Análisis Exploratorio

Camilo Vega

Este documento contiene un análisis exploratorio de los datos contenidos en el dataframe vivienda4 del paquete paqueteMET, los cuales corresponden a los precios de vivienda en millones de pesos colombianos por zona, estrato, área construida y tipo de vivienda. El objetivo de este análisis es evaluar la calidad de los datos y detectar posibles relaciones entre las variables, con el fin de desarrollar un modelo de regresión lineal que explique de la mejor manera posible el precio en millones de pesos colombianos.

Carga de librerias y funciones personalizadas

En este análisis, se utilizan las librería listadas en el siguiente código. Además, para facilitar la creación de visualizaciones, se emplean lasfunciones personalizadas gg_rain_cloud, gg_bar y summary_table, las cuales se encuentran en el archivo funciones_personalizadas.R.

```
# Carga de paquetes necesarios para el código
library(tidyverse) # Conjunto de paquetes para manipulación de datos
library(ggside) # Extiende ggplot2 con gráficos adicionales
library(GGally) # Extiende ggplot2 con gráficos de matriz
library(ggdist) # Extiende ggplot2 con gráficos de distribución
library(tidyquant) # Paquete de finanzas para análisis cuantitativo de datos
library(paqueteMET) # Paquete para el análisis de series temporales
library(skimr) # Paquete para el análisis exploratorio de datos
library(knitr) # Paquete para creación de tablas en formato de salida

# Se crea un objeto "datos_vivienda" que contiene los datos de vivienda4
datos_vivienda <- vivienda4

# Se cargan funciones personalizadas
source("funciones_personalizadas.R")
```

Calidad de los datos

Con el fin de identificar posibles datos faltantes y obtener una descripción general de los datos, se utilizará la función skim.

```
datos_vivienda |>
    skim()
```

Table 1: Data summary

Name	datos_vivienda
Number of rows	1706
Number of columns	5
Column type frequency:	
factor	3
numeric	2
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
zona	0	1	FALSE	5	Zon: 1344, Zon: 288, Zon: 60, Zon: 8
estrato	0	1	FALSE	1	4: 1706, 3: 0, 5: 0, 6: 0
tipo	0	1	FALSE	2	Apa: 1363, Cas: 343

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
preciom	0	1	225.37	85.89	78	160	210	265	760	
areaconst	0	1	87.63	36.35	40	60	75	98	200	

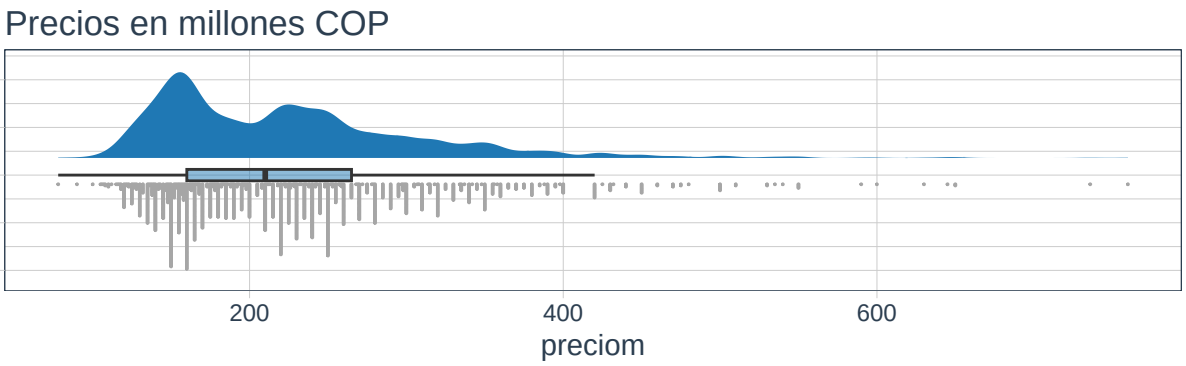
Se observa que el dataframe vivienda4 cuenta con 1706 observaciones y 5 variables, de las cuales 2 son numéricas y 3 son categóricas. No se han encontrado datos faltantes en el conjunto de datos. Es importante mencionar que la variable estrato solo presenta un valor, el cual es 4. Por lo tanto, se realizará un análisis y modelo únicamente para viviendas de estrato 4, y se tomará en cuenta esta acotación en los análisis posteriores.

Analisis univariado.

A continuation se realizará un analisis univariado de las distintas variables del `vivienda4`.

preciom

```
gg_rain_cloud(datos_vivienda, preciom) +
    ggtitle(
        "Precios en millones COP"
        ); summary_table(datos_vivienda, preciom)
```



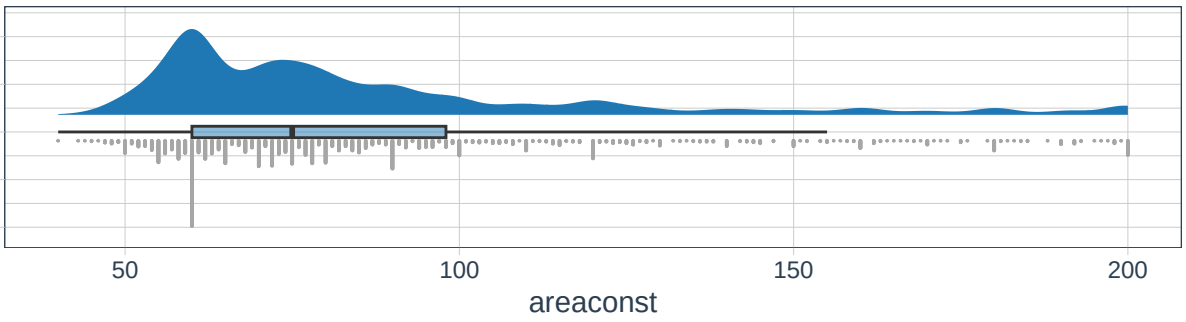
min	q1	median	mean	q3	max	skewness
78	160	210	225.37	265	760	1.49

Se puede observar que la distribución presenta una fuerte asimetría positiva, lo cual indica que la mayoría de las viviendas tienen precios bajos y hay algunas viviendas con precios muy altos. Además, se puede apreciar una bimodalidad en la distribución, lo cual sugiere la presencia de variables categóricas que influyen en los precios de las viviendas.

---

areaconst

```
gg_rain_cloud(datos_vivienda, areaconst) +
    ggtitle(
        "Área construida metros cuadrados"
        ) ; summary_table(datos_vivienda, areaconst)
```

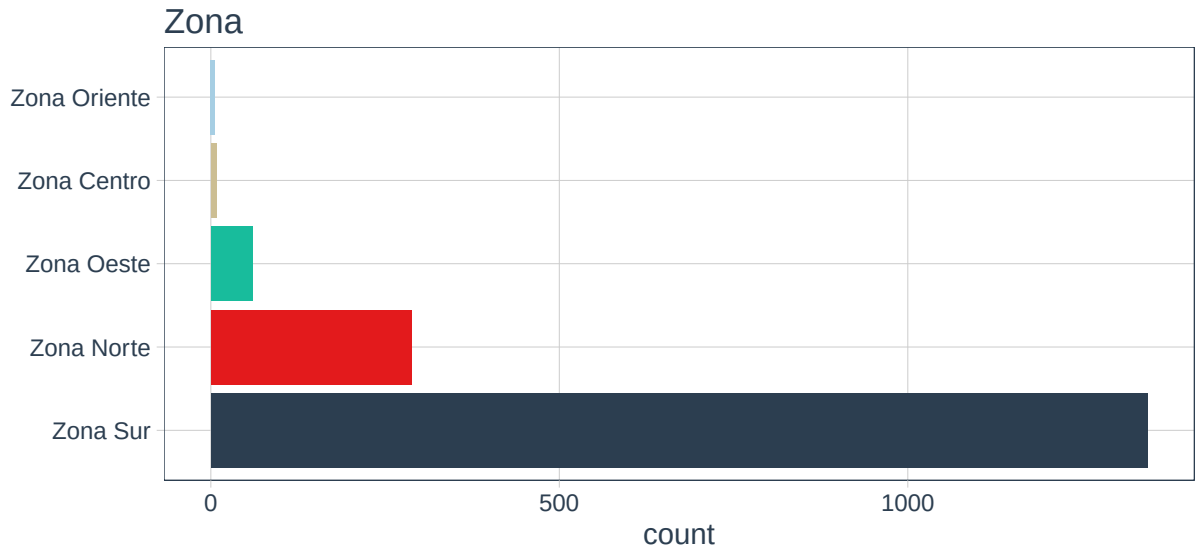


min	q1	median	mean	q3	max	skewness
40	60	75	87.63	98	200	1.53

Al igual que el precio, la variable correspondiente al área construida presenta una distribución con fuerte asimetría positiva, lo cual indica que la mayoría de las viviendas tienen áreas bajas y hay algunas viviendas con áreas altos. Además, se puede apreciar una bimodalidad en la distribución, lo cual sugiere la presencia de variables categóricas que influyen en lás áreas de las viviendas.

zona

```
gg_bar(datos_vivienda, zona) +
    ggtitle("Zona"); table(datos_vivienda$zona) |>
    kable()
```

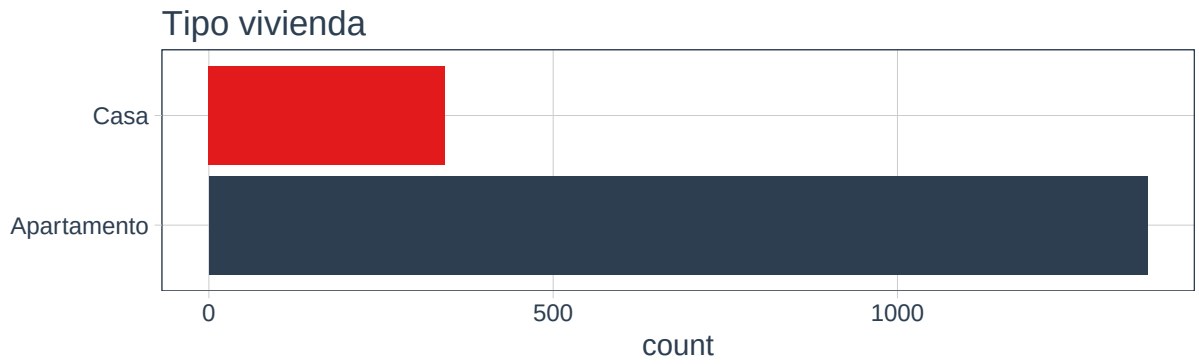


Var1	Freq
Zona Centro	8
Zona Norte	288
Zona Oeste	60
Zona Oriente	6
Zona Sur	1344

Se puede observar una fuerte predominancia de observaciones correspondientes a la zona sur, la cual representa aproximadamente el 79% del conjunto de datos.

tipo

```
gg_bar(datos_vivienda, tipo)  +
    ggtitle("Tipo vivienda"); table(datos_vivienda$tipo) |>
    kable()
```



Var1	Freq
Apartamento	1363
Casa	343

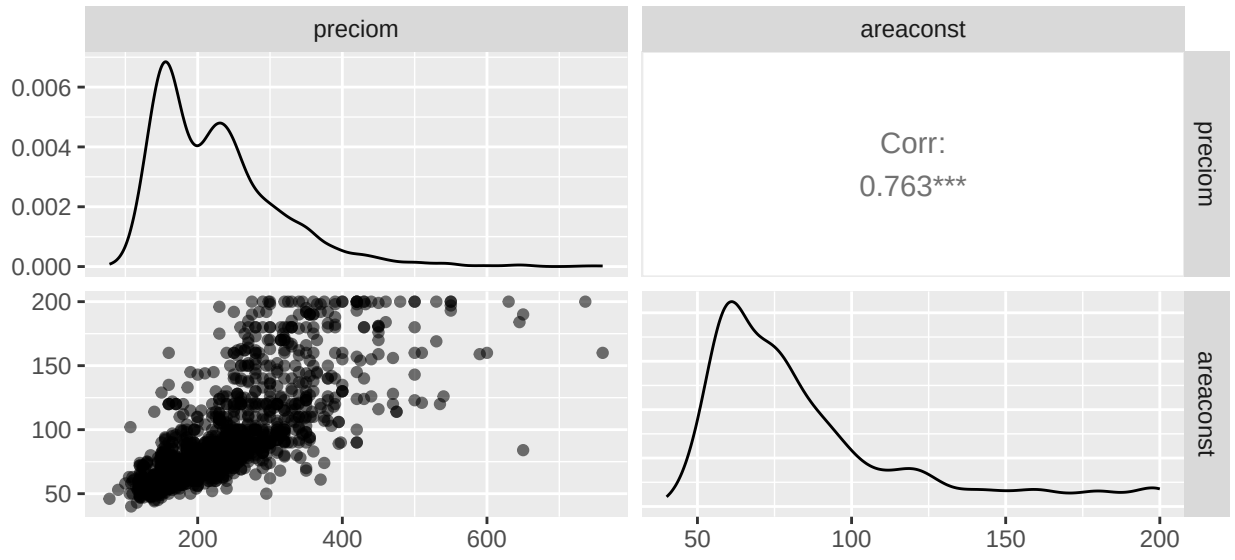
Se puede observar una fuerte predominancia de observaciones correspondientes a apartamentos, los cuales representan aproximadamente el 80% del conjunto de datos.

Análisis multi-variado

A continuación, se realizará un análisis multivariado enfocado en la relación entre las variables numéricas preciom y areaconst.

preciom vs. areaconst

```
ggpairs(datos_vivienda, columns = c(3,4), aes(alpha = 0.5))
```



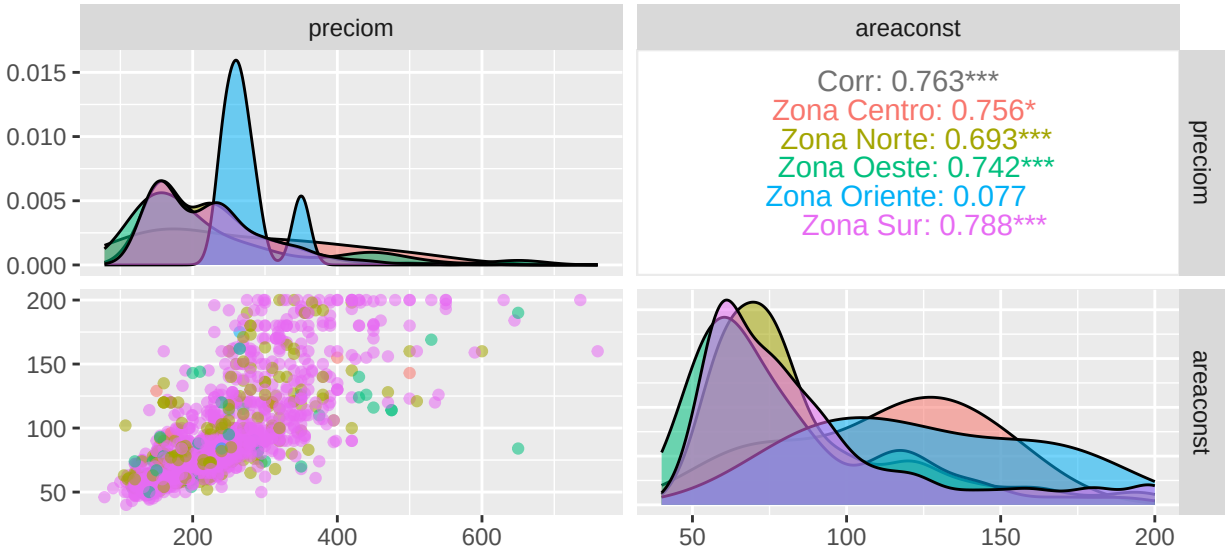
Se puede observar una correlación positiva entre las variables preciom y areaconst, la cual se vuelve cada vez más dispersa a medida que ambas variables aumentan.

En el análisis univariado exploratorio, se observó que tanto preciom como areaconst presentaban bimodalidad. Por lo tanto, a continuación se repetirá el análisis agrupando por zona y tipo para analizar cómo estas dos variables afectan la relación entre preciom y areaconst.

---

preciom vs. areaconst por zona

```
ggpairs(datos_vivienda, columns = c(3,4), aes(colour = zona, alpha = 0.5))
```



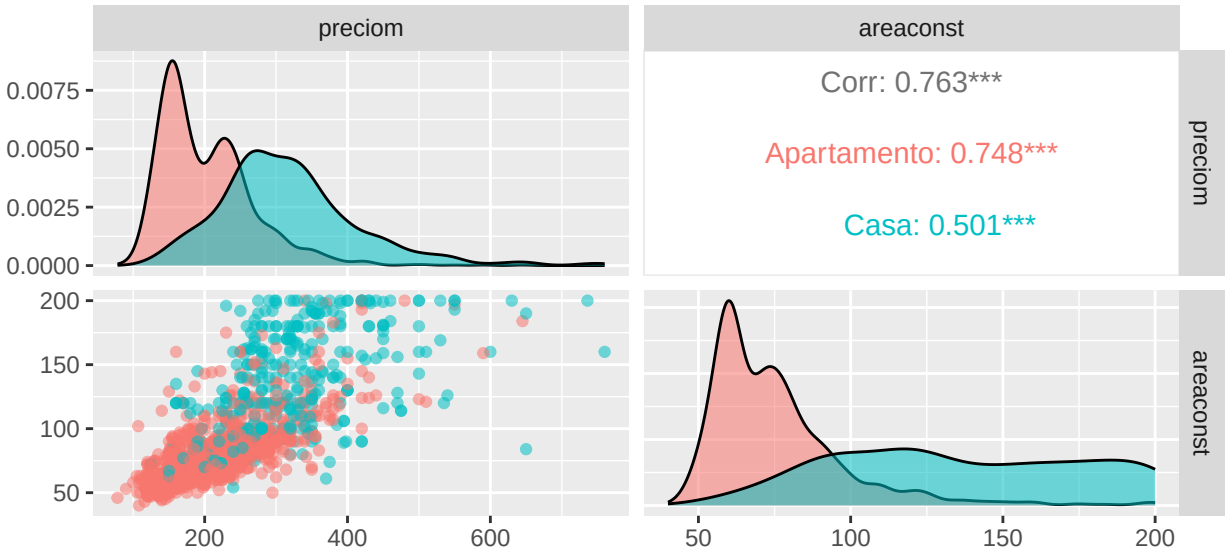
Es evidente que la variable zona contribuye a explicar parte de la multimodalidad de preciom y areaconst, siendo la Zona Sur la que presenta una mayor correlación entre ambas variables.

En el gráfico de dispersión se puede apreciar cómo la relación entre preciom y areaconst es más fuerte en los valores inferiores de ambas variables para la Zona Sur, y se vuelve más dispersa a medida que estas crecen.

---

preciom vs. areaconst por tipo

```
ggpairs(datos_vivienda, columns = c(3,4), aes(colour = tipo, alpha = 0.7))
```



Es evidente que la variable tipo contribuye a explicar parte de la multimodalidad de preciom y areaconst, siendo los apartamentos los que presentan una mayor correlación entre ambas variables.

En el gráfico de dispersión se puede apreciar cómo la relación entre preciom y areaconst es más clara para los apartamentos que para las casas.

Conclusiones y Recomendaciones.

Se concluye de este estudio exploratorio de datos que existe una relación entre las variables preciom y areaconst, lo que sugiere que se puede avanzar en la construcción de un modelo lineal que explique preciom a partir de areaconst. Sin embargo, se recomienda tener en cuenta las siguientes recomendaciones para mejorar la calidad del modelo:

Se sugiere filtrar los datos únicamente para la combinación de variables "Zona Sur" y "Tipo Apartamentos" si se desea construir un modelo lineal simple, ya que estas categorías demostraron tener la mayor correlación en la relación entre preciom y areaconst. Además, esta combinación de variables tiene la mayor proporción dentro del conjunto de datos, como se puede observar en las tablas de conteo y proporciones presentadas a continuación:.

	Apartamento	Casa
Zona Centro	7	1
Zona Norte	237	51
Zona Oeste	52	8
Zona Oriente	2	4
Zona Sur	1065	279

	Apartamento	Casa
Zona Centro	0.0041	0.0006
Zona Norte	0.1389	0.0299
Zona Oeste	0.0305	0.0047

	Apartamento	Casa
Zona Oriente	0.0012	0.0023
Zona Sur	0.6243	0.1635

Sin embargo, es importante tener en cuenta que al filtrar los datos de esta manera, el modelo solo será válido para inferir los precios de apartamentos de la zona sur. Aunque esta limitación puede reducir la generalidad del modelo, crear un modelo lineal simple con datos específicos puede reducir el ruido y mejorar la consistencia del modelo. Por lo tanto, la selección de estas variables específicas se justifica para cumplir con el objetivo de crear un modelo lineal simple con los datos proporcionados.

Como alternativa, se sugiere crear un modelo lineal multivariado que explique la variable preciom a partir de areaconst, tipo y zona. Se recomienda excluir las zonas Centro, Oriente y Oeste debido a la baja cantidad de observaciones, lo que podría afectar la calidad del modelo.

Por último, se recomienda la transformación (aproximando a una distribución normal) y eliminación de valores extremos de las variables preciom. Esto permitirá mejorar la calidad de los modelos creados, independientemente de si se utiliza un modelo lineal simple o uno multivariado.