Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

EDA y Modelos Bayesianos

Informe Tarea 1

Integrantes: Miguel Espinoza
Sebastián López
Pedro Pérez
Profesor: Nicolás Caro.
Auxiliar: Rodrigo Lara
Fecha de entrega: 17 de Mayo de 2020
Santiago, Chile

Índice de Contenidos

Índice de Figuras

Índice de Tablas

1. Carga y Limpieza de Datos

1.1. Parte 1

En primer lugar se accede a las subcarpetas dentro de `data/raw`. Se carga cada uno de los dataframes de tipo `all` y se combinan entre ellos con la función `merge` de Pandas. Luego se revisan las entradas duplicadas y se eliminan. Posteriormente se realiza lo mismo con los dataframes de tipo `furnished`.

Luego se combinan ambos dataframes, los `all` con los `furnished`, con la función `merge`. Para saber de donde proviene cada dato en la combinación se ocupa el parámetro `indicator` en la función. Este parámetro agrega una columna al dataframe combinado. Sus entradas indican si un dato proviene del dataframe izquierdo con *left*, del derecho con *right* y de ambos con *both*.

Al revisar esta columna se observo que ningún dato provenía de ambos dataframes, no existe intersección entre `all` con `furnished`.

1.2. Parte 2

1.2.1. Parte A

Primeramente se limpia la columna `price`. Se elimina el signo $ y se quitan los puntos separadores de miles. Finalmente se convierte a `float`.

Para limpiar la columna `surface` primero se elimina la unidad `m2` y luego se convierte a `float`. En esta columna ocurre que hay algunas viviendas con área nula, lo cual no tiene sentido. Por lo tanto se decide eliminar estas entradas, que sólo eran 16, por lo que afectan mínimamente la cantidad de datos.

Para el número de habitaciones en la columna `n_rooms`, ocurre que hay 9 entradas que contienen el valor `5+`. Para evitar eliminar estas columnas se decide cambiar el valor `5+` por `6`. Luego es posible pasar los datos a `float`.

Por último, en la columna `n_bath` ocurre lo mismo. Hay 8 entradas que contienen el valor `5+`, las cuales se cambian nuevamente por `6`. Luego se cambia el tipo a `float`.

1.2.2. Parte B

Para separar la columna `property_type|rent_type|location|`, se ocupa la función `str.split` de Pandas. La cual permite separar el texto de cada fila de acuerdo a cierto separador. Los datos están en el estilo `'Apartamento en Arriendo, EL NOGAL Bogotá D.C..'`, por lo que se separan por una coma en primer lugar. Para extraer el barrio y colocarlo en la columna `location`, se toman sólo las palabras mayúsculas del lugar, que representan el barrio, y se elimina la ciudad.

Continuando, para separar el tipo de renta y el tipo de propiedad, se separa la columna `property_type|rent_type`, que contiene cadenas del estilo `'Apartamento en Arriendo'`, con el separador `en`. Luego cada una de estas columnas se agrega a los datos.

Finalmente la columna `property_type|rent_type|location|`, quedo separada en las columnas `property`, `rent` y `location`.

1.3. Parte 3

1.3.1. Parte A

Para calcular el precio por m^2 de la vivienda, simplemente se divide la columna `price` por la columna `surface`. El resultado se agrega a la data como `Precio m2`.

1.3.2. Parte B

Para agregar la columna que tenga el número de garajes de la vivienda, se debe trabajar con la columna `url`. Las url tienen la forma, `https://www.metrocuadrado.com/inmueble/arriendo-casa-bogota-la-soledad-norte-3-habitaciones-2-banos-1-garajes/35-M1919`, se ve que al final del último *slash* `/`, aparece la cadena `1-garajes`. Por lo tanto se deduce que si esta escrito `n-garajes`, `n` es el número de garajes de la propiedad.

En base a lo deducido, se decide separar la url en base al carácter slash `/`. De la lista que aparece, se toma el elemento que contiene la cadena `arriendo-casa-bogota-la-soledad-norte-3-habitaciones-2-banos-1-garajes`. Luego, se separa la cadena anterior por el guión `-`. Se revisa si el último elemento de la lista es `garajes`, y si es así, se añade el penúltimo elemento a la columna como el número de garajes. Si no lo es, se coloca un 0.

Por último, ocurre que que existen 9 entradas con valor `4+`, éstas al igual que antes, se cambian por `5`. Por último se convierte la columna a `float`.

1.4. Parte 4

Se añade una nueva columna llamada `tipo de producto`, la cual tiene distintos valores de acuerdo al tipo de vivienda y a su área. Para poder clasificar los datos, se utiliza el método `query` de Pandas. Este método permite extraer las entradas que cumplan cierta condición.

Se extraen los 8 tipos de productos según la clasificación de la tarea en 8 series distintas. Mediante el método `concat` se unen estas series y se agregan a la data. Sin embargo se encuentra que un 14.9 % de las viviendas no entra en ninguna clasificación. Dado que es una cantidad importante de datos, se decide estudiarla más a fondo.

Se tiene que los datos que no entraron en la clasificación anterior pertenecen a uno de los 4 grupos que aparecen en la tabla 1. Asimismo se muestra la cantidad de datos de cada grupo.

Tabla 1: Grupos que no entraron en ninguna categoría de producto.

Grupo	Cantidad	Nuevo tipo
Casa con área menor a $80m^2$	188	9
Casa con área mayor a $460m^2$	166	10
Apartamento con área menor a $40m^2$	1084	11
Apartamento con área mayor a $120m^2$	1308	12

Finalmente como se vió que eran tantos datos, se deciden agregar como nuevos tipos a la columna `tipo de producto`, de acuerdo a la columna `Nuevo Tipo` que aparece en la tabla 1.

1.5. Parte 5

Para añadir el código UPZ a cada barrio se debe revisar archivo `data/asignacion_upz/barrio-upz.csv`. De este archivo las columnas que interesan son `pro_location` y `UPlCodigo`.

En primer lugar se revisa `pro_location`. En esta columna hay 12 valores faltantes. Es necesario que no haya ningún valor `NaN`, por lo tanto se cambian los `NaN` por `'NaN'`. Luego se convierte la columna a tipo string.

Se observa que `pro_location` contiene string del tipo `'palermo'`, `'molinos norte'` o `'chico norte ii'`. Estos string son iguales a los que tiene la columna `location`, de la data. La única diferencia es que en esta última están escritos en mayúsculas.

Para agregar el código UPZ a la data, se itera sobre toda las filas de `location`, y mediante el método `str.contains` se revisa si el texto está en algunas de las filas de `pro_location`. Si es así, se busca el índice en el que está el string. Con el índice se puede acceder al código UPZ del barrio en `UPlCodigo`, el cual se agrega a la data en la columna `CodigoUPZ`.

Por último, se observa que a 1741 de los datos les falta código UPZ. Es decir, un 90.6 % de los datos tiene código, esto es muy cercano al porcentaje que se entrega en el enunciado de la tarea. En total son 168 los barrios a los que no se les puede adjuntar el código UPZ.

1.6. Parte 6

Se accede a la carpeta `data/estadisticas_upz/` y se cargan los archivos que hay en ella. Estos archivos son: `estadisticas_poblacion.csv`, `indice_inseguridad.csv`, `porcentaje_areas_verdes.csv` y `readme.md`. Se trabajará sólo con los 3 primeros. Para combinar los dataframes se utiliza el método `merge`.

Para agregar la columna con la densidad de población por UPZ, se debe dividir la cantidad de población del barrio por el área del barrio. El área se encuentra en el dataframe `barrio-upz.csv`, en la columna `UPlArea`. Se añade el área de cada barrio según su código UPZ a la data mediante

el método `merge`. Luego simplemente se dividen las columnas `personas` y `UPlArea`, y se añade el resultado a la columna `pob_density` de la data.

Finalmente se cambia el nombre a algunas columnas para que sean más cortos. Se detallan las columnas del dataset en la tabla 2 del anexo.

2. EDA

En esta sección se buscará hacer un análisis exploratorio de los datos con los que se está trabajando, y así obtener y filtrar información para la última parte del laboratorio.

2.1. Parte 1

En esta sección se crea una función llamada `estilo()` que definirá la configuración visual de los gráficos que se hagan en el notebook.

El estilo se define 'ticks', que asigna mantiene los bordes del gráfico y asigna lineas a los distintos valores respresentados en los ejes. Se escogió el contexto 'notebook', que configura el tamaño relativo entre las lineas de los gráficos de forma adecuada a la visualización en el `.ipynb`.

2.2. Parte 2

Luego, se separaron las variables por su naturaleza a través de un multiíndice, y se realizó un perfil univariado de cada variable.

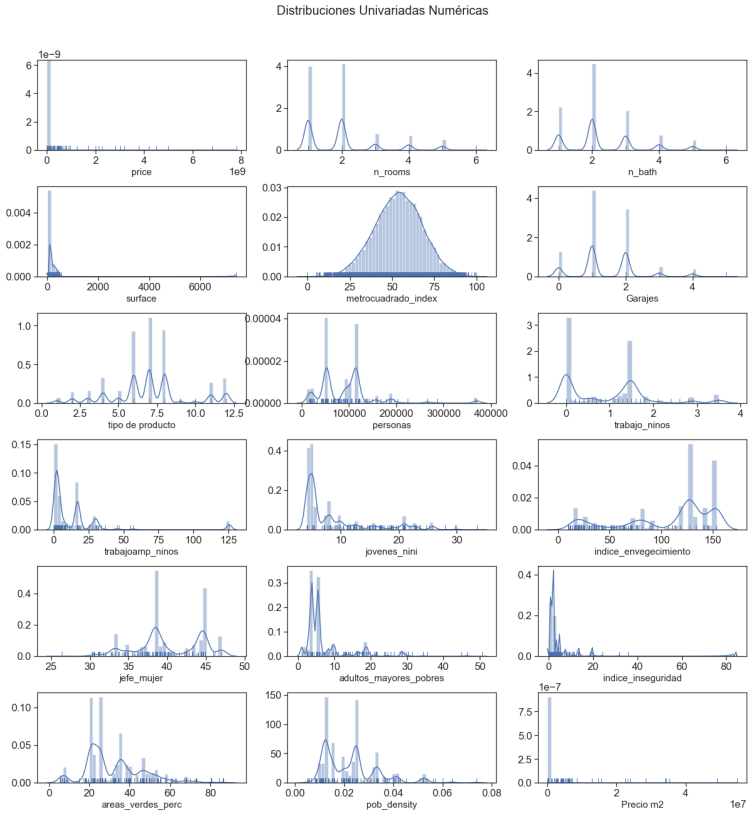

Figura 1: Distribuciones univariadas para variables numéricas.

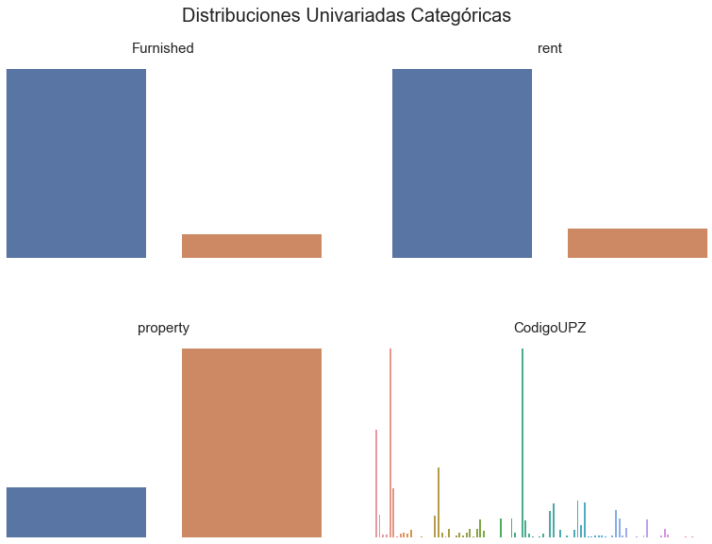

Figura 2: Distribuciones univariadas para variables categóricas.

Además se realizó un perfil bivariado entre las variables y se compararon las varibles categóricas contra la variable de respuesta en un gráfico tipo violín, además de estudiar las correlaciones entre las distintas variables.

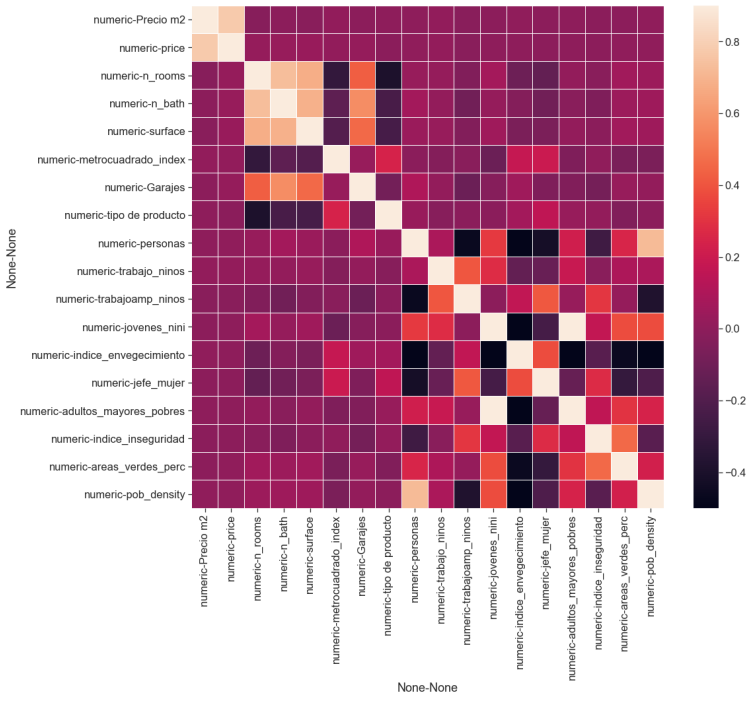

Figura 3: Tabla de correlaciones.

De la figura 1 se puede destacar que '`metrocuadrado_index`' podría ser una variable que vale la pena estudiar y también podría serlo '`tipo de producto`'.

Además, en primera instancia no se destacan correlaciones muy grandes con la variable de respuesta, excepto por 'price', pero esta variable se descarta ya que lo que se quiere predecir es el precio/m^2.

2.3. Parte 3

Para estudiar los datos faltantes, se reemplazaron los 'nan' por `np.nan` y se visualizó la distribución de datos faltantes.

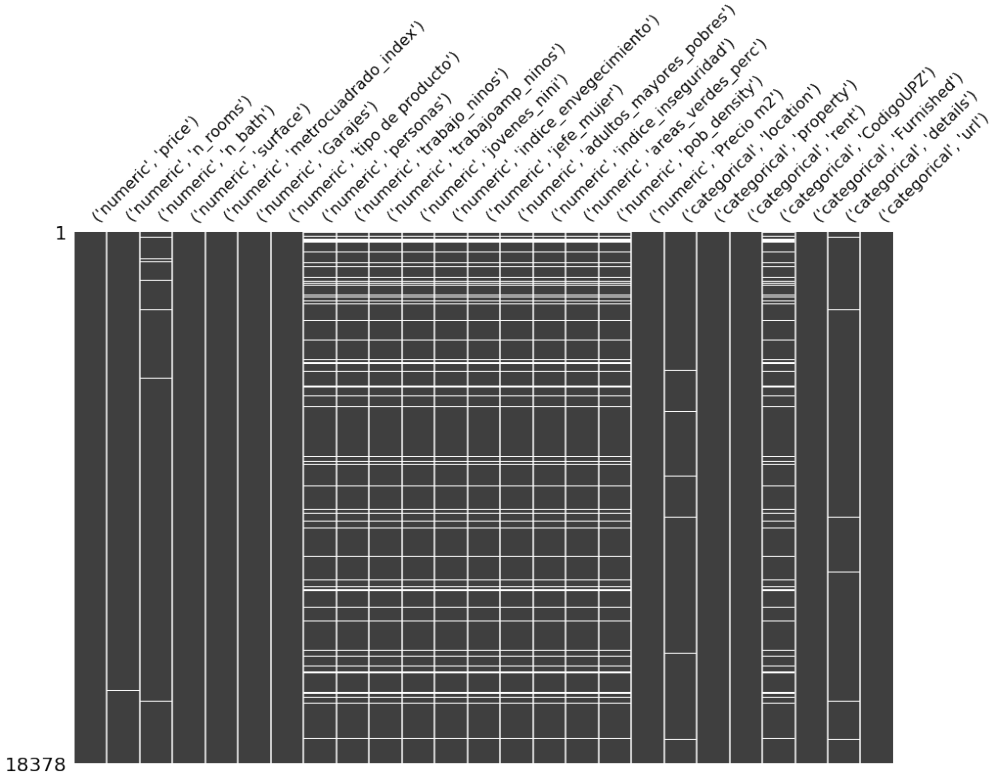

Figura 4: Tabla indicadora de valores faltantes, las líneas blancas indican en que columnas hay valores NaN para un dato.

De la siguiente visualización se infiere que si a un dato le falta información en la columna 'CodigoUPZ', entonces le faltará información en las columnas 'personas', 'trabajo_ninos', 'trabajoamp_ninos', 'jovenes_nini','indice_envegecimiento', 'jefe_mujer', 'adultos_mayores_pobres', 'indice_inseguridad', 'areas_verdes_perc', 'pob_density'.

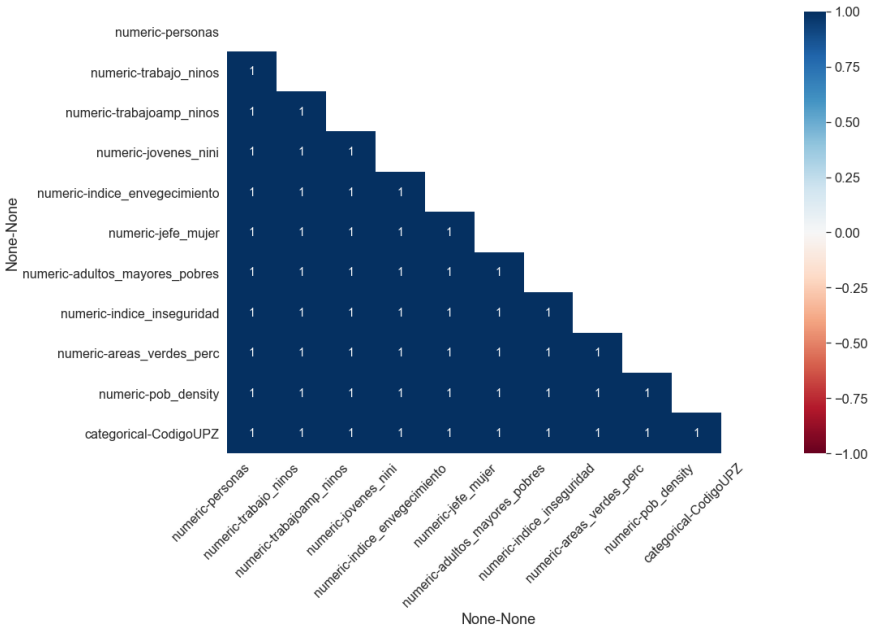

Figura 5: Tabla de correlaciones para variables con datos faltantes.

Debido a esto concluimos que la información faltante es del tipo MNAR, ya que depende de una variable que tiene pérdida de información.

2.4. Parte 4

Luego se procede a recategorizar la variable `'CodigoUPZ'`, para esto los datos faltantes se modifican en una nueva categoría llamada 'UPZ0'. Estas variables al ser categóricas, se transforman a numéricas a través de `OneHotEncoder`, agregando una dimensión por cada categoría distinta y luego se procede a hacer k-means para poder distinguir distintos grupos sobre la variable de respuesta `'Precio m2'`.

Viendo la curva de scores asociado a la distinta cantidad de clusters, se realiza k-means con 4 y 5 clusters y finalmente el que dió mejor resultado fue la cantidad de 4 clusters. Estos clusters se testearon bajo el test F, para ver si generaban grupos suficientemente distintos en la variable de respuesta, donde el test arrojó un valor de $p = 9.3152e - 05$, que es menor a una significacncia del 5 % y así se concluye que los nuevos grupos generan diferencias suficientes en `'Precio m2'`.

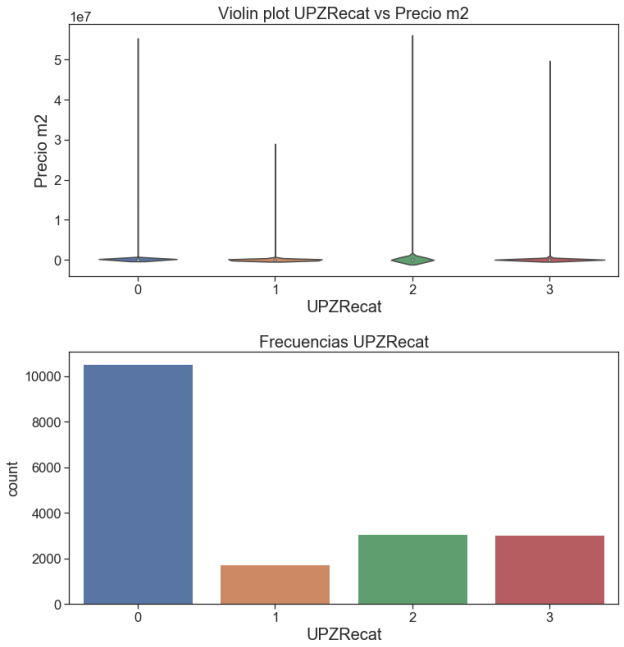

Figura 6: Gráfico de tipo violín para la nueva variable 'UPZRecat' y la variable de respuesta.

2.5. Parte 5

Se escogen 10 variables más la variable de respuesta para estudiar sus relaciones en particular. Las variables son 'n_rooms', 'n_bath', 'Garajes', 'metrocuadrado_index', 'tipo de producto', 'indice_envegecimiento', 'jefe_mujer', 'UPZRecat', 'property', 'Furnished' y la variable de respuesta 'Precio m2'.

Estas variables se escogieron porque en primera instancia son las que tienen mejor relación con 'Precio m2'.

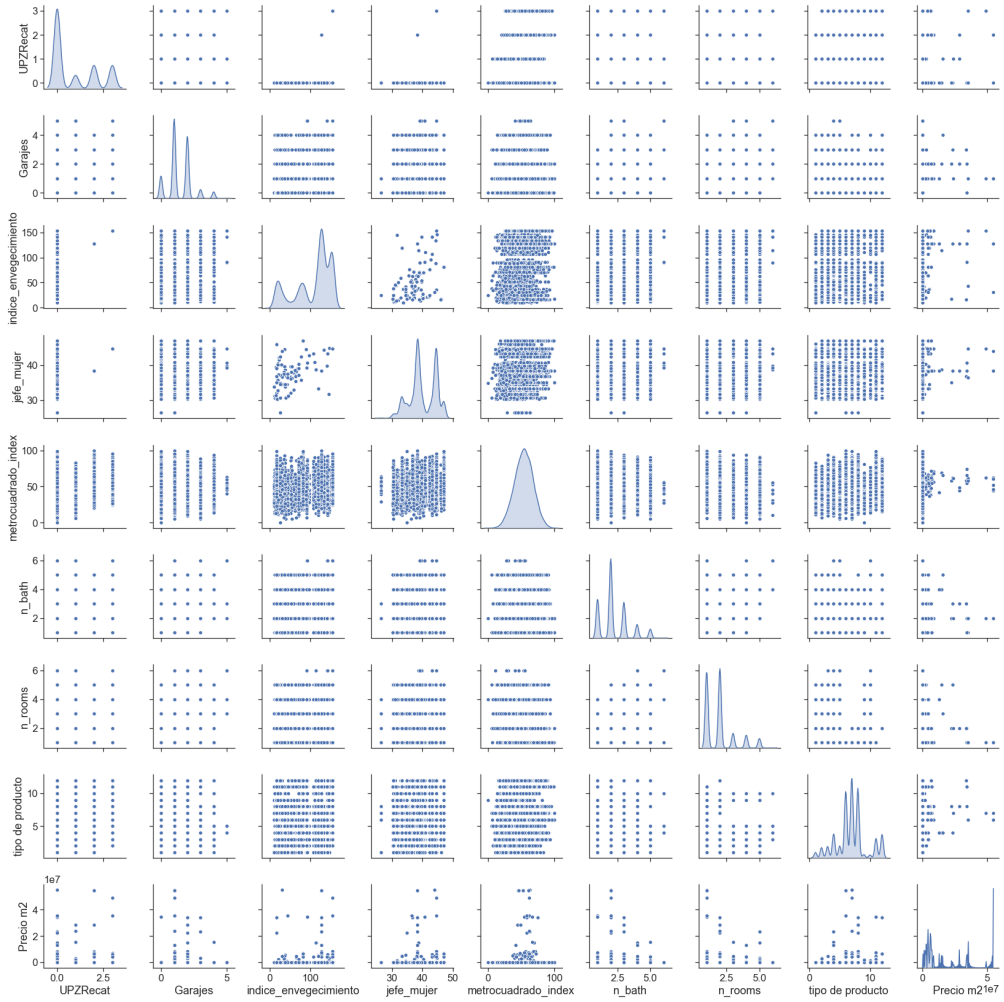

Figura 7: Distibuciones bivariadas para las variables escogidas

En esta instancia no se logra apreciar bien relaciones entre la variables escogidas y 'Precio m2', ya que la mayoría datos tienen valores cercanos a 0 en esta variable. Se logra ver una relación positiva entre las variables 'n_bath' y 'n_rooms', por lo que sería adecuado quedarse con una de estas variables y no ambas.

2.6. Parte 6

Una vez realizado un primer análisis de los datos se procede a estudiar los valores anómalos en la data.

En primera instancia se estudia la variable de respuesta que en el primer análisis se ve que tiene valores muy alejados del la mayoría de los datos. Esta variable, por lo que se ve en las gráficas, parece tener una distribución exponencial, pero al hacer un filtrado poco riguroso, se logra ver que la gran parte de los datos se comporta relativamente como una normal, en ambos casos se puede usar la relación interquartílica para detectar valores anómalos, por lo que se usa esta herramienta.

La cota superior para 'Precio m2' es 69478, por lo que se filtra a ese nivel, se destaca que menos del 2 % de los datos son anómalos, por lo que se acepta este filtrado; y se estudia nuevamente el comportamiento de las variables ahora sin considerar los valores anómalos en la data.

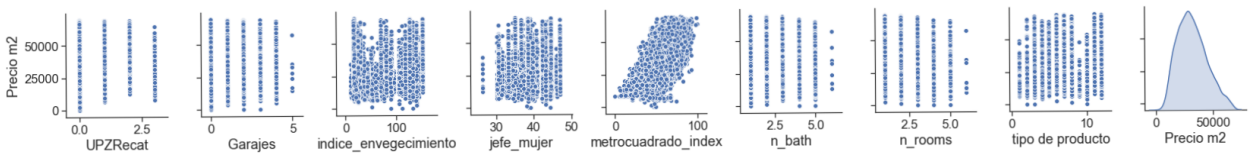

Figura 8: Última fila de la matriz de distribuciones bivariadas, una vez realizado el filtrado de valores anómalos.

A partir del nuevo estudio se concluye que filtrar respecto a la variable de respuesta es suficiente para mejorar de manera significativa las relaciones entra variables sin quitar demasiada información de la data, por lo que no se estudian las demás variables en esta ocasión.

Además, se ve la distribución de los valores anómalos respecto a las variables 'UPZRecat', 'tipo de producto' como se ve en la siguiente figura

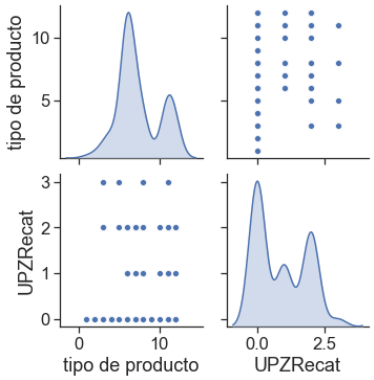

Figura 9: Distribución de los datos filtrados con respecto a las variables 'UPZRecat' y 'tipo de producto'

Se puede apreciar, en relacion a las distibuciones bivariadas respecto a 'Precio m2' de ambas variable vistas en la figura 7, que los valores anómalos siguen distribuciones muy similares a las que muestran estas variables sobre el total de datos, es decir, los datos filtrados son sacados más o menos al azar desde el punto de vista de las variables 'UPZRecat', 'tipo de producto',

2.7. Parte 7

Finalmente, se decide que las variables que permiten estimar mejor son 'n_rooms', 'metrocuadrado_index', 'tipo de producto', 'indice_envegecimiento', 'jefe_mujer', 'UPZRecat', 'property', 'Furnished', ya que al realizar una segunda etapa de todo lo mencionado en esta pregunta, se obtienen mejores correlaciones, separaciones y distribuciones en relación a la variable de respuesta.

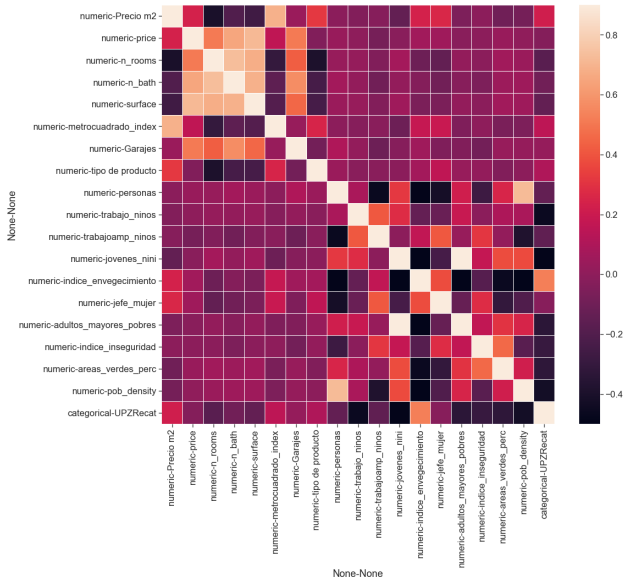

Figura 10: Matriz de correlaciones entre las variables, una vez realizado el filtrado de datos con valores anómalos.

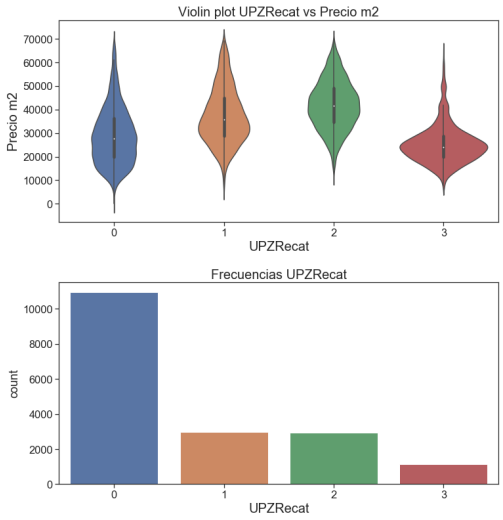

Figura 11: Gráfico de violín para la nueva variable 'UPZRecat', una vez realizado el filtrado de datos con valores anómalos.

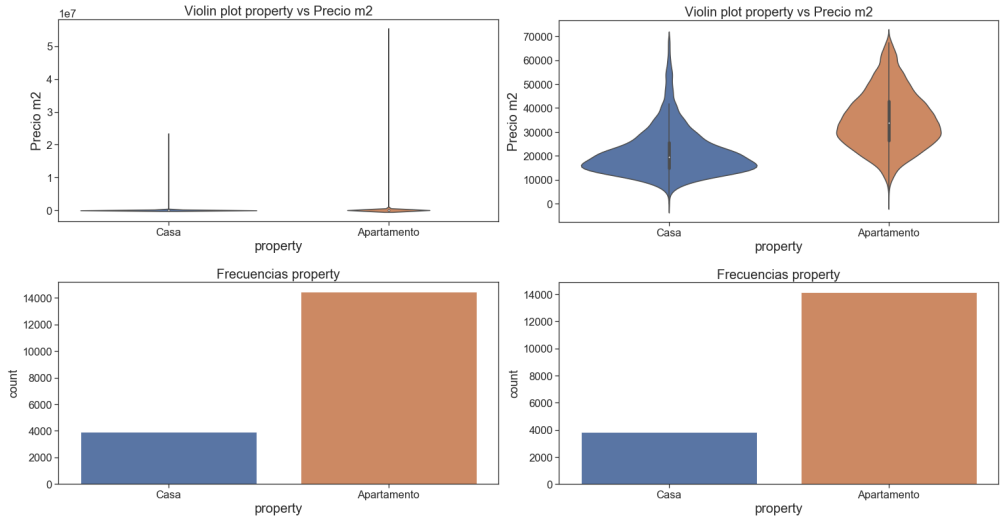

Figura 12: Gráfico de violín para la variable '`porperty`', antes (a la izquierda) y después (a la derecha) de haber realizado el filtrado de datos con valores anómalos.

3. Regresión Lineal Bayesiana

3.1. Desarrollo Teórico

3.1.1. Parte 1

Duante todo el problema, utilizaremos $X \in \mathbb{R}^{N\times d}$, $y \in \mathbb{R}^N$, $w \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$.
Se pide mostrar que la densidad

$$p(w|y, X, \alpha, \beta) \sim N(m_N, S_N)$$

Donde $S_N^{-1} = \alpha I + \beta X^T X$ y $m_N = \beta S_N X^T y$. Para esto, utilizaremos que:

$$p(y_i|x_i, w, \beta) \sim N(w^T x_i, \beta^{-1})$$

$$p(w|\alpha) \sim N(0, \alpha^{-1} I)$$

$$p(w|y, X, \alpha, \beta) = \frac{p(y|X, w, \beta)p(w|\alpha)}{p(y)}$$

Para efectos de esta pregunta, asumiremos que $p(y)$ es constante.
Primero, podemos ver que:

$$\begin{aligned} p(w|y, X, \alpha, \beta) &= \frac{1}{p(y)} \cdot \frac{1}{(2\pi)^{N/2}|\beta^{-1}I|^{1/2}} e^{-\frac{1}{2}(y-Xw)^T \beta I (y-Xw)} \cdot \frac{1}{(2\pi)^{d/2}|\alpha^{-1}I|^{1/2}} e^{-\frac{1}{2}w^T \alpha I w} \\ &= K \cdot \frac{1}{(2\pi)^{d/2}} e^{-\frac{\beta}{2}y^T y} e^{-\frac{1}{2}(w^T(\beta X^T X + \alpha I)w - 2w^T \beta X^T y)} \end{aligned}$$

Aquí podemos notar que:

$$w^T(\beta X^T X + \alpha I)w - 2w^T \beta X^T y = (w - m_N)^T S_N^{-1}(w - m_N) - m_N^T S_N^{-1} m_N$$

Por lo tanto, volviendo a la expresón anterior:

$$\begin{aligned} p(w|y, X, \alpha, \beta) &= K \frac{1}{(2\pi)^{d/2}} e^{-\frac{\beta}{2}y^T y} e^{-\frac{1}{2}((w-m_N)^T S_N^{-1}(w-m_N) - m_N^T S_N^{-1} m_N)} \\ &= K' \frac{1}{(2\pi)^{d/2}|S_N|^{1/2}} e^{-\frac{1}{2}((w-m_N)^T S_N^{-1}(w-m_N))} \end{aligned}$$

Con esto, queda demostrado que $p(w|y, X, \alpha, \beta)$ es proporcional, por una cte K' que depende de X, y, α, β, a una distribución $N(m_N, S_N)$.

3.1.2. Parte 2

Queremos demostrar que la densidad $p(y'|X', y, X, \alpha, \beta)$ definida como:

$$p(y'|x', y, X, \alpha, \beta) = \int p(y'|x', w, \beta)p(w|y, X, \alpha, \beta)dw$$

cumple que es proporcional a una normal, es decir:

$$p(y'|x', y, X, \alpha, \beta) \sim N(m_N^T x', \sigma_N^2(x'))$$

$$\sigma_N^2(x') = \frac{1}{\beta} + x'^T S_N x'$$

Notemos que por enunciado y parte anterior:

$$\begin{aligned} p(y'|x', y, X, \alpha, \beta) &= \int \frac{\beta^{1/2}}{(2\pi)^{1/2}} e^{-\frac{\beta(y'-w^T x')^2}{2}} \cdot K' \frac{1}{(2\pi)^{d/2}|S_N|^{1/2}} e^{-\frac{1}{2}((w-m_N)^T S_N^{-1}(w-m_N))} dw \\ &= K' \frac{\beta^{1/2}}{(2\pi)^{1/2}} \frac{1}{(2\pi)^{d/2}|S_N|^{1/2}} \int e^{-\frac{1}{2}(\beta(y'-w^T x')^2 + (w-m_N)^T S_N^{-1}(w-m_N))} dw \end{aligned}$$

Trabajemos con el término dentro de la integral:

$$e^{-\frac{1}{2}(\beta(y'-w^T x')^2 + (w-m_N)^T S_N^{-1}(w-m_N))} = e^{-\frac{1}{2}(\beta y'^2 + m_N^T S_N^{-1} m_N)} e^{-\frac{1}{2}(w^T(\beta x' x'^T + S_N^{-1})w - 2w^T(\beta y' x' + S_N^{-1} m_N))}$$

Definamos

$$L^{-1} = \beta x' x'^T + S_N^{-1}$$

$$L^{-1}\mu = \beta y' x' + S_N^{-1} m_N$$

Luego

$$\mu = L(\beta y' x' + S_N^{-1} m_N)$$

Volviendo a la ecuación anterior

$$e^{-\frac{1}{2}(\beta(y'-w^T x')^2 + (w-m_N)^T S_N^{-1}(w-m_N))} = e^{-\frac{1}{2}(\beta y'^2 + m_N^T S_N^{-1} m_N - \mu^T L^{-1}\mu)} e^{-\frac{1}{2}((w-\mu)^T L^{-1}(w-\mu))}$$

Luego volviendo a la integral

$$\begin{aligned} p(y'|x', y, X, \alpha, \beta) &= K' \frac{\beta^{1/2}}{(2\pi)^{1/2}} \frac{1}{(2\pi)^{d/2}|S_N|^{1/2}} \int e^{-\frac{1}{2}(\beta y'^2 + m_N^T S_N^{-1} m_N - \mu^T L^{-1}\mu)} e^{-\frac{1}{2}((w-\mu)^T L^{-1}(w-\mu))} dw \\ &= K' \frac{\beta^{1/2}}{(2\pi)^{1/2}} \frac{|L|^{1/2}}{|S_N|^{1/2}} e^{-\frac{1}{2}(\beta y'^2 + m_N^T S_N^{-1} m_N - \mu^T L^{-1}\mu)} \int \frac{1}{(2\pi)^{d/2}|L|^{1/2}} e^{-\frac{1}{2}((w-\mu)^T L^{-1}(w-\mu))} dw \end{aligned}$$

Esta última integral vale 1, pues es integrar la densidad de una normal para w. Luego

$$p(y'|x', y, X, \alpha, \beta) = K'' \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(\beta y'^2 - \mu^T L^{-1}\mu)}$$

Ahora desarrollemos el término $\mu^T L^{-1}\mu$

$$\begin{aligned} \mu^T L^{-1}\mu &= (\beta y' x' + S_N^{-1} m_N)^T L L^{-1} L (\beta y' x' + S_N^{-1} m_N) \\ &= (\beta^2 x'^T L x') y'^2 + (2\beta x'^T L S_N^{-1} m_N) y' + m_N^T S_N^{-1} L S_N^{-1} m_N \end{aligned}$$

Entonces

$$\beta y'^2 - \mu^T L^{-1}\mu = \beta(1 - \beta x'^T L x')y'^2 - 2(\beta x'^T L S_N^{-1} m_N)y' - m_N^T S_N^{-1} L S_N^{-1} m_N$$

Definimos

$$\frac{1}{\lambda^2} = \beta(1 - \beta x'^T L x')$$
$$u = \lambda^2(\beta x'^T L S_N^{-1} m_N)$$

Entonces

$$\begin{aligned}\beta y'^2 - \mu^T L^{-1}\mu &= \frac{1}{\lambda^2}y'^2 - 2\frac{1}{\lambda^2}uy' - m_N^T S_N^{-1} L S_N^{-1} m_N \\ &= \frac{1}{\lambda^2}(y' - u)^2 - \frac{1}{\lambda^2}u^2 - m_N^T S_N^{-1} L S_N^{-1} m_N\end{aligned}$$

Volviendo a la exponencial

$$\begin{aligned}p(y'|x', y, X, \alpha, \beta) &= K'' \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(\frac{(y'-u)^2}{\lambda^2} - \frac{1}{\lambda^2}u^2 - m_N^T S_N^{-1} L S_N^{-1} m_N)} \\ &= K''' \frac{1}{(2\pi)^{1/2}\lambda} e^{-\frac{1}{2}\frac{(y'-u)^2}{\lambda^2}}\end{aligned}$$

Por lo tanto, $p(y'|x', y, X, \alpha, \beta)$ es proporcional, por K''', a la densidad de una normal. Solo falta ver que

$$u = m_N^T x'$$
$$\lambda^2 = \frac{1}{\beta} + x'^T S_N x'$$

Por la **identidad de Sherman-Morrison**[1] tenemos que

$$L = (S_N^{-1} + \beta x' x'^T)^{-1} = S_N - \frac{\beta S_N x' x'^T S_N}{1 + \beta x'^T S_N x'}$$

Entonces

$$\frac{1}{\lambda^2} = \beta(1 - \beta x'^T(S_N - \frac{\beta S_N x' x'^T S_N}{1 + \beta x'^T S_N x'})x') = \frac{\beta}{1 + \beta x'^T S_N x'}$$

Luego

$$\lambda^2 = \frac{1}{\beta} + x'^T S_N x'$$

Por último, para u

$$u = \lambda^2(\beta x'^T L S_N^{-1} m_N) = (\frac{1}{\beta} + x'^T S_N x')(\beta x'^T(S_N - \frac{\beta S_N x' x'^T S_N}{1 + \beta x'^T S_N x'})S_N^{-1} m_N) = x'^T m_N$$

Con esto, queda mostrado que $p(y'|x', y, X, \alpha, \beta)$ es proporcional a $N(m_N^T x', \frac{1}{\beta} + x'^T S_N x')$

3.1.3. Parte 3

Partamos de la identidad

$$p(y|\alpha,\beta) = \int p(y|\alpha,\beta,w)p(w|\alpha)dw$$

Y ya sabemos que $p(y|\alpha,\beta,w) \sim N(Xw,\beta^{-1}I)$ y $p(w|\alpha) \sim N(0,\alpha^{-1}I)$. Luego

$$\begin{aligned} p(y|\alpha,\beta) &= \int \frac{1}{(2\pi)^{N/2}\beta^{-N/2}} e^{-\frac{1}{2}\beta(y-Xw)^T(y-Xw)} \frac{1}{(2\pi)^{d/2}\alpha^{-d/2}} e^{-\frac{1}{2}\alpha w^T w} \\ &= \frac{|S_N|^{1/2} e^{-\frac{1}{2}(\beta y^T y - m_N^T S_N^{-1} m_N)}}{(2\pi)^{N/2}\beta^{-N/2}\alpha^{-d/2}} \int \frac{1}{(2\pi)^{d/2}|S_N|^{1/2}} e^{-\frac{1}{2}((w-m_N)^T S_N^{-1}(w-m_N))} \end{aligned}$$

Donde podemos notar que la integral vale 1, pues es integrar una densidad normal. Entonces tomando logaritmo

$$log(p(y|\alpha,\beta)) = \frac{d}{2}log(\alpha) + \frac{N}{2}log(\beta) - \frac{1}{2}log(|S_N^{-1}|) - \frac{N}{2}log(2\pi) - \frac{1}{2}(\beta y^T y - m_N^T S_N^{-1} m_N)$$

Desarrollando el último término, notemos

$$m_N^T S_N^{-1} m_N = \alpha m_N^T m_N + \beta m_N^T X^T X m_N$$

Luego

$$\begin{aligned} \beta y^T y - m_N^T S_N^{-1} m_N &= \beta(y^T y - (Xm_N)^T X m_N) - \alpha m_N^T m_N \\ &= \beta||y - Xm_N||_2^2 + \alpha m_N^T m_N = 2E(m_N) \end{aligned}$$

Volviendo a la expresión anterior, concluimos que

$$log(p(y|\alpha,\beta)) = \frac{d}{2}log(\alpha) + \frac{N}{2}log(\beta) - \frac{1}{2}log(|S_N^{-1}|) - \frac{N}{2}log(2\pi) - E(m_N)$$

3.1.4. Parte 4

Para maximizar la log-verosimilitud de la parte 3, con respecto a α y β, vamos a derivarla marginalmente e igualar a 0. Durante el desarrollo, asumiremos que m_N es independiente de α y β.
Primero, para α

$$\frac{\partial}{\partial\alpha} log(p(y|\alpha,\beta)) = \frac{d}{2\alpha} - \frac{1}{2}\frac{\partial}{\partial\alpha} log(|S_N^{-1}|) - \frac{\partial}{\partial\alpha} E(m_N)$$

Primero, resolvamos la derivada de $E(m_N)$

$$\frac{\partial}{\partial\alpha} E(m_N) = \frac{m_N^T m_N}{2}$$

Y para $log(|S_N^{-1}|)$, notemos primero que X^TX es una matriz simética, por lo tanto diagonalizable, digamos $X^TX = PDP^{-1}$. Luego

$$S_N^{-1} = P(\alpha I + \beta D)P^{-1}$$

Por lo tanto, los valores propios de la matriz S_N^{-1} son de la forma $(\alpha + \beta\delta_i)$, con δ_i los valores propios de la matriz X^TX. También podemos notar que los valores propios de la matriz βX^TX, llamémoslos λ_i, son de la forma $\lambda_i = \beta\delta_i$.
Luego por teorema de álgebra lineal

$$|S_N^{-1}| = \prod_{i=1}^{d}(\alpha + \lambda_i) \Rightarrow log(|S_N^{-1}|) = \sum_{i=1}^{d} log(\alpha + \lambda_i) \Rightarrow \frac{\partial}{\partial\alpha} log(|S_N^{-1}|) = \sum_{i=1}^{d} \frac{1}{\alpha + \lambda_i}$$

Volviendo a la ecuación principal, y definiendo $\gamma = \sum_{i=1}^{d} \frac{\lambda_i}{\alpha+\lambda_i}$

$$\frac{\partial}{\partial\alpha} log(p(y|\alpha, \beta)) = \frac{d}{2\alpha} - \frac{1}{2}\sum_{i=1}^{d} \frac{1}{\alpha + \lambda_i} - \frac{m_N^T m_N}{2}$$

Imponemos $\frac{\partial}{\partial\alpha} log(p(y|\alpha, \beta)) = 0$ para maximizar α

$$0 = \frac{d}{2\alpha} - \frac{1}{2}\sum_{i=1}^{d} \frac{1}{\alpha + \lambda_i} - \frac{m_N^T m_N}{2}$$
$$\alpha = \frac{\gamma}{m_N^T m_N}$$

Ahora, para β

$$\frac{\partial}{\partial\beta} log(p(y|\alpha, \beta)) = \frac{N}{2\beta} - \frac{1}{2}\frac{\partial}{\partial\beta} log(|S_N^{-1}|) - \frac{\partial}{\partial\beta} E(m_N)$$

Resolvamos la derivada de $E(m_N)$

$$\frac{\partial}{\partial\beta} E(m_N) = \frac{||y - Xm_N||_2^2}{2}$$

Ahora, notemos que

$$\frac{\partial(\alpha + \lambda_i)}{\partial\beta} = \frac{\partial\beta\delta_i}{\partial\beta} = \delta_i = \frac{\lambda_i}{\beta}$$

Por lo tanto, por un desarrollo análogo al hecho para α

$$\frac{\partial}{\partial\beta} log(|S_N^{-1}|) = \sum_{i=1}^{d} \frac{\lambda_i/\beta}{\alpha + \lambda_i}$$

Luego

$$\frac{\partial}{\partial\beta}log(p(y|\alpha,\beta)) = \frac{N}{2\beta} - \frac{1}{2}\sum_{i=1}^{d}\frac{\lambda_i/\beta}{\alpha+\lambda_i} - \frac{||y - Xm_N||_2^2}{2}$$

Imponemos $\frac{\partial}{\partial\beta}log(p(y|\alpha,\beta)) = 0$

$$0 = \frac{1}{2}(\frac{N}{\beta} - \frac{1}{\beta}\sum_{i=1}^{d}\frac{\lambda_i}{\alpha+\lambda_i} - ||y - Xm_N||_2^2)$$
$$\frac{1}{\beta} = \frac{||y - Xm_N||_2^2}{N-\gamma}$$

3.2. Implementación

3.2.1. Parte 3

Se utiliza la matriz de datos, siendo df_y la columna *"Precio m2"* que indica el precio por metro cuadrado de la vivienda. La matriz df_X son todo el resto de los datos, procesados según corresponda por columna.
Se entrena un modelo de **Regresión Bayesiana Empírica**, que utiliza la metodología descrita en la parte teórica, para la cuál se obtienen los siguientes scores:

$$RMSE = 7184.75$$
$$R^2 = 0.68$$

En general, el modelo es decente, pues entrega predicciones con un RMSE de 7000, frente a viviendas que valen en promedio 31000 por metro cuadrado. Además, explica gran parte de la varianza de los datos, como indica el valor 0.68 de R^2.

3.2.2. Parte 4

Se realiza el mismo procedimiento de la parte anterior, a excepción de que df_X ahora es un DataFrame que contiene solo aquellas columnas seleccionadas a partir del EDA realizado en la Pregunta 2.
El modelo entrega los siguientes scores:

$$RMSE = 8083.61$$
$$R^2 = 0.59$$

Podemos ver que el modelo empeoró, puesto que ambos scores (RMSE y R^2) son peores que antes. Este comportamiento es esperable, pues un modelo con menos información eventualmente entregará peores predicciones.
Sin embargo, nuestra selección de variables parece ser acertada, puesto que se descartaron 11 de ellas y la calidad de las predicciones no empeoró demasiado con respecto a la parte anterior.

3.2.3. Parte 5

A modo de comparación, se entrena un regresor de tipo **BayesianRidge** implementado en el módulo *sklearn.linear_model*.
Los scores para este regresor son:

$$RMSE = 8083.35$$
$$R^2 = 0.59$$

Podemos ver que el rendimiento de este clasificador es prácticamente igual al implementado por nosotros.
Esto es esperable puesto que ambos modelos son casi idénticos, como se aprecia al comparar el desarrollo teórico expuesto en este informe con la documentación[2] de **BayesianRidge**, donde vemos que la única diferencia es que BayesianRidge impone una distribución prior Gamma sobre α y β.

Referencias

[1] Identidad de Sherman Morrison.
https://en.wikipedia.org/wiki/Sherman%E2%80%93Morrison_formula

[2] Documentación de BayesianRidge del módulo sklearn.
https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression

4. Anexo

4.1. Columnas del dataset

En la tabla 2 se detallan las columnas.

Tabla 2: Columnas del dataset.

Columna	**Tipo**	**Significado**
price	float64	Precio
`n_rooms`	float64	Número de habitaciones
`n_bath`	float64	Número de baños
surface	float64	Área de la vivienda
details	object	Detalles de la vivienda
url	object	Url de la página
`metrocuadrado_index`	float64	Índice de la página
Furnished	object	Indica si proviene de `furnished`
location	object	Barrio de la vivienda
property	object	Tipo de propiedad
rent	object	Tipo de renta
Precio m2	float64	Precio por m^2
Garajes	int32	Cantidad de garajes
tipo de producto	int32	Tipo en base a clasificación
CodigoUPZ	object	Código UPZ del barrio
personas	float64	N° personas en el barrio
`trabajo_ninos`	float64	% de niños que trabajan
`trabajoamp_ninos`	float64	% de niños que trabajan ampliado
`jovenes_nini`	float64	% de jóvenes que no trabajan ni estudian
`indice_envegecimiento`	float64	Cuociente de adultos mayores entre niños
`jefe_mujer`	float64	% de mujeres jefes de hogar
`adultos_mayores_pobres`	float64	% de adultos mayores pobres
`indice_inseguridad`	float64	% de lugares inseguros del barrio
`areas_verdes_perc`	float64	% de áreas verdes
`pob_density`	float64	Densidad de población