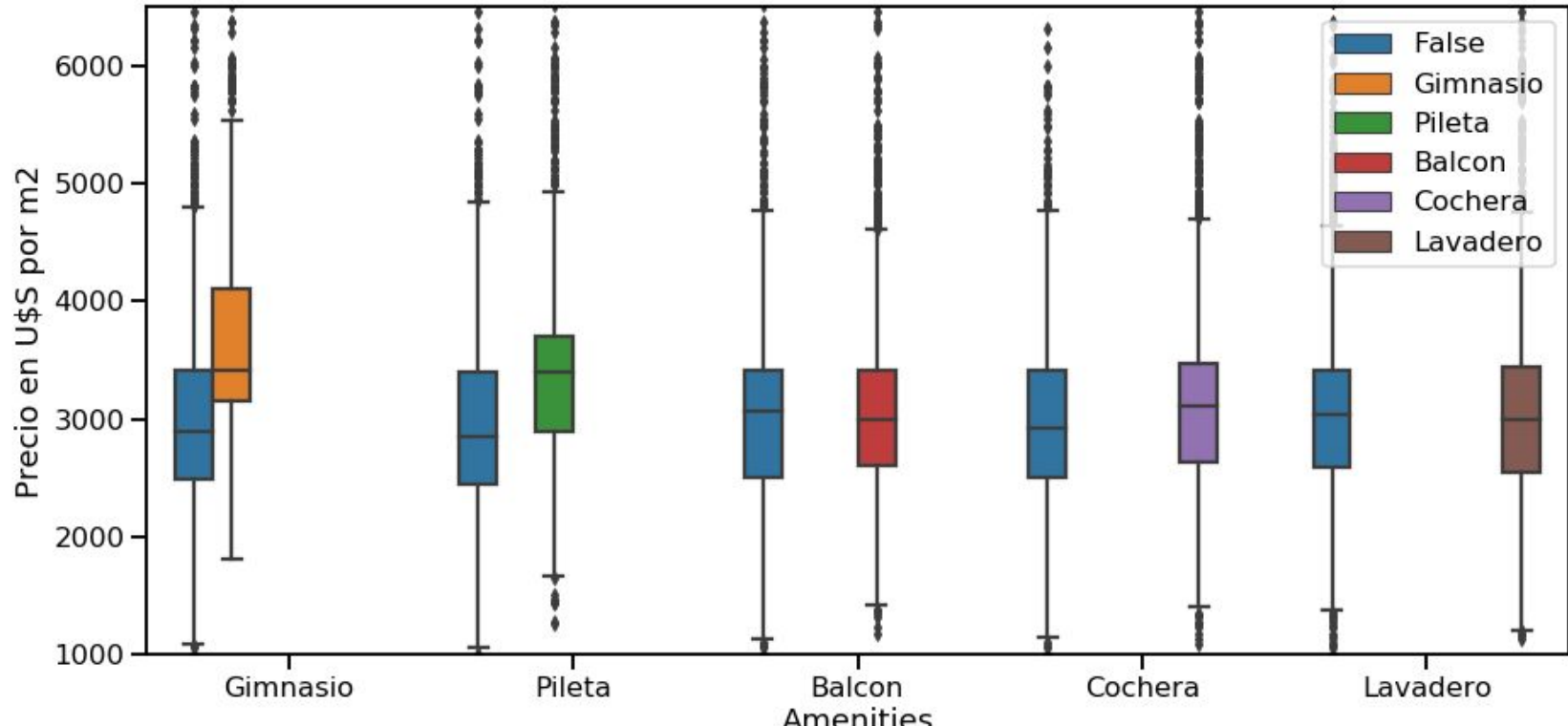


En relación a la cercanía a una estación de Tren, parece tener una influencia positiva en el precio, pero en sentido inverso(más alejada mayor el precio). Esto se puede ver en Nuñez (más significativa) y Belgrano(más débil) para departamentos.



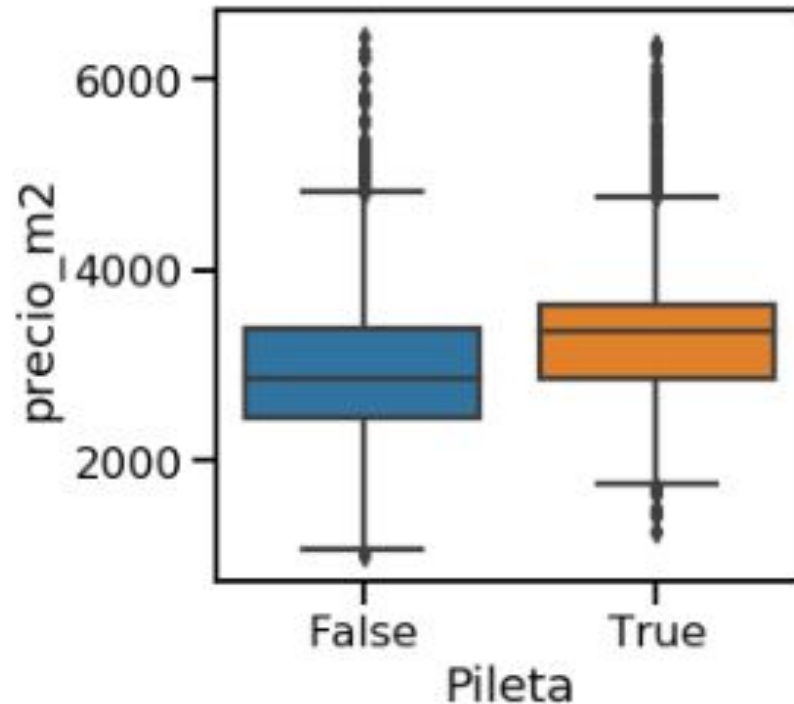
# Variación del precio por m2 con y sin otras palabras clave

Box-plots Amenities

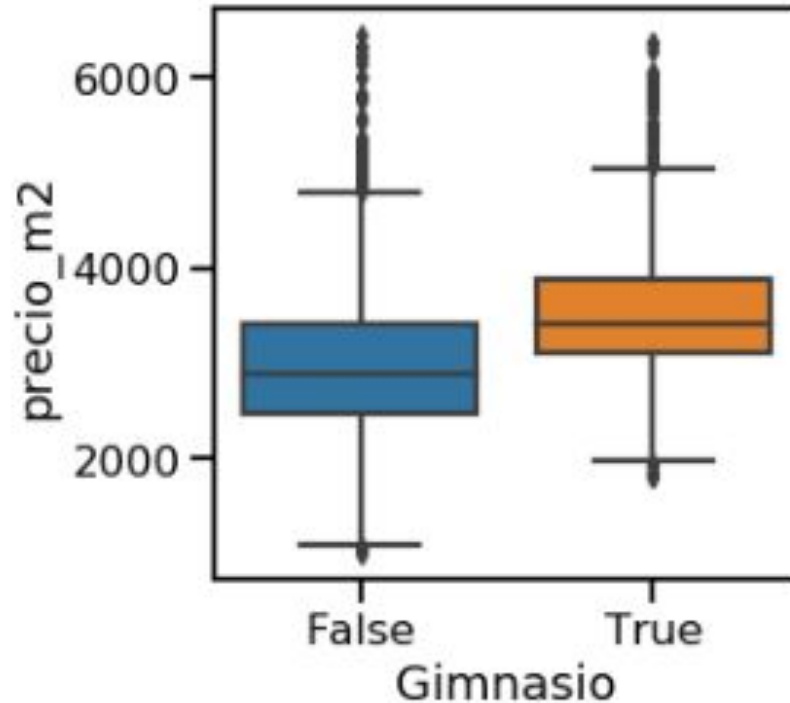




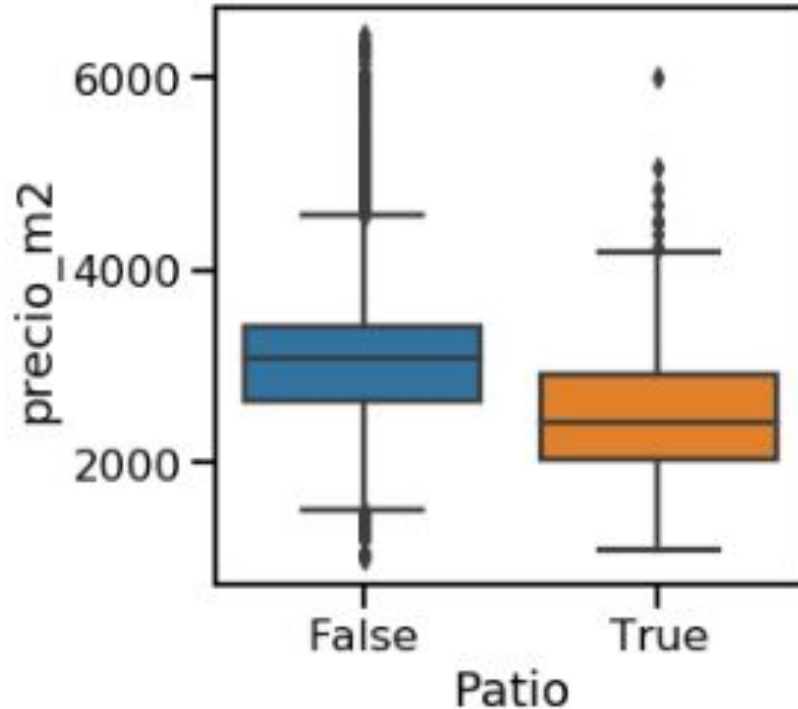
PILETA es otra palabra clave que está relacionada a un precio por m2 más alto



La palabra clave GIMNASIO está relacionada a un precio por m2 más alto



La palabra clave PATIO está relacionada a un precio por m2 más bajo



# Conclusiones

El precio del m<sup>2</sup> en la Comuna 13 parece estar conectado a palabras clave con las que describen a las propiedades. Las que más relación con el precio por m<sup>2</sup> parecen tener distancia al subte/tren, pileta, gimnasio y patio

Las propiedades que tendían más valor por m<sup>2</sup> serían departamentos con amenities que incluyen gimnasio y pileta.

La distancia con una estación de subte se relaciona positivamente, pero no con una gran influencia a nivel Comuna, aunque si para es más fuerte para Belgrano.

En tanto parece que la cercanía a una estación de Tren influye negativamente.

# Conclusiones

Que una propiedad tenga balcón o terraza parece no estar fuertemente relacionado al precio por m<sup>2</sup> con respecto a otras características

Baulera, cochera y parrilla son características conectadas con el precio por m<sup>2</sup> de forma más secundaria que el resto de las palabras clave.

En relación a la distancia con una estación de subte, su influencia parece ser poco relevante. Pero al observar el precio del m<sup>2</sup> con una estación de tren, se observa que el precio disminuye (Nuñez y Belgrano) posiblemente podría ser por la contaminación sonora.

## PARTE 2

Utilizar Machine Learning para elegir un modelo que prediga el precio por m2 de las propiedades según sus características

Luego de hacer un análisis exploratorio, podemos empezar a afinarlo el dataset para armar un modelo que prediga el precio por m2 según tipo de propiedad

-Agregamos más extracción de datos + dropeamos outliers

## Split

```
[114]: feature_cols=data_comuna.columns.drop(['precio_m2', 'log_pm2'])
X=data_comuna.loc[:,feature_cols]
y=data_comuna.precio_m2
X_train,X_test,y_train,y_test=train_test_split(X,y,random_state=42)
```

```
[115]: print(data_comuna.shape)
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
```

```
(3412, 98)
```

```
(2559, 96)
```

```
(2559,)
```

```
(853, 96)
```

-Imputamos todos los NaN que quedaron en train

## Imputacion train de Ambientes

```
[107]: group1 = X_train.groupby(['place_name', 'property_type', "superficie_fill"])[ "Ambientes"].mean()  
group2 = X_train.groupby(['property_type', "superficie_fill"])[ "Ambientes"].mean()  
group3 = X_train.groupby(["superficie_fill"])[ "Ambientes"].mean()  
group9 = X_train.groupby(["property_type"])[ "Ambientes"].mean()  
X_train["Ambientes"] = X_train["Ambientes"].fillna(X_train[['place_name', 'property_type', "superficie_fill"]].apply(tuple,  
X_train["Ambientes"] = X_train["Ambientes"].fillna(X_train[['property_type', "superficie_fill"]].apply(tuple, axis=1).map  
X_train["Ambientes"] = X_train["Ambientes"].fillna(X_train["superficie_fill"].map(group3))
```

```
[108]: X_train["Ambientes"].isnull().sum()
```

```
[108]: 20
```

```
[109]: X_train['Ambientes'].fillna((X_train['Ambientes'].mean()), inplace=True)
```

```
[110]: X_train["Ambientes"].isnull().sum()
```

```
[110]: 0
```



```
[150]: #Escalamos las columnas
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
```

```
[151]: poly=PolynomialFeatures(2,include_bias=False)
X_train=poly.fit_transform(X_train)    #generamos Features polinomicas
```

## Cross-Validation

```
[152]: from sklearn.linear_model import LassoCV, RidgeCV
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LinearRegression
```

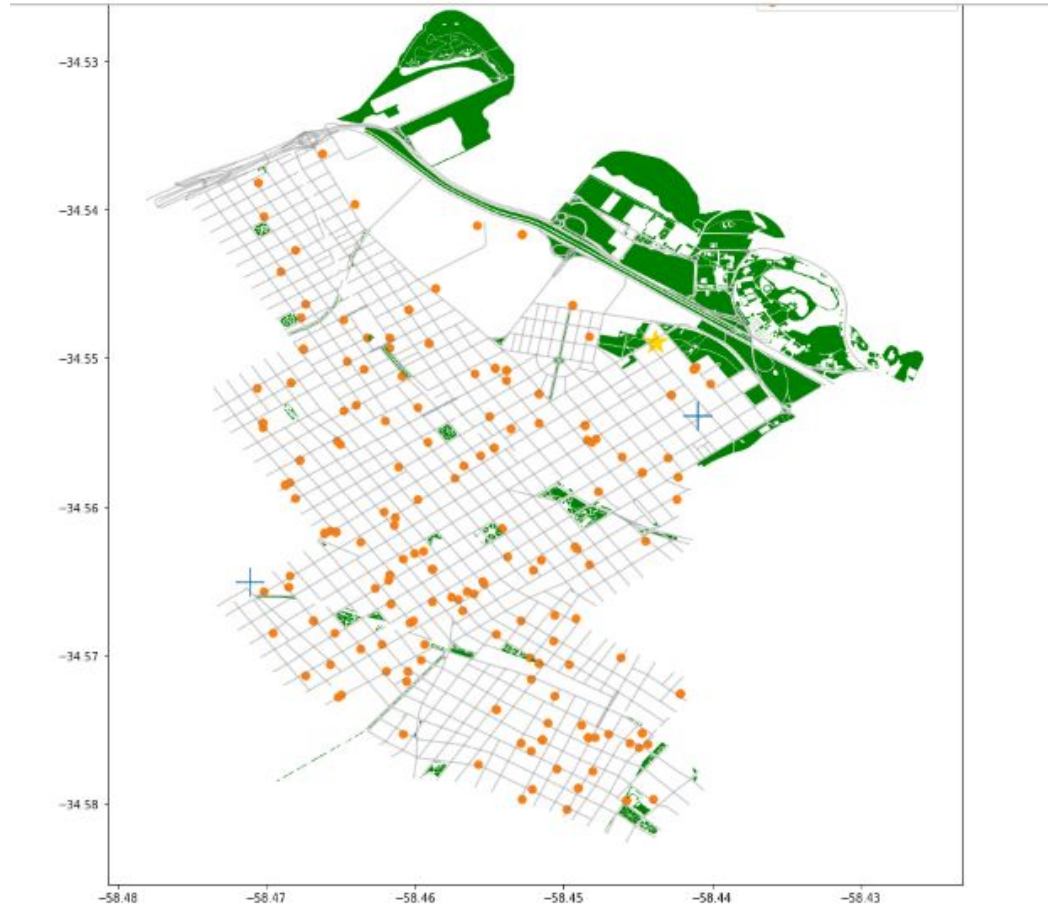
```
[153]: alpha_lasso=np.linspace(8,20,1000)
alpha_ridge=np.linspace(800,3500,1000)
```

```
#alpha_lasso = 13.04
#alpha_ridge = 1970.17
```

```
lasso=LassoCV(alphas=alpha_lasso)
ridge=RidgeCV(alphas=alpha_ridge)
```

```
lasso.fit(X_train,y_train)
ridge.fit(X_train,y_train)
```

Agregamos distancia de  
centros educativos,  
espacios verdes y  
hospitales de cada  
propiedad



# Lasso

```
[155]: from sklearn.linear_model import Lasso  
model = Lasso(alpha=lasso.alpha_, max_iter=2500)  
  
print(cross_val_score(model,X_train,y_train).mean())  
  
0.4847888084562646
```

# Ridge

```
[156]: print(cross_val_score(ridge,X_train,y_train).mean())  
  
0.4965148824694176
```

# Elastic Net

```
[135]: print(cross_val_score(estimator=elastic, X=X_train, y=y_train, cv=5).mean())  
0.4960222133926262
```

-Imputamos NaN en Test

-Usamos Ridge para predecir

```
[178]: from sklearn.metrics import r2_score  
y_pred = ridge.predict(X_test)  
r2_score(y_test, y_pred)
```

```
[178]: 0.5070418387709704
```

---

Propiedades con precio m2 de predicción más alto que precios originales  
(son una oportunidad para invertir)

```
[180]: [<matplotlib.lines.Line2D at 0x17fc69576d0>]
```

