Instituto Tecnológico de Costa Rica
Centro Académico de Alajuela
Bases de datos 2

Proyecto 3 Bases de Datos II

Antonio Enrique Fernández García – Carné: 2022075006
Josué David Mena González – Carné: 2022138381
Marco Vinicio Rodríguez Vargas - Carné: 2022149445
Maximilian Latysh – Carné: 2022091544

Profesor:
Alberto Shum Chan

II Semestre, 2023

Descripción general del sistema

Este sistema tiene como objetivo la aplicación de técnicas de extracción, transformación y carga de datos provenientes de dos instituciones públicas de Costa Rica, precisamente del Instituto Nacional de Estadística y Censos (INEC) y del Organismo de Investigación Judicial (OIJ), todo esto para analizar los diferentes factores que influyen en la cantidad de delitos de un distrito en específico.

Para procesar y extraer los datos se utiliza el lenguaje de programación Python junto con la herramienta Spark, esta combinación ofrece distintas operaciones que facilitan el manejo de los mismos, además brinda eficiencia a la hora de hacerlo con un volumen de datos muy grande, lo anterior es de gran utilidad pues la cantidad de información que se maneja es considerable, además que al ser datos reales provenientes de múltiples fuentes, el formato no coincide, siendo necesario una transformación antes de su análisis.

Como se mencionó, la relación entre ambos conjuntos de datos se hará mediante el distrito, en el caso del OIJ, existe una columna para esto, sin embargo, los datos provenientes del INEC utilizan una misma columna para referirse al país, provincia, cantón y distrito, de modo que no sería posible hacer dicha unión por la diferencia de formatos, para esto es necesario hacer una función que obtenga solamente el distrito de dicha columna, una vez hecho esto, simplemente quedaría quitar en ambas tablas los espacios en blanco del nombre de los distritos, ponerlos en minúscula y eliminar tildes o ñ.

Una vez los datos estén limpios y con una estructura similar, se guardan en una base de datos PostgreSQl, este paso es importante pues asegura que se guarden en un lugar confiable, además que este sistema de gestión de bases de datos tiene una alta capacidad para manejar volúmenes muy grandes. Lo anterior tiene como propósito facilitar el acceso a estos datos, de manera que, si se desea realizar un análisis o investigación estos estén disponibles y listos para trabajar.

La manera de culminar un análisis de datos es mediante su representación visual, esta permite una mayor comprensión de las relaciones o patrones que se desean demostrar por medio de dicho estudio, en este caso se utilizan las bibliotecas matplotlib y seaborn para generar gráficos en dos dimensiones que sean intuitivos y estéticos a la vista, además se usa ipywidgets y IPython para lograr que dichos gráficos puedan ser iterativos en un jupyter notebook, esto es necesario pues en una de las visualizaciones se le permite al usuario elegir el distrito.

Descripción de las funciones

a. Una función que elimine los espacios en blanco de la columna distrito para usarse en ambos conjuntos de datos.

- Descripción general: La presente función elimina los espacios en blanco a la derecha e izquierda del texto de una columna.
- Descripción de parámetros de entrada: Se reciben dos parámetros, el dataframe y la columna donde se desea realizar la operación.
- Descripción de salida: Un dataframe con una columna sin espacios en blanco.
- Descripción de bloques relevantes: Lo más relevante es el uso de la función trim(), esta función es parte del conjunto de funciones SQL que posee spark.

b. Una función que convierte a minúsculas el contenido de la columna distrito para usarse en ambos conjuntos de datos.

- Descripción general: La presente función pone en minúscula en texto de una columna.
- Descripción de parámetros de entrada: Se reciben dos parámetros, el dataframe y la columna donde se desea realizar la operación.
- Descripción de salida: Un dataframe con una columna donde sus registros están en minúscula.
- Descripción de bloques relevantes: Lo más relevante es el uso de la función lower(), esta función es parte del conjunto de funciones SQL que posee spark.

c. Una función que devuelva la lista de distritos del conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.

- Descripción general: Función que hace un left join entre dos dataframes
- para encontrar las ubicaciones que no están en el en el lado derecho
- Descripción de parámetros de entrada: Se reciben dos parámetros, df_left que es el dataframe que contiene las ubicaciones que se quieren buscar y df_right que es el otro dataframe.
- Descripción de salida: Dataframe con las ubicaciones que no están en el lado derecho
- Descripción de bloques relevantes: La parte más relevante o importante es la sentencia "*SELECT a.ubi FROM a LEFT JOIN b ON a.ubi = b.ubi WHERE b.ubi IS NULL*", gracias a esta es posible obtener los distritos que no coinciden.

d. Una función que devuelva la cantidad de registros en el conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.

- Descripción general: Función que cuenta las ubicaciones que no están en el df_right
- Descripción de parámetros de entrada: Se reciben dos parámetros, df_left que es el dataframe que contiene las ubicaciones que se quieren buscar y df_right que es el otro dataframe.
- Descripción de salida: Cantidad de ubicaciones que no están en el df_right
- Descripción de bloques relevantes: Lo más relevante es el uso de la función count(), esta función es parte del conjunto de funciones SQL que posee spark.

e. Edite, utilizando SparkSQL, los nombres de los distritos del INEC para que coincidan con algunos de los del OIJ.

- Descripción general: Función que quita las tildes y la ñ de una columna.
- Descripción de parámetros de entrada: Se reciben dos parámetros, el dataframe y la columna donde se desea realizar la operación.
- Descripción de salida: Un dataframe con una columna sin tildes o ñ.
- Descripción de bloques relevantes: Lo más relevante es el uso de la función quitar_tildes_eñes(), dentro de esta función se recibe una columna y se utiliza la función de sparksql translate(), donde básicamente se le dice que cambie las letras "áéíóúüñ" por "aeiouun".

Descripción de cada visualización

1. Compare la cantidad de delitos y la tasa de ocupación para los 10 distritos con más delitos en el país:

- En la creación de esta visualización, se comenzó reuniendo información relevante sobre delitos y tasas de ocupación (porcentaje) por distrito. Utilizando el conjunto de datos combinado del OIJ y el INEC, se agruparon los datos por ubicación (distrito) y se calculó tanto el recuento total de delitos como el porcentaje de ocupación para cada distrito. En los 10 distritos con la mayor incidencia de delitos.

- La visualización final se compone de un histograma de barras que muestra la cantidad de delitos en cada uno de los 10 distritos. Cada barra está etiquetada con la cantidad correspondiente de delitos y se representa en tono gris. Además, se incorporó una línea que refleja el promedio de ocupación para cada distrito, añadiendo una dimensión adicional al gráfico. Esta combinación de elementos proporciona una representación visual intuitiva de la relación entre la cantidad de delitos y la tasa de ocupación en los distintos distritos durante el año 2011.

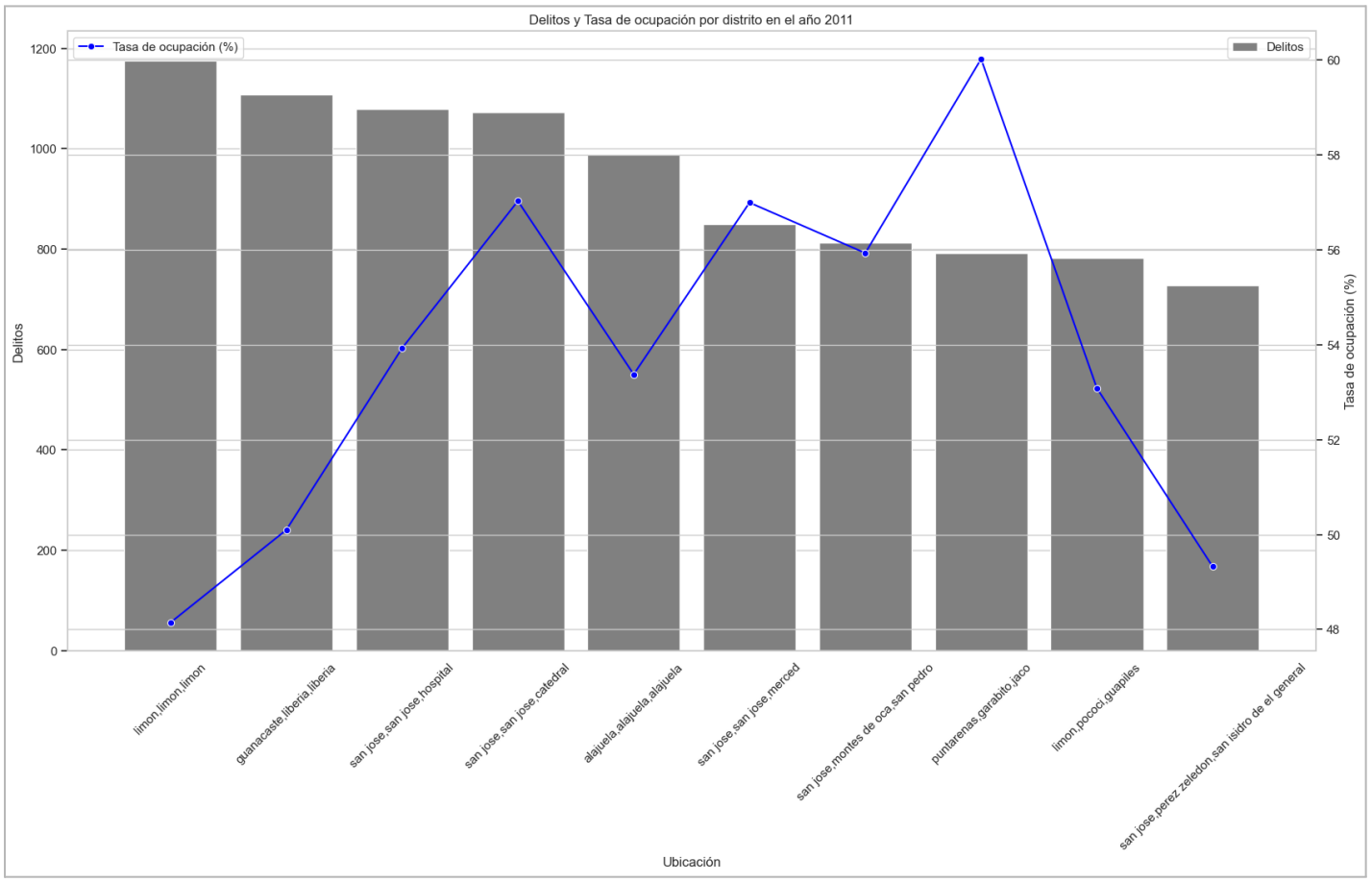

Ilustración 1. Delitos y Tasa de ocupación por distrito en el año 2011

2. Grafique la cantidad de delitos por día de la semana para el distrito con más delitos:

- Para la construcción de esta visualización, se inició filtrando el Data Frame para obtener los datos del distrito con la mayor cantidad de delitos. Se utilizó el conjunto de datos generado a partir del OIJ y el INEC, agrupando la información por distrito y calculando el recuento total de delitos para cada uno.

- Se extrajo el día de la semana de cada registro y se contó la cantidad de delitos para cada día. El gráfico resultante es un histograma de barras que muestra la cantidad de delitos para cada día de la semana en el distrito con más delitos. Cada barra está coloreada en tono salmón para una mejor distinción. La visualización ofrece una representación clara de la distribución de delitos a lo largo de la semana durante el año 2011.

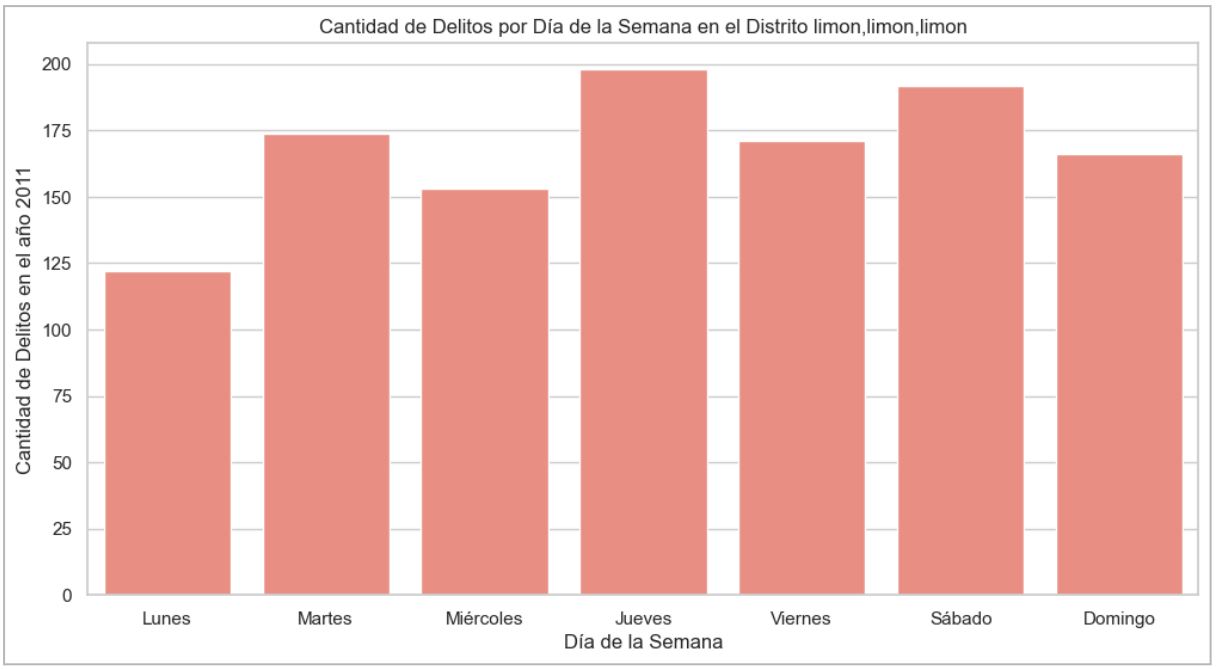

Ilustración 2. Cantidad de Delitos por Día de la Semana en el Distrito más peligroso del año 2011

3. Grafique la cantidad de delitos por tipo y por distrito. Es decir, para el distrito seleccionado se debe graficar la cantidad de delitos por tipo:

- Esta visualización se creó para analizar la distribución de delitos por tipo en un distrito específico seleccionado. El proceso de construcción comenzó obteniendo una lista de distritos distintos presentes en el conjunto de datos y ordenándose alfabéticamente. Se creó un widget de menú desplegable (Drop Down) que permite al usuario seleccionar el distrito de interés.

- La función "update_grafico" se diseñó para actualizar el gráfico de barras interactivo cuando se cambia el distrito seleccionado en el menú desplegable. Se filtra el Data Frame para obtener los datos del distrito seleccionado y se agrupan por tipo de delito. El resultado se convierte en un DataFrame de pandas para su visualización con Seaborn.

- La visualización final es un gráfico de barras en el que cada barra representa la cantidad de delitos para un tipo específico en el distrito seleccionado. Las barras están coloreadas en tono azul marino para resaltar las diferencias. El eje y muestra los diferentes tipos de delitos, mientras que el eje x indica la cantidad correspondiente de delitos.

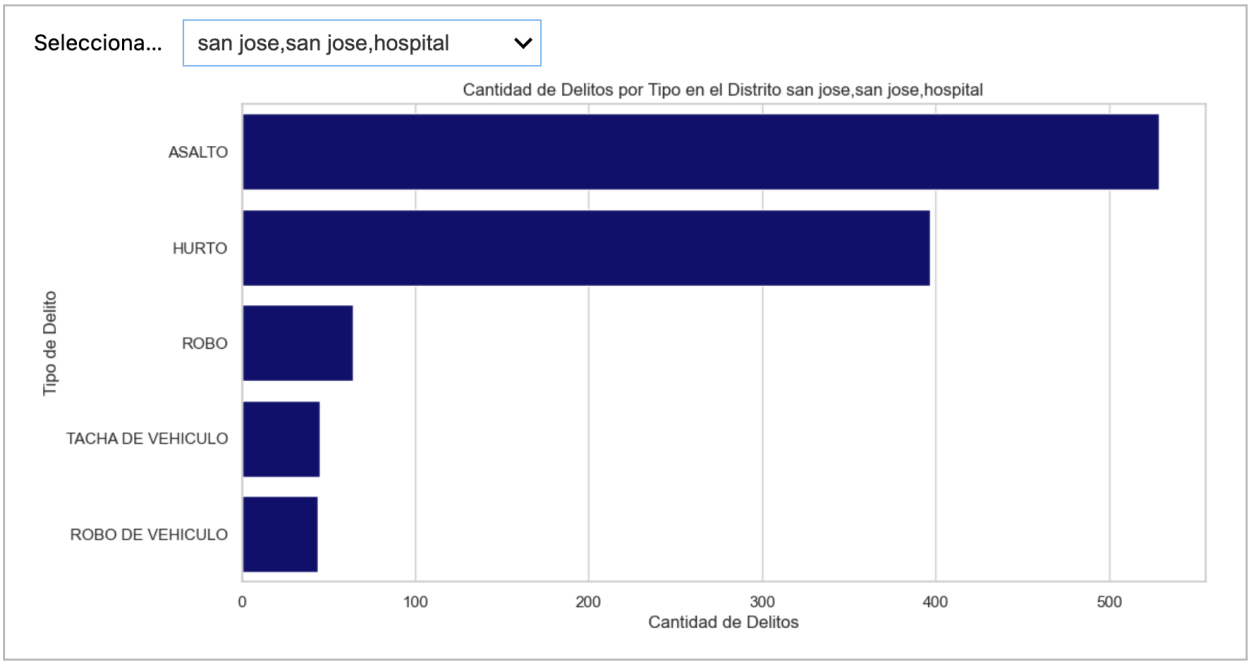

Ilustración 3. Cantidad de Delitos por Tipo en un distrito seleccionado

4. Grafique la cantidad de delitos por sexo para todo el conjunto de datos:

- Esta visualización tiene como objetivo representar la cantidad de delitos por género en el conjunto de datos. Para su construcción, se filtró el Data Frame para obtener los datos utilizando la función “group by” para cada género.

- El gráfico final es un histograma de barras que muestra la cantidad de delitos para cada género. Cada barra está coloreada en tono celeste para una mejor visualización. El eje x representa los diferentes géneros, mientras que el eje y indica la cantidad de delitos.

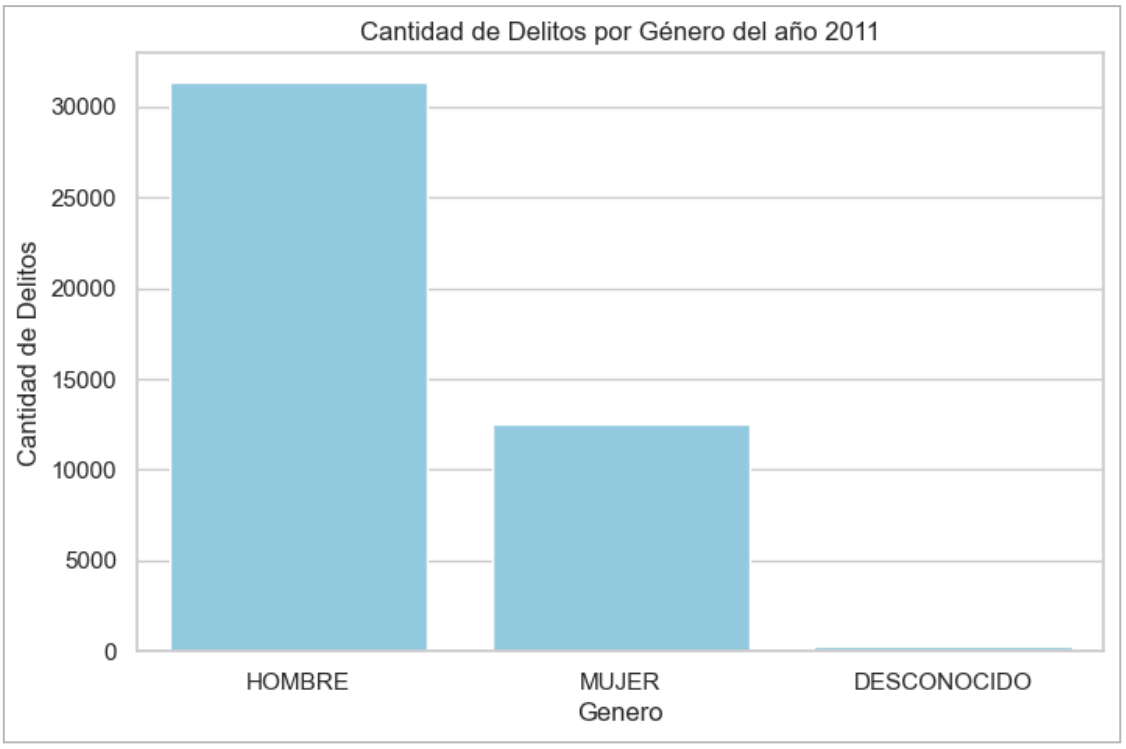

Ilustración 4. Cantidad de Delitos por Género del año 2011

5. Grafique la cantidad de delitos por las 10 nacionalidades más peligrosas (Proponga una visualización de su interés):

- Esta visualización presenta la cantidad de delitos por las 10 nacionalidades más peligrosas en el año 2011. El proceso comenzó filtrando el Data Frame para obtener los datos de delitos por nacionalidad. Se utilizó la función “group by” para cada nacionalidad y se ordenaron en orden descendente según su cantidad de delitos.

- Para enfocarse en las nacionalidades más significativas, se eliminaron las entradas correspondientes a "COSTA RICA" y "DESCONOCIDO", y se seleccionaron las 10 principales nacionalidades extranjeras. El gráfico resultante es un histograma de barras donde cada barra representa la cantidad de delitos para una nacionalidad específica. Las barras están coloreadas en tono naranja para una mejor visualización.

- Además, se agregaron las cifras exactas al lado de cada barra para ofrecer una referencia numérica precisa. El eje y representa las diferentes nacionalidades, mientras que el eje x la cantidad correspondiente de delitos.

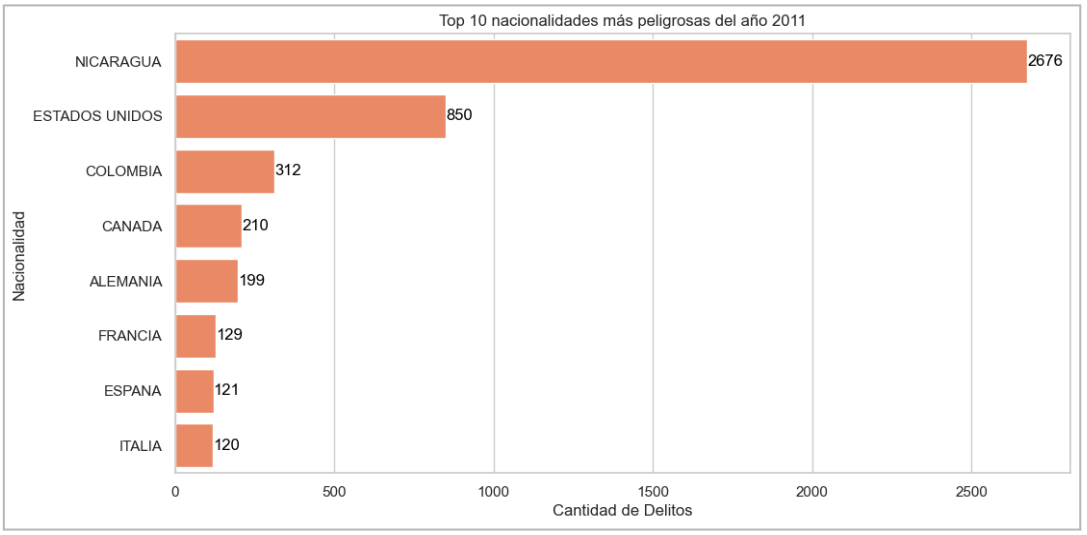

Ilustración 5. Top 10 nacionalidades más peligrosas del año 2011

Conclusiones

1. En general los hombres en el año 2011 tuvieron una tendencia muchísimo mayor a cometer delitos que las mujeres, lo cual es un dato preocupante acerca de nuestra sociedad.
2. Excluyendo la nacionalidad costarricense, la nacionalidad con mayor cantidad de delitos es Nicaragua y con mucha diferencia del segundo. Este es un dato alarmante para el país.
3. En la provincia de San José se dieron gran cantidad de delitos en el año 2011, teniendo 4 distritos en el top 10 de distritos con más delitos del país en ese año. Este debe ser un dato alarmante para la provincia ya que la criminalidad en ese año fue muy persistente.
4. En los dos distritos con más criminalidad se puede ver una tendencia de tasa de ocupación MUY baja, lo que se puede interpretar como un problema a nivel social que implica que el poco trabajo conduce a una alta criminalidad por distrito en esos casos.

Referencias

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

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

Suhong, K., Param, J., Parminder, K. y Pooya, T. (2018). *Crime Analysis Through Machine Learning. IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).* 10.1109/IEMCON.2018.8614828. Recuperado de https://www.researchgate.net/publication/330475412

freeCodeCamp.org. (2021). *PySpark Tutorial* [Video]. YouTube. https://www.youtube.com/watch?v=_C8kWso4ne4

GK Codelabs. (2019). *Data Visualization with Spark Dataframes & Python Matplotlib | From Scratch* [Video]. YouTube. https://www.youtube.com/watch?v=lZvs-YNk4V0