Bachillerato en Ingeniería en
Ciencia de Datos

2023-lll BCD9221 Optativa

Proyecto Final – Grupo 2

Profesor: Nathalia Herrera

Alumnos:
Solanlly Barker
Luis Diego Chacon

Diciembre 2023

Contenidos

Definición del proyecto:

Objetivo general:

Proveer de una opción de visualización de datos de la CGR para proporcionar información rápida y accesible sobre las compras institucionales, mejorando la experiencia del usuario.

Objetivos específicos:

1. Funcionalidades de Análisis Avanzado:
    a. Implementar herramientas analíticas avanzadas en la plataforma para permitir a los usuarios realizar análisis más detallados de los datos de compra. Esto puede incluir la integración de gráficos interactivos, herramientas de filtrado avanzado o generación automática de informes personalizados.
2. Facilitar la Interpretación de Datos:
    a. Desarrollar herramientas de resumen que simplifiquen la información compleja para los usuarios, como resúmenes mensuales o anuales de compras, para una comprensión rápida.
3. Optimizar la Interactividad:
    a. Incorporar funciones interactivas simples, como filtros desplegables o herramientas de zoom, para permitir a los usuarios explorar datos específicos con mayor detalle.

Análisis de Requisitos y Diseño:

Investigación de requerimientos:

Se reúne información detallada sobre las necesidades de los usuarios y los problemas y requisitos que como institución la Contraloría General de la República (CGR) ha estado teniendo en el desarrollo de esta herramienta: Entre la información requerida la institución genera sus códigos utilizando el lenguaje de programación en R así como una exclusiva compatibilidad con herramientas de trabajo del paquete proveído por Google. Por lo tanto, se utilizará la herramienta Looker studio desarrollada por google para crear las visualizaciones.

Limpieza de Datos:

Una vez que se generaron los CVS del código proveído por la CGR se inicia un análisis exploratorio de los 13 archivos generados. A continuación de mencionan los problemas encontrados en los datos:

1. Datos Faltantes:
    a. Dentro de algunos archivos se encontraron en promedio un 0.07% hasta un 0.07% de datos faltantes o en blanco por columna.
2. Formatos inconsistentes:
    a. Problemas especialmente en tildes y caracteres especiales del español que no lograron transcribirse correctamente en los archivos.

Por lo cual, para poder utilizar estos datos se debió ejecutar una limpieza de datos donde por medio de las siguientes técnicas:

1. Sustitución de Datos:
    a. Sustitución por valores faltantes: Cuando existen valores faltantes (nulos) en un conjunto de datos, la sustitución implica reemplazar esos valores con datos válidos. Esto puede hacerse asignando un valor predeterminado (como la media, mediana o moda de la columna) o utilizando técnicas más avanzadas como la imputación estadística.

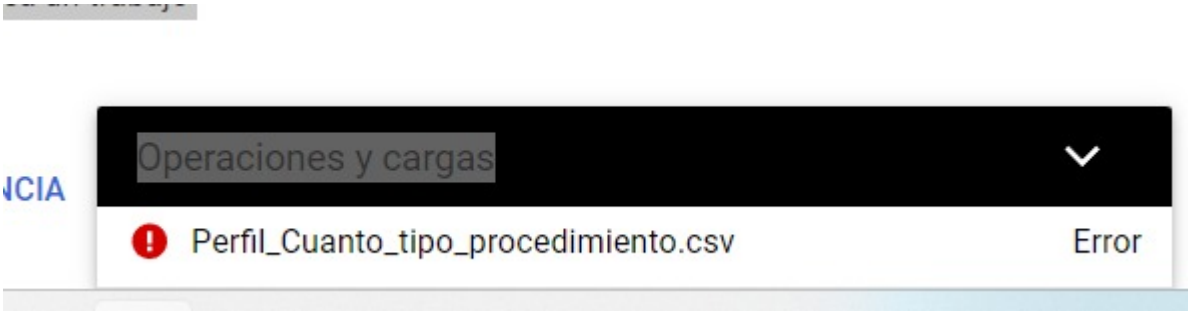
2. Eliminación de datos:

    a. Cuando no es significativo, los registros que contengan campos nulos o en blanco, se pueden eliminar toda vez que la cantidad de registros no representen un sesgo para los resultados finales.

Limitaciones:

1. Looker Studio:
    a. Límites de la subida de archivos:
        i. 1000 conjuntos de datos por usuario
        ii. 2 GB de almacenamiento por usuario
        iii. 100 subidas por conjunto de datos al día
        iv. 100 MB de tamaño de archivo máximo de datos

La información anterior fue recopilada directamente de Google help. Debido al tamaño de los data set que utiliza el presente proyecto así como sus tamaño variables, la subida de datos fue uno de los grandes obstáculos que se tuvo durante el proceso ya que no se lograron subir los archivos CVS a la plataforma para su respectivo uso.



2. Nivel de detalle disponible:
    a. Cantidad de empresas:
        a. Solamente del sector público descentralizado institucional, se pueden visualizar 34 instituciones autónomas, sus 13 órganos adscritos, 8 semiautónomas, 25 públicas, 46 entes públicos no estatales entre muchas otras instituciones.
        b. En los datasets las instituciones están referenciadas por medio de ID’s no por nombre de la institución, asimismo existen hasta 1 millón de IDs únicos en los datasets lo cual imposibilidad brindar un nivel de detalle deseado para el público meta que corresponde a los ciudadanos costarricenses.

3. Calidad de los datos

    En el momento que se identificó cuáles archivos se utilizarían para la visualización de datos, se identificó que los datos finales tenían problemas de calidad, por ejemplo, registros con datos nulos, caracteres especiales no reconocidos por el idioma español, formatos de números, entre otros. Se trató de limpiarlos utilizando python, pero no fue posible, debido a que las primeras líneas de los archivos, tenían registros en blanco. Al ser formato csv, al momento de cargar los archivos, python reconoce una cierta cantidad de columnas, pero por los registros en blanco, daba error de carga ya que no coincidía la cantidad de registros en una línea respecto a la cantidad de columnas existentes.

Por lo anterior, se tuvo que hacer limpieza de datos, directamente en el archivo csv lo que representó un tiempo que no se tenía estipulado.

4. Otras herramientas de procesamiento de datos

Conforme a la situación de problemas de calidad de datos y la limitante para utilizar locker studios, se exploraron otras alternativas como librerías de R y Python para generar dashboards.

En el caso de Python, existe Dash para generar dashboards interactivos, sin embargo, para poder desarrollar algún proyecto con esta biblioteca, es necesario también generar caracterizaciones en CSS y HTML. Debido a lo complejo de esta biblioteca y que requería capacitarse en ella, no fue viable utilizarla por razones de curva de aprendizaje y tiempo para poder dar un entregable en el tiempo que se tenía estipulado.

Con R, la dificultad presentada fue la versión que se utiliza, se hizo una pequeña prueba de desarrollar una visualización de un gráfico interactivo, pero por motivos de versión, cambio en las bibliotecas y funciones, no se adaptó a cualquier equipo, por lo que el riesgo de desarrollar un dashboard con esta herramienta y que no se adaptara a una versión distinta era muy alta, por ello se decidió no utilizar R.

Conclusiones:

1. Se implementaron técnicas matemáticas y estadísticas para subsanar los inconvenientes de calidad en las fuentes de datos para poder hacer tratamiento de estos.

2. Debido a las limitaciones encontradas, se utilizó Power BI como herramienta para la visualización de datos, lo cual es más amigable para el usuario y además, se pudo conformar relaciones entre las distintas tablas, lo que permitió generar información con datos cruzados.

Bibliografía:

- *Tutorial: Crear un informe - ayuda de Looker Studio*. (s. f.). Recuperado de: https://support.google.com/looker-studio/answer/6292570?hl=ES#zippy=%2Csecciones-de-este-art%C3%ADculo

- López, B. R. (s. f.). *Librerías para animar tus datos en R*. Cursos GIS | TYC GIS Formación. Recuperado de: https://www.cursosgis.com/librerias-para-animar-tus-datos-en-r/

- *RPUbS - Taller R Markdown*. (s. f.). Recuperado de: https://rpubs.com/tracelac/Rmarkdown

- *DASH Documentation & User Guide | Plotly*. (s. f.). Recuperado de: https://dash.plotly.com/