**Analítica de Grandes Datos**

**Departamento de Ciencias de la Computación y la Decisión**

**Facultad de Minas**

**Universidad Nacional de Colombia**

Trabajo Nociones de Arquitectura de la Información

Versión: aaaa.mm.dd HH:mm

*Observación: Cada vez que agregue nuevos elementos al documento, o que modifique algún componente del informe, revise la coherencia y consistencia con los otros elementos que hacen parte del mismo.*

## **Responsables**

|  |
| --- |
| Nombre Completo – Documento de Identificación |
| 1.Santiago Palacio Velásquez - 1.017.194.675 |
| 2. Gustavo Adolfo González Hernández - 1.152.214.298 |
| 3. Simon Andres Arias Valencia - 1.040.748.605 |
| 4. Andres Felipe Obando Trejo - 1.152.186.183 |
| 5. Johan Manuel Ortega Caro- 1.047.381.524 |
| **REPO EN GITHUB:** https://github.com/spalaciov0819/Analitica-de-Grandes-Datos |

*Realiza este trabajo considerando los datos que generan los sistemas transaccionales e información no estructurada de tu dominio (si trabajas por ejemplo para TCC tu dominio es la mensajería; también puedes explorar en la página* [*https://www.kaggle.com/datasets*](https://www.kaggle.com/datasets) *o* [*https://arxiv.org/*](https://arxiv.org/)*). Considera tener acceso a esta información, de al menos 10 MB (puede ser uno o varios archivos de texto), y* ***tener al menos cuatro clases conceptuales. Este documento también debe almacenarse en el REPO. Plazo Máximo de Entrega 23 de Mayo, NO SE recibirá por correo electrónico, envío por*** [***https://forms.gle/h7ty3yZykaUq5m7y6***](https://forms.gle/h7ty3yZykaUq5m7y6)

# **1** **Comprensión del negocio**

## **1.1** **Descripción del contexto del negocio.**

El área de dominio es una cadena de almacenes de grandes superficies de retail, que se encarga de comercializar productos como alimentos, ropa, productos para el hogar, tecnología, entre otros. El enfoque principal de los datos, es ser insumo de la vicepresidencia de marketing para la generación de insights que permita tomar decisiones basadas en datos y generar estrategias de ventas que ayuden a mejorar los resultados.

La cadena cuenta con 2 marcas (Almacenes A y B), con 63 tiendas distribuidas en las regiones de Costa y Santanderes, y cuenta con 3 canales de ventas, (físicas, Domicilios y ventas online). Para efectos de nuestra investigación, se tomará como eje central, la información que se almacena en cada venta ya sea por el canal físico, que se realiza directamente en los puntos de venta, la venta de domicilios que hacen los clientes a través de teléfono o WhatsApp, o la venta online a través de la página web.

La estructura de la base de datos se toma con base en un referenciación realizada por integrantes del equipo de trabajo que laboran en una cadena de almacenes de grandes superficies. Sin embargo, por ser esta una información confidencial, para efectos académicos se trabajará con datos dummy.

## **1.2** **Identificación del problema:**

Las directivas del almacén requieren conocer las categorías con menores ventas por canal en el 2020, con el fin de definir estrategias para optimizar el uso de recursos y potenciar las ventas de estas categorías identificadas en cada canal para el año 2021.

## **1.3** **Determinación de objetivos:**

* Proponer un modelo de datos o analítica aplicando los conceptos de arquitectura de la información y de ciencia de datos para entregar insights de valor a las directivas de la compañía que ayuden a la toma de decisiones.
* Identificar las categorías menos compradas por canal en los supermercados en el 2020 haciendo uso de herramientas de analítica para realizar campañas que ayuden a potencializar sus resultados.

## **1.4** **Evaluación de la situación actual:**

La compañía actualmente no cuenta con un modelo de datos estructurado que le permita analizar su ventas en los diversos niveles de la información con la que cuenta, por lo cual hace uso de herramientas que no permiten manejar ágilmente grandes volúmenes de datos. Es así como procedemos a conocer el contexto del negocio y los datos de la compañía, para poder generar el modelo de dominio y de entidad-relación que nos permitió armar la infraestructura del modelo de datos para esta compañía. Teniendo lista la infraestructura, y haciendo uso de bases de datos relacional en SQL y bases NoSQL (orientada a documentos) de MongoDB, procedemos a realizar las respectivas consultas que permitan dar respuesta a la dirección sobre el problema planteado. Es así como llevamos a que la compañía cuente con su información estructurada, a la cual puede acceder ágilmente y que le permite tomar decisiones basadas en datos.

# **2** **Comprensión de los datos**

## **2.1** **Recolección de datos**

En principio se tienen 6 tablas: venta, producto, almacén, canal, tipo\_documento y cliente; donde se guarda la información relacionada con las ventas del retail, el producto de compra, los almacenes de compra o despacho, los canales de compra, el documento y la información del cliente.

El Retail captura la información de los clientes a través del programa de fidelización, en donde los clientes pueden registrarse y actualizar su información personal permitiendo generar las tablas ‘cliente’ y ‘tipo\_documento’.

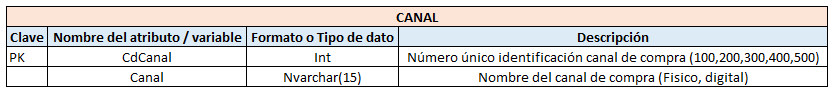
La tabla de venta es obtenida a partir de las transacciones realizadas por cada uno de los clientes en nuestros almacenes y de acuerdo a los productos comprados por este. Esta tabla de venta se asocia con la tabla de almacén y canal para hacer relación con el lugar y canal usado en la compra. Así como con la tabla producto para identificar qué fue lo que compró el cliente.

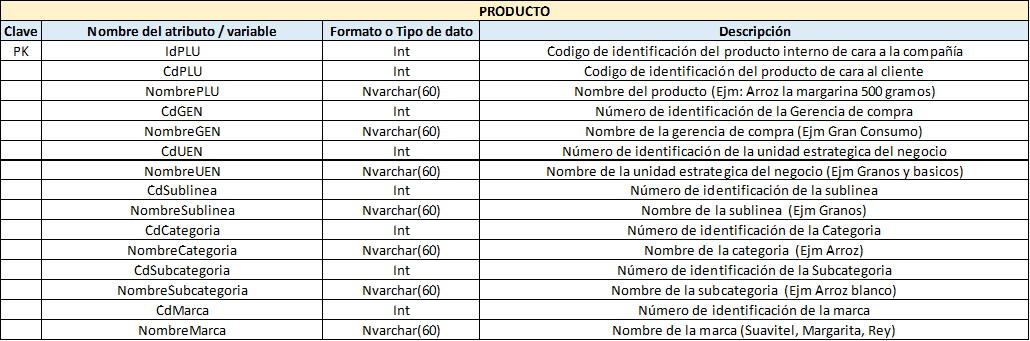
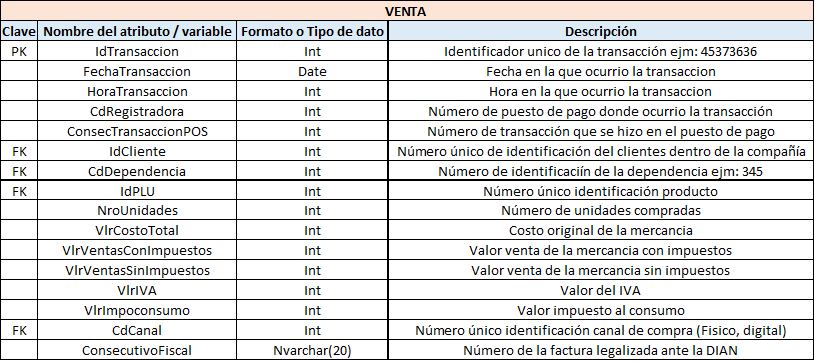
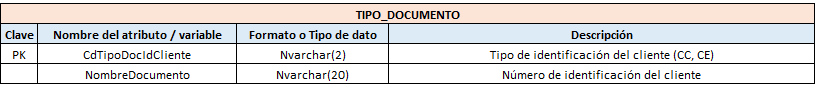
Dado que la información es confidencial y se debe respetar la protección de datos personales (Habeas Data) se usaron datos dummy para estas tablas.

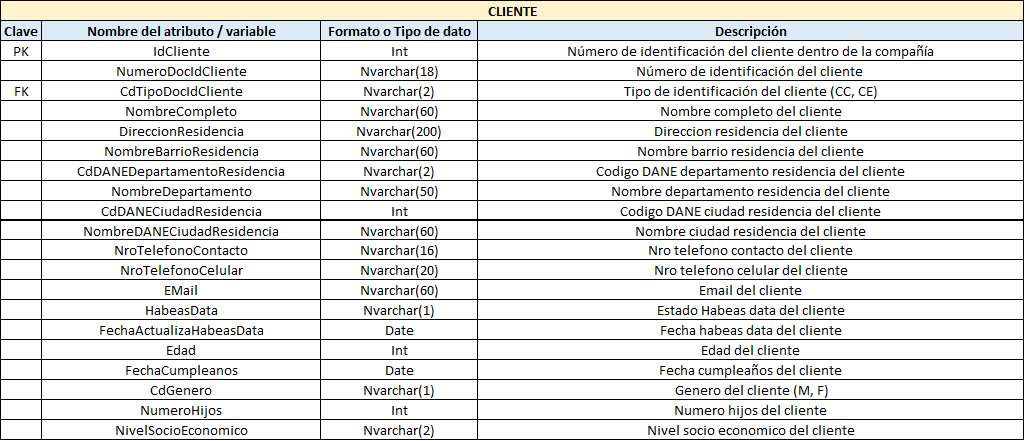
## **2.2** **Descripción de datos (diccionario):**

Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

## 

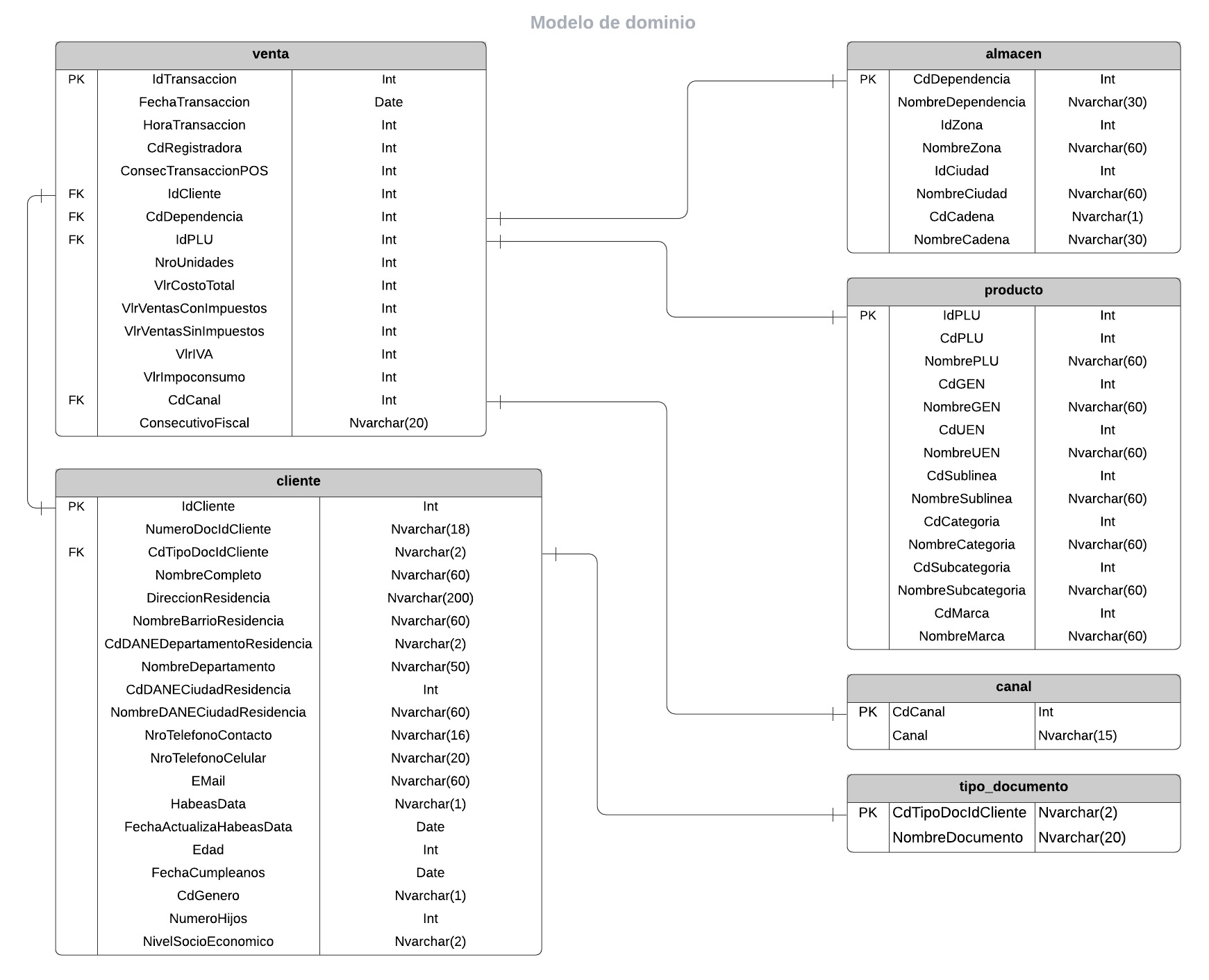


****



**2.3** **Modelo del dominio**

***Observación****: Incluya el gráfico del modelo del dominio que representa la estructura de datos de su problema.*

**

<https://lucid.app/lucidchart/invitations/accept/inv_cee01e1d-1f3f-4a2a-b5f6-ec4d7d8191b3?viewport_loc=-189%2C-3%2C2378%2C1022%2C0_0>

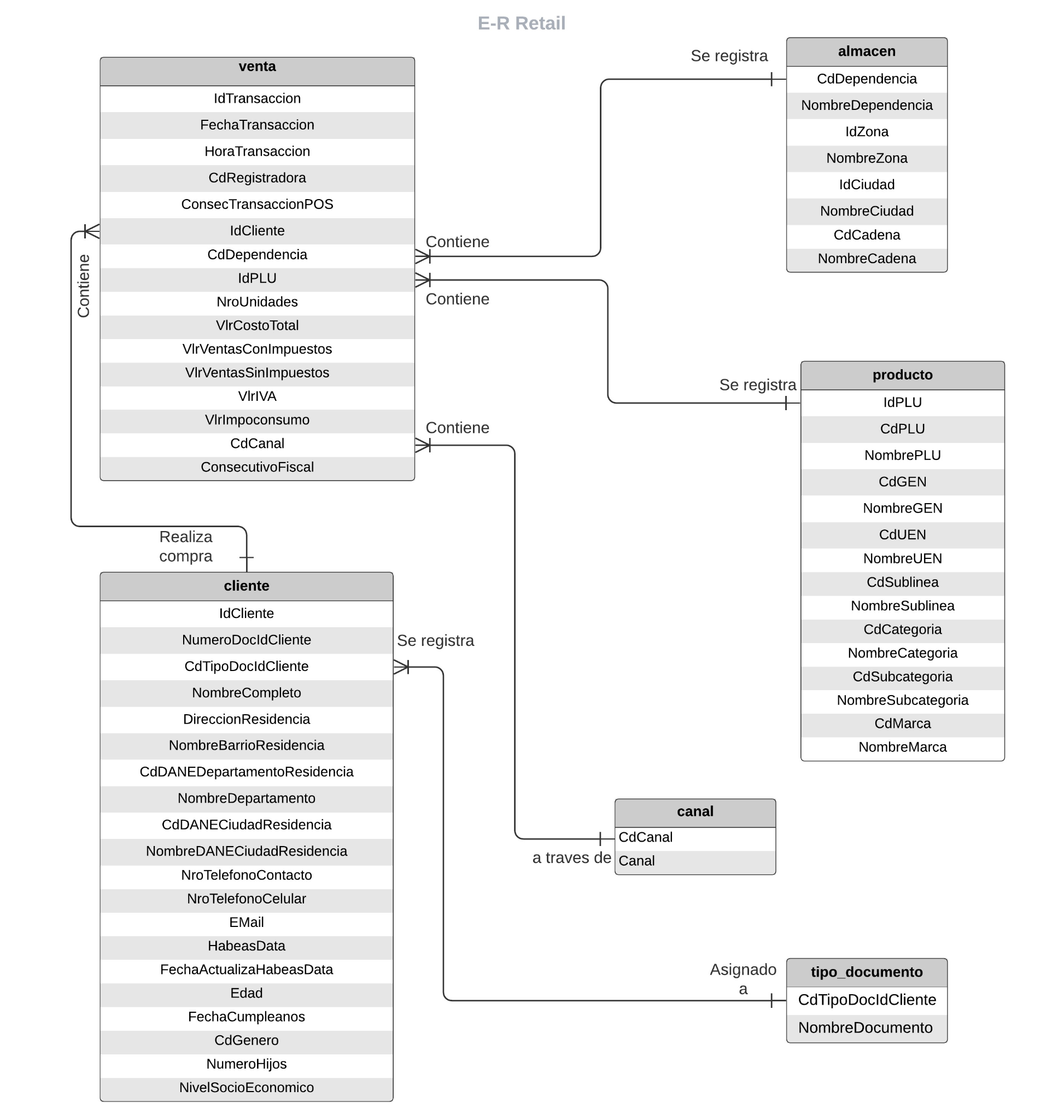
# 

# 

# 

# **3** **Modelo Entidad-Relación**

## **3.1** **Toma de pantalla del modelo E-R**

******

[https://lucid.app/lucidchart/cbeb6518-71ea-4dab-a197-c434136fd6be/edit?page=fwQcDkAJnwES#](https://lucid.app/lucidchart/cbeb6518-71ea-4dab-a197-c434136fd6be/edit?page=fwQcDkAJnwES)

## **3.2** **Sentencia o consulta de creación del tabla(s)**

Las sentencias para crear las 6 tablas se incluyeron en el repositorio de GitHub *T1.3.2.Creacion\_Tablas.sql*

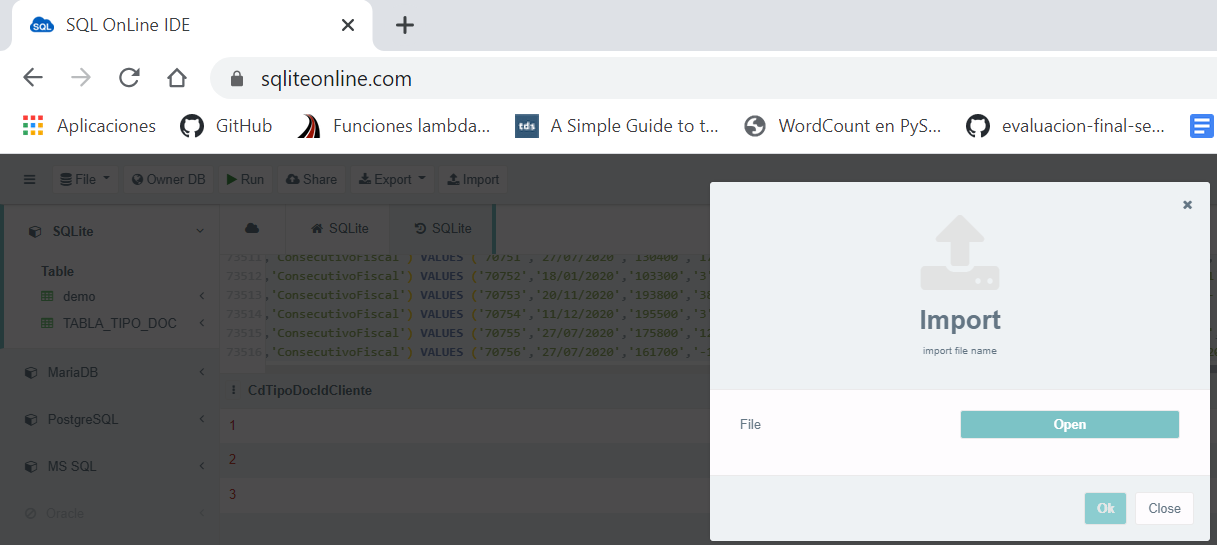
[*https://github.com/spalaciov0819/Analitica-de-Grandes-Datos/blob/main/T1.3.2.Creacion\_Tablas.sql.sql*](https://github.com/spalaciov0819/Analitica-de-Grandes-Datos/blob/main/T1.3.2.Creacion_Tablas.sql.sql)

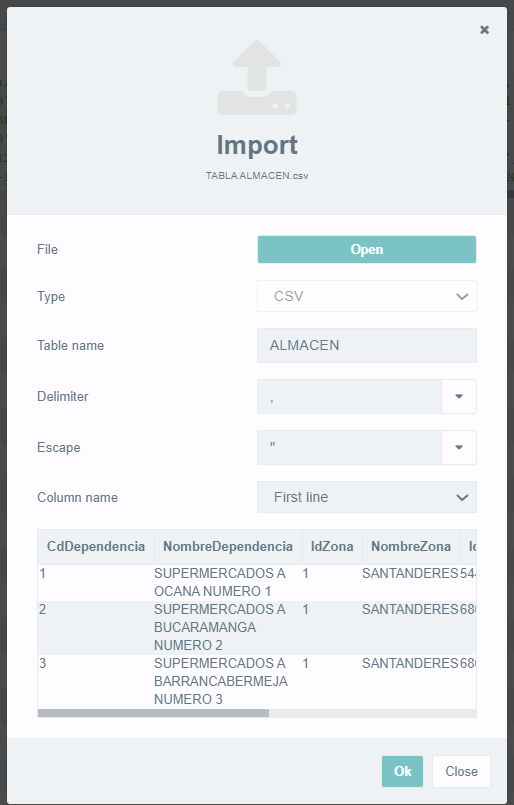
## **3.3** **Sentencias para Insertar datos**

Las bases de datos las teníamos en formato .CSV, por esto usamos sqliteonline.com para insertar los CSV en las bases de datos, que a su vez nos entregaba las sentencias “INSERT” de SQL

Las sentencias SQL para insertar los datos a las 6 tablas se incluyeron en el repositorio en formato .zip (T1.3.3.Insertar\_Datos.zip) puesto que pesaba más de las 25Mb permitidas por GitHub.

<https://github.com/spalaciov0819/Analitica-de-Grandes-Datos/blob/main/T1.3.3.Insertar_Datos.zip>

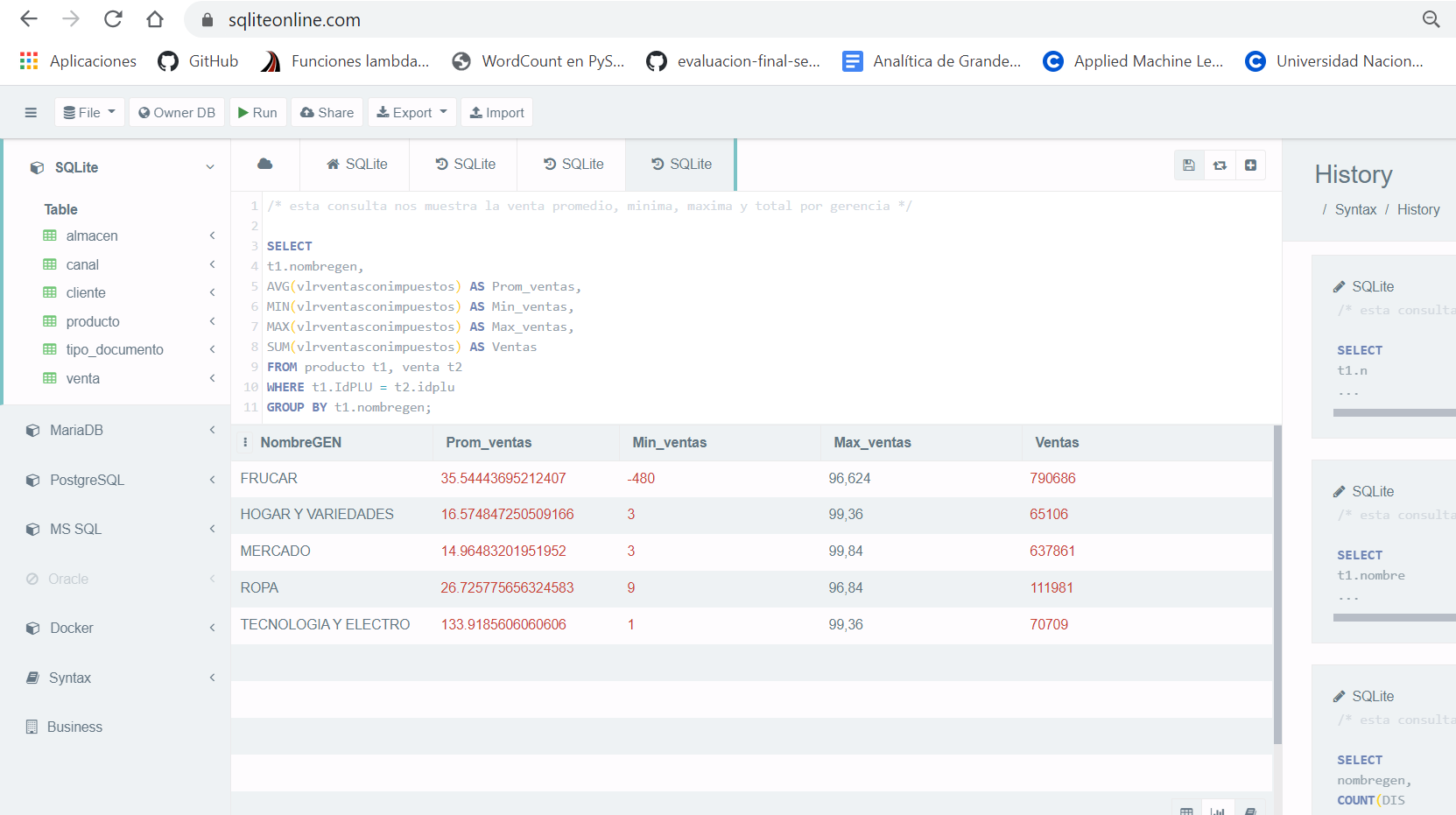


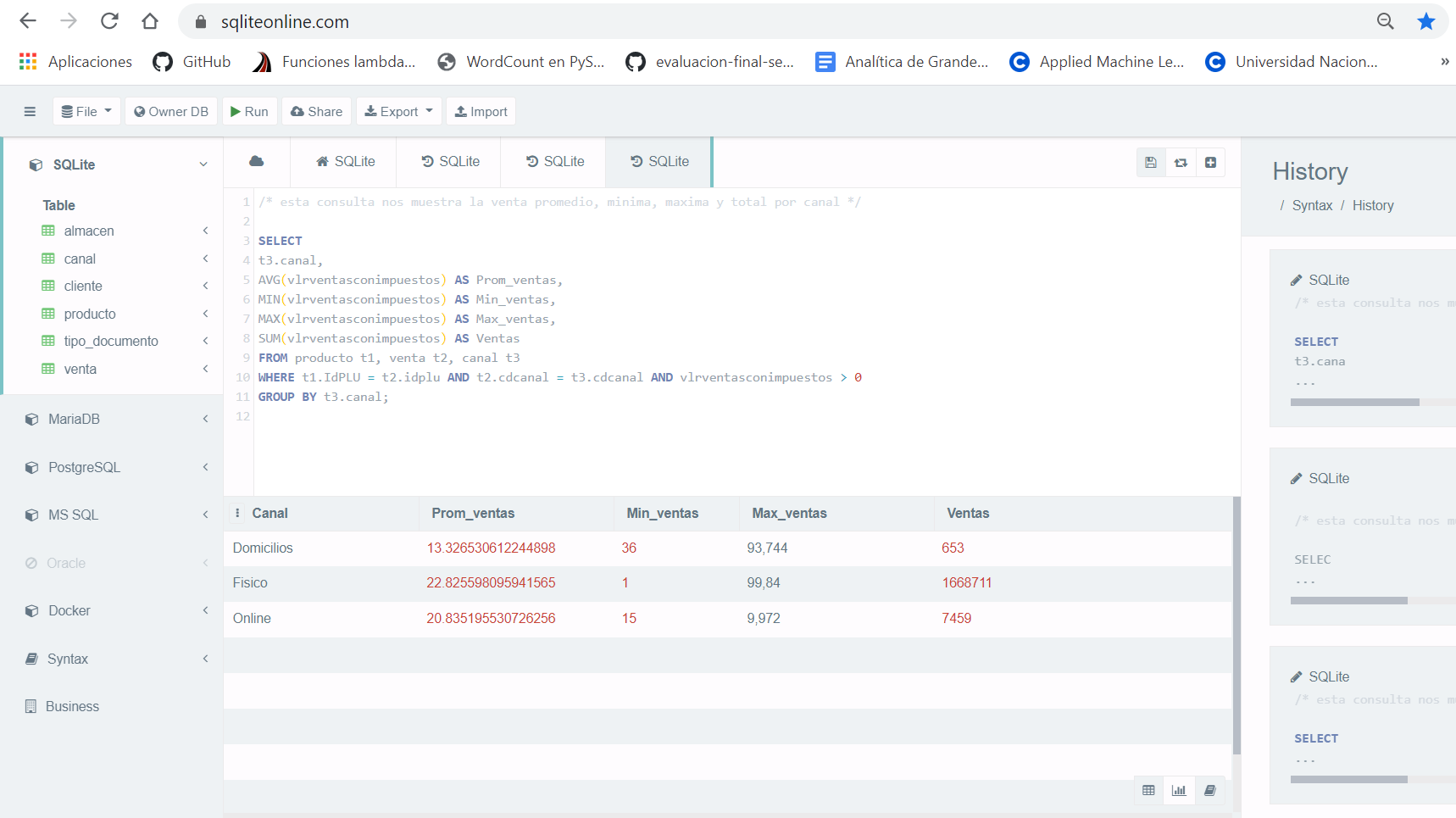


## **3.4** **Sentencia de consulta**

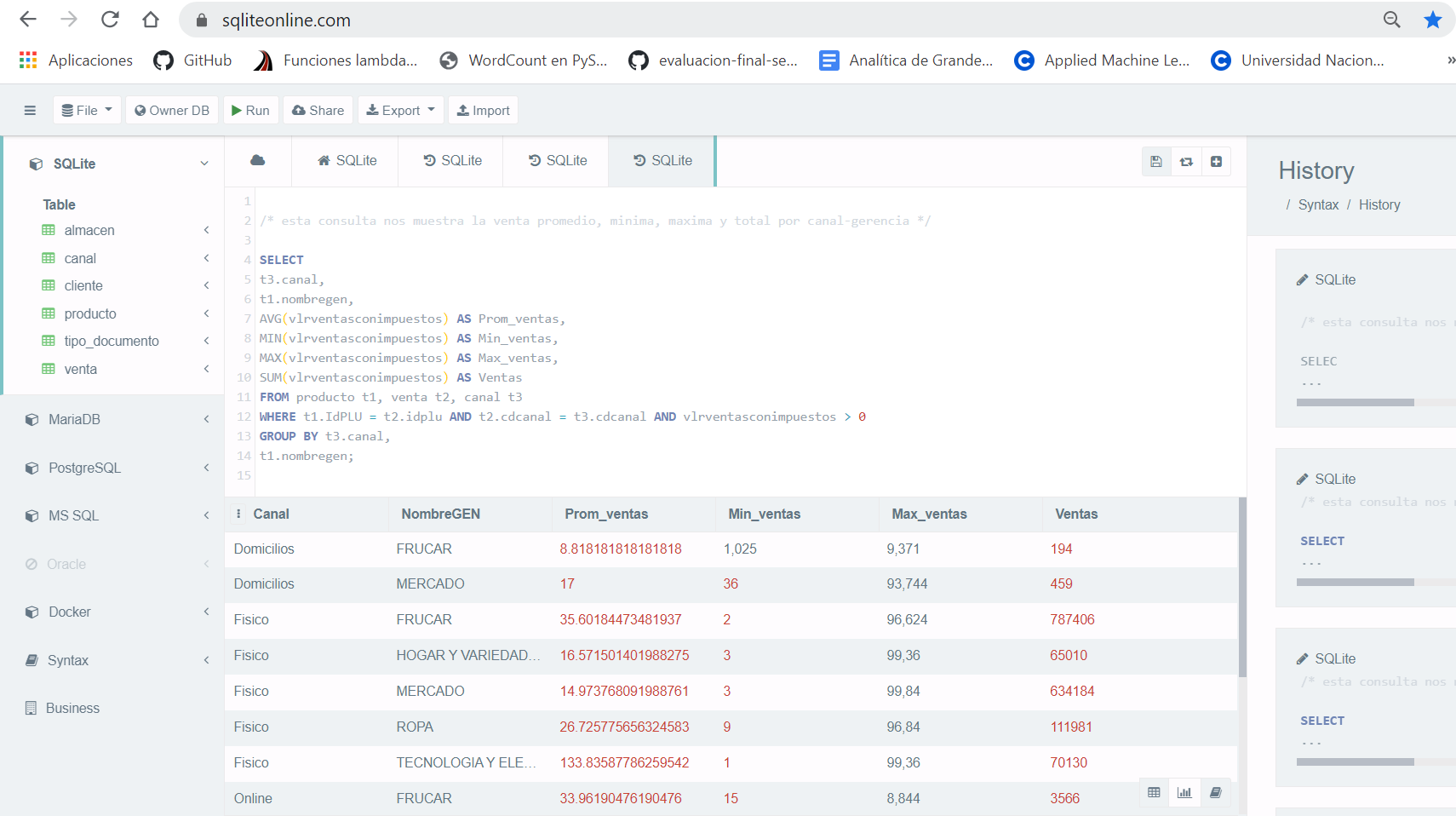
Para dar respuesta al problema identificado por la gerencia de la compañía, hicimos un diagnóstico de las ventas de las gerencias y de los canales en el 2020, para encontrar cuales son las categorías que debemos impulsar en el 2021 para obtener mejores resultados.

Inicialmente se hizo un conteo de la cantidad de registros de la información contenida en las bases de datos, así como comprender cada campo (de acuerdo al diccionario) que nos ayudará a realizar las consultas SQL. Posteriormente revisamos cómo es la venta de cada gerencia, encontrando que Hogar y tecnología son las de menores ventas, a su vez logramos identificar que el canal de domicilios es el de menor venta de los 3 canales y que solo vende 2 gerencias (Frucar y mercado).

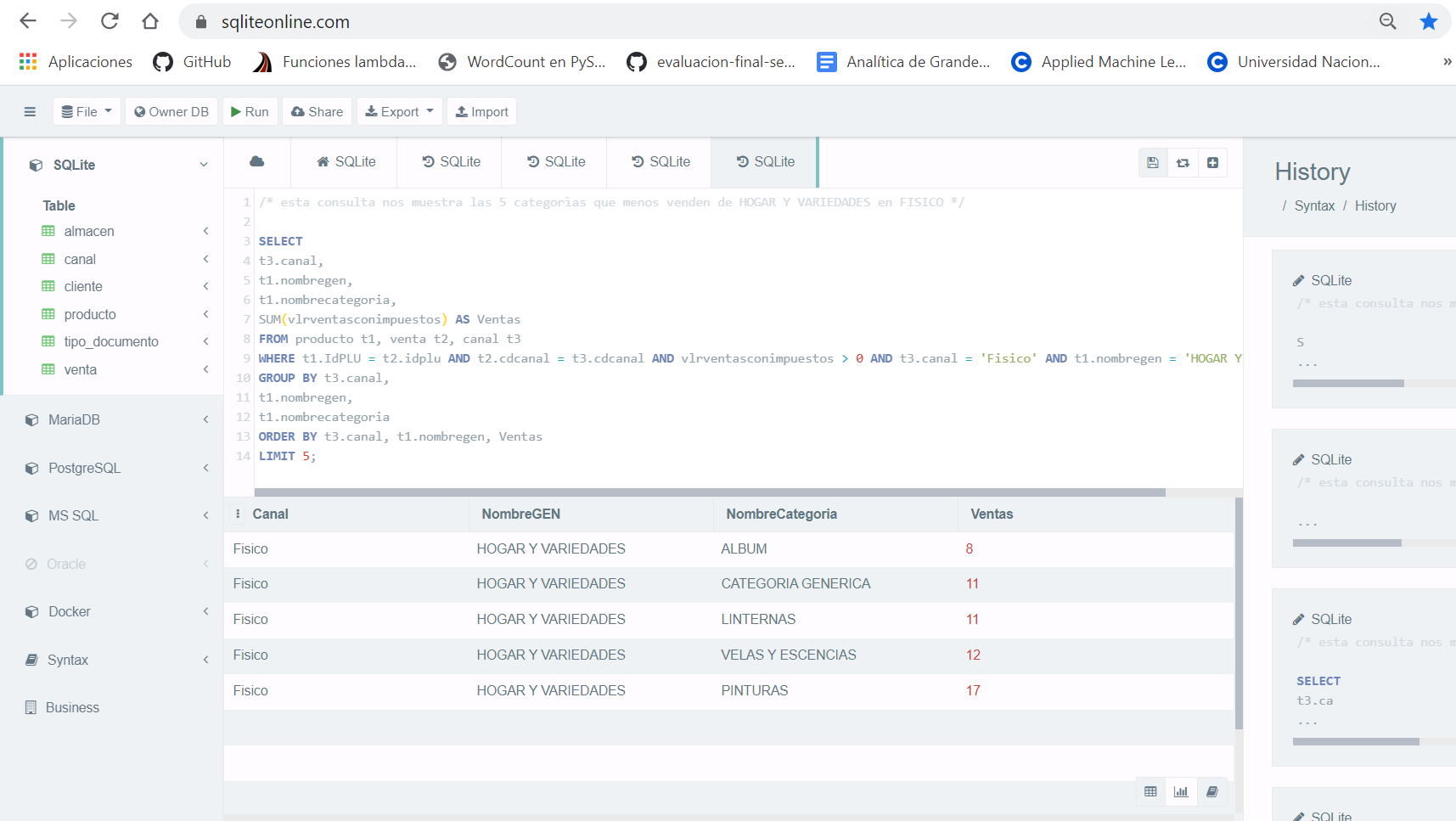


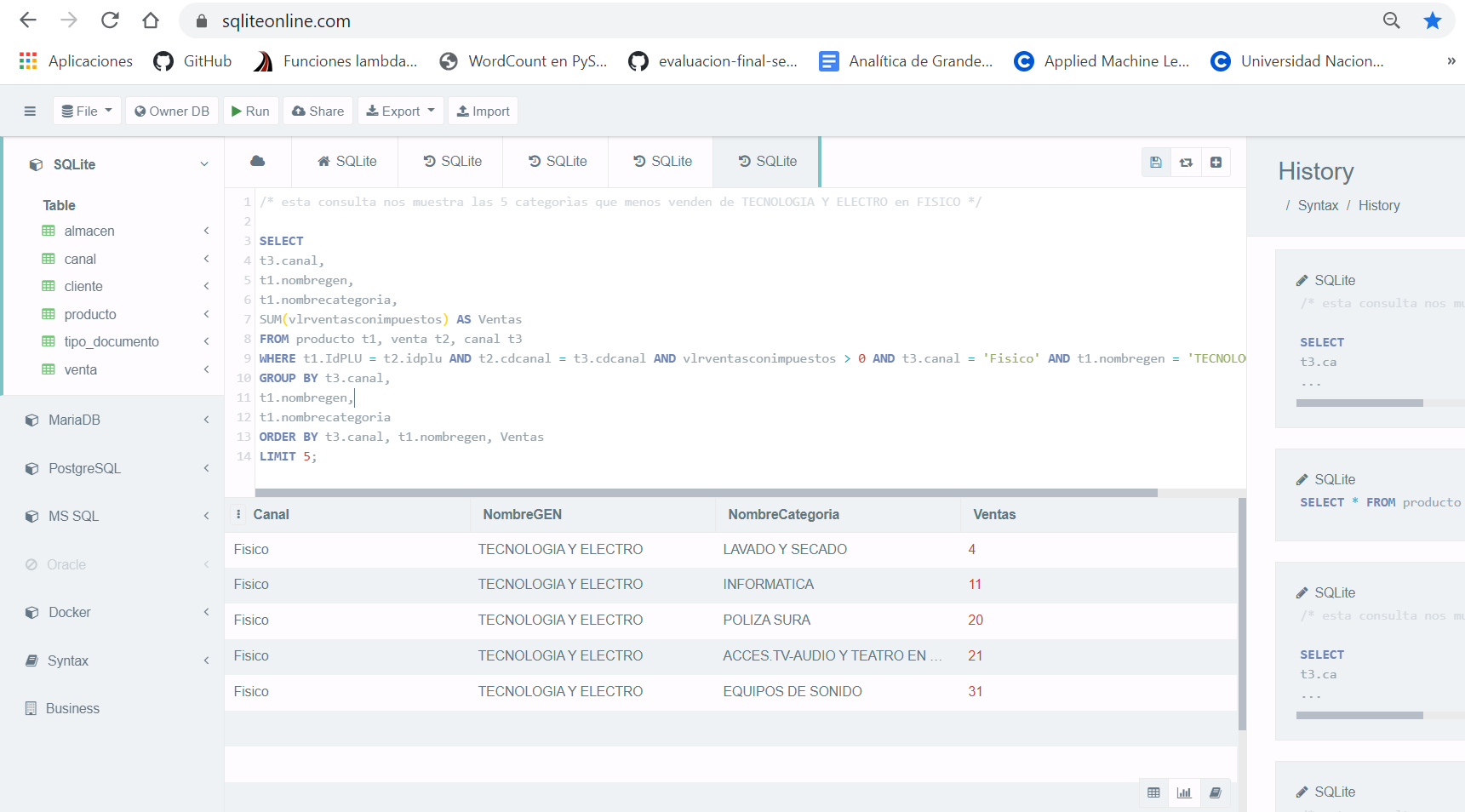


Al revisar la venta de cada gerencia por canal, encontramos cuáles eran las gerencia-canal que menores ventas tuvieron en el 2020, para posteriormente consultar en estas cuales eran las 5 categorías de menores ventas. Estas 5 son las que le proponemos a la gerencia trabajar de cara al 2021 para mejorar nuestras ventas.

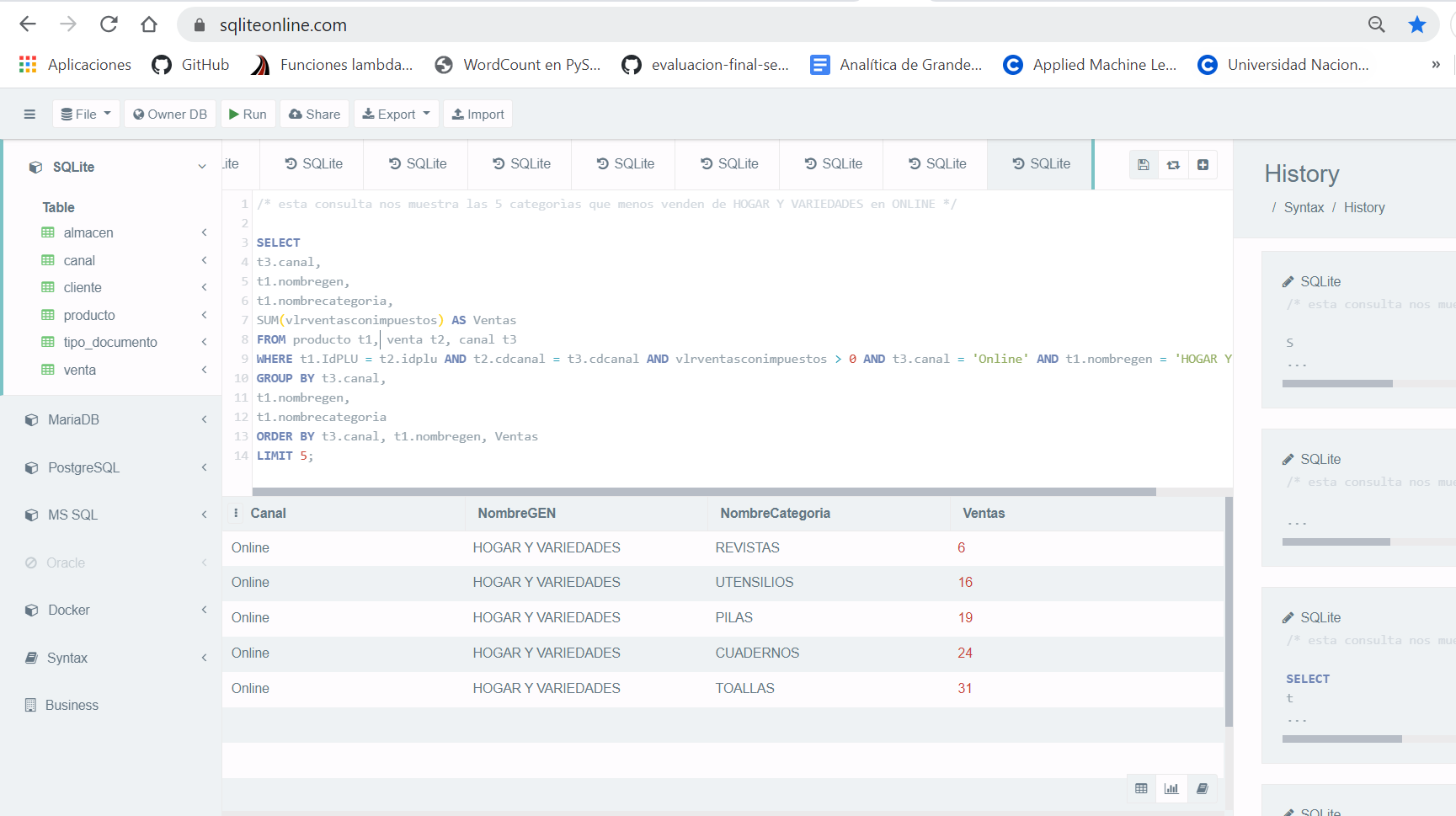


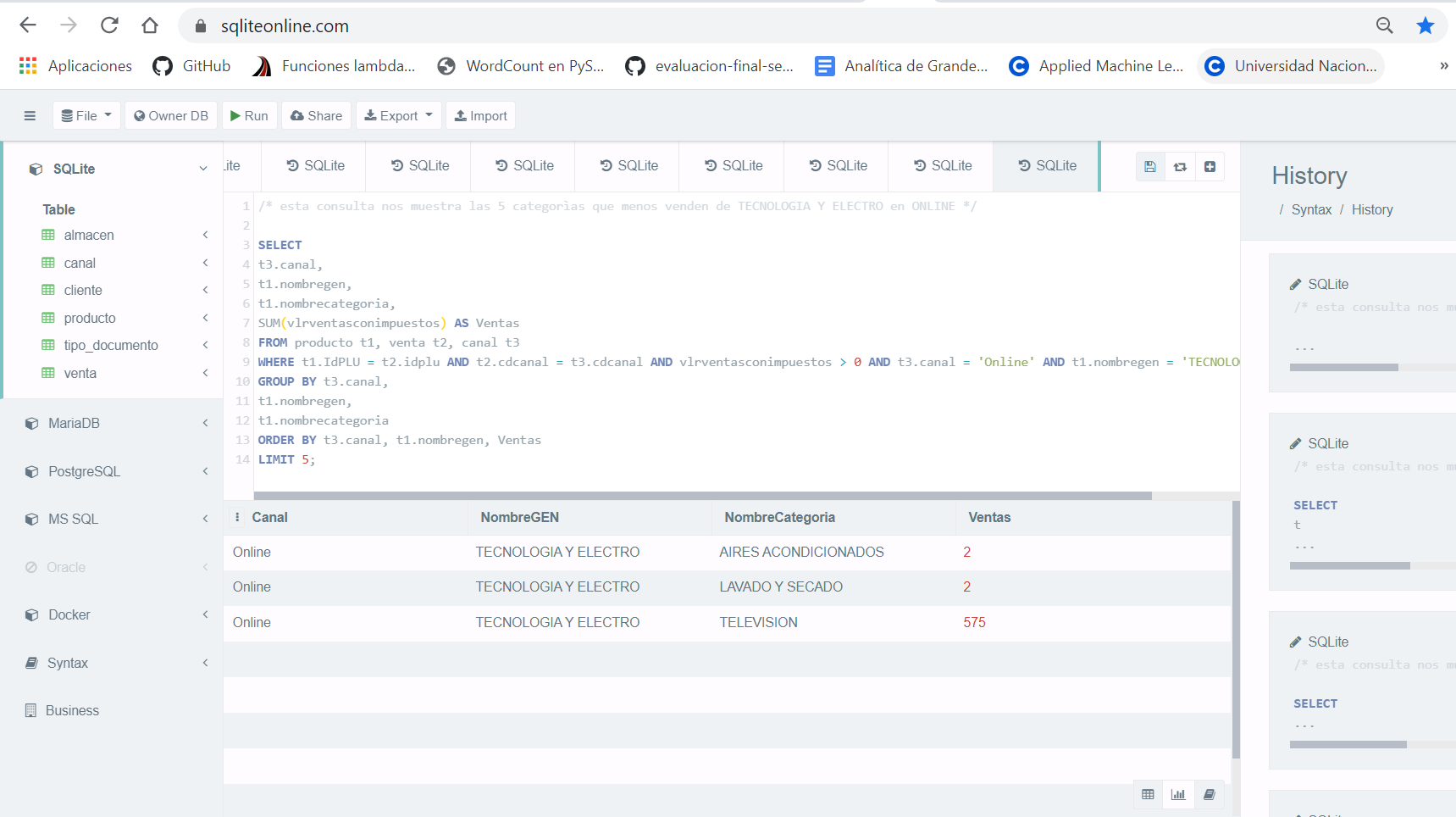
En el canal físico encontramos que debemos trabajar los negocios “Hogar y variedades” y “Tecnología y electro”. Las categorías a trabajar son Álbum, categoría genérica, linternas, velas y linternas, y pinturas, para el negocio de “Hogar y variedades”. En el negocio de “Tecnología y electro” debemos trabajar en lavado y secado, informática, póliza, accesorios tvs y equipos de sonido.



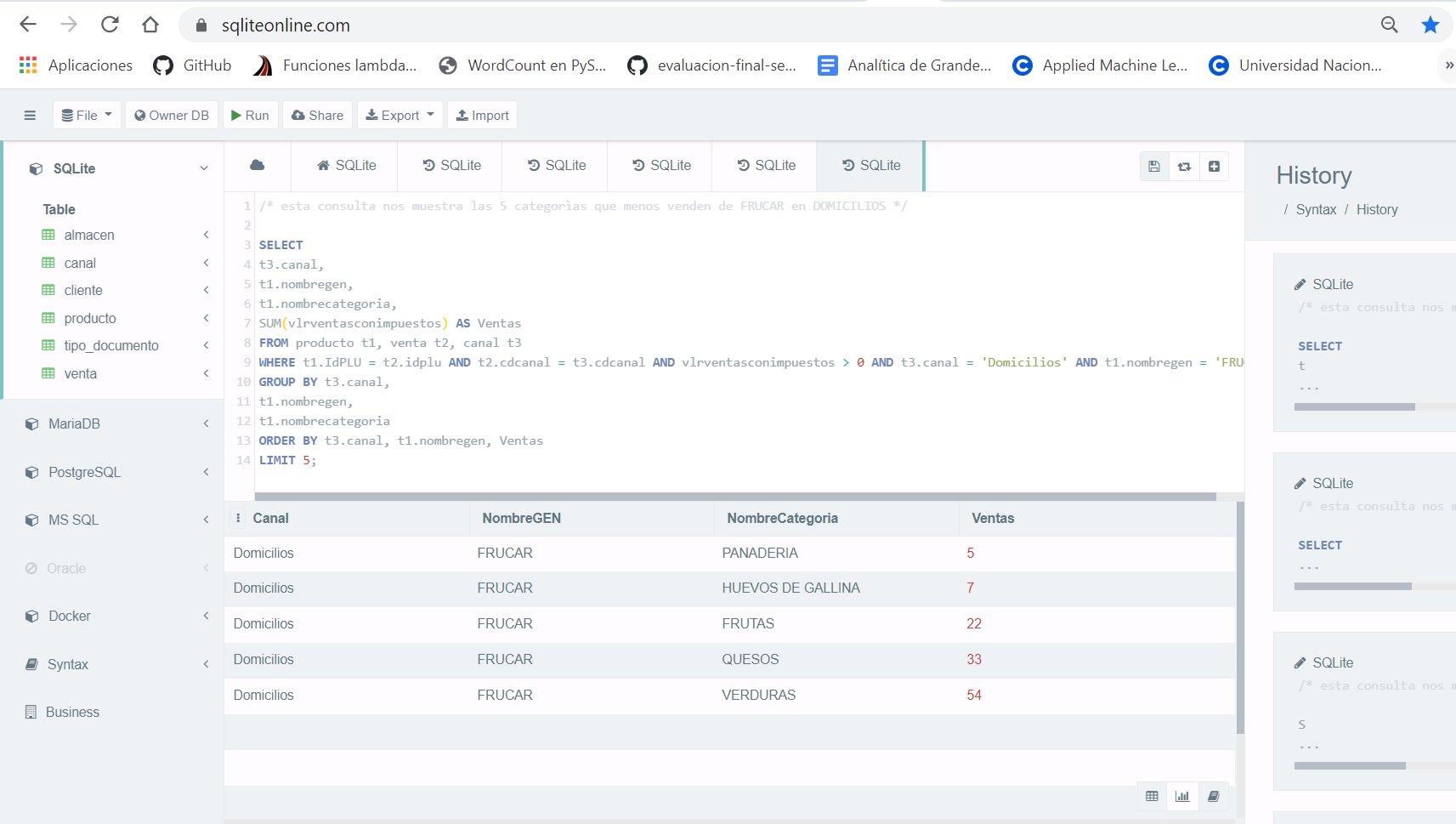


En el canal Online debemos trabajar el negocio “Hogar y variedades” en las categorías de Revistas, utensilios, pilas, cuadernos y toallas. En el negocio de “Tecnología y electro” debemos mejorar las ventas Aires acondicionados, lavado y secado, y televisión.





Finalmente, en el canal de Domicilios debemos trabajar el negocio “FRUCAR” en las categorías de Panadería, huevos de gallina, frutas, quesos y verduras.



En el trabajo realizamos más consultas, pero con estas específicamente logramos darle respuesta a la gerencia para el trabajo de cara al 2021. El resto de las consultas quedaron alojadas en repositorio de GitHub

<https://github.com/spalaciov0819/Analitica-de-Grandes-Datos/blob/main/T1.3.4.Consultar_Datos.sql>

# **4** **MongoDB**

## **4.1** **Sentencia o consulta de creación del documento(s)**

Las sentencias de creación de los documentos en Mongo quedaron alojadas en el repositorio de GitHub.

<https://github.com/spalaciov0819/Analitica-de-Grandes-Datos/blob/main/T1.4.1%20Creacion_Documentos.sql.txt>

## **4.2** **Sentencia de consulta**

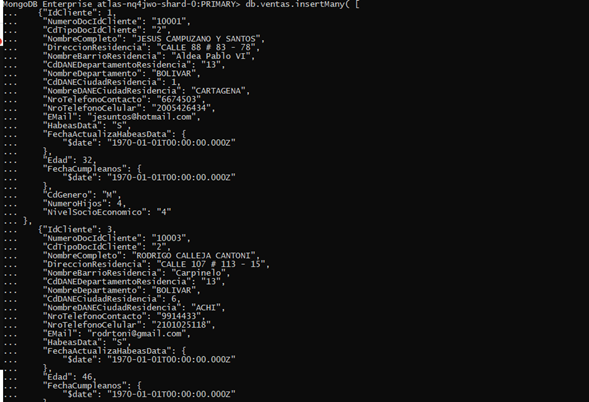
El trabajo de Mongo se realiza creando un database compuesto de 20 documentos en formato JSON, al cual le hacemos ejercicios de conteos, máximos, promedios y mínimos sobre la muestra.

Para comenzar con el ejercicio, se tiene instalado Mongo, se usa la siguiente sentencia para conectar con mongo y se inserta la clave.

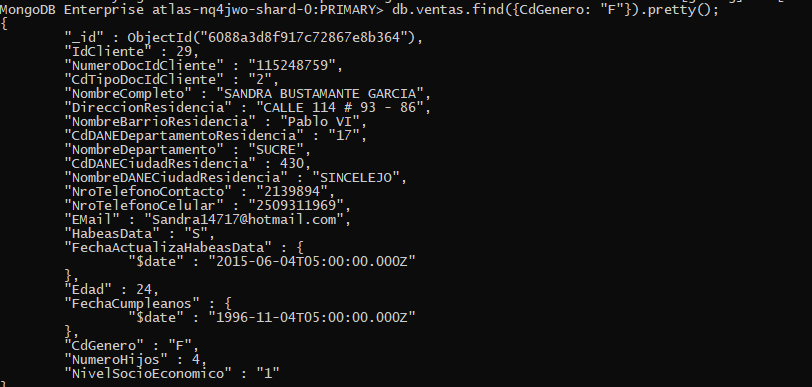
1) Inicio de mongo en consola



2) Utilizando **insertMany** se agregan los documentos al database que llamamos **ventas.**



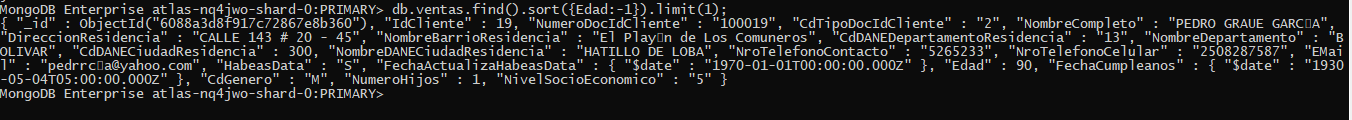
3) Filtramos los clientes de género femenino de una forma ordenada con la siguiente sentencia:

****

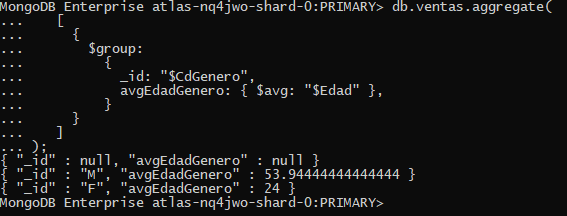
4) Contamos cuántos clientes existen en el database:

;

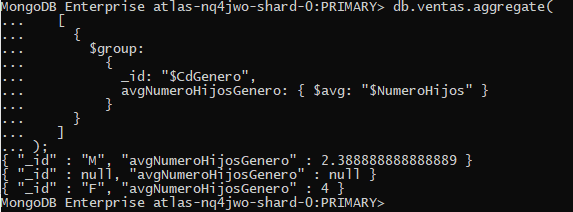
5) Encontramos la persona de mayor edad en el database:

****

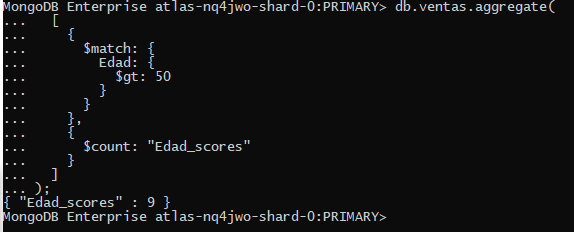
6) Generamos la edad promedio por Genero (Hombres y Mujeres) :

****

**7)** Generamos el número de hijos promedio por Genero (Hombres y Mujeres) :



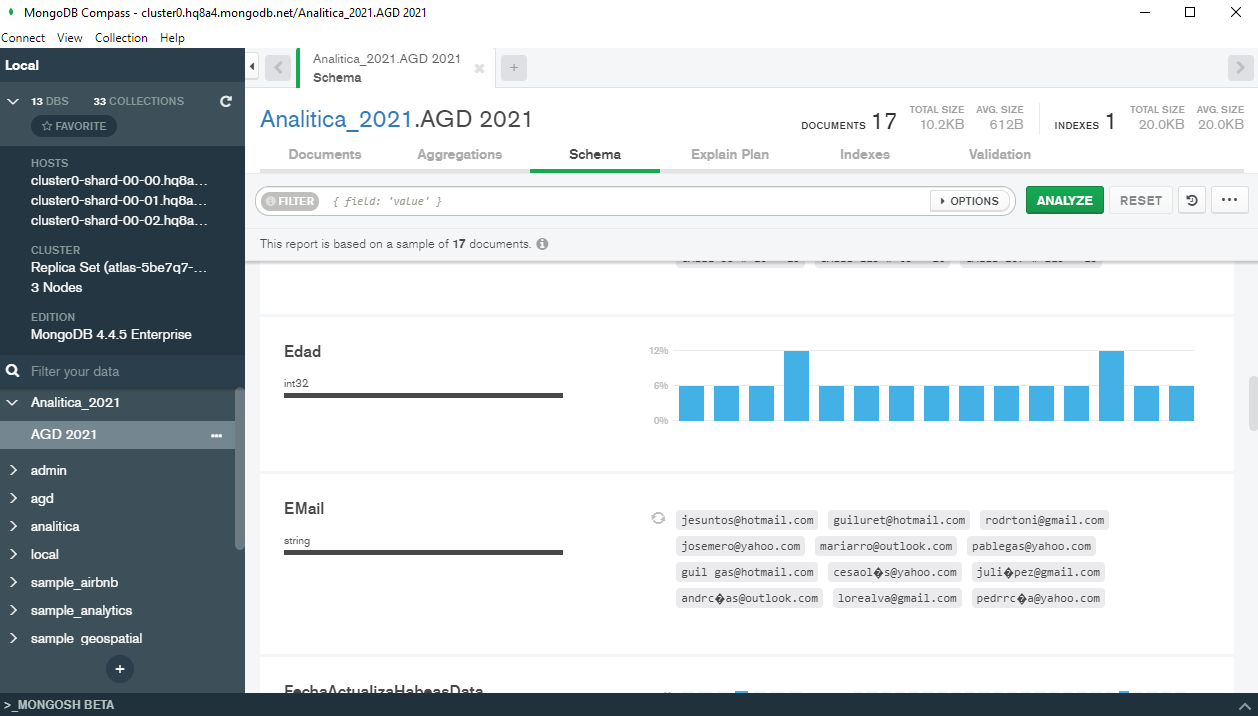
8) Contamos cuántos clientes son mayores de 50 años



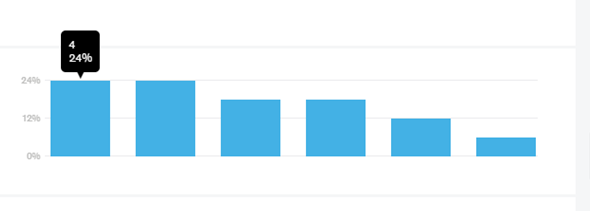
Al realizar las consultas de Mongo evidenciamos:

* Del total de 20 clientes 4 son mujeres y 16 hombres
* 9 clientes son mayores de 50 años.
* El promedio de edad para los hombros es de 54 años y del género femenino es de 24 años.
* El cliente de mayor edad es de 90 años.
* El género femenino tiene en promedio 4 hijos y masculino en promedio 2 hijos.

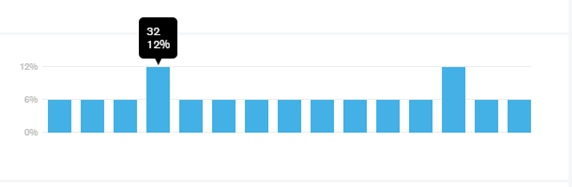
Haciendo uso de la herramienta Mongo DB Compass se sacaron algunos datos adicionales, tales como el porcentaje de clientes por cada una de las edades y por cada uno de los estratos socioeconómicos; donde el estrato 4 y 6 tuvieron un 24% de participación cada uno.



Se evidencia que el 48% de la muestra es estrato 4 ó 6.



Se evidencia que el 24% de la muestra hay personas con 32 ó 84 años .



# **5** **Análisis de lectura**

***Observación****: Considerando el artículo: “*The Definitive Guide to Graph Databases for the RDBMS Developer*” de Neo4J. Compartido en las carpeta de lecturas recomendadas. Analice y responda cada pregunta en máximo 150 palabras:*

1. ¿Cuáles son las limitaciones, que se pueden inferir de la lectura, para migrar los conjuntos de datos relacionales a NoSQL?

El paso de una base de datos SQL a NoSQL exige evaluar qué alternativa se va a tomar en la migración; así mismo se debe mapear a nivel general el negocio para que el equipo involucrado comprenda tanto las limitantes como las ventajas que se obtendrán con el nuevo esquema de datos. A continuación, hablaremos de las diferentes opciones para realizar su migración:

* Migración y switcheo completo: Al momento de realizar el export y el import de los datos se asume el riesgo de tener pérdidas parciales de la información; se puede afectar la disponibilidad del sistema en caso de tener fallas en el piloto.
* Esquema híbrido: Con esta solución se mitiga la pérdida de información de forma espontánea, no se afecta la disponibilidad del sistema en caso de tener fallas en el piloto.
* Duplicar información: Se definen las herramientas a trabajar en SQL y NoSQL; se puede afectar la integridad de la información, pero no se afecta la disponibilidad del sistema en caso de tener fallas en el piloto.

1. ¿Cuáles limitaciones adicionales se deben considerar, a parte de las mencionadas en el artículo?

Se ha identificado que en algunos casos, las bases de datos NoSQL, pueden tener problemas en la ***atomicidad****(*propiedad que asegura que una operación se ha realizado o no*)*y sus patrones de estandarización. La integridad de la información podría verse comprometida, según el tipo de negocio al que se esté atendiendo.

Falta de estandarización, teniendo en cuenta que aunque ya existen muchas bases de datos NoSQL, aún no hay un estándar generalizado como sí ocurre en las BD relacionales.

Problemas de compatibilidad entre instrucciones SQL. Las nuevas bases de datos utilizan sus propias características en el lenguaje de consulta y no son 100% compatibles con el SQL de las bases de datos relacionales. El soporte a problemas con los queries de trabajo en una base de datos NoSQL es más complicado.

Soporte multiplataforma. Aún quedan muchas mejoras en algunos sistemas para que soporten sistemas operativos que no sean Linux.

1. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?

La principal razón para realizar una migración de una base de datos relacional a una base de datos NoSQl se da principalmente cuando modelar y almacenar datos conectados comienza a tener una complejidad extrema, que representa un gran costo de máquina, operación o de sintaxis de consultas SQL anidadas que fácilmente permiten perder el foco de la extracción inicial.

Otros criterios son:

* Si las estructuras de los datos deben cambiar (como sucede hoy en el mundo cambiante que vivimos), ya que SQL es bueno cuando la estructura es fija y estuvo bien definida desde el principio; mientras que NoSQL es ágil para adaptarse a cambios en las estructuras.
* Si el volumen de los datos crece acelerada y exponencialmente.
* Si las consultas están siendo lentas en su procesamiento.
* Si se tiene que conectar constantemente muchas tablas con Joins.

REFERENCIAS:

<https://pandorafms.com/blog/es/nosql-vs-sql-diferencias-y-cuando-elegir-cada-una/>

<https://www.grapheverywhere.com/nosql-de-grafos/>