UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MAESTRÍA EN ANÁLISIS Y VISUALIZACIÓN DE DATOS

Herramienta de Visualización

**Análisis exploratorio con Apache Hive sobre HDFS**

Presenta

**Carmen Sayuri Maldonado Pinto**

**Eduardo García García**

**Juan Carlos de Jesús Garcés Carrillo**

**Kevin David Farinango Cinilin**

Agosto 2022

Tabla de contenido

[Introducción 1](#_Toc110520788)

[Objetivo 1](#_Toc110520789)

[Desarrollo 1](#_Toc110520790)

[Fuentes de Información por utilizar 1](#_Toc110520791)

[Información del dataset 1](#_Toc110520792)

[Pasos 1](#_Toc110520793)

[Resultados 1](#_Toc110520794)

[Conclusiones 4](#_Toc110520795)

[Hoja de control 4](#_Toc110520796)

Introducción

En el presente análisis, se trabajó con la información proporcionada en los csv adjuntos (features, stores, sales).La practica la realizamos principalmente con la herramienta databricks, enfocándonos en utilizar el lenguaje de SQL con Spark.

Objetivo

Realizar un análisis utilizando las herramientas y teoría vistas en la materia. Realizar un proyecto de colaboración y aplicación como científicos de datos.

Desarrollo

Se utiliza el repositorio de Github para compartir el código con los miembros del equipo como se puede ver en Ilustración 1

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración Codigo y archivos en Github

Esto nos permitió trabajar de manera conjunta y ordenada con el fin de que cada integrante pudiera descargar los archivos en su disco local. Los archivos correspondientes a la información por analizar fueron cargados en la herramienta de Databricks y se creo un notebook en el cual se verán reflejados cada uno de los pasos solicitados, así como de los resultados obtenidos.

Fuentes de Información por utilizar

Como se menciono en las secciones anteriores nuestras fuentes de informaciòn son los archivos features, stores y sales.

Información del dataset

El archivo de features se compone de las variables de Store, Date, Temperature, Fuel\_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, IsHoliday. Los cuales

El archivo de sales se compone de las variables de Store, Dept, Date, Weekly\_Sales, IsHoliday.

El archivo de stores se compone de las variables de Store, Type, Size

Como se puede observar la información que se puede extraer de features son características adicionales de la tienda y la información de la región donde se ubica como lo es la temperatura media, el precio de combustible, índice de precio al consumo, tasa de desempleo, así también información sobre la semana en que se registraron dicha información. En la tabla de sales es la información semanal de las ventas por tienda, departamento. Por último esta la tabla de stores que contiene información general de la tiendo como el tipo y el tamaño.

Pasos

Como se menciono anteriormente la mayoría del trabajo se llevo a cabo en Databricks, sin embargo, para la muestra del funcionamiento con HDFS se trabajo directamente en consola. Por lo cual, dicho proceso se ve reflejado en la siguiente Ilustración 2

(AQUÍ VA LA IMAGEN)

Ilustración 2Captura de pantalla de la consola con el ejemplo del funcionamiento de HDFS

Con databricks se hizo un proceso muy similar pero utilizando las herramientas disponibles. Primeramente se cargo la información de los archivos csv a databricks dicho proceso se muestra en la siguiente imagen (PENDIENTE), caba resaltar que estos archivos quedaron almacenados en la ruta “/FileStore/tables”

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración Carga de la información en databricks

Después se creo un cluster y en dicho cluster se genero el Notebook “actividad\_3” el cual se utilizo para poder plasmar nuestros pasos y mostrar la información solicitada en cada uno de los incisos de la actividad.

### Proceso en el Notebook

Estando en el notebook primeramente se decidió crear una función que pudiera realizar la lectura de los archivos csv y después crear las tablas correspondientes con el esquema adecuado, estos pasos se muestran en la Ilustración 4

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración Proceso de lectura de csv y creación de tablas de cada archivo

Mostrar las cinco primeras filas de cada tabla cargada. Para este caso se realizo la selección de 5 registros por tabla con el query

***“Select \* from features limit 5***

***Select \* from sales limit 5***

***Select \* from stores limit 5”***

El detalle de estas tablas se muestra en las siguientes imágenes (Ilustración 5 y Ilustración 6)

Tabla

Descripción generada automáticamente

Ilustración Las cinco primeras filas de la tabla features

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ilustración Las cinco primeras filas de la tabla sales y stores

Contar el número de filas de cada tabla. Indicar este número en el informe. Para este caso se realizo el conteo por tabla con el query

***“SELECT COUNT(\*) as feature FROM feature***

***SELECT COUNT(\*) as sales FROM sales***

***SELECT COUNT(\*) as stores FROM stores”***

El detalle de estas consultas se muestra en la imagen (Ilustración 7) obteniendo como resultado que la tabla features contiene 8,190 filas, la tabla sales tiene 421,570 filas y la tabla stores cuenta con 45 filas.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración Contabilización de filas por tabla

Estudiar el rango (máximo y mínimo) de cada variable numérica. Indicar estos datos en el informe. Para este caso se utilizo nuevamente una query de SQL para obtener el máximo y mínimo de cada variable numérica sin embargo algunas variables contenían datos como NA y se modifico con el fin de que la variable pudiera mendirse correctamente por lo cual la query de sql es la siguiente.

*SELECT MIN(Store) as Min\_Store, MAX(case when Store='NA' then null else cast( Store as double) end ) as Max\_Store FROM feature*

Para este caso solo se mostrara una consulta ya que existen en total 10 variables numéricas entre las 3 tablas. Adicionalmente y con el propósito de realizar la consulta de manera más automática se creo una función para detectar de manera automática cuales eran las variables numéricas y de esa misma manera poder realizar una presentación de todas las variables. Los pasos se pueden ver en las siguientes imágenes.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración Función para detectar las variables numericas.

Aplicación

Descripción generada automáticamente con confianza baja

Ilustración Presentación de las 3 primeras variables numéricas con sus mínimos y máximos

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración Presentación de la cuarta, quinta y sexta variables numéricas con sus mínimos y máximos

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

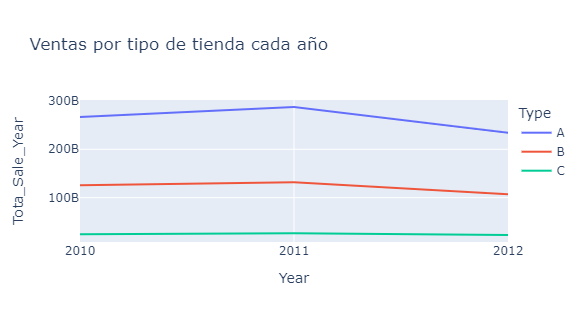
Ilustración Presentación de las ultimas 3 variables numéricas con sus mínimos y máximos

Ejemplo importación hdfs, leer los datos, queries, categorías principales, valores inexistentes, anómalos etc

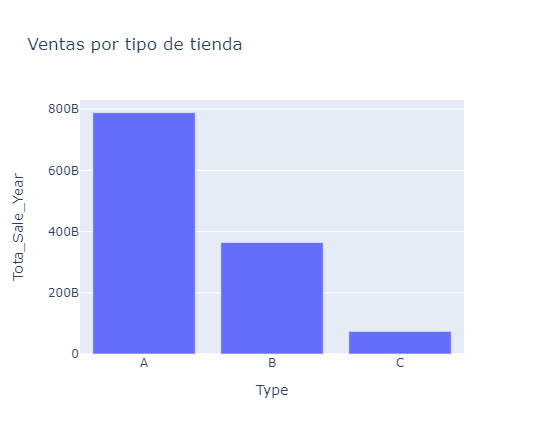
Resultados

**Preguntas de negocio**

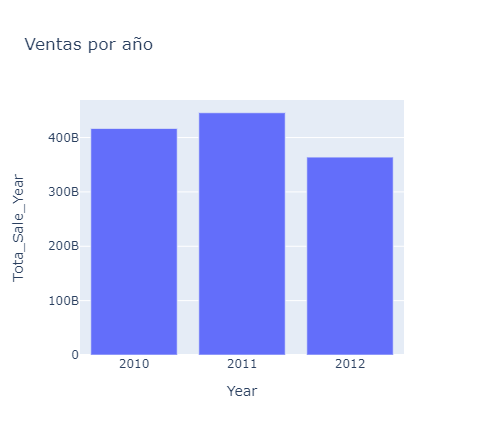
Realizar el análisis de cada imagen



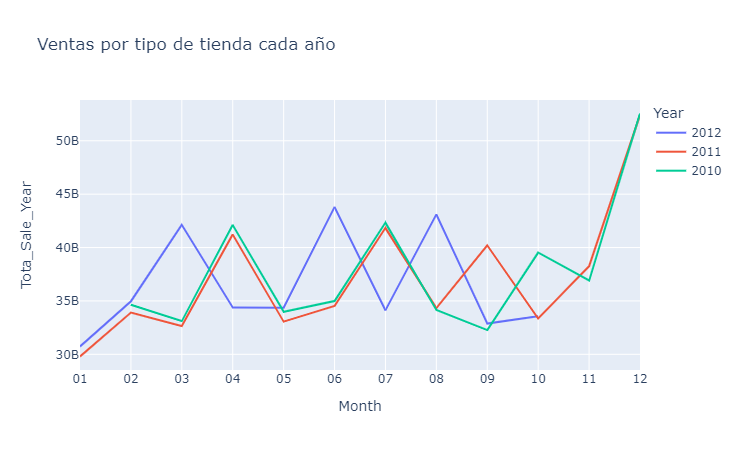
Análisis



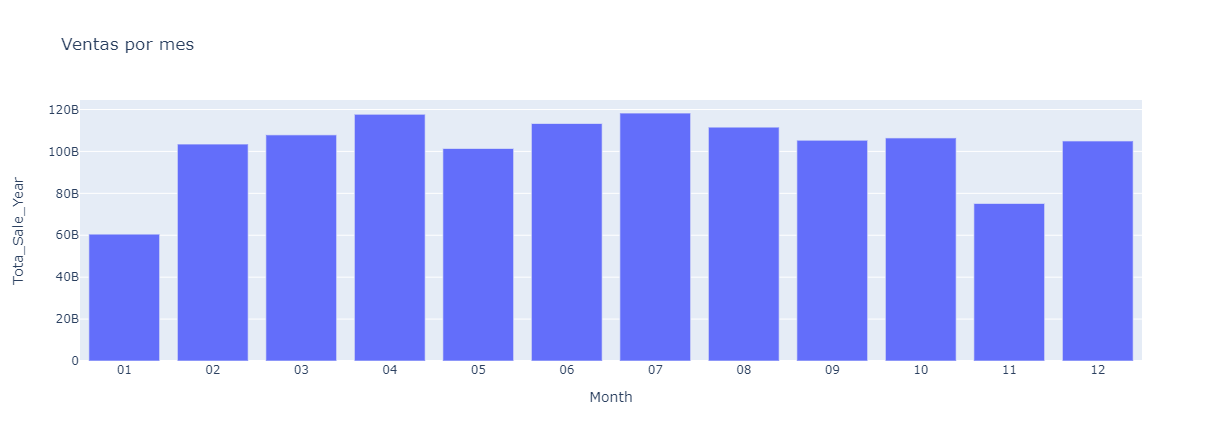
Análisis



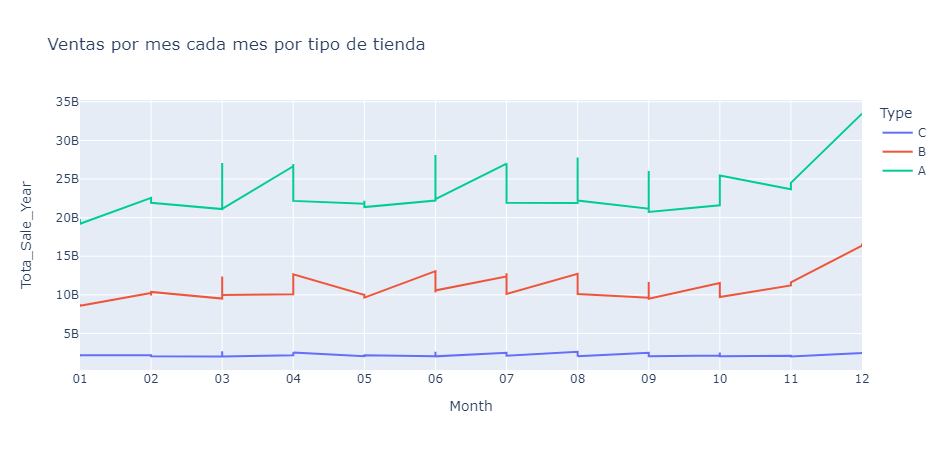
Análisis



Análisis



Análisis



Análisis

Conclusiones

Hoja de control

La siguiente tabla presenta en lo que cada uno de los miembros del equipo ayudaron.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Documentación** | **Desarrollo** | **Investigación** |
| Carmen Sayuri Maldonado Pinto | Si | Si | No |
| Eduardo García García | Sí | Sí | Sí |
| Juan Carlos de Jesús Garcés Carrillo | No | No | No |
| Kevin David Farinango Cinilin | Sí | Sí | Sí |

Se recomienda únicamente colocar la calificación a los estudiantes:

Eduardo García García, Kevin David Farinango Cinilin y Carmen Sayuri Maldonado Pinto quienes fueron los únicos que aportaron con la creación del documento y el proyecto.