SEGUNDA ENTREGA DE PROYECTO

POR:

Danilo Tovar Arias

Daniel Rosas Mendoza

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raúl Ramos Pollán

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERIA

MEDELLIN 2023

1. EXPLORACIÓN DE DATOS

Para iniciar, realizamos una exploración inicial del dataset, identificando las columnas de interés para el desarrollo.

1.1. Análisis de la variable objetivo

Se construyó una gráfica que nos permite observar la distribución que tiene la variable objetivo, como se muestra en la *Figura 1*, donde se puede apreciar que posee una asimetría muy baja hacia la izquierda. Sin embargo, posee un comportamiento adecuado para la realización del entrenamiento y las pruebas de algoritmo.

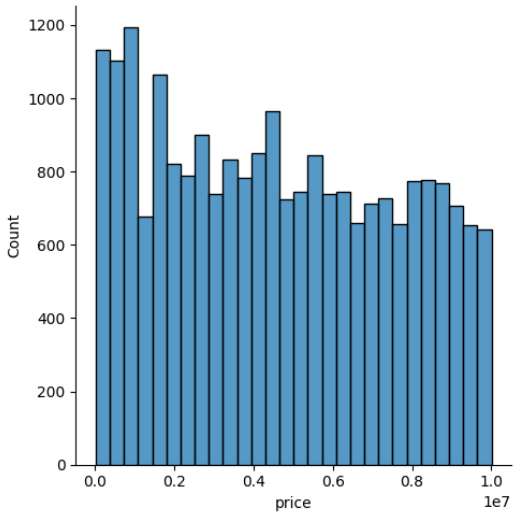

Figura 1. Distribución de la variable objetivo

1.2. Secciones de la ciudad

Entre las variables de interés se identificó *cityPartRange*, que nos dice la sección de la ciudad donde se encuentra ubicado el edificio. A esta variable se le realizó un análisis de frecuencia y de precio promedio por cada valor único, como se puede observar en las *Figura 2* y *Figura 3*, respectivamente.

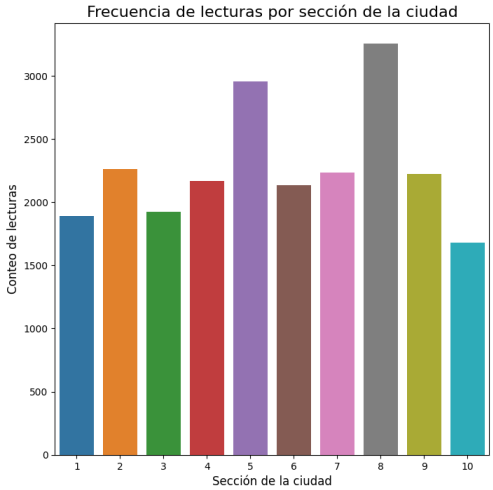

Figura 2. Frecuencia de los valores para la variable cityPartRange.

Se pudo observar que la sección 8 es aquella a la que pertenece el mayor porcentaje de edificios, mientras que la sección 10 se aquel al que pertenece la menor cantidad.

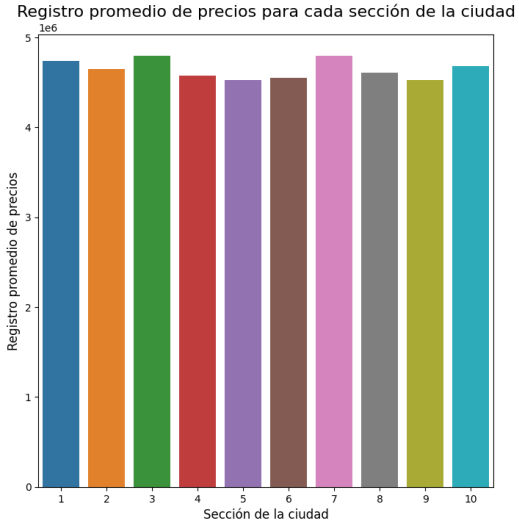

Figura 3. Promedio del precio por cada valor de la variable cityPartRange.

Se pudo observar que el promedio de la variable objetivo es relativamente constante entre las diferentes secciones de la ciudad, donde la sección 3 con el mayor valor se separa de la sección 9, correspondiente al menor, aproximadamente por un 5,6% (0.270909e+06).

1.3. Año de construcción

Otra de las variables de interés identificadas fue *made*, que nos dice el año de construcción del edificio. A esta variable se le realizó el

mismo análisis que a la variable anterior. Los resultados se pueden observar en la *Figura 4* y *Figura 5*.

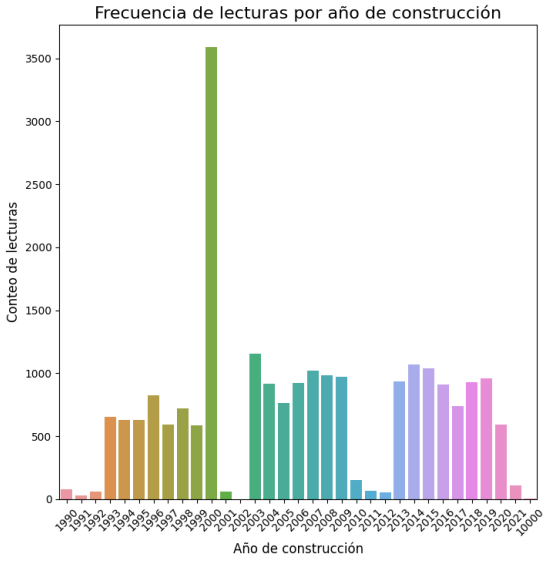

Figura 4. Frecuencia de los valores para la variable made.

Se pudo observar que el año 2000 es aquel en el que fueron construidos el mayor porcentaje de edificios, mientras que el año 2002 es aquel en el que fueron construidos la menor cantidad.

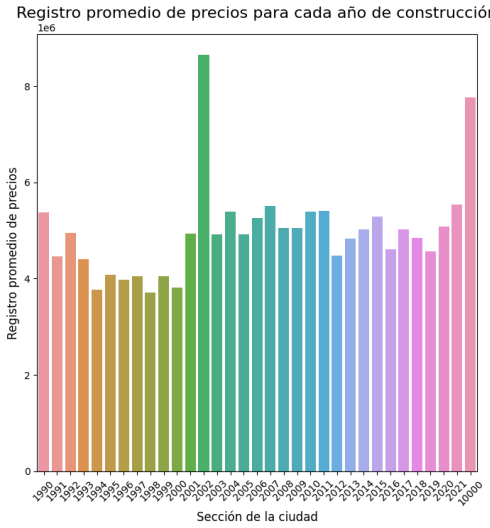

Figura 5. Promedio del precio por cada valor de la variable made.

Se pudo observar que el promedio de la variable objetivo es relativamente constante entre los diferentes años de construcción, a excepción del año 2000 y 10000.

1.4. Datos faltantes

Un elemento importante para el análisis son los datos faltantes. Luego de realizar un recorrido y conteo a través del database se logró comprobar que no existen datos faltantes dentro del mismo. Sin embargo, por motivos de aprendizaje y cumplir con los requerimientos del proyecto, se realizó una simulación para generar datos faltantes dentro del database la cual será explicada en la sección de preprocesado.

1.5. Correlación entre las diferentes variables

Otro factor importante que se tuvo en cuenta para el análisis fue la correlación entre las diferentes variables con la variable objetivo. A través de la *tabla 1* se puede observar que la variable squareMeters es aquella que posee la mayor correlación.

price	
price	1.000000
squareMeters	0.591749
numberOfRooms	0.091681
floors	0.038374
made	0.024270
cityCode	0.021986
hasStormProtector	0.020512
isNewBuilt	0.008080
hasPool	0.006023
hasStorageRoom	0.001567
hasYard	-0.002545
attic	-0.006851
id	-0.008060
numPrevOwners	-0.008546
hasGuestRoom	-0.009309
cityPartRange	-0.009366
basement	-0.034940
garage	-0.120137

Tabla 1. Correlación entre variable con price.

1.6. Distribución de las variables numéricas

Se realizó una muestra de las distribuciones de cada una de las variables. Sin embargo, no se observaron problemas relevantes dentro de las mismas.

2. TRATAMIENTO DE DATOS

2.1. Simulación de datos faltantes

Se tomaron aleatoriamente 3 columnas diferentes a id y precio, a las cuales se les aplicó una eliminación aleatoria de datos, de tal manera que entre 5 a 70% de los datos para dichas columnas serían ahora datos faltantes.

Este nuevo dataset será utilizado como el nuevo train, donde se realizará el adecuado tratamiento de datos para su posterior aplicación en los diferentes algoritmos.