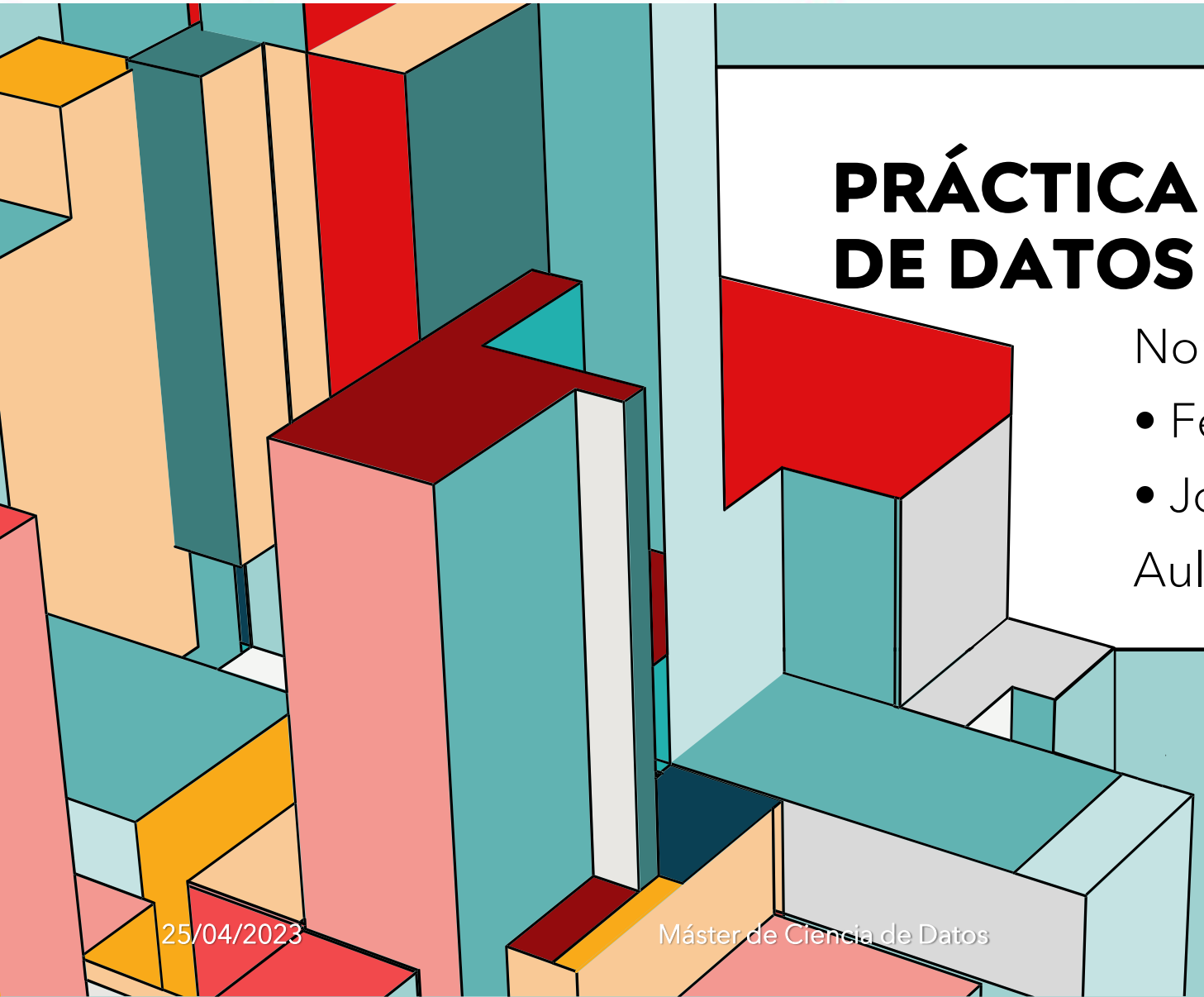


PRÁCTICA 2: PROCESAMIENTO DE DATOS (DATASET)

Nombres de los estudiantes:

- Félix Antonio Mucha Morales
- José Carlos Enriquez Lira

Aula 2

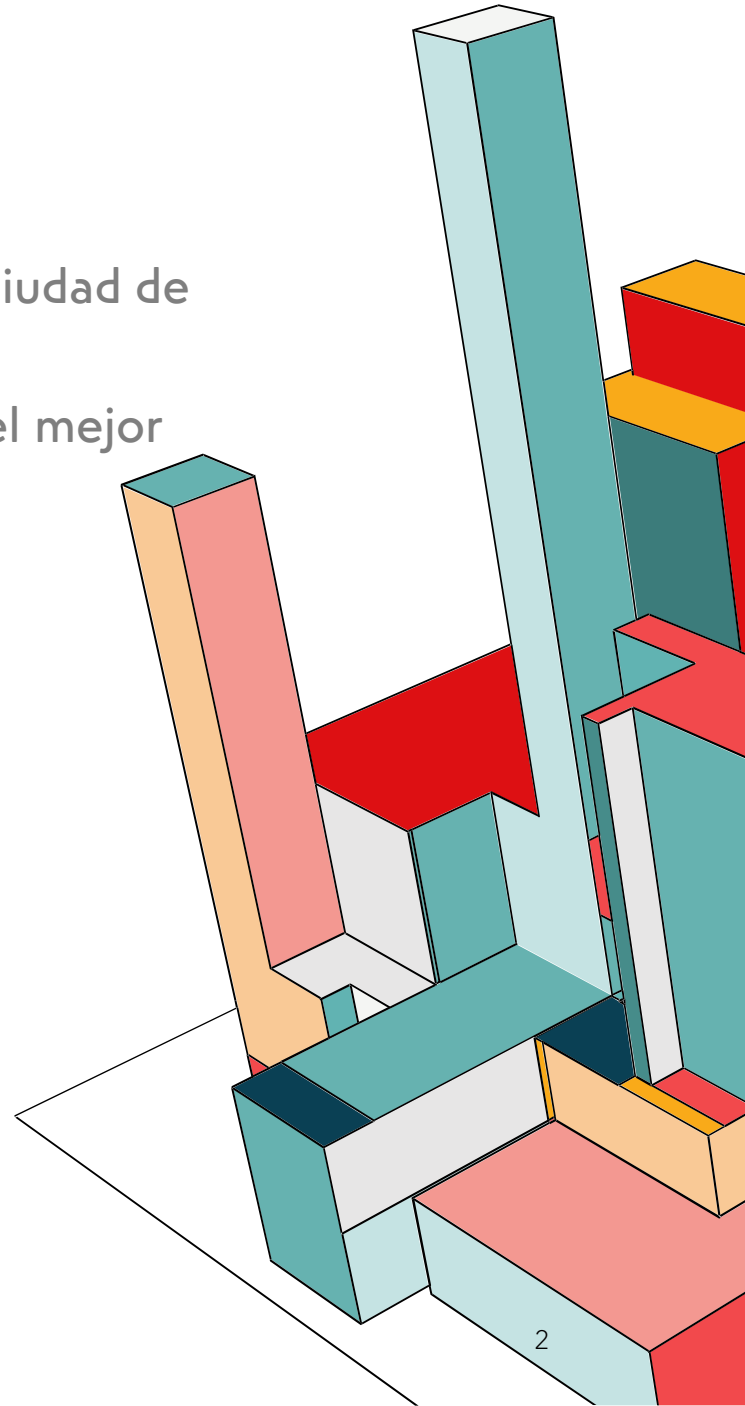
1. DESCRIPCIÓN DEL DATASET

El dataset contiene información de venta de departamentos de la ciudad de Buenos Aires, dicha información nos va a permitir realizar análisis, investigación, predicción y toma de decisiones para la selección del mejor inmueble

Las variables del dataset son:

- created_on: fecha de publicación de aviso
- operation: operación de venta
- property_type: tipo de propiedad
- place_with_parent_names: lugar de la propiedad
- lat.lon: Latitud y longitud
- lat: Latitud
- lon: Longitud
- price_aprox_usd: Precio aprox en usd
- surface_total_in_m2: superficie_total_en_m2
- surface_covered_in_m2: superficie cubierta en m2
- price_usd_per_m2: precio usd por m2
- floor: piso
- rooms: habitaciones
- expenses: gastos
- properati_url: URL_propiedad
- barrio: barrio
- comuna: comuna

Origende datos Kaggle (https://www.kaggle.com/datasets/gastonmichelotti/properati-data-set).

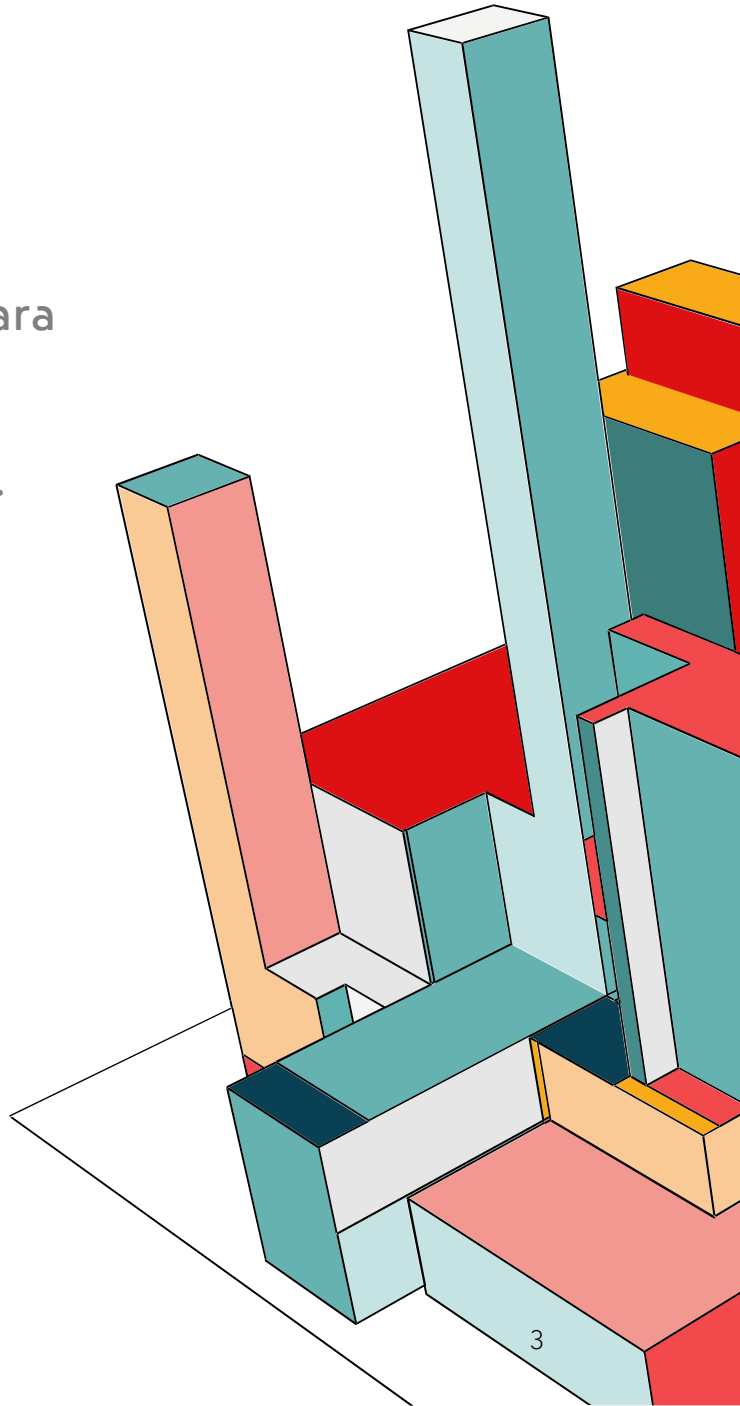


2. OBJETIVOS

- Determinar si el modelo de regresión lineal múltiple es el adecuado para predecir los precios aproximados de los inmuebles.
- Determinar los puntos geográficos de los inmuebles de mayor interés.

3. INTEGRACIÓN Y SELECCIÓN

created_on	property_type	lat.lon
"Date"	"factor"	"character"
lat	lon	price_aprox_usd
"numeric"	"numeric"	"numeric"
surface_total_in_m2	surface_covered_in_m2	price_usd_per_m2
"numeric"	"numeric"	"numeric"
floor	rooms	expenses
"numeric"	"integer"	"numeric"
properati_url	barrio	comuna
"character"	"factor"	"factor"



4. LIMPIEZA DE LOS DATOS

a)

```
# Filtramos solo departamentos
Dpto <- filter(datos, property_type == 'apartment')

porc_null_dpto <- calcular_porcentaje_nulos(Dpto)
porc_null_dpto
```

```
##   created_on property_type lat.lon lat lon price_aprox_usd surface_total_in_m2
## 1          0             0       0   0   0        7.293848             10.17762
##   surface_covered_in_m2 price_usd_per_m2    floor    rooms expenses
## 1              9.731884         12.96684 83.24441 22.48936 75.93706
```

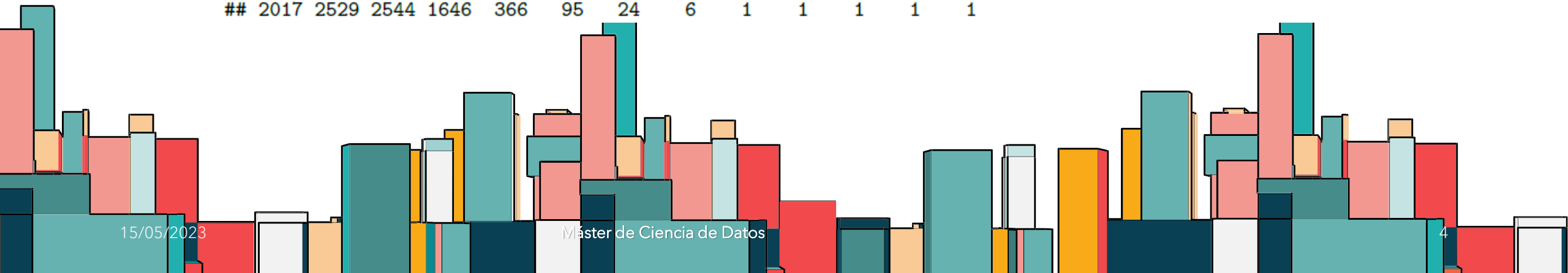
b)

```
# Validación del requisito
dfDptoNN$val <- dfDptoNN$surface_total_in_m2 >= dfDptoNN$surface_covered_in_m2
dfDptoNN <- dfDptoNN[dfDptoNN$val == 'TRUE',]
dfDptoNN <- select(dfDptoNN, -val)
```

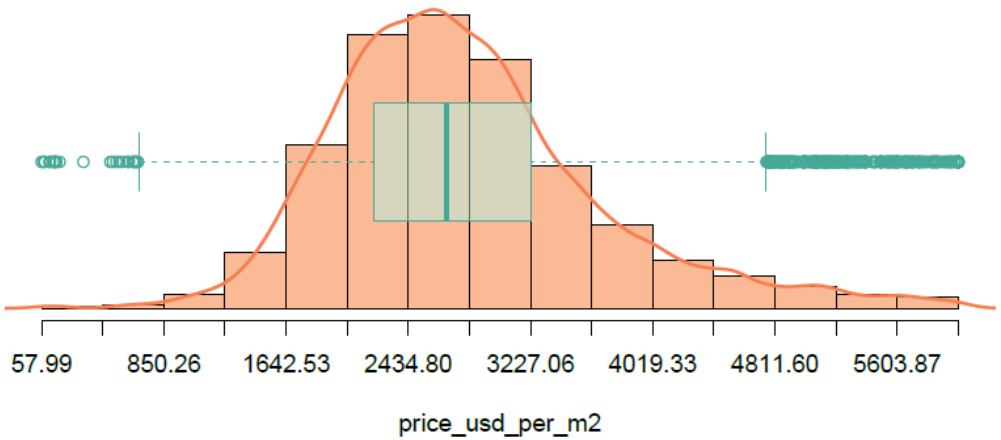
c)

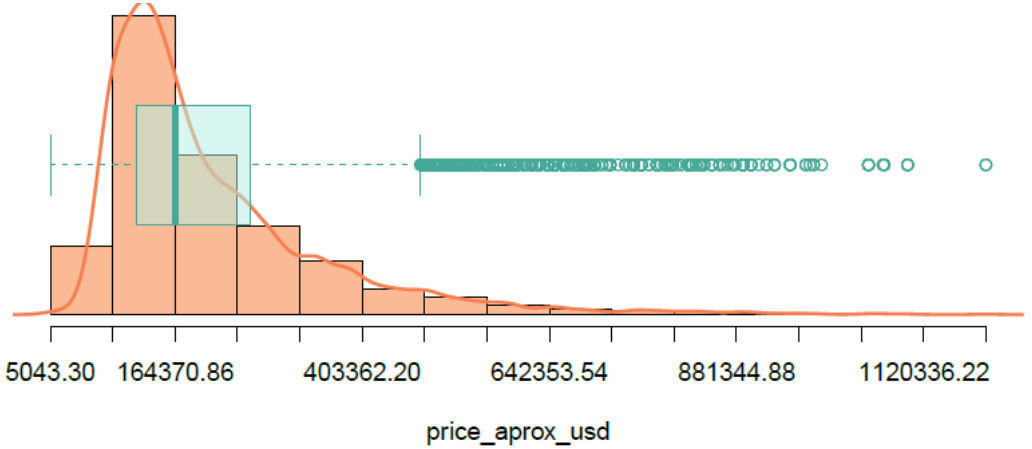
```
# Filtro de datos
dfDptoN <- dfDptoNN[(dfDptoNN$surface_total_in_m2 <= 200) & (dfDptoNN$price_usd_per_m2 <= 6000),]
```

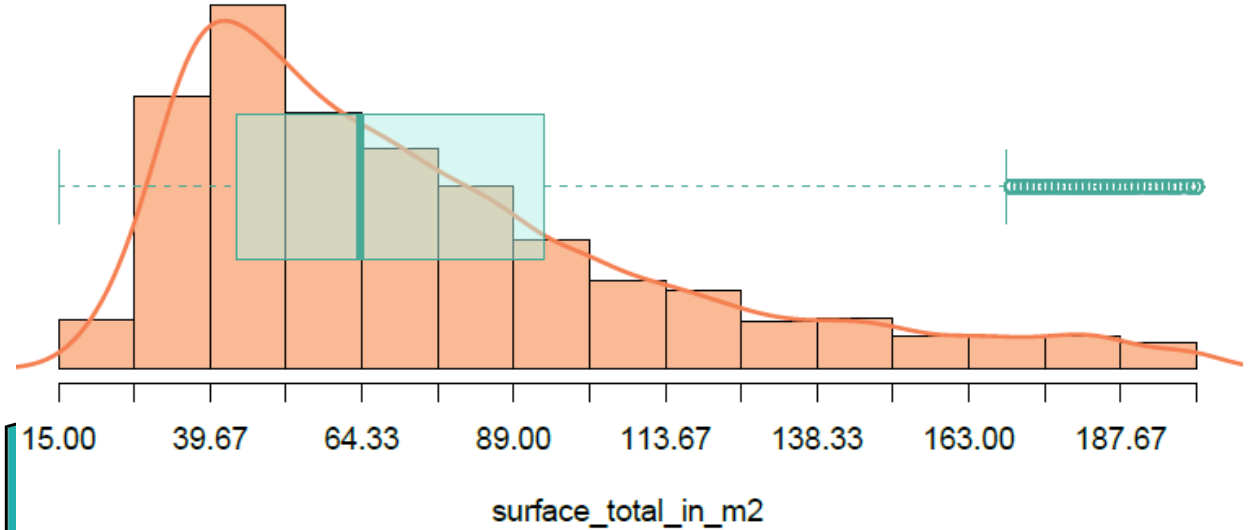
```
##    1    2    3    4    5    6    7    8    9   10   11   13   30
## 2017 2529 2544 1646  366   95   24    6    1    1    1    1    1
```



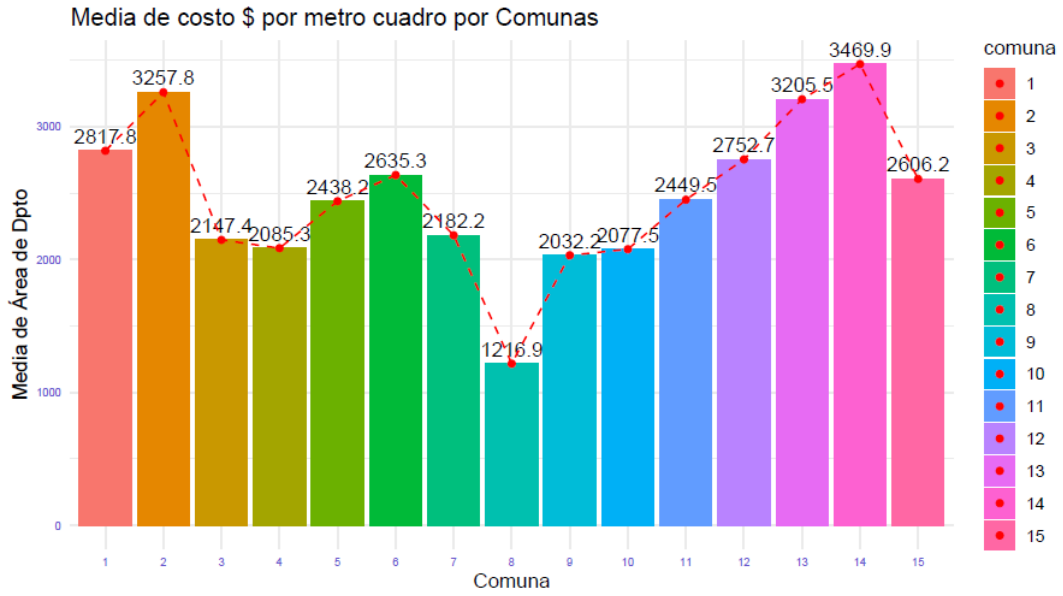
4. LIMPIEZA DE LOS DATOS

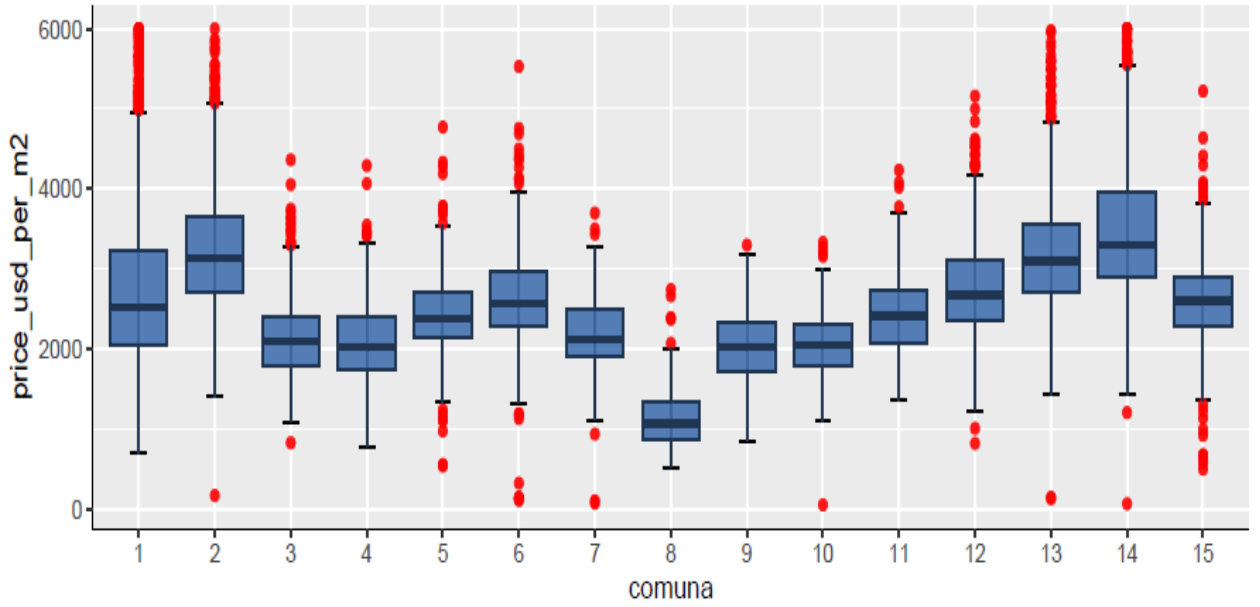






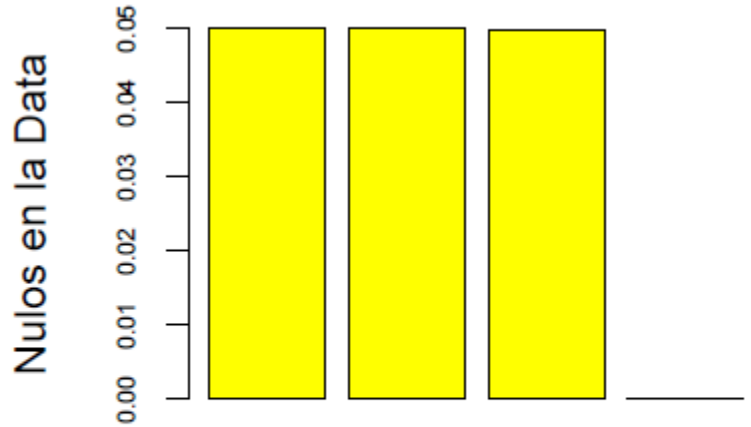
4. LIMPIEZA DE LOS DATOS

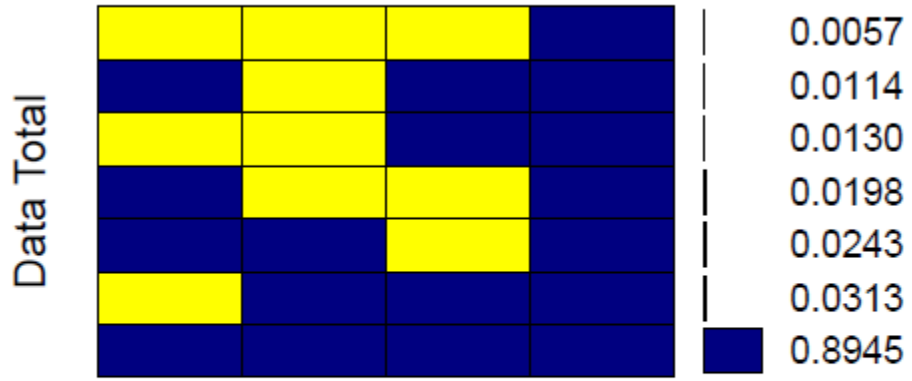


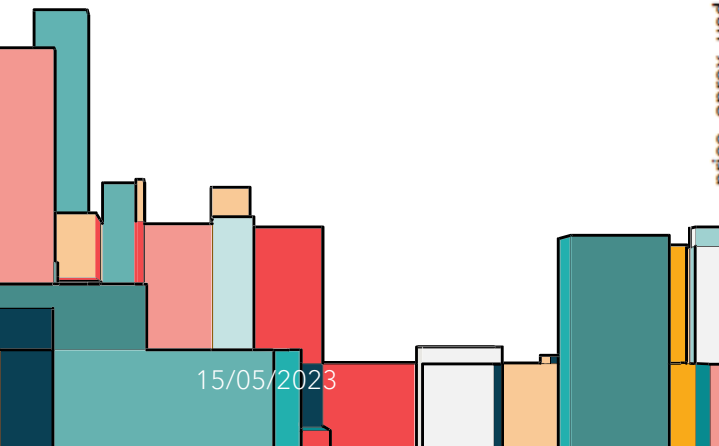


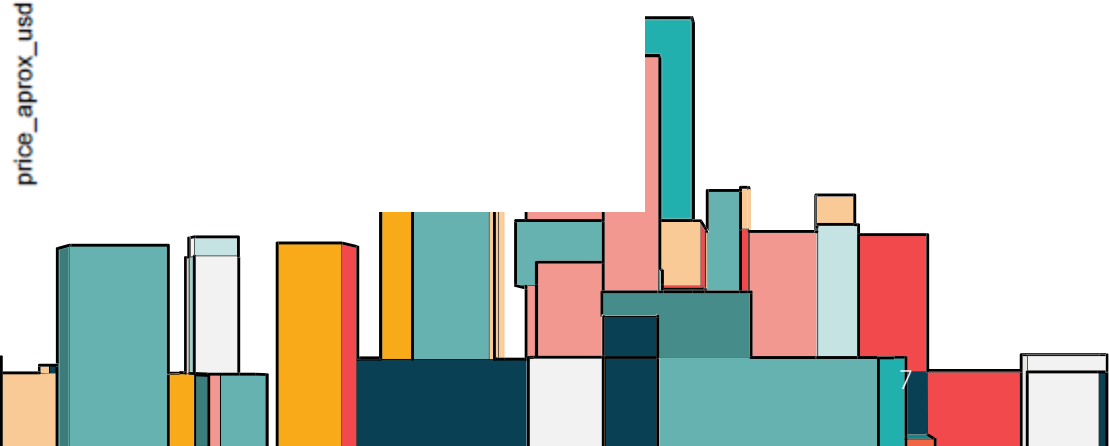
4. LIMPIEZA DE LOS DATOS

- price_aprox_usd
- surface_total_in_m2
- price_usd_per_m2

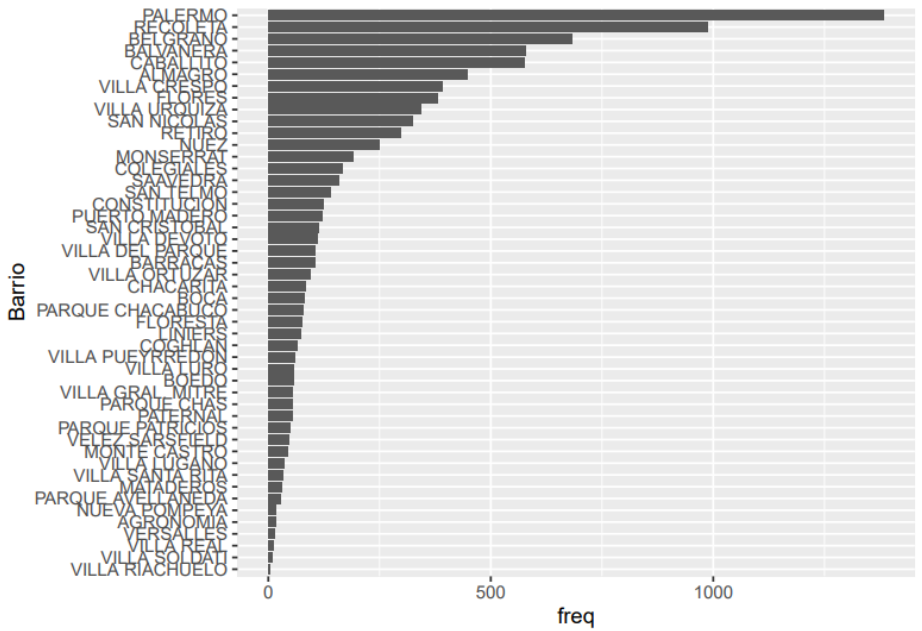


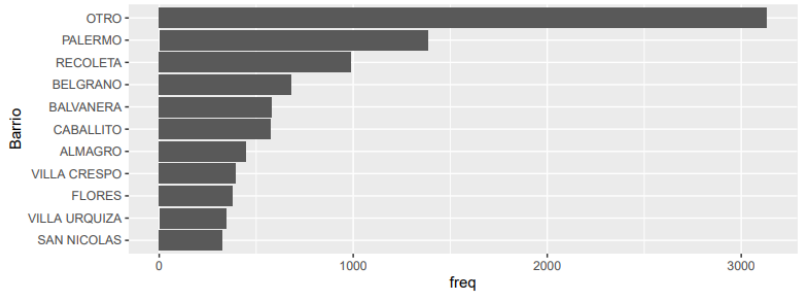


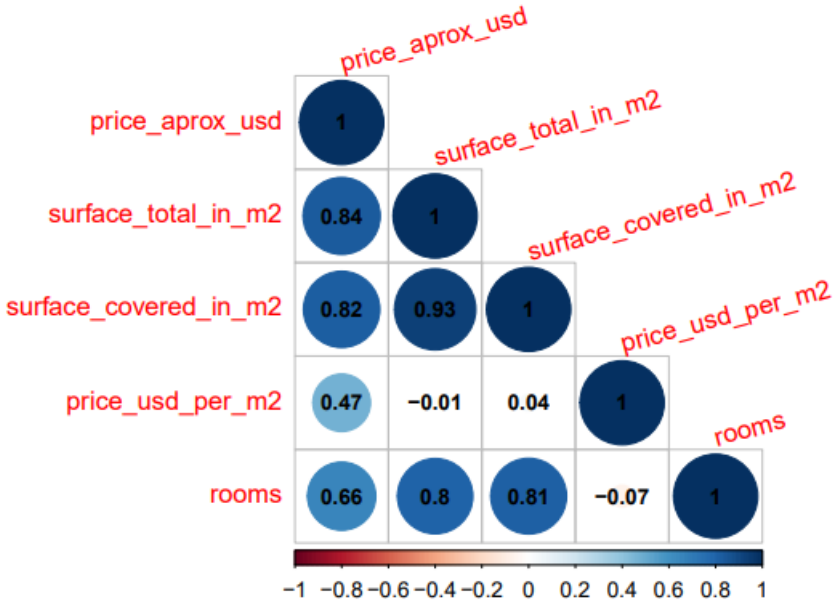




5. ANÁLISIS








```
# Test de independencia
tabla <- table(df_model_pre$n_barrio, df_model_pre$n_comuna)
chisq.test(tabla)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 53654, df = 80, p-value < 2.2e-16
```

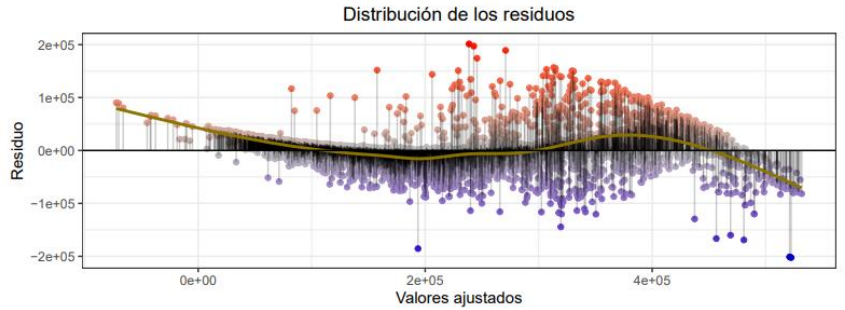
Se retira la variable habitaciones y la superficie cubierta (m2)

Se retira la variable comuna

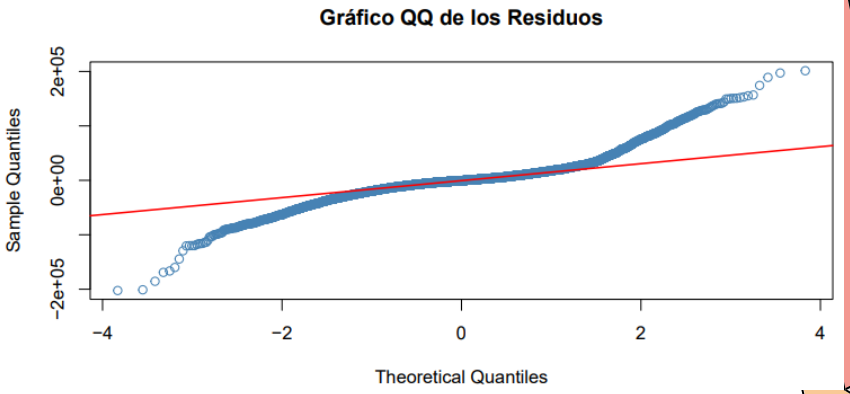
5.1 MODELAMIENTO

```
ModelF <- lm(price_aprox_usd ~ surface_total_in_m2 + price_usd_per_m2 + n_barrio ,
summary(ModelF)
```

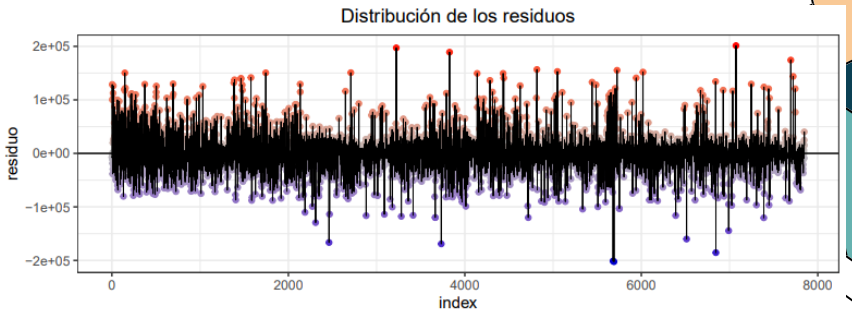
```
##
## Call:
## lm(formula = price_aprox_usd ~ surface_total_in_m2 + price_usd_per_m2 +
##     n_barrio, data = dftrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -202414  -11068    -598    9951  201360
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.991e+05  2.067e+03 -96.316  < 2e-16 ***
## surface_total_in_m2    2.706e+03  9.413e+00 287.525  < 2e-16 ***
## price_usd_per_m2       7.185e+01  5.080e-01 141.447  < 2e-16 ***
## n_barrioBALVANERA      9.961e+02  1.973e+03   0.505 0.613761
## n_barrioBELGRANO       6.362e+03  1.926e+03   3.303 0.000960 ***
## n_barrioCABALLITO      3.191e+03  1.957e+03   1.631 0.102935
## n_barrioFLORES         3.828e+03  2.181e+03   1.755 0.079267 .
## n_barrioOTRO           7.001e+03  1.565e+03   4.474 7.77e-06 ***
## n_barrioPALERMO        6.087e+03  1.746e+03   3.486 0.000493 ***
## n_barrioRECOLETA       8.636e+03  1.812e+03   4.767 1.91e-06 ***
## n_barrioSAN NICOLAS    2.537e+03  2.258e+03   1.124 0.261080
## n_barrioVILLA CRESPO   4.380e+03  2.145e+03   2.042 0.041213 *
## n_barrioVILLA URQUIZA  2.335e+03  2.247e+03   1.039 0.298924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28780 on 7835 degrees of freedom
## Multiple R-squared:  0.9357, Adjusted R-squared:  0.9356
## F-statistic:  9499 on 12 and 7835 DF,  p-value: < 2.2e-16
```




```
##
##  studentized Breusch-Pagan test
##
## data:  ModelF
## BP = 707, df = 12, p-value < 2.2e-16
```

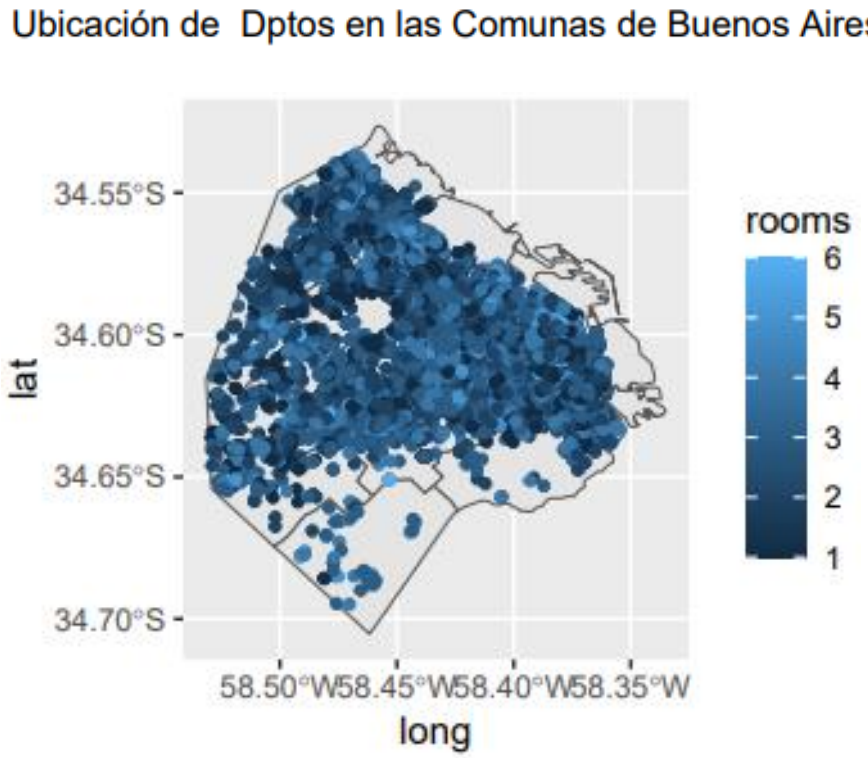


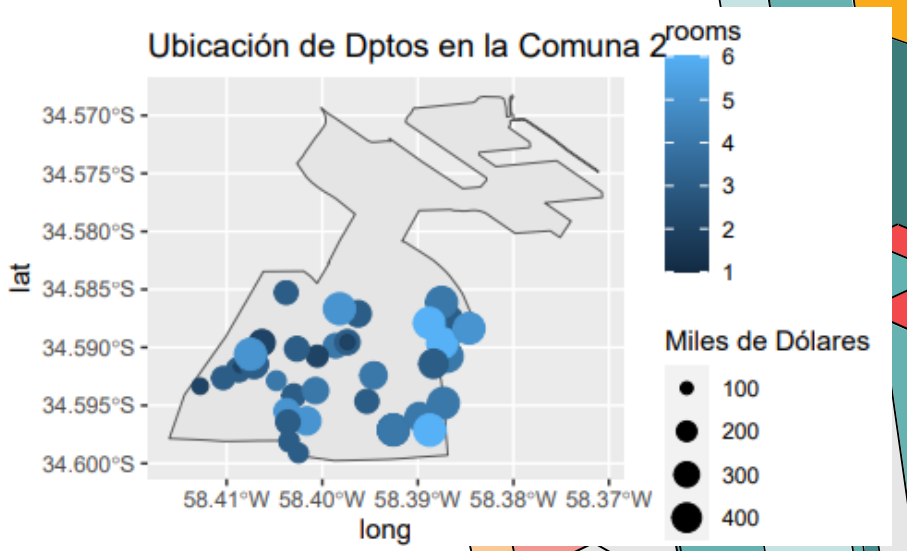

```
##
##  Anderson-Darling normality test
##
## data:  residuo
## A = 250.7, p-value < 2.2e-16
```




```
##  Durbin-Watson test
##
## data:  ModelF
## DW = 1.8929, p-value = 5.81e-07
```

6. REPRESENTACIÓN DE RESULTADOS

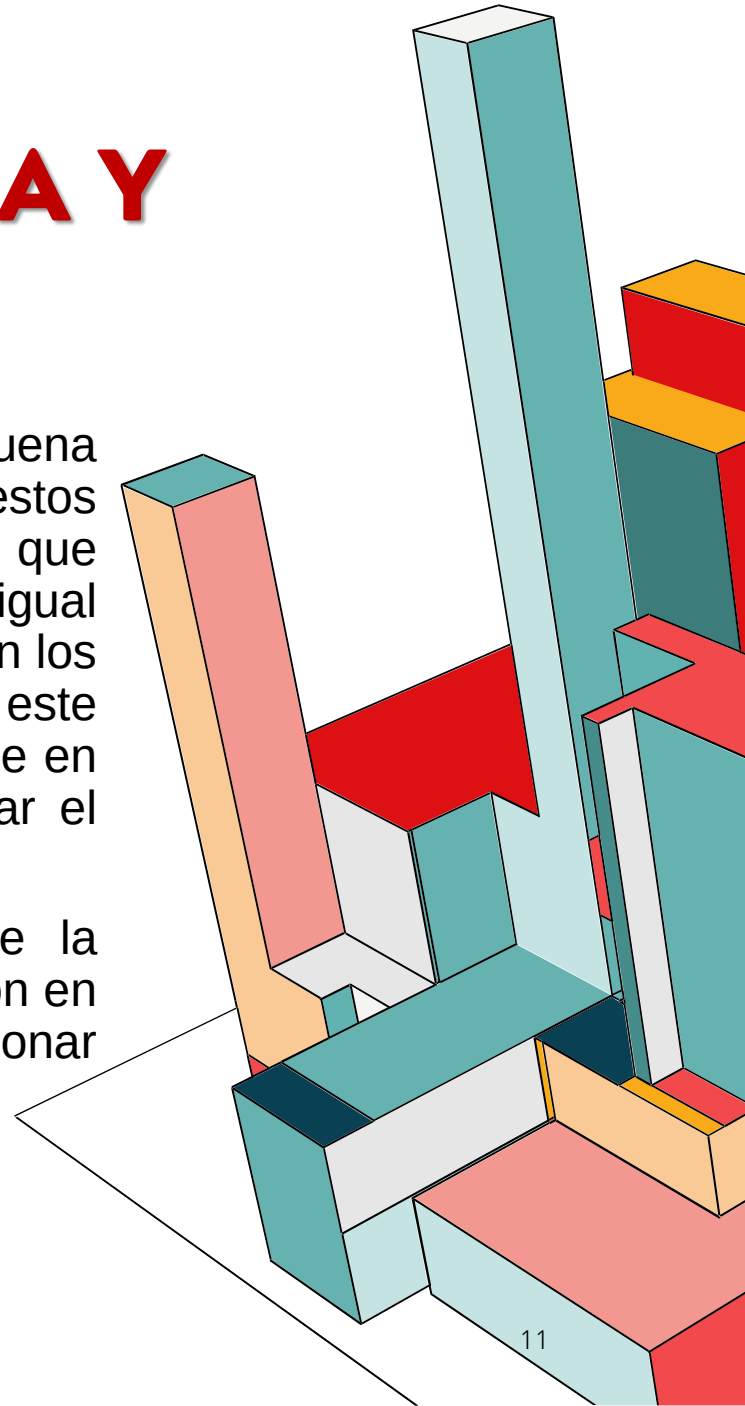




7. RESOLUCIÓN DEL PROBLEMA Y CONCLUSIONES

El modelo de regresión lineal múltiple no es considerada como una buena opción en este proyecto. Esto se debe a que no cumple con los supuestos necesarios para poder tener una consistencia en el modelo. Es así que vez que los errores no se ajustan a una distribución normal ($p - value < 0.05$). Al igual que los supuesto de homocedasticidad y heterocedasticidad no cumplen con los suficientes requisitos para que puedan satisfacer dichos supuestos. Por este motivo, llegamos a la conclusión de que aplicar una regresión lineal múltiple en estos datos no es una opción viable, a menos que se puedan solucionar el problema con los supuestos.

Por otro lado, una vía para poder continuar el análisis es mediante la categorización de los precios aproximados y convertir el modelo de regresión en un modelo de clasificación. Esta transformación podría solucionar probablemente algunas métricas al igual que los supuestos.



MUCHAS GRACIAS

Félix Antonio Mucha Morales

José Carlos Enriquez Lira