Resumen ejecutivo

El dataset que se analizó contiene 62.628 registros y 31 columnas con información sobre quejas o consultas registradas, probablemente de la región de Cundinamarca, esto lo digo por el nombre del archivo CSV que se usó en el notebook. En general, los datos están bastante bien, con pocas o ninguna pérdida en la mayoría de las columnas que revisamos, y abarcan desde 2009 hasta 2023. Las pruebas que hicimos para ver si las variables numéricas seguían una distribución normal mostraron que no es así. Después de usar el método IQR para buscar y eliminar valores atípicos, el número de registros se mantuvo en 62.628, lo que sugiere que no se eliminaron filas por valores atípicos según los criterios que utilizamos.

Metodología (qué contiene el notebook)

En el notebook, realizamos las siguientes tareas principales de forma secuencial:

Importamos las librerías que usamos normalmente, como pandas, numpy, matplotlib y seaborn, y cargamos un archivo CSV, que en el entorno original era /Quejascundinamarca.csv.

Mostramos información general del DataFrame usando df.info(), mostramos las primeras filas y usamos df.describe() para obtener estadísticas de las variables numéricas.

Calculamos y mostramos cuántos valores faltantes hay en cada columna, y también creamos un mapa de calor para visualizar los datos faltantes.

Seleccionamos los subconjuntos que nos interesaban, es decir, las columnas que queríamos analizar, mostramos los encabezados y los tipos de variables.

Creamos visualizaciones como histogramas, gráficos de barras y diagramas de caja (boxplots) para las variables más importantes, como AÑO, MES, ITEM, TIPO PETICIÓN, etc., y también generamos matrices de correlación y mapas de calor.

Detectamos valores atípicos utilizando el método IQR (Q1/Q3) en las variables numéricas y los filtramos.

Realizamos pruebas de normalidad, como Shapiro-Wilk, Kolmogorov–Smirnov, Anderson–Darling y Jarque–Bera, en las variables numéricas que habíamos seleccionado.

Generamos varias figuras, como histogramas, boxplots y mapas de calor, que guardamos y mostramos en el notebook como resultados.

Hallazgos clave (detallados)

1. Tamaño y estructura

Registros: 62.628 filas.

Columnas: 31 columnas.

Algunos ejemplos de las columnas que encontramos son: ITEM, FECHA RADICACIÓN, AÑO, MES, TIPO PETICIÓN, CANAL, USUARIO, FECHA VENCIMIENTO, TIPO RESPUESTA, FECHA RESPUESTA FINAL, DEPENDENCIA PRINCIPAL, COLOR ESTADO, ESTADO RESPUESTA, MUNICIPIO, VEREDA, SUELO, FLORA, FAUNA, AIRE, AGUA, etc.

2. Estadística descriptiva (variables numéricas)

ITEM: el promedio es de aproximadamente 31314.5, el valor mínimo es 1 y el máximo es 62628. Probablemente sea un identificador que se incrementa con cada registro.

AÑO: el promedio es de aproximadamente 2016.51 y los valores están entre 2009 y 2023.

MES: el promedio es de aproximadamente 6.46 y los valores están entre 1 y 12.

Vemos que las desviaciones estándar de AÑO y MES son razonables, por ejemplo, la desviación estándar de AÑO es de aproximadamente 4.32.

3. Valores faltantes

En el notebook, mostramos cuántos valores faltantes hay en cada columna. En la mayoría de las columnas, las que imprimimos en la salida, el conteo es 0. Parece que algunas columnas tienen espacios en blanco o entradas vacías, como AGUA, donde parece faltar contenido en la línea, pero en general no hay faltantes importantes en las columnas principales.

4. Distribución y normalidad

Aplicamos varias pruebas de normalidad, como Shapiro–Wilk, Kolmogorov–Smirnov, Anderson–Darling y Jarque–Bera.

El resultado fue que la mayoría de las variables numéricas que probamos NO siguen una distribución normal, ya que los valores p son cercanos a 0 en las pruebas. Por lo tanto, no es adecuado asumir normalidad para los procedimientos que lo requieran sin hacer transformaciones.

5. Outliers

Calculamos el IQR (Q1, Q3) y filtramos los valores atípicos según el criterio clásico. Después de eliminar los valores atípicos, el conteo final fue de 62.628 datos restantes, que es el mismo que el dataset original. Esto sugiere que:

No se detectaron valores atípicos con ese criterio,

O la implementación del filtro no eliminó filas, lo que podría indicar un error lógico en el filtrado,

O las columnas numéricas que analizamos no tenían valores extremos fuera de los límites IQR definidos.

6. Visualizaciones

Histogramas de AÑO, MES y otras variables numéricas.

Boxplots de AÑO, ITEM y otras variables categóricas versus numéricas.

Mapa de calor de valores faltantes, que muestra que casi no hay faltantes.

Matriz de correlación y mapa de calor, aunque al haber muchas variables categóricas, la correlación entre las numéricas puede ser limitada.

Interpretación y consecuencias prácticas

Amplia cobertura temporal: los datos abarcan desde 2009 hasta 2023, lo que permite hacer análisis temporales y ver tendencias anuales o mensuales.

Calidad de datos aceptable: la ausencia de faltantes en las columnas clave facilita el análisis cuantitativo sin necesidad de hacer limpiezas exhaustivas.

No normalidad: para hacer comparaciones o pruebas estadísticas que asumen normalidad, como la prueba t paramétrica o ANOVA, es mejor usar pruebas no paramétricas o transformar las variables.

Outliers: la aparente ausencia de valores atípicos extremos puede ser real, pero deberíamos revisar el código que aplicó el filtro para asegurarnos de que el filtrado se aplique a las columnas y filas correctas.

Recomendaciones y próximos pasos (priorizadas)

Verificar el análisis de fechas y crear métricas de tiempo

Convertir FECHA RADICACIÓN, FECHA VENCIMIENTO y FECHA RESPUESTA FINAL a formato datetime.

Calcular tiempo_respuesta = FECHA RESPUESTA FINAL - FECHA RADICACIÓN (en días) para analizar los tiempos de respuesta por año, mes, tipo de petición, dependencia y municipio.

Esto permite calcular KPIs como la mediana o los percentiles de tiempo de respuesta, la proporción de respuestas fuera de término, etc.

Revisar el procedimiento de detección de outliers

Confirmar en qué columnas se calculó el IQR.

Verificar que el filtrado se aplicó correctamente, ya que podría haber errores lógicos que hicieron que se devolviera el dataset original.

Considerar los valores atípicos por variable y también detectar valores atípicos multivariantes, por ejemplo, usando Mahalanobis o clustering.

Análisis temporal y estacional

Crear series temporales: contar mensualmente las quejas por TIPO PETICIÓN y por MUNICIPIO.

Identificar picos o caídas en ciertos años o meses y relacionarlos con eventos o regulaciones locales.