INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE INGENIERÍA EN COMPUTACIÓN

IC4302 BASES DE DATOS II GRUPO 20

Proyecto 3: Extraer, Transformar y Cargar Datos

SparkSQL

Profesor:

Alberto Shum Chan

Elaborado por:

Andrés Felipe Arias Corrales

2015028947

Diego Esteban Castro Chaves

200419896

Fabián José Fernández Fernández

2022144383

Hengerlyn Araya Nash

2021051418

Fecha de Entrega:

16 de noviembre del 2024

II semestre, 2024

Índice

Descripción General del Sistema

El sistema desarrollado es una aplicación escrita en Python que utiliza el framework Apache Spark para la manipulación, análisis y despliegue de datos, integrándose con el entorno de desarrollo Jupyter Lab en formato de Jupyter Notebook. Adicionalmente, se emplea una base de datos en PostgreSQL para el almacenamiento y análisis estructurado de los datos procesados. Como entorno de desarrollo, se eligió Visual Studio Code (VS Code) debido a su compatibilidad con Jupyter Lab y herramientas avanzadas para programación. Además, se utiliza GitHub como repositorio central para el sistema de control de versiones, facilitando la colaboración y la gestión del proyecto.

El objetivo principal de la aplicación es analizar la posible correlación entre datos demográficos, de empleo y participación económica, en comparación con datos de actividad criminal en distintas zonas del país delimitadas por distritos. Las fuentes de datos empleadas son el censo nacional de 2011 proporcionado por el INEC y los registros de actividad criminal recopilados por el OIJ para el mismo año.

Objetivo del Proyecto

El propósito del proyecto es integrar y analizar datos recopilados por dos entidades distintas, que emplean metodologías diferentes, pero que comparten un marco común de referencia: la división geográfica del país. A pesar de que ambos conjuntos de datos utilizan la misma delimitación administrativa, pueden existir discrepancias en la forma en que los datos son presentados (nombres de distritos, formatos, etc.). Por esta razón, es necesario realizar procesos de limpieza, estandarización y unificación de los datos para asegurar una comparación adecuada y precisa.

Procesos de Transformación y Análisis

Para lograr la integración y análisis de los datos, se implementaron diversas funciones que incluyen:

1. Limpieza y estandarización:
    - Eliminación de caracteres innecesarios, como tildes, espacios y mayúsculas, para garantizar la uniformidad.
    - Normalización de nombres de distritos y variables clave para asegurar la interoperabilidad entre los conjuntos de datos.
2. Emparejamiento de datos:
    - Desarrollo de funciones para verificar el grado de correspondencia entre las divisiones administrativas en ambos conjuntos de datos.
    - Algoritmos para maximizar el número de datos coincidentes entre las dos fuentes.

3. Almacenamiento:
    - Una vez procesados, los datos se almacenan en una base de datos PostgreSQL, asegurando su disponibilidad para análisis posteriores y facilitando la persistencia de los datos depurados.
4. Visualización y análisis:
    - Dentro del entorno de Jupyter Notebook, se desarrollaron gráficos que permiten explorar visualmente las posibles correlaciones entre diferentes variables. Ejemplos de análisis incluyen:
        - Relación entre la cantidad de delitos y la tasa de ocupación económica.
        - Distribución de delitos por día de la semana.

Beneficios del Sistema

El sistema propuesto permite:

- Integrar y limpiar datos complejos provenientes de fuentes heterogéneas.
- Detectar correlaciones significativas entre indicadores sociales y actividad criminal.

Esta solución aporta valor al abordar problemas de calidad de datos en investigaciones basadas en múltiples fuentes y permite explorar hipótesis relevantes para la sociedad.

Descripción de las Transformaciones

Conexión a Base de Datos PostgreSQL

Para la conexión con la base de datos PostgreSQL utilizamos el controlador Psycopg 3, el cual supone una conexión sencilla y transparente con la base de datos. Consiste en brindarle los datos de conexión (Nombre de la base de datos, usuario, password, host y puerto) con los cuales se crea la conexión mediante la función psycopg.connect(**conn_params).

Eliminación de espacio en blanco

Tiene como objetivo limpiar y estandarizar los valores de las columnas relacionadas con distritos en los DataFrames del OIJ e INEC. El proceso elimina los espacios en blanco dentro de los nombres de distritos, asegurando una uniformidad en los datos y facilitando su uso en comparaciones y análisis.

Se obtiene como resultado:

- DataFrame del OIJ:
  - La columna Distrito está limpia de espacios en blanco.
  - Se elimina la columna _C11 si está presente.
- DataFrame del INEC:
  - La columna distrito está limpia de espacios en blanco.

Conversión a minúsculas

Para la conversión a minúsculas se utiliza la función lower para transformación de cadenas de caracteres a minúsculas.

En nuestro caso creamos una función llamada convertir_minusculas(df, columna) la cual permite convertir cualquier columna de texto en un dataframe a minúsculas.

Creación de Tablas

En la creación de tablas vamos a pasar por dos procesos para cada uno de los data frames: la creación de la tabla vacía, y el llenado con los datos.

En este caso usamos dos funciones básicas que aplica para cada uno de estos procesos:

crear_tabla_sql(df, nombre_tabla)

Esta función va a retornar un string con el script en sql con los datos que va a llevar la tabla. Para esto mapea los headers del data frame e identifica el tipo de dato que contiene. Con esto va creando una lista que contiene las columnas correspondientes al data frame que se van a crear en la tabla dentro de la base de datos.

Insertar_datos_postgres(df, nombre_tabla, conn_params)

Esta función va a recorrer el data frame fila por fila recopilando los datos, captura los datos de los nombres de las columnas del data frame y crea los scripts para insertar los datos en la base de datos. Al final se conecta a la base de datos e inserta todos los datos recopilados dentro de la tabla correspondiente y muestra un mensaje de terminación exitosa cuando se pudo insertar todo.

Luego en este bloque de código se procede a crear la tabla y llenar los datos correspondientes a los data frames en el sistema.

Eliminación de tildes

Se enfoca en la normalización de los valores de la columna distrito del conjunto de datos INEC, eliminando tildes de los nombres de los Distritos. Esto es fundamental para garantizar la uniformidad de los datos y evitar discrepancias causadas por diferencias ortográficas al realizar análisis o comparaciones con el conjunto de datos del OIJ.

Después de aplicar esta transformación, los valores de la columna distrito ya no contienen tildes, facilitando la uniformidad y la interoperabilidad con otros conjuntos de datos. Garantizando que nombres como San José y San Jose sean tratados como equivalentes, lo cual es crucial en operaciones de comparación o integración de datos.

Actualización de la Base de Datos

Tiene como objetivo normalizar los valores de la columna distrito en la tabla INEC, almacenada en una base de datos PostgreSQL. El proceso elimina tildes directamente en la base de datos mediante una consulta SQL, asegurando que los datos estén listos para comparaciones y análisis uniformes.

Listado de Distritos que no coinciden entre las Tablas

Tiene como propósito identificar los distritos presentes en el conjunto de datos del OIJ que no tienen una coincidencia correspondiente en el conjunto de datos del INEC. Para ello, se implementa una función que limpia las tildes de las columnas relevantes y realiza una operación de "*left anti join*" para obtener un subconjunto único de distritos del OIJ que no están en el INEC.

El resultado final de esta transformación es un DataFrame con los nombres únicos de los distritos en el conjunto de datos del OIJ que no tienen una contraparte en el conjunto del INEC. Esto facilita la identificación de inconsistencias o la incorporación de valores faltantes en futuras integraciones.

Cantidad de Registros de Distritos que no coinciden entre las Tablas

Tiene como objetivo identificar y contar los distritos presentes en el conjunto de datos del OIJ que no tienen una coincidencia en el conjunto de datos del INEC. Para lograrlo, la función *contar_distritos_no_coincidentes* realiza varias normalizaciones de

texto y utiliza una operación de "*left anti join*" para encontrar las discrepancias entre los dos conjuntos de datos.

Esta transformación garantiza:

- Una comparación confiable entre dos conjuntos de datos geográficos.
- La identificación de inconsistencias que pueden ser corregidas o analizadas posteriormente.
- La normalización de los datos para facilitar futuras integraciones y análisis.

Edición de nombres de los Distritos en la Tabla INEC

Tiene como objetivo alinear los nombres de los distritos en los conjuntos de datos del OIJ e INEC para facilitar comparaciones y análisis. Se logra mediante la normalización de los nombres de distritos, eliminando diferencias como tildes, mayúsculas y caracteres especiales como la letra ñ.

Se obtienen:

- *DataFrames* Normalizados: Los nombres de los distritos en los *DataFrames* del OIJ e INEC son consistentes, eliminando tildes, espacios y mayúsculas.
- Coincidencias Específicas: Una lista de distritos que cumplen con criterios particulares, como incluir términos relevantes (cañas, peñas blancas).

Descripción de las Visualizaciones

Cantidad de Delitos y Tasa de Ocupación para los 10 Distritos con más Delitos

En esta visualización mostramos un gráfico de barras con un gráfico de líneas sobrepuestos en donde podemos ver la relación de cantidad de delitos con la tasa de ocupación. Este último dato corresponde al porcentaje de la población ocupada con respecto a la población de 12 años o más (INEC, 2009).

Particularmente en esta visualización vemos que en Quepos no se registran datos de tasa de ocupación. Esto se debe a que en los datos suministrados por las dos fuentes hay una discrepancia en el nombre del cantón: originalmente se llamó Aguirre en su fundación en 1948, pero en febrero del 2015 cambió su nombre a Quepos (Colaboradores de Wikipedia, 2024) Al generar los datos desde el sistema del OIJ la información geográfica está actualizada a la fecha. Esto a diferencia de la del INEC que no es generada y nada más se descarga un archivo viejo.

Cantidad de Delitos por día de la Semana

Se muestra la distribución de delitos por día de la semana en el distrito con la mayor cantidad de delitos, identificado a partir de los datos del OIJ. La visualización resalta los días con mayor incidencia criminal en dicho distrito, proporcionando una perspectiva clave para el análisis de tendencias temporales y posibles patrones.

Se observa que:

- Los días entre lunes y viernes presentan la mayor cantidad de delitos, alcanzando su punto máximo los miércoles y jueves.
- Los domingos y sábados tienen una menor incidencia, posiblemente reflejando patrones relacionados con la actividad económica o social en el distrito.

Esto podría sugerir que los días laborales concentran mayor actividad delictiva en el distrito, lo cual podría estar relacionado con horarios laborales, flujo de personas, o incluso comportamiento rutinario de los residentes.

Cantidad de Delitos por tipo en el Distrito determinado

En esta visualización podemos determinar el distrito en que nos queremos enfocar. Se muestra un gráfico de barras horizontales en donde se muestra el tipo de delitos en el eje Y y la cantidad de los mismos en el X para poder comparar la magnitud de estos en cada comunidad.

Dentro del código se puede digitar la provincia, el cantón y el distrito deseado que corresponda a los datos dentro del sistema. En caso de que no se encuentre el distrito se le muestra al usuario un mensaje indicándole el error.

Nosotros elegimos el distrito Carmen, San José, San José como ejemplo de esta visualización sin ninguna razón específica.

Cantidad de Delitos por Sexo

Se ilustra la cantidad de delitos registrados, categorizados por el sexo de las víctimas. La gráfica se construyó a partir de los datos procesados en el conjunto de datos del OIJ, donde se agruparon los registros por la columna sexo y se contó la cantidad de delitos en cada categoría. Este análisis permite identificar tendencias relacionadas con el género de las víctimas y explorar posibles disparidades.

- El gráfico muestra que la categoría "HOMBRE" representa la mayor cantidad de delitos registrados, seguida de "MUJER".
- Una pequeña proporción de delitos tiene el sexo clasificado como "DESCONOCIDO".

Esta visualización proporciona una perspectiva clara sobre la distribución de delitos por sexo en los datos del OIJ, lo que es crucial para entender patrones y diseñar políticas públicas más inclusivas. Además, sirve como base para analizar posibles correlaciones entre género y otros factores, como el tipo de delito o la ubicación geográfica.

Cantidad de Delitos por Tipo y Hora del día

Se muestra un mapa de calor que relaciona los tipos de delitos con la hora del día en que se cometieron. Utiliza un esquema de colores para resaltar las frecuencias de delitos por combinación de hora y tipo, permitiendo identificar patrones y tendencias temporales en la actividad criminal.

- Se observa una concentración elevada de delitos tipo asalto y robo durante la noche, específicamente entre las 18:00 y las 00:00 horas.
- Los delitos como hurto muestran una distribución más uniforme a lo largo del día, con un ligero incremento en las tardes.
- Las horas de la madrugada (03:00 a 06:00) presentan la menor cantidad de delitos en general, independientemente del tipo.

El mapa de calor permite identificar claramente las horas de mayor actividad criminal para cada tipo de delito, proporcionando información clave para la planificación de estrategias de seguridad pública. Además, destaca la necesidad de explorar correlaciones con otras variables, como días de la semana o ubicaciones específicas, para profundizar en el análisis de los datos.

Particularmente hay una diferencia marcada con los asaltos en el tramo de las 18 horas y las 21. Podríamos suponer que corresponde a la salida del trabajo de muchas personas. Si nos sorprende el dato del robo entre las 0 horas y las 3 de la mañana ya que sobrepasa por mucho a los demás delitos en otros momentos del día.

Conclusiones

La calidad de los datos afecta directamente el análisis. Las discrepancias en la nomenclatura administrativa y las diferencias en la calidad de los datos entre las fuentes resaltaron la importancia de los procesos de limpieza y normalización para garantizar una integración adecuada. Sin estas transformaciones, habría sido imposible realizar un análisis confiable de las correlaciones entre factores socioeconómicos y criminalidad (Instituto Nacional de Estadística y Censos [INEC], 2011; Organismo de Investigación Judicial [OIJ], 2011).

El análisis también identificó un aumento significativo en la actividad delictiva durante las noches y días laborales, especialmente entre las 18:00 y 00:00 horas. Este patrón podría estar relacionado con la rutina laboral de las personas, sugiriendo que los horarios de mayor flujo de población coinciden con los picos delictivos (Fallas-Valenciano & Treminio, 2016).

Además, los datos del OIJ evidencian que los hombres son las principales víctimas de delitos, mientras que las mujeres representan un porcentaje menor, aunque significativo. Las diferencias en el tipo de delitos por género subrayan la necesidad de diseñar políticas públicas que consideren las dinámicas específicas de género en la actividad criminal (OIJ, 2011).

Finalmente, los resultados mostraron una posible correlación entre la tasa de ocupación económica y la incidencia de delitos en algunos distritos. Sin embargo, las variaciones geográficas, como en el caso de Quepos, destacan la importancia de contar con datos actualizados y uniformes para evitar interpretaciones erróneas y mejorar la toma de decisiones informadas (INEC, 2011; OIJ, 2011).

Referencias

1. colaboradores de Wikipedia. (2024, January 9). *Cantón de Quepos*. Wikipedia, La Enciclopedia Libre. https://es.wikipedia.org/wiki/Cant%C3%B3n_de_Quepos
2. INEC(2009) Conceptos y Definiciones https://inec.cr/wwwisis/documentos/INEC/EHPM/EHPM_2009/EHPM_2009_Conceptos_Definiciones.pdf
3. Instituto Nacional de Estadística y Censos (INEC). (2011). *Censo Nacional de Población y Vivienda*.
4. Organismo de Investigación Judicial (OIJ). (2011). *Registros de actividad criminal en Costa Rica*.
5. Fallas-Valenciano, E., & Treminio, C. (2016). Violencia de mujeres en Costa Rica: un problema para hacer conciencia. *Revista Hispanoamericana De Ciencias De La Salud*, *2*(2), 190–191. Recuperado a partir de https://uhsalud.com/index.php/revhispano/article/view/154