Programación para analítica de datos

Semestre: 2025 - S1

Profesor: Jorge Victorino

Trabajo 2, Calidad de los datos

El objetivo es realizar el ejercicio grupal en donde se puedan aplicar las diferentes técnicas y prácticas en la *Calidad de los datos*. A continuación se describen la base de datos a utilizar en el trabajo:

Datos:

A continuación se presenta el diccionario de datos de el dataset de ejemplo:

Tabla Incidentes:

- **Tipo**: Determina en algunos casos la naturaleza del tipo de incidente
- **Creación**: Fecha de creación del registro del incidente
- **Inicio**: Fecha de inicio de atención al incidente
- **Estado**: Siglas de los diferentes estados por los que pasa un incidente
- **Prioridad**: Tiempo en el cual se debería resolver el incidente
- **Incidente**: código del sistema para el incidente
- **Estructura**: código de la estructura afectada

Tabla Estructura:

- **Serie**: número de serie del equipo
- **Latitud**: posición relativa vertical
- **Longitud**: posición relativa horizontal
- **Estructura**: código de la estructura existente
- **Depto**: Departamento donde se ubica la estructura
- **Municipio**: Municipio donde se ubica la estructura
- **Vereda**: Vereda donde se ubica la estructura
- **Cia**: Compañía dueña de las estructuras

Calidad de datos

La base de datos tiene como pilar fundamental analizar los incidentes que ocurren en las estructuras de las empresas de energía. Pero no todos los incidentes son de interés solo se analizará los que están asociados a problemas de invasión: Estos se pueden identificar en la tabla de incidentes con los estados: GPRE, PRER, TAMB, GEAM, AMPO, y GERE. La tabla estructuras está relacionadas con los incidentes y se deben filtrar con los incidentes de interés. A continuación se enumeran las actividades del ejercicio que se debe realizar:

1 Análisis preliminar

Lo primero es realizar un análisis de los duplicados, los valores nulos y la información errónea en cada tabla antes de cruzarlas. Luego se debe verificar que los campos a utilizar como llaves para cruzar la información sean consistentes, es decir que no tengan valores nulos o duplicados. Al final debe quedar un incidente por fila en la tabla incidente y una estructura por fila en la tabla estructura.

Tabla incidentes:

- Filtrar los incidentes que contengan alguno de los siguientes estados:
    - GPRE: Gestión predial
    - PRER: Predio revisado
    - TAMB: Trámite ambiental
    - GEAM: Gestión ambiental
    - AMPO: Amparo policivo
    - GERE: Gestión realizada
- Analice los duplicados y tome una decisión argumentada de cuál registro debería quedar. Determine cuáles son los duplicados y cuántas veces se repiten. Mostrar un resumen del análisis con la información relevante.
- Análisis de valores nulos después de filtrar la información de interés. Mostrar un resumen del análisis con la información relevante de los valores nulos y sus posibles causas.
- Verificar el tipo de dato adecuado para cada variable y corregir valores erróneos, como quitar los espacios en blanco.

- Agregar variables que indiquen: si un incidente se encuentra abierto o cerrado, si está a tiempo o está atrasado, cuántos días abierto, si cumplió o no la prioridad, el año, el mes.

Tabla Estructuras:

- Al parecer hay estructuras que no tienen número. Analizar si corresponden a estructuras válidas y asignar un número.
- Definir que sería una estructura duplicada, detectar duplicados, eliminar y corregir en caso de ser necesario
- Analizar los datos nulos y determinar si se puede corregir
- Corregir problemas de codificación de caracteres y espacio en blanco sobrantes
- Corregir valores erróneos de municipios, departamentos o veredas.
- Agregar variables para la geolocalización como ubicación (latitud y longitud), municipio con departamento y país.

Realizar los cruces de información (no se deben cruzar todas las tablas)

- Incidentes - Estructuras

Al realizar los cruces, debe analizar la información con los datos que quedan por fuera de la combinación (por ejemplo: números de estructuras que aparecen en incidentes y que no aparecen en la tabla de estructuras, y estructuras que no tienen incidentes). Resultado, tablas con registros filtrados.

2 Dividir las variables por tipo

Dividir variables en categóricas, numéricas y de fecha. Hacer una descripción detallada de cada variable después del filtrado del punto 1 según aplique: cantidad de valores diferentes, valores nulos, valor máximo, valor mínimo, valor más frecuente, frecuencia del valor más frecuente, media, mediana, desviación estándar, tiempo entre fecha mínima y máxima. Entienda las variables en el contexto del negocio y su importancia, argumente en cada caso por qué se pueden presentar los resultados mostrados en cada tabla. Resultado: tablas que describen las variables organizadas por tipo.

3 Depurar la escritura de la base de datos

Corregir errores de digitación o de codificación para el caso de las tildes, busque espacios en blanco que no corresponden en valores y nombres de columna. Analice las categorías resultantes y verifique que guarden coherencia. Tablas de las bases de datos depuradas.

4 Hacer análisis univariado

Con variables numéricas y de fecha y hora hacer: gráficas de distribución (violines y cajas) analizar cada gráfica, ubicar valores atípicos y valorar la ocurrencia de ellos. Con variables categóricas hallar tablas de frecuencia con porcentaje y graficar con barras. Resultados: tablas y gráficas de las variables con su respectivo análisis relacionado con su entendimiento del negocio.

5 Análisis bivariado

Para las numéricas hacer análisis de correlación con gráficas de dispersión (con líneas de tendencia) utilizando color con el tipo de incidente o estado del caso (abierto o cerrado), y hacer el test de Chi-cuadrado para las categóricas. Es importante valorar las relaciones bivariadas en los gráficos de dispersión y detectar los valores atípicos bivariados. Resultado: análisis de las relaciones a partir de los gráficos bi-variados para identificar reglas del negocio.

6 Análisis de fechas

La fecha de corte de los datos es 6 de mayo de 2019, los incidentes que no tienen fecha de cierre están sin resolver, es decir están abiertos. El objetivo en este punto es hacer un análisis de fecha y prioridad, y responder las siguientes preguntas calculando nuevas columnas y gráficamente: qué incidentes están abiertos, cuál es el tiempo de vida de un incidente, cuánto tiempo tardó en iniciar la solución, determinar si se cumplió la prioridad, cuánto tiempo de desfase hubo (entre lo que duró y lo que debería durar según prioridad), calcular fecha límite del cierre, cuántos incidentes abiertos están vencidos o por vencerse, determinar inconsistencias en las fechas. Hacer gráficas de cumplimiento de la prioridad discriminar por tipo de incidente. Resultado: reporte con gráficas, tablas y análisis.

7 Análisis espacial de los incidentes

Analizar la ocurrencia de incidentes por tipo de incidente, y espacialmente por departamento, municipio, vereda y por la posición de cada estructura. Interesa revisar por unidad espacial cuántos incidentes están abiertos, cerrados o cuántos están vencidos o activos. Además identificar zonas en las cuales hay mayor incidencia de algún tipo de incidente. Realizar mapas coropléticos mostrando el estado de las variables (mayor

incidencia, mayor cantidad de un tipo de incidentes). Resultado: reporte con gráficas, tablas y análisis.

8 Organice los datos para el modelado

A partir del entendimiento del negocio y de los datos proponga un hallazgo que consideren interesante para presentar como elemento diferenciador del trabajo.

Pautas del Trabajo

1. La entrega es un notebook de python (solo un archivo) que ustedes descargan y que adjuntan a la actividad en la plataforma con la fecha límite que aparece en la plataforma. Los trabajos entregados después de esta fecha y hora no serán calificados.
2. Realizar una presentación de 20 minutos de los aspectos más relevantes y del último punto. Cada exposición y cada trabajo tendrá un énfasis específico en los puntos 6, 7, 8, o 9 previamente acordados. Todos los integrantes del grupo deben tener una participación similar en la exposición.
3. El notebook lleva un encabezado con la portada del trabajo (logo ucentral, integrantes, profesor, materia y grupo). El trabajo se organiza en secciones una de cada punto con su respectivo desarrollo.
4. En el reporte se deben usar elementos de interfaz gráfica que facilitan la interacción para el análisis de los datos. Las gráficas deben tener títulos, etiquetas y leyendas si es necesario.
5. Cumplir las normas de calidad, como correcta ortografía, correcta redacción, documentación de las funciones, calidad de los gráficos y una excelente presentación. Verifique que todos los gráficos tienen los 3 títulos: (eje X, eje Y, gráfica) y deben tener descripción que indica cuál es la intención del gráfico y los hallazgos más importantes.
6. El notebook solo debe usar librerías de python válidas en “google colab” y los archivos de entrada deben ser los originales (no generar archivos parciales en excel). Antes de entregar verifique que todas las celdas de código se ejecutan sin errores, para esto reinicie el entorno de ejecución y vuelva a ejecutar todas las celdas en orden para verificar.

Evaluación del trabajo.

1. Problemas de calidad en el desarrollo de cada punto reciben una penalización independiente de 10% por cada uno de los siguientes literales:
    a. Un error de ortografía o redacción.
    b. Gráficas sin títulos o sin la respectiva descripción.
    c. Si aparece algún error de ejecución en una de las celdas de código.
2. El uso de elementos de interfaz gráfica en el notebook puede subir la calificación hasta en el 15% (hasta un máximo del 100%, obviamente)
3. El último punto puede aportar 10% adicional al trabajo si lo amerita.