

1/09/2024



Proyecto Programación

C++

**Integrantes:**

Juan David Delgado Burbano

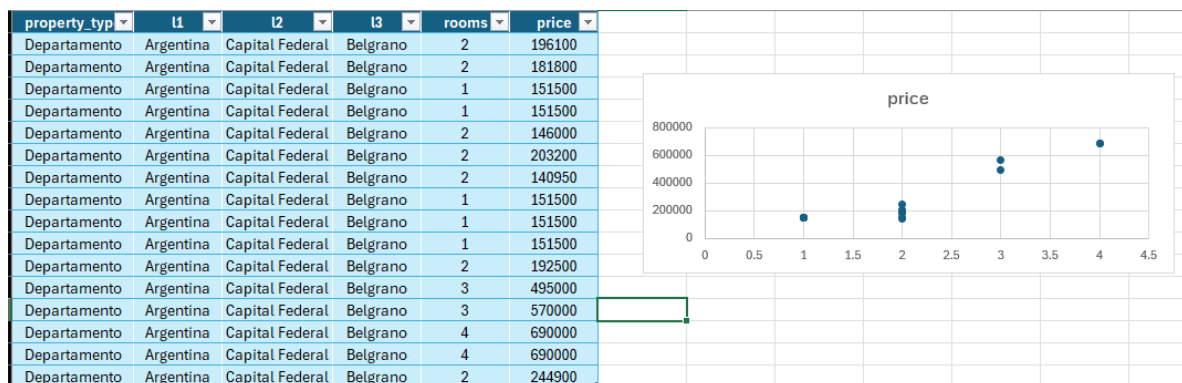
Nicolas Martínez

Camila Ariza

## Reporte análisis:

Para empezar, el archivo csv el cual el proyecto e c++ lee, es un archivo el cual contiene 414.120 filas de datos separados por comas, de los cuales, algunos de dichos datos no eran útiles para la realización de una regresión lineal en donde se predica el precio de alguna propiedad en base a alguno de estos datos como el numero de habitaciones, o los metros cuadrado de la vivienda.

Para ello, el equipo realizo el siguiente análisis, para ver datos que tuviesen relación entre ellos para la regresión lineal, en donde tomando 16 datos iniciales del csv, los cuales pertenecieran a la misma ciudad, mismo barrio, y sean el mismo tipo de vivienda (departamentos en este caso), se realizó una tabla con ellos utilizando el número de habitaciones como eje x, y el precio propiedad como eje y, (utilizando las funciones básicas de Excel).



De esta forma se comprobó una relación de a mayor numero de habitaciones, mayor valor posee la propiedad por lo cual se utilizaron estos datos como punto de partida.

# ● Limpieza de Datos:

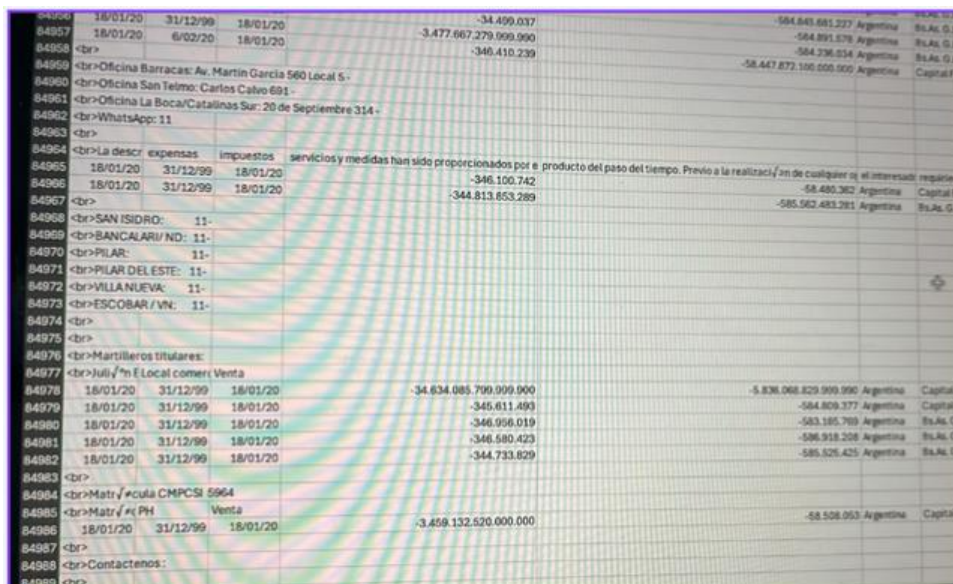
La base de datos del csv, tenia varios problemas, como:

1.

Apartir de la línea 80 mil, el csv contenía datos que interrumpían la lectura del código, como los que se muestran en la siguiente imagen, los cuales no correspondían al patron de lectura del csv siendo (

- Fecha de inicio,
- fecha termiando
- ,creado en ,
- lat,
- lon ,
- pais,
- ciudad,
- barrio ,
- #habitaciones
- #baños,
- #alcobas,
- superficie total,
- superficie cubierta ,
- precio ,
- usd ,
- nombre
- ,descripcion de la propiedad ,
- tipo de propiedad.

) al romper este ciclo de lectura, el código empezaba a tomar datos de tipo carácter como (tipo de propiedad) como valores enteros ( INT ), lo cual producía errores,



84959	18/01/20	31/12/99	18/01/20	-34.499.037	-584.843.681.237	Argentina	Capital F
84957	18/01/20	6/02/20	18/01/20	-3.477.667.279.999.990	-584.891.578	Argentina	BUA G.B.
84958	 			-340.410.239	-584.236.034	Argentina	BUA G.B.
84959	 Oficina Barracas: Av. Martin Garcia 560 Local 5 -				-58.447.872.100.000.000	Argentina	Capital F
84960	 Oficina San Telmo: Carlos Calvo 691 -						
84961	 Oficina La Boca/Catalinas Sur: 20 de Septiembre 314 -						
84962	 WhatsApp: 11						
84963	 						
84964	 La descr. expensas impuestos servicios y medidas han sido proporcionados por e						
84965	18/01/20	31/12/99	18/01/20	-346.100.742	-58.480.362	Argentina	Capital F
84966	18/01/20	31/12/99	18/01/20	-344.813.853.289	-585.562.483.281	Argentina	BUA G.B.
84967	 						
84968	 SAN ISIDRO: 11-						
84969	 BANCAJAR/ND: 11-						
84970	 PILAR: 11-						
84971	 PILAR DEL ESTE: 11-						
84972	 VILLA NUEVA: 11-						
84973	 ESCOBAR / VN: 11-						
84974	 						
84975	 						
84976	 Martilleros titulares:						
84977	 Julio F Local comeri Venta						
84978	18/01/20	31/12/99	18/01/20	-34.634.085.799.999.990	-5.836.068.829.999.990	Argentina	Capital F
84979	18/01/20	31/12/99	18/01/20	-345.611.493	-584.809.377	Argentina	Capital F
84980	18/01/20	31/12/99	18/01/20	-346.956.019	-583.185.769	Argentina	BUA G.B.
84981	18/01/20	31/12/99	18/01/20	-346.580.423	-586.918.208	Argentina	BUA G.B.
84982	18/01/20	31/12/99	18/01/20	-344.733.829	-585.525.425	Argentina	BUA G.B.
84983	 						
84984	 Matr/ ecula CMPCSI 5964						
84985	 Matr/ PH Venta						
84986	18/01/20	31/12/99	18/01/20	-3.459.132.520.000.000	-58.508.053	Argentina	Capital F
84987	 						
84988	 Contactenos:						
84989	 						

## 2. caracteres en valores numéricos:

Palermo	3	73	310000
Palermo	3	69	265000
Palermo	3	63	175000
Palermo	3	58	158000
Palermo	3	87	220000
Palermo	3	68	335000
Palermo	3	66	184900
Palermo	3	55	240000
Palermo	3	70	220000
Palermo	3	NA	236000
Palermo	3	66	184900
Palermo	3	65	134000
Palermo	3	90	399000
Palermo	3	64	145000
Palermo	3	82	334000
Palermo	3	213	318000
Palermo	3	66	174000
Palermo	3	72	215000
Palermo	3	56	189000
Palermo	3	78	2,00E+05
Palermo	3	78	2,00E+05

62	228000
71	245000
63	175000
197	6,00E+05
45	89000
52	120000

Completamente aleatorio en el csv, existían datos de carácter numérico, los cuales contenían caracteres, por lo cual esto producía error en la lectura del código; por otro lado a lo largo del csv existían datos que no se tomaron los cuales aparecían como ( NA ), y en otros casos, aparecían simplemente en blanco, porque no fueron llenados.

Este problema fue muy importante resolverlo puesto que, la forma de lectura del csv, era leer carácter por carácter, y cuando encontrase una coma significaba que hasta ese ultimo carácter antes de la coma era parte de una variable. Y si por alguna razón alguna variable no estaba en la fila, entonces el código emparejaría datos erróneamente con sus respectivas variables de lectura en el código.

## 3. Eliminación de datos no necesarios:

Por ultimo, solo de dejo datos que serian necesarios para la implementación del código, los cuales fueron: 7

property_type	l1	l2	l3	rooms	surface_total	price
Departamento	Argentina	Capital Fede	Belgrano	2	57	196100

Estos datos eran necesarios para hacer la regresión lineal, puesto que variables como país, ciudad barrio de procedencia de la propiedad, ayudaban a entender la valorización de las propiedades dependiendo su ciudad u barrio. Además el numero de habitaciones y el precio fueron los datos seleccionados para la regresión, y datos adicionales como tipo de propiedad, y superficie total, eran necesarios para poder diferencia las propiedades, ( si eran departamentos o casas, o porque el valor de un apartamento con 2 habitaciones era mas caro que otro de dos habitaciones, esto debido a los metros cuadrados ).

## Solución de problemas:

Se elaboraron dos código en “ R “ para la limpieza de datos del csv,

I.

Este código busca eliminar, en la columnas los elementos NA y los números con notación científica.

```

8
9 library(dplyr)
10
11 file_path <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati.csv"
12 data <- read.csv(file_path, stringsAsFactors = FALSE)
13
14 data <- data %>%
15   mutate(across(everything(), ~ trimws(.)))
16
17 data <- data %>% select(property_type, l1, l2, l3, rooms, surface_total, price)
18
19 output_file <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati_Limpio.csv"
20 write.csv(data, output_file, row.names = FALSE)
21

```

4.14 (Top Level) ±

Console Terminal Background Jobs

```

R 4.4.0 - ~/Desktop/Datos/
Cannot open file "/Users/nicolasmartinezdelgadillo/Desktop/DS_Proyecto_01_Datos_Properati.csv": No such file
library(dplyr)

# Cargar el archivo CSV
file_path <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati.csv"
data <- read.csv(file_path, stringsAsFactors = FALSE)

# Limpiar los espacios en blanco de las columnas
data <- data %>%
  mutate(across(everything(), ~ trimws(.)))

# Seleccionar solo las columnas necesarias
data <- data %>% select(property_type, l1, l2, l3, rooms, surface_total, price)

# Guardar el archivo limpio
output_file <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati_Limpio.csv"
write.csv(data, output_file, row.names = FALSE)

```

## II.

Este código busca en el csv en la carpeta del escritorio, luego limpia la base de datos eliminando las columnas y filas vacías y luego elimina todas las columnas , menos las columnas de los datos escogidos:

```
library(dplyr)
|
file_path <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati_Limpio.csv"
df <- read.csv(file_path, stringsAsFactors = FALSE)
df$price <- as.numeric(df$price)
df_cleaned <- df %>%
  na.omit() %>%
  filter(price < 1e6)
cleaned_file_path <- "/Users/nicolasmartinezdelgadillo/Desktop/Limpia/DS_Proyecto_01_Datos_Properati_Limpio2.csv"
write.csv(df_cleaned, cleaned_file_path, row.names = FALSE)
```

Dejando únicamente estos datos por fila:

property_type	l1	l2	l3	rooms	surface_total	price
Departamento	Argentina	Capital Fede	Belgrano	2	57	196100

## Resultado final csv:

Finalmente, la tabla quedaría con 123517 filas, cada fila contiene 7 datos, separados por comas como se muestra en la siguiente imagen:

property_type,l1,l2,l3,rooms,surface_total,price	
Departamento,Argentina,Capital Federal,San Cristobal,7,140,153000	
PH,Argentina,Capital Federal,Boedo,2,70,159000	
PH,Argentina,Capital Federal,Palermo,2,45,125000	
PH,Argentina,Capital Federal,Palermo,2,85,295000	
PH,Argentina,Bs.As. G.B.A. Zona Sur,La Plata,2,50,40000	
PH,Argentina,Capital Federal,Villa Crespo,2,56,150000	
PH,Argentina,Capital Federal,Villa Crespo,2,70,159500	
PH,Argentina,Capital Federal,Villa Crespo,2,70,159500	
PH,Argentina,Capital Federal,Parque Patricios,1,45,89000	
PH,Argentina,Capital Federal,Parque Patricios,1,45,89000	
PH,Argentina,Capital Federal,Villa Pueyrredón,2,66,170000	
Departamento,Argentina,Capital Federal,Boedo,2,68,149000	
Departamento,Argentina,Capital Federal,Boedo,2,50,115000	

## Análisis de Validación de modelo / Modelo:

( la explicación del código de c++ de validación del modelo, y el modelo de regresión lineal, esta detallada en el github del proyecto, en apartado presentación )

### Análisis:

regresión original	regresión prueba	regresión entrenamiento
<div>Epocas   20000</div> <div>Ecuación de la recta: <math>y = 3791.6887 * \text{habitaciones} + 200955.9219</math></div> <div>Error cuadrático medio (MSE): 0.0219</div>	<div>Epocas   20000</div> <div>Ecuación de la recta: <math>y = 7857.1094 * \text{habitaciones} + 201363.5781</math></div> <div>Error cuadrático medio (MSE): 0.0240</div>	<div>Epocas   20000</div> <div>Ecuación de la recta: <math>y = 3835.5237 * \text{habitaciones} + 197617.8906</math></div> <div>Error cuadrático medio (MSE): 0.0214</div>
<div>Habitaciones   Precio estimado (\$)</div> <div>7.00   \$ 227497.75</div>	<div>Habitaciones   Precio estimado (\$)</div> <div>7.00   \$ 256363.34</div>	<div>Habitaciones   Precio estimado (\$)</div> <div>7.00   \$ 224466.56</div>

En la imagen se muestra las impresiones finales del modelo original (100% datos csv) , el modelo de prueba (20% datos csv ), y el modelo de entrenamiento (80% datos csv ),

- Ecuación de la recta:

Como se ve en la imagen las ecuaciones de las rectas del modelo de regresión lineal original, y el modelo de entrenamiento son muy cercanas, sobre todo en la pendiente de la recta, sus valores son muy similares, y diferencian un poco en el corte con el eje y, sin embargo, su cercanía comprueba que el modelo original funciona correctamente.

Por otro lado frente a la pendiente, las ecuaciones del modelo original, y el modelo de prueba, diferencian bastante, sin embargo, en el caso de la pendiente, este valor tiene sentido puesto que dentro del rango del 20% que toma el modelo de prueba, puede tener datos que están más dispersos en proporción al 100% que posee el original. Prueba de ello, es que ambos tienen un corte con el eje y, muy similar.

- MSE:

Todos los modelos poseen un error cuadrático promedio de 0.02, variando únicamente en el decimal después del 2. Lo que significa que la predicción es casi perfecta.

- PREDICCION:

Todos los precios predichos para #7 habitaciones, son muy similares, especialmente entre el modelo original y el modelo de entrenamiento, mientras que el modelo de

prueba varia significativamente 2000 dólares mas alto que los demás, pero esto nuevamente se debe a la pendiente de su recta, la cual ya se explico en el punto anterior la razón de ello.

- EPOCAS:

Todos los modelos fueron probados con 20 mil épocas.

### Ejemplo de impresión:

```
Resultados de la regresión lineal:
-----
Epocas           | 20000
-----
Ecuación de la recta:
y = 3791.6887 * habitaciones + 200955.9219
-----
Error cuadrático medio (MSE): 0.0219
-----
```

Habitaciones	Precio estimado (\$)
1.00	\$ 204747.61
2.00	\$ 208539.30
3.00	\$ 212330.98
4.00	\$ 216122.67
5.00	\$ 219914.36
6.00	\$ 223706.06
7.00	\$ 227497.75
8.00	\$ 231289.44
9.00	\$ 235081.12
10.00	\$ 238872.81
11.00	\$ 242664.50
12.00	\$ 246456.19
13.00	\$ 250247.88
14.00	\$ 254039.56
15.00	\$ 257831.25
16.00	\$ 261622.94
17.00	\$ 265414.62
18.00	\$ 269206.31
19.00	\$ 272998.00
20.00	\$ 276789.69