

LABORATORIO 3

**DANIEL GUTIERREZ 90378
JASON RODRIGUEZ 99229**

SEMINARIO BIG DATA

ELIAS BUITRAGO BOLIVAR

**BOGOTA D.C.
30 DE JUNIO DE 2024
UNIVERSIDAD ECCI**

Recolectar datos iniciales

Según IBM, se deben tener en cuenta una serie de preguntas en el momento de realizar la recolección inicial de datos. En esta sección se listan dichas preguntas. Puede consultar la publicación detallada mediante este hipervínculo.

¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?

RTA: Estado, ubicación, precio, estrato, antigüedad, administración, nombre, administración.

¿Qué variables parecen irrelevantes y pueden ser excluidos?

RTA: Precio_m2, área_construida, chimenea, barra_estilo_americano,

¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?

RTA: Si, contamos con variables de precio, estado, antigüedad y información descriptiva del inmueble que nos permite realizar un análisis de costo/beneficio, viabilidad para un posible comprador.

¿Hay demasiadas variables para el método de modelado de su elección?

RTA: No, las variables que tenemos aportan bastante para lograr dar un modelo acertado que nos permita tener un contexto general del inmueble y así el comprador pueda tomar una decisión acertada de acuerdo a su presupuesto y necesidad.

¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?

RTA: No, Todos los datos con los que contamos actualmente hacen referencia al inmueble, por lo que tienen una correlación directa que permite tener un contexto amplio, adicional todo viene de una fuente de información.

¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?

RTA: Hemos evidenciado que hay varias columnas que no cuentan con datos o se encuentran nulos, por lo que por medio de la estadística se puede utilizar la media, promedio, top, bottom, que nos permite tener una mayor visibilidad del ruido que hay en los datos o si genera una afectación al momento de plantear el modelo para tomar una decisión de eliminarlos o mantenerlos.

Describir los datos

Según IBM, es recomendable “enfocarse en la cantidad y calidad de los datos, y realizar un reporte de descripción de los datos; puede consultar la publicación detallada mediante este hipervínculo.
De

igual manera, las preguntas que sugiere IBM para responder en esta sección son las siguientes:

¿Cuál es el formato de los datos?

RTA: Tenemos datos de tipo booleano, varchar, char y integer

¿Cuál es el método utilizado para capturar los datos?

RTA: Los datos parecen haber sido capturados mediante observación directa o recopilación de registros de propiedades, probablemente de un sistema de administración inmobiliaria.

¿Qué tamaño tiene la base de datos (en número de filas y columnas)?

RTA: la base de datos cuenta con 30 columnas y 8428 filas o registros

¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?

¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?

RTA: Encontramos datos numéricos, de tipo texto, con símbolos y booleanos (1/0)

¿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?

Explorar los datos

Según IBM, es recomendable “utilizar herramientas de visualización” para explorar los datos. En la práctica esto significa aplicar estadística descriptiva buscando gráficos para visualizar de manera resumida los hallazgos. De esta manera el equipo de analítica de datos podrá entender mejor la naturaleza de los datos. Por ejemplo, definir si los datos se corresponden con una distribución de probabilidad determinada. Además, tendrán insumos valiosos para realizar entregar un reporte de descripción de los datos, en caso de que sea requerido. Una guía detallada sobre este punto se puede
consultar aquí. De igual manera, las preguntas sugeridas para responder en esta sección son las siguientes:

¿Qué tipo de hipótesis se ha formado sobre los datos?

RTA: Al revisar las unidades de medida en el dataset, se observaron las siguientes inconsistencias y posibles errores:

- `area_construida` y `area_privada`: Ambas variables están en metros cuadrados (m^2), pero los valores parecen tener caracteres adicionales como espacios y símbolos (13 m^2 , 92 m^2). Estos deberían limpiarse para asegurar que los datos numéricos se manejen correctamente.
- `precio` y `precio_m2`: Los precios tienen caracteres especiales y símbolos monetarios (\$, ,, COP, * m^2). Esto puede causar problemas en análisis cuantitativos y debe estandarizarse a valores numéricos sin caracteres especiales.

Pasos recomendados para corregir los errores de medición:

- Eliminar caracteres no numéricos en las columnas `area_construida` y `area_privada`.
- Convertir las columnas `precio` y `precio_m2` a valores numéricos eliminando símbolos y caracteres innecesarios.
- Verificar la coherencia de las unidades de medida a lo largo del dataset para asegurarse de que todas las áreas están en metros cuadrados y todos los precios están en la misma moneda.

¿Qué variables parecen prometedoras para un análisis más profundo?

RTA:

- `habitaciones`: Número de habitaciones.
- `baños`: Número de baños.
- `parqueaderos`: Número de parqueaderos.
- `area_construida` y `area_privada`: Superficie construida y privada.
- `estrato`: Nivel socioeconómico.
- `precio` y `precio_m2`: Precio total y precio por metro cuadrado.
- `ubicación`: Localización de las propiedades.

¿Sus exploraciones han revelado nuevas características sobre los datos?

RTA: Sí, se revelaron nuevas características:

- La variable `estado` tiene valores como "No definida" y "Excelente".
- La antigüedad está clasificada en categorías como "menor a 1 año", "1 a 8 años", etc.
- Existen múltiples variables binarias que indican la presencia de características específicas (`Cocina_integral`, `Terraza`, `Vigilancia`, etc.).

¿Cómo han cambiado estas exploraciones su hipótesis inicial?

RTA: Inicialmente, podría haberse asumido que las variables precio y area_construida serían las más relevantes. La presencia de muchas variables categóricas y binarias sugiere que un análisis más detallado de estas podría ser necesario.

¿Considera que debería reformular el alcance del proyecto?

RTA: No, ya que el análisis de las variables con las que actualmente se cuenta es el alcance inicial, ya de acuerdo a comportamientos o resultados iríamos viendo si es necesario realizar ajustes.

¿Esta exploración ha alterado los objetivos?

RTA: Sí, los objetivos del proyecto podrían ajustarse para analizar cómo diferentes características de las propiedades afectan el precio y el valor por metro cuadrado.

¿Puede identificar subconjuntos particulares de datos para su uso posterior?

RTA: Sí, se pueden identificar varios subconjuntos:

- Propiedades con estado específico.
- Diferentes rangos de antigüedad.
- Propiedades en distintas ubicaciones.
- Análisis por estrato para observar diferencias socioeconómicas.

Verificar la calidad de los datos

Según IBM, es recomendable verificar la calidad de los datos con un enfoque en los siguientes aspectos:

Identificar datos faltantes

RTA: Hay datos que se identificaron como “No definidos” en diferentes columnas, que nos indican que no fueron llenados en la página donde se extrajo la información

Identificar errores tipográficos en los datos

RTA: Se logró identificar que hay columnas que mezclaban mayúsculas y minúsculas, adicionalmente algunos campos de precio *m2 se encontraban en un formato incorrecto, se hizo los respectivos replace y conversiones para ajustarlos.

Identificar errores de medición (en las unidades de medida)

Identificar inconsistencias en la codificación

RTA: El archivo fuente es un CSV, por lo que tuvimos que tomar los datos de cada una de las columnas para almacenarlos en una tabla de SQL y realizar la limpieza, no contenía código.

Complementariamente, IBM sugiere dar respuesta a las siguientes preguntas:

¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?

RTA: Hay varios campos que tienen datos como “No Definidos”, por lo que debemos validar que tanto podrían afectar en el modelo que se pueda implementar o si llegaría a hacer ruido en los datos, las encontramos en columnas como estado, baños, antigüedad, administración, etc..

¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?

RTA: Si, hemos encontrado en la ubicación una mezcla de mayúsculas y minúsculas que nos pueden causar problemas a futuro en la filtración de datos.

¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?

RTA: Encontramos que el 44% de los datos que tenemos en la columna de estado están “No definidos” esto nos genera un ruido al momento de conocer las condiciones del inmueble, además de que es una columna principal que da una gran visibilidad.

¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).

RTA: No se han detectado conflictos aparentes en una inspección preliminar. Se recomienda una exploración más detallada para detectar valores implausibles como adolescentes con altos niveles de ingresos.

¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?

RTA: Podría considerarse excluir variables como **nombre** si no aportan valor analítico significativo.

¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores coherentes entre los archivos?

RTA: Al archivo se le tuvo que hacer unas conversiones en los delimitadores para que al momento de cargar la información a SQL pudiera reconocer cada una de las columnas, lo estaba tomando como “,” y se debía leer en “;”.

¿Cada registro contiene el mismo número de campos?

Sí, cada registro contiene el mismo número de campos, lo cual indica consistencia en la estructura del archivo.

DESCRIPCIÓN GENERAL

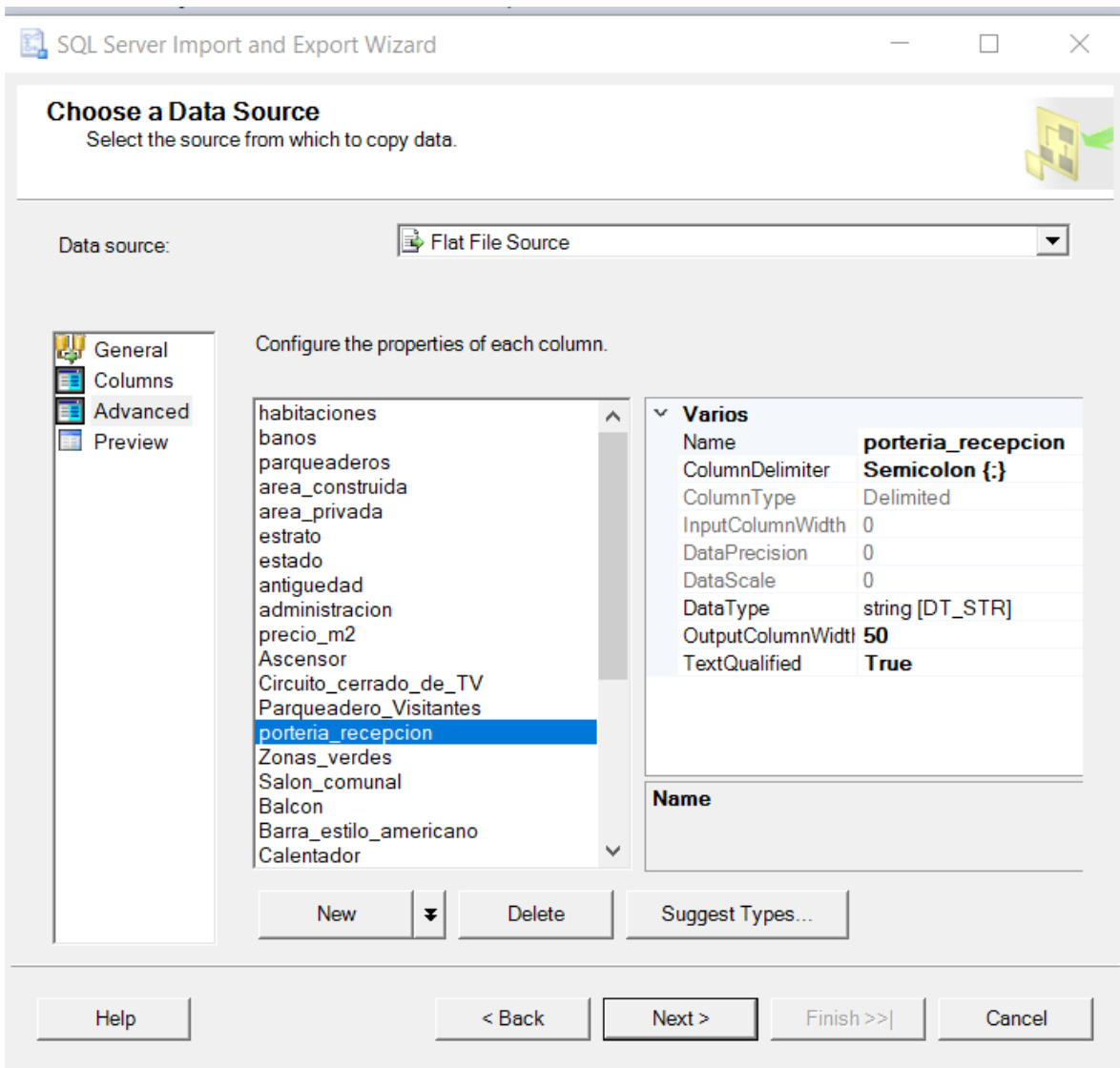
Se conformarán grupos de máximo 3 estudiantes asignados por cada computador disponible en la sala de sistemas asignada. El taller en clase se apoya en las siguientes actividades:

1. Leer con detalle la lectura 6.
2. Leer con detalle la lectura 7.

3. Desarrollar los lineamientos de IBM en lo relacionado con “data understanding” para un el caso de estudio que se plasma mediante el conjunto de datos “housing_fincaraiz.csv”. Esto implica hacer limpieza de los datos iniciales, describirlos y explorarlos, así como realizar la respectiva verificación de calidad de estos. Opcional, podrían utilizar un dataset que descarguen durante la actividad de web scraping.
4. Deben preparar un informe en PDF que evidencie el trabajo realizado. Así mismo, debe quedar evidencia del código en su perfil de Github.

CODIGO Y MODIFICACIONES EN SQL

Modificación de estructura de las columnas desde SQL



SQL Server Import and Export Wizard

Choose a Data Source
Select the source from which to copy data.

Data source: Flat File Source

Configure the properties of each column.

Columns:

- habitaciones
- banos
- parqueaderos
- area_construida
- area_privada
- estrato
- estado
- antiguedad
- administracion
- precio_m2
- Ascensor
- Circuito_cerrado_de_TV
- Parqueadero_Visitantes
- porteria_recepcion**
- Zonas_verdes
- Salon_comunal
- Balcon
- Barra_estilo_americano
- Calentador

Properties for **porteria_recepcion**:

Name	Value
ColumnDelimiter	Semicolon (;)
ColumnType	Delimited
InputColumnWidth	0
DataPrecision	0
DataScale	0
DataType	string [DT_STR]
OutputColumnWidth	50
TextQualified	True

Buttons: New, Delete, Suggest Types...

Navigation: Help, < Back, Next >, Finish >>, Cancel

LIMPIEZA DE DATA

```

SELECT
-- se eliminan duplicados en la data
distinct
habitaciones
,banos
,parqueaderos
-- Se quitan medidas y puntos de la columna
,replace(replace(area_construida,'mÂ²',''),'.','') area_construida
-- Se quitan medidas y puntos de la columna
,replace(replace(area_privada,'mÂ²',''),'.','') area_privada
,estrato
,estado
-- Se hace un replace para corregir el error de la palabra años, más y año
,replace(replace(replace(antiguedad,' aÃ±os',' Años'),'mÃ¡s','Mas'),'aÃ±o','Año')
antiguedad
-- Se hace un replace para corregir errores de estructura, adicionalmente se usa el
TRIM para quitar espacios dentro del campo
,trim(replace(replace(replace(administracion,'$Â',''),' COP',''),'.',''))
administracion
-- Se hace un replace para corregir errores de estructura
,replace(replace(replace(replace(precio_m2,'$Â',''),'mÂ²',''),'*',''),'.','')
precio_m2
,Ascensor
,Circuito_cerrado_de_TV
,Parqueadero_Visitantes
,porteria_recepcion
,Zonas_verdes
,Salon_comunal
,Balcon
,Barra_estilo_americano
,Calentador
,Chimenea
,Citofono
,Cocina_integral
,Terraza
,Vigilancia
,Parques_cercanos
,Estudio
,Patio
,Deposito_bodega
,nombre
-- Se realiza un replace para quitar los errores de tildes de las ubicaciones,
adicionalmente un Lower para manejar todo en minúscula
,lower(replace(replace(replace(replace(replace(ubicacion,'Ã¡','a'),'Ã³','o'),'ã@','e'),
'ä±','u'),'ã-','i')) ubicacion
,precio
FROM [Seminariobigdata].[dbo].[finca_raiz]

```