Calidad de Datos
Universidad de Buenos Aires
Primer Cuatrimestre de 2020
Trabajo Práctico Final

Informe Gerencial

Diego Santos, Alvaro Machicado & Maximiliano Cabezón Alvarez

Se utilizo el dataset de `Obras Registradas Junio Septiembre 2019` provisto por el `gobierno de la Ciudad de Buenos Aires`. El cual contiene infomación catastral sobre las obras de construcción realizadas en la Ciudad de Buenos Aires durante el período mencionado,

long	lat	nro exp	nomenclacion par	fecha registro p	direccion	direccion normalizada

tipo obra	metros cuadrados obra	calle nombre	calle altura	barrio	comuna	comuna

codigo postal	codigo postal argentino

El dataset no tiene un procesamiento previo, ni posee un diccionario de datos asociado. Los principales problemas detectados son columnas con valores de registro duplicados, datos faltantes, información concatenada dentro de un mismo campo que dificulta su explotación en insights y un orden de las columnas que dificulta la rápida interpretación del set.

Para mejorar la calidad del dataset se tomaron las siguientes medidas

1. Se creo un diccionario de datos, utilizando la información de dataset similares.
2. Se renombraron las columnas con nombres declarativos.
3. Se reordenaron las columnas para facilitar la comprensión de los datos representados.
4. Las columnas duplicadas fueron reorganizadas sin perdidas de información.
5. Las columnas que representaban más de un atributo se separaron.

El resultado fue un set de datos con la estructura.

Expediente	FechaRegistro	Calle	Numero	AlturaObra	Zona	Seccion	Manzana

Parcela	Barrio	Comuna	CodigoPostal	Tipo	MetrosCuadrados	Longitud	Latitud

Se realizó un análisis univariado de los atributos identificandose datos erroneos y missing. Para remediar la situación.

1. Se corrigieron los errores de carga en los campos numéricos

2. Donde fue posible se completaron los valores `missing` utilizando servicios externos.

Posteriormente se realizó un análisis bivariado de los atributos para establecer correlaciones entre los datos.

Los criterios utilizados para la limpieza del dataset mejoran la calidad de los datos ya que

1. Se eliminó la ambiguedad en la notación de las calles.
2. Se desglozó la información catastral.
3. Se ordenaron las columnas de forma tal que pueden identificarse de forma clara lo que los datos representan.
4. Mantener solo el código postal argentino sirve para facilitar la comprensión de usuarios no familiarizados con la nomenclatura interna de la Ciudad.

Y permite distinguir las reglas del negocio facilitando la detección de insights. De manera ordenada pueden identificarse:

1. La ubicación exacta de una obra conociendose no solo la calle y la altura.
2. Permite identificar especificamente la seccion, manzana y parcela afectadas por la obra.
3. La columna alturaObra permite identificar las parcelas con más de una numeración asociada.
4. Identificar comuna, barrio, calle, sección, manzana y parcela con más obras.
5. La cantidad de metros cuadrados por comuna, barrio, calle, sección, manzana y parcela.
6. Detallar los tipos de obra que se llevan a cabo por comuna, barrio, calle, sección, manzana y parcela.