Instituto Tecnológico de Costa Rica

Centro académico de Alajuela

IC-4302. Bases de datos II

Semestre II – 2022

Proyecto III

Prof. Maria Auxiliadora Mora Cross.

Estudiantes:

Brandon Sthuar Retana Chacón.

Carné: 2021121141

Correo: sthuar@estudiantec.cr

Kevin Cubillo Chacón

Carné: 202123138

Correo: kevincubillo@estudiantec.cr

Fecha de Entrega: 25/11/2022

Descripción del sistema:

El sistema es una implementación de un ETL, esta implementación está creada en el lenguaje de programación Python, con el framework de Spark y para su escritura utilizamos JupyterNotebooks. El sistema se encarga de cargar datos de dos diferentes csv, uno proveniente de las bases de datos del INEC, y otros del OIJ, los cuales se cargan en distintos DataFrames para el debido procesamiento de los datos. Lo primero que realiza el sistema es la carga de los csv para posteriormente eliminar los espacios en blanco al inicio y al final de los textos, se pasa todo a letras minúsculas, se encuentra una lista con las palabras las cuales no hace match entre los dos DataFrames, se cambian todas las letras que contienen acentos en español y se separa en columnas.

Una vez tratados los DataFames se guardan en una base de datos Postgres, que será consultada con Pandas para generar algunas visualizaciones sobre la criminalidad en el país.

Descripción las funciones:

remove_spaces(df):

Esta función recibe un DataFrame, luego recorre cada una de sus filas y le aplica un trim a las columnas ‘Provincia’, ‘Canton’ y ‘Distrito’ y de este modo eliminar los espacios en blaco al inicio y fin de la palabra. Finalmente retorna el DataFrame modificado.

to_lower_case(df):

Esta función recibe un DataFrame, luego recorre cada una de sus filas y le aplica la función lower() de Spark a las columnas ‘Provincia’, ‘Canton’, ‘Distrito’ y de este modo pasar a minúscula todas las letras de cada palabra. Finalmente retorna el DataFrame modificado.

replace_accents(df):

Esta función recibe un DataFrame, luego recorre cada una de sus filas y le aplica la función regexp_replace() de Spark a las columnas ‘Provincia’, ‘Canton’ y ‘Distrito’ para eliminar las tildes y la letra ‘ñ’ cambiarla por la ‘n’ y de este modo los datos

hagan match con su contraparte del OIJ. Finalmente retorna el DataFrame modificado.

generate_new_columns(df):

Esta función separa la columna ‘Provincia, cantón’ y distrito’ del DataFrame del INEC en tres diferentes, de modo que se pueda realizar el join con las columnas del OIJ.

find_non_matches(df1, df2):

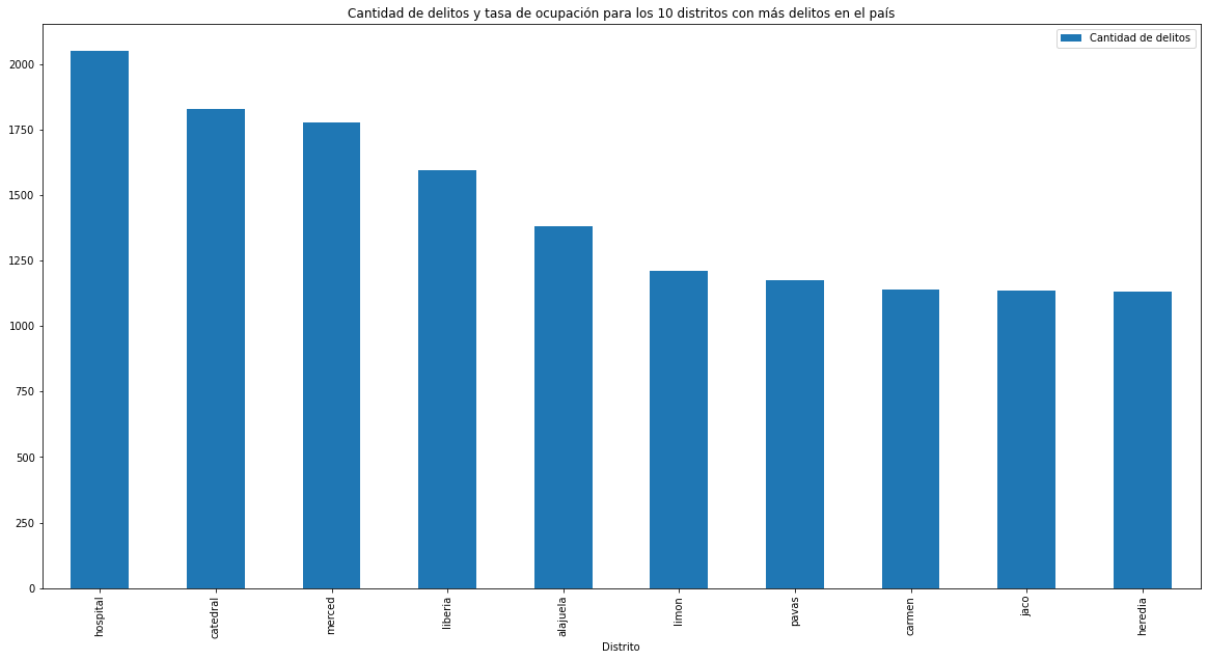
Esta función recibe los dataframes del OIJ e INEC, luego con la función subtract() de spark se le eliminan los registros que hacen match en las tres columnas, finalmente se recorre el DataFrame generado y se insertan los registros en una lista que será retornada.

number_of_non_matches(df1, df2):

Esta función recibe los dataframes del OIJ e INEC, luego con la función subtract() de spark se le eliminan los registros que hacen match en la columna ‘Distrito’, finalmente se recorre el DataFrame generado y se cuentan los registros para luego retornar el valor.

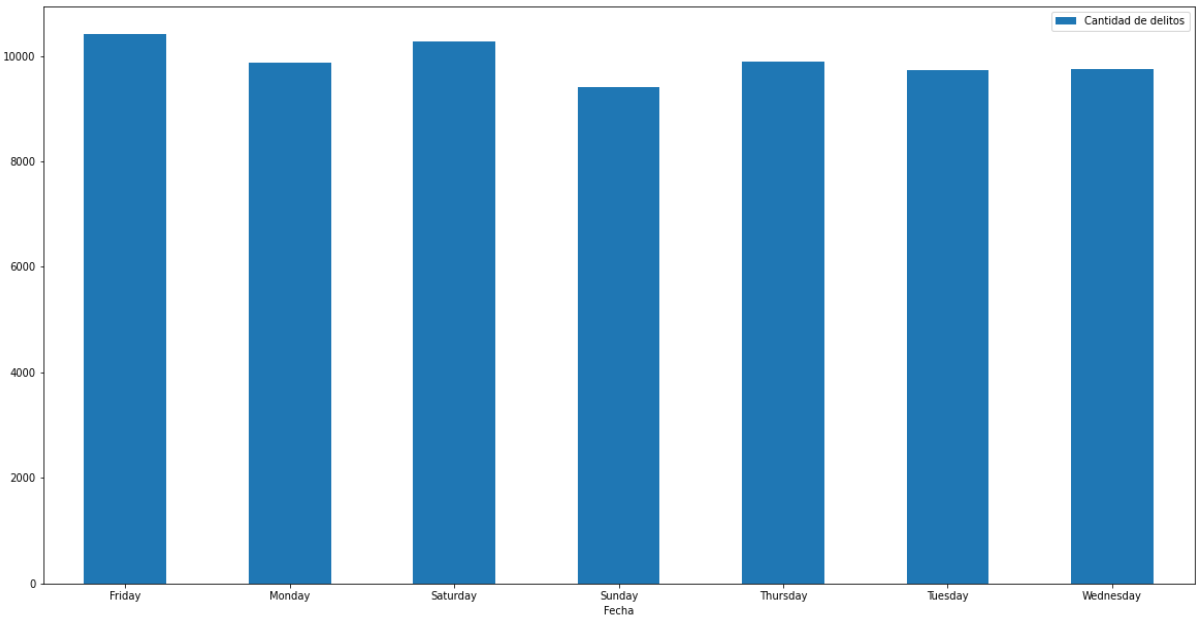
Descripción de las visualizaciones

1.



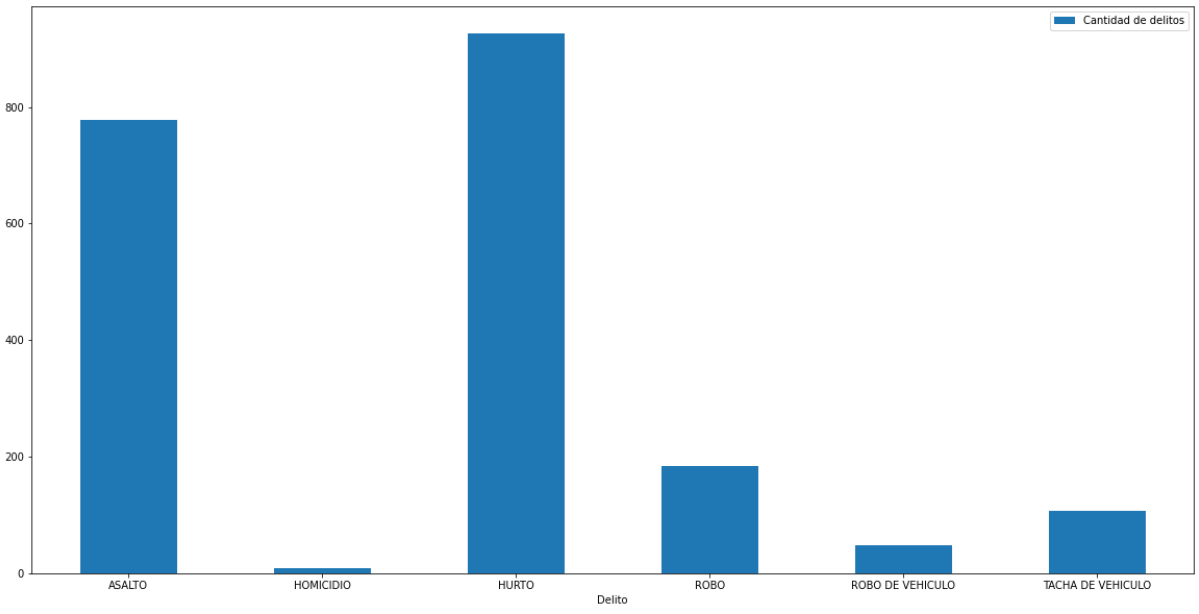
Esta visualización permite comparar la cantidad de delitos y la tasa de ocupación de los diez distritos más conflictivos del país.

2.



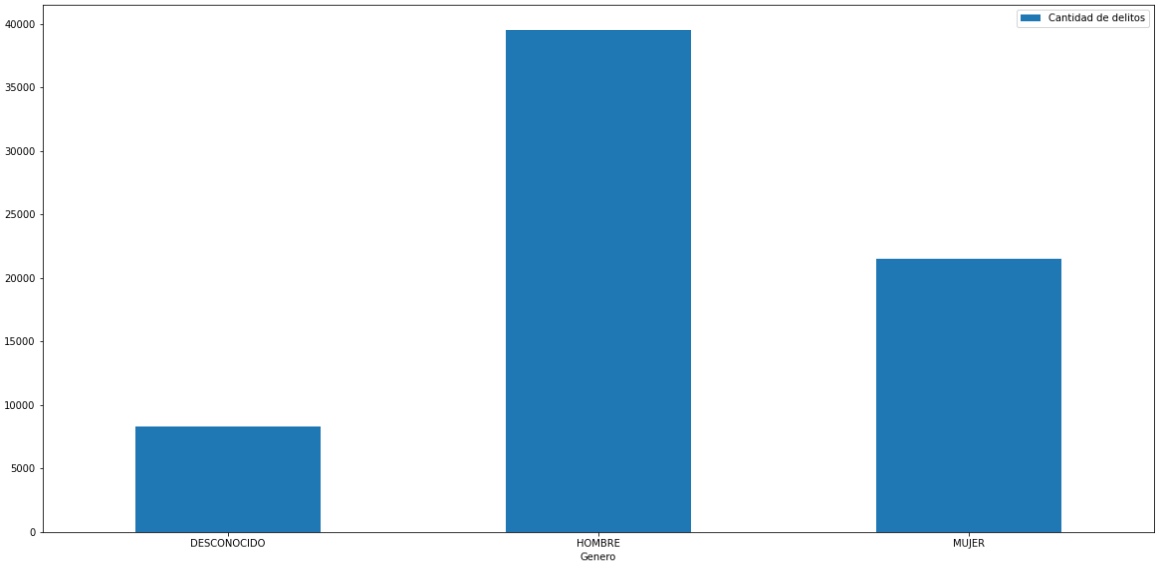
Esta visualización permite observar la cantidad de delitos por día de la semana en el periodo del 01/06/2021 al 15/11/2022 del distrito de Hospital, el cual es el más conflictivo a nivel nacional.

3



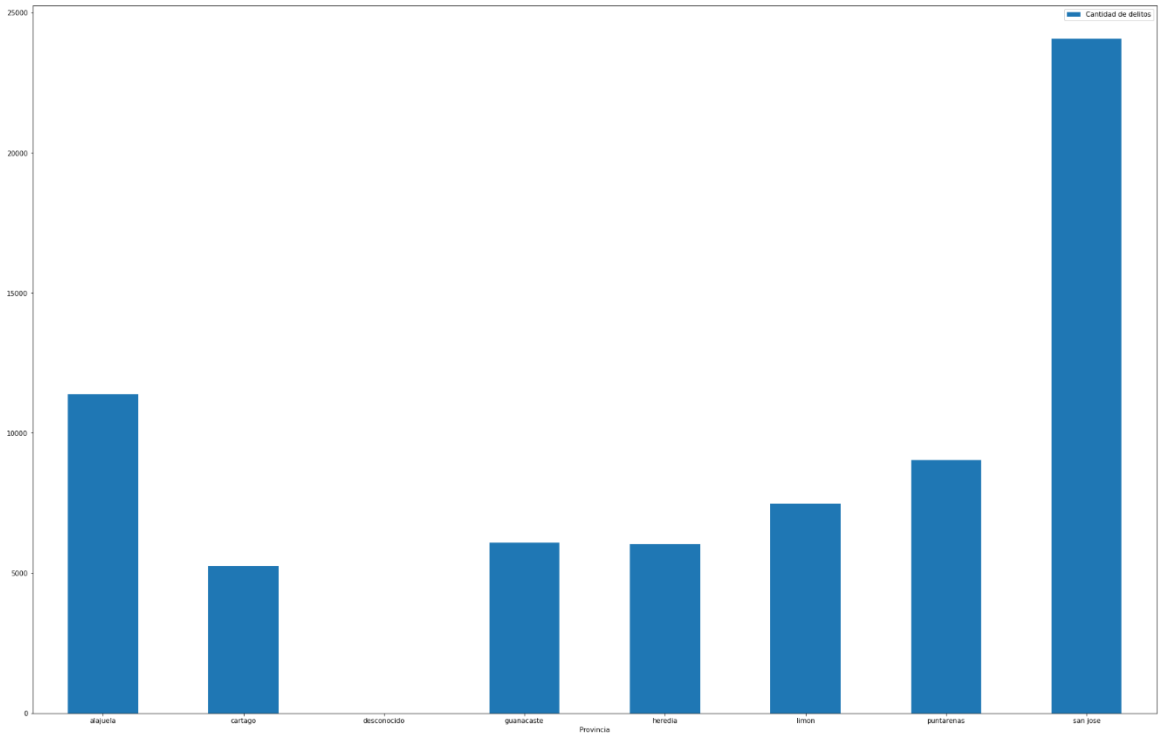
Esta visualización permite observar la cantidad de delitos según su tipo para un distrito en específico en el periodo del 01/06/2021 al 15/11/2022.

4.



Esta visualización permite observar la cantidad de delitos según el sexo del delincuente en el periodo del 01/06/2021 al 15/11/2022 y en todo el territorio nacional.

5.



Esta visualización permite observar la cantidad de delitos por provincia en el periodo del 01/06/2021 al 15/11/2022, para todo el territorio nacional.

Conclusiones

- Se logró adquirir conocimiento sobre SparkSQL, aplicando una extracción, transformación y carga de datos nacionales sobre criminalidad y aspectos económicos.
- A través de diferentes funciones se logró limpiar los datos del INEC de modo que se pudieran unir con los del OIJ para obtener información valiosa.
- Se crearon 5 visualizaciones que facilitan la interpretación de información sobre la criminalidad en el país.
- Gracias a Jupyter se logró segmentar y documentar las diferentes funciones requeridas en el proyecto, de modo que la lectura y el flujo del problema sean más compresibles.

Referencias

[1] Instituto Nacional de Estadísticas y Censos (2011). Censo 2011: Indicadores económicos, según provincia, cantón y distrito. Recuperado de https://admin.inec.cr/sites/default/files/media/reempleocenso2011-22.xls_2.xls 4

[2] Organismo de Investigación Judicial (2018). Estadísticas policiales. Recuperado de https://sitiooij.poder-judicial.go.cr/index.php/apertura/transparencia/estadisticas-policiales