Checkpoint 2 - Grupo 08

Análisis Exploratorio

Nuestro dataset está compuesto por 460.154 publicaciones (filas) y 20 variables (columnas)

Columnas del dataset:

#	Column	Non-Null Count	Dtype
0	id	460154 non-null	object
1	start_date	460154 non-null	object
2	end_date	460154 non-null	object
3	created_on	460154 non-null	object
4	latitud	419740 non-null	float64
5	longitud	419740 non-null	float64
6	place_l2	460154 non-null	object
7	place_l3	437665 non-null	object
8	place_l4	139020 non-null	object
9	place_l5	2430 non-null	object
10	place_l6	0 non-null	float64
11	operation	460154 non-null	object
12	property_type	460154 non-null	object
13	property_rooms	368498 non-null	float64
14	property_bedrooms	344113 non-null	float64
15	property_surface_total	397813 non-null	float64
16	property_surface_covered	427916 non-null	float64
17	property_price	442153 non-null	float64
18	property_currency	441590 non-null	object
19	property_title	460154 non-null	object

Las variables **cuantitativas** observadas en el dataset son:

- `latitud` (continua): Latitud en la que se encuentra la propiedad.
- `longitud` (continua): Longitud en la que se encuentra la propiedad.
- `property_rooms` (discreta): Cantidad de ambientes con los que cuenta de la propiedad.
- `property_bedrooms` (discreta): Cantidad de dormitorios con los que cuenta la propiedad.
- `property_surface_total` (continua): Superficie total que ocupa la propiedad.
- `property_surface_covered` (continua): Superficie de terreno cubierta con que cuenta la propiedad.
- `property_price` (continua): Precio de la propiedad
- `start_date` (discreta): Fecha de alta del aviso.
- `end_date` (discreta): Fecho de baja del aviso.
- `created_on` (discreta): Fecha de alta de la primera versión del aviso.

En cuanto a variables **cualitativas**, se tiene:

- `id` (nominal): ID de la propiedad.
- `operation` (nominal): Tipo de operación (venta, alquiler, etc.)
- `place_l2` (nominal): Nivel de division administrativa 2, correspondiente a la provincia donde se encuentra la propiedad.
- `place_l3` (nominal): Nivel de division administrativa 3, correspondiente a la ciudad donde se encuentra la propiedad.
- `place_l4` (nominal): Nivel de division administrativa 4, correspondiente al barrio donde se encuentra la propiedad.
- `place_l5` (nominal): Nivel de division administrativa 5. No tiene una equivalencia definida por documentación.
- `place_l6` (nominal): Nivel de division administrativa 6. No tiene una equivalencia definida por documentación.
- `property_type` (nominal): Tipo de propiedad (Casa, Departamento, PH)
- `property_currency` (nominal): Moneda correspondiente al precio publicado.
- `property_title` (nominal): Titulo del anuncio.

Al realizar un filtrado inicial del dataset, conservamos únicamente las propiedades publicadas para la venta, en dólares y que sean Departamentos, Casas o PHs en la CABA.

Preprocesamiento de Datos

Procedemos a eliminar las siguientes variables, dado que las mismas carecen de sentido para el estudio que estamos realizando.

- `id` - No tiene ningún sentido conservar esta columna con los códigos con los que se identificaba a las publicaciones.
- `property_currency` - Dado que hemos filtrado el dataset, dejando solo las publicaciones en dólares, carece de sentido conservar esta columna.
- `operation` - Dado que solo dejamos las operaciones de Venta de propiedades, carece de sentido conservar esta variable.
- `place_l2` - Esta variable contiene la Provincia donde está ubicada la publicación, y como solo hemos dejado las publicaciones ubicadas en CABA, carece de sentido conservarla.
- `place_l4` - Esta variable tiene un 96,13% de valores nulos, carece de sentido conservarla.
- `place_l5` - Esta variable tiene un 100% de valores nulos, carece de sentido conservarla.
- `place_l6` - Esta variable tiene un 100% de valores nulos, carece de sentido conservarla.
- `property_title` - No vemos ninguna utilidad práctica en conservar el título con el que fue publicada la vivienda.
- `created_on` - Esta variable contiene la fecha de alta de la primera versión del aviso, pero al revisar se vio que tiene exactamente los mismos valores que la variable `start_date`, con lo cual se la elimina.

Variables que se conservaron

- `latitud` (continúa): Latitud en la que se encuentra la propiedad.
- `longitud` (continúa): Longitud en la que se encuentra la propiedad.
- `property_rooms` (discreta): Cantidad de ambientes con los que cuenta de la propiedad.
- `property_bedrooms` (discreta): Cantidad de dormitorios con los que cuenta la propiedad.
- `property_surface_total` (continua): Superficie total que ocupa la propiedad.
- `property_surface_covered` (continua): Superficie de terreno cubierta con que cuenta la propiedad.

- `property_price` (continua): Precio de la propiedad
- `start_date` (discreta): Fecha de alta del aviso.
- `end_date` (discreta): Fecha de baja del aviso.
- `place_l3` (nominal): Nivel de división administrativa 3, correspondiente al barrio donde se encuentra la propiedad.
- `property_type` (nominal): Tipo de propiedad (Casa, Departamento, PH)

Se detectaron las siguientes correlaciones entre variables al realizar un heatmap con las variables cuantitativas.

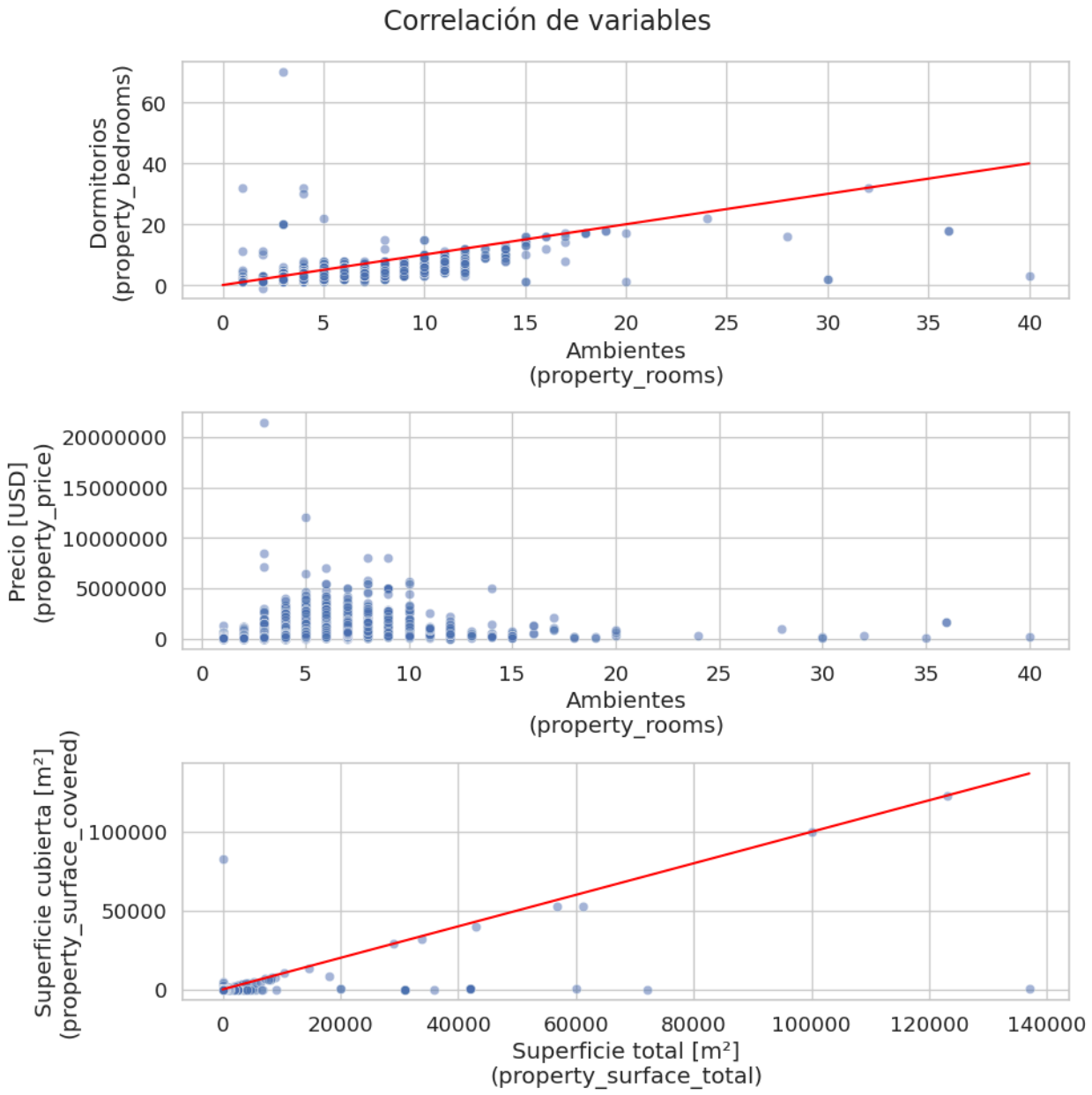
- `property_rooms` y `property_bedrooms (0.85)`
  Podemos observar una correlación positiva, lo cual por supuesto es algo esperable, dado que la cantidad de dormitorios de un inmueble siempre estará acotado superiormente por la cantidad de ambientes del mismo.

- `property_surface_covered` y `property_surface_total (0.6)`
  Al igual que en el caso anterior, observamos una correlación positiva esperable, dado que la superficie total cubierta siempre estará acotada superiormente por la superficie total del inmueble.

- `property_rooms` y `property_price (0.49)`
  Finalmente, observamos una correlación positiva entre la cantidad de ambientes y el precio de la propiedad. Lo cual, era algo esperable.

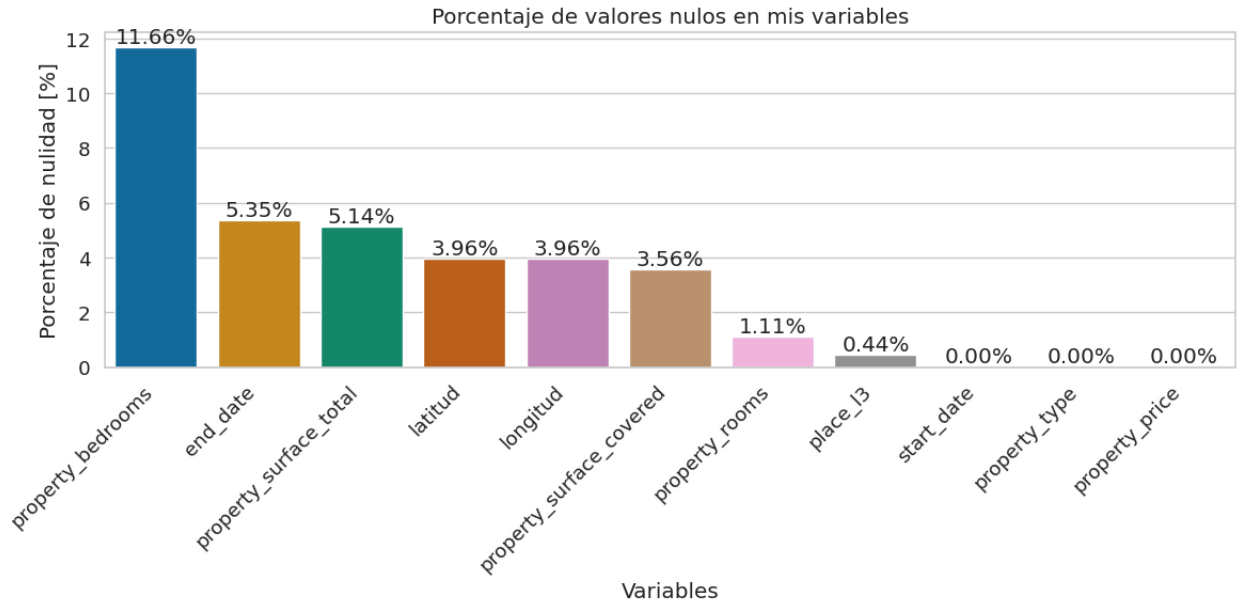
Además, analizamos estas variables mediante gráficas de dispersión, lo cual nos llevó rápidamente a observar que las mismas tenían varios valores atípicos.

- Hay propiedades con más habitaciones que ambientes.
- Hay propiedades con más superficie cubierta que superficie total
- Hay propiedades con un enorme número de ambientes.
- Hay propiedades con una superficie desmesurada.
- Hay propiedades que por su cantidad de ambientes parecen tener un precio desorbitado.



Cómo se manejaron estos valores atípicos se explica más adelante.

2.3.3. Análisis de valores faltantes



1) Se eliminaron las filas que tuvieran valores nulos en las variables Latitud o Longitud. Esto es debido a que posteriormente necesitaremos la ubicación precisa de cada propiedad.
2) Se eliminó la variable end_date dado que a esta altura vimos que carecía de utilidad conservarla.
3) Se eliminaron las propiedades cuya latitud y longitud estuviera fuera de los límites de CABA.
4) Se obtuvo el valor de los barrios que faltaban (variable place_l3) usando los valores de latitud y longitud.

2.3.4. Imputación de datos

1) Imputamos los valores nulos de las columnas `property_bedrooms` (dormitorios) y `property_rooms` (ambientes) mediante un modelo de regresión lineal utilizando IterativeImputer.

   Al modelo le pasamos, además de las columnas mencionadas arriba, las columnas `latitud`, `longitud` y `property_type` para que este pueda utilizarlas como referencia, y así aprovechar mejor la información disponible para imputar los valores faltantes de manera más precisa.

2) Imputamos los valores nulos de las columnas `property_surface_total` (Superficie total) y `property_surface_covered` (Superficie total cubierta) mediante un modelo de regresión lineal utilizando IterativeImputer.

   Al modelo le pasamos, además de las columnas mencionadas arriba, las columnas `latitud`, `longitud`, `property_type` y `property_rooms` para que este pueda utilizarlas como referencia, y así aprovechar mejor la información disponible para imputar los valores faltantes de manera más precisa.

Una vez hecho esto, nos hemos quedado sin valores faltantes.

2.4. Valores atípicos

2.4.1. Analisis Univariado

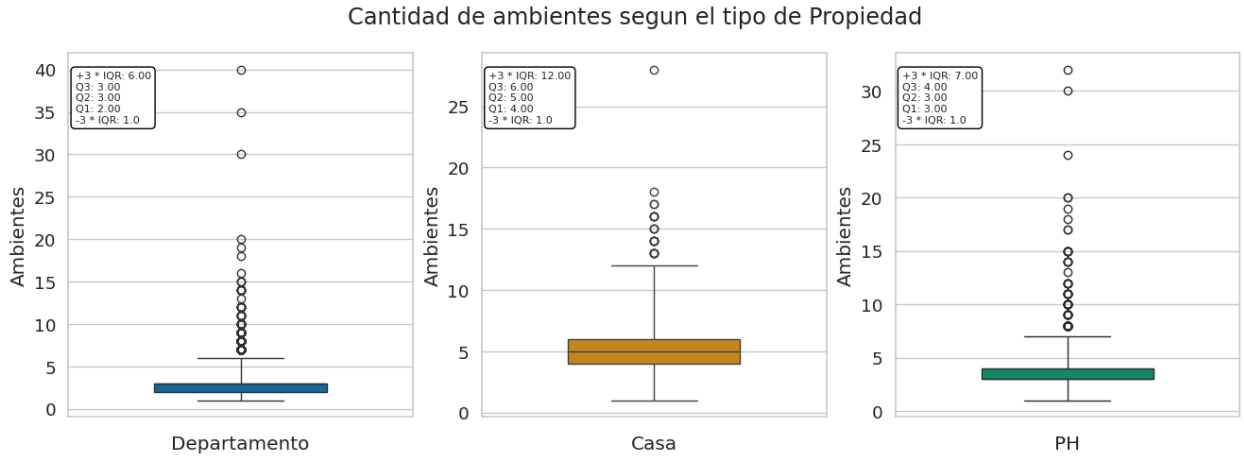
Para la detección de valores atípicos en el dataset se recurrió primero a un análisis univariado de las columnas `property_rooms`, `property_bedrooms`, `property_surface_total`, `property_surface_covered` y `property_price` mediante el uso de boxplots. En este análisis vamos a hacer una distinción según el tipo de propiedad estudiada (Depto, Casa, PH)

Criterios elegidos

En el análisis de los boxplots, consideraremos como outliers a aquellas observaciones que se encuentren a una distancia de 3 veces el rango intercuartílico, en lugar de 1,5 veces (como viene por defecto).

Tomamos esta decisión debido a que observamos que al usar el valor por defecto, perdíamos gran parte de las observaciones que se encontraban dentro de rangos de valores relativamente "comunes" en lo que a propiedades se refiere. De esta forma nos aseguramos que solo se consideran outliers a aquellas observaciones que se alejan en demasía de la media.

2.4.1.1. Variable `property_rooms` (ambientes)



En base a las gráficas arriba desplegada, se obtuvieron las siguientes observaciones:

- En las propiedades de tipo **Departamento** vemos que la gran mayoría de los outliers se ubican por encima de los 6 y hasta los 20 ambientes. Además, vemos que tenemos tres excepciones con 30, 35 y 40 ambientes respectivamente.

- Luego, en las propiedades de tipo **Casa** podemos observar que la gran mayoría de los outliers se ubican por encima de los 12 y por debajo de los 20 ambientes. Ademas, tenemos una propiedad en particular por encima de los 25 ambientes.

- Finalmente, en las propiedades de tipo **PH**, vemos como la mayoría de los outliers oscilan por encima de los 7 y hasta los 15 ambientes. Luego, hay un grupo menor que oscila por encima de los 15 y los 20 ambientes. Y finalmente, algunos casos esporádicos por encima de los 20 ambientes.

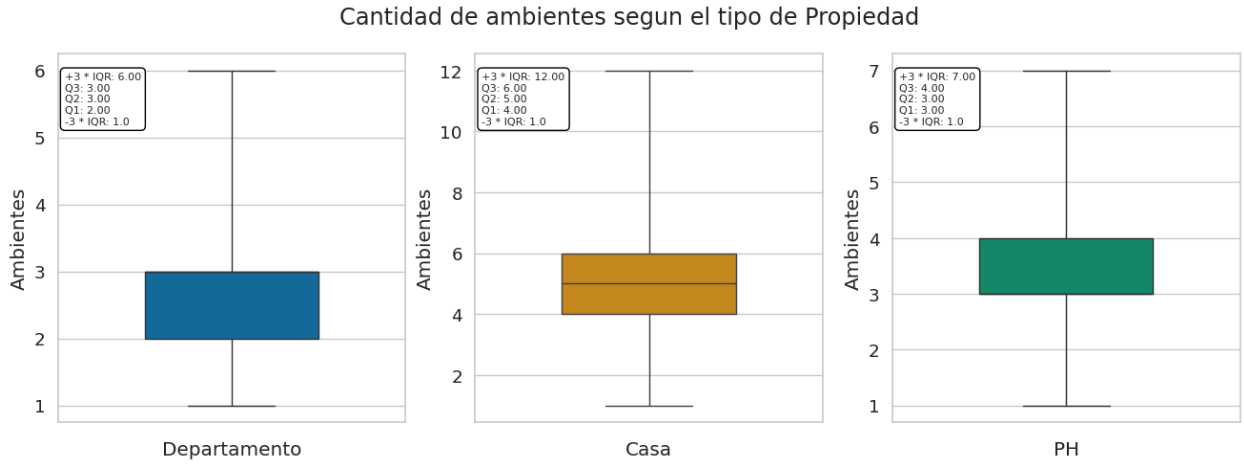
Conclusión:

Concluimos que las propiedades que cuentan con una cantidad de ambientes por encima del valor de su bigote superior es debido a que, o bien ocurrió un error durante la carga de los datos, o bien, se trata de propiedades cuyas características exceden en demasía lo que se consideraría una cantidad de ambientes "normal" para una vivienda.

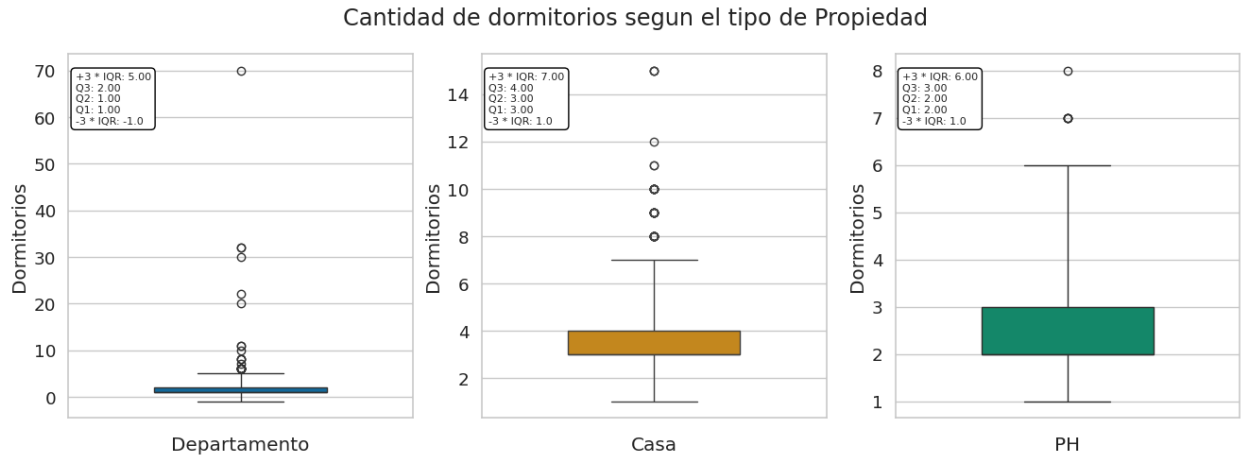
Dado que, consideramos que los valores que estén por encima de esto no son representativos del mercado que se busca estudiar, se decidió eliminar los siguientes registros:

- Departamentos con más de 6 ambientes.
- Casas con más de 12 ambientes.
- PHs con más de 7 ambientes.

Gráfica luego del filtrado de outliers



2.4.1.2. Variable `property_bedrooms` (dormitorios)



Dado que en la sección anterior restringimos el número de ambientes que podían tener los distintos tipos de propiedades, es obvio que estamos ante la presencia de outliers en los tres casos. Pues es imposible que existan viviendas con un número de dormitorios mayor a su número de ambientes.

Además, vemos que hay un caso que llama poderosamente la atención: En los departamentos parece haber propiedades con una cantidad de dormitorios **negativa**.

Al hacer un análisis exhaustivo de este caso, vimos que claramente se trataba de un dato mal cargado, pues era una vivienda de 2 ambientes que decía tener -1 dormitorios. Se corrigió el valor mal cargado.

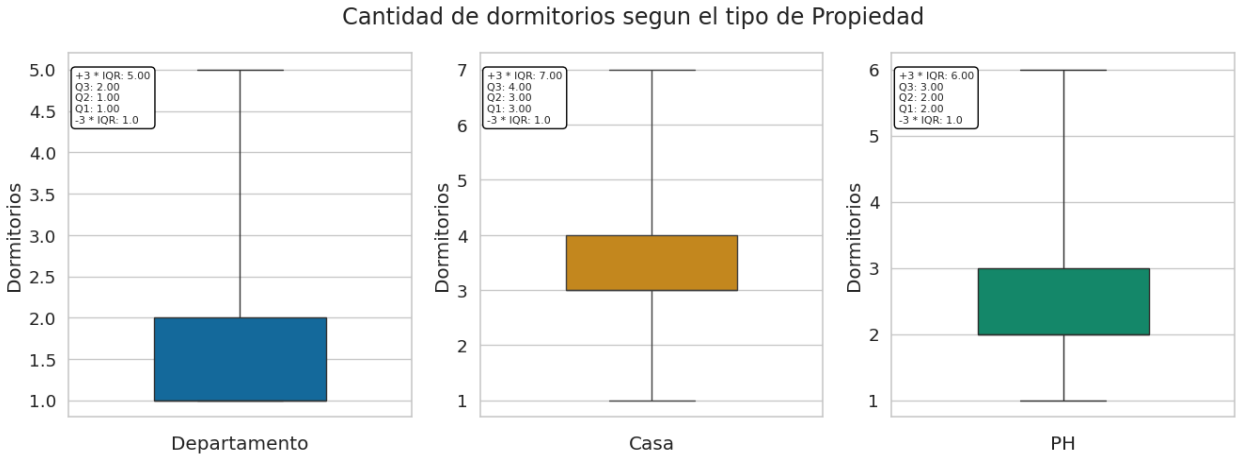
Resuelto ese caso en particular, tenemos que analizar lo siguiente:

En este estudio mediante boxplots puede haber outliers que queden solapados. Por ejemplo: Una casa de 2 ambientes que tenga 3 dormitorios no saldrá como un outlier en el boxplot (esto lo resolveremos más adelante cuando hagamos el análisis multivariado).

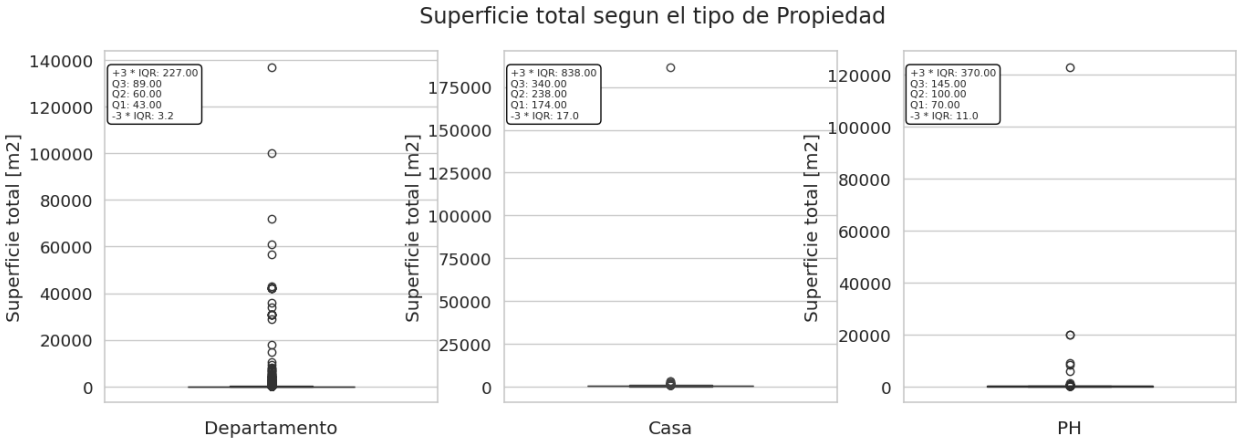
De momento, para poner un límite superior al número de dormitorios, procederemos a eliminar las siguientes propiedades:

- Departamentos que cuenten con más de 5 dormitorios.
- Casas que cuenten con más de 7 dormitorios.
- PHs que cuenten con más de 6 dormitorios.

Gráfica luego del filtrado



2.4.1.3. Variable `property_surface_total` (superficie total)

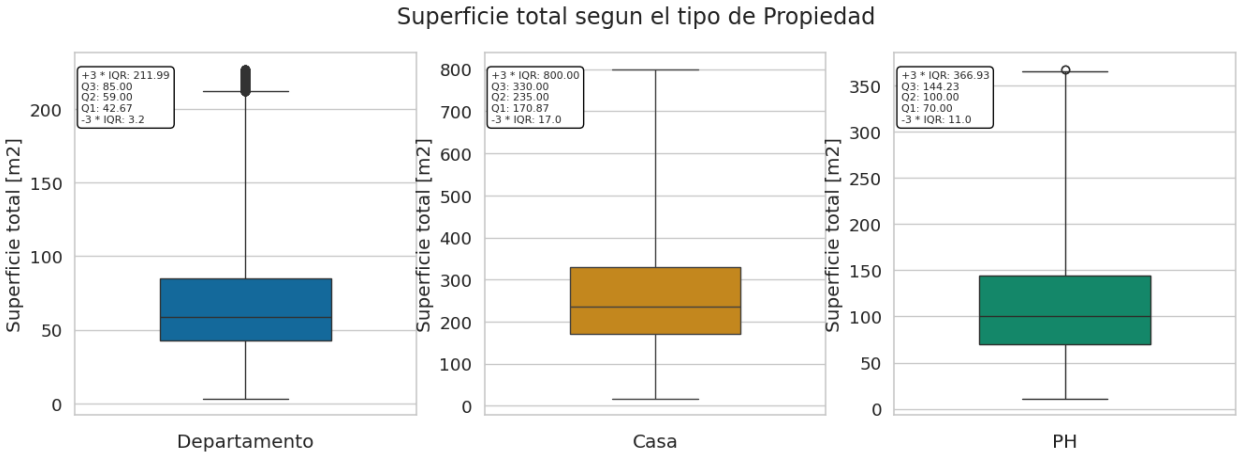


Como podemos observar en las gráficas, tenemos un gran número de observaciones cuya superficie no se condice con lo que uno esperaría encontrar en la realidad. Sobre todo en el caso de los departamentos, donde hay observaciones con una superficie de 100.000 m2 (aproximadamente 14 canchas de fútbol)

De momento estableceremos como cota superior para la superficie de nuestras propiedades el valor de sus bigotes superiores.

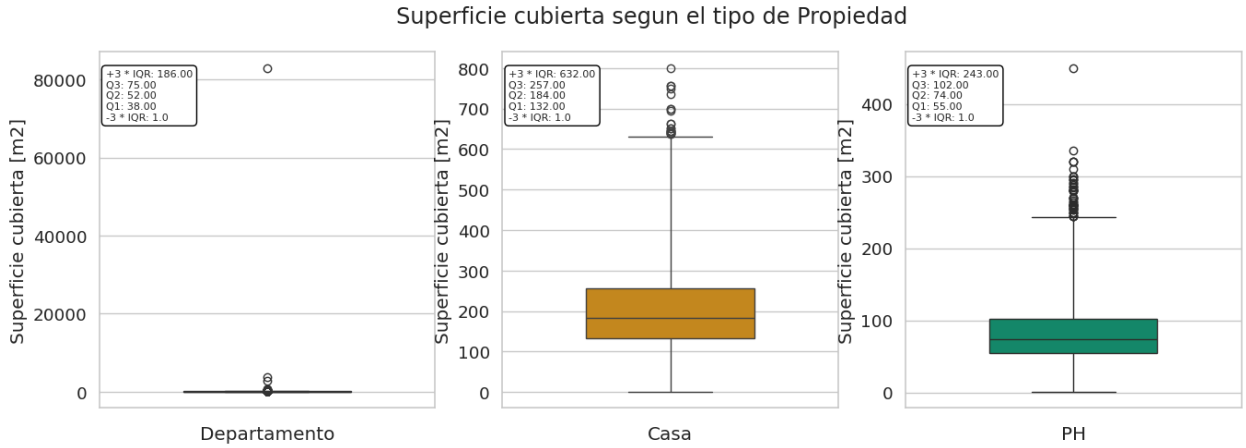
- Departamentos ▸ Máximo 227 m2
- Casas ▸ Máximo 834 m2
- PHs ▸ Máximo 370 m2

Gráfica luego del filtrado



Más adelante, en el análisis multivariado, haremos un hilado más fino. Dado que puede haber outliers que queden solapados en este análisis mediante boxplots (por ejemplo, un monoambiente con 200 m2)

2.4.1.4. Variable `property_surface_covered` (superficie cubierta)

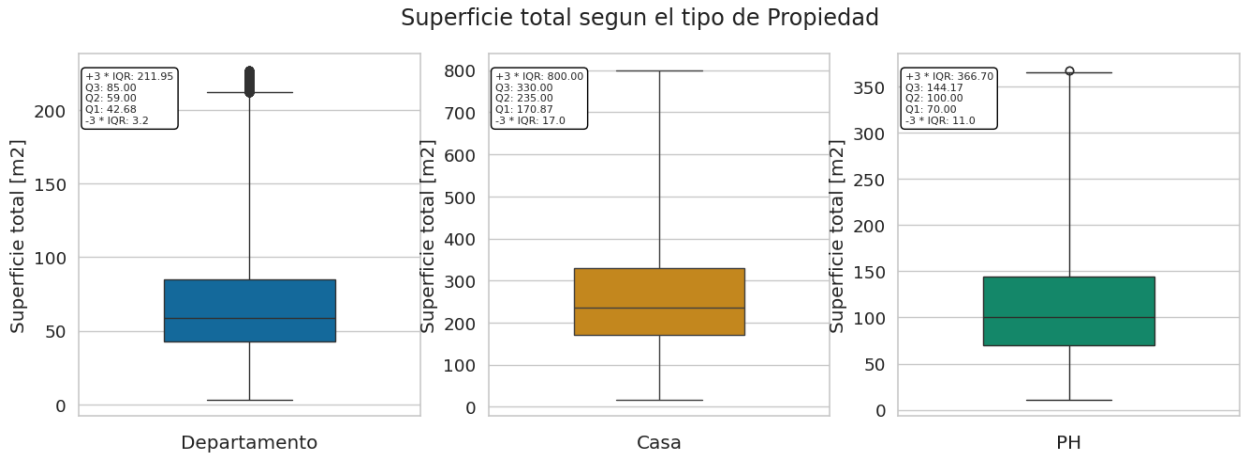


Como podemos observar en las gráficas, ocurre algo muy similar a lo que veíamos con la variable `property_surface_total`.

De momento estableceremos como cota superior para la superficie cubierta los mismos valores que habíamos asignado a la variable `property_surface_total`. Dado que me está dando límites muy bajos el boxplot y no quiero perder tanta data de golpe.

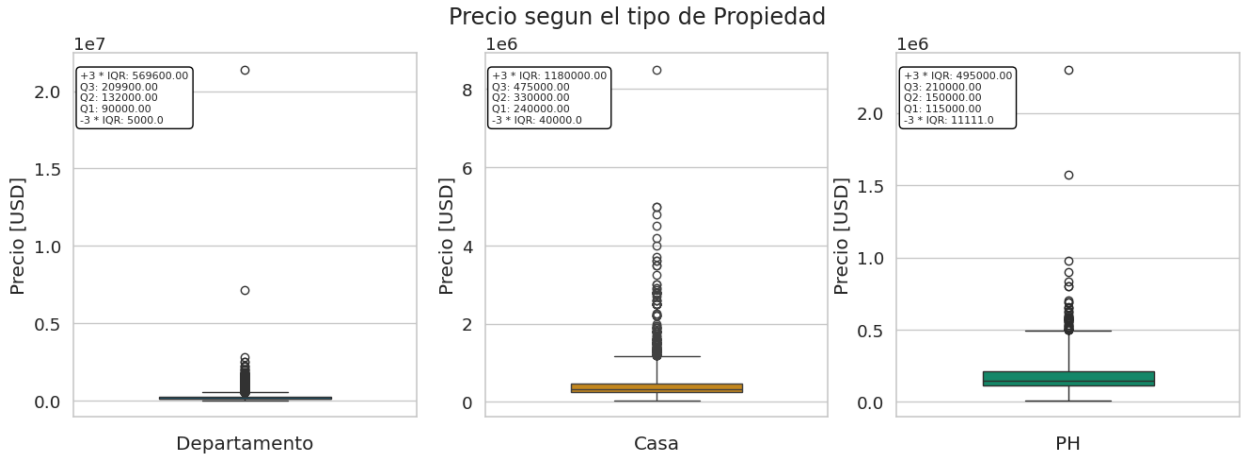
- Departamentos ▸ Máximo 246 m2
- Casas ▸ Maximo 841 m2
- PHs ▸ Maximo 367 m2

Gráfica luego del filtrado



Por ahora lo vamos a dejar así y vamos a hacer un ajuste más fino en análisis multivariado.

2.4.1.5. Variable `property_price` (Precio)



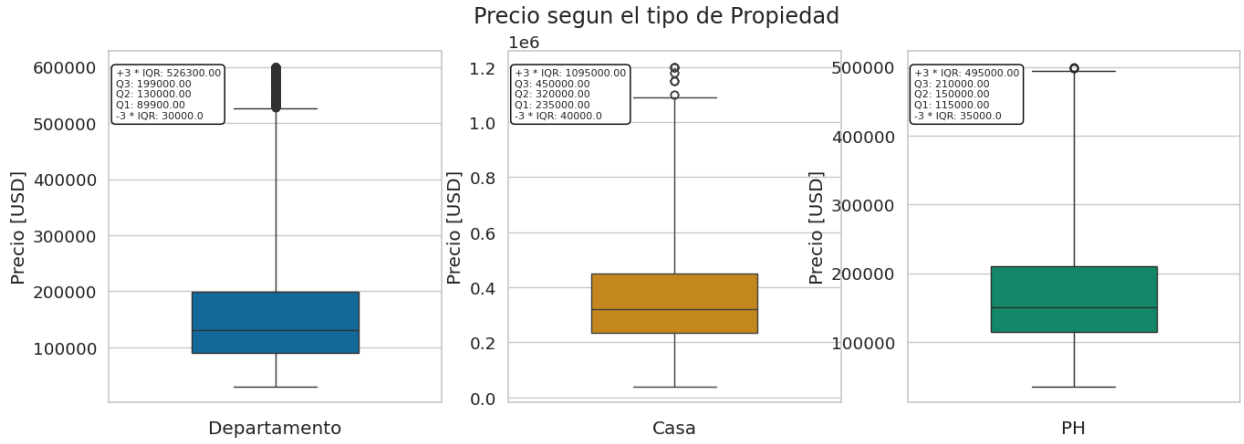
En esta variable es en donde, probablemente, sea más inutil hacer un análisis univariado. Dado que el precio de una propiedad varía sustancialmente en base a parámetros como la ubicación, la superficie y la cantidad de ambientes de la misma.

En primera instancia solo vamos a conservar:

* Departamentos ▸ Precio por debajo de 600 K
* Casas ▸ Precio por debajo de 1,2 M
* PHs ▸ Precio por debajo de 500 K

Además, vamos a poner un límite inferior de 30 K a todas las propiedades, pues consideramos que es ilógico que haya propiedades en venta por un precio tan bajo.

Gráfica luego del filtrado



2.4.2. Análisis Multivariado

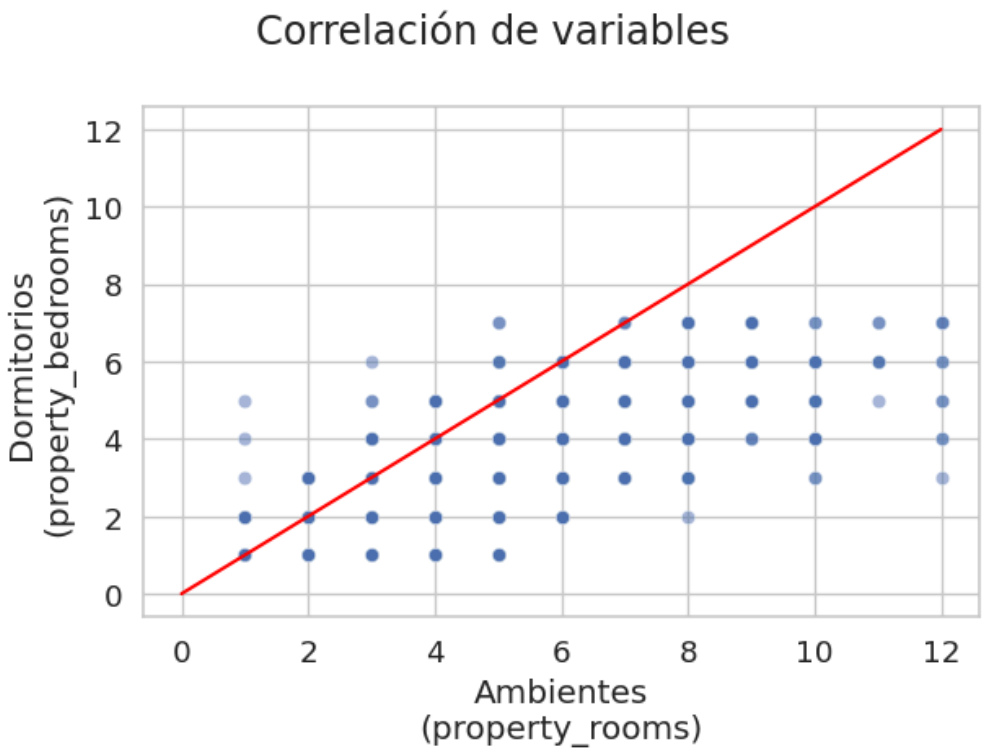
2.4.2.1. `property_rooms` (ambientes) vs `property_bedrooms` (dormitorios)

Dado que, como dijimos en la sección 2.4.1.2, pueden haber quedado outliers solapados en mis datos. Por ejemplo: Una casa de 2 ambientes que tenga 3 dormitorios. Haremos lo siguiente:

- Las propiedades de 1 ambiente, deberán tener 1 dormitorio.
- Las propiedades de n ambientes, podrán tener como máximo n-1 dormitorios (n > 1)

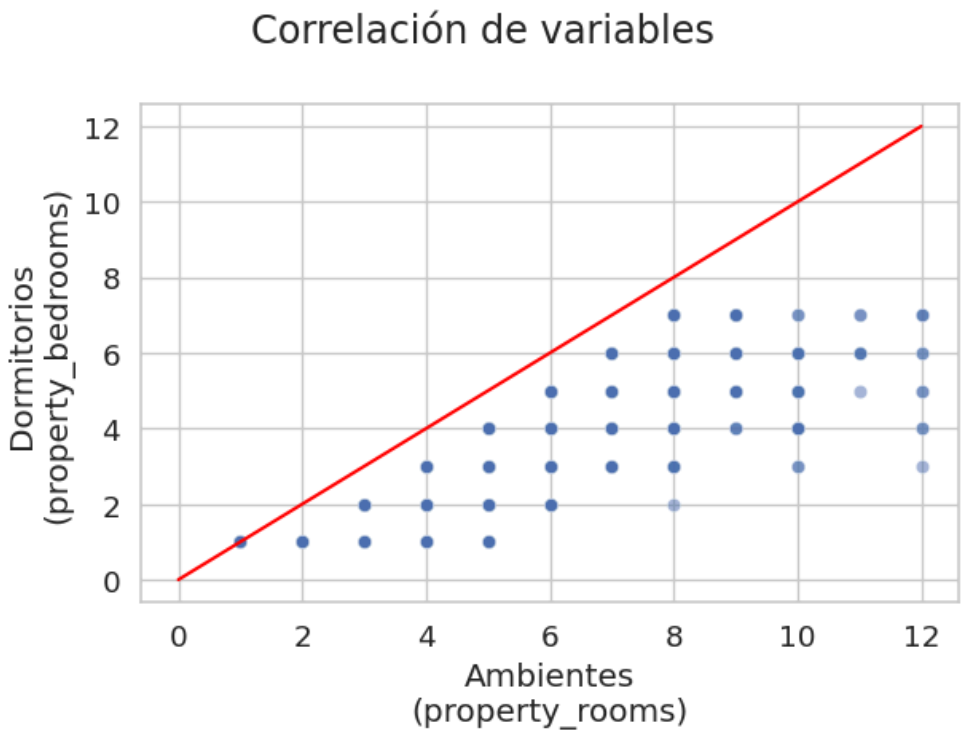
Considerando que, si bien el dataframe indica `property_bedrooms` (es decir, cantidad de dormitorios) en la venta de propiedades lo que se suele relevar es el número de **habitaciones** con que cuenta la propiedad sobre el número de ambientes.

Es decir:
- En una casa de 2 ambientes, tiene 1 habitación.
- En una casa de 3 ambientes, tiene 2 habitaciones.
- En una casa de 4 ambientes, tiene 3 habitaciones.
- etc.



Como podemos observar, hay propiedades en las cuales tenemos más dormitorios que ambientes. Vamos a corregir eso.

Gráfica luego del filtrado



2.4.2.2. `property_rooms` (ambientes) vs `property_surface_total` (superficie)

En la sección 2.4.1.3 al realizar el análisis univariado de la variable property_surface_total lo que hicimos fue establecer una cota superior para la superficie máxima que podían tener los distintos tipos de propiedades.

Esto en el caso de las Casas y los PHs no es tan problemático, pues una casa puede tener tranquilamente 2 ambientes, pero 100 m2 de patio.

El problema de este filtrado surge con los Departamentos donde se puede dar, por ejemplo, el caso de tener un monoambiente que con 200 m2 de superficie.

Para resolver esto recurriremos a un dataset del gobierno de la CABA, en el cual hay un relevamiento de la superficie promedio (en metros cuadrados) en función de la cantidad de ambientes de departamentos publicados entre 2010 y 2024.

Dado que este dataset solo contiene data de deptos de 1 a 3 ambientes, tendremos que estimar la superficie de los deptos con un número de ambientes mayor por otro medio.

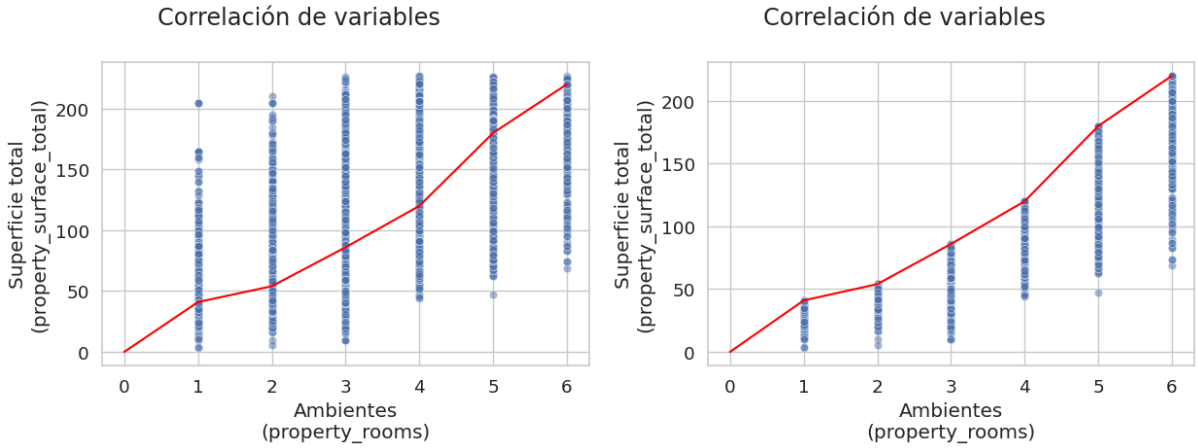
Como regla práctica, cada tipo de Departamento podrá tener como máximo un 25% de superficie por encima del valor que estimo nuestro modelo. (redondeando valores para mayor practicidad)

Para departamentos de 4 a 6 ambientes decidimos tomar los siguientes valores máximos:

* Para 4 ambientes, la superficie máxima será 120 m2
* Para 5 ambientes, la superficie máxima será 180 m2
* Para 6 ambientes, la superficie máxima será 220 m2

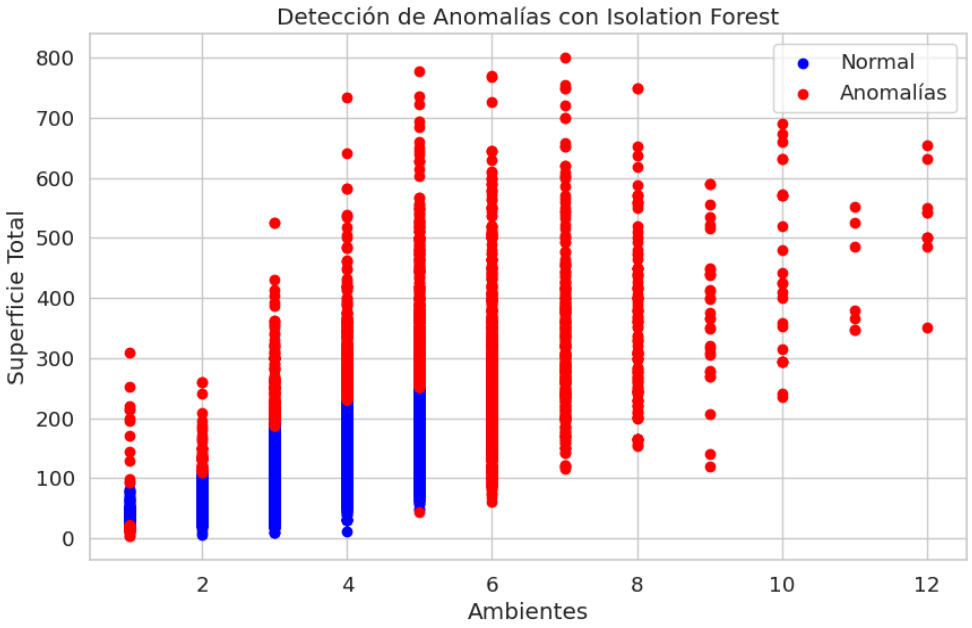
Estos valores se estimaron en base a observaciones realizadas sobre las viviendas publicadas en los sitios Zonaprop y Argenprop.

Gráfica antes de filtrar y luego de filtrar



La línea roja es el límite de superficie que fijamos en función del número de ambientes de los departamentos.
Si bien esto ayudó a corregir los máximos, aún falta purgar bastante los mínimos. Pues como podemos ver, tengo deptos de 6 ambientes con una superficie de 50 m2

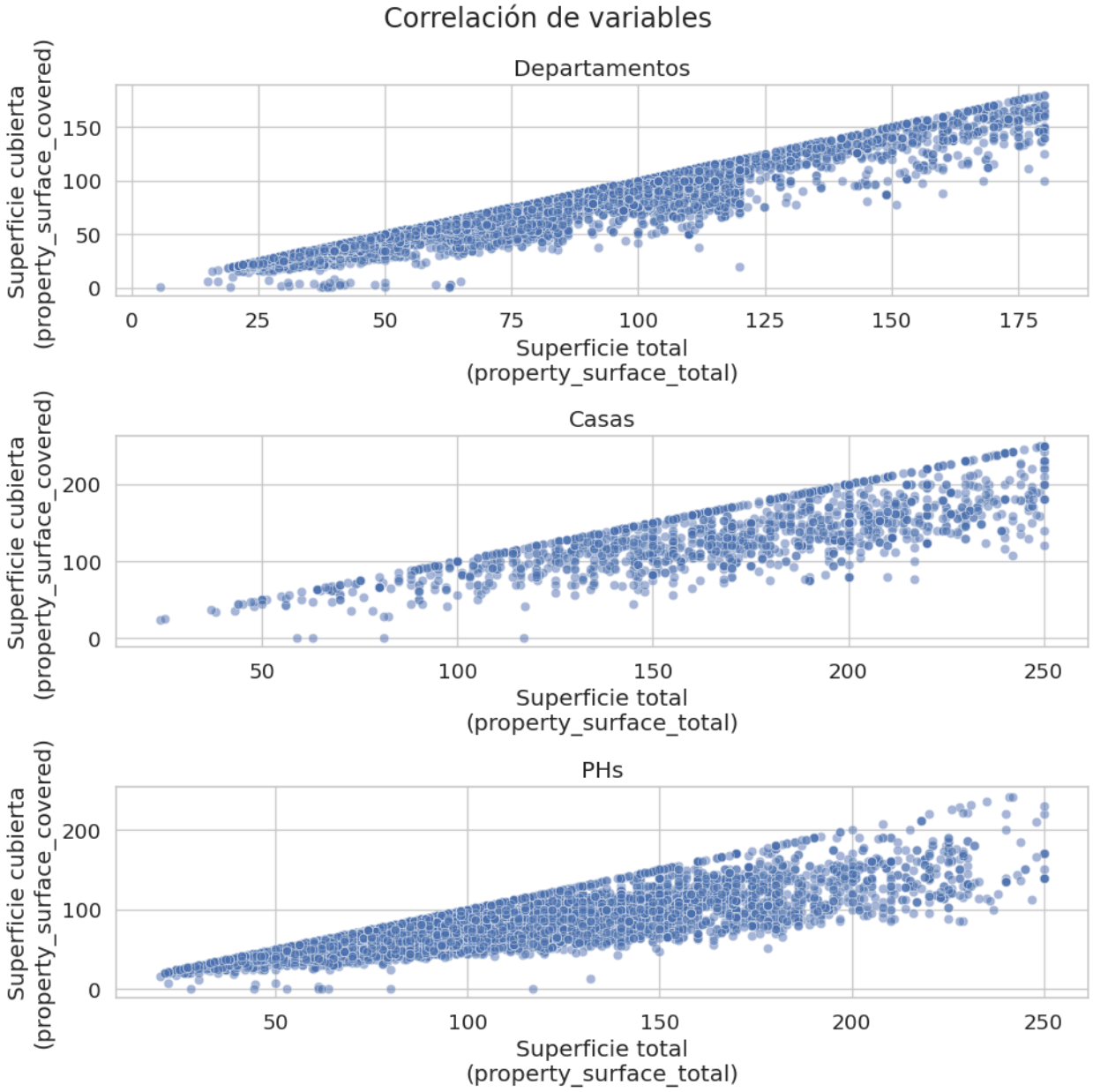
Para un filtrado más fino recurrimos a un Insolation Forest



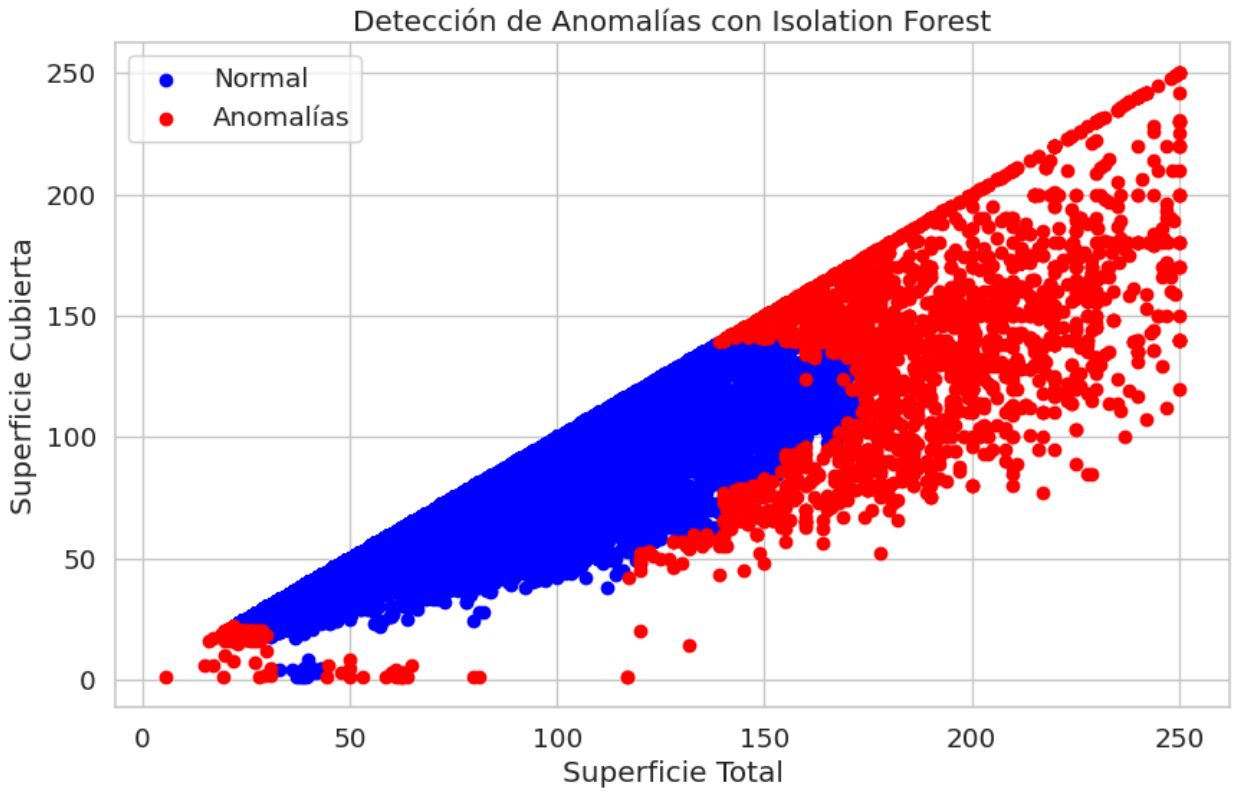
2.4.2.3. `property_surface_total` (superficie total) vs `property_surface_covered` (superficie cubierta)

Primero que nada graficamos una variable versus la otra y eliminamos todas las propiedades cuya superficie cubierta sea mayor a su superficie total.

Gráfica luego del filtrado inicial.



Luego, para un filtrado más fino, recurrimos a un Insolation Forest



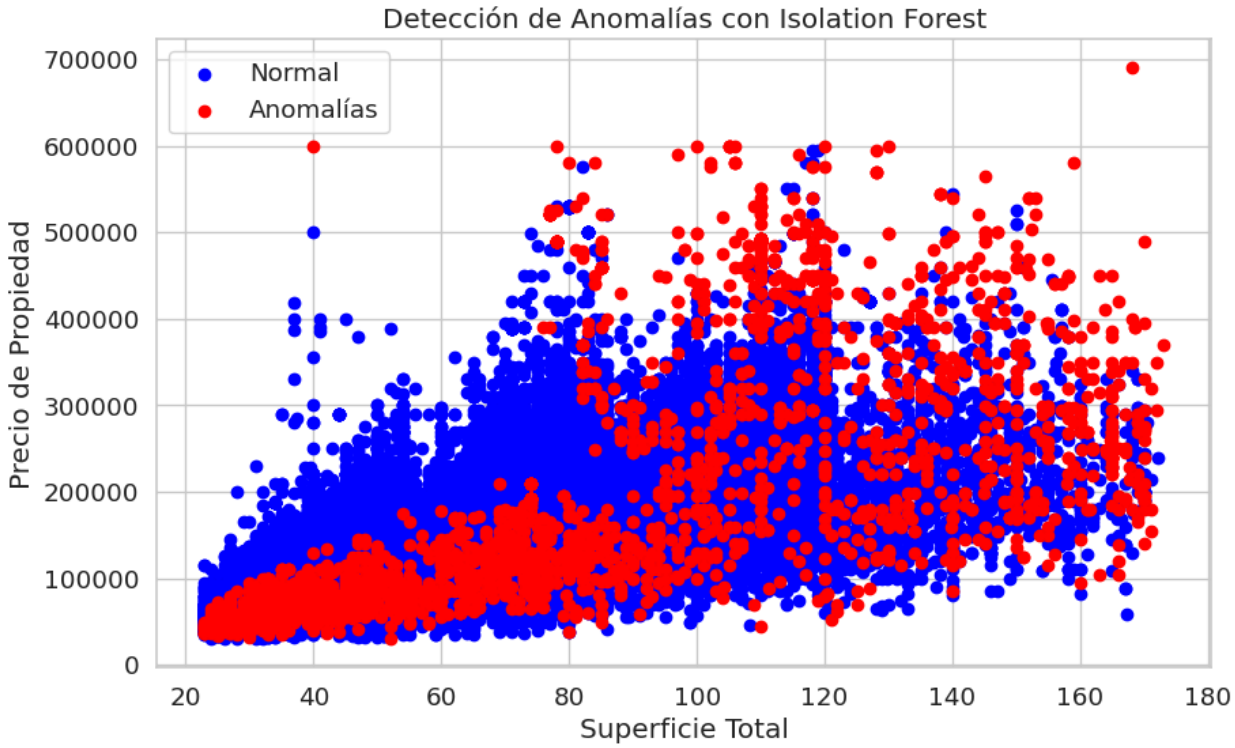
2.4.2.4. `property_surface_total` (Superficie total) vs `property_price` (Precio)

Vamos a analizar el precio de las propiedades comparándolo contra la superficie de las mismas para buscar anomalías.

Pero, hay un problema que salta a la vista en cuanto pensamos en relacionar estas dos variables ▸ El precio por metro cuadrado puede llegar a variar considerablemente en función del barrio en el cual nos encontremos.

Para intentar compensar esto, vamos a tener en cuenta también el barrio en el cual esté ubicada la propiedad.

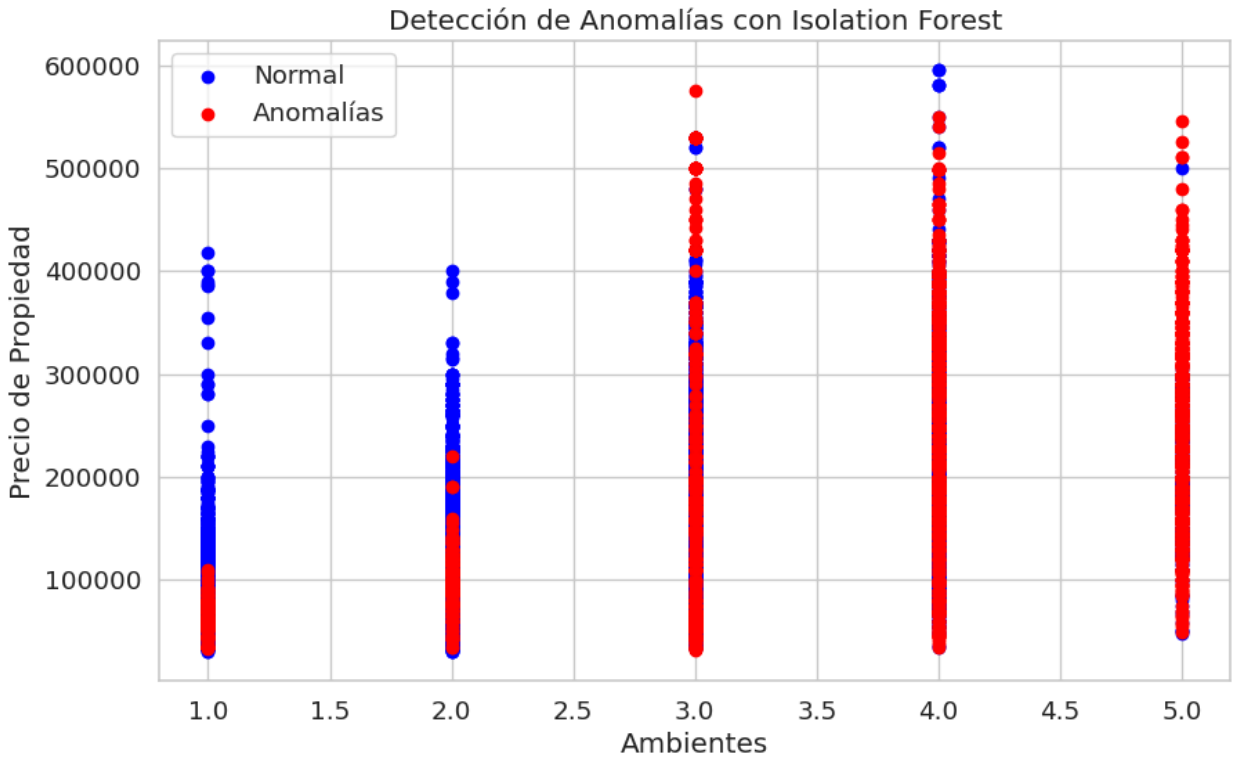
Gráfica luego del filtrado



2.4.2.5. `property_rooms`(ambientes) vs `property_price` (precio)

Similar a lo que hicimos en la sección anterior, ahora vamos a analizar el precio de las propiedades comparándolo contra la cantidad de ambientes para buscar anomalías.

Al igual que antes, también vamos a tener en cuenta el barrio en el cual se encuentra la propiedad ubicada, dado que no cuesta lo mismo un 2 ambientes en Recoleta que uno en Floresta.



Visualizaciones

Esto se realizó en la sección de análisis y procesamiento.

Clustering

Aún no se ha avanzado en esta sección.

Clasificación

Aún no se ha avanzado en esta sección.

Regresión

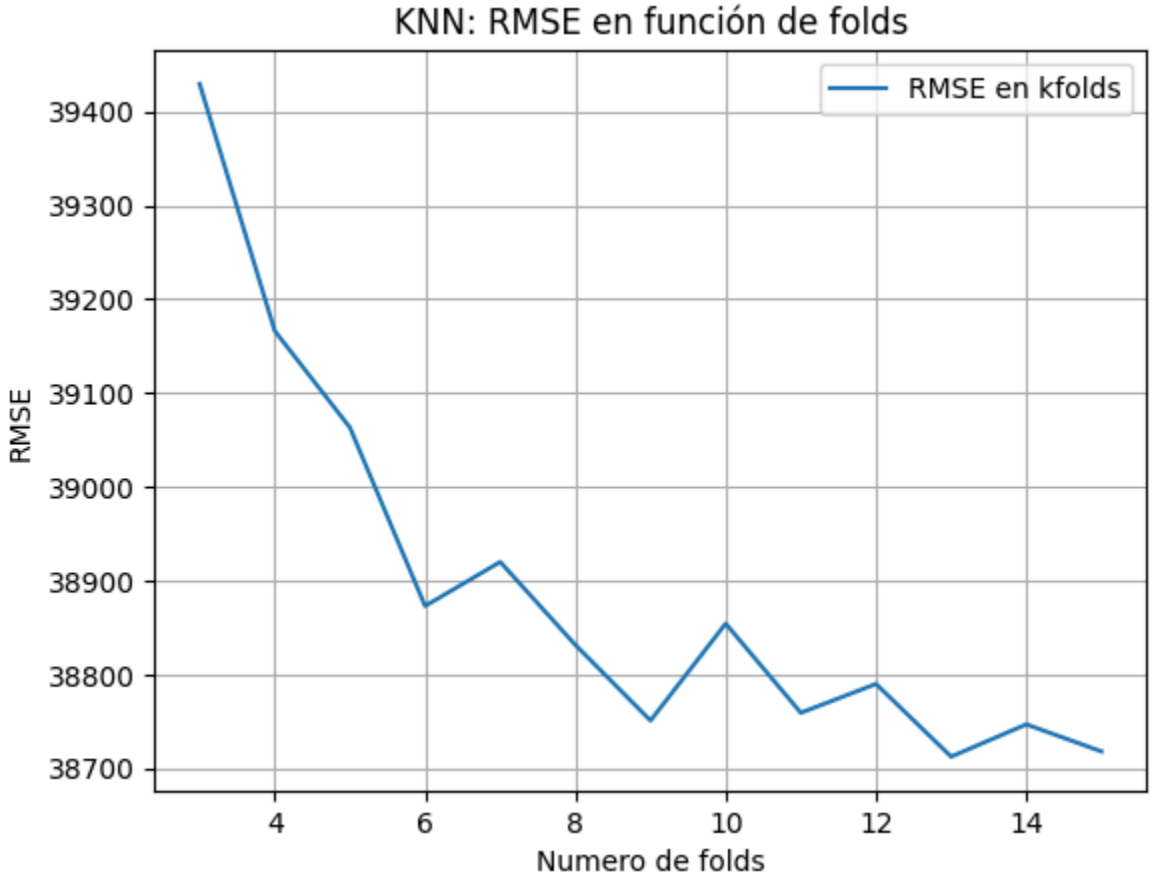
Si llegaron a entrenar alguno de los modelos, mencionar cuáles y qué métricas obtuvieron en test y si realizaron nuevas transformaciones sobre los datos (encoding, normalización, etc) completando los ítems a y b:

a. Construcción del modelo

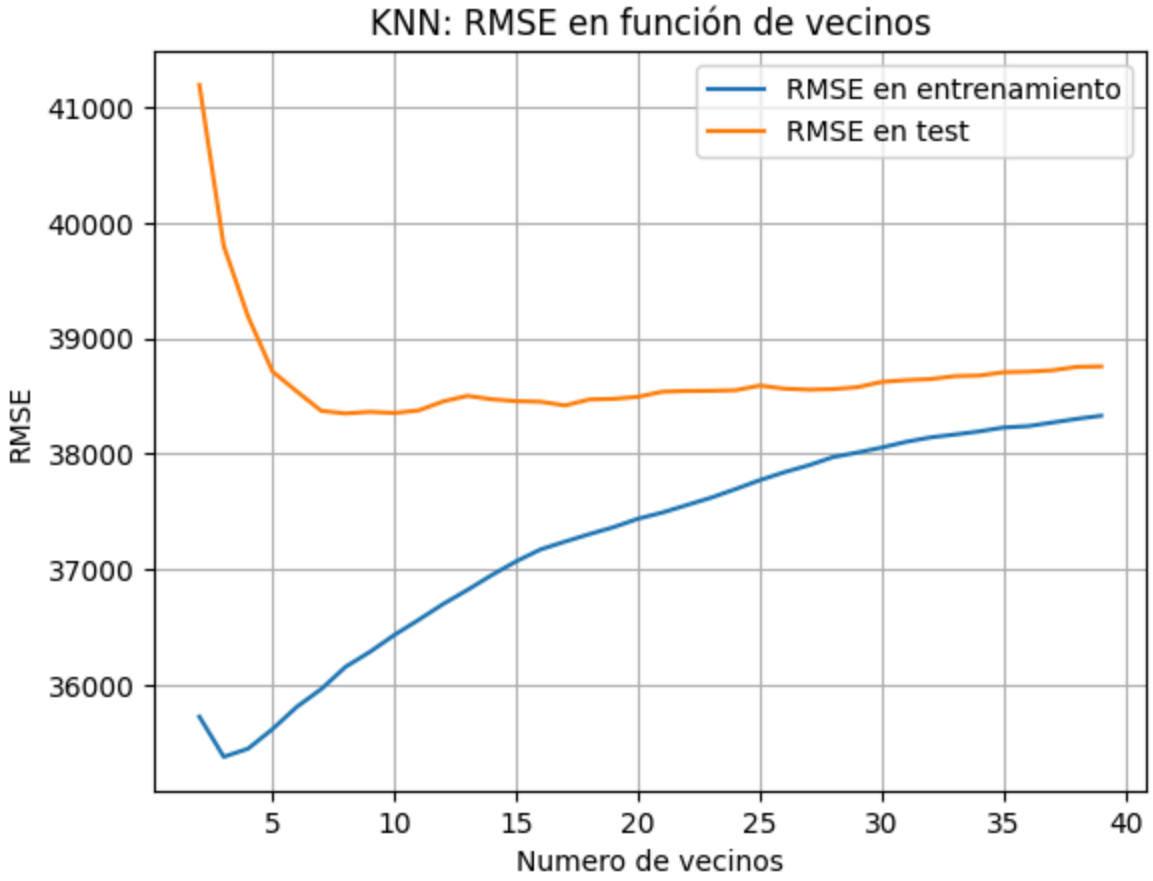
KNN

- ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
- ¿Qué métrica utilizaron para buscar los hiper parámetros?
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Se utilizó cross validation. La cantidad de folds óptima fue de **trece folds**. Para optimizar los hiper parámetros se utilizó el round mean squared error por medio de la función de scikit learn cross_val_score, teniendo en consideración también que la desviación del error entre los folds no difiera de los demás casos.



Se dejaron los parámetros por determinado del modelo (algoritmo automático, métrica minkowski con potencia de dos, peso uniforme) y se buscó la cantidad óptima de vecinos, siendo el resultado **ocho vecinos** con los errores del cuadro de resultados.



Se evidencia la diferencia de valores entre entrenamiento y evaluación, tanto en los resultados del cuadro como en el progreso del error en el número de vecinos. Esto se debe a que el modelo tiene cierto grado de overfit en la cantidad de ocho vecinos, pero con una generalización razonable respecto de una cantidad de vecinos mayor.
La tendencia en entrenamiento es a reducir el overfitting mientras más vecinos hay, mientras que en entrenamiento para una baja cantidad de vecinos realiza una mala generalización, siendo el punto crítico la cantidad de vecinos elegida.
Como última conclusión podemos observar como el modelo parecería converger a una misma tendencia para más de treinta vecinos, lo cual nos muestra que el modelo logra una buena generalización para los valores considerados.

XGBoost

- ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
- ¿Qué métrica utilizaron para buscar los hiperparámetros?
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo a elección:

- ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
- ¿Qué métrica utilizaron para buscar los hiperparámetros?
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

b. Cuadro de Resultados

Realizar un cuadro de resultados comparando los modelos que entrenaron (entre ellos debe figurar cuál es el que seleccionaron como mejor predictor).

Medidas de rendimiento en el conjunto de TEST:

- MSE
- RMSE
- XXX: si seleccionaron alguna métrica adicional...

Confeccionar el siguiente cuadro con esta información:

Modelo	MSE Test	RMSE Test	RMSE cross-Val Test	RMSE cross-Val Train	Score Test
KNN	1.21e9	34748	38345	36159	0.677
XGBoost					
Modelo					

En cada caso ¿Cómo resultó la performance respecto al set de entrenamiento?

Nota: indicar brevemente en qué consiste cada modelo de la tabla

Estado de Avance

1. Análisis Exploratorio y Preprocesamiento de Datos

Porcentaje de Avance: 100%/100%

Tareas en curso: trabajo terminado.

Tareas planificadas: ninguna.

Impedimentos: ninguno.

a) Exploración Inicial: análisis concluido.
b) Visualización de los datos: análisis concluido.
c) Datos Faltantes: análisis concluido.
d) Valores atípicos: análisis concluido.
e) Opcional: -

2. Agrupamiento

Porcentaje de Avance: 0%/100%

Tareas en curso: aún sin comenzar.

Tareas planificadas: -

Impedimentos: Estamos terminando el análisis exploratorio y preprocesamiento.

3. Clasificación

Porcentaje de Avance: 0%/100%

Tareas en curso: aún sin comenzar.

Tareas planificadas: -

Impedimentos: Estamos terminando el análisis exploratorio y preprocesamiento.

4. Regresión

Porcentaje de Avance: 25%/100%

Tareas en curso: N/A

Tareas planificadas: XGBoost y definición del último modelo

Impedimentos: Finalizar etapas anteriores

Tiempo dedicado

Indicar brevemente en qué tarea trabajó cada integrante del equipo durante estas semanas. Si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte). Deben indicar el promedio de horas semanales que dedicaron al trabajo práctico. En esta tabla solo deben incluir las tareas que realizaron luego de entregar el CHP1.

Integrante	Tarea	Prom. Hs Semana
Testa, Santiago Tomas	Regresión	3
Pratto, Federico Nicolás	Análisis Exploratorio y Preprocesamiento de Datos e informe	12
Ramirez, Jose Israel	-	-
Torres, Santiago/Danny	-	-