

LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

LABORATORIO 3

SEMINARIO DE BIG DATA Y CIENCIAS DE DATOS

CRISTHIAN CAMILO BALAGUERA SOTO - 88913
DANIEL MAURICIO GORDILLO ALVAREZ - 89400

PROFESOR

ELIAS BUITRAGO BOLIVAR

UNIVERSIDAD ECCI

BOGOTÁ D.C., COLOMBIA

2024

Laboratorio de comprensión de los datos

Introducción

En este documento se procederá a realizar la limpieza de datos del conjunto de nombre “*housing_fincaraiz.csv*”. Así mismo se desarrollarán los lineamientos de IBM relacionados con “*data understanding*”.

Fase 1: Recolectar datos iniciales

habitacioni	baños	parqueade	area_const	area_priva	estrato	estado	antiguedad	administraci	precio_m2	Ascensor	Circuito cerrado	Parqueadero	Portería	Zonas Verdes	Salón Comunal	Balcón	Barra estilo
2	2	1	92 m²	92 m²	4	No definida	9 a 15 años	\$A 622.000 C/ \$A 6.521.739,	1	1	1	1	1	0	0	1	1
1	2	1	56 m²	56 m²	6	No definida	1 a 8 años	\$A 523.000 C/ \$A 8.392.857,	1	0	1	1	1	0	0	0	1
3	4	2	144 m²	144 m²	6	No definida	16 a 30 años	\$A 620.000 C/ \$A 6.597.222,	1	0	0	0	0	0	0	0	1
1	1	0	31 m²	31 m²	4	Excelente	menor a 1 a	\$A 130.000 C/ \$A 7.419.354,	1	1	1	1	0	0	0	0	0
3	2	1	52 m²	52 m²	4	No definida	1 a 8 años	\$A 219.000 C/ \$A 5.576.923,	1	1	1	1	0	0	0	0	1
3	3	1	150 m²	150 m²	6	Bueno	más de 30 a	\$A 872.000 C/ \$A 6.533.333,	1	1	1	1	0	0	0	0	0
3	2	1	110 m²	100 m²	3	No definida	16 a 30 años	\$A 135.000 C/ \$A 4.181.818,	0	1	0	0	0	1	0	0	0
3	2	0	53 m²	47 m²	3	No definida	9 a 15 años	\$A 125.000 C/ \$A 3.207.547,	0	0	1	1	1	1	1	0	0
3	3	1	111 m²	0 m²	4	Excelente	16 a 30 años	No definida \$A 3.873.873,	0	1	1	1	0	0	0	1	0
4	4	2	264 m²	264 m²	5	Bueno	más de 30 a	\$A 836.000 C/ \$A 5.303.030,	1	0	0	0	0	0	0	0	0
3	2	2	97 m²	0 m²	4	No definida	9 a 15 años	\$A 272.000 C/ \$A 6.175.257,	1	1	1	0	0	0	0	0	1
2	3	2	87 m²	82 m²	4	No definida	9 a 15 años	\$A 350.000 C/ \$A 4.712.643,	1	1	1	1	1	0	1	0	0
3	4	2	175 m²	175 m²	4	No definida	16 a 30 años	No definida \$A 4.285.714,	0	0	0	0	0	1	0	0	0
10	5	2	391 m²	0 m²	3	No definida	16 a 30 años	No definida \$A 1.994.884,	0	0	0	0	0	0	0	0	0
6	4	3	218 m²	0 m²	4	Bueno	más de 30 a	\$A 3.027.522,	0	1	0	0	0	0	0	0	0
1	2	1	50 m²	0 m²	5	Bueno	9 a 15 años	\$A 310.000 C/ \$A 6.800.000'	0	0	0	0	1	0	0	0	1
3	3	1	140 m²	140 m²	4	Bueno	más de 30 a	\$A 460.000 C/ \$A 5.071.428,	0	1	0	1	0	1	0	1	0
4	3	1	90 m²	0 m²	4	Bueno	más de 30 a	\$A 369.000 C/ \$A 4.000.000'	1	0	0	1	1	0	1	0	0
2	3	2	111 m²	111 m²	6	Bueno	9 a 15 años	\$A 786.000 C/ \$A 11.261.261,	1	1	1	1	0	0	0	0	1
12	5	0	375 m²	320 m²	3	Bueno	16 a 30 años	No definida \$A 1.493.333,	0	0	0	0	0	0	0	0	1
2	3	2	111 m²	73 m²	4	No definida	16 a 30 años	\$A 483.000 C/ \$A 3.423.423,	1	0	1	1	1	0	1	0	0
4	4	2	307 m²	307 m²	3	No definida	más de 30 a	\$A 2.931.596,	0	1	1	1	0	1	0	1	0
3	2	0	61 m²	0 m²	3	No definida	16 a 30 años	\$A 130.000 C/ \$A 2.622.950,	0	0	1	0	0	0	1	0	0
2	3	4	322 m²	277 m²	6	No definida	1 a 8 años	\$A 2.304.000' \$A 9.000.000'	1	1	0	0	0	0	0	0	1
5	5	4	280 m²	280 m²	6	No definida	16 a 30 años	\$A 2.200.000' \$A 7.142.857,	0	1	1	1	1	1	0	0	0
3	2	1	59 m²	54 m²	4	Bueno	16 a 30 años	\$A 150.000 C/ \$A 4.220.338,	0	1	1	1	0	0	1	0	0
3	3	2	146 m²	133 m²	5	Excelente	16 a 30 años	\$A 660.000 C/ \$A 5.410.958,	1	1	0	0	1	1	0	0	1

- ¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?
 - habitaciones, baños, parqueadero, area_construida, area_privada, estrato, antigüedad, administracion, precio_m2, zonas verdes, calentador, cocina integral, vigilancia, parques cercanos, nombre, ubicación, precio.
- ¿Qué variables parecen irrelevantes y pueden ser excluidos?
 - Ascensor, Circuito cerrado de TV, Parqueadero Visitantes, portería/recepción, Salón Comunal, Balcon, barra estilo americano, chimenea, citofono, terraza, estudio, patio, deposito/bodega.
- ¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?
 - En el grupo de datos se encuentran 8429 filas de datos por lo que se considera viable para predicciones o conclusiones.
- ¿Hay demasiadas variables para el método de modelado de su elección?
 - En el grupo de datos se encuentran 31 variables diferentes, por lo que si se consideraría reducirlas para obtener resultados más precisos.
- ¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?
 - Esta es una base de datos única que no incluye fusión alguna.

- ¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?
 - Al momento de la revisión de datos, no se encontraron espacios en blanco por lo que no se revisará este asunto.

Fase 2: Describir los datos

- ¿Cuál es el formato de los datos?
 - Incluyen datos numéricos, alfanuméricos y caracteres.
- ¿Cuál es el método utilizado para capturar los datos?
 - Registros de casas en finca raíz.
- ¿Qué tamaño tiene la base de datos (en número de filas y columnas)?
 - 8.429 filas y 31 columnas para un total de 261.299 datos.
- ¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?
 - Variables como el precio, valor del metro cuadrado, estrato, valor de la administración y ubicación se consideran relevantes para la pregunta de negocio.
- ¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?
 - Numéricos, alfanuméricos, caracteres y simbólicos.
- ¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?
- ¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?

Fase 3: Describir los datos

- ¿Qué tipo de hipótesis se ha formado sobre los datos?
 - Son datos sobre los valores de apartamentos o casas y si estas incluyen cosas específicas como vigilancia, zonas verdes, etc.
- ¿Qué variables parecen prometedoras para un análisis más profundo?
 - Precio, estrato, valor del metro cuadrado, estado, antigüedad, habitaciones, valor de la administración y ubicación se consideran útiles para el análisis profundo de los datos.
- ¿Sus exploraciones han revelado nuevas características sobre los datos?
 - Los datos incluyen detalles bastante específicos sobre las características que incluyen las casas/apartamentos por lo que se explica con detalle lo que viene con cualquier propiedad.
- ¿Cómo han cambiado estas exploraciones su hipótesis inicial?
 - En este caso se mantiene la hipótesis ya que desde un inicio se explicó que este era un registro de propiedades con detalles de cada una.
- ¿Considera que debería reformular el alcance del proyecto?
- ¿Esta exploración ha alterado los objetivos?

- ¿Puede identificar subconjuntos particulares de datos para su uso posterior?
 - Los datos que hablan de los detalles de cada propiedad. Estos datos se pueden usar para un análisis de las propiedades más completas con mejor precio.

Fase 4: Verificar la calidad de los datos

Se cambia el nombre de la casilla “baños” por “banos” (casilla B2)



Se cambia el nombre de la casilla “Ascensor” por “ascensor” (casilla K1)



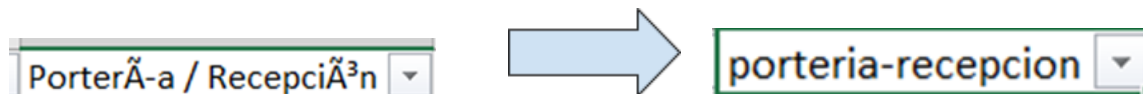
Se cambia el nombre de la casilla “Circuito cerrado de TV” por “circuito_cerrado_tv” (casilla L1)



Se cambia el nombre de la casilla “Parqueadero Visitantes” por “parqueadero_visitantes” (casilla M1)



Se cambia el nombre de la casilla “Portería / Recepción” por “porteria-recepcion” (casilla N1)



Se cambia el nombre de la casilla “Zonas Verdes” por “zonas_verdes” (casilla O1)



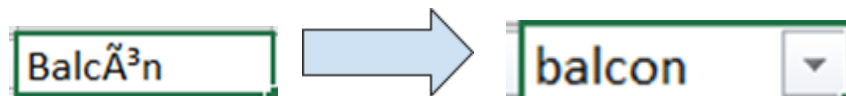
LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

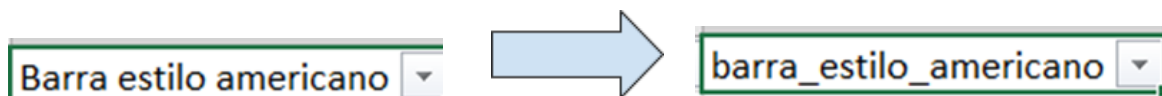
Se cambia el nombre de la casilla “SalÃ³n Comunal” por “salon_comunal” (casilla P1)



Se cambia el nombre de la casilla “BalcÃ³n” por “balcon” (casilla P1)



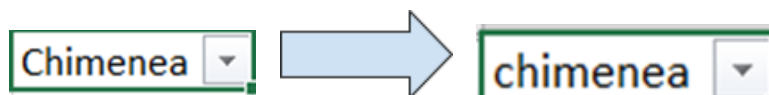
Se cambia el nombre de la casilla “Barra estilo americano” por “barra_estilo_americano” (casilla R1)



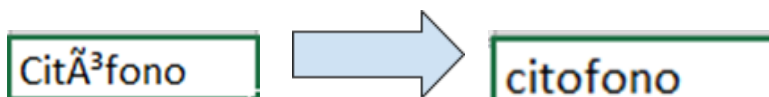
Se cambia el nombre de la casilla “Calentador” por “calentador” (casilla S1)



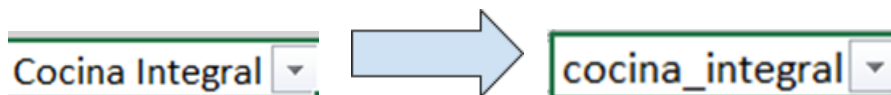
Se cambia el nombre de la casilla “Chimenea” por “chimenea” (casilla T1)



Se cambia el nombre de la casilla “CitÃ³fono” por “citofono” (casilla U1)



Se cambia el nombre de la casilla “Cocina Integral” por “cocina_integral” (casilla V1)



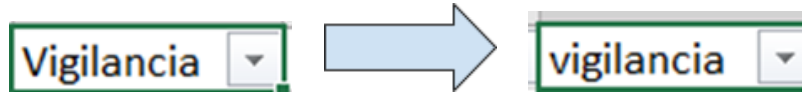
Se cambia el nombre de la casilla “Terraza” por “terrazza” (casilla W1)



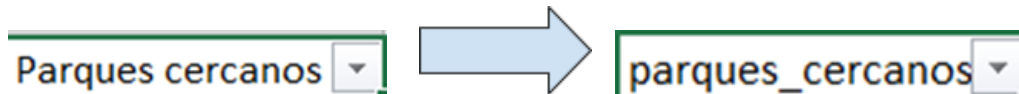
LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

Se cambia el nombre de la casilla “Vigilancia” por “vigilancia” (casilla X1)



Se cambia el nombre de la casilla “Parques cercanos” por “parques_cercanos” (casilla Y1)



Se cambia el nombre de la casilla “Estudio” por “estudio” (casilla Z1)



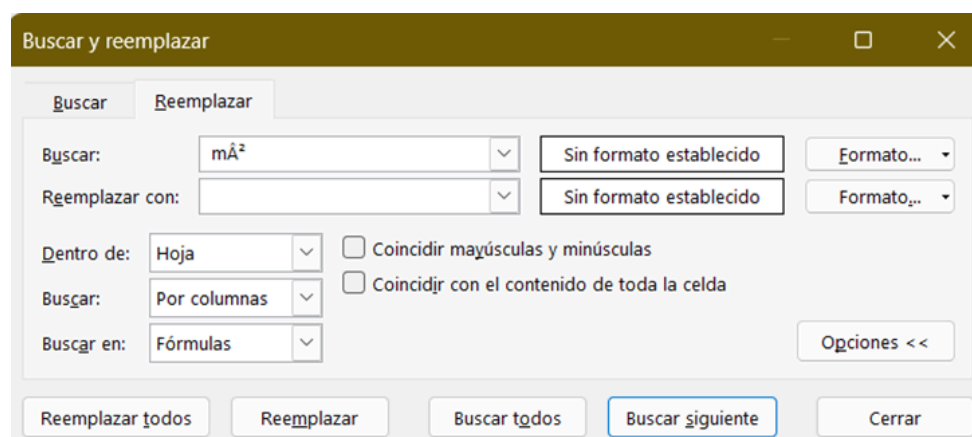
Se cambia el nombre de la casilla “Patio” por “patio” (casilla AA1)



Se cambia el nombre de la casilla “Depósito / Bodega” por “deposito-bodega” (casilla AB1)



A partir de este punto, se empezaran a limpiar los datos en columnas con la función “Buscar y reemplazar”. Cuando se requiera de cambio de todos los datos se usara la opcion de “Reemplazar todos”.



LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

Para la columna “habitaciones” (columna A) se cambiaron los valores que aparecían como “No definida” por “0”. Esto con el fin de que únicamente se manejen valores numéricos en esa columna.

The diagram illustrates a transformation of a table. On the left, a table with 10 rows and 1 column is shown. The header is 'habitacion' and all 10 rows contain the text 'No definida'. A large blue arrow points from this table to a second table on the right. The second table also has 10 rows and 1 column, with the same header 'habitacion', but all 10 rows contain the value '0'.

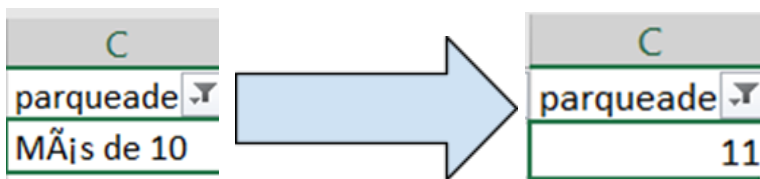
Para la columna “banos” (columna B) Se cambiaron los valores que aparecían como “No definida” por “0”. Esto con el fin de que únicamente se manejen valores numéricos en esa columna.

The diagram illustrates a transformation process. On the left, a table with the header 'banos' contains 15 rows, all of which are 'No definida'. A large blue arrow points from this table to a second table on the right. The second table also has the header 'banos' and contains 15 rows with numerical values: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, and 0.

LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

Para la columna “parqueaderos” (columna C) se cambiaron los valores que aparecían como “Más de 10” por “11”. Esto con el fin de que únicamente se manejen valores numéricos en esa columna. Se entenderá que cuando se ingrese 11 en esta columna es porque se encuentran más de 10 baños dentro de la residencia.

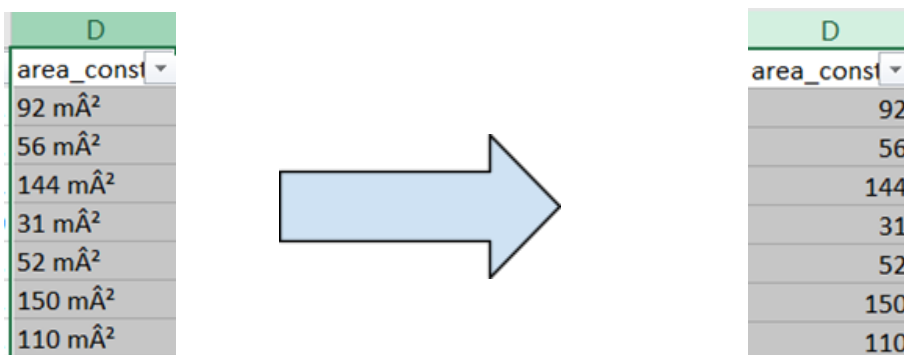


The diagram illustrates a change in the data for column C. On the left, a table snippet shows column C with a dropdown menu labeled 'parqueade' and a value 'Más de 10'. A large blue arrow points to the right, where the same table snippet is shown, but the value in column C has been changed to '11'.

C
parqueade
Más de 10

C
parqueade
11

Dado a que para la columna “area_construida” (columna D) se entiende que todas las medidas se encuentran en metros cuadrados (m^2) se optó por retirar la medida de los valores y únicamente dejar los valores numéricos. Mas precisamente, se borra la parte de “ m^2 ” dentro de las casillas.

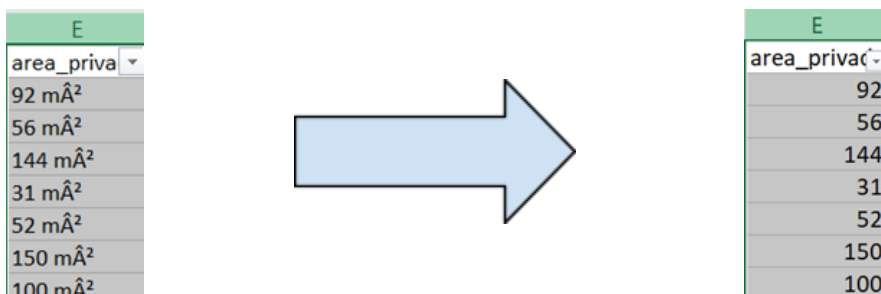


The diagram illustrates the removal of the unit ' m^2 ' from the values in column D. On the left, a table snippet shows column D with a dropdown menu labeled 'area_const' and values including '92 m^2 ', '56 m^2 ', '144 m^2 ', '31 m^2 ', '52 m^2 ', '150 m^2 ', and '110 m^2 '. A large blue arrow points to the right, where the same table snippet is shown, but the unit ' m^2 ' has been removed from all values, leaving only the numbers: '92', '56', '144', '31', '52', '150', and '110'.

D
area_const
92 m^2
56 m^2
144 m^2
31 m^2
52 m^2
150 m^2
110 m^2

D
area_const
92
56
144
31
52
150
110

Dado a que para la columna “area_privada” (columna E) se entiende que todas las medidas se encuentran en metros cuadrados (m^2) se optó por retirar la medida de los valores y únicamente dejar los valores numéricos. Mas precisamente, se borra la parte de “ m^2 ” dentro de las casillas.



The diagram illustrates the removal of the unit ' m^2 ' from the values in column E. On the left, a table snippet shows column E with a dropdown menu labeled 'area_privada' and values including '92 m^2 ', '56 m^2 ', '144 m^2 ', '31 m^2 ', '52 m^2 ', '150 m^2 ', and '100 m^2 '. A large blue arrow points to the right, where the same table snippet is shown, but the unit ' m^2 ' has been removed from all values, leaving only the numbers: '92', '56', '144', '31', '52', '150', and '100'.

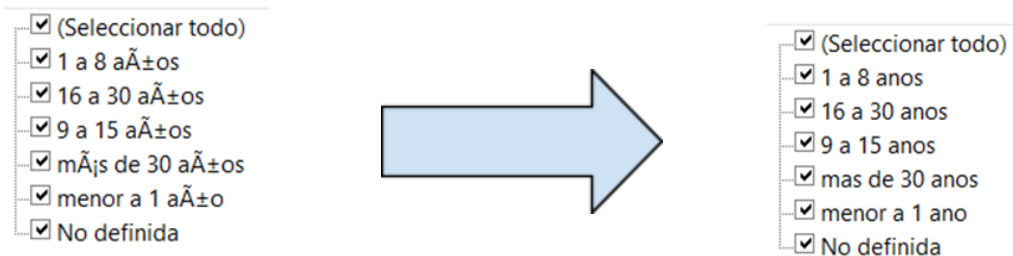
E
area_privada
92 m^2
56 m^2
144 m^2
31 m^2
52 m^2
150 m^2
100 m^2

E
area_privada
92
56
144
31
52
150
100

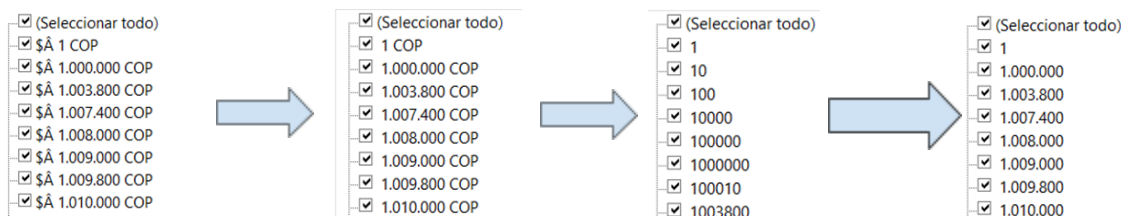
LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

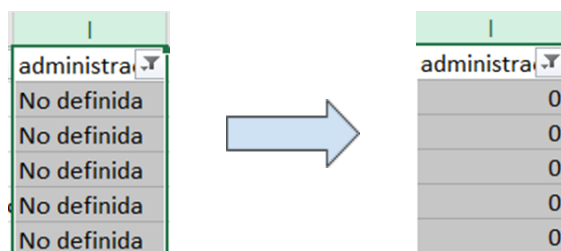
En el caso de la columna “antigüedad” (columna H) se encontraron conflictos de caracteres como en la palabra “años” la cual en la columna está guardada como “aÃ±os” o en otro caso, la palabra “más” la cual está guardada como “mÃ¡s”. Por estas razones, se cambiarán estas palabras en donde se incluyan por “anos” y “mas” respectivamente.



En el caso de la columna “administracion” (columna I) se encontró que los valores en estas casillas contienen este mismo formato en general: “\$Â 622.000 COP” por lo que se entiende que se están refiriendo a un valor en pesos colombianos. Por esta razón se toma la decisión de cambiar el formato para dejar solo los valores numéricos. En conclusión, el formato se vería así: “622000”



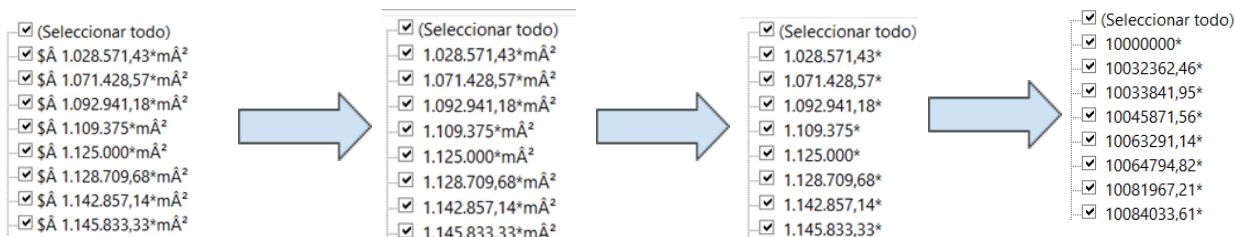
Como adicional en la columna “administracion” (columna I) también se encuentran algunas casillas con el contenido de “No definida” por lo que estas se reemplazarán con 0 con el fin de mantener solo valores numéricos dentro de la columna.



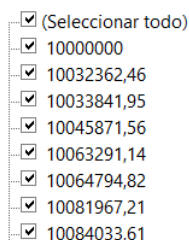
LABORATORIO 3

Integrantes: Cristhian Balaguera - Daniel Gordillo

En el caso de la columna “precio_m2” (columna J) se encontró que los valores en estas casillas contienen este mismo formato en general: “\$Â 6.521.739,13*mÂ²” por lo que se entiende que se están refiriendo al precio del metro cuadrado (m²) para cada residencia. Por esta razón se toma la decisión de cambiar el formato para dejar solo los valores numéricos. En conclusión, el formato ahora se vería así: “6521739,13”.

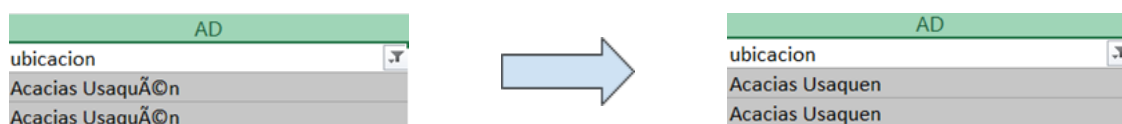


Para el caso específico de tener que retirar el “*” de los datos, se tiene que ingresar “~*” para que borre únicamente el “*”. Esto dará a entender que borre el carácter literal y no termine por borrar todo el dato.



En el caso de la columna “ubicacion” (columna AD), se encontraron varias vocales con tilde que entraron en conflicto y cambiaron sus caracteres, por lo que a continuación, se mostrará los caracteres en conflicto y por qué otro carácter será reemplazado.

“Ã©” por “e”



“Ã³” por “o”



“Ã±” por “n”



“Ã” por “a”

“-” por “i”

AD
ubicacion
Ciudad JardÃ-n

→

AD
ubicacion
Bogota

- ¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?
 - No se encuentran campos en blanco
- ¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?
 - Así como se describió anteriormente en el procedimiento de limpieza de los datos, se encontraron varios errores que eran vocales con tildes o veces en las que se incluyen la “ñ”.
- ¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?
- ¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).
 - Se han encontrado conflictos con propiedades con 1 metro cuadrados de propiedad privada.
 - Se han encontrado conflictos con propiedades con 1 metro cuadrados de propiedad construida.
 - Se han encontrado propiedades con valor de 1 COP.
- ¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?
 - Se ha propuesto excluir estos datos con conflictos mencionados anteriormente.
- ¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores coherentes entre los archivos?
 - Los datos están almacenados en un archivo plano delimitado por comas. No se encontró con ningún conflicto entre estas delimitaciones.
- ¿Cada registro contiene el mismo número de campos?
 - Se podrían catalogar los registros como simétricos ya que cada uno están completos desde la primera hasta la última columna.